THE WASSERSTEIN BELIEVER

LEARNING BELIEF UPDATES FOR PARTIALLY OBSERVABLE ENVIRONMENTS THROUGH RELIABLE LATENT SPACE MODELS

Raphael Avalos[1] * **Florent Delgrange**[1,2] *
Ann Nowé[1] † **Guillermo A. Pérez**[2,3] † **Diederik M. Roijers**[1,4] †
[1] AI Lab, Vrije Universiteit Brussel (Belgium) [2] University of Antwerp (Belgium)
[3] Flanders Make (Belgium) [4] City of Amsterdam (The Netherlands)
{raphael.avalos, florent.delgrange}@vub.be

ABSTRACT

Partially Observable Markov Decision Processes (POMDPs) are used to model environments where the state cannot be perceived, necessitating reasoning based on past observations and actions. However, remembering the full history is generally intractable due to the exponential growth in the history space. Maintaining a probability distribution that models the belief over the current state can be used as a sufficient statistic of the history, but its computation requires access to the model of the environment and is often intractable. While SOTA algorithms use Recurrent Neural Networks to compress the observation-action history aiming to learn a sufficient statistic, they lack guarantees of success and can lead to sub-optimal policies. To overcome this, we propose the Wasserstein Belief Updater, an RL algorithm that learns a latent model of the POMDP and an approximation of the belief update under the assumption that the state is observable during training. Our approach comes with theoretical guarantees on the quality of our approximation ensuring that our latent beliefs allow for learning the optimal value function.

1 INTRODUCTION

Partially Observable Markov Decision Processes (POMDPs) define a powerful framework for modeling decision-making in uncertain environments where the state is not fully observable. These problems are common in many real-world applications, such as robotics (Lauri et al., 2023), and recommendation systems (Wu et al., 2021). In contrast to *Markov Decision Processes* (MDPs), in a POMDP the agent perceives an imperfect observation of the state that does not suffice as conditioning signal for an optimal policy. As such, optimal policies must take the entire interaction history into account. As the space of possible histories scales exponentially in the length of the episode, using histories to condition policies is generally intractable. An alternative is the notion of *belief*, which is defined as a probability distribution over states based on the agent history. Beliefs are a sufficient statistic of the history (Kaelbling et al., 1998) but the computation of their closed-form expression require the access to a model of the environment and is in general intractable.

To overcome those challenges, SOTA algorithms compress the history into a fixed-size vector with the help of *Recurrent Neural Networks* (RNNs) (Hausknecht & Stone, 2015). Nonetheless, this may lead to information loss, resulting in suboptimal policies. To improve the likelihood of obtaining a sufficient statistic, RNNs can be combined with regularization techniques, including generative models (Chen et al., 2022; Hafner et al., 2019; 2021), particle filtering (Igl et al., 2018; Ma et al., 2020), and predicting distant observations (Gregor et al., 2018; 2019). However, *none of these techniques guarantee that the representation of histories induced by RNNs is suitable for optimizing the return.* Additionally, many algorithms assume that beliefs are simple distributions (e.g., Gaussian), which limits their applicability (Gregor et al., 2018; Lee et al., 2020; Hafner et al., 2021).

In this paper, we propose *Wasserstein Belief Updater* (WBU), a model-based reinforcement learning (RL) algorithm for POMDPs that allows learning the belief space over the unobservable states. Specifically, WBU learns an approximation of the belief update rule through a latent space model

* Both authors contributed equally to this research, alphabetic order. † Equal supervision.

whose behaviors (expressed as expected returns) are close to those of the original environment. Furthermore, we show that WBU is guaranteed to induce a suitable representation of the history to optimize the return. WBU consists of three components that are learned in a round-robin fashion: the model, the belief learner, and the policy (Fig. 1). Harnessing only histories to learn a model whose dynamics can be provably linked to the original unobservable environment poses a considerable challenge. Therefore, in the same spirit as the *Centralized Training with Decentralized Execution* paradigm in *multi-agent* RL (MARL) (Oliehoek et al., 2008; Avalos et al., 2022), where leveraging additional information such as the true state of the environment is a common practice, *we assume that the POMDP states can be accessed during training*. While this might seem restrictive at first sight, this assumption is typically met in simulation-based training and can be applied in real-world settings such as robotics, where extra sensors can be used during training in a laboratory setting.

Our core contribution is the *development of a sound framework equipped with theoretical guarantees in the context of RL within partial observability*. While SOTA algorithms primarily concentrate on enhancing the overall return — potentially resulting in substantial performance gains — we contend that performance is not the exclusive goal and that possessing guarantees is equally important, as the balance between these two aspects varies based on the specific application. By tackling POMDPs with a formal approach, we offer theoretical guarantees that other methods cannot provide: we ensure that *our latent model is able to replicate the dynamics of the original, partially observable environment*, which further *yields a belief representation suitable for learning the value function*.

We learn the latent model of the POMDP via a *Wasserstein auto-encoded MDP* (WAE-MDP, Delgrange et al. 2022), which embeds bisimulation metrics — intuitively leading to our guarantees. In parallel, we maintain a belief distribution over the latent state space via a *belief update network*: we minimize its *Wasserstein distance* to the exact belief update rule, through a tractable variational proxy. To allow for complex belief distributions, we use *normalizing flows* (Kobyzev et al., 2021). In contrast to SOTA algorithms, the beliefs are only optimized towards accurately replicating its update rule. While we call recursively the belief network to maintain the belief distribution, we do not back-propagate through time and thus implement it as a simple feed forward network. The policy is learned on the latent belief space using a vector integrating the parameters of the belief distribution. Our experimental results are promising and show the ability of our algorithm to learn to encode the history into a representation useful to learn a policy, *without using RNNs*.

Other related work. DVRL (Igl et al., 2018) extends A2C (Mnih et al., 2016) combined with RNNs (R-A2C) with auxiliary losses aiming to learn beliefs via a variational autoencoder and particle filtering, but it lacks guarantees. DVRL further assumes independent normal distributions for beliefs, limiting its applicability. FORBES (Chen et al., 2022) use normalizing flows but learn policies conditioned on latent states, which is suboptimal as the state distribution is approximated with a single sample. Some works focus on specific POMDP types, like compact image representations (e.g., visual motor tasks, e.g., Lee et al. 2020) or states masked with Gaussian noise (Wang & Tan, 2021). While accessing states is common in MARL, it is less common in single-agent but has been explored in kernel-POMDPs (Nishiyama et al., 2012), and more recently for sample efficient learning (Lee et al., 2023). Leveraging additional information available during training (not necessarily the states) has also been explored by Lambrechts et al. (2023), but RNNs remain crucial while no abstraction nor representation quality guarantee are provided. Finally, other works (Gelada et al., 2019; Delgrange et al., 2022) study similar value difference bounds to ours and connect them to bisimulation theory (Larsen & Skou, 1989; Givan et al., 2003), but in fully observable environments.

2 Background

2.1 Probability Distributions and Discrepancy Measures

We write $\Sigma(\mathcal{X})$ for the set of all *measurable* subsets of a complete, separable space $\mathcal{X}$, $\Delta(\mathcal{X})$ for the set of measures on $\mathcal{X}$, and $\delta_a \in \Delta(\mathcal{X})$ for the *Dirac measure* with impulse $a \in \mathcal{X}$. Let $P, Q \in \Delta(\mathcal{X})$ be measures with densities p, q, the divergence between P and Q can be measured via:

- the solution of the *optimal transport* problem (OT), defined as $\mathcal{W}_c(P, Q) = \inf_\lambda \mathbb{E}_{x,y\sim\lambda}\, c(x, y)$, which is the *minimum cost of changing P into Q* (Villani, 2009), where $c\colon \mathcal{X} \times \mathcal{X} \to [0, \infty)$ is a cost function and the infimum is taken over the set of all *couplings* of P and Q. When c is equal to a distance metric d over $\mathcal{X}$, $\mathcal{W}_d$ is the *Wasserstein distance* between the two distributions.

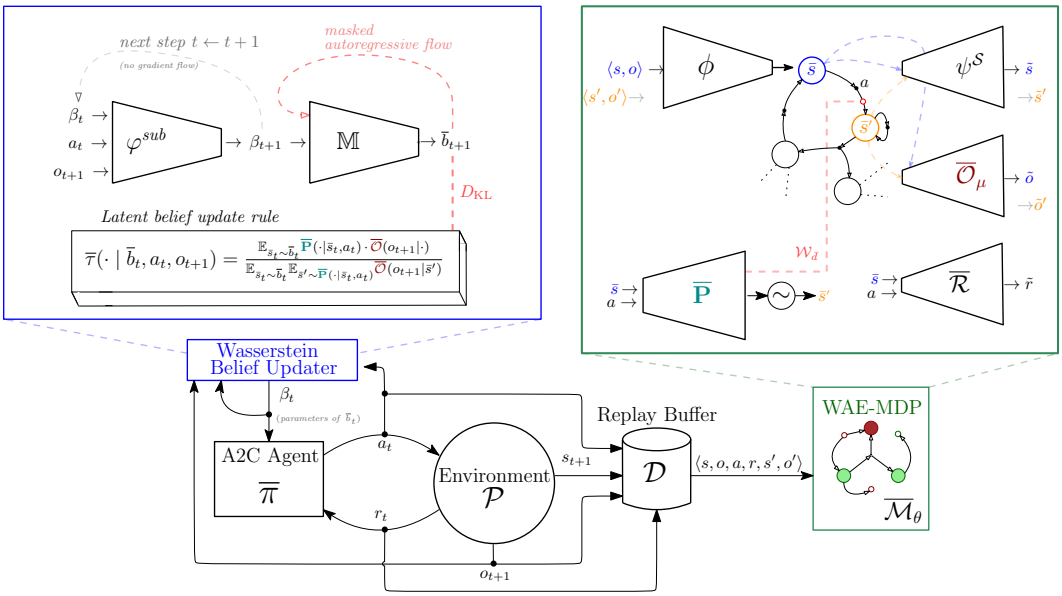


Figure 1: *WBU framework*. The WAE-MDP is presented in Sect. 3, and WBU in Sect. 4. Learning the different components is done in a round-robin fashion. The WAE-MDP learns from data collected by the RL agent and stored in a Replay Buffer. WBU uses the transition function $\bar{\mathbf{P}}$ and observation decoder $\bar{\mathcal{O}}$ of the WAE-MDP to learn to approximate the belief update rule. The agent learns a policy conditioned on the resulting *sub-belief* β_t (i.e., the parameters of the latent belief $\bar{b}_t$).

- the *total variation distance* (TV), defined as $d_{TV}(P, Q) = \sup_{A \in \Sigma(\mathcal{X})} |P(A) - Q(A)|$. If $\mathcal{X}$ is equipped with the discrete metric $\mathbf{1}_{\neq}$, TV coincides with the Wasserstein metric.
- the *Kullback-Leibler* (KL) divergence, defined as $D_{\mathrm{KL}}(P, Q) = \mathbb{E}_{x \sim P}\left[\log(p(x)/q(x))\right]$.

2.2 Decision Making under Uncertainty

Markov Decision Processes (MDPs) are tuples $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathbf{P}, \mathcal{R}, s_I, \gamma \rangle$ where $\mathcal{S}$ is a set of *states*; $\mathcal{A}$, a set of *actions*; $\mathbf{P} \colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, a *probability transition function* that maps the current state and action to a *distribution* over the next states; $\mathcal{R} \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, a *reward function*; $s_I \in \mathcal{S}$, the *initial state*; and $\gamma \in [0, 1)$ a discount factor. We refer to MDPs with continuous state or action spaces as *continuous MDPs*. In that case, we assume $\mathcal{S}$ and $\mathcal{A}$ are complete separable metric spaces equipped with a Borel σ-algebra. An agent interacting in $\mathcal{M}$ produces *trajectories*, i.e., sequences of states and actions $\langle s_{0:T}, a_{0:T-1} \rangle$ where $s_0 = s_I$ and $s_{t+1} \sim \mathbf{P}(\cdot \mid s_t, a_t)$ for $t < T$.

Policies and probability measure. A *stationary policy* $\pi \colon \mathcal{S} \to \Delta(\mathcal{A})$ prescribes which action to choose at each step of the interaction. A policy π and $\mathcal{M}$ induce a unique probability measure $\mathbb{P}_\pi^{\mathcal{M}}$ on the Borel σ-algebra over (measurable) infinite trajectories (Puterman, 1994). The typical goal of an RL agent is to learn a policy that maximizes the *expected return*, given by $\mathbb{E}_\pi^{\mathcal{M}}\left[\sum_{t=0}^{\infty} \gamma^t \cdot \mathcal{R}(s_t, a_t)\right]$, by interacting with $\mathcal{M}$. We omit the superscript when the context is clear.

Partially Observable MDPs (POMDPs) are tuples $\mathcal{P} = \langle \mathcal{M}, \Omega, \mathcal{O} \rangle$ where $\mathcal{M}$ is an MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$; Ω is a set of *observations*; and $\mathcal{O} \colon \mathcal{S} \times \mathcal{A} \to \Delta(\Omega)$ is an *observation function* that defines the distribution of observations that may occur when the MDP $\mathcal{M}$ transitions to a state upon the execution of a particular action. An agent in $\mathcal{P}$ actually interacts in $\mathcal{M}$, but *without directly observing the states* of $\mathcal{M}$: instead, the agent perceives observations, which yields *histories*, i.e., sequences of actions and observations $\langle a_{0:T-1}, o_{1:T} \rangle$ that can be associated to an (unobservable) trajectory $\langle s_{0:T}, a_{0:T-1} \rangle$ in $\mathcal{M}$, where $o_{t+1} \sim \mathcal{O}(\cdot \mid s_{t+1}, a_t)$ for all $t < T$.

Beliefs. Unlike in MDPs, where it is sufficient to take decisions based on the current state (Puterman, 1994), policies solely based on the current observation of $\mathcal{P}$ are not sufficient to optimize the return. Intuitively, due to the partial observability, the agent must take into account full histories in order

to make an informed decision on its next action. Alternatively, the agent can maintain a *belief* $b_t \in \Delta(\mathcal{S}) = \mathcal{B}$ over the current state of $\mathcal{M}$ (Åström, 1965). Given the next observation o_{t+1}, the next belief b_{t+1} is computed according to the *belief update function* $\tau \colon \mathcal{B} \times \mathcal{A} \times \Omega \to \mathcal{B}$, where $\tau(b_t, a_t, o_{t+1}) = b_{t+1}$ iff the belief over any next state $s_{t+1} \in \mathcal{S}$ has for density

$$b_{t+1}(s_{t+1}) = \frac{\mathbb{E}_{s_t \sim b_t} \mathbf{P}(s_{t+1} \mid s_t, a_t) \cdot \mathcal{O}(o_{t+1} \mid s_{t+1}, a_t)}{\mathbb{E}_{s_t \sim b_t} \mathbb{E}_{s' \sim \mathbf{P}(\cdot \mid s_t, a_t)} \mathcal{O}(o_{t+1} \mid s', a_t)}. \tag{1}$$

Each belief b_{t+1} constructed this way is a *sufficient statistic* for the history $\langle a_{0:t}, o_{1:t+1} \rangle$ to optimize the return (Kaelbling et al., 1998). We write $\tau^*(a_{0:t}, o_{1:t+1}) = \tau(\cdot\,, a_t, o_{t+1}) \circ \cdots \circ \tau(\delta_{s_I}, a_0, o_1) = b_{t+1}$ for the recursive application of τ along the history. The belief update rule derived from τ allows to formulate $\mathcal{P}$ as a continuous *belief MDP* $\mathcal{M}_{\mathcal{B}} = \langle \mathcal{B}, \mathcal{A}, \mathbf{P}_{\mathcal{B}}, \mathcal{R}_{\mathcal{B}}, b_I, \gamma \rangle$, where $\mathbf{P}_{\mathcal{B}}(b' \mid b, a) = \mathbb{E}_{s \sim b} \mathbb{E}_{s' \sim \mathbf{P}(\cdot \mid s, a)} \mathbb{E}_{o' \sim \mathcal{O}(\cdot \mid s', a)} \delta_{\tau(b, a, o')}(b')$; $\mathcal{R}_{\mathcal{B}}(b, a) = \mathbb{E}_{s \sim b} \mathcal{R}(s, a)$; and $b_I = \delta_{s_I}$. As for all MDPs, $\mathcal{M}_{\mathcal{B}}$ and any stationary policy for $\mathcal{M}_{\mathcal{B}}$ (thus conditioned on beliefs) induce a well-defined probability space over trajectories of $\mathcal{M}_{\mathcal{B}}$, which allows optimizing the expected return in $\mathcal{P}$.

2.3 Latent Space Modeling

Latent MDPs. Given the original (continuous or very large, possibly unknown) environment $\mathcal{M}$, a *latent space model* is another (tractable, explicit) MDP $\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \mathcal{A}, \overline{\mathbf{P}}, \overline{\mathcal{R}}, \bar{s}_I, \gamma \rangle$ with state space linked to the original one via a *state embedding function*: $\phi \colon \mathcal{S} \to \overline{\mathcal{S}}$.

Wasserstein Auto-encoded MDPs (WAE-MDPs, Delgrange et al. 2023) are latent space models that are trained based on the OT from trajectories resulting from the execution of the RL agent policy in the real environment $\mathcal{M}$, to that reconstructed from the latent model $\overline{\mathcal{M}}$. The optimization process relies on a temperature $\lambda \in (0, 1)$ that controls the continuity of the latent space learned, the zero-temperature limit corresponding to a discrete latent state space (see Appendix E for a discussion). This procedure guarantees $\overline{\mathcal{M}}$ to be probably approximately *bisimilarly close* (Larsen & Skou, 1989; Givan et al., 2003; Delgrange et al., 2022) to $\mathcal{M}$ as $\lambda \to 0$: in a nutshell, *bisimulation metrics* imply the closeness of the two models in terms of probability measures and expected return (Desharnais et al., 2004; Ferns et al., 2011). Specifically, a WAE-MDP learns the following components:

a *state embedding function*	$\phi \colon \mathcal{S} \to \overline{\mathcal{S}}$	a *latent transition function*	$\overline{\mathbf{P}} \colon \overline{\mathcal{S}} \times \mathcal{A} \to \Delta(\overline{\mathcal{S}})$
a *latent reward function*	$\overline{\mathcal{R}} \colon \overline{\mathcal{S}} \times \mathcal{A} \to \mathbb{R}$	a *state decoder*	$\psi \colon \overline{\mathcal{S}} \to \mathcal{S}$.

(2)

3 Learning the Dynamics

The agent is assumed to operate within a POMDP. In an RL setting, the former have no explicit access to the environment dynamics: instead, it reinforces its behaviors through interactions and experiences without directly accessing the transition, reward, and observation functions of the environment. To provide the aforementioned guarantees, we henceforth adhere the following assumption.

Assumption 1 (Access to the state during training)**.** *In addition to the observation, the agent is able to observe the true state of the environment, but only during the training phase.*

Remark 1. Seemingly restrictive, Assumption 1 can actually be met in a broad range of training scenarios, in particular those relying on simulators, where one could merely consider the RAM as the state. Otherwise, additional sensors with higher fidelity could be considered to obtain the state. Other applicable scenarios are model-based design or model-predictive control, where a model is accessible during training, and situations where accessing the state is costly (Bulychev et al., 2012).

Concretely, when the RL agent interacts in a POMDP $\mathcal{P} = \langle \mathcal{M}, \Omega, \mathcal{O} \rangle$ with underlying MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathbf{P}, \mathcal{R}, s_I, \gamma \rangle$, *we leverage this access to allow the agent to learn the dynamics of the environment*, i.e., those of $\mathcal{M}$, as well as those related to the observation function $\mathcal{O}$. To do so, we learn an internal, explicit representation of the experiences gathered, through a latent space model. *We then use this model as a teacher for the agent to make it learn how to perform its belief updates*. Hence, acquiring an accurate environment model is crucial to learn a reliable belief update function. In Sect. 3.2, we further demonstrate that the resulting model is guaranteed to closely replicate the original environment behaviors. The trick we use to learn such a model is to reason on an equivalent POMDP, where the underlying MDP is refined to encode all the crucial dynamics.

3.1 THE LATENT POMDP ENCODING

We enable learning the dynamics of $\mathcal{P}$ via a WAE-MDP by considering the POMDP $\mathcal{P}^{\uparrow} = \langle \mathcal{M}_{\Omega}, \Omega, \mathcal{O}^{\uparrow} \rangle$, where (i) the state space of the underlying MDP is refined to encode the observations: $\mathcal{M}_{\Omega} = \langle \mathcal{S}_{\Omega}, \mathcal{A}, \mathbf{P}_{\Omega}, \mathcal{R}_{\Omega}, \langle s_I, o_I \rangle, \gamma \rangle$ with $\mathcal{S}_{\Omega} = \mathcal{S} \times \Omega$, $\mathbf{P}_{\Omega}(s', o' \mid s, o, a) = \mathbf{P}(s' \mid s, a) \cdot \mathcal{O}(o' \mid s', a)$, $\mathcal{R}_{\Omega}(\langle s, o \rangle, a) = \mathcal{R}(s, a)$, and o_I is an observation from Ω linked to the initial state s_I; (ii) the observation function $\mathcal{O}^{\uparrow} \colon \mathcal{S}_{\Omega} \to \Omega$ is the *deterministic* projection of the refined state on the observation space, with $\mathcal{O}^{\uparrow}(\langle s, o \rangle) = o$. The POMDPs $\mathcal{P}$ and $\mathcal{P}^{\uparrow}$ are equivalent (Chatterjee et al., 2016): $\mathcal{P}^{\uparrow}$ captures the stochasticity of $\mathcal{O}$ in its transition function through the refinement of the state space, further *yielding a deterministic observation function*, only dependent on refined states.

Henceforth, the goal is to learn a latent space model $\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \mathcal{A}, \overline{\mathbf{P}}, \overline{\mathcal{R}}, \bar{s}_I, \gamma \rangle$ linked to $\mathcal{M}_{\Omega}$ via the embedding $\phi \colon \mathcal{S}_{\Omega} \to \overline{\mathcal{S}}$, and we achieve this via the WAE-MDP framework. Not only does the latter allow learning the observation dynamics through $\overline{\mathbf{P}}$, but it also enables to learn the deterministic observation function $\mathcal{O}^{\uparrow}$ through the use of the state decoder ψ, by decomposing the latter in two networks $\psi^{\mathcal{S}} \colon \overline{\mathcal{S}} \to \mathcal{S}$ and $\overline{\mathcal{O}}_{\mu} \colon \overline{\mathcal{S}} \to \Omega$, which yield $\psi(\bar{s}) = \langle \psi^{\mathcal{S}}(\bar{s}), \overline{\mathcal{O}}_{\mu}(\bar{s}) \rangle$. This way, the WAE-MDP learns all the components of $\mathcal{P}^{\uparrow}$, the latter being equivalent to $\mathcal{P}$. With this model, we construct a *latent POMDP* $\overline{\mathcal{P}} = \langle \overline{\mathcal{M}}, \Omega, \overline{\mathcal{O}} \rangle$, where $\overline{\mathcal{O}}$ is a Dirac measure with impulse $\overline{\mathcal{O}}_{\mu}$. Figure 6 in Appendix F presents a visualization of the relationship between the different models.

As with any POMDP, the belief update function $\bar{\tau}$ of $\overline{\mathcal{P}}$ allows to reason on the belief space to optimize the return. The belief update procedure is illustrated in Appendix B. Formally, assuming the latent belief at time step $t \geqslant 0$ is $\bar{b}_t \in \Delta(\overline{\mathcal{S}}) = \overline{\mathcal{B}}$, a_t is executed, and then o_{t+1} observed, $\bar{b}_t$ is updated according to $\bar{\tau}(\bar{b}_t, a_t, o_{t+1}) = \bar{b}_{t+1}$ iff, for any (unobservable) next state $\bar{s}_{t+1} \in \overline{\mathcal{S}}$,

$$\bar{b}_{t+1}(\bar{s}_{t+1}) = \frac{\mathbb{E}_{\bar{s}_t \sim \bar{b}_t} \overline{\mathbf{P}}(\bar{s}_{t+1} \mid \bar{s}_t, a_t) \cdot \overline{\mathcal{O}}(o_{t+1} \mid \bar{s}_{t+1})}{\mathbb{E}_{\bar{s}_t \sim \bar{b}_t} \mathbb{E}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot \mid \bar{s}_t, \bar{a}_t)} \overline{\mathcal{O}}(o_{t+1} \mid \bar{s}')}. \tag{3}$$

Latent policies. Given *any* history $h \in (\mathcal{A} \cdot \Omega)^*$, running a latent policy $\bar{\pi} \colon \overline{\mathcal{B}} \to \Delta(\mathcal{A})$ in $\mathcal{P}$ is possible by converting h into a latent belief $\bar{\tau}^*(h) = \bar{b}$ and executing the action prescribed by $\bar{\pi}(\cdot \mid \bar{b})$. Training $\overline{\mathcal{M}}$ grants access to the dynamics required to update the belief through its closed form (Eq. 3). However, integrating over the full latent space remains computationally intractable.

> As a solution, we propose to leverage the access to the dynamics of $\overline{\mathcal{M}}$ to learn a *latent belief encoder* $\varphi \colon \overline{\mathcal{B}} \times \mathcal{A} \times \overline{\mathcal{S}} \to \overline{\mathcal{B}}$ that approximates the belief update function by minimizing $D(\bar{\tau}^*(h), \varphi^*(h))$ for *some* discrepancy D and $h \in (\mathcal{A} \cdot \Omega)^*$ drawn from *some* distribution. The belief encoder φ thus enables to learn a policy $\bar{\pi}$ conditioned on latent beliefs to optimize the return in $\mathcal{P}$: *given the current history h, the next action to play is given by $a \sim \bar{\pi}(\cdot \mid \varphi^*(h))$.*

Two key questions arise: (i) does the WAE-MDP encoding induces a latent POMDP with behaviors close to $\mathcal{P}$? (ii) is the history representation induced by φ suitable for optimizing the expected return in $\mathcal{P}$? Our guarantees hinge on the history distribution and chosen discrepancy. The next section details our theoretical analysis of the required distributions and losses to answer (i) and (ii).

3.2 LOSSES AND THEORETICAL GUARANTEES

To yield the guarantees, we specifically target the *episodic RL process* setting for drawing histories.

Assumption 2 (Episodic RL process)**.** *The environment $\mathcal{P}$ embeds a special* reset state *so that (i) under any policy, the environment is almost surely eventually reset; (ii) when reset, the environment transitions to the initial state; and (iii) the reset state is observable.*

Lemma 3.1. *There exists a well defined stationary distribution $\mathcal{H}_{\bar{\pi}} \in \Delta\big((\mathcal{A} \cdot \Omega)^*\big)$ over histories likely to be seen at the limit of the interaction when $\bar{\pi}$ is executed in $\mathcal{P}$* (proof in Appendix D).

Remark 2. Assumption 2 is present in a vast majority of RL scenarios, where it is common practice to reset the environment and start anew from an initial state when the agent succeeds, fails, or after a finite number of time steps (Brockman et al., 2016; Pardo et al., 2018). A notable exception is in the domain of continual RL. However, the existence of a stationary distribution (Lem. 3.1) is often assumed in such scenarios (see, e.g., Huang 2020), which allows to relax the episodic assumption.

Local losses. The objective function of the WAE-MDP incorporates *local losses* (Gelada et al., 2019) that minimize the expected distance between the original and latent reward and transition functions:

$$L_{\mathcal{R}} = \mathop{\mathbb{E}}_{s,o,a\sim\mathcal{H}_{\overline{\pi}}} \left|\mathcal{R}(s,a) - \overline{\mathcal{R}}(\phi(s,o),a)\right|, \quad L_{\mathbf{P}} = \mathop{\mathbb{E}}_{s,o,a\sim\mathcal{H}_{\overline{\pi}}} \mathcal{W}_{\bar{d}}\left(\phi\mathbf{P}_{\Omega}(\cdot \mid s,o,a), \overline{\mathbf{P}}(\cdot \mid \phi(s,o),a)\right);$$

and both are optimized *locally*, i.e., under $\mathcal{H}_{\overline{\pi}}$, where $s,o,a \sim \mathcal{H}_{\overline{\pi}}$ is a shorthand for (i) $h \sim \mathcal{H}_{\overline{\pi}}$ so that o is the last observation of h, (ii) $s \sim \tau^*(h)$, and (iii) $a \sim \overline{\pi}(\cdot \mid \varphi^*(h))$. Furthermore, $\phi\mathbf{P}(\cdot \mid s,a)$ is the distribution of transitioning to $s' \sim \mathbf{P}(\cdot \mid s,a)$, then embedding it to the latent space $\bar{s}' = \phi(s')$, and $\bar{d}$ is a metric on $\overline{\mathcal{S}}$. In practice, the ability of observing states during learning enables the optimization of those local losses without the need of explicitly storing histories. Instead, we simply store the transitions of $\mathcal{M}_{\Omega}$ encountered while executing $\overline{\pi}$. We also introduce an *observation loss* in addition to the reconstruction loss of the decoder, which allows learning $\overline{\mathcal{O}}$:

$$L_{\mathcal{O}} = \mathop{\mathbb{E}}_{s,o,a\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} d_{TV}\left(\mathcal{O}\big(\cdot \mid s',a\big), \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)} \overline{\mathcal{O}}\big(\cdot \mid \phi\big(s',o'\big)\big)\right). \tag{4}$$

Belief Losses. We set D as the Wassertein distance between the true latent belief update and our belief encoder. In addition, we argue that the following reward and transition regularizers are required to bound the gap between the fully observable model $\overline{\mathcal{M}}$ and the partially observable one $\overline{\mathcal{P}}$:

$$L_{\overline{\tau}} = \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathcal{W}_{\bar{d}}\left(\overline{\tau}^*(h), \varphi^*(h)\right), \quad L^{\varphi}_{\overline{\mathcal{R}}} = \mathop{\mathbb{E}}_{h,s,o,a\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{\bar{s}\sim\varphi^*(h)} \left|\overline{\mathcal{R}}(\phi(s,o),a) - \overline{\mathcal{R}}(\bar{s},a)\right|,$$

$$L^{\varphi}_{\overline{\mathbf{P}}} = \mathop{\mathbb{E}}_{h,s,o,a\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{\bar{s}\sim\varphi^*(h)} \mathcal{W}_{\bar{d}}\left(\overline{\mathbf{P}}(\cdot \mid \phi(s,o),a), \overline{\mathbf{P}}(\cdot \mid \bar{s},a)\right). \tag{5}$$

$L^{\varphi}_{\overline{\mathcal{R}}}$ and $L^{\varphi}_{\overline{\mathbf{P}}}$ aim at regularizing φ and minimize the gap between the rewards (resp. transition probabilities) that are expected when drawing states from the current belief compared to those actually observed. Again, the ability to observe states during training enables optimizing those losses while the states are not required to execute the policy. The belief loss and the related two regularizers can be optimized *on-policy*, i.e., coupled with the optimization of $\overline{\pi}$ that is used to generate the episodes.

Value difference bounds. We provide guarantees concerning the *agent behaviors in* $\mathcal{P}$, when the policies are *conditioned on latent beliefs*. To do so, we formalize the behaviors of the agent through *value functions*. For a specific policy π, the value of a history is the expected return that would result from continuing to follow the policy from the latest point reached in that history: $V_{\pi}(h) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t\, r_t \mid b_I = \overline{\tau}^*(h)\right]$. Similarly, we write $\overline{V}_{\overline{\pi}}$ for the values of the latent policy $\overline{\pi}$ in $\overline{\mathcal{P}}$.

Suppose the agent uses a latent policy whose inputs are produced by φ, we claim that when the losses are minimized to zero, then (i) the latent model almost surely mimics the original environment, and (ii) our belief representation almost surely captures the value function. Precisely, let $\mathcal{L} = L_{\mathcal{R}} + L^{\varphi}_{\overline{\mathcal{R}}} + \overline{\mathcal{R}}^{\star} L_{\overline{\tau}} + \gamma K_{\overline{V}} \cdot (L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} + L_{\mathcal{O}})$, with $\overline{\mathcal{R}}^{\star} = \left\|\overline{\mathcal{R}}\right\|_{\infty}$ and $K_{\overline{V}} = {\overline{\mathcal{R}}^{\star}}/{1-\gamma}$, then:

Theorem 3.2 (Model quality)**.** *For any latent policy $\overline{\pi}\colon \overline{\mathcal{B}} \to \Delta(\mathcal{A})$, the values of $\mathcal{P}$ and $\overline{\mathcal{P}}$ are bounded by the local and belief losses in average when $\overline{\pi}$ is executed in $\mathcal{P}$ via $a \sim \overline{\pi}(\cdot \mid \varphi^*(h))$:*

$$\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \left|V_{\overline{\pi}}(h) - \overline{V}_{\overline{\pi}}(h)\right| \leqslant \frac{\mathcal{L}}{1-\gamma}. \tag{6}$$

Theorem 3.3 (Representation quality)**.** *Let $\overline{\pi}^{\star}$ be an optimal policy of $\overline{\mathcal{P}}$, then for any $\epsilon > 0$, there is a $K \geqslant 0$ so that for any histories h_1, h_2 measurable under $\mathcal{P}$ and $\overline{\mathcal{P}}$ with $\varphi^*(h_1) = \bar{b}_1$ and $\varphi^*(h_2) = \bar{b}_2$, the representation induced by φ almost surely yields:*

$$|V_{\overline{\pi}^{\star}}(h_1) - V_{\overline{\pi}^{\star}}(h_2)| \quad \leqslant \quad K\mathcal{W}_{\bar{d}}\left(\bar{b}_1, \bar{b}_2\right) \;+\; \epsilon \;+\; \frac{KL_{\overline{\tau}} + \mathcal{L}}{1-\gamma}\left(\frac{1}{\mathcal{H}_{\overline{\pi}^{\star}}(h_1)} + \frac{1}{\mathcal{H}_{\overline{\pi}^{\star}}(h_2)}\right). \tag{7}$$

While Thm. 3.2 asserts that training a WAE-MDP as a latent space model of the environment results in similar behaviors (i.e., close expected returns) compared to the original environment when they are measured under the agent policy — which justifies the usage of $\overline{\mathcal{P}}$ as model of the environment — Thm. 3.3 states that our learned update procedure yields a belief representation which is well-suited to optimize the policy: execution traces leading to close latent beliefs (via our learned updater φ) are guaranteed to yield close expected returns as well (proofs in Appendix F).

Analogous to τ^*, we define φ^* as the recursive application of φ along histories.

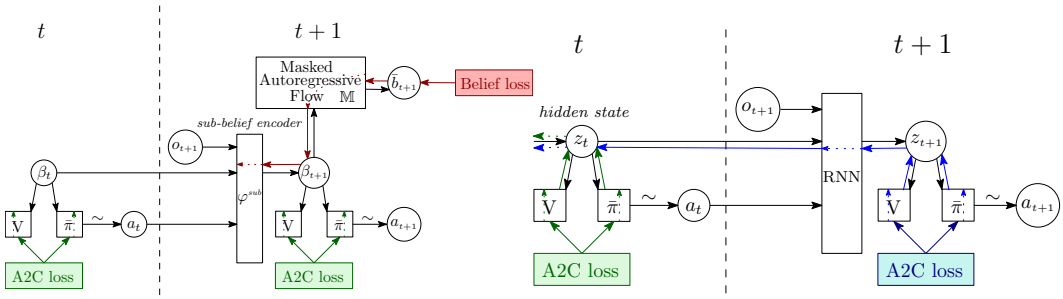


Figure 2: WBU (*left*) learns to encode the history into a sub-belief solely by optimizing the belief loss. The policy, being conditioned on the sub-belief, is learned via A2C and does not back-propagate through the sub-belief encoder. The R-A2C agent (*right*) uses BPTT: the RNN leverages gradients from future time-steps to improve its compression of the history for learning a policy and value function. In both plots, the colored arrows represent the gradient flows of the different losses.

4 Learning to Believe

In the following, we assume that we have access to the latent model learned by the WAE-MDP.

Architecture. Our latent belief encoder φ aims at generalizing to *any* POMDP. Therefore, *we do not make any assumption about the underlying belief distribution*. To accommodate complex belief distributions, we use a *Masked Auto-Regressive Flows* (MAF) (Papamakarios et al., 2017), a type of normalizing flow built on the auto-regressive property. Precisely, to fit with the WAE-MDP framework and leverage the guarantees presented in Sect. 3.2, we use the MAF of Delgrange et al. (2023) that learns relaxed multivariate latent distributions.

The *sub-belief* β_t is the vector that embeds the parameters of the belief distribution, which is converted into a belief via the MAF $\mathbb{M}(\beta_t) = \bar{b}_t$. We use a *sub-belief encoder* φ^{sub} to recursively update β_t via $\varphi^{\text{sub}}(\beta_t, a_t, o_{t+1}) = \beta_{t+1}$, so that $\varphi\big(\bar{b}_t, a_t, o_{t+1}\big) = \mathbb{M} \circ \varphi^{\text{sub}}(\beta_t, a_t, o_{t+1})$. RNNs are trained via back-propagation through time (BPTT), which is challenging (Pascanu et al., 2013). In contrast, albeit sub-beliefs are updated recursively in the same spirit as RNN hidden states, *we do not need to use BPTT and use a simple feed-forward network for* φ^{sub}, as illustrated in Fig. 2. In R-A2C, RNN hidden states serve as compact representations of histories for the policy. Since values of time-steps closer to the end of an episode are easier to learn, the gradients of future time-steps tend to be more accurate; thus BPTT helps learning. This is in stark contrast with learning the belief update rule: the beliefs of early time-steps are easier to infer, so BPTT is unnecessary and might even be armful. Additionally, viewing the belief update function as the transition function of $\mathcal{M}_{\mathcal{B}}$, disabling BPTT aligns with the literature on model-based RL for learning Markovian transition functions (Gelada et al., 2019; François-Lavet et al., 2019). See Appendix G for a more detailed discussion on BPTT.

Training. We aim to train φ^{sub} and $\mathbb{M}$ to approximate the update rule by minimizing the Wasserstein between the belief update rule $\bar{\tau}$ of the latent POMDP, and the belief encoder φ (Eq. 5), to leverage the theoretical learning guarantees of Thm. 3.2 and 3.3. However, Wasserstein optimization is known to be challenging, often requiring the use of additional networks, Lipschitz constraints, and a min-max optimization procedure (Arjovsky et al., 2017), similar to how WAE-MDPs are trained. Also, sampling from both distributions is necessary for optimizing Wasserstein and, while sampling from our belief approximation is straightforward, sampling from the update rule (Eq. 3) is non-trivial.

As an alternative to the Wasserstein optimization, we minimize the KL divergence between the two distributions. D_{KL} is easier to optimize and only requires sampling from one of the two distributions (in our case, the belief encoder). However, unlike the Wasserstein distance, guarantees can only be derived when the divergence approaches zero. Nonetheless, in the WAE-MDP zero-temperature limit, D_{KL} bounds Wasserstein by the Pinsker's inequality (Borwein & Lewis, 2005).

On-policy KL divergence. Using D_{KL} as a proxy for $\mathcal{W}_{\bar{d}}$ allows to narrow the gap between φ and $\bar{\tau}$. We train φ *on-policy*, with the same samples as used for $\bar{\pi}$, which aids learning despite gradients are not allowed to flow between the networks. At any time-step $t \geqslant 0$, given the current belief $\bar{b}_t$,

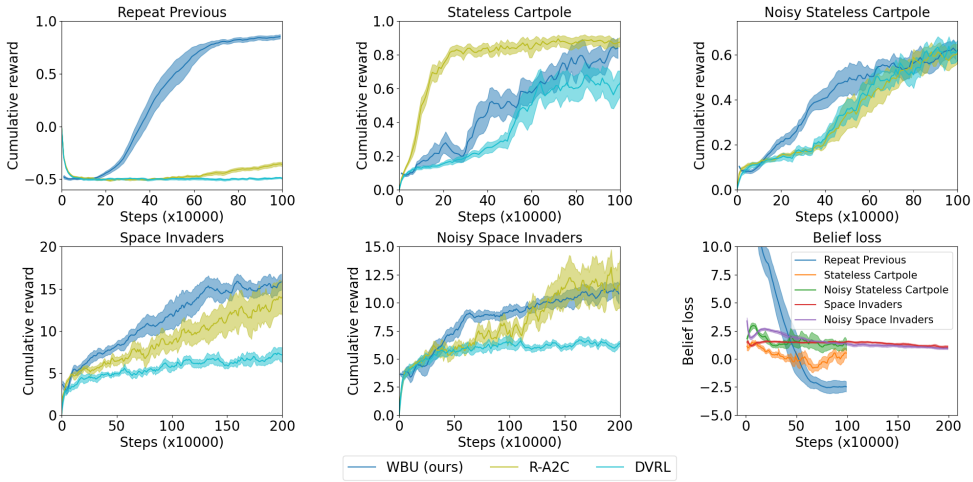


Figure 3: Evolution of the (i) undiscounted cumulative return for WBU, R-A2C and DVRL, and (ii) estimated belief loss during learning for WBU (mean and standard error). We report 5 instances of each algorithm. Appendix I details the hyperparameter search performed.

the action a_t played by the agent, and the next perceived observation o_{t+1}, the belief proxy loss is:

$$D_{\mathrm{KL}}\big(\varphi(\bar{b}_t, a_t, o_{t+1}) \parallel \bar{\tau}(\bar{b}_t, a_t, o_{t+1})\big) = \log\Big(\mathop{\mathbb{E}}_{\bar{s}\sim\bar{b}_t}\ \mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\bar{s},a_t)} \overline{\mathcal{O}}(o_{t+1} \mid \bar{s}')\Big) +$$

$$\mathop{\mathbb{E}}_{\bar{s}_{t+1}\sim\varphi(\bar{b}_t,a_t,o_{t+1})}\left[\log\varphi(\bar{s}_{t+1} \mid \bar{b}_t, a_t, o_{t+1}) - \log \mathop{\mathbb{E}}_{\bar{s}\sim\bar{b}_t} \overline{\mathbf{P}}(\bar{s}_{t+1} \mid \bar{s}, a_t) - \log\overline{\mathcal{O}}(o_{t+1} \mid \bar{s}_{t+1})\right]. \quad (8)$$

Eq. 8 consists of 4 terms: a normalization factor, negative entropy of φ, belief update conformity with the latent MDP's state transition function, and filtration of latent states unrelated to o_{t+1}.

Remark 3 (Smoothness of latent observations). The WAE-MDP learns from the augmented POMDP $\mathcal{P}^{\uparrow}$ (Sect. 3.1), which is equivalent to $\mathcal{P}$ and possesses a deterministic observation function. Similarly, the WAE-MDP also learns such a deterministic mapping through $\overline{\mathcal{O}}_\mu$. However, when deterministic, the observation terms of Eq. 3 and 8 are Dirac, which prevents learning: deterministically filtering out states from the next belief that do not share the next observation boils down to constructing the full belief distribution, which is usually intractable. In practice, to alleviate this, we model the observation function as a normal distribution: $\overline{\mathcal{O}}(\cdot \mid \bar{s}) = \mathcal{N}(\overline{\mathcal{O}}_\mu(\bar{s}), \sigma^2)$ where the variance can, e.g., be learned via $L_{\mathcal{O}}$. Notice that we can enforce annealing the variance to zero to recover a Dirac.

Policy learning is enabled by inputting the sub-belief into the policy, while the optimization of the belief encoder parameters by the RL agent is not allowed. Our method is applicable to *any* on-policy algorithm, and we employ A2C in our experiments. We provide the final algorithm in Appendix H.

5 Experiments

To evaluate our approach, we identify three types of POMDPs: those requiring *long-term memory*, those where *features of the state space are hidden* (and may be inferred from short-term memory), and those with *noisy* observations. Notably, we stress that long-term memory is crucial in POMDPs, whereas short-term memory could be mitigated by stacking frames (e.g., Mnih et al. 2015). We compare our agent to R-A2C and DVRL (Fig. 3), trained in environments from POPGym (Morad et al., 2023) and our own partially observable version of MinAtar (Young & Tian, 2019).

Memorization. The RepeatPrevious environment involves shuffling two decks of cards at the start of each episode and presenting the agent with a card at each time step. The goal is to identify the suit of the card seen 8 time steps earlier. *Our algorithm stands out as the sole method demonstrating mid- to long-term memorization capabilities.* Unlike other methods, notably DVRL which also attempts to learn a belief distribution, WBU provably acquires a suitable representation of the history by learning to maintain a sufficient statistic, thereby explaining its ability to retain past information.

Hidden features. We employ a cart pole scenario (STATELESSCARTPOLE) where velocity components of the system are hidden. R-A2C excels rapidly here, capitalizing on short-term memory to infer velocities from the preceding observation, while DVRL is overtaken. Still, WBU eventually reaches R-A2C final performance. We also explore the SPACEINVADERS environment, where the agent takes command of a cannon with the objective of shooting at groups of moving aliens. In the observation, we intentionally concealed the direction of alien movement and confounded friendly and enemy fires. In this more challenging setting, WBU excels by earning the highest rewards.

Noise. We explore two types of noise. First, we introduce *Gaussian noise* to the observations of STATELESSCARTPOLE. Second, for SPACEINVADERS, *binary noise* is injected via a radar-like mask obscuring the position of each alien with high probability. Hence, the agent must infer their positions based on previous observations. By leveraging its ability to maintain a belief over the noiseless (latent) state space, WBU demonstrates its resilience to noise and swiftly provides superior solutions, whereas R-A2C eventually achieves comparable performance but with more variance.

Belief representation. Ideally, close policy inputs should lead to close values, which would ease its optimization. Thm. 3.3 provides such a representation guarantee and ensures that the representation induced by φ captures the value function. To support this, we performed a t-SNE (van der Maaten & Hinton, 2008) on our belief representation at the late stage of training, which projects latent beliefs on a 2D space (Fig. 4). Interestingly, *latent beliefs clustered together have indeed close values*, in line with Thm. 3.3. We reported the belief loss throughout the training phases (Fig. 3). Importantly, unlike other baselines, *our approach distinctly separates the optimization of φ and $\overline{\pi}$*. Consequently, *the policy optimization does not influence the representation* which is solely learned via φ. The decrease in this loss thus relates to improved representation quality for RL.

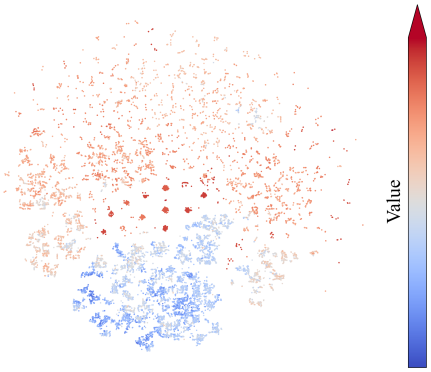


Figure 4: Two-dimensional t-SNE of our belief representation for SPACEINVADERS

6 CONCLUSION

WBU provides a principled approach that approximates directly the belief update for POMDPs, in contrast to SOTA methods that uses the RL objective and regularization to attempt to turn the history into a sufficient statistic. By learning the belief and its update rule, we provide strong guarantees on the quality of the belief, its ability to condition the optimal value function, and ultimately, the effectiveness of our algorithm. Our theoretical analysis and experimental results demonstrate the potential of our approach. Overall, WBU provides a promising new direction for RL in POMDPs, with potential applications in a wide range of settings where decision-making is complicated by uncertainty and partial observability, or when guarantees on the agent behaviors are required.

Future work. The theory we developed is not limited to the algorithm proposed in our paper. It opens diverse avenues for future work, e.g., on formally verifiable policies for POMDPs, by leveraging the guarantees presented in our framework. We also leave the study and the adaptation of our framework under relaxed assumptions (e.g., in settings akin to the work of Lambrechts et al. 2022) to future work. In addition, Thm 3.2 enables policy optimization through planning in the learned model, as demonstrated in successful model-based RL methods (e.g., Hafner et al. 2021). Scaling to high-dimensional observations (e.g., images) may potentially be computation-intensive due to observation filtering. For this further challenge, we suggest either to modify the WAE-MDP framework by using a stochastic decoder (see Tolstikhin et al. 2018, e.g., via PixelCNN, van den Oord et al. 2016), or learning a lower-dimensional latent observation space synced with the policy with a normalized or (relaxed) discrete prior (e.g., via a WAE-GAN, Tolstikhin et al. 2018). Finally, incorporating bisimulation metrics (Desharnais et al., 2004; Ferns et al., 2011) will strengthen guarantees for belief learning, even though bisimulation is challenging in POMDPs (Castro et al., 2009).

ACKNOWLEDGEMENTS

This research was supported by funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program and was supported by the DESCARTES iBOF project. R. Avalos is supported by the Research Foundation – Flanders (FWO), under grant number 11F5721N. G.A. Perez is also supported by the Belgian FWO "SAILor" project (G030020N). We thank Mathieu Reymond, Denis Steckelmacher, and Mustafa Mert Çelikok for their valuable feedback.

REPRODUCIBILITY STATEMENT

We referenced in the main text the parts of the Appendix presenting the proofs of our Lemma (Appendix D) and Theorems (Appendix F). We also provide the pseudo-code of our algorithm (Appendix H) as well as extra details required to compute our losses (Appendix B, E, and F.2). The code is available at `https://github.com/raphaelavalos/wbu`. Additionally, we provide the details of our hyperparameter search (Appendix I).

REFERENCES

Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 2017. URL `http://proceedings.mlr.press/v70/arjovsky17a.html`.

Raphael Avalos, Mathieu Reymond, Ann Nowé, and Diederik M. Roijers. Local Advantage Networks for Cooperative Multi-Agent Reinforcement Learning. In *AAMAS '22: Proceedings of the 21st International Conference on Autonomous Agents and MultiAgent Systems (Extended Abstract)*, 2022.

J.M. Borwein and A.S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. CMS Books in Mathematics. Springer New York, 2005. ISBN 9780387295701. URL `https://books.google.be/books?id=TXWzqEkAa7IC`.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *CoRR*, abs/1606.01540, 2016. URL `http://arxiv.org/abs/1606.01540`.

Peter E. Bulychev, Franck Cassez, Alexandre David, Kim Guldstrand Larsen, Jean-François Raskin, and Pierre-Alain Reynier. Controllers with minimal observation power (application to timed systems). In Supratik Chakraborty and Madhavan Mukund (eds.), *Automated Technology for Verification and Analysis - 10th International Symposium, ATVA 2012, Thiruvananthapuram, India, October 3-6, 2012. Proceedings*, volume 7561 of *Lecture Notes in Computer Science*, pp. 223–237. Springer, 2012. doi: 10.1007/978-3-642-33386-6_19. URL `https://doi.org/10.1007/978-3-642-33386-6_19`.

Pablo Samuel Castro, Prakash Panangaden, and Doina Precup. Equivalence relations in fully and partially observable markov decision processes. In Craig Boutilier (ed.), *IJCAI 2009, Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA, July 11-17, 2009*, pp. 1653–1658, 2009. URL `http://ijcai.org/Proceedings/09/Papers/276.pdf`.

Pablo Samuel Castro, Tyler Kastner, Prakash Panangaden, and Mark Rowland. Mico: Improved representations via sampling-based state similarity for markov decision processes. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 30113–30126, 2021. URL `https://proceedings.neurips.cc/paper/2021/hash/fd06b8ea02fe5b1c2496fe1700e9d16c-Abstract.html`.

Krishnendu Chatterjee, Luca de Alfaro, Rupak Majumdar, and Vishwanath Raman. Algorithms for game metrics (full version). *Log. Methods Comput. Sci.*, 6(3), 2010. URL http://arxiv.org/abs/0809.4326.

Krishnendu Chatterjee, Martin Chmelik, Raghav Gupta, and Ayush Kanodia. Optimal cost almost-sure reachability in pomdps. *Artif. Intell.*, 234:26–48, 2016. doi: 10.1016/j.artint.2016.01.007. URL https://doi.org/10.1016/j.artint.2016.01.007.

Xiaoyu Chen, Yao Mark Mu, Ping Luo, Shengbo Li, and Jianyu Chen. Flow-based recurrent belief state learning for pomdps. In *International Conference on Machine Learning*, pp. 3444–3468. PMLR, 2022.

Luca de Alfaro, Thomas A. Henzinger, and Rupak Majumdar. Discounting the future in systems theory. In Jos C. M. Baeten, Jan Karel Lenstra, Joachim Parrow, and Gerhard J. Woeginger (eds.), *Automata, Languages and Programming, 30th International Colloquium, ICALP 2003, Eindhoven, The Netherlands, June 30 - July 4, 2003. Proceedings*, volume 2719 of *Lecture Notes in Computer Science*, pp. 1022–1037. Springer, 2003. doi: 10.1007/3-540-45061-0_79. URL https://doi.org/10.1007/3-540-45061-0_79.

Florent Delgrange, Ann Nowé, and Guillermo A. Pérez. Distillation of rl policies with formal guarantees via variational abstraction of markov decision processes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6497–6505, Jun. 2022. doi: 10.1609/aaai.v36i6.20602. URL https://ojs.aaai.org/index.php/AAAI/article/view/20602.

Florent Delgrange, Ann Nowe, and Guillermo Perez. Wasserstein auto-encoded MDPs: Formal verification of efficiently distilled RL policies with many-sided guarantees. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JLLTtEdh1ZY.

Josée Desharnais, Vineet Gupta, Radha Jagadeesan, and Prakash Panangaden. Metrics for labelled markov processes. *Theor. Comput. Sci.*, 318(3):323–354, 2004. doi: 10.1016/j.tcs.2003.09.013. URL https://doi.org/10.1016/j.tcs.2003.09.013.

Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation metrics for continuous markov decision processes. *SIAM J. Comput.*, 40(6):1662–1714, 2011. doi: 10.1137/10080484X. URL https://doi.org/10.1137/10080484X.

Vincent François-Lavet, Yoshua Bengio, Doina Precup, and Joelle Pineau. Combined reinforcement learning via abstract representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3582–3589, 2019.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. Deepmdp: Learning continuous latent space models for representation learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2170–2179. PMLR, 2019. URL http://proceedings.mlr.press/v97/gelada19a.html.

Robert Givan, Thomas L. Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artif. Intell.*, 147(1-2):163–223, 2003. doi: 10.1016/S0004-3702(02)00376-4. URL https://doi.org/10.1016/S0004-3702(02)00376-4.

Karol Gregor, George Papamakarios, Frederic Besse, Lars Buesing, and Théophane Weber. Temporal Difference Variational Auto-Encoder. *7th International Conference on Learning Representations, ICLR 2019*, 6 2018. doi: 10.48550/arxiv.1806.03107. URL https://arxiv.org/abs/1806.03107v3.

Karol Gregor, Danilo Jimenez Rezende, Frederic Besse, Yan Wu, Hamza Merzic, and Aäron van den Oord. Shaping Belief States with Generative Environment Models for RL. *Advances in Neural Information Processing Systems*, 32, 6 2019. ISSN 10495258. doi: 10.48550/arxiv.1906.09237. URL https://arxiv.org/abs/1906.09237v2.

Danijar Hafner, Timothy Lillicrap Deepmind, Jimmy Ba, Mohammad Norouzi, and Google Brain. Dream to Control: Learning Behaviors by Latent Imagination. 12 2019. doi: 10.48550/arxiv.1912.01603. URL https://arxiv.org/abs/1912.01603v3.

Danijar Hafner, Timothy P Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0oabwyZbOu.

Matthew Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable MDPs. In *AAAI Fall Symposium - Technical Report*, volume FS-15-06, pp. 29–37. AI Access Foundation, 2015. ISBN 9781577357520.

Milos Hauskrecht. Value-function approximations for partially observable markov decision processes. *J. Artif. Intell. Res.*, 13:33–94, 2000. doi: 10.1613/jair.678. URL https://doi.org/10.1613/jair.678.

Bojun Huang. Steady state analysis of episodic reinforcement learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/69bfa2aa2b7b139ff581a806abf0a886-Abstract.html.

Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for POMDPs. In *35th International Conference on Machine Learning, ICML 2018*, volume 5, 2018.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, nov 2021. doi: 10.1109/tpami.2020.2992934. URL https://doi.org/10.1109%2Ftpami.2020.2992934.

Gaspard Lambrechts, Adrien Bolland, and Damien Ernst. Recurrent networks, hidden states and beliefs in partially observable environments. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=dkHfV3wB2l.

Gaspard Lambrechts, Adrien Bolland, and Damien Ernst. Informed POMDP: Leveraging additional information in model-based RL. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023. URL https://openreview.net/forum?id=Em4sFgoZk7.

Kim Guldstrand Larsen and Arne Skou. Bisimulation through probabilistic testing. In *Conference Record of the Sixteenth Annual ACM Symposium on Principles of Programming Languages, Austin, Texas, USA, January 11-13, 1989*, pp. 344–352. ACM Press, 1989. doi: 10.1145/75277.75307. URL https://doi.org/10.1145/75277.75307.

Mikko Lauri, David Hsu, and Joni Pajarinen. Partially observable markov decision processes in robotics: A survey. *IEEE Transactions on Robotics*, 39(1):21–40, 2023. doi: 10.1109/TRO.2022.3200138.

Alex X Lee, Anusha Nagabandi, Pieter Abbeel, and Sergey Levine. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. *Advances in Neural Information Processing Systems*, 33:741–752, 2020.

Jonathan Lee, Alekh Agarwal, Christoph Dann, and Tong Zhang. Learning in pomdps is sample-efficient with hindsight observability. In *International Conference on Machine Learning*, pp. 18733–18773. PMLR, 2023.

Xiao Ma, Peter Karkus, David Hsu, and Wee Sun Lee. Particle filter recurrent neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5101–5108, 2020.

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. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2 2015. ISSN 14764687. doi: 10.1038/nature14236.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous Methods for Deep Reinforcement Learning. *33rd International Conference on Machine Learning, ICML 2016*, 4:2850–2869, 2 2016. URL http://arxiv.org/abs/1602.01783.

Steven Morad, Ryan Kortvelesy, Matteo Bettini, Stephan Liwicki, and Amanda Prorok. POPGym: Benchmarking partially observable reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=chDrutUTs0K.

Yu Nishiyama, Abdeslam Boularias, Arthur Gretton, and Kenji Fukumizu. Hilbert Space Embeddings of POMDPs. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI2012)*, 10 2012. doi: 10.48550/arxiv.1210.4887. URL https://arxiv.org/abs/1210.4887v1.

Frans A. Oliehoek, Matthijs T.J. Spaan, and Nikos Vlassis. Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research*, 32:289–353, 10 2008. ISSN 10769757. doi: 10.1613/jair.2447. URL http://dx.doi.org/10.1613/jair.2447.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/6c1da886822c67822bcf3679d04369fa-Paper.pdf.

Fabio Pardo, Arash Tavakoli, Vitaly Levdik, and Petar Kormushev. Time limits in reinforcement learning. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4042–4051. PMLR, 2018. URL http://proceedings.mlr.press/v80/pardo18a.html.

Razvan Pascanu, Tomás Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 1310–1318. JMLR.org, 2013. URL http://proceedings.mlr.press/v28/pascanu13.html.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1994. ISBN 978-0-47161977-2. doi: 10.1002/9780470316887. URL https://doi.org/10.1002/9780470316887.

Sebastian Raschka, Joshua Patterson, and Corey Nolet. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *arXiv preprint arXiv:2002.04803*, 2020.

Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable markov processes over a finite horizon. *Oper. Res.*, 21(5):1071–1088, 1973. doi: 10.1287/opre.21.5.1071. URL https://doi.org/10.1287/opre.21.5.1071.

Edward J. Sondik. The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Oper. Res.*, 26(2):282–304, 1978. doi: 10.1287/opre.26.2.282. URL https://doi.org/10.1287/opre.26.2.282.

R.S. Sutton and A.G. Barto. Reinforcement Learning: An Introduction. *IEEE Transactions on Neural Networks*, 1998. ISSN 1045-9227. doi: 10.1109/tnn.1998.712192.

Ilya O. Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. 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, 2018. URL https://openreview.net/forum?id=HkL7n1-0b.

Aäron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Koray Kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4790–4798, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/b1301141feffabac455e1f90a7de2054-Abstract.html.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Elise van der Pol, Thomas Kipf, Frans A. Oliehoek, and Max Welling. Plannable approximations to MDP homomorphisms: Equivariance under actions. In Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith (eds.), *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020*, pp. 1431–1439. International Foundation for Autonomous Agents and Multiagent Systems, 2020. doi: 10.5555/3398761.3398926. URL https://dl.acm.org/doi/10.5555/3398761.3398926.

Cédric Villani. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-540-71050-9. doi: 10.1007/978-3-540-71050-9_6. URL https://doi.org/10.1007/978-3-540-71050-9_6.

Yuhui Wang and Xiaoyang Tan. Deep recurrent belief propagation network for pomdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10236–10244, 2021.

Yaxiong Wu, Craig Macdonald, and Iadh Ounis. Partially observable reinforcement learning for dialog-based interactive recommendation. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pp. 241–251, 2021.

Kenny Young and Tian Tian. Minatar: An atari-inspired testbed for thorough and reproducible reinforcement learning experiments. *arXiv preprint arXiv:1903.03176*, 2019.

Hongyu Zang, Xin Li, and Mingzhong Wang. Simsr: Simple distance-based state representations for deep reinforcement learning. 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*, pp. 8997–9005. AAAI Press, 2022. doi: 10.1609/AAAI.V36I8.20883. URL https://doi.org/10.1609/aaai.v36i8.20883.

Tianhao Zhang, Yueheng Li, Chen Wang, Guangming Xie, and Zongqing Lu. FOP: Factorizing Optimal Joint Policy of Maximum-Entropy Multi-Agent Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 2021.

K.J Åström. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965. ISSN 0022-247X. doi: https://doi.org/10.1016/0022-247X(65)90154-X. URL https://www.sciencedirect.com/science/article/pii/0022247X6590154X.

APPENDIX

A ADDITIONAL RELATED WORK

In the same spirit as in our work, the observability assumption of Lee et al. (2023) aligns perfectly with ours, but is leveraged in a different way, to provide results oriented towards sample complexity in the tabular case, as well as regret bounds when particular types of function approximators are used. Several complexity results and regrets bounds were developed for sub-classes of POMDPS in finite horizon, such as when the observability is well-behaved (weakly revealed POMDPs, Liu et al., 2022), or when the sequential decision making problem admits low rank structure (Liu et al., 2022, Huang et al., 2024). On the contrary, our results hold for general POMDPs; precisely, our constraints rather rely on the training procedure and not on the nature of the POMDP. Finally, we note that RQL-AIS (SeyedSalehi et al., 2023) presents an analysis of recurrent Q-learning for the tabular setting which also includes value difference bounds where their latent state is the (discrete) hidden state of an RNN. While theoretically interesting, their bounds are generally intractable as they rely on global metrics requiring them to be minimized over the whole history and action space. This limitation is made clear in their algorithm as they use the local version of those metrics as regularization losses. In our case, because we focus on on-policy metrics (local losses) our theoretical guarantees hold in practice.

Bisimulation. One of our result involves abstraction quality guarantees (Theorem 3.2), formalized as the closeness of the value function of the original and latent models. In spirit, this result can be related to *bisimulation*. Bisimulation is a behavioral *equivalence relation* between states and models whose properties include trajectory and value *equivalence* (Larsen & Skou, 1991; Givan et al., 2003). However, due to their "all or nothing" nature, bisimulation relations are often too restrictive; intuitively, states that share close rewards and dynamics but exhibit slight numerical differences are treated as being completely different. Bisimulation pseudometrics (de Alfaro et al., 2003; Desharnais et al., 2004) generalize bisimulation by assigning a distance between states and quantifying their bisimilarity instead of yielding binary signals indicating whether they are bisimilar or not. Specifically, real-valued (discounted) quantitative properties (de Alfaro et al., 2003; Chatterjee et al., 2010) as well as expected returns (Ferns et al., 2011) are guaranteed to be Lipschitz continuous w.r.t. the related bisimulation distance (compare with the *almost-Lipschitzness* of our Theorem 3.3).

Works on representation learning for RL (e.g., Gelada et al., 2019; van der Pol et al., 2020; Castro et al., 2021; Zhang et al., 2021; Zang et al., 2022; Delgrange et al., 2023) use bisimulation metrics to optimize the representation quality offered by a latent space, and aim at learning representations which capture the bisimulation distance, so that two states close in the representation provide close and relevant information for learning the policy. Precisely, *such works seek the Lipschitz continuity of the optimal value function with regards to the distance between pairs of points in the latent space*.

Note that while our work builds upon the WAE-MDP framework (Delgrange et al., 2023), which learns a latent model based on bisimulation metrics, our work does not learn any bisimulation-like metric in the underlying latent space of the POMDP. Instead, *we prioritize what truly matters in the context of representation for RL: the Lipschitz-continuity of the optimal value function*. In particular, the attentive reader may notice that, in contrast to usual representation bounds in MDPs (e.g., Gelada et al., 2019; Delgrange et al., 2022), our bound involves the almost Lipschitzness the value function, as well as various additional losses linked to the belief update and observation functions, which underlines the challenge of providing such guarantees. Indeed, bisimulation in the context of POMDPs represents a major challenge (e.g., Castro et al., 2009): intuitively, in the context of POMDPs, the bisimulation equivalence is not defined on the state space but rather on the belief space induced by the different components of the POMDP. Notably, even if the POMDP is finite, i.e., state, action, and observation spaces are finite, the belief space can be represented as the usual (uncountable space) simplex, which prevents the application of bisimulation algorithms developed in the context of countable MDPs.

B THE BELIEF UPDATE RULE

Let $\overline{\mathcal{P}} = \langle \overline{\mathcal{M}}, \Omega, \overline{\mathcal{O}} \rangle$ be a latent POMDP with underlying MDP $\overline{\mathcal{M}} = \langle \bar{\mathcal{S}}, \mathcal{A}, \overline{\mathbf{P}}, \overline{\mathcal{R}}, \bar{s}_I, \gamma \rangle$. At step $t \geqslant 0$, assume that the current latent belief is $\bar{b}_t \in \overline{\mathcal{B}}$. Then, when $a_t \in \mathcal{A}$ is executed and $o_{t+1} \in \Omega$

is observed, $\bar{b}_t$ is updated according to $\bar{\tau}\big(\bar{b}_t, a_t, o_{t+1}\big)$ as follows, so that for any (believed) state $\bar{s}_{t+1} \in \bar{\mathcal{S}}$,

$$\bar{b}_{t+1}(\bar{s}_{t+1}) = \frac{\mathbb{E}_{\bar{s}_t \sim \bar{b}_t} \overline{\mathbf{P}}(\bar{s}_{t+1} \mid \bar{s}_t, a_t) \cdot \overline{\mathcal{O}}(o_{t+1} \mid \bar{s}_{t+1})}{\mathbb{E}_{\bar{s}_t \sim \bar{b}_t} \mathbb{E}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot \mid \bar{s}_t, \bar{a}_t)} \overline{\mathcal{O}}(o_{t+1} \mid \bar{s}')}. \tag{9}$$

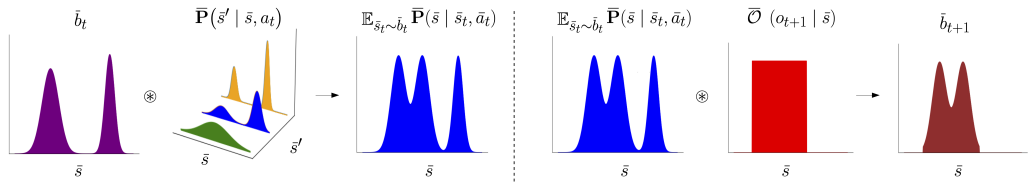


Figure 5: The belief update rule: (*left*) transformation of the current belief $\bar{b}_t$ with the transition probability function $\overline{\mathbf{P}}$, evaluated on the current action a_t, into the next state probability density; (*right*) filtering out the next states that could not have produced the next observation o_{t+1}.

The belief update rule $\bar{\tau}$ is divided in two steps (cf. Fig. 5). First, the current belief distribution $\bar{b}_t$ is used to marginalize the latent transition function $\overline{\mathbf{P}}$ over the believed latent states, to further infer the distribution over the possible next states. This first part corresponds to computing the probability of going to the next believed state $\bar{s}_{t+1}$ from the possible states $\bar{s}_t$ witnessed by the current belief $\bar{b}_t$. Second, the next observation o_{t+1} is used to filter the resulting density based on the observation function: believed latent states that do not share the observation o_{t+1} perceived are filtered out according to $\overline{\mathcal{O}}(o_{t+1} \mid \bar{s}_{t+1})$. It is worth noting that the latent model is learned from $\mathcal{P}^{\uparrow}$, whose observation function is deterministic. Without modelling the latent observation function $\overline{\mathcal{O}}$ as a normal distribution, the second part of the belief update would need to eliminate all next states with different observations — which is not gradient descent friendly. The third operation (not present in Fig. 5) normalizes the output of the observation filtering to obtain a probability density.

C Dirac Measures

In this work, we consider the *Dirac delta function* δ as a *measure*. Specifically, this means that for any complete, separable space $\mathcal{X}$ and point $a \in \mathcal{X}$, the *Dirac measure* with impulse a is $\delta_a \in \Delta(\mathcal{X})$ and satisfies $\delta_a(A) = 1$ if $a \in A$ and $\delta_a(A) = 0$ otherwise, for any $A \in \Sigma(\mathcal{X})$. Interesting properties of the Dirac measure include $\delta_a = \lim_{\sigma \to 0} \mathcal{N}(a, \sigma^2)$, where $\mathcal{N}(a, \sigma^2)$ is the normal distribution with mean a and variance σ^2, and $\int_{\mathcal{X}} \delta_a(x) f(x)\, dx = f(a)$ for any compactly supported function f.

D Proof of Lemma 3.1: Stationarity over Histories

Let us formally restate the Lemma:

Lemma D.1. *Let $\mathcal{P}$ be an episodic POMDP with action space $\mathcal{A}$ and observation space Ω, and $\bar{\pi}$ be a latent policy conditioned over the beliefs of a latent POMDP $\overline{\mathcal{P}}$ that shares the action and observation spaces of $\mathcal{P}$ while being executed via the belief encoder, i.e., $a \sim \bar{\pi}(\cdot \mid \varphi^*(h))$ for any $h \in (\mathcal{A} \cdot \Omega)^*$. Then, there is a well defined probability distribution $\mathcal{H}_{\bar{\pi}} \in \Delta\big((\mathcal{A} \cdot \Omega)^*\big)$ over histories drawn at the limit from the interaction of the RL agent with $\mathcal{P}$, when it operates under the latent policy $\bar{\pi}$.*

Proof sketch. Build a *history unfolding* as the MDP whose state space consists of all histories and keeps track of the current history of $\mathcal{P}$ at any time of the interaction. The resulting MDP remains episodic since it is equivalent to $\mathcal{P}$: the former mimics the behaviors of the latter under $\bar{\pi}$. All episodic processes are *ergodic* (e.g., Huang 2020), which guarantees the existence of such a distribution. □

We dedicate this Section to formally detailing and proving every claim of this proof sketch. Before going further, we formally define the notion of *episodic process*, and we further introduce the notions of *memory-based policies*, *Markov Chains*, and *limiting distributions in Markov Chains*.

D.1 Preliminaries

Recall by Assumption 2 that the environment $\mathcal{P}$ is an episodic process. We *formally* recall the notion of *episodic process*:

Definition D.2 (Episodic RL process)**.** *The RL procedure is* episodic *iff the environment $\mathcal{P}$ embeds a special* reset state $s_{\mathsf{reset}} \in \mathcal{S}$ *so that*

(i) under any policy π, the environment is almost surely eventually reset:

$$\mathbb{P}_{\pi}^{\mathcal{M}}(\{\, s_{0:\infty}, a_{0:\infty} \mid \exists t > 0, s_t = s_{\mathsf{reset}} \,\}) = 1;$$

(ii) when reset, the environment transitions to the initial state: $\mathbf{P}(s_I \mid s_{\mathsf{reset}}, a) > 0$ and $\mathbf{P}(\mathcal{S} \setminus \{\, s_I, s_{\mathsf{reset}} \,\} \mid s_{\mathsf{reset}}, a) = 0$ for all $a \in \mathcal{A}$, and the distribution is the same for all actions, i.e., for all $a, a' \in \mathcal{A}$, $\mathbf{P}(\cdot \mid s_{\mathsf{reset}}, a) = \mathbf{P}(\cdot \mid s_{\mathsf{reset}}, a')$;

(iii) the reset state is observable*: there is an observation $o^\star \in \Omega$ so that $\mathcal{O}(o^\star \mid s', a) = 0$ when $s' \neq s_{\mathsf{reset}}$, and $\mathcal{O}(\cdot \mid s_{\mathsf{reset}}, a) = \delta_{o^\star}$ for $a \in \mathcal{A}$.*

An episode *is a history $\langle a_{0:T-1}, o_{1:T} \rangle$ where $\mathcal{O}(o_1 \mid s_I, a_0) > 0$ and $o_T = o^\star$.*

Encoding memory through Mealy machines. *Policies* are building blocks to define the probability space of any MDP. To deal with policies whose decisions are based on unrolling histories, we formally define the notion of *memory* in policies.

Definition D.3 (Policy as Mealy Machine)**.** *Given an MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathbf{P}, \mathcal{R}, s_I, \gamma \rangle$, any policy for $\mathcal{M}$ can be encoded as a* stochastic Mealy machine *$\pi = \langle Q, \pi_\alpha, \pi_\mu, q_I \rangle$ as follows: Q is a set of* memory states*; $\pi_\alpha \colon \mathcal{S} \times Q \to \Delta(\mathcal{A})$ is the* next action function*; $\pi_\mu \colon \mathcal{S} \times Q \times \mathcal{A} \times \mathcal{S} \to \Delta(Q)$ is the* memory update function*; and q_I is the initial memory state.*

Example 1 (Stationary policy). A stationary policy π can be encoded as any Mealy machine π with memory space Q where $|Q| = 1$.

Example 2 (Latent policy). Let $\mathcal{P} = \langle \mathcal{M}, \Omega, \mathcal{O} \rangle$ with underlying MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathbf{P}, \mathcal{R}, s_I, \gamma \rangle$ and the latent space model $\overline{\mathcal{P}}$ with initial state $\bar{s}_I$ be the POMDPs of Lemma D.1. Then, any latent (stationary) policy $\overline{\pi} \colon \overline{\mathcal{B}} \to \Delta(\mathcal{A})$ conditioned on the belief space $\overline{\mathcal{B}}$ of $\overline{\mathcal{P}}$ can be executed in the belief MDP $\mathcal{M}_{\mathcal{B}}$ of $\mathcal{P}$ via the Mealy machine $\overline{\pi}' = \langle \overline{\mathcal{B}}, \overline{\pi}_\alpha, \overline{\pi}_\mu, \delta_{\bar{s}_I} \rangle$, keeping track in its memory of the current latent belief $\bar{b} \in \overline{\mathcal{B}}$ inferred by our belief encoder φ. This enables the agent to take its decisions solely based on the latter: $\overline{\pi}_\alpha(\cdot \mid b, \bar{b}) = \overline{\pi}(\cdot \mid \bar{b})$. When the belief MDP transitions to the next belief b', the memory is then updated according to the observation dynamics:

$$\overline{\pi}_\mu(\bar{b}' \mid b, \bar{b}, a, b') = \frac{\mathbb{E}_{s \sim b} \mathbb{E}_{s' \sim \mathbf{P}(\cdot \mid s, a)} \mathbb{E}_{o' \sim \mathcal{O}(\cdot \mid s', a)} \delta_{\varphi(\bar{b}, a, o')}(\bar{b}') \cdot \delta_{\tau(b, a, o')}(b')}{\mathbb{E}_{s \sim b} \mathbb{E}_{s' \sim \mathbf{P}(\cdot \mid s, a)} \mathbb{E}_{o' \sim \mathcal{O}(\cdot \mid s', a)} \delta_{\tau(b, a, o')}(b')} \text{ if } b' \neq \delta_{s_{\mathsf{reset}}},$$

$$\overline{\pi}_\mu(\cdot \mid b, \bar{b}, a, \delta_{s_{\mathsf{reset}}}) = \delta_{\bar{s}_{\mathsf{reset}}} \qquad \text{otherwise (to fulfil the episodic constraint).}$$

Note that $\overline{\pi}_\mu$ is simply obtained by applying the usual conditional probability rule: $\overline{\pi}_\mu(\bar{b}' \mid b, \bar{b}, a, b') = \Pr(b', \bar{b}' \mid b, \bar{b}, a) / \Pr(b' \mid b, \bar{b}, a)$, where $\Pr(b', \bar{b}' \mid b, \bar{b}, a) = \mathbb{E}_{s \sim b} \mathbb{E}_{s' \sim \mathbf{P}(\cdot \mid s, a)} \mathbb{E}_{o' \sim \mathcal{O}(\cdot \mid s', a)} \delta_{\varphi(\bar{b}, a, o')}(\bar{b}') \cdot \delta_{\tau(b, a, o')}(b')$ and $\Pr(b' \mid b, \bar{b}, a) = \mathbf{P}_{\mathcal{B}}(b' \mid b, a)$ since the next *original* belief state is independent of the current *latent* belief state.

Definition D.4 (Markov Chain)**.** *A* Markov Chain *(MC) is an MDP whose action space $\mathcal{A}$ consists of a singleton, i.e., $|\mathcal{A}| = 1$. Any MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathbf{P}, \mathcal{R}, s_I, \gamma \rangle$ and memory-based policy $\pi = \langle Q, \pi_\alpha, \pi_\mu, q_I \rangle$ induces a Markov Chain*

$$\mathcal{M}^\pi = \langle \mathcal{S} \times Q, \mathbf{P}_\pi, \mathcal{R}_\pi, \langle s_I, q_I \rangle, \gamma \rangle,$$

where:

- *the state space consists of the product of the original state space and the memory of π;*

- *the transition function embeds the next action and the policy update functions from the policy, i.e.,*
$$\mathbf{P}_\pi\big(\langle s', q'\rangle \mid \langle s, q\rangle\big) = \mathop{\mathbb{E}}_{a\sim\pi_\alpha(\cdot\mid s,q)} \pi_\mu\big(q' \mid s, q, a, s'\big) \cdot \mathbf{P}\big(s' \mid s, a\big), \text{ and}$$
- *the rewards are averaged over the possible actions produced by the next action function, i.e.,* $\mathcal{R}_\pi(\langle s, q\rangle) = \mathbb{E}_{a\sim\pi_\alpha(\cdot\mid s,q)}\, \mathcal{R}(s, a)$.

The probability measure $\mathbb{P}_\pi^{\mathcal{M}}$ *is actually the unique probability measure defined over the measurable infinite trajectories of the MC* $\mathcal{M}^\pi$ (Baier & Katoen, 2008). We omit the memory dimension Q when π is stationary (i.e., memoryless).

Limiting distributions. We now formally define the distribution over states encountered at the limit when an agent operates in an MDP under a given policy, as well as the conditions of existence of such a distribution. To that aim, we distinctly separate the discrete and the continuous cases, where the conditions to reach the stationarity of the system and the properties that hold at the limit of the interaction are different. In the following, we write $\mathcal{M}[s]$ for the MDP where we change the initial state s_I of $\mathcal{M}$ by $s \in \mathcal{S}$.

(a) *Finite state spaces.* Let $\mathcal{M}$ be an MDP with *finite* state space $\mathcal{S}$ and π be a stationary policy for $\mathcal{M}$. The measure $\xi_\pi^t : \mathcal{S} \to \Delta(\mathcal{S})$ with $\xi_\pi^t(s' \mid s) = \mathbb{P}_\pi^{\mathcal{M}[s]}\big(\{\, s_{0:\infty}, a_{0:\infty} \mid s_t = s' \,\}\big)$ is the distribution giving the probability for the agent of being in each state of $\mathcal{M}[s]$ after exactly t steps.

Definition D.5 (BSCC)**.** *The subset $B \subseteq \mathcal{S}$ is a* strongly connected component *(SCC) of $\mathcal{M}^\pi$ if for any pair of states $s, s' \in B$, there is a $t \in \mathbb{N}$ so that $\xi_\pi^t(s' \mid s) > 0$. It is a* bottom SCC *(BSCC) if (i) B is a maximal SCC, and (ii) for each $s \in B$, $\mathbf{P}_\pi(B \mid s) = 1$.*

The unique *stationary distribution* of B is $\xi_\pi \in \Delta(B)$, defined as $\xi_\pi(s) = \mathbb{E}_{\dot s\sim\xi_\pi} \mathbf{P}_\pi(s \mid \dot s) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} \xi_\pi^t(s \mid s_\perp)$ for any $s_\perp \in B$ (Baier & Katoen, 2008).

The *period* of a state $s \in \mathcal{S}$ is the greatest common divisor of the set $\{\, t > 0 \mid \xi_\pi^t(s \mid s) > 0 \,\}$. A BSCC B is *aperiodic* iff all states of B have period one.

An MDP $\mathcal{M}$ is *ergodic* under the policy π if the state space of $\mathcal{M}^\pi$ consists of a unique aperiodic BSCC. In that case, $\xi_\pi = \lim_{t\to\infty} \xi_\pi^t(\cdot \mid s)$ for all $s \in \mathcal{S}$.

(b) *General state spaces.* For convenience, we focus on a subset of well-behaved Markov chains, namely *Harris chains* (e.g., Athreya & Ney 1978; Revuz 1984), which are guaranteed to converge to a stationary distribution at the limit under some conditions. Let $\mathcal{M}$ be an MDP with a *complete, separable* state space $\mathcal{S}$ *endowed with a Borel σ-algebra*. We denote by $\Sigma(\mathcal{S})$ the set of Borel-measurable events of $\mathcal{S}$. Let π be a stationary policy for $\mathcal{M}$. The measure $\xi_\pi^t : \mathcal{S} \to \Delta(\mathcal{S})$ with $\xi_\pi^t(S' \mid s) = \mathbb{P}_\pi^{\mathcal{M}[s]}\big(\{\, s_{0:\infty}, a_{0:\infty} \mid s_t \in S' \,\}\big)$ is the distribution giving the probability for the agent of being in a state $s' \in S$ after exactly t steps, starting from $s \in \mathcal{S}$, for any $S' \in \Sigma(\mathcal{S})$. Denote the *reachability* event to $H \in \Sigma(\mathcal{S})$ as

$$\Diamond H = \{\, \langle s_{0:\infty}, a_{0:\infty}\rangle \in \mathit{Traj}_{\mathcal{M}} \mid \text{there is a } t \in \mathbb{N} \text{ so that } s_t \in H \,\}\,.$$

Definition D.6 (Recurrent Harris chain)**.** *The induced MC $\mathcal{M}^\pi$ is a* recurrent Harris chain *if there exist a measurable set $H \in \Sigma(\mathcal{S})$, an $\epsilon \in (0, 1]$, and a probability measure $\rho \in \Delta(\mathcal{S})$ so that*

(i) H is recurrent*, i.e., $\mathbb{P}_\pi^{\mathcal{M}[s]}(\Diamond H) = 1$ for all $s \in \mathcal{S}$; and*
(ii) H is a small set*, i.e., for any $s \in H$ and $C \in \Sigma(\mathcal{S})$, $\mathbf{P}_\pi(C \mid s) \geqslant \epsilon\rho(C)$.*

Recurrent Harris chains are guaranteed to have a stationary measure $\xi_\pi \in \Delta(\mathcal{S})$ (Durrett, 2010, Theorem 5.8.9), so that $\xi_\pi(S) = \mathbb{E}_{\dot s\sim\xi_\pi} \mathbf{P}_\pi(S \mid \dot s)$ for any $S \in \Sigma(\mathcal{S})$.

Similarly to the discrete case, we say that a Harris chain is *aperiodic* if and only if there is a moment from which the agent is able to return to the region H *at any time*: for any $s \in H$, there exists a moment $i \in \mathbb{N}$ so that for any $j \geqslant i$, $\xi_\pi^j(H \mid s) > 0$.

Assume $\mathcal{M}_\pi$ is an aperiodic recurrent Harris chain with stationary distribution ξ_π, then ξ_π^t converges to ξ_π *in total variation* as t goes to ∞ (Durrett, 2010, Theorem 5.8.12), i.e.,

$$\lim_{t\to\infty} d_{TV}\big(\xi_\pi^t(\cdot \mid s), \xi_\pi\big) = 0$$

for all $s \in \mathcal{S}$.

To provide such a stationary distribution over histories, we define a *history unfolding* MDP, where the state space keeps track of the current history of $\mathcal{P}$ during the interaction. We then show that this history MDP is *equivalent* to $\mathcal{P}$ under $\overline{\pi}$.

D.2 History Unfolding

Let us define the *history unfolding* MDP $\mathcal{M}_{\mathcal{H}}$, which consists of the tuple $\langle \mathcal{S}_{\mathcal{H}}, \mathcal{A}, \mathbf{P}_{\mathcal{H}}, \mathcal{R}_{\mathcal{H}}, \star, \gamma \rangle$, where:

- the state space consists of the set of all the possible histories (i.e., sequence of actions and observations) that can be encountered in $\mathcal{P}$, i.e., $\mathcal{S}_{\mathcal{H}} = (\mathcal{A} \cdot \Omega)^* \cup \{ \star, h_{\mathsf{reset}} \}$, which additionally embeds a special symbol $\star$ indicating that no observation has been perceived yet with $\tau^*(\star) = \delta_{s_I}$, as well as a special reset state h_{reset};
- the transition function maps the current history to the belief space to infer the distribution over the next possible observations, i.e.,

$$\mathbf{P}_{\mathcal{H}}\big(h' \mid h, a\big) = \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s' \sim \mathbf{P}(\cdot \mid s,a)} \mathop{\mathbb{E}}_{o' \sim \mathcal{O}(\cdot \mid s',a)} \delta_{h \cdot a \cdot o'}\big(h'\big) \qquad \text{if } \tau^*(h) \neq \delta_{s_{\mathsf{reset}}}, \text{ and}$$

$$\mathbf{P}_{\mathcal{H}}\big(h' \mid h, a\big) = \mathbf{P}_{\mathcal{B}}(\delta_{s_{\mathsf{reset}}} \mid \delta_{s_{\mathsf{reset}}}, a) \cdot \delta_{h_{\mathsf{reset}}}\big(h'\big) + \mathbf{P}_{\mathcal{B}}(\delta_{s_I} \mid \delta_{s_{\mathsf{reset}}}, a) \cdot \delta_{\star}\big(h'\big) \qquad \text{otherwise,}$$

  where $h \cdot a \cdot o'$ is the concatenation of a, o' with the history $h = \langle a_{0:T-1}, o_{1:T} \rangle$, resulting in the history $\langle a_{0:T}, o_{1:T+1} \rangle$ so that $a_T = a$ and $o_{T+1} = o'$; and
- the reward function maps the history to the belief space as well, which enables to infer the expected rewards obtained in the states over the this belief, i.e., $\mathcal{R}_{\mathcal{H}}(h, a) = \mathbb{E}_{s \sim \tau^*(h)} \mathcal{R}(s, a)$.

We now aim at showing that, under the latent policy $\overline{\pi}$, the POMDP $\mathcal{P}$ and the MDP $\mathcal{M}_{\mathcal{H}}$ are *equivalent*. More formally, we are looking for an equivalence relation between two probabilistic models, so that the latter induce the same behaviors, or in other words, the same expected return. We formalize this equivalence relation as a *stochastic bisimulation* between $\mathcal{M}_{\mathcal{B}}$ (that we know being an MDP formulation of $\mathcal{P}$) and $\mathcal{M}_{\mathcal{H}}$.

Definition D.7 (Bisimulation)**.** *Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathbf{P}, \mathcal{R}, s_I, \gamma \rangle$ be an MDP. A stochastic* bisimulation *$\equiv$ on $\mathcal{M}$ is a behavioral equivalence between states $s_1, s_2 \in \mathcal{S}$ so that, $s_1 \equiv s_2$ iff*

1. *$\mathcal{R}(s_1, a) = \mathcal{R}(s_2, a)$, and*
2. *$\mathbf{P}(T \mid s_1, a) = \mathbf{P}(T \mid s_2, a)$,*

for each action $a \in \mathcal{A}$ and equivalence class $T \in \mathcal{S}/\equiv$.

Properties of bisimulation include trajectory equivalence and the equality of their optimal expected return (Larsen & Skou, 1989; Givan et al., 2003). The relation can be extended to compare two MDPs by considering the disjoint union of their state space.

Lemma D.8. *Let $\mathcal{P}$ be the POMDP of Lemma D.1, and $\overline{\pi} \colon \overline{\mathcal{B}} \to \Delta(\mathcal{A})$ be a latent policy conditioned on the beliefs of a latent space model of $\mathcal{P}$. Define the stationary policy $\overline{\pi}^{\clubsuit} \colon \mathcal{S}_{\mathcal{H}} \to \Delta(\mathcal{A})$ for $\mathcal{M}_{\mathcal{H}}$ as $\overline{\pi}^{\clubsuit}(\cdot \mid h) = \overline{\pi}(\cdot \mid \varphi^*(h))$, and the policy $\overline{\pi}^{\diamondsuit}$ for $\mathcal{M}_{\mathcal{B}}$ encoded by the Mealy machine detailed in Example 2. Then, $\mathcal{M}_{\mathcal{H}}^{\overline{\pi}^{\clubsuit}}$ and $\mathcal{M}_{\mathcal{B}}^{\overline{\pi}^{\diamondsuit}}$ are in stochastic bisimulation.*

Proof. First, note that the MC $\mathcal{M}_{\mathcal{B}}^{\overline{\pi}^{\diamondsuit}}$ is defined as the tuple $\big\langle \mathcal{B} \times \overline{\mathcal{B}}, \mathbf{P}_{\overline{\pi}^{\diamondsuit}}, \mathcal{R}_{\overline{\pi}^{\diamondsuit}}, \big\langle b_I, \bar{b}_I \big\rangle, \gamma \big\rangle$ so that

$$\begin{aligned}
\mathbf{P}_{\overline{\pi}^{\diamondsuit}}\big(b', \bar{b}' \mid b, \bar{b}\big) &= \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \bar{b})} \overline{\pi}_{\mu}\big(\bar{b}' \mid b, \bar{b}, a, b'\big) \cdot \mathbf{P}_{\mathcal{B}}\big(b' \mid b, a\big) \\
&= \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \bar{b})} \mathop{\mathbb{E}}_{s \sim b} \mathop{\mathbb{E}}_{s' \sim \mathbf{P}(\cdot \mid s,a)} \mathop{\mathbb{E}}_{o' \sim \mathcal{O}(\cdot \mid s,a)} \delta_{\varphi(\bar{b},o,a)}\big(\bar{b}'\big) \cdot \delta_{\tau(b,o,a)}\big(b'\big), \text{and} \\
\mathcal{R}_{\overline{\pi}^{\diamondsuit}}(b, \bar{b}) &= \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \bar{b})} \mathop{\mathbb{E}}_{s \sim b} \mathcal{R}(s, a). \qquad \text{(cf. Definition D.4)}
\end{aligned}$$

Define the relation $\rightrightarrows_{\varphi}^{\tau}$ as the set $\{ \big\langle h, \langle b, \bar{b} \rangle \big\rangle \mid \tau^*(h) = b \text{ and } \varphi^*(h) = \bar{b} \} \subseteq \mathcal{S}_{\mathcal{H}} \times \mathcal{B} \times \overline{\mathcal{B}}$. We show that $\rightrightarrows_{\varphi}^{\tau}$ is a bisimulation relation between the states of $\mathcal{M}_{\mathcal{H}}^{\overline{\pi}^{\clubsuit}}$ and $\mathcal{M}_{\mathcal{B}}^{\overline{\pi}^{\diamondsuit}}$. Let $h \in \mathcal{S}_{\mathcal{H}}$, $b \in \mathcal{B}$, and $\bar{b} \in \overline{\mathcal{B}}$ so that $h \rightrightarrows_{\varphi}^{\tau} \langle b, \bar{b} \rangle$:

1. $\mathcal{R}_{\overline{\pi}^{\clubsuit}}(h) = \mathbb{E}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\,\mathbb{E}_{s\sim\tau^*(h)}\,\mathcal{R}(s,a) = \mathbb{E}_{a\sim\overline{\pi}\left(\cdot|\bar{b}\right)}\,\mathbb{E}_{s\sim b}\,\mathcal{R}(s,a) = \mathcal{R}_{\overline{\pi}^{\diamondsuit}}(b,\bar{b})$;

2. Each equivalence class $T \in \left(\mathcal{S}_{\mathcal{H}} \times \mathcal{B} \times \overline{\mathcal{B}}\right)/\rightrightarrows^{\tau}_{\varphi}$ consists of histories sharing the same belief and latent beliefs. Then, each equivalence class T can be associated to a single belief and latent belief pair. Concretely, let $b' \in \mathcal{B}$, $\bar{b}' \in \overline{\mathcal{B}}$, an equivalence class of $\rightrightarrows^{\tau}_{\varphi}$ has the form $T = \left[\langle b', \bar{b}'\rangle\right]_{\rightrightarrows^{\tau}_{\varphi}}$ so that

   (a) the projection of $\left[\langle b', \bar{b}'\rangle\right]_{\rightrightarrows^{\tau}_{\varphi}}$ on $\mathcal{S}_{\mathcal{H}}$ is the set $\{\, h' \in \mathcal{S}_{\mathcal{H}} \mid \tau^*(h') = b' \text{ and } \varphi^*(h') = \bar{b}' \,\}$, and

   (b) the projection of $\left[\langle b', \bar{b}'\rangle\right]_{\rightrightarrows^{\tau}_{\varphi}}$ on the state space of $\mathcal{M}_{\mathcal{B}}^{\overline{\pi}^{\diamondsuit}}$ is merely the pair $\langle b', \bar{b}'\rangle$.

   Therefore,

$$
\begin{aligned}
&\mathbf{P}_{\overline{\pi}^{\clubsuit}}\left(\left[\langle b', \bar{b}'\rangle\right]_{\rightrightarrows^{\tau}_{\varphi}} \mid h\right) \\
&= \int_{\left[\langle b', \bar{b}'\rangle\right]_{\rightrightarrows^{\tau}_{\varphi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))} \mathop{\mathbb{E}}_{s\sim\tau^*} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)} \delta_{h\cdot a\cdot o'}\left(h'\right)\, dh' \\
&= \int_{\mathcal{S}_{\mathcal{H}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))} \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)} \delta_{h\cdot a\cdot o'}\left(h'\right) \cdot \delta_{\tau^*(h')}\left(b'\right) \cdot \delta_{\varphi^*(h')}\left(\bar{b}'\right)\, dh' \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by definition of } \left[\langle b', \bar{b}'\rangle\right]_{\rightrightarrows^{\tau}_{\varphi}}) \\
&= \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))} \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)} \delta_{\tau^*(h\cdot a\cdot o')}\left(b'\right) \cdot \delta_{\varphi^*(h\cdot a\cdot o')}\left(\bar{b}'\right) \\
&= \mathop{\mathbb{E}}_{a\sim\overline{\pi}\left(\cdot|\bar{b}\right)} \mathop{\mathbb{E}}_{s\sim b} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)} \delta_{\tau(b,a,o')}\left(b'\right) \cdot \delta_{\varphi\left(\bar{b},a,o'\right)}\left(\bar{b}'\right) \qquad \text{(since } h \rightrightarrows^{\tau}_{\varphi} \langle b, \bar{b}\rangle) \\
&= \mathbf{P}_{\overline{\pi}^{\diamondsuit}}\left(b', \bar{b}' \mid b, \bar{b}\right) \\
&= \mathbf{P}_{\overline{\pi}^{\diamondsuit}}\left(\left[\langle b', \bar{b}'\rangle\right]_{\rightrightarrows^{\tau}_{\varphi}} \mid b, \bar{b}\right)
\end{aligned}
$$

By 1 and 2, we have that $\mathcal{M}_{\mathcal{H}}$ and $\mathcal{M}_{\mathcal{B}}$ are in bisimulation under the equivalence relation $\rightrightarrows^{\tau}_{\varphi}$, when the policies $\overline{\pi}^{\clubsuit}$ and $\overline{\pi}^{\diamondsuit}$ are respectively executed in the two models. □

Corollary D.9. *The agent behaviors, formulated through the expected return, that are obtained by executing the policies respectively in the two models are the same:* $\mathbb{E}^{\mathcal{M}_{\mathcal{H}}}_{\overline{\pi}^{\clubsuit}}\left[\sum_{t=0}^{\infty}\gamma^t \cdot \mathcal{R}_{\mathcal{H}}(a_{0:t}, o_{1:t})\right] = \mathbb{E}^{\mathcal{M}_{\mathcal{B}}}_{\overline{\pi}^{\diamondsuit}}\left[\sum_{t=0}^{\infty}\gamma^t \cdot \mathcal{R}_{\mathcal{B}}(b_t, a_t)\right]$.

Proof. Follows directly from (Larsen & Skou, 1989; Givan et al., 2003): the bisimulation relation implies the equality of the maximum expected return in the two models, where the maximum is taken over the set of all stationary policies of the two models. Since we consider MCs and not MDPs, the models are purely stochastic, so there is no nondeterministic choice linked to the choice of action, and then there is only one possible expected return, which yields the result. □

Note that we omitted the super script of $\overline{\pi}^{\clubsuit}$ in the main text; we directly considered $\overline{\pi}$ as a policy conditioned over histories, by using the exact same definition.

D.3 Existence of a Stationary Distribution over Histories

Now that we have proven that the history unfolding is equivalent to the belief MDP, we now have all the ingredients to prove Lemma D.1.

Proof. By definition of $\mathcal{M}_{\mathcal{H}}$, the execution of $\overline{\pi}^{\clubsuit}$ is guaranteed to remain an episodic process.

(a) Assume that the number of branching in the history space is finite when unrolling all possible episodes of the environment when $\overline{\pi}^{\clubsuit}$ is executed. Then, $\mathcal{M}_{\mathcal{H}}^{\overline{\pi}^{\clubsuit}}$ is finite as well. Every episodic process in MDP with countable state spaces is ergodic (Huang, 2020), there is thus a unique stationary distribution $\mathcal{H}_{\overline{\pi}^{\clubsuit}} = \lim_{t\to\infty} \xi^t_{\overline{\pi}^{\clubsuit}}(\cdot \mid \star)$ defined over the state space of $\mathcal{M}_{\mathcal{H}}$.

(b) More generally, assume now that the history space is endowed with a Borel σ-algebra over $\mathcal{S}_{\mathcal{H}}$. Then, it means that Ω is itself endowed with a Borel σ-algebra, while $\mathcal{A}$ is discrete by definition. We first show that $\mathcal{M}_{\mathcal{H}}^{\overline{\pi}\clubsuit}$ defines a recurrent Harris chain. Take $H = \{ h_{\mathsf{reset}} \}$; intuitively the event of visiting the reset state is triggered infinitely often with probability one under any policy, which ensures the history unfolding to be Harris recurrent. To see why, notice first that by Assumption 2, $\mathcal{M}_{\mathcal{H}}$ is episodic, meaning that under any policy, the environment almost surely reaches s_{reset} (and so, h_{reset} by definition of $\mathcal{M}_{\mathcal{H}}$), thus condition (i) of Definition D.6 is fulfilled. Second, since H merely consists of a singleton, we can take $\rho \in \Delta(\mathcal{S}_{\mathcal{H}})$ so that $\rho(h) = \mathbb{E}_{a \sim \overline{\pi}\clubsuit(\cdot \mid h_{\mathsf{reset}})} \mathbf{P}_{\mathcal{H}}(h \mid h_{\mathsf{reset}}, a)$, which almost surely consists in either staying in s_{reset} or transitioning from the reset state to the initial state, by Definition D.2. Then, condition (ii) of Definition D.6 is trivially satisfied. Therefore, $\mathcal{M}_{\mathcal{H}}^{\overline{\pi}\clubsuit}$ is a recurrent Harris chain, so we know there exists a stationary measure $\mathcal{H}_{\overline{\pi}\clubsuit}$ over the histories of $\mathcal{P}$ that are likely to be seen under $\overline{\pi}\clubsuit$ (Durrett, 2010, Theorem 5.8.9). Then, the *ϵ-perturbation trick* of Huang 2020 can be used as is to enforce the aperiodicity in $\mathcal{M}_{\mathcal{H}}^{\overline{\pi}\clubsuit}$. So, we have that $\xi^t_{\overline{\pi}\clubsuit}(\cdot \mid \star)$ converges to $\mathcal{H}_{\overline{\pi}\clubsuit}$ in total variation as t goes to ∞ (Durrett, 2010, Theorem 5.8.12).

The resulting stationary distribution is thus the limiting distribution over the histories of $\mathcal{P}$ when the latter operates under $\overline{\pi}$, or equivalently, the limiting distribution of the MC $\mathcal{M}_{\mathcal{B}}^{\overline{\pi}\diamond}$ by Corollary D.9. □

E Discrete Latent Variables and Temperatures

As mentioned in Section 2.3, the optimization process of the WAE-MDP relies on a temperature parameter, $\lambda \in (0, 1)$. The latter controls the continuity of the latent space learned. The purpose of the parameter is primarily to learn a discrete latent space model: precisely, we use continuous relaxation of discrete random variables (Maddison et al., 2017). This is essentially the Bernoulli version of the Gumbel softmax trick (Jang et al., 2017). The technique yields re-parameterizable and convex densities, which is compliant with stochastic gradient descent. The zero-temperature limit (i.e., passing from the continuous to the discrete setting) is theoretically enabled via simple rounding of continuous random variables. Furthermore, the logits of discrete densities are guaranteed to be identical to those of their relaxed counterparts.

Alternatively, one could use the *straight-through gradients estimator* (Bengio et al., 2013), as used by van den Oord et al. (2017); Fajtl et al. (2020) and Hafner et al. (2021). This consists in using discrete variables and non-differentiable functions in the forward pass of the input through the neural networks, while continuous variables and surrogate functions are used in the backward pass, i.e., during the backpropagation of the gradients. This yields low-variance, but biased gradients. In contrast, continuous relaxations allow to interpolate between these phenomena: a higher temperature produces low-variance but biased gradients, while lowering the temperature to zero increases the variance but ends up producing unbiased gradients.

Therefore, at any time, the discrete densities (zero-temperature limit) of the WAE-MDP are used, except when the gradients of the objective are computed, where their relaxed counterpart are used $(\lambda > 0)$. As such, the WBU training procedure follows the same principle as WAE-MDPs: using continuous relaxation of the discrete random variables (precisely, multivariate Bernoullis) when the belief loss (Eq. 8) is minimized while the actual discrete (autoregressive) density is used otherwise. We chose temperature values by following the guidelines from the original paper (Maddison et al., 2017). In the latter, it is mentioned that setting up annealing schemes (to the zero temperature limit) is a good practice but is not necessary for obtaining good results, which is confirmed experimentally in the experimental evaluation of WAE-MDPs (Delgrange et al., 2023, Appendix B.8).

F Value Difference Bounds

This section is dedicated to proving Theorems 3.2 and 3.3. Both Theorems bound the value difference of histories, in the original and latent space models via our local and belief losses, to provide model and representation quality guarantees. Before proving the Theorems, we first formally define

the *value function* of any POMDP, and then illustrate intuitively the meaning of each loss used to bound the value differences.

F.1 Value Functions

We start by formally defining the value function of any MDP.

Definition F.1 (Value function)**.** *Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathbf{P}, \mathcal{R}, s_I, \gamma \rangle$ be an MDP, and π be a policy for $\mathcal{M}$. Write $\mathcal{M}[s]$ for the MDP obtained by replacing s_I by $s \in \mathcal{S}$. Then, the value of the state $s \in \mathcal{S}$ is defined as the expected return obtained from that state by running π, i.e., $V_\pi(s) = \mathbb{E}_\pi^{\mathcal{M}[s]}\left[\sum_{t=0}^{\infty} \gamma^t \cdot \mathcal{R}(s_t, a_t)\right]$. Let $\mathcal{M}^\pi = \langle \mathcal{S}_\pi, \mathbf{P}_\pi, \mathcal{R}_\pi, s_I, \gamma \rangle$ be the Markov Chain induced by π (cf. Definition D.4). Then, the value function can be defined as the unique solution of the Bellman's equation (Puterman, 1994): $V_\pi(s) = \mathcal{R}_\pi(s) + \mathbb{E}_{s' \sim \mathbf{P}_\pi(s)}\left[\gamma \cdot V_\pi(s')\right]$. The typical goal of an RL agent is to learn a policy $\pi^\star$ that maximizes the value of the initial state of $\mathcal{M}$: $\max_{\pi^\star} V_{\pi^\star}(s_I)$.*

Property F.2 (POMDP values)**.** *We obtain the value function of any POMDP $\mathcal{P} = \langle \mathcal{M}, \Omega, \mathcal{O} \rangle$ by considering the values obtained in its belief MDP $\mathcal{M}_{\mathcal{B}} = \langle \mathcal{B}, \mathcal{A}, \mathbf{P}_{\mathcal{B}}, \mathcal{R}_{\mathcal{B}}, b_I, \gamma \rangle$. Therefore, the value of any history $h \in (\mathcal{A} \cdot \Omega)^*$ is obtained by mapping h to the belief space: let π be a policy conditioned on the beliefs of $\mathcal{P}$, then we write $V_\pi(h)$ for $V_\pi(\tau^*(h))$. Therefore, we have in particular for any latent policy $\overline{\pi} \colon \overline{\mathcal{B}} \to \Delta(\mathcal{A})$:*

$$
\begin{aligned}
V_{\overline{\pi}}(h) &= \mathbb{E}_{\overline{\pi}^\diamondsuit}^{\mathcal{M}_{\mathcal{B}}[\tau^*(h)]}\left[\sum_{t=0}^{\infty} \gamma^t \cdot \mathcal{R}_{\mathcal{B}}(b_t, a_t)\right] && \text{(cf. Lemma D.8 for definitions of } \overline{\pi}^\diamondsuit \text{ and } \overline{\pi}^\clubsuit\text{)} \\
&= \mathbb{E}_{\overline{\pi}^\clubsuit}^{\mathcal{M}_{\mathcal{H}}[h]}\left[\sum_{t=0}^{\infty} \gamma^t \cdot \mathcal{R}_{\mathcal{H}}(h_t, a_t)\right] && \text{(cf. Corollary D.9)} \\
&= \mathop{\mathbb{E}}_{a \sim \overline{\pi}^\clubsuit(\cdot|h)}\left[\mathcal{R}_{\mathcal{H}}(h, a) + \mathop{\mathbb{E}}_{h' \sim \mathbf{P}_{\mathcal{H}}(\cdot|h,a)}\left[\gamma \cdot V_{\overline{\pi}}(h')\right]\right] && \text{(by Definition F.1)} \\
&= \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot|\varphi^*(h))}\left[\mathcal{R}_{\mathcal{H}}(h, a) + \mathop{\mathbb{E}}_{h' \sim \mathbf{P}_{\mathcal{H}}(\cdot|h,a)}\left[\gamma \cdot V_{\overline{\pi}}(h')\right]\right] && \text{(by definition of } \overline{\pi}^\clubsuit\text{)} \\
&= \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot|\varphi^*(h))} \mathop{\mathbb{E}}_{s \sim \tau^*(h)}\left[\mathcal{R}(s, a) + \mathop{\mathbb{E}}_{s' \sim \mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o' \sim \mathcal{O}(\cdot|s',a)}\left[\gamma \cdot V_{\overline{\pi}}(h \cdot a \cdot o')\right]\right]. \\
& && \text{(by definition of } \mathcal{M}_{\mathcal{H}}\text{)}
\end{aligned}
$$

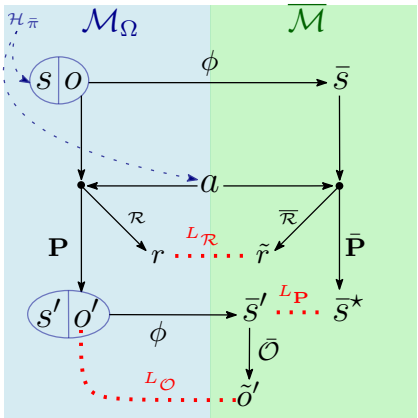
Similarly, we write $\overline{V}_{\overline{\pi}}$ for the values of a latent POMDP $\overline{\mathcal{P}}$.

F.2 Local and Belief Losses

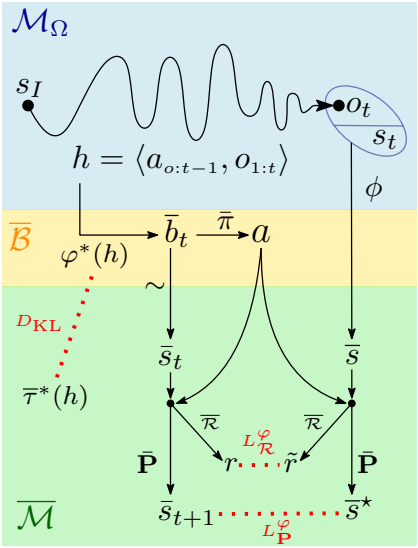
Theorems 3.2 and 3.3 involve the minimization of *local* ($L_{\mathcal{R}}, L_{\mathbf{P}}, L_{\mathcal{O}}$) and *belief* ($L_{\overline{\tau}}, L^{\varphi}_{\overline{\mathbf{P}}}, L^{\varphi}_{\overline{\mathcal{R}}}$) losses. We intuitively describe how these losses are minimized via the latent flows depicted in Fig. 6.

The procedure allowing to minimize the local losses is depicted in Fig. 6a. At each step, a state s, an observation o of s, and an action a are drawn from the distribution $\mathcal{H}_{\overline{\pi}}$ of experiences encountered while executing $\overline{\pi}$. First, $\langle s, o \rangle$ is mapped to the latent space via the state embedding function of the WAE-MDP: $\phi(s, o) = \bar{s}$. Then, the action a is executed both in the original and latent space models (respectively from $\langle s, o \rangle$ and $\bar{s}$), which allows to quantify the distance between the next reward and transition produced in the two models. Finally, the original model transitions to the next state-observation pair $\langle s', o' \rangle$, and mapping it again to the latent space through $\phi(s', o') = \bar{s}'$ allows to quantify the distance between the original observation o' and the one that is produced in the latent space model, from $\bar{s}'$ via $\bar{o}' \sim \overline{\mathcal{O}}(\cdot \mid \bar{s}')$. Notice that in practice, $\overline{\mathcal{O}}$ is not necessarily Dirac (cf. Remark 3).

The procedure allowing to minimize the belief losses is depicted in Fig. 6b. This time, the optimization is performed *on-policy*, which means that it is performed while executing the policy in the original environment. At time step $t \geqslant 1$, the current history h ends up in the observation o_t of state s_t. First, the discrepancy between the latent belief obtained via our belief encoder $\bar{b}_t = \varphi^*(h)$



(a) Optimization of the latent space model parameters (i.e., $\overline{\mathcal{R}}, \overline{\mathbf{P}}$, and $\overline{\mathcal{O}}$) by minimizing local losses.



(b) Optimization of the belief encoder φ by minimizing the (proxy) belief loss, as well as the reward and transition regularizers.

Figure 6: Latent flows used to compute the local and belief losses. Arrows represent (stochastic) mappings, the original state-observation (resp. latent state) space is spread along the blue (resp. green) area, and the latent belief space is spread along the yellow area. Distances (and discrepancies) are depicted in red. Notice that the blue area corresponds to the state-observation space $\mathcal{S}_\Omega$, which is accessible during training (Assumption 1).

and the one obtained via the true latent belief update function $\bar{b}' = \bar{\tau}^*(h)$ is evaluated. Second, we compute the reward and transition regularizers by minimizing the distance between rewards and transitions produced from believed states $\bar{s}_t \sim \bar{b}_t$ (i.e., states expected from the current belief $\bar{b}_t$) and those produced by mapping the current state-observation pair into the latent space, via $\phi(s_t, o_t) = \bar{s}$, when the action $a \sim \bar{\pi}(\cdot \mid \bar{b}_t)$ is produced. Finally, a is executed in the original environment and the process is repeated until the end of the episode.

F.3 Warm Up: Some Wasserstein Properties

In the following, we elaborate on properties and definitions related to the Wasserstein metrics that will be useful to prove the main claims. In particular, Wasserstein can be reformulated as the maximum mean discrepancy of 1-*Lipschitz functions*. The claim holds when the temperature is at the zero-limit, i.e., when the discrete variables and densities of the models are used (cf. Appendix E). This further makes the distance metric $\bar{d}$ associated with the latent state space go to the discrete metric $\mathbf{1}_{\neq} \colon \mathcal{X} \to \{1, 0\}$ (Delgrange et al., 2023), formally defined as $\mathbf{1}_{\neq}(x1, x2) = 1$ iff $x_1 \neq x_2$.

Definition F.3 (Lipschitz continuity)**.** *Let $\mathcal{X}, \mathcal{Y}$ be two measurable set and $f \colon \mathcal{X} \to \mathcal{Y}$ be a function mapping elements from $\mathcal{X}$ to $\mathcal{Y}$. If otherwise specified, we consider that f is real-valued function, i.e., $\mathcal{Y} = \mathbb{R}$. Assume that $\mathcal{X}$ is equipped with a metric $d \colon \mathcal{X} \to [0, \infty)$. Then, given a constant $K \geqslant 0$, we say that f is K-Lipschitz iff, for any $x_1, x_2 \in \mathcal{X}$, $|f(x_1) - f(x_2)| \leqslant K \cdot d(x_1, x_2)$. We write $\mathcal{F}_d^K$ for the set of K-Lipschitz functions.*

Definition F.4 (Wasserstein dual)**.** *The Kantorovich-Rubinstein duality (Villani, 2009) allows formulating the Wasserstein distance between P and Q as $\mathcal{W}_d(P, Q) = \sup_{f \in \mathcal{F}_d^1} |\mathbb{E}_{x \sim P} f(x) - \mathbb{E}_{y \sim Q} f(y)|$.*

Property F.5 (Lipschitz constant)**.** *Let $f \colon \mathcal{X} \to \mathbb{R}$, so that d is a metric on $\mathcal{X}$. Assume that f is K-Lipschitz, i.e., $f \in \mathcal{F}_d^K$, then for any two distributions $P, Q \in \Delta(\mathcal{X})$, $|\mathbb{E}_{x_1 \sim P} f(x_1) - \mathbb{E}_{x_2 \sim Q} f(x_2)| \leqslant K \cdot \mathcal{W}_d(P, Q)$.*

*In particular, for **any** bounded function $g\colon \mathcal{X} \to Y$ with $Y \subseteq \mathbb{R}$, when the distance metric associated with $\mathcal{X}$ is the discrete metric, i.e., $d = \mathbf{1}_{\neq}$, we have $|\mathbb{E}_{x_1\sim P}\, g(x_1) - \mathbb{E}_{x_2\sim Q}\, g(x_2)| \leqslant K_Y \cdot \mathcal{W}_{\mathbf{1}_{\neq}}(P,Q) = K_Y \cdot d_{TV}(P,Q)$, where $K_Y \geqslant \sup_{x\in\mathcal{X}} |g(x)|$ (see, e.g., Gelada et al. 2019, Sect. 6, for a discussion).*

Property F.5 intuitively implies the emergence of the $K_{\overline{V}}$ constant in the Theorem's inequality: we know that the latent value function is bounded by $\sup_{\bar{s},a}|\overline{\mathcal{R}}(\bar{s},a)|/1-\gamma$, so given two distributions P, Q over $\bar{\mathcal{S}}$, the maximum mean discrepancy of the latent value function is bounded by $K_{\overline{V}} \cdot \mathcal{W}_{\bar{d}}(P,Q)$ when the temperature goes to zero.

Finally, since the value difference is computed in expectation, we introduce the following useful property:

Lemma F.6 (Wasserstein in expectation)**.** *For any $f\colon \mathcal{Y}\times\mathcal{X} \to \mathbb{R}$ so that $\mathcal{X}$ is equipped with the metric d, consider the function $g_y\colon \mathcal{X} \to \mathbb{R}$ defined as $g_y(x) = f(y,x)$. Assume that for any $y \in \mathcal{Y}$, g_y is K-Lipschitz, i.e., $g_y \in \mathcal{F}_d^K$. Then, let $\mathcal{D} \in \Delta(\mathcal{Y})$ be a distribution over $\mathcal{Y}$ and $P, Q \in \Delta(\mathcal{X})$ be two distributions over $\mathcal{X}$, we have $\mathbb{E}_{y\sim\mathcal{D}}\, |\mathbb{E}_{x_1\sim P}\, f(y,x_1) - \mathbb{E}_{x_2\sim Q}\, f(y,x_2)| \leqslant K \cdot \mathcal{W}_d(P,Q)$.*

Proof. The proof is straightforward by construction of g_y:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{y\sim\mathcal{D}} \left| \mathop{\mathbb{E}}_{x_1\sim P} f(y,x_1) - \mathop{\mathbb{E}}_{x_2\sim Q} f(y,x_2) \right| \\
=& \mathop{\mathbb{E}}_{y\sim\mathcal{D}} \left| \mathop{\mathbb{E}}_{x_1\sim P} g_y(x_1) - \mathop{\mathbb{E}}_{x_2\sim Q} g_y(x_2) \right| \\
\leqslant& \mathop{\mathbb{E}}_{y\sim\mathcal{D}} \left[K \cdot \mathcal{W}_d(P,Q) \right] && \text{(by Property F.5, since } g_y \text{ is } K\text{-Lipschitz)} \\
=& K \cdot \mathcal{W}_d(P,Q)
\end{aligned}
$$

□

F.4 Model Quality Bound: Time to Raise your Expectations

Let us restate Theorem 3.2:

Theorem F.7. *Let $\mathcal{P}$, $\overline{\mathcal{P}}$, and $\overline{\pi}\colon \overline{\mathcal{B}} \to \Delta(\mathcal{A})$ be respectively the original and the latent POMDP, as well as the latent policy of Lemma D.1, so that the latent POMDP is learned through a WAE-MDP, via the minimization of the local losses $L_{\mathcal{R}}, L_{\mathbf{P}}$. Assume that the WAE-MDP is at the zero-temperature limit (i.e., $\lambda \to 0$, see Appendix E) and let $K_{\overline{\mathrm{V}}} = \|\overline{\mathcal{R}}\|_\infty/1-\gamma$, then for any such latent policy $\overline{\pi}$, the values of $\mathcal{P}$ and $\overline{\mathcal{P}}$ are guaranteed to be bounded by the local losses in average:*

$$
\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \left| V_{\overline{\pi}}(h) - \overline{V}_{\overline{\pi}}(h) \right| \leqslant \frac{L_{\mathcal{R}} + L_{\overline{\mathcal{R}}}^{\varphi} + \overline{\mathcal{R}}^{\star} L_{\overline{\tau}} + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L_{\overline{\mathbf{P}}}^{\varphi} + L_{\overline{\tau}} + L_{\mathcal{O}} \right)}{1-\gamma}. \tag{10}
$$

Proof. The plan of the proof is as follows:

1. We exploit the fact that the value function can be defined as the fixed point of the Bellman's equations;
2. We repeatedly apply the triangular and the Jenson's inequalities to end up with inequalities which reveal mean discrepancies for either rewards or value functions;
3. We exploit the fact that the temperature goes to zero to bound those discrepancies by Wasserstein (see Porperty F.5 and the related discussion);
4. The last two points allow highlighting the L_1 norm and Wasserstein terms in the local and belief losses;
5. Finally, we set up the inequalities to obtain a discounted next value difference term, and we exploit the stationary property of $\mathcal{H}_{\overline{\pi}}$ to fall back on the original, discounted, absolute value difference term;

6. Putting all together, we end up with an inequality only composed of constants, multiplied by losses that we aim at minimizing.

Concretely, the absolute value difference can be bounded by:

$$\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}}\left|V_{\overline{\pi}}(h)-\overline{V}_{\overline{\pi}}(h)\right|$$

$$= \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}}\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Big[\mathcal{R}(s,a)+\gamma\mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)}\mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)}V_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\Big]$$
$$-\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Big[\overline{\mathcal{R}}(\bar{s},a)+\gamma\mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\bar{s},a)}\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\Big]\Bigg|$$

(see Property F.2)

$$= \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\Big[\mathcal{R}(s,a)+\gamma\mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)}\mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)}V_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\Big]$$
$$-\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\Big[\overline{\mathcal{R}}(\bar{s},a)+\gamma\mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\bar{s},a)}\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\Big]\Bigg|$$

(Jensen's inequality)

$$\leqslant \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Bigg[\,\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathcal{R}(s,a)-\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\overline{\mathcal{R}}(\bar{s},a)\Bigg|$$
$$+\gamma\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)}\mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)}V_{\overline{\pi}}\big(h\cdot a\cdot o'\big)-\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\bar{s},a)}\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\Bigg|\,\Bigg]$$

(Triangular inequality)

For the sake of clarity, we split the inequality in two parts.

Part 1: Reward bounds

$$\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathcal{R}(s,a)-\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\overline{\mathcal{R}}(\bar{s},a)\Bigg|$$

$$= \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\big[\mathcal{R}(s,a)-\overline{\mathcal{R}}(\phi(s,o),a)\big]+\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\big[\overline{\mathcal{R}}(\phi(s,o),a)-\overline{\mathcal{R}}(\bar{s},a)\big]\Bigg|$$

(o is the last observation of h;
the state embedding function ϕ that links the original and latent state spaces comes into play)

$$\leqslant \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Bigg[\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\big[\mathcal{R}(s,a)-\overline{\mathcal{R}}(\phi(s,o),a)\big]\Bigg|+\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\big[\overline{\mathcal{R}}(\phi(s,o),a)-\overline{\mathcal{R}}(\bar{s},a)\big]\Bigg|\Bigg]$$

(Triangular inequality)

$$\leqslant \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Bigg[\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\big|\mathcal{R}(s,a)-\overline{\mathcal{R}}(\phi(s,o),a)\big|+\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\big[\overline{\mathcal{R}}(\phi(s,o),a)-\overline{\mathcal{R}}(\bar{s},a)\big]\Bigg|\Bigg]$$

(Jensen's inequality)

$$= \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\big|\mathcal{R}(s,a)-\overline{\mathcal{R}}(\phi(s,o),a)\big|$$
$$+\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\big[\overline{\mathcal{R}}(\phi(s,o),a)-\overline{\mathcal{R}}(\bar{s},a)\big]\Bigg|$$

$$= L_{\mathcal{R}}+\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\big[\overline{\mathcal{R}}(\phi(s,o),a)-\overline{\mathcal{R}}(\bar{s},a)\big]\Bigg| \quad \text{(by definition of } L_{\mathcal{R}}\text{)}$$

$$= L_{\mathcal{R}}+\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\mathop{\mathbb{E}}_{\bar{s}_\perp\sim\varphi^*(h)}\Big[\big[\overline{\mathcal{R}}(\phi(s,o),a)-\overline{\mathcal{R}}(\bar{s}_\perp,a)\big]+\big[\overline{\mathcal{R}}(\bar{s}_\perp,a)-\overline{\mathcal{R}}(\bar{s},a)\big]\Big]\Bigg|$$

(the belief encoder φ comes into play)

$$
\begin{aligned}
=& L_{\mathcal{R}} + \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\bar\pi}} \mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot\mid\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{\bar s\sim\varphi^*(h)} \left[\overline{\mathcal{R}}(\phi(s,o),a) - \overline{\mathcal{R}}(\bar s,a)\right] \\
&\qquad + \mathop{\mathbb{E}}_{\bar s\sim\bar\tau^*(h)} \mathop{\mathbb{E}}_{\bar s_\perp\sim\varphi^*(h)} \left[\overline{\mathcal{R}}(\bar s_\perp,a) - \overline{\mathcal{R}}(\bar s,a)\right]\Bigg| \\
\leqslant& L_{\mathcal{R}} + \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\bar\pi}} \mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot\mid\varphi^*(h))} \Bigg[\Bigg| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{\bar s\sim\varphi^*(h)} \left[\overline{\mathcal{R}}(\phi(s,o),a) - \overline{\mathcal{R}}(\bar s,a)\right]\Bigg| \\
&\qquad + \Bigg| \mathop{\mathbb{E}}_{\bar s\sim\bar\tau^*(h)} \mathop{\mathbb{E}}_{\bar s_\perp\sim\varphi^*(h)} \left[\overline{\mathcal{R}}(\bar s_\perp,a) - \overline{\mathcal{R}}(\bar s,a)\right]\Bigg|\Bigg] && (\text{ Triangular inequality}) \\
\leqslant& L_{\mathcal{R}} + \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\bar\pi}} \mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot\mid\varphi^*(h))} \Bigg[\mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{\bar s\sim\varphi^*(h)} \left|\overline{\mathcal{R}}(\phi(s,o),a) - \overline{\mathcal{R}}(\bar s,a)\right| \\
&\qquad + \Bigg| \mathop{\mathbb{E}}_{\bar s\sim\bar\tau^*(h)} \mathop{\mathbb{E}}_{\bar s_\perp\sim\varphi^*(h)} \left[\overline{\mathcal{R}}(\bar s_\perp,a) - \overline{\mathcal{R}}(\bar s,a)\right]\Bigg|\Bigg] && (\text{ Jensen's inequality}) \\
=& L_{\mathcal{R}} + \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\bar\pi}} \mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot\mid\varphi^*(h))} \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{\bar s\sim\varphi^*(h)} \left|\overline{\mathcal{R}}(\phi(s,o),a) - \overline{\mathcal{R}}(\bar s,a)\right| \\
&+ \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\bar\pi}} \mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot\mid\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{\bar s\sim\bar\tau^*(h)} \mathop{\mathbb{E}}_{\bar s_\perp\sim\varphi^*(h)} \overline{\mathcal{R}}(\bar s_\perp,a) - \overline{\mathcal{R}}(\bar s,a)\Bigg| \\
=& L_{\mathcal{R}} + L_{\overline{\mathcal{R}}}^{\varphi} + \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\bar\pi}} \mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot\mid\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{\bar s\sim\bar\tau^*(h)} \mathop{\mathbb{E}}_{\bar s_\perp\sim\varphi^*(h)} \left[\overline{\mathcal{R}}(\bar s_\perp,a) - \overline{\mathcal{R}}(\bar s,a)\right]\Bigg| \\
& && (\text{ by definition of } L_{\overline{\mathcal{R}}}^{\varphi}, \text{Eq. 5}) \\
\leqslant& L_{\mathcal{R}} + L_{\overline{\mathcal{R}}}^{\varphi} + \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\bar\pi}} \overline{\mathcal{R}}^{\star} \mathcal{W}_{\bar d}\left(\bar\tau^*(h), \varphi^*(h)\right) && (\text{ as } \lambda\to 0, \text{ by Lem. F.6 and Prop. F.5}) \\
=& L_{\mathcal{R}} + L_{\overline{\mathcal{R}}}^{\varphi} + \overline{\mathcal{R}}^{\star} L_{\bar\tau};
\end{aligned}
$$

Part 2: Next value bounds

$$
\begin{aligned}
& \gamma\cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\bar\pi}} \mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot\mid\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot\mid s,a)} \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot\mid s',a)} V_{\bar\pi}\left(h\cdot a\cdot o'\right) \\
&\qquad - \mathop{\mathbb{E}}_{\bar s\sim\bar\tau^*(h)} \mathop{\mathbb{E}}_{\bar s'\sim\overline{\mathbf{P}}(\cdot\mid\bar s,a)} \mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot\mid\bar s')} \overline{V}_{\bar\pi}\left(h\cdot a\cdot o'\right)\Bigg| \\
=\ & \gamma\cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\bar\pi}} \mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot\mid\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot\mid s,a)} \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot\mid s',a)} \Bigg[V_{\bar\pi}\left(h\cdot a\cdot o'\right) - \mathop{\mathbb{E}}_{\hat o'\sim\overline{\mathcal{O}}(\cdot\mid\phi(s',o'))} \overline{V}_{\bar\pi}\left(h\cdot a\cdot \hat o'\right)\Bigg] \\
& + \Bigg[\mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot\mid s,a)} \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot\mid s',a)} \mathop{\mathbb{E}}_{\hat o'\sim\overline{\mathcal{O}}(\cdot\mid\phi(s',o'))} \overline{V}_{\bar\pi}\left(h\cdot a\cdot \hat o'\right) \\
&\qquad - \mathop{\mathbb{E}}_{\bar s\sim\bar\tau^*(h)} \mathop{\mathbb{E}}_{\bar s'\sim\overline{\mathbf{P}}(\cdot\mid\bar s,a)} \mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot\mid\bar s')} \overline{V}_{\bar\pi}\left(h\cdot a\cdot o'\right)\Bigg]\Bigg|
\end{aligned}
$$

(the state embedding function ϕ comes into play, as well as the latent observation function $\overline{\mathcal{O}}$)

$$
\begin{aligned}
\leqslant\ & \gamma\cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\bar\pi}} \mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot\mid\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot\mid s,a)} \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot\mid s',a)} \Bigg[V_{\bar\pi}\left(h\cdot a\cdot o'\right) - \mathop{\mathbb{E}}_{\hat o'\sim\overline{\mathcal{O}}(\cdot\mid\phi(s',o'))} \overline{V}_{\bar\pi}\left(h\cdot a\cdot \hat o'\right)\Bigg]\Bigg| \\
& + \gamma\cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\bar\pi}} \mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot\mid\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot\mid s,a)} \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot\mid s',a)} \mathop{\mathbb{E}}_{\hat o'\sim\overline{\mathcal{O}}(\cdot\mid\phi(s',o'))} \overline{V}_{\bar\pi}\left(h\cdot a\cdot \hat o'\right)
\end{aligned}
$$

$$- \mathop{\mathbb{E}}_{\bar{s} \sim \bar{\tau}^*(h)} \mathop{\mathbb{E}}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o' \sim \overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\overline{\pi}}\left(h \cdot a \cdot o'\right) \Bigg|$$

(Triangular inequality)

$$\begin{aligned}
= \quad & \gamma \cdot \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot|\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s' \sim \mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o' \sim \mathcal{O}(\cdot|s',a)} \left[V_{\overline{\pi}}\left(h \cdot a \cdot o'\right) - \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot|\phi(s',o'))} \overline{V}_{\overline{\pi}}\left(h \cdot a \cdot \hat{o}'\right) \right] \Bigg| \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h, o \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot|\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \Bigg[\mathop{\mathbb{E}}_{s', o' \sim \mathbf{P}_\Omega(\cdot|s,o,a)} \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot|\phi(s',o'))} \overline{V}_{\overline{\pi}}\left(h \cdot a \cdot \hat{o}'\right) \\
& \qquad - \mathop{\mathbb{E}}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot|\phi(s,o),a)} \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\overline{\pi}}\left(h \cdot a \cdot \hat{o}'\right) \Bigg] \\
& \qquad + \Bigg[\mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot|\phi(s,o),a)} \mathop{\mathbb{E}}_{o' \sim \overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\overline{\pi}}\left(h \cdot a \cdot o'\right) \\
& \qquad - \mathop{\mathbb{E}}_{\bar{s} \sim \bar{\tau}^*(h)} \mathop{\mathbb{E}}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o' \sim \overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\overline{\pi}}\left(h \cdot a \cdot o'\right) \Bigg] \Bigg|
\end{aligned}$$

(o is the last observation of h; the latent MDP dynamics, modeled by $\overline{\mathbf{P}}$, come into play)

$$\begin{aligned}
\leqslant \quad & \gamma \cdot \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot|\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s' \sim \mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o' \sim \mathcal{O}(\cdot|s',a)} \left[V_{\overline{\pi}}\left(h \cdot a \cdot o'\right) - \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot|\phi(s',o'))} \overline{V}_{\overline{\pi}}\left(h \cdot a \cdot \hat{o}'\right) \right] \Bigg| \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h, o \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot|\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \Bigg[\mathop{\mathbb{E}}_{s', o' \sim \mathbf{P}_\Omega(\cdot|s,o,a)} \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot|\phi(s',o'))} \overline{V}_{\overline{\pi}}\left(h \cdot a \cdot \hat{o}'\right) \\
& \qquad - \mathop{\mathbb{E}}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot|\phi(s,o),a)} \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\overline{\pi}}\left(h \cdot a \cdot \hat{o}'\right) \Bigg] \Bigg| \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h, o \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot|\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot|\phi(s,o),a)} \mathop{\mathbb{E}}_{o' \sim \overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\overline{\pi}}\left(h \cdot a \cdot o'\right) \\
& \qquad - \mathop{\mathbb{E}}_{\bar{s} \sim \bar{\tau}^*(h)} \mathop{\mathbb{E}}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o' \sim \overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\overline{\pi}}\left(h \cdot a \cdot o'\right) \Bigg|
\end{aligned}$$

(Triangular inequality)

$$\begin{aligned}
\leqslant \quad & \gamma \cdot \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot|\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s' \sim \mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o' \sim \mathcal{O}(\cdot|s',a)} \left[V_{\overline{\pi}}\left(h \cdot a \cdot o'\right) - \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot|\phi(s',o'))} \overline{V}_{\overline{\pi}}\left(h \cdot a \cdot \hat{o}'\right) \right] \Bigg| \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h, o \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot|\varphi^*(h))} \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \Bigg| \mathop{\mathbb{E}}_{\bar{s}' \sim \phi\mathbf{P}_\Omega(\cdot|s,o,a)} \mathop{\mathbb{E}}_{o' \sim \overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\overline{\pi}}\left(h \cdot a \cdot o'\right) \\
& \qquad - \mathop{\mathbb{E}}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot|\phi(s,o),a)} \mathop{\mathbb{E}}_{o' \sim \overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\overline{\pi}}\left(h \cdot a \cdot o'\right) \Bigg| \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h, o \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot|\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot|\phi(s,o),a)} \mathop{\mathbb{E}}_{o' \sim \overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\overline{\pi}}\left(h \cdot a \cdot o'\right) \\
& \qquad - \mathop{\mathbb{E}}_{\bar{s} \sim \bar{\tau}^*(h)} \mathop{\mathbb{E}}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o' \sim \overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\overline{\pi}}\left(h \cdot a \cdot o'\right) \Bigg|
\end{aligned}$$

(by definition of $\phi\mathbf{P}_\Omega$ and the Jensen's inequality)

$$\begin{aligned}
\leqslant \quad & \gamma \cdot \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot|\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s' \sim \mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o' \sim \mathcal{O}(\cdot|s',a)} \left[V_{\overline{\pi}}\left(h \cdot a \cdot o'\right) - \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot|\phi(s',o'))} \overline{V}_{\overline{\pi}}\left(h \cdot a \cdot \hat{o}'\right) \right] \Bigg| \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h, o \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot|\varphi^*(h))} \mathop{\mathbb{E}}_{s \sim \tau^*(h)} K_{\overline{V}} \cdot \mathcal{W}_{\bar{d}}\left(\phi\mathbf{P}_\Omega(\cdot \mid s, a), \overline{\mathbf{P}}(\cdot \mid \phi(s,o), a)\right)
\end{aligned}$$

$$
\begin{aligned}
&+\gamma\cdot\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\phi(s,o),a)}\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\\
&\qquad\qquad-\mathop{\mathbb{E}}_{\bar{s}\sim\overline{\tau}^*(h)}\mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\bar{s},a)}\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\Bigg|\\
&\qquad\qquad\qquad\text{(as }\lambda\to 0\text{, by Lem. F.6)}\\
\leqslant\;&\gamma\cdot\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)}\mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)}\Bigg[V_{\overline{\pi}}\big(h\cdot a\cdot o'\big)-\mathop{\mathbb{E}}_{\hat{o}'\sim\overline{\mathcal{O}}(\cdot|\phi(s',o'))}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot\hat{o}'\big)\Bigg]\Bigg|\\
&+\gamma K_{\overline{V}}L_{\mathbf{P}}\\
&+\gamma\cdot\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\phi(s,o),a)}\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\\
&\qquad\qquad-\mathop{\mathbb{E}}_{\bar{s}\sim\overline{\tau}^*(h)}\mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\bar{s},a)}\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\Bigg|\\
&\qquad\qquad\qquad\text{(by definition of }L_{\mathbf{P}}\text{)}\\
=\;&\gamma\cdot\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)}\mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)}\Bigg[V_{\overline{\pi}}\big(h\cdot a\cdot o'\big)-\mathop{\mathbb{E}}_{\hat{o}'\sim\overline{\mathcal{O}}(\cdot|\phi(s',o'))}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot\hat{o}'\big)\Bigg]\Bigg|\\
&+\gamma K_{\overline{V}}L_{\mathbf{P}}\\
&+\gamma\cdot\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Bigg|\Bigg[\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\phi(s,o),a)}\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\\
&\qquad\qquad-\mathop{\mathbb{E}}_{\bar{s}\sim\varphi^*(h)}\mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\bar{s},a)}\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\Bigg]\\
&\qquad+\Bigg[\mathop{\mathbb{E}}_{\bar{s}\sim\varphi^*(h)}\mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\bar{s},a)}\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)-\mathop{\mathbb{E}}_{\bar{s}\sim\overline{\tau}^*(h)}\mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\bar{s},a)}\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\Bigg]\Bigg|\\
&\qquad\qquad\qquad\text{(the belief encoder }\varphi\text{ comes into play)}\\
\leqslant\;&\gamma\cdot\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)}\mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)}\Bigg[V_{\overline{\pi}}\big(h\cdot a\cdot o'\big)-\mathop{\mathbb{E}}_{\hat{o}'\sim\overline{\mathcal{O}}(\cdot|\phi(s',o'))}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot\hat{o}'\big)\Bigg]\Bigg|\\
&+\gamma K_{\overline{V}}L_{\mathbf{P}}\\
&+\gamma\cdot\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\phi(s,o),a)}\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\\
&\qquad\qquad-\mathop{\mathbb{E}}_{\bar{s}\sim\varphi^*(h)}\mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\bar{s},a)}\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\Bigg|\\
&+\gamma\cdot\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Bigg|\mathop{\mathbb{E}}_{\bar{s}\sim\varphi^*(h)}\mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\bar{s},a)}\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\\
&\qquad\qquad-\mathop{\mathbb{E}}_{\bar{s}\sim\overline{\tau}^*(h)}\mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\bar{s},a)}\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\Bigg|\\
&\qquad\qquad\qquad\text{(triangular inequality)}\\
\leqslant\;&\gamma\cdot\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)}\mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)}\Bigg[V_{\overline{\pi}}\big(h\cdot a\cdot o'\big)-\mathop{\mathbb{E}}_{\hat{o}'\sim\overline{\mathcal{O}}(\cdot|\phi(s',o'))}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot\hat{o}'\big)\Bigg]\Bigg|\\
&+\gamma K_{\overline{V}}L_{\mathbf{P}}\\
&+\gamma\cdot\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))}\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar{s}\sim\varphi^*(h)}\Bigg|\mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\phi(s,o),a)}\Bigg[\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')}\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
& \qquad\qquad - \mathop{\mathbb{E}}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot|\bar{s},a)} \left[\mathop{\mathbb{E}}_{o' \sim \overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\bar{\pi}}\left(h \cdot a \cdot o'\right) \right] \Bigg| \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\bar{\pi}}} \mathop{\mathbb{E}}_{a \sim \bar{\pi}(\cdot|\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{\bar{s} \sim \varphi^*(h)} \mathop{\mathbb{E}}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o' \sim \overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\bar{\pi}}\left(h \cdot a \cdot o'\right) \\
& \qquad - \mathop{\mathbb{E}}_{\bar{s} \sim \bar{\tau}^*(h)} \mathop{\mathbb{E}}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o' \sim \overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\bar{\pi}}\left(h \cdot a \cdot o'\right) \Bigg| \qquad \text{(Jensen's inequality)} \\
\leqslant\ & \gamma \cdot \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\bar{\pi}}} \mathop{\mathbb{E}}_{a \sim \bar{\pi}(\cdot|\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s' \sim \mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o' \sim \mathcal{O}(\cdot|s',a)} \left[V_{\bar{\pi}}\left(h \cdot a \cdot o'\right) - \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot|\phi(s',o'))} \overline{V}_{\bar{\pi}}\left(h \cdot a \cdot \hat{o}'\right) \right] \Bigg| \\
& + \gamma K_{\overline{V}} L_{\mathbf{P}} \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h,o \sim \mathcal{H}_{\bar{\pi}}} \mathop{\mathbb{E}}_{a \sim \bar{\pi}(\cdot|\varphi^*(h))} \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{\bar{s} \sim \varphi^*(h)} K_{\overline{V}} \mathcal{W}_{\bar{d}}\left(\overline{\mathbf{P}}(\cdot \mid \phi(s,o),a), \overline{\mathbf{P}}(\cdot \mid \bar{s},a)\right) \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\bar{\pi}}} \mathop{\mathbb{E}}_{a \sim \bar{\pi}(\cdot|\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{\bar{s} \sim \varphi^*(h)} \mathop{\mathbb{E}}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o' \sim \overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\bar{\pi}}\left(h \cdot a \cdot o'\right) \\
& \qquad - \mathop{\mathbb{E}}_{\bar{s} \sim \bar{\tau}^*(h)} \mathop{\mathbb{E}}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o' \sim \overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\bar{\pi}}\left(h \cdot a \cdot o'\right) \Bigg| \qquad \text{(as } \lambda \to 0\text{, by Lem. F.6)} \\
=\ & \gamma \cdot \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\bar{\pi}}} \mathop{\mathbb{E}}_{a \sim \bar{\pi}(\cdot|\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s' \sim \mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o' \sim \mathcal{O}(\cdot|s',a)} \left[V_{\bar{\pi}}\left(h \cdot a \cdot o'\right) - \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot|\phi(s',o'))} \overline{V}_{\bar{\pi}}\left(h \cdot a \cdot \hat{o}'\right) \right] \Bigg| \\
& + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}}\right) \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\bar{\pi}}} \mathop{\mathbb{E}}_{a \sim \bar{\pi}(\cdot|\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{\bar{s} \sim \varphi^*(h)} \left[\mathop{\mathbb{E}}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o' \sim \overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\bar{\pi}}\left(h \cdot a \cdot o'\right) \right] \\
& \qquad - \mathop{\mathbb{E}}_{\bar{s} \sim \bar{\tau}^*(h)} \left[\mathop{\mathbb{E}}_{\bar{s}' \sim \overline{\mathbf{P}}(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o' \sim \overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\bar{\pi}}\left(h \cdot a \cdot o'\right) \right] \Bigg| \qquad \text{(by definition of } L^{\varphi}_{\overline{\mathbf{P}}}\text{, Eq. 5)} \\
\leqslant\ & \gamma \cdot \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\bar{\pi}}} \mathop{\mathbb{E}}_{a \sim \bar{\pi}(\cdot|\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s' \sim \mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o' \sim \mathcal{O}(\cdot|s',a)} \left[V_{\bar{\pi}}\left(h \cdot a \cdot o'\right) - \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot|\phi(s',o'))} \overline{V}_{\bar{\pi}}\left(h \cdot a \cdot \hat{o}'\right) \right] \Bigg| \\
& + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}}\right) \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\bar{\pi}}} \mathop{\mathbb{E}}_{a \sim \bar{\pi}(\cdot|\varphi^*(h))} K_{\overline{V}} \mathcal{W}_{\bar{d}}\left(\bar{\tau}^*(h), \varphi^*(h)\right)
\end{aligned}
$$

(as $\lambda \to 0$, by Lem. F.6; note that Wasserstein is symmetric since it is a distance metric (Villani, 2009))

$$
\begin{aligned}
\leqslant\ & \gamma \cdot \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\bar{\pi}}} \mathop{\mathbb{E}}_{a \sim \bar{\pi}(\cdot|\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s' \sim \mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o' \sim \mathcal{O}(\cdot|s',a)} \left[V_{\bar{\pi}}\left(h \cdot a \cdot o'\right) - \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot|\phi(s',o'))} \overline{V}_{\bar{\pi}}\left(h \cdot a \cdot \hat{o}'\right) \right] \Bigg| \\
& + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\bar{\tau}}\right) \qquad \text{(by definition of } L_{\bar{\tau}}\text{, Eq. 5)} \\
=\ & \gamma \cdot \mathop{\mathbb{E}}_{h,o \sim \mathcal{H}_{\bar{\pi}}} \mathop{\mathbb{E}}_{a \sim \bar{\pi}(\cdot|\varphi^*(h))} \Bigg| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s',o' \sim \mathbf{P}_{\Omega}(\cdot|s,o,a)} \Bigg[\left(V_{\bar{\pi}}\left(h \cdot a \cdot o'\right) - \overline{V}_{\bar{\pi}}\left(h \cdot a \cdot o'\right)\right) \\
& \qquad + \left(\overline{V}_{\bar{\pi}}\left(h \cdot a \cdot o'\right) - \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot|\phi(s',o'))} \overline{V}_{\bar{\pi}}\left(h \cdot a \cdot \hat{o}'\right) \right) \Bigg] \Bigg| \\
& + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\bar{\tau}}\right)
\end{aligned}
$$

$$
\begin{aligned}
\leqslant\ & \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s',o'\sim\mathbf{P}_\Omega(\cdot|s,o,a)} \left[V_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) \right] \right| \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s',o'\sim\mathbf{P}_\Omega(\cdot|s,o,a)} \left[\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{\mathcal{O}}(\cdot|\phi(s',o'))} \overline{V}_{\overline{\pi}}\big(h\cdot a\cdot \hat{o}'\big) \right] \right| \\
& + \gamma K_{\overline{V}} \cdot \Big(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} \Big) \qquad \text{(triangular inequality)} \\
\leqslant\ & \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s',o'\sim\mathbf{P}_\Omega(\cdot|s,o,a)} \left[V_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) \right] \right| \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))} \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \left| \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)} \left[\overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{\mathcal{O}}(\cdot|\phi(s',o'))} \overline{V}_{\overline{\pi}}\big(h\cdot a\cdot \hat{o}'\big) \right] \right| \\
& + \gamma K_{\overline{V}} \cdot \Big(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} \Big) \qquad \text{(by definition of } \mathbf{P}_\Omega \text{ and the Jensen's inequality)} \\
\leqslant\ & \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s',o'\sim\mathbf{P}_\Omega(\cdot|s,o,a)} \left[V_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) \right] \right| \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))} \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} K_{\overline{V}} d_{TV}\Big(\mathcal{O}\big(\cdot \mid s',a\big), \mathop{\mathbb{E}}_{o'\sim s',a} \overline{\mathcal{O}}\big(\cdot \mid \phi\big(s',o'\big)\big) \Big) \\
& + \gamma K_{\overline{V}} \cdot \Big(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} \Big) \qquad \text{(cf. Prop. F.5 and Lem F.6)} \\
=\ & \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s',o'\sim\mathbf{P}_\Omega(\cdot|s,o,a)} \left[V_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) \right] \right| \\
& + \gamma K_{\overline{V}} \cdot \Big(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} + L_{\mathcal{O}} \Big) \qquad \text{(by definition of } L_{\mathcal{O}}\text{, Eq. 4)} \\
\leqslant\ & \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))} \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s',o'\sim\mathbf{P}_\Omega(\cdot|s,o,a)} \left| V_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) \right| \\
& + \gamma K_{\overline{V}} \cdot \Big(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} + L_{\mathcal{O}} \Big) \qquad \text{(Jensen's inequality)} \\
=\ & \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \left| V_{\overline{\pi}}(h) - \overline{V}_{\overline{\pi}}(h) \right| + \gamma K_{\overline{V}} \cdot \Big(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} + L_{\mathcal{O}} \Big) \\
& \qquad \text{(}\mathcal{H}_{\overline{\pi}} \text{ is a stationary distribution (Lem. D.1) which allows us to apply the stationary property)}
\end{aligned}
$$

Putting all together. To recap, by Part 1 and 2, we have:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \left| V_{\overline{\pi}}(h) - \overline{V}_{\overline{\pi}}(h) \right| &\leqslant L_{\mathcal{R}} + L^{\varphi}_{\overline{\mathcal{R}}} + \overline{\mathcal{R}}^{\star} L_{\overline{\tau}} + \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \left| V_{\overline{\pi}}(h) - \overline{V}_{\overline{\pi}}(h) \right| \\
&\quad + \gamma K_{\overline{V}} \cdot \Big(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} + L_{\mathcal{O}} \Big) \\
\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \left| V_{\overline{\pi}}(h) - \overline{V}_{\overline{\pi}}(h) \right| \cdot (1-\gamma) &\leqslant L_{\mathcal{R}} + L^{\varphi}_{\overline{\mathcal{R}}} + \overline{\mathcal{R}}^{\star} L_{\overline{\tau}} + \gamma K_{\overline{V}} \cdot \Big(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} + L_{\mathcal{O}} \Big) \\
\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \left| V_{\overline{\pi}}(h) - \overline{V}_{\overline{\pi}}(h) \right| &\leqslant \frac{L_{\mathcal{R}} + L^{\varphi}_{\overline{\mathcal{R}}} + \overline{\mathcal{R}}^{\star} L_{\overline{\tau}} + \gamma K_{\overline{V}} \cdot \Big(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} + L_{\mathcal{O}} \Big)}{1-\gamma}
\end{aligned}
$$

which finally concludes the proof. □

F.5 Representation Quality Bound

We start by showing that the optimal *latent* value function is *almost* Lipschitz continuous in the latent belief space. Coupled with Theorem F.7, this result allows to show that whenever *two pairs of histories are encoded to close representations, their values (i.e., the return obtained from that history points) are guaranteed to be close as well* whenever the losses introduced in Sec. 3.2 are minimized and go to zero. Phrased differently, this Theorem ensures that the representation induced by our encoder is suitable to optimize the value function since the distance between beliefs in the latent space characterizes the distance of behaviors of the agent in the original environment. The latent

belief space thus captures the necessary information to learn a policy that optimizes the expected return.

Definition F.8 (Almost Lipschitzness)**.** *Let $\mathcal{X}$ be a measurable set equipped with a metric $d\colon \mathcal{X} \to [0, \infty)$ and $f\colon \mathcal{X} \to \mathbb{R}$. We say that f is* almost *Lipschitz continuous (e.g., Vanderbei 1991) iff for all $\epsilon > 0$, there is a constant $K \geqslant 0$ so that $|f(x_1) - f(x_2)| \leqslant Kd(x_1, x_2) + \epsilon$ for any $x_1, x_2 \in \mathcal{X}$.*

Notation 1 (Optimal value function). For any MDP $\mathcal{M}$, let $\pi^\star$ be an optimal policy of $\mathcal{M}$, then we write $V^\star$ for $V_{\pi^\star}$ (see Property F.2 for the values of a POMDP).

Lemma F.9. *Let $\mathcal{P} = \langle \mathcal{M}, \Omega, \mathcal{O} \rangle$ be a POMDP with underlying MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathbf{P}, \mathcal{R}, s_I, \gamma \rangle$. Assume that $\mathcal{P}$ is discrete, i.e., $\mathcal{S}$, $\mathcal{A}$, and Ω are finite sets. Then, $V^\star$ is almost Lipschitz continuous.*

Proof. Define $\mathcal{V}$ as the set of real-valued bounded functions $V\colon \mathcal{B} \to \mathbb{R}$ and $U\colon \mathcal{B} \times \mathcal{A} \times \mathcal{V} \to \mathcal{V}$ as

$$U(b, a, V) = \mathcal{R}_{\mathcal{B}}(b, a) + \mathop{\mathbb{E}}_{b' \sim \mathbf{P}_{\mathcal{B}}(\cdot|b,a)} \left[\gamma V\left(b'\right)\right].$$

The Bellman update operator is defined as $\mathcal{U}\colon \mathcal{V} \to \mathcal{V}$ as $(\mathcal{U}V)(b) = \max_{a \in \mathcal{A}} U(b, a, V)$ and is an isotone mapping that is a contraction under the supremum norm with fixed point $V^\star$, i.e., $V^\star = \mathcal{U}V^\star$ (Puterman, 1994; Hauskrecht, 2000; Sutton & Barto, 1998). Furthermore, for any initial value function $V_0 \in \mathcal{V}$, the sequence resulting from value iteration (VI), $V_{i+1} = \mathcal{U}V_i$, converges to $V^\star$ (with linear convergence rate γ Puterman (1994)): for any $\epsilon' > 0$, there is a $i \in \mathbb{N}$ so that for all $j \geqslant i$, $\left\|V_j - V^\star\right\|_\infty \leqslant \epsilon'$. Now, let $\epsilon > 0$; in particular, the latter statement holds for $\epsilon' = \epsilon/2$. Since the convergence of VI holds for any initial value, we assume that $V_0 \in \mathcal{V}$ has been chosen as a *piecewise linear convex* (PWLC) function. Then, for all $i \geqslant 0$, V_i is also PWLC (Sondik, 1978; Smallwood & Sondik, 1973; Hauskrecht, 2000). Since $\mathcal{S}$ is discrete, the belief space $\mathcal{B}$ is the standard $|\mathcal{S}|$-dimensional simplex, so the domain of V_i is compact, meaning that it is defined as a finite collection of linear functions. Thus, V_i is also $2K$-Lipschitz: one just need take $2K$ as the higher slope of these functions (in absolute value). In consequence, for any pair of beliefs $b_1, b_2 \in \mathcal{B}$,

$$\begin{aligned}
&|V^\star(b_1) - V^\star(b_2)| \\
=&|V^\star(b_1) - V_i(b_1) + V_i(b_1) - V_i(b_2) + V_i(b_2) - V^\star(b_2)| \\
\leqslant&|V^\star(b_1) - V_i(b_1)| + |V_i(b_1) - V_i(b_2)| + |V_i(b_2) - V^\star(b_2)| && \text{(Triangular inequality)} \\
\leqslant&2\epsilon' + |V_i(b_1) - V_i(b_2)| && \text{(by the convergence of VI)} \\
=&\epsilon + |V_i(b_1) - V_i(b_2)| \\
\leqslant&\epsilon + K \cdot d_{TV}(b_1, b_2), && \text{(since } d_{TV}(b_1, b_2) = {}^1\!/\!_2 \left\|b_1 - b_2\right\|_1\text{)}
\end{aligned}$$

which means that $V^\star$ is almost Lipschitz, by definition. □

Corollary F.10. *When the temperature of the WAE-MDP (see Appendix E) and the variance of $\overline{\mathcal{O}}$ go to zero (see Remark 3), the optimal latent value function of $\overline{\mathcal{P}}$ is almost Lipschitz-continuous.*

Proof. Assuming the WAE-MDP temperature goes to zero, the state space of $\overline{\mathcal{P}}$ is discrete, $\bar{d} = \mathbf{1}_{\neq}$, and $\mathcal{W}_{\bar{d}} = d_{TV}$. Furthermore, $\overline{\mathcal{O}}$ is deterministic as its variance goes to zero; therefore the set of observations of $\overline{\mathcal{P}}$ can be limited to the set of images of $\overline{\mathcal{O}}_\mu$, which is finite since $\bar{\mathcal{S}}$ is finite. Then Lemma F.9 can be applied. □

Theorem F.11. *Let $\overline{\pi}^\star$ be an optimal policy of the POMDP $\overline{\mathcal{P}}$. Then, for any $\epsilon > 0$, there is a constant $K \geqslant 0$ so that for any pair of measurable histories under the original and latent models h_1, h_2 that are mapped to latent beliefs through $\varphi^*(h_1) = \bar{b}_1$ and $\varphi^*(h_2) = \bar{b}_2$, the belief representation induced by φ almost surely yields:*

$$\begin{aligned}
&|V_{\overline{\pi}^\star}(h_1) - V_{\overline{\pi}^\star}(h_2)| \leqslant K\mathcal{W}_{\bar{d}}\left(\bar{b}_1, \bar{b}_2\right) + \epsilon + \\
&\frac{L_{\mathcal{R}} + L_{\overline{\mathcal{R}}}^{\varphi} + \left(K + \gamma K_{\overline{V}} + \overline{\mathcal{R}}^\star\right) L_{\overline{\tau}} + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L_{\overline{\mathbf{P}}}^{\varphi} + L_{\mathcal{O}}\right)}{1 - \gamma} \left(\mathcal{H}_{\overline{\pi}^\star}(h_1)^{-1} + \mathcal{H}_{\overline{\pi}^\star}(h_2)^{-1}\right)
\end{aligned}$$

when the WAE-MDP temperature as well as the variance of the observation decoder go to zero.

Proof. First, observe that for any measurable history h, $\left|V_{\overline{\pi}^\star}(h) - \overline{V}_{\overline{\pi}^\star}(h)\right| \leqslant \mathcal{H}_{\overline{\pi}^\star}(h)^{-1} \cdot \mathbb{E}_{h' \sim \mathcal{H}_{\overline{\pi}^\star}} \left|V_{\overline{\pi}^\star}(h') - \overline{V}_{\overline{\pi}^\star}(h')\right|$ (cf. Gelada et al. 2019). Therefore, we have:

$$
\begin{aligned}
&\left|V_{\overline{\pi}^\star}(h_1) - V_{\overline{\pi}^\star}(h_2)\right| \\
=&\left|V_{\overline{\pi}^\star}(h_1) - \overline{V}_{\overline{\pi}^\star}(h_1) + \overline{V}_{\overline{\pi}^\star}(h_1) - \overline{V}_{\overline{\pi}^\star}(h_2) + \overline{V}_{\overline{\pi}^\star}(h_2) - V_{\overline{\pi}^\star}(h_2)\right| \\
\leqslant&\left|V_{\overline{\pi}^\star}(h_1) - \overline{V}_{\overline{\pi}^\star}(h_1)\right| + \left|\overline{V}_{\overline{\pi}^\star}(h_1) - \overline{V}_{\overline{\pi}^\star}(h_2)\right| + \left|\overline{V}_{\overline{\pi}^\star}(h_2) - V_{\overline{\pi}^\star}(h_2)\right| \quad \text{(Triangular inequality)} \\
\leqslant&\mathcal{H}_{\overline{\pi}^\star}(h_1)^{-1} \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\overline{\pi}^\star}} \left|V_{\overline{\pi}^\star}(h) - \overline{V}_{\overline{\pi}^\star}(h)\right| + \left|\overline{V}_{\overline{\pi}^\star}(h_1) - \overline{V}_{\overline{\pi}^\star}(h_2)\right| \\
&+ \mathcal{H}_{\overline{\pi}^\star}(h_2)^{-1} \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\overline{\pi}^\star}} \left|V_{\overline{\pi}^\star}(h) - \overline{V}_{\overline{\pi}^\star}(h)\right| \\
\leqslant&\left|\overline{V}_{\overline{\pi}^\star}(h_1) - \overline{V}_{\overline{\pi}^\star}(h_2)\right| \\
&+ \frac{L_{\mathcal{R}} + L^{\varphi}_{\mathcal{R}} + \overline{\mathcal{R}}^\star L_{\overline{\tau}} + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\mathbf{P}} + L_{\overline{\tau}} + L_{\mathcal{O}}\right)}{1-\gamma} \left(\mathcal{H}_{\overline{\pi}^\star}(h_1)^{-1} + \mathcal{H}_{\overline{\pi}^\star}(h_2)^{-1}\right) \\
&\hfill \text{(Thm. 3.2)}
\end{aligned}
$$

Consider $h_\perp = \arg\min\left\{ \mathcal{H}_{\overline{\pi}^\star}(h) \colon h \in \operatorname{supp}(\mathcal{H}_{\overline{\pi}^\star}) \cap \mathbf{H}\left(\mathcal{P}, \overline{\mathcal{P}}\right) \right\}$, where $\operatorname{supp}(\mathcal{H}_{\overline{\pi}^\star})$ denotes the support of $\mathcal{H}_{\overline{\pi}^\star}$ and $\mathbf{H}\left(\mathcal{P}, \overline{\mathcal{P}}\right)$ the set of histories compliant for $\mathcal{P}$ and $\overline{\mathcal{P}}$ (i.e., histories whose observations are in the intersection of the reachable parts of the original and latent observation spaces). Notice that $h_\perp$ exists almost surely, i.e., with probability one. First, $\mathcal{H}_{\overline{\pi}}$, which is defined over histories from the history unfolding (cf. Section D.3), is solely defined over episodes, i.e., histories for which the POMDP did not restart following a reset (Section D.2), so we only need consider such histories. By Assumption 2, the reset state is almost surely triggered, which means that the length of histories ending in $o^\star$ (the "reset" observation, cf. Definition D.2) is also almost surely finite (otherwise, the POMDP does not reset and this event has a zero probability). Second, we consider histories which are measurable in the two models (i.e., the original and latent POMDPs) and the observation space of the latent POMDP is finite (cf. the Proof of Corollary F.10). So, the set of histories from which we extract the minimum is almost surely finite.

Now, let $\epsilon > 0$. Recall that the latent value function (defined over the latent belief space) is almost Lipschitz continuous (Corollary F.10). In particular, for $\delta = \frac{\epsilon}{\left(1 + 2\mathcal{H}_{\overline{\pi}^\star}(h_\perp)^{-1}\right)}$, there is a $K \geqslant 0$ so that for any $\bar{b}, \bar{b}' \in \overline{\mathcal{B}}$, $\left|\overline{V}^\star(b) - \overline{V}^\star(b')\right| \leqslant K \mathcal{W}_{\bar{d}}\left(b, b'\right) + \delta$. Then:

$$
\begin{aligned}
&\left|\overline{V}^\star(h_1) - \overline{V}^\star(h_2)\right| \\
=&\left|\overline{V}^\star(\bar{\tau}^*(h_1)) - \overline{V}^\star(\bar{\tau}^*(h_2))\right| \\
=&\left|\overline{V}^\star(\bar{\tau}^*(h_1)) - \overline{V}^\star(\varphi^*(h_1)) + \overline{V}^\star(\varphi^*(h_1)) - \overline{V}^\star(\varphi^*(h_2)) + \overline{V}^\star(\varphi^*(h_2)) - \overline{V}^\star(\bar{\tau}^*(h_2))\right| \\
\leqslant&\left|\overline{V}^\star(\bar{\tau}^*(h_1)) - \overline{V}^\star(\varphi^*(h_1))\right| + \left|\overline{V}^\star(\varphi^*(h_1)) - \overline{V}^\star(\varphi^*(h_2))\right| + \left|\overline{V}^\star(\varphi^*(h_2)) - \overline{V}^\star(\bar{\tau}^*(h_2))\right| \\
&\hfill \text{(Triangular inequality)} \\
\leqslant&\mathcal{H}_{\overline{\pi}^\star}(h_1)^{-1} \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\overline{\pi}^\star}} \left|\overline{V}^\star(\bar{\tau}^*(h)) - \overline{V}^\star(\varphi^*(h))\right| + \left|\overline{V}^\star(b_1) - \overline{V}^\star(b_2)\right| \\
&+ \mathcal{H}_{\overline{\pi}^\star}(h_2)^{-1} \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\overline{\pi}^\star}} \left|\overline{V}^\star(\bar{\tau}^*(h)) - \overline{V}^\star(\varphi^*(h))\right| \\
=&\left(\mathcal{H}_{\overline{\pi}^\star}(h_1)^{-1} + \mathcal{H}_{\overline{\pi}^\star}(h_2)^{-1}\right) \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\overline{\pi}^\star}} \left|\overline{V}^\star(\bar{\tau}^*(h)) - \overline{V}^\star(\varphi^*(h))\right| + \left|\overline{V}^\star(b_1) - \overline{V}^\star(b_2)\right| \\
\leqslant&\left(\mathcal{H}_{\overline{\pi}^\star}(h_1)^{-1} + \mathcal{H}_{\overline{\pi}^\star}(h_2)^{-1}\right) \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\overline{\pi}^\star}} \left[K \mathcal{W}_{\bar{d}}\left(\bar{\tau}^*(h), \varphi^*(h)\right) + \delta\right] + K \mathcal{W}_{\bar{b}}\left(\bar{b}_1, \bar{b}_2\right) + \delta \\
&\hfill (\overline{V}^\star \text{ is almost Lipschitz}) \\
=&\left(\mathcal{H}_{\overline{\pi}^\star}(h_1)^{-1} + \mathcal{H}_{\overline{\pi}^\star}(h_2)^{-1}\right)(K L_{\bar{\tau}} + \delta) + K \mathcal{W}_{\bar{b}}\left(\bar{b}_1, \bar{b}_2\right) + \delta \qquad \text{(by definition of } L_{\bar{\tau}}) \\
=&\left(\mathcal{H}_{\overline{\pi}^\star}(h_1)^{-1} + \mathcal{H}_{\overline{\pi}^\star}(h_2)^{-1}\right) K L_{\bar{\tau}} + K \mathcal{W}_{\bar{b}}\left(\bar{b}_1, \bar{b}_2\right) + \delta\left(1 + \mathcal{H}_{\overline{\pi}^\star}(h_1)^{-1} + \mathcal{H}_{\overline{\pi}^\star}(h_2)^{-1}\right) \\
\leqslant&\left(\mathcal{H}_{\overline{\pi}^\star}(h_1)^{-1} + \mathcal{H}_{\overline{\pi}^\star}(h_2)^{-1}\right) K L_{\bar{\tau}} + K \mathcal{W}_{\bar{b}}\left(\bar{b}_1, \bar{b}_2\right) + \delta\left(1 + 2\mathcal{H}_{\overline{\pi}^\star}(h_\perp)^{-1}\right) \\
=&\left(\mathcal{H}_{\overline{\pi}^\star}(h_1)^{-1} + \mathcal{H}_{\overline{\pi}^\star}(h_2)^{-1}\right) K L_{\bar{\tau}} + K \mathcal{W}_{\bar{b}}\left(\bar{b}_1, \bar{b}_2\right) + \epsilon
\end{aligned}
$$

Putting all together, we have that for any $\epsilon > 0$, there almost surely exists a constant $K \geqslant 0$ so that:

$$|V_{\overline{\pi}^\star}(h_1) - V_{\overline{\pi}^\star}(h_2)| \leqslant K\mathcal{W}_{\bar{d}}\left(\bar{b}_1, \bar{b}_2\right) + \epsilon +$$
$$\frac{L_{\mathcal{R}} + L^{\varphi}_{\overline{\mathcal{R}}} + \left(K + \gamma K_{\overline{V}} + \overline{\mathcal{R}}^\star\right) L_{\overline{\tau}} + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\mathcal{O}}\right)}{1-\gamma} \left(\mathcal{H}_{\overline{\pi}^\star}(h_1)^{-1} + \mathcal{H}_{\overline{\pi}^\star}(h_2)^{-1}\right).$$

□

G Additional Details on Back Propagation Through Time

In this section, we elaborate on why BPTT is not required, nor recommended, to learn a representation of the belief when using our algorithm.

In RNN-A2C (Figure 2, right) the hidden state is computed recursively, which is *in the same spirit* as how the belief is computed through our WBU (Figure 2, left). The motivation behind activating BPTT for RNN-A2C and disabling it for WBU lies in the difference between beliefs and hidden states and what they represent.

On the one hand, beliefs are state distributions which can be inferred from the previous belief and be themselves seen as Markovian states (in the belief MDP $\mathcal{M}_{\mathcal{B}}$). Therefore, one can see the belief update function as the transition function of an MDP, where an observation is additionally provided. Disabling BPTT is thus consistent with the literature of model based RL for learning Markovian transition functions (e.g., Gelada et al., 2019; François-Lavet et al., 2019; Delgrange et al., 2022). Notably, our goal is not targeted towards learning directly a history representation (from sequences) but rather on learning to reproduce the belief update rule, based on the latent stochastic transition function $\overline{\mathbf{P}}$, and observation function $\overline{\mathcal{O}}$ of the latent model. The latter are learned separately and considered fixed when learning the belief update (Section 4). Therefore, stopping the gradients does not produce approximate gradients. Furthermore, the error linked to this recursive update can be fully determined through the error of our latent model (Theorem 3.2). Then, enabling BPTT increases the risk of accumulating errors across time.

On the other hand, RNN hidden states are latent representations of the history and it is up to the RL algorithm to learn a useful representation of the history for the control task. This is in stark contrast to our approach: we separately learn a representation which is guaranteed to be well suited for the RL agent, and we use it as input of the policy. In RNN-A2C the main objective is to find a policy that optimizes the return conditioned on the full history, making it sound to use BPTT in that case. By contrast, in WBU, the RL agent uses the belief as a given state-signal to condition the value function and the policy, and therefore, the RL objective is not backpropagated through the belief encoder.

Finally, it is worth mentioning that as beliefs are computed with a forward recurrence, errors may accumulate over time. This means that if the belief b_{t+k} diverges significantly from its correct value, then b_{t+k+1} will diverge even more. Then, allowing gradients from the learning of b_{t+k} to be computed based on an incorrect b_{t+k+1} and further flow to b_t may hinders the learning procedure.

H Algorithm

We describe the final WBU learning procedure in Algorithm 1. Note that they keyword **Update** means that we compute the gradients of the input loss, and update the parameters of the neural networks of the pointed function/model accordingly.

Normalizing term. Given the set of parameters ι of φ, we minimize the KL divergence D_{KL} by gradient descent on the Monte-Carlo estimate of the divergence:

$$\nabla_\iota D_{\mathrm{KL}}\big(\varphi\big(\bar{b}_t, a_t, o_{t+1}; \iota\big) \parallel \bar{\tau}\big(\bar{b}_t, a_t, o_{t+1}\big)\big) =$$
$$\nabla_\iota \mathop{\mathbb{E}}_{\bar{s}_{t+1} \sim \varphi\left(\bar{b}_t, a_t, o_{t+1}; \iota\right)} \left[\log \varphi\left(\bar{s}_{t+1} \mid \bar{b}_t, a_t, o_{t+1}; \iota\right) - \log \mathop{\mathbb{E}}_{\bar{s} \sim \bar{b}_t} \overline{\mathbf{P}}\left(\bar{s}_{t+1} \mid \bar{s}, a_t\right) - \log \overline{\mathcal{O}}\left(o_{t+1} \mid \bar{s}_{t+1}\right)\right]$$

Notice that the first term of the divergence (the belief normalization term) of Eq. 8 does not depend on φ and thus yields zero gradient. Nevertheless, we observed during our experiments that adding the normalizing term allows to stabilize and reduce the variance of the belief loss.

Optimizing Wasserstein. To optimize the Wasserstein term of the belief losses, we follow the same learning procedure than (Delgrange et al., 2023, Appenix A.5): we introduce neural networks $\mathcal{F}_\spadesuit$ (for $\spadesuit \in \{ \overline{\mathbf{P}}, \Omega \}$) that are trained to attain the supremum of the dual formulation of the Wasserstein distance. To do so, we need enforce the Lipschitzness of $\mathcal{F}_\spadesuit$ and, as in WAE-MDPs (Delgrange et al., 2023), we do so via the gradient penalty approach of Gulrajani et al. (2017), leveraging that any differentiable function is 1-Lipschitz iff it has gradients with norm at most 1 everywhere. Finally, notice that we do not directly optimize the total variation distance of $L_{\mathcal{O}}$, but rather the Wasserstein; we take the usual Euclidean distance as metric over Ω which is proven to be Lipschitz equivalent to a distance converging to the discrete metric as the temperature of the WAE-MDP goes to zero (Delgrange et al., 2023, Appendix A.6) to finally recover d_{TV}.

I HYPERPARAMETERS

Table 1 provides the range of hyperparameters used in the search, along with the selected values for each environment. The hyperparameter search was performed using OPTUNA (Akiba et al., 2019). Pre-training of the WAE-MDP involved collecting 10240 transitions with a random policy and performing 200 training steps. These pre-training transitions are taken into account in the reported results.

Additionally, Table 2 (for R-A2C) and Table 3 (for DVRL) present the specific hyperparameters used for each algorithm. A grid search was conducted over all possible combinations for both baselines. The hidden size of all neural networks was set to 128 neurons and two hidden layers (except for the sub-belief encoder, and the policy which uses one) without further tuning. The experiments were carried out using 16 parallel environments. The original implementation of DVRL and their version of R-A2C were used in this study.

We ran the experiments on a cluster composed of Intel Xeon Gold 6148 CPU.

In the following tables the enviromnents are represented as follows: STATELESSCARTPOLE $\heartsuit$, NOISYSTATELESSCARTPOLE $\clubsuit$, REPEATPREVIOUS $\diamondsuit$, SPACEINVADERS $\spadesuit$, NOISYSPACEINVADERS $\bigstar$.

The t-SNE (van der Maaten & Hinton, 2008) presented in the main text, in Fig. 4, was obtained using cuML (Raschka et al., 2020). Table 4 reports the hyperparameters that were used. We used 50,000 sub-belief-value pairs, sampled from a dataset of 1 million timesteps collected at the end of the training.

Algorithm 1: WASSERSTEIN BELIEF UPDATER

Input: Batch sizes B_{WBU}, B_{WAE}, B_{NEXT}; global learning steps N; no. of model updates per iteration N_{MODEL}; your favorite collect strategy π_{init}; replay buffer (RB) $\mathcal{D}$; Lipschitz networks $\mathcal{F}_{\overline{\mathbf{P}}} \colon \bar{\mathcal{S}} \to \mathbb{R}$, $\mathcal{F}_{\Omega} \colon \Omega \to \mathbb{R}$; observation variance network $\overline{\mathcal{O}}_{\sigma}$; and loss weights $w_{\overline{\mathcal{R}}}, w_{\overline{\mathbf{P}}}$

collect $\theta = \{ s_i, o_i, a_i, r_i, s'_i, o'_i \}_{i=1}^{N_{\text{init}}}$ by executing π_{init} for N_{init} steps; **store** θ in $\mathcal{D}$

▷ *Use the exploration policy π_{init} to collect transitions and initialize the RB*

repeat N **times**

  repeat N_{MODEL} **times**

    ▷ *Update the WAE-MDP model for N_{MODEL} consecutive training steps*

    for $i \leftarrow 1$ *to* B_{WAE} **do**

      $\langle s_i, o_i, a_i, r_i, s'_i, o'_i \rangle \sim \mathcal{D}$ ▷ *Sample a transition from the RB*

      $\bar{s}' \leftarrow \phi(s'_i, o'_i)$ ▷ *Embed $\langle s'_i, o'_i \rangle$ to the latent space*

      $\tilde{o}'_i \sim \overline{\mathcal{O}}(\cdot \mid \bar{s}')$ ▷ *Observe the resulting latent state via $\overline{\mathcal{O}}$*

    $\mathcal{L}_{\text{WAE}} \leftarrow$ **compute the WAE-MDP loss** on transition batch $\{ s_i, o_i, a_i, r_i, s'_i, o'_i \}_{i=1}^{B_{\text{WAE}}}$

    Update the WAE-MDP components (*in particular, those of Eq. 2*) by minimizing $\mathcal{L}_{\text{WAE}}$

    $\mathcal{L}_{\mathcal{O}} \leftarrow {}^{1}\!/_{B_{\text{WAE}}} \cdot \sum_{i=1}^{B_{\text{WAE}}} [\mathcal{F}_{\Omega}(o'_i) - \mathcal{F}_{\Omega}(\tilde{o}'_i)]$ ▷ *Observation loss*

    Update $\mathcal{F}_{\Omega}$ by maximizing $\mathcal{L}_{\mathcal{O}}$ **and** enforcing its 1-Lipschizness w.r.t. metric d_{Ω}

    Update $\overline{\mathcal{O}}_{\sigma}$ by minimizing $\mathcal{L}_{\mathcal{O}}$

  for $i \leftarrow 1$ *to* B_{WBU} **do**

    $s_0 \leftarrow s_I$; $\bar{s}_0 \leftarrow \bar{s}_I$; $\bar{b}_0 \leftarrow \delta_{\bar{s}_0}$; $\beta_0 \leftarrow \beta_I$ ▷ *β_I is arbitrary, e.g., zeroes*

    for $t \leftarrow 0$ *to* T **do**

      $a_t \sim \bar{\pi}(\cdot \mid \beta_t)$ ▷ *Produce the action a_t according to the sub-belief β_t*

      execute a in the environment, **receive reward** r_t, and **perceive** the next state-observation $\langle s_{t+1}, o_{t+1} \rangle$

      store the transition $\langle s_t, o_t, a_t, r_t, s_{t+1}, o_{t+1} \rangle$ into $\mathcal{D}$

      $\beta_{t+1} \leftarrow \varphi^{sub}(\text{sg}(\beta_t), a_t, o_{t+1})$ ▷ *Update the sub-belief;* sg *is stop gradients*

      $\bar{b}_{t+1} \leftarrow \mathbb{M}(\beta_{t+1})$ ▷ *Retrieve the belief distribution b_{t+1} via the MAF $\mathbb{M}$*

      $\bar{s}_{t+1} \sim \bar{b}_{t+1}$ ▷ ***Believe** the next latent state*

      ▷ *Marginalize the next latent state distribution w.r.t. the current belief*

      for $j \leftarrow 1$ *to* B_{NEXT} **do**

        $\bar{s} \sim \bar{b}_t$; $\mathcal{L}^{j}_{\log \overline{\mathbf{P}}} \leftarrow \left[\log \overline{\mathbf{P}}(\bar{s}_{t+1} \mid \bar{s}, a_t) - \log B_{\text{NEXT}}\right]$

      $\mathcal{L}^{i,t}_{\text{KL}} \leftarrow \log \bar{b}_{t+1}(\bar{s}_{t+1}) - \text{LSE}(\{ \mathcal{L}^{j}_{\log \overline{\mathbf{P}}} \}_{j=1}^{B_{\text{NEXT}}}) - \log \overline{\mathcal{O}}(o_{t+1} \mid \bar{s}_{t+1})$

        ▷ *Pointwise decomposition of Eq. 8: divergence with the belief update rule*

      $\mathcal{L}^{i,t}_{\overline{\mathcal{R}}} \leftarrow \left| \overline{\mathcal{R}}(\phi(s_t, o_t), a_t) - \overline{\mathcal{R}}(\bar{s}_t, a_t) \right|$ ▷ *Latent reward regularizer*

      $\bar{s}' \sim \overline{\mathbf{P}}(\cdot \mid \bar{s}_t, a_t)$

        ▷ *Transition to the next latent state from the current believed latent state*

      $\mathcal{L}^{i,t}_{\overline{\mathbf{P}}} \leftarrow [\mathcal{F}_{\overline{\mathbf{P}}}(\phi(s_{t+1}, o_{t+1})) - \mathcal{F}_{\overline{\mathbf{P}}}(\bar{s}')]$ ▷ *Latent transition regularizer*

  Update $\mathcal{F}_{\overline{\mathbf{P}}}$ by maximizing $\sum_{i=1}^{B_{\text{WBU}}} \sum_{t=0}^{T-1} \mathcal{L}^{i,t}_{\overline{\mathbf{P}}}$ **and** enforcing its 1-Lipschitzness w.r.t. latent metric $\bar{d}$

  Update φ^{sub} and $\mathbb{M}$ by minimizing ${}^{1}\!/_{(B_{\text{WBU}}+T)} \sum_{i=1}^{B_{\text{WBU}}} \sum_{t=0}^{T-1} (\mathcal{L}^{i,t}_{\text{KL}} + w_{\overline{\mathcal{R}}} \cdot \mathcal{L}^{i,t}_{\overline{\mathcal{R}}} + w_{\overline{\mathbf{P}}} \cdot \mathcal{L}^{i,t}_{\overline{\mathbf{P}}})$

  Update $\bar{\pi}$ by minimizing the A2C loss on the batch $\{ \beta^{i}_{0:T}, a^{i}_{0:T-1}, r^{i}_{0:T-1} \}_{i=1}^{B_{\text{WBU}}}$

function $\overline{\mathcal{O}}(\cdot \mid \bar{s})$ ▷ (*smooth*) *Observation Filter*

  $\mu \leftarrow \overline{\mathcal{O}}_{\mu}(\bar{s}); \sigma \leftarrow \overline{\mathcal{O}}_{\sigma}(\bar{s})$ ▷ *Decode $\bar{s}$; get the standard deviation of the reconstruction*

  return $\mathcal{N}(\mu, \sigma^2)$

Table 1: Range of hyperparameter search and selection per environment.

	Range	♡	♣	♢	♠	★
WAE Updates per Belief Update	1-2	1	1	2	2	2
Activation function	leaky relu, elu	leaky relu	leaky relu	elu	elu	relu
Activation function lipshitz	leaky relu, tanh	tanh	tanh	tanh	tanh	leaky relu
Activation CNN	leaky relu, elu				elu	elu
Policy config						
Learning rate	1.e-4, 3.e-4, 5.e-4	3.e-4	1.e-4	1.e-4	5.e-4	5.e-4
Clip norm	1, 10	10	10	10	10	10
Belief config						
Same opt as policy	Yes, No	Yes	Yes	Yes	No	No
Learning rate	1.e-4, 3.e-4, 5.e-4				1.e-4	3.e-4
Loss factor	1.e-1, 1.e-2, 1.e-3, 1.e-4, 1.e-5	1.e-1	1.e-4	1.e-5		
Filter variance min	1.e-2, 1.e-3	1.e-2	1.e-2	1.e-3	1.e-2	1.e-2
Normalize log obs filter	True, False	False	False	True	True	True
Sub belief prior temperature	.5, 0.66, .75, .9, .99	0.66	0.9	0.66	0.99	0.99
Reward loss scale factor	0.1, 1., 10., 20., 50., 100.	0.1	20	100	100	100
Transition loss scale factor	0.1, 1., 10., 20., 50., 100.	50	100	50	50	100
Buffer size	4096, 8192, 16384	4096	4096	16384	4096	4096
N Critic Update	5, 10	10	10	5	5	5
N State samples	16, 32, 64	32	64	32	32	32
N next state samples	16, 32, 64	32	64	64	16	32
WAE config						
Latent state size	♡, ♣ : $5 \rightarrow 10$, Others: $18 \rightarrow 25$	5	6	20	18	19
Minimizer learning rate	1.e-3, 3.e-4,1.e-4, 5.e-5, 1.e-5	5.e-5	5.e-5	3.e-4	1.e-5	1.e-5
Maximizer learning rate	1.e-3, 3.e-4,1.e-4, 5.e-5, 1.e-5	5.e-5	5.e-5	1.e-3	3.e-4	1.e-4
State encoder temperature	0.33, .5, 0.66, .75, .9, .99	0.33	0.75	0.66	0.5	0.5
State prior temperature	0.33, .5, 0.66, .75, .9, .99	0.75	0.75	0.5	0.75	0.75
Local transition loss scaling	10., 25., 50., 75., 80.	50	80	80	10	10
Steady state scaling	10., 25., 50., 75., 80.	80	80	100	25	25
N critic update	5, 10	10	10	5	5	10
Batch size	128, 256	128	128	256	128	128
Clip grad	1, 10. 100.	100	100	100	10	10
State vs Obs reconstruction weight	1, 2	2	2	2	1	1
Obs reg. min learning rate	1.e-3, 3.e-4,1.e-4, 5.e-5, 1.e-5	5.e-5	1.e-4	1.e-3	5.e-5	5.e-5
Obs reg. max learning rate	1.e-3, 3.e-4,1.e-4, 5.e-5, 1.e-5	5.e-5	3.e-4	5.e-5	1.e-5	3.e-4
Obs reg. gradient penalty	50, 100, 500, 1000	50	50	1000	100	1000

Table 2: R-A2C hyperparameters

	Range	♡	♣	♢	♠	★
Opitmizer	Adam, RMSProp	RMSProp	RMSProp	Adam	RMSProp	RMSProp
Clip grad norm	0.5, 1., 10.	1.0	10	10.	10	10
Learning rate	3.e-5, 1.e-4, 3.e-4, 5.e-4, 1.e-3	5.e-4	1.e-4	5.e-4	1.e-3	1.e-3

Table 3: DVRL hyperparameters

	Range	♡	♣	♢	♠	★
Optimizer	Adam, RMSProp	Adam	Adam	RMSProp	RMSProp	RMSProp
Clip grad norm	0.5, 1., 10.	10.0	1.0	1.0	10	0.5
Learning rate	3.e-5, 1.e-4, 3.e-4, 5.e-4, 1.e-3	5.e-4	1.e-3	3.e-4	1.e-3	1.e-3
Encoding loss factor	1,.1,.5,.05	0.5	1.0	1.0	1.0	0.05
Number of particles	5,10,15	10	3	10	5	5

Table 4: t-SNE hyperparameters

Method	FFT
Perplexity	10
N Neighbors	50
Metric	Manhattan
Learning rate	2400
Early exaggeration	50
Late exaggeration	3
Exaggeration iter	300

J INDEX OF NOTATIONS

Divergences

D_{KL} KL divergence

d_{TV} Total variation distance

$\mathcal{W}_d$ Wasserstein distance

Extended POMDP

$\mathcal{O}^{\uparrow}\colon \mathcal{S}_\Omega \to \Omega$ Extended observation deterministic function (projection on the observation space)

$\mathcal{P}^{\uparrow}$ Extended POMDP

$\mathcal{M}_\Omega$ Extended MDP

$\mathbf{P}_\Omega\colon \mathcal{S}_\Omega \times \mathcal{A} \to \Delta(\mathcal{S}_\Omega)$ Extended probability transition

$\mathcal{R}_\Omega\colon \mathcal{S}_\Omega \times \mathcal{A} \to \mathbb{R}$ Extended reward function

$\mathcal{S}_\Omega = \mathcal{S} \times \Omega$ Extended state space

Losses

$L_{\bar{\tau}}$ Belief loss

$L_{\mathcal{R}}$ Local reward loss

$L_{\mathbf{P}}$ Local transition loss

$\mathcal{L}$ Combined local loss: $\mathcal{L} = L_{\mathcal{R}} + L^{\varphi}_{\overline{\mathcal{R}}} + \overline{\mathcal{R}}^{\star} L_{\bar{\tau}} + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\bar{\tau}} + L_{\mathcal{O}} \right)$

$L_{\mathcal{O}}$ Observation loss

$L^{\varphi}_{\overline{\mathcal{R}}}$ On-policy reward loss

$L^{\varphi}_{\overline{\mathbf{P}}}$ On-policy transition loss

MDP

$a, \mathcal{A}$ Action, action space

$\gamma \in [0, 1)$ Discount factor

$\mathcal{M}$ MDP

$\mathbf{P}\colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ Transition function

$r, \mathcal{R}$ Reward, reward function

$s, \mathcal{S}$ State, state space

POMDP

$b, \mathcal{B}$ Belief, belief space

$\tau\colon \mathcal{B} \times \mathcal{A} \times \Omega \to \mathcal{B}$ Belief update function

$\tau^*\colon (\mathcal{A} \cdot \Omega)^* \to \mathcal{B}$ Recursive application of the belief update rule along histories

$h, (\mathcal{A} \cdot \Omega)^*$ History, history space

$\mathcal{H}_\pi$ Stationary distribution over histories based on a policy π

o, Ω Observation, observation space

$\mathcal{O}\colon \mathcal{S} \times \mathcal{A} \to \Delta(\Omega)$ Observation stochastic function

$\pi\colon \mathcal{B} \to \Delta(\mathcal{A})$ POMDP policy conditioned on beliefs

$\mathcal{P}$ POMDP

V_π Value function for the policy π

Latent (PO)MDP

$\varphi\colon \overline{\mathcal{B}} \times \mathcal{A} \times \Omega \to \overline{\mathcal{B}}$ Belief encoder

$\varphi^*\colon (\mathcal{A} \cdot \Omega)^* \to \overline{\mathcal{B}}$ Recursive application of the belief encoder along histories

φ^{sub} Sub-belief encoder, so that $\mathbb{M} \circ \varphi^{\text{sub}}(\beta, a, o) = \varphi(\mathbb{M}(\beta), a, o)$

β sub-belief, so that $\mathbb{M}(\beta) = \bar{b} \in \overline{\mathcal{B}}$

$K_{\overline{V}}$ $\frac{\overline{\mathcal{R}}^{\star}}{1-\gamma}$

$\bar{b}, \overline{\mathcal{B}}$ Latent belief, latent belief space

$\bar{\tau} \colon \overline{\mathcal{B}} \times \mathcal{A} \times \Omega \to \overline{\mathcal{B}}$ Latent belief update

$\bar{\tau}^* \colon (\mathcal{A} \cdot \Omega)^* \to \overline{\mathcal{B}}$ Recursive application of the latent belief update rule along histories

$\overline{\mathcal{M}}$ Latent MDP

$\overline{\mathcal{O}} : \bar{\mathcal{S}} \to \Delta(\Omega)$ Latent observation function

$\bar{\pi} \colon \overline{\mathcal{B}} \to \Delta(\mathcal{A})$ POMDP latent policy conditioned on latent beliefs

$\overline{\mathcal{P}}$ Latent POMDP

$\overline{\mathbf{P}} \colon \bar{\mathcal{S}} \times \mathcal{A} \to \Delta(\bar{\mathcal{S}})$ Latent probability transition

$\bar{s}, \bar{\mathcal{S}}$ Latent state, latent state space

$\overline{V}_{\bar{\pi}}$ Latent value function for the policy $\bar{\pi}$

$\mathbb{M}$ MAF; allows to pass from sub-beliefs to latent beliefs $\bar{b}$ via $\mathbb{M}(\beta) = \bar{b} \in \overline{\mathcal{B}}$

$\bar{r}, \overline{\mathcal{R}}$ Latent reward, latent reward function

$\overline{\mathcal{R}}^{\star}$ $\left\|\overline{\mathcal{R}}\right\|_{\infty} = \sup_{\bar{s} \in \bar{\mathcal{S}}, a \in \mathcal{A}} \left|\overline{\mathcal{R}}(\bar{s}, a)\right|$

WAE-MDP Components

$\psi \colon \bar{\mathcal{S}} \to \mathcal{S} \times \Omega$ Extended state decoder, defined as $\psi(\bar{s}) = \langle \psi^{\mathcal{S}}(\bar{s}), \overline{\mathcal{O}}_{\mu}(\bar{s}) \rangle$

$\psi^{\mathcal{S}} \colon \bar{\mathcal{S}} \to \mathcal{S}$ State decoder

$\phi \colon \mathcal{S}_{\Omega} \to \bar{\mathcal{S}}$ State embedding function

$\overline{\mathcal{O}}_{\mu} \colon \bar{\mathcal{S}} \to \Omega$ Deterministic observation decoder

ADDITIONAL REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. *CoRR*, abs/1907.10902, 2019. URL http://arxiv.org/abs/1907.10902.

K. B. Athreya and P. Ney. A new approach to the limit theory of recurrent markov chains. *Transactions of the American Mathematical Society*, 245:493–501, 1978. ISSN 00029947. URL http://www.jstor.org/stable/1998882.

Christel Baier and Joost-Pieter Katoen. *Principles of model checking*. MIT Press, 2008. ISBN 978-0-262-02649-9.

Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *CoRR*, abs/1308.3432, 2013. URL http://arxiv.org/abs/1308.3432.

Luca de Alfaro, Thomas A. Henzinger, and Rupak Majumdar. Discounting the future in systems theory. In Jos C. M. Baeten, Jan Karel Lenstra, Joachim Parrow, and Gerhard J. Woeginger (eds.), *Automata, Languages and Programming, 30th International Colloquium, ICALP 2003, Eindhoven, The Netherlands, June 30 - July 4, 2003. Proceedings*, volume 2719 of *Lecture Notes in Computer Science*, pp. 1022–1037. Springer, 2003. doi: 10.1007/3-540-45061-0_79. URL https://doi.org/10.1007/3-540-45061-0_79.

Josée Desharnais, Vineet Gupta, Radha Jagadeesan, and Prakash Panangaden. Metrics for labelled markov processes. *Theor. Comput. Sci.*, 318(3):323–354, 2004. doi: 10.1016/j.tcs.2003.09.013. URL https://doi.org/10.1016/j.tcs.2003.09.013.

Rick Durrett. *Probability: Theory and Examples, 4th Edition*. Cambridge University Press, 2010. ISBN 9780511779398. doi: 10.1017/CBO9780511779398. URL https://doi.org/10.1017/CBO9780511779398.

Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Latent bernoulli autoencoder. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2964–2974. PMLR, 2020. URL http://proceedings.mlr.press/v119/fajtl20a.html.

Robert Givan, Thomas L. Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artif. Intell.*, 147(1-2):163–223, 2003. doi: 10.1016/S0004-3702(02)00376-4. URL https://doi.org/10.1016/S0004-3702(02)00376-4.

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

Ruiquan Huang, Yingbin Liang, and Jing Yang. Provably efficient UCB-type algorithms for learning predictive state representations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jId5PXbBbX.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 11 2017. URL https://arxiv.org/abs/1611.01144.

Kim Guldstrand Larsen and Arne Skou. Bisimulation through probabilistic testing. *Inf. Comput.*, 94 (1):1–28, 1991. doi: 10.1016/0890-5401(91)90030-6. URL https://doi.org/10.1016/0890-5401(91)90030-6.

Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? In Po-Ling Loh and Maxim Raginsky (eds.), *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pp. 5175–5220. PMLR, 2022. URL https://proceedings.mlr.press/v178/liu22f.html.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=S1jE5L5gl.

D. Revuz. *Markov Chains*. North-Holland mathematical library. Elsevier Science Publishers B.V., second (revised) edition, 1984. ISBN 9780444864000. URL https://books.google.be/books?id=zIDWlAEACAAJ.

Erfan SeyedSalehi, Nima Akbarzadeh, Amit Sinha, and Aditya Mahajan. Approximate information state based convergence analysis of recurrent q-learning. In *Sixteenth European Workshop on Reinforcement Learning*, 2023. URL https://openreview.net/forum?id=co8W7pvSrS.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6306–6315, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html.

Robert J. Vanderbei. Uniform continuity is almost lipschitz continuity. Technical Report SOR-91–11, Statistics and Operations Research Series, Princeton University, 1991.