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 (BE) [2] University of Antwerp (BE)
[3] Flanders Make (BE) [4] Urban Innovation and R&D, City of Amsterdam (NL)
{raphael.avalos, florent.delgrange}@vub.be

Abstract

Partially Observable Markov Decision Processes (POMDPs) are useful tools to model environments where the full state cannot be perceived by an agent. As such the agent needs to reason taking into account the past observations and actions. However, simply remembering the full history is generally intractable due to the exponential growth in the history space. Keeping a probability distribution that models the belief over what the true state is can be used as a sufficient statistic of the history, but its computation requires access to the model of the environment and is also intractable. State-of-the-art algorithms use Recurrent Neural Networks to compress the observation-action history aiming to learn a sufficient statistic, but 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. Our approach comes with theoretical guarantees on the quality of our approximation ensuring that our outputted beliefs allow for learning the optimal value function.

1 Introduction

Partially Observable Markov Decision Processes (POMDPs) [37] define a powerful framework for modeling decision-making in uncertain environments where the state is not fully observable. These problems are a common occurrence in many real-world applications, such as robotics [26], and recommendation systems [41]. In contrast to in a *Markov Decision Process* (MDP), in a POMDP, the agent observes a noisy observation of the state that does not suffice as a signal to condition an optimal policy on. As such, optimal policies need to take the entire action-observation 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's history. The beliefs are a sufficient statistic of the history for control [23] and, when used as states, define a *belief MDP* equivalent to the original POMDP [3]. While closed-form expressions of the belief exists they requires access to a model of the environment. The computation is also in general intractable, as it requires to integrate over the full state space and therefore only applicable to smaller problems.

To overcome those challenges, SOTA algorithms focus on compressing the history into a fixed-size vector with the help of *Recurrent Neural Networks* (RNNs) [19]. However, compressing the history using RNNs can lead to loss of information, resulting in suboptimal policies. To improve the likelihood of obtaining a sufficient statistic, RNNs can be combined with regularization techniques. These techniques include generative models [7, 17, 18], particle filtering [22, 28], and predicting

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

16th European Workshop on Reinforcement Learning (EWRL 2023).

distant observations [15, 16]. It is important to note that *none of these techniques guarantee that the representation of histories induced by RNNs is suitable for optimizing the return* Additionally, a limitation of many algorithms is their assumption that beliefs are simple distributions like Gaussian distributions, which limits their applicability [15, 27, 18].

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 (partially observable) latent space model whose behaviors (expressed as expected returns) are close to the original model. Furthermore, we show that WBU is guaranteed to induce a suitable representation of the history to optimize the return. WBU is composed of three components that are learned in a round-robin fashion: the model, the belief learner, and the policy (Fig. 1). As histories are not enough to learn the full environment model, 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 also be applied in real-world settings such as robotics, where additional sensors can be used during training in a laboratory setting. In *multi-agent* RL, using additional information, such as the state, during training is a common practice [32, 4].

We learn the latent model of the POMDP via a *Wasserstein auto-encoded MDP* (WAE-MDP) [10]. We then learn the *belief update network* (BUN) by minimizing the Wasserstein distance with the exact belief update rule in the latent POMDP through a tractable variational proxy. To allow for complex belief distributions, we use *Normalizing Flows* [24]. Unlike the current SOTA algorithms, the beliefs are only optimized towards accurately representing the current state distribution and following the belief update rule. While we use a recursive network in our belief update architecture we do not back-propagate through time and therefore implement it as a simple feed forward network. The policy is then learned on the latent belief space by using as input a vector embedding the parameters of the belief (*sub-beliefs*).

We note that using Normalizing Flows for the belief distribution as been experienced in FORBES [7]. However, FORBES does not condition its policy on the beliefs but rather on sample latent states, which is sub-optimal as it approximates the state distribution with one sample.

Our contributions are two-fold. First, we present WBU, a novel algorithm that approximates the belief update of a learned latent environment from *any* POMDP, and allows the learning of a policy conditioned on those beliefs. Second, we provide theoretical guarantees ensuring that our latent belief learner, on top of learning the dynamics of the POMDP and replicating the belief update function, outputs a belief encoding suitable for learning the value function. Our experimental results are promising as they show that our algorithm is able to learn to encode the history into a representation useful to learn a policy, without using RNNs.

Other Related Work. Some other works also focus on specific types of POMDPs, such as building compact latent representation of images for visual motor tasks [27], or environment where the observation are masked states with Gaussian noise [40]. While accessing the state is common in partially observable deep multi-agent RL, it is not a common practice in single-agent but has already been explored in kernel-POMDPs [31] that uses the states to build models based on RKHSs.

2 Background

2.1 Probability Distributions and Discrepancy Measures

We write $\Sigma(\mathcal{X})$ for the set of all Borel 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 $P, Q \in \Delta(\mathcal{X})$, the divergence between P and Q can be measured according to the following discrepancies:

- 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* [39], 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.
- the *Kullback-Leibler* (KL) divergence, defined as $D_{\mathrm{KL}}(P, Q) = \mathbb{E}_{x\sim P}\left[\log(P(x)/Q(x))\right]$.
- 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 measure.

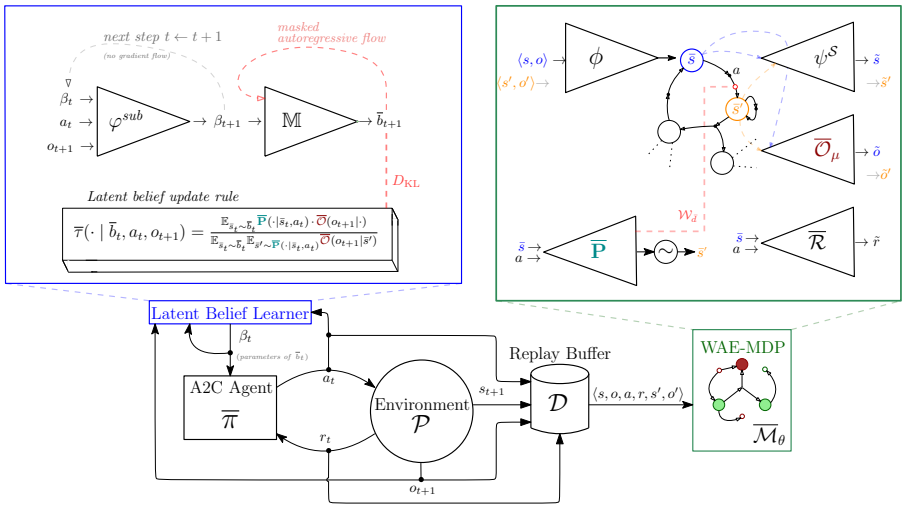


Figure 1: High-level picture of our *WBU framework*. The *WAE-MDP* is presented in Sect. 3, and the *Latent Belief Learner* is presented in Sect. 4. Learning the different components is performed in a round-robin fashion. The WAE-MDP learns from data collected by the RL agent and stored in a Replay Buffer.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

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. Any policy π and $\mathcal{M}$ induce a unique probability measure $\mathbb{P}_\pi^{\mathcal{M}}$ on the Borel σ-algebra over (measurable) infinite trajectories [34]. 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 may drop the superscript when the context is clear.

Partially Observable MDPs (POMDPs) [37] 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, stationary policies that are based solely on the current observation of $\mathcal{P}$ *do not induce any probability space* on trajectories of $\mathcal{M}$. Intuitively, due to the partial observability of the current state $s_t \in \mathcal{S}$ at each interaction step $t \geqslant 0$, the agent must take into account full histories in order to infer the distribution of rewards accumulated up to the current time step t, and make an informed decision on its next action $a_t \in \mathcal{A}$. Alternatively, the agent can maintain a *belief* $b_t \in \Delta(\mathcal{S}) = \mathcal{B}$ over the current state of $\mathcal{M}$ [42]. 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 [35]. 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[2] *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 $\mathcal{P}$ [3].

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) [9] 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 corresponding to a discrete latent state space. This procedure guarantees $\overline{\mathcal{M}}$ to be probably approximately *bisimilarly close* [25, 14, 10] 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 [11, 12]. 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

An RL agent does not have explicit access to the environment dynamics. Instead, it can reinforce its behaviors through its interactions and experiences without having direct access to the environment transition, reward, and observation functions. In this setting, the agent is assumed to operate within a partially observable environment. The key of our approach lies in *granting the RL agent access to the true state of the environment during its training*, while *its perception of the environment is only limited to actions and observations when the learned policy is finally deployed*. Therefore, 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. The latter serves as a teacher for the agent to make it learn how to perform its belief updates. The trick to learn this latent space model is to reason on an equivalent POMDP, where the underlying MDP is refined to encode all the crucial dynamics. We further demonstrate that the resulting model is guaranteed to closely replicate the original environment behavior when the agent interacts with it.

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 now deterministic and defined as the 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

[2] even if $\mathcal{S}$ is finite, there is an infinite, uncountable number of measures in $\Delta(\mathcal{S}) = \mathcal{B}$.

[6]: $\mathcal{P}^{\uparrow}$ captures the stochasticity of $\mathcal{O}$ in the transition function through the refinement of the state space, further resulting in 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 for the learning of the observation dynamics through $\overline{\mathbf{P}}$, but it also enables the learning of the deterministic observation function $\bar{\mathcal{O}}_\mu$ 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 $\bar{\mathcal{O}}_\mu \colon \overline{\mathcal{S}} \to \Omega$, which yield $\psi(\bar{s}) = \langle \psi^{\mathcal{S}}(\bar{s}), \bar{\mathcal{O}}_\mu(\bar{s}) \rangle$. This way, the WAE-MDP procedure 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 the observation function outputs a normal distribution centered in $\bar{\mathcal{O}}_\mu$: $\overline{\mathcal{O}}(\cdot \mid \bar{s}) = \mathcal{N}(\bar{\mathcal{O}}_\mu(\bar{s}), \sigma^2)$. Note that the deterministic function is retrieved as the variance approaches zero. However, it is worth mentioning that the smoothness of $\overline{\mathcal{O}}$ is favorable for gradient descent when learning distributions, unlike Dirac measures (see Eq. 3 below). As with any POMDP, the belief update function $\bar{\tau}$ of $\overline{\mathcal{P}}$ allows to reason on the belief space to optimize the expected return. Formally, at any time step $t \geqslant 0$ of the interaction with latent belief $\bar{b}_t \in \Delta(\overline{\mathcal{S}}) = \overline{\mathcal{B}}$, the latent belief update is given by $\bar{b}_{t+1} = \bar{\tau}(\bar{b}_t, a_t, o_{t+1})$ when a_t is executed and o_{t+1} is observed iff, for any 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 $h \in (\mathcal{A} \cdot \Omega)^*$, executing a latent policy $\bar{\pi} \colon \overline{\mathcal{B}} \to \Delta(\mathcal{A})$ in $\mathcal{P}$ is possible by converting h into a belief $\bar{\tau}^*(h) = \bar{b}$ over the latent state space and executing the action given by $\bar{\pi}(\cdot \mid \bar{b})$. Training $\overline{\mathcal{M}}$ gives access to the dynamics that compute the belief through the closed form of the updater $\bar{\tau}$ (Eq. 3). However, the integration 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

$$D(\bar{\tau}^*(h), \varphi^*(h)) \tag{4}$$

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 main questions arise: *"Does the latent POMDP induced by our WAE-MDP encoding yields a model whose behaviors are close to $\mathcal{P}$?"* and *"Is the history representation induced by φ suitable to optimize the expected return in $\mathcal{P}$?"*. Clearly, the obtained guarantees depend on the history distribution and chosen discrepancy. The following section provides a detailed theoretical analysis of the required distribution and losses to achieve these learning guarantees.

3.2 Losses and Theoretical Guarantees

To provide the guarantees, we assume that histories drawn from the interaction follow an *episodic RL process*: the environment $\mathcal{P}$ is assumed to embed 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 is a well defined probability distribution $\mathcal{H}_{\bar{\pi}} \in \Delta\big((\mathcal{A} \cdot \Omega)^*\big)$ over histories likely to be perceived at the limit by the agent when it executes $\bar{\pi}$ in $\mathcal{P}$* (proof in Appendix B).

Local losses. The objective function of the WAE-MDP incorporates *local losses* [13] that minimize the expected distance between the original and latent reward and transition functions:

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

and both are optimized *locally*, i.e., under $\mathcal{H}_{\bar{\pi}}$, where $s, o, a \sim \mathcal{H}_{\bar{\pi}}$ is a shorthand for (i) $h \sim \mathcal{H}_{\bar{\pi}}$ so that o is the last observation of h, (ii) $s \sim \tau^*(h)$, and (iii) $a \sim \bar{\pi}(\cdot \mid \varphi^*(h))$. 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 $\overline{\mathcal{O}}$:

$$L_{\mathcal{O}} = \mathop{\mathbb{E}}_{s,o,a\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} d_{TV}\left(\mathcal{O}\big(\cdot \mid s',a\big), \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)} \overline{\mathcal{O}}\big(\cdot \mid \phi\big(s',o'\big)\big)\right); \tag{5}$$

$L_{\mathcal{O}}$ provides a way to gauge the variation between the observations generated in the latent space and those actually observed, which allows to set the variance of $\overline{\mathcal{O}}$ (while $\bar{\mathcal{O}}_\mu$ allows to set its mean).

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 $\overline{\mathcal{P}}$:

$$L_{\overline{\tau}} = \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathcal{W}_{\bar{d}}\left(\overline{\tau}^*(h), \varphi^*(h)\right), \quad L^{\varphi}_{\overline{\mathcal{R}}} = \mathop{\mathbb{E}}_{h,s,o,a\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{\bar{s}\sim\varphi^*(h)} \left|\overline{\mathcal{R}}(\phi(s,o),a) - \overline{\mathcal{R}}(\bar{s},a)\right|,$$
$$L^{\varphi}_{\overline{\mathbf{P}}} = \mathop{\mathbb{E}}_{h,s,o,a\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{\bar{s}\sim\varphi^*(h)} \mathcal{W}_{\bar{d}}\left(\overline{\mathbf{P}}(\cdot \mid \phi(s,o),a), \overline{\mathbf{P}}(\cdot \mid \bar{s},a)\right). \tag{6}$$

$L^{\varphi}_{\overline{\mathcal{R}}}$ 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 $\overline{\mathcal{P}}$. The following Theorems assert that when the agent employs a latent policy conditioned on the latent belief, while simultaneously minimizing both the local and belief losses to zero (i) the behaviors exhibited in the original environment align perfectly with those observed in the latent POMDP, and (ii) any pair of histories whose belief representations are close have close values as well.

Theorem 3.2. *Assume that the WAE-MDP is at the zero-temperature limit and let* $\overline{\mathcal{R}}^\star = \left\|\overline{\mathcal{R}}\right\|_\infty$, $K_{\overline{V}} = \overline{\mathcal{R}}^\star/_{1-\gamma}$, *then for any latent policy* $\overline{\pi}\colon \overline{\mathcal{B}} \to \Delta(\mathcal{A})$, *the values of* $\mathcal{P}$ *and* $\overline{\mathcal{P}}$ *are guaranteed to be bounded by the local and belief losses in average when* $\overline{\pi}$ *is executed in* $\mathcal{P}$ *via* $a \sim \overline{\pi}(\cdot \mid \varphi^*(h))$*:*

$$\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \left|V_{\overline{\pi}}(h) - \overline{V}_{\overline{\pi}}(h)\right| \leqslant \frac{L_{\mathcal{R}} + L^{\varphi}_{\overline{\mathcal{R}}} + \overline{\mathcal{R}}^\star L_{\overline{\tau}} + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} + L_{\mathcal{O}}\right)}{1-\gamma}. \tag{7}$$

Theorem 3.3. *Assume that the temperature of the WAE-MDP and the variance of* $\overline{\mathcal{O}}$ *go to zero, and let* $\overline{\pi}^\star$ *be an optimal policy of* $\overline{\mathcal{P}}$*, then for any* $\epsilon > 0$*, there is a constant* $K \geqslant 0$ *so that, for any histories* h_1, h_2 *mapped to latent beliefs via* $\varphi^*(h_1) = \bar{b}_1$ *and* $\varphi^*(h_2) = \bar{b}_2$*, the representation induced by* φ *yields:*

$$\left|V_{\overline{\pi}^\star}(h_1) - V_{\overline{\pi}^\star}(h_2)\right| \leqslant K\mathcal{W}_{\bar{d}}\left(\bar{b}_1, \bar{b}_2\right) + \epsilon +$$
$$\frac{L_{\mathcal{R}} + L^{\varphi}_{\overline{\mathcal{R}}} + \left(K + \gamma K_{\overline{V}} + \overline{\mathcal{R}}^\star\right)L_{\overline{\tau}} + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\mathcal{O}}\right)}{1-\gamma}\left(\frac{1}{\mathcal{H}_{\overline{\pi}^\star}(h_1)} + \frac{1}{\mathcal{H}_{\overline{\pi}^\star}(h_2)}\right). \tag{8}$$

We prove those theorems in Appendix C. While Thm. 3.2 guarantees the *average equivalence* of the two models and justifies the usage of $\overline{\mathcal{P}}$ as model of the environment, Thm. 3.3 shows that φ induces a suitable representation of the history to learn a policy that optimizes the expected return since *the closeness of the encoding of histories in our latent belief space implies the closeness of their values in the original POMDP*.

4 Learning to Believe

In this section, we assume access to the latent model learned by the WAE-MDP. The belief updater's goal is to compute the belief states of the latent POMDP $\overline{\mathcal{P}}$ so that an RL agent can learn to optimize a latent policy based on those latent belief.

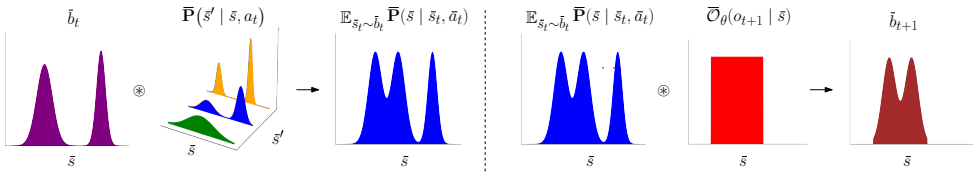


Figure 2: The Belief Update rule: (left) transformation of the current belief $\bar{b}_t$ with the transition probability function $\overline{\mathbf{P}}$, evaluated on the current action a_t, into the next state probability density; (right) filtering out the next states that could not have produced the next observation o_{t+1}.

The belief update rule. The belief update rule $\bar{\tau}$ (Eq. 3) outlines how to update the current belief based on the current action and the next observation. The former is divided into two steps (cf. Fig. 2). 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 looking at the different states that can be reached from the states that have a non-zero probability based on the latent belief. Second, the next observation o_{t+1} is used to filter the previous density based on the observation probability. 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. 2) normalizes the output of the observation filtering to obtain a probability density.

Architecture. Since our method generalizes to *any* POMDP, *we do not make any assumption about the belief distribution*. This means that we cannot assume, for example, that the belief is a multi-modal normal distribution. To accommodate complex belief distributions, we use *Masked Auto-Regressive Flows* (MAF) [33], a type of normalizing flow built on the auto-regressive property. Precisely, to accommodate with the WAE-MDP framework and leverage the guarantees presented in Sect. 3.2, we use the MAF presented in [9] that learns multivariate, latent, relaxed distributions which become discrete (binary) in the zero-temperature limit of the WAE-MDP.

We define the *sub-belief* β_t as the vector that embed the parameters of the belief distribution, the MAF allowing the transformation of sub-beliefs into beliefs, $\mathbb{M}(\beta_t) = \bar{b}_t$. The sub-belief functions similarly to the hidden states in an RNN as it is updated recursively: $\varphi^{\text{sub}}(\beta_t, a_t, o_{t+1}) = \beta_{t+1}$. However, as we do not allow gradients to back-propagate through time (BPTT), we use a feed-forward network instead of an RNN. This choice is motivated by the difference between the nature of the RNN hidden states in the partially observable version of A2C [29] (R-A2C), and sub-beliefs.

On the one hand, the goal of the RNN hidden states is to compress the history into a finite vector that can be used to compute the policy and value that maximize returns. As the policy and 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. Using the gradients of future time steps with BPTT thus helps the learning. On the other hand, the sub-belief is the vector embedding the parameters of any latent belief generated from our belief encoder, which is learned to follow the belief update rule. Disabling BPTT improves sub-belief learning as gradients from future time steps are typically of lower quality, due to beliefs from earlier steps being easier to compute as the history is smaller. Fig. 3 illustrates the distinctions in gradient flow between the two methods.

Training. We aim to train the sub-belief encoder and the MAF to approximate the update rule by minimizing the Wasserstein Distance between the belief update rule $\bar{\tau}$ of the latent POMDP, and our belief encoder φ (Eq. 6) 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

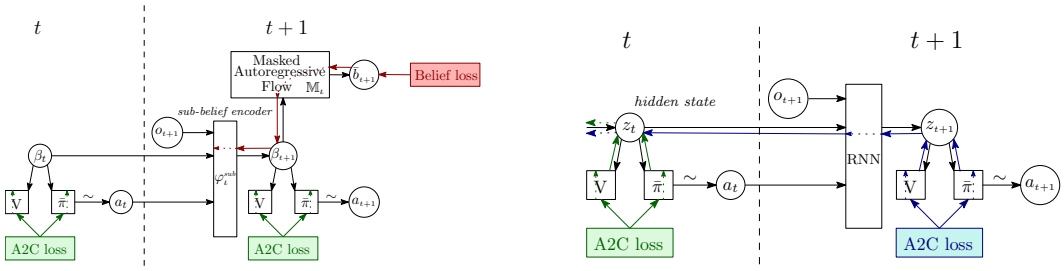


Figure 3: The Belief A2C agent (*left*) learns to encode the history into a (sub-)belief solely by optimizing the belief loss, the A2C component uses the sub-belief as state and does not back-propagate through the sub-belief encoder. Both gradients do not back-propagate through time. The RNN A2C agent (*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.

additional networks, Lipschitz constraints, and a min-max optimization procedure (e.g., [2]), similar to the WAE-MDP training procedure. Also, sampling from both distributions is necessary for the Wasserstein optimization and, while sampling from our belief approximation is straightforward, sampling from the update rule (Eq. 3) is a non-trivial task. Monte Carlo Markov Chain [1] techniques such as Metropolis-Hastings [8] could be considered, but accessing a function proportional to the density is not possible as the expectation would need to be approximated.

As an alternative to the Wasserstein optimization, we minimize the KL divergence between the two distributions. 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, KL bounds Wasserstein by the Pinsker's inequality [5, 10].

On-policy KL divergence. Using D_{KL} as a proxy for the Wassertein distance allows to close the gap between $\bar{\tau}$ and φ while optimizing the policy; at any time step $t \geqslant 0$, given the current belief $\bar{b}_t$, the action a_t played by the agent, and the next perceived observation o_{t+1}, the belief proxy loss is:

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

$$\mathop{\mathbb{E}}_{\bar{s}_{t+1}\sim\varphi\left(\bar{b}_t,a_t,o_{t+1}\right)}\left[\log\varphi\big(\bar{s}_{t+1} \mid \bar{b}_t, a_t, o_{t+1}\big) - \log \mathop{\mathbb{E}}_{\bar{s}\sim\bar{b}_t} \overline{\mathbf{P}}(\bar{s}_{t+1} \mid \bar{s}, a_t) - \log\overline{\mathcal{O}}(o_{t+1} \mid \bar{s}_{t+1})\right]. \quad (9)$$

Eq. 9 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}.

We train the belief-updater with on-policy data. Using data from the replay buffer to train the belief updater, as is done in DRQN, would require sampling full trajectories as the belief representation may change after multiple updates. Additionally, training the policy and belief updater on the same samples can facilitate learning, even though gradients are not allowed to flow between the networks.

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 algorihtm in Appendix D.

5 Experiments

Previous works on POMDPs typically evaluated algorithms on a modified Atari benchmark that excludes frame stacking and may simulate a flickering screen. However, we contend that this benchmark is limited to POMDPs where a memory of just 4 frames is adequate for state recovery. POPGym [30] addresses these limitations with environments designed to assess crucial features for generalization in POMDPs, including short-term memory for control and long-term memory.

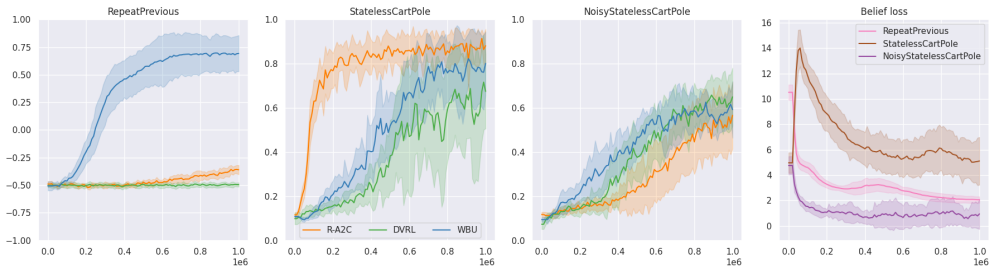


Figure 4: Evolution of the un-discounted cumulative return for WBU, R-A2C and DVRL, and evolution of the belief loss during learning for WBU (mean and standard deviation).

Our algorithm is evaluated on three distinct environments. The *Repeat Previous* environment tests the agent ability to maintain and retrieve long-term memory. It 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 for the agent to identify the suit of the card it saw 8 time steps earlier. The episode continues until all the cards have been seen, with positive rewards given for correct cards and negative otherwise. The *Stateless Cart-Pole* environment challenges the agent to control a cart on a rail, maintaining an attached pole within a specific angle range. The system state includes cart position, velocity, pole angular position, and velocity. In this partially observable variant, the agent relies on short-term memory to estimate hidden velocity components. Positive rewards are given at each time step. Finally, the *Noisy Stateless Cart-Pole* environment is a more challenging version of the previous one featuring both partial observability and added Gaussian noise on the positions and velocities of the cart and pole. The rewards of all the environments are scaled so that the maximum return is 1.

We compare the performance of WBU in those three environments with R-A2C and DVRL [22] (Fig. 4). DVRL is an algorithm derived from R-A2C, utilizing a combination of Variational Auto-Encoder and particle filtering to sustain a state distribution that acts as a proxy of the latent belief, without providing any guarantee of being the true belief. We train 10 instances of each algorithm for 1 million time steps. Appendix E presents the hyper-parameters used and the range of the search.

In *Repeat Previous*, WBU stands out by achieving a positive return within a quarter of the learning process, and by the end of the learning demonstrates near-perfect memorization of the last 8 cards with appropriate actions selection (return of 0.75). In contrast, R-A2C only obtains -0.3 and DVRL does not show any learning. This experiment demonstrates WBU's ability to effectively remember and recall previous observations when needed. In *Stateless Cart-Pole*, we observe that R-A2C rapidly reaches a return of 0.8 and achieves a final performance of 0.9. WBU outperforms DVRL (0.7), achieving a final performance of 0.8. Both algorithms show ongoing signs of learning. This experiment demonstrates that WBU effectively utilizes short-term memory for control, albeit requiring more interaction compared to R-A2C. WBU's ability to map complete histories to beliefs may not be necessary for optimal policy, as the optimal policy might only depend on recent history. Focusing on recent history may aid in generalization, which could explain R-A2C's performance in this environment. In *Noisy Stateless Cart-Pole*, the three algorithms perform similarly, with R-A2C having the lowest performance and WBU initially learning faster. Notably, DVRL is less affected by noisy observations compared to WBU and R-A2C. The final plot in Figure 4 shows the satisfactory decrease in WBU's belief loss.

6 Conclusion

WBU provides a novel 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, our WBU algorithm 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.

In future work, we aim to explore the use of simulated trajectories for policy learning, which is theoretically enabled through the model and representation quality guarantees (Thm 3.2 and 3.3). Furthermore, the works of [13, 10] study similar value difference bounds to ours in the context of fully observable environments. They further link their bounds with bisimulation theory (e.g., [25, 14]). We defer as future work the study of bisimulation metrics [11, 12] in POMDPs.

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 and Denis Steckelmacher for their valuable feedback.

References

[1] Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1-2):5–43, 2003.

[2] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *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*, pages 214–223. PMLR, 2017.

[3] Karl Johan Åström. Optimal control of markov processes with incomplete state information. *Journal of mathematical analysis and applications*, 10(1):174–205, 1965.

[4] 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.

[5] J.M. Borwein and A.S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. CMS Books in Mathematics. Springer New York, 2005.

[6] Krishnendu Chatterjee, Martin Chmelik, Raghav Gupta, and Ayush Kanodia. Optimal cost almost-sure reachability in pomdps. *Artif. Intell.*, 234:26–48, 2016.

[7] 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*, pages 3444–3468. PMLR, 2022.

[8] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

[9] 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.

[10] 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.

[11] Josée Desharnais, Vineet Gupta, Radha Jagadeesan, and Prakash Panangaden. Metrics for labelled markov processes. *Theor. Comput. Sci.*, 318(3):323–354, 2004.

[12] Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation metrics for continuous markov decision processes. *SIAM J. Comput.*, 40(6):1662–1714, 2011.

[13] 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, editors, *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*, pages 2170–2179. PMLR, 2019.

[14] 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.

[15] 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.

[16] 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.

[17] Danijar Hafner, Timothy Lillicrap Deepmind, Jimmy Ba, Mohammad Norouzi, and Google Brain. Dream to Control: Learning Behaviors by Latent Imagination. 12 2019.

[18] Danijar Hafner, Timothy P Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021.

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

[20] Milos Hauskrecht. Value-function approximations for partially observable markov decision processes. *J. Artif. Intell. Res.*, 13:33–94, 2000.

[21] Bojun Huang. Steady state analysis of episodic reinforcement learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[22] 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.

[23] 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.

[24] 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.

[25] 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*, pages 344–352. ACM Press, 1989.

[26] 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.

[27] 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.

[28] 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, pages 5101–5108, 2020.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[34] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1994.

[35] 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.

[36] Edward J. Sondik. The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Oper. Res.*, 26(2):282–304, 1978.

[37] Matthijs TJ Spaan. Partially observable markov decision processes. *Reinforcement learning: State-of-the-art*, pages 387–414, 2012.

[38] R.S. Sutton and A.G. Barto. Reinforcement Learning: An Introduction. *IEEE Transactions on Neural Networks*, 1998.

[39] Cédric Villani. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

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

[41] 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*, pages 241–251, 2021.

[42] K.J Åström. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965.

Appendix

A 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.

B Proof of Lemma 3.1: Stationarity over Histories

Let us restate the Lemma:

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

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 $\overline{\pi}$. All episodic processes are *ergodic* [21], 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*.

B.1 Preliminaries

We *formally* recall the notion of *episodic process*:

Definition B.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 (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 $o_T = o^\star$.*

Assumption B.3. *The environment $\mathcal{P}$ is an episodic process.*

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-based* policies.

Definition B.4 (Memory-based policies)**.** *Given an MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathbf{P}, \mathcal{R}, s_I, \gamma \rangle$, a* memory-based *policy for $\mathcal{M}$ is a policy that can be encoded as a* stochastic Mealy machine *$\pi = \langle Q, \pi_\alpha, \pi_\mu, q_I \rangle$, where 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 B.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:

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

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

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 B.5 (Markov Chain)**.** *A* Markov Chain *(MC) is an MDP whose action space 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., $\mathcal{R}_\pi(\langle s, q \rangle) = \mathbb{E}_{a \sim \pi_\alpha(\cdot \mid s, q)} \mathcal{R}(s, a)$.*

Furthermore, the probability measure $\mathbb{P}_\pi^{\mathcal{M}}$ is actually the unique probability measure defined over the measurable infinite trajectories of the MC $\mathcal{M}^\pi$ *[34].*

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.

Definition B.6 (Bottom strongly connected components and limiting distributions)**.** *Let $\mathcal{M}$ be an MDP with state space $\mathcal{M}$ and π be a policy for $\mathcal{M}$. Write $\mathcal{M}[s]$ for the MDP where we change the initial state s_I of $\mathcal{M}$ by $s \in \mathcal{S}$. 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. The subset $B \subseteq \mathcal{S}$ is a* strongly connected component *(SCC) of $\mathcal{M}^\pi$ if for any pair of states $s, s' \in B$, $\xi_\pi^t(s' \mid s) > 0$ for some $t \in \mathbb{N}$. 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 $s_\perp \in B$. 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 $s \in \mathcal{S}$.*

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 $\bar{\pi}$.

B.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 $\bar{\pi}$, 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 B.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 [25, 14]. The relation can be extended to compare two MDPs by considering the disjoint union of their state space.

Lemma B.8. *Let $\mathcal{P}$ be the POMDP of Lemma B.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 memory-based 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

$$\begin{aligned}
\mathbf{P}_{\overline{\pi}^{\diamondsuit}}\big(b', \bar{b}' \mid b, \bar{b}\big) &= \mathop{\mathbb{E}}_{a \sim \overline{\pi}\left(\cdot \mid \bar{b}\right)} \overline{\pi}_{\mu}\big(\bar{b}' \mid b, \bar{b}, a, b'\big) \cdot \mathbf{P}_{\mathcal{B}}\big(b' \mid b, a\big) \\
&= \mathop{\mathbb{E}}_{a \sim \overline{\pi}\left(\cdot \mid \bar{b}\right)} \mathop{\mathbb{E}}_{s \sim b} \mathop{\mathbb{E}}_{s' \sim \mathbf{P}(\cdot \mid s, a)} \mathop{\mathbb{E}}_{o' \sim \mathcal{O}(\cdot \mid s, a)} \delta_{\varphi\left(\bar{b}, o, a\right)}\big(\bar{b}'\big) \cdot \delta_{\tau(b, o, a)}\big(b'\big), \text{and} \\
\mathcal{R}_{\overline{\pi}^{\diamondsuit}}(b, \bar{b}) &= \mathop{\mathbb{E}}_{a \sim \overline{\pi}\left(\cdot \mid \bar{b}\right)} \mathop{\mathbb{E}}_{s \sim b} \mathcal{R}(s, a). \qquad \text{(cf. Definition B.5)}
\end{aligned}$$

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

2. Each equivalence class $T \in \big(\mathcal{S}_{\mathcal{H}} \times \mathcal{B} \times \overline{\mathcal{B}}\big) / \rightrightarrows_{\varphi}^{\tau}$ consists of histories sharing the same belief and latent beliefs. Since $\tau^*\colon \mathcal{S}_{\mathcal{H}} \to \mathcal{B}$ and $\varphi^*\colon \mathcal{S}_{\mathcal{H}} \to \overline{\mathcal{B}}$ are surjective, 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 $\rightrightarrows_{\varphi}^{\tau}$ has the form $T = \big[\langle b', \bar{b}' \rangle\big]_{\rightrightarrows_{\varphi}^{\tau}}$ so that

   (a) the projection of $\big[\langle b', \bar{b}' \rangle\big]_{\rightrightarrows_{\varphi}^{\tau}}$ 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 $\big[\langle b', \bar{b}' \rangle\big]_{\rightrightarrows_{\varphi}^{\tau}}$ 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}}\Big(\big[\langle b', \bar{b}' \rangle\big]_{\rightrightarrows_{\varphi}^{\tau}} \mid h\Big) \\
&= \int_{\left[\langle b', \bar{b}' \rangle\right]_{\rightrightarrows_{\varphi}^{\tau}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi^*(h))} \mathop{\mathbb{E}}_{s \sim \tau^*} \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)\, dh' \\
&= \int_{\mathcal{S}_{\mathcal{H}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi^*(h))} \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) \cdot \delta_{\tau^*(h')}\big(b'\big) \cdot \delta_{\varphi^*(h')}\big(\bar{b}'\big)\, dh' \quad \text{(by definition of } \big[\langle b', \bar{b}' \rangle\big]_{\rightrightarrows_{\varphi}^{\tau}}\text{)} \\
&= \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi^*(h))} \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_{\tau^*(h \cdot a \cdot o')}\big(b'\big) \cdot \delta_{\varphi^*(h \cdot a \cdot o')}\big(\bar{b}'\big) \\
&= \mathop{\mathbb{E}}_{a \sim \overline{\pi}\left(\cdot \mid \bar{b}\right)} \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_{\tau(b, a, o')}\big(b'\big) \cdot \delta_{\varphi\left(\bar{b}, a, o'\right)}\big(\bar{b}'\big) \qquad \text{(since } h \rightrightarrows_{\varphi}^{\tau} \langle b, \bar{b} \rangle\text{)} \\
&= \mathbf{P}_{\overline{\pi}^{\diamondsuit}}\big(b', \bar{b}' \mid b, \bar{b}\big) \\
&= \mathbf{P}_{\overline{\pi}^{\diamondsuit}}\Big(\big[\langle b', \bar{b}' \rangle\big]_{\rightrightarrows_{\varphi}^{\tau}} \mid b, \bar{b}\Big)
\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 $\rightrightarrows_{\varphi}^{\tau}$, when the policies $\overline{\pi}^{\clubsuit}$ and $\overline{\pi}^{\diamondsuit}$ are respectively executed in the two models. □

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

Proof. Follows directly from [25, 14]: the bisimulation relation implies the equivalence of the optimal policies in the two models, i.e., the maximum expected returns are the same in the two models. Since we consider MCs and not MDPs, the models are purely stochastic, and the behavior equality follows. □

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.

B.3 Existence of a Stationary Distribution over Histories

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

Proof. By definition of $\mathcal{M}_{\mathcal{H}}$, the execution of $\overline{\pi}^{\clubsuit}$ is guaranteed to remain an episodic process. Every episodic process is ergodic (see [21]), 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 $\mathcal{M}^{\overline{\pi}^{\clubsuit}}_{\mathcal{H}}$, which actually consists of histories of $\mathcal{P}$ when the latter operates under $\overline{\pi}$, or equivalently, the execution of the MC $\mathcal{M}^{\overline{\pi}^{\diamondsuit}}_{\mathcal{B}}$. □

C 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.

C.1 Value Functions

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

Definition C.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 B.5). Then, the value function can be defined as the unique solution of the Bellman's equation [34]: $V_\pi(s) = \mathcal{R}_\pi(s) + \mathbb{E}_{s'\sim\mathbf{P}_\pi(s)}\left[\gamma \cdot V_\pi(s')\right]$. 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 C.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 $\overline{\pi}\colon \overline{\mathcal{B}} \to \Delta(\mathcal{A})$:*

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

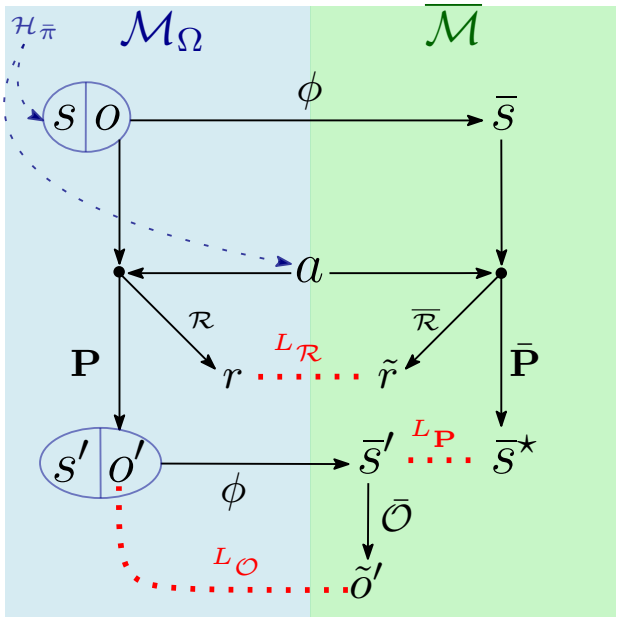
Similarly, we write $\overline{V}_{\overline{\pi}}$ for the values of a latent POMDP $\overline{\mathcal{P}}$.

C.2 Local and Belief Losses

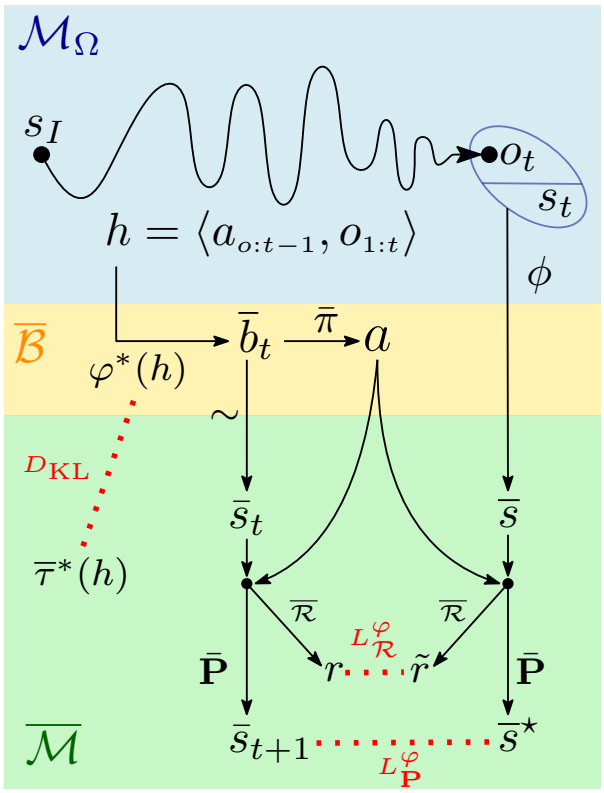
Theorem C.7 involves the minimization of *local* ($L_{\mathcal{R}}, L_{\mathbf{P}}, L_{\mathcal{O}}$) and *belief* ($L_{\overline{\tau}}, L^{\varphi}_{\overline{\mathbf{P}}}, L^{\varphi}_{\overline{\mathcal{R}}}$) losses. We intuitively describe how these losses are minimized via the latent flows depicted in Fig. 5.

The procedure allowing to minimize the local losses is depicted in Fig. 5a. 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 $\tilde{o}' \sim \overline{\mathcal{O}}(\cdot \mid \bar{s}')$.

The procedure allowing to minimize the belief losses is depicted in Fig. 5b. 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



(a) Optimization of the latent space model parameters (i.e., $\overline{\mathcal{R}}, \bar{\mathbf{P}}$, and $\bar{\mathcal{O}}$) 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 5: 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.

h ends up in the observation o_t of state s_t. First, the discrepancy between the latent belief obtained via our belief encoder $\bar{b}_t = \varphi^*(h)$ 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.

C.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 main trick to prove the claim is to decay the temperature to the zero-limit, which makes the distance $\bar{d}$ metric associated with the latent state space converge to the discrete metric $\mathbf{1}_{\neq} \colon \mathcal{X} \to \{1, 0\}$ [9], formally defined as $\mathbf{1}_{\neq}(x1, x2) = 1$ iff $x_1 \neq x_2$.

Definition C.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 C.4 (Wasserstein dual)**.** *The Kantorovich-Rubinstein duality [39] 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 C.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., [13, Sect. 6] for a discussion).*

The latter property 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 $\sup_{\bar{s},a}\left|\overline{\mathcal{R}}(\bar{s},a)\right|/1-\gamma$, so given two distributions P, Q over $\overline{\mathcal{S}}$, 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 C.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 $\mathbb{E}_{y\sim\mathcal{D}}\left|\mathbb{E}_{x_1\sim P} f(y,x_1) - \mathbb{E}_{x_2\sim Q} f(y,x_2)\right| \leqslant K \cdot \mathcal{W}_d(P,Q)$.*

Proof. The proof is straightforward by construction of g_y:

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

□

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

Let us restate Theorem 3.2:

Theorem C.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 B.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., $\lambda \to 0$) 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:*

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

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 C.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:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \left|V_{\overline{\pi}}(h) - \overline{V}_{\overline{\pi}}(h)\right| \\
&= \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \Bigg| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))} \left[\mathcal{R}(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)} V_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\right] && \text{(see Property C.2)} \\
&\qquad - \mathop{\mathbb{E}}_{\bar{s}\sim\overline{\tau}^*(h)} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))} \left[\overline{\mathcal{R}}(\bar{s},a) + \gamma \mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\right] \Bigg| \\
&\leqslant \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*(h))} \Bigg[\left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathcal{R}(s,a) - \mathop{\mathbb{E}}_{\bar{s}\sim\overline{\tau}^*(h)} \overline{\mathcal{R}}(\bar{s},a)\right| \\
&\qquad + \gamma \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)} V_{\pi}\big(h\cdot a\cdot o'\big) - \mathop{\mathbb{E}}_{\bar{s}\sim\overline{\tau}^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\right| \Bigg] \\
& && \text{(Triangular inequality)}
\end{aligned}
$$

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

Part 1: Reward bounds

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\bar\pi}}\mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot|\varphi^*(h))}\left|\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)\right| \\
=&\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\bar\pi}}\mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot|\varphi^*(h))}\left|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\left[\mathcal{R}(s,a)-\overline{\mathcal{R}}(\phi(s,o),a)\right]+\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar s\sim\bar\tau^*(h)}\left[\overline{\mathcal{R}}(\phi(s,o),a)-\overline{\mathcal{R}}(\bar s,a)\right]\right| \\
&\qquad(o\text{ is the last observation of }h\text{; the state embedding function }\phi\text{ that links the original and latent state spaces comes into play}) \\
\leqslant&\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\bar\pi}}\mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot|\varphi^*(h))}\left[\left|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\left[\mathcal{R}(s,a)-\overline{\mathcal{R}}(\phi(s,o),a)\right]\right|+\left|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar s\sim\bar\tau^*(h)}\left[\overline{\mathcal{R}}(\phi(s,o),a)-\overline{\mathcal{R}}(\bar s,a)\right]\right|\right] \\
&\qquad\text{(Triangular inequality)} \\
\leqslant&\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\bar\pi}}\mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot|\varphi^*(h))}\left[\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\left|\mathcal{R}(s,a)-\overline{\mathcal{R}}(\phi(s,o),a)\right|+\left|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar s\sim\bar\tau^*(h)}\left[\overline{\mathcal{R}}(\phi(s,o),a)-\overline{\mathcal{R}}(\bar s,a)\right]\right|\right] \\
&\qquad\text{(Jensen's inequality)} \\
=&\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)}\left|\mathcal{R}(s,a)-\overline{\mathcal{R}}(\phi(s,o),a)\right|+\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\bar\pi}}\mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot|\varphi^*(h))}\left|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar s\sim\bar\tau^*(h)}\left[\overline{\mathcal{R}}(\phi(s,o),a)-\overline{\mathcal{R}}(\bar s,a)\right]\right| \\
=&L_{\mathcal{R}}+\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\bar\pi}}\mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot|\varphi^*(h))}\left|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar s\sim\bar\tau^*(h)}\left[\overline{\mathcal{R}}(\phi(s,o),a)-\overline{\mathcal{R}}(\bar s,a)\right]\right| \qquad\text{(by definition of } L_{\mathcal{R}}\text{)} \\
=&L_{\mathcal{R}}+\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\bar\pi}}\mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot|\varphi^*(h))}\left|\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)}\left[\left[\overline{\mathcal{R}}(\phi(s,o),a)-\overline{\mathcal{R}}(\bar s_\perp,a)\right]+\left[\overline{\mathcal{R}}(\bar s_\perp,a)-\overline{\mathcal{R}}(\bar s,a)\right]\right]\right| \\
&\qquad\text{(the belief encoder }\varphi\text{ comes into play)} \\
=&L_{\mathcal{R}}+\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\bar\pi}}\mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot|\varphi^*(h))}\left|\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}}_{\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]\right| \\
\leqslant&L_{\mathcal{R}}+\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\bar\pi}}\mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot|\varphi^*(h))}\left[\left|\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]\right|+\left|\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]\right|\right] \\
&\qquad\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|\varphi^*(h))}\left[\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|+\left|\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]\right|\right] \\
&\qquad\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|\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|\varphi^*(h))}\left|\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)\right| \\
=&L_{\mathcal{R}}+L^{\varphi}_{\overline{\mathcal{R}}}+\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\bar\pi}}\mathop{\mathbb{E}}_{a\sim\bar\pi(\cdot|\varphi^*(h))}\left|\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)\right| \qquad\text{(by definition of } L^{\varphi}_{\overline{\mathcal{R}}}\text{, Eq. 6)} \\
\leqslant&L_{\mathcal{R}}+L^{\varphi}_{\overline{\mathcal{R}}}+\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\bar\pi}}\overline{\mathcal{R}}^{\star}\mathcal{W}_{\bar d}\left(\bar\tau^*(h),\varphi^*(h)\right) \qquad\text{(as }\lambda\to 0\text{, by Lem. C.6 and Prop. C.5)} \\
=&L_{\mathcal{R}}+L^{\varphi}_{\overline{\mathcal{R}}}+\overline{\mathcal{R}}^{\star}L_{\bar\tau};
\end{aligned}
$$

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

Part 2: Next value bounds

$$
\begin{aligned}
& \gamma \cdot \mathop{\mathbb{E}}_{h\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'\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\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) \right| \\
= \; & \gamma \cdot \mathop{\mathbb{E}}_{h\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'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)} \left[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 o'\big) \right] \right. \\
& \qquad \left. + \left[\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)} \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) - \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) \right] \right| \\
& \qquad\qquad \text{(the state embedding function } \phi \text{ comes into play, as well as the latent observation function } \overline{\mathcal{O}}) \\
\leqslant \; & \gamma \cdot \mathop{\mathbb{E}}_{h\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'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)} \left[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| \\
& \qquad + \gamma \cdot \mathop{\mathbb{E}}_{h\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'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)} \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) - \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) \right| \\
& \qquad\qquad \text{(Triangular inequality)} \\
= \; & \gamma \cdot \mathop{\mathbb{E}}_{h\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'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)} \left[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| \\
& \qquad + \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[\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] - \mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\phi(s,o),a)} \left[\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\bar{s}')} \overline{V}_{\overline{\pi}}\big(h\cdot a\cdot \hat{o}'\big) \right] \right. \\
& \qquad\qquad \left. + \left[\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) - \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) \right] \right| \\
& \qquad\qquad (o \text{ is the last observation of } h\text{; the latent MDP dynamics, modeled by } \overline{\mathbf{P}}, \text{ come into play}) \\
\leqslant \; & \gamma \cdot \mathop{\mathbb{E}}_{h\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'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)} \left[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| \\
& \quad + \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[\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] - \mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\phi(s,o),a)} \left[\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\overline{\pi}}\big(h\cdot a\cdot \hat{o}'\big) \right] \right| \\
& \quad + \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}}_{\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) - \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) \right| \\
& \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))} \left| \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}}\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| \\
& \quad + \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)} \left| \mathop{\mathbb{E}}_{\bar{s}'\sim\phi\mathbf{P}_\Omega(\cdot|s,o,a)} \left[\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) \right] - \mathop{\mathbb{E}}_{\bar{s}'\sim\overline{\mathbf{P}}(\cdot|\phi(s,o),a)} \left[\mathop{\mathbb{E}}_{o'\sim\overline{\mathcal{O}}(\cdot|\bar{s}')} \overline{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) \right] \right| \\
& \quad + \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}}_{\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) - \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) \right| \\
& \qquad\qquad \text{(Jensen's inequality)} \\
\leqslant \; & \gamma \cdot \mathop{\mathbb{E}}_{h\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'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim\mathcal{O}(\cdot|s',a)} \left[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| \\
& \quad + \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}}\big(\phi\mathbf{P}_\Omega(\cdot \mid s,a), \overline{\mathbf{P}}(\cdot \mid \phi(s,o),a)\big) \\
& \quad + \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}}_{\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) - \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) \right| \\
& \qquad\qquad \text{(as } \lambda \to 0\text{, by Lem. C.6)}
\end{aligned}
$$

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

$$
\begin{aligned}
=\;& \gamma \cdot \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi^*(h))} \left| \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)} \left[V_{\overline{\pi}}(h \cdot a \cdot o') - \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot \mid \phi(s', o'))} \overline{V}_{\overline{\pi}}(h \cdot a \cdot \hat{o}') \right] \right| \\
& + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} \right) \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi^*(h))} \left| \mathop{\mathbb{E}}_{\bar{s} \sim \varphi^*(h)} \left[\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}_{\overline{\pi}}(h \cdot a \cdot o') \right] - \mathop{\mathbb{E}}_{\bar{s} \sim \overline{\tau}^*(h)} \left[\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}_{\overline{\pi}}(h \cdot a \cdot o') \right] \right| \\
& \hfill \text{(by definition of } L^{\varphi}_{\overline{\mathbf{P}}}\text{, Eq. 6)} \\
\leqslant\;& \gamma \cdot \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi^*(h))} \left| \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)} \left[V_{\overline{\pi}}(h \cdot a \cdot o') - \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot \mid \phi(s', o'))} \overline{V}_{\overline{\pi}}(h \cdot a \cdot \hat{o}') \right] \right| \\
& + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} \right) \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi^*(h))} K_{\overline{V}} \mathcal{W}_{\bar{d}}\left(\overline{\tau}^*(h), \varphi^*(h) \right) \\
& \hfill \text{(as } \lambda \to 0\text{, by Lem. C.6; note that Wasserstein is symmetric since it is a distance metric [39])} \\
\leqslant\;& \gamma \cdot \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi^*(h))} \left| \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)} \left[V_{\overline{\pi}}(h \cdot a \cdot o') - \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot \mid \phi(s', o'))} \overline{V}_{\overline{\pi}}(h \cdot a \cdot \hat{o}') \right] \right| \\
& + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} \right) \hfill \text{(by definition of } L_{\overline{\tau}}\text{, Eq. 6)} \\
=\;& \gamma \cdot \mathop{\mathbb{E}}_{h, o \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi^*(h))} \left| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s', o' \sim \mathbf{P}_{\Omega}(\cdot \mid s, o, a)} \left[\left(V_{\overline{\pi}}(h \cdot a \cdot o') - \overline{V}_{\overline{\pi}}(h \cdot a \cdot o') \right) + \left(\overline{V}_{\overline{\pi}}(h \cdot a \cdot o') - \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot \mid \phi(s', o'))} \overline{V}_{\overline{\pi}}(h \cdot a \cdot \hat{o}') \right) \right] \right| \\
& + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} \right) \\
\leqslant\;& \gamma \cdot \mathop{\mathbb{E}}_{h, o \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi^*(h))} \left| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s', o' \sim \mathbf{P}_{\Omega}(\cdot \mid s, o, a)} \left[V_{\overline{\pi}}(h \cdot a \cdot o') - \overline{V}_{\overline{\pi}}(h \cdot a \cdot o') \right] \right| \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h, o \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi^*(h))} \left| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s', o' \sim \mathbf{P}_{\Omega}(\cdot \mid s, o, a)} \left[\overline{V}_{\overline{\pi}}(h \cdot a \cdot o') - \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot \mid \phi(s', o'))} \overline{V}_{\overline{\pi}}(h \cdot a \cdot \hat{o}') \right] \right| \\
& + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} \right) \hfill \text{(triangular inequality)} \\
\leqslant\;& \gamma \cdot \mathop{\mathbb{E}}_{h, o \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi^*(h))} \left| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s', o' \sim \mathbf{P}_{\Omega}(\cdot \mid s, o, a)} \left[V_{\overline{\pi}}(h \cdot a \cdot o') - \overline{V}_{\overline{\pi}}(h \cdot a \cdot o') \right] \right| \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h, o \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi^*(h))} \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s' \sim \mathbf{P}(\cdot \mid s, a)} \left| \mathop{\mathbb{E}}_{o' \sim \mathcal{O}(\cdot \mid s', a)} \left[\overline{V}_{\overline{\pi}}(h \cdot a \cdot o') - \mathop{\mathbb{E}}_{\hat{o}' \sim \overline{\mathcal{O}}(\cdot \mid \phi(s', o'))} \overline{V}_{\overline{\pi}}(h \cdot a \cdot \hat{o}') \right] \right| \\
& + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} \right) \hfill \text{(Jensen's inequality)} \\
\leqslant\;& \gamma \cdot \mathop{\mathbb{E}}_{h, o \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi^*(h))} \left| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s', o' \sim \mathbf{P}_{\Omega}(\cdot \mid s, o, a)} \left[V_{\overline{\pi}}(h \cdot a \cdot o') - \overline{V}_{\overline{\pi}}(h \cdot a \cdot o') \right] \right| \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h, o \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi^*(h))} \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s' \sim \mathbf{P}(\cdot \mid s, a)} K_{\overline{V}} d_{TV}\left(\mathcal{O}(\cdot \mid s', a), \mathop{\mathbb{E}}_{o' \sim s', a} \overline{\mathcal{O}}(\cdot \mid \phi(s', o')) \right) \\
& + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} \right) \hfill \text{(cf. Prop. C.5 and Lem C.6)} \\
=\;& \gamma \cdot \mathop{\mathbb{E}}_{h, o \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi^*(h))} \left| \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s', o' \sim \mathbf{P}_{\Omega}(\cdot \mid s, o, a)} \left[V_{\overline{\pi}}(h \cdot a \cdot o') - \overline{V}_{\overline{\pi}}(h \cdot a \cdot o') \right] \right| \\
& + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} + L_{\mathcal{O}} \right) \hfill \text{(by definition of } L_{\mathcal{O}}\text{, Eq. 5)} \\
\leqslant\;& \gamma \cdot \mathop{\mathbb{E}}_{h, o \sim \mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi^*(h))} \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \mathop{\mathbb{E}}_{s', o' \sim \mathbf{P}_{\Omega}(\cdot \mid s, o, a)} \left| V_{\overline{\pi}}(h \cdot a \cdot o') - \overline{V}_{\overline{\pi}}(h \cdot a \cdot o') \right| + \gamma K_{\overline{V}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} + L_{\mathcal{O}} \right) \\
& \hfill \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 \left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\overline{\tau}} + L_{\mathcal{O}} \right) \\
& \hfill \text{(} \mathcal{H}_{\overline{\pi}} \text{ is a stationary distribution (Lem. B.1) which allows us to apply the stationary property (Def. B.6))}
\end{aligned}
$$

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

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

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

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

which finally concludes the proof. □

C.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 C.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 C.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., [45]) 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 $V_{\pi^\star}$.

Lemma C.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 $U\colon\mathcal{B}\times\mathcal{A}\times\mathcal{V}$ as

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

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., $V^\star=\mathcal{U}V^\star$ [34, 20, 38]. 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 [34]): 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, V_i is also PWLC [36, 35, 20]. Since $\mathcal{S}$ is discrete, the belief space $\mathcal{B}$ is the standard $|\mathcal{S}|$-dimensional simplex, so the domain of $(V_i)_{i\in\mathbb{N}}$ is compact, meaning that it is defined as a finite collection of linear functions. Thus, $(V_i)_{i\in\mathbb{N}}$ is also $2K'$-Lipschitz: one just need take K' as the higher slope of these functions (in absolute value). In consequence, for any pair of beliefs $b_1,b_2\in\mathcal{B}$,

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

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

Corollary C.10. *When the temperature of the WAE-MDP and the variance of $\overline{\mathcal{O}}$ go to zero, 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 $\overline{\mathcal{S}}$ is finite. Then Lemma C.9 can be applied. □

Theorem C.11. *Let $\bar{\pi}^\star$ be an optimal policy of the POMDP $\overline{\mathcal{P}}$, then for any couple of histories $h_1,h_2\in(\mathcal{A}\cdot\Omega)^*$ mapped to latent beliefs through $\varphi^*(h_1)=\bar{b}_1$ and $\varphi^*(h_2)=\bar{b}_2$ and any arbitrary error term $\epsilon>0$, the belief representation induced by φ*

yields the existence of a constant $K \geqslant 0$ so that:

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

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

Proof. First, observe that for any history $h \in (\mathcal{A} \cdot \Omega)^*$, $\left|V_{\bar{\pi}^\star}(h) - \overline{V}_{\bar{\pi}^\star}(h)\right| \leqslant \mathcal{H}_{\bar{\pi}^\star}(h)^{-1} \cdot \mathbb{E}_{h' \sim \mathcal{H}_{\bar{\pi}^\star}} \left|V_{\bar{\pi}^\star}(h') - \overline{V}_{\bar{\pi}^\star}(h')\right|$ (cf. [13]). Therefore, we have:

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

Now, let $\epsilon > 0$. Recall that the latent value function (defined over the latent belief space) is almost Lipschitz continuous (Corollary C.10). In particular, for $\delta = \frac{\epsilon}{\left(1 + \mathcal{H}_{\bar{\pi}^\star}(h_1)^{-1} + \mathcal{H}_{\bar{\pi}^\star}(h_2)^{-1}\right)}$, there is a $K \geqslant 0$ so that for any $\bar{b}, \bar{b}' \in \bar{\mathcal{B}}$, $\left|\overline{V}^\star(b) - \overline{V}^\star(b')\right| \leqslant K\mathcal{W}_{\bar{d}}(b, b') + \delta$. 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| && \text{(Triangular inequality)} \\
\leqslant&\mathcal{H}_{\bar{\pi}^\star}(h_1)^{-1} \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\bar{\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}_{\bar{\pi}^\star}(h_2)^{-1} \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\bar{\pi}^\star}} \left|\overline{V}^\star(\bar{\tau}^*(h)) - \overline{V}^\star(\varphi^*(h))\right| \\
=&\left(\mathcal{H}_{\bar{\pi}^\star}(h_1)^{-1} + \mathcal{H}_{\bar{\pi}^\star}(h_2)^{-1}\right) \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\bar{\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}_{\bar{\pi}^\star}(h_1)^{-1} + \mathcal{H}_{\bar{\pi}^\star}(h_2)^{-1}\right) \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\bar{\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 && (\overline{V}^\star \text{ is almost Lipschitz}) \\
=&\left(\mathcal{H}_{\bar{\pi}^\star}(h_1)^{-1} + \mathcal{H}_{\bar{\pi}^\star}(h_2)^{-1}\right)(KL_{\bar{\tau}} + \delta) + K\mathcal{W}_{\bar{b}}\left(\bar{b}_1, \bar{b}_2\right) + \delta \\
=&\left(\mathcal{H}_{\bar{\pi}^\star}(h_1)^{-1} + \mathcal{H}_{\bar{\pi}^\star}(h_2)^{-1}\right)KL_{\bar{\tau}} + K\mathcal{W}_{\bar{b}}\left(\bar{b}_1, \bar{b}_2\right) + \delta\left(1 + \mathcal{H}_{\bar{\pi}^\star}(h_1)^{-1} + \mathcal{H}_{\bar{\pi}^\star}(h_2)^{-1}\right) \\
=&\left(\mathcal{H}_{\bar{\pi}^\star}(h_1)^{-1} + \mathcal{H}_{\bar{\pi}^\star}(h_2)^{-1}\right)KL_{\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 exists a constant $K \geqslant 0$ so that:

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

□

D 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:

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

Notice that the first term of the divergence (the belief normalization term) of Eq. 9 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 [9, 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 [9], we do so via the gradient penalty approach of [44], 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 [9, Appendix A.6] to recover d_{TV}.

Algorithm 1: WASSERSTEIN BELIEF UPDATER

Input: Batch sizes B_{WBU}, B_{WAE}; 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**

      $a_t \sim \overline{\pi}(\cdot \mid \beta_t)$ ▷ *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**

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

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

      ▷ *Pointwise decomposition of Eq. 9: 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 $\overline{\pi}$ 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 the observation of $\bar{s}$; get the standard deviation of the reconstruction*

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

E 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 [43]. 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 which uses three) 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.

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

	Range	**RepeatPrevious**	**StatelessCartpole**	**NoisyStatelessCartpole**
WAE Updates per Belief Update	1-2	2	1	1
Activation function	leaky relu, elu	elu	leaky relu	leaky relu
Activation function lipshitz	leaky relu, smooth elu	leaky relu	leaky relu	leaky relu
Policy config				
Learning rate	1.e-4, 3.e-4, 5.e-4, 1.e-3	1.e-4	1.e-4	1.e-4
λ	0.95, 1.	0.95	0.95	1.
Clip norm	1, 10	1	1	10
Belief config				
Loss factor	1.e-5, 1.e-4, 1.e-3, 1.e-2, .1, 1.	1.e-5	1.e-4	1.e-2
Clip Norm	1, 10.	1	1	10
Filter variance min	1.e-2, 1.e-3, 1.e-4, 5.e-5	1.e-2	1.e-4	1.e-2
Normalize log obs filter	True, False	True	True	False
Sub belief prior temperature	0.33, .5, 0.66, .75, .9, .99	0.99	0.5	0.99
Reward loss scale factor	0, 0.1, 1., 10., 20., 50., 100.	50	0	20
Transition loss scale factor	0, 0.1, 1., 10., 20., 50., 100.	50	0	100
Buffer size	4096, 8192, 16384, 32768	32768	4096	4096
Use normalization term	True, False	True	True	True
WAE config				
Latent state size	RP: 18→25; Cartpole: 5→10	22	7	8
Minimizer learning rate	1.e-4, 3.e-4, 5.e-5, 1.e-3	1.e-4	5.e-5	1.e-4
Maximizer learning rate	1.e-4, 3.e-4, 5.e-5, 1.e-3	1.e-3	5.e-5	1.e-3
State encoder temperature	0.33, .5, 0.66, .75, .9, .99	0.33	0.33	0.5
State prior temperature	0.33, .5, 0.66, .75, .9, .99	0.99	0.99	0.75
Local transition loss scaling	10., 25., 50., 75., 80.	75	10	25
Steady state scaling	10., 25., 50., 75., 80.	75	25	75
N critic update	5, 10	5	10	10
Batch size	128, 256	256	128	128
Clip grad	1, 10. 100.	10	1	100
Cost function	l22, l2	l2	l2	l2
State reconstruction weight vs obs	1, 2, 5, 10	10	5	5
Observation regularizer	True, False	True	False	True
Observation reg. same optimizer	True, False	True		False
Observation reg. min learning rate	1.e-4, 3.e-4, 5.e-5, 1.e-3			1.e-4
Observation reg. max learning rate	1.e-4, 3.e-4, 5.e-5, 1.e-3			1.e-3
Observation reg. gradient penalty	50, 100, 500, 1000	100		100

Table 2: R-A2C hyperparameters

	Range	**RepeatPrevious**	**StatelessCartpole**	**NoisyStatelessCartpole**
Optimizer	Adam, RMSProp	Adam	RMSProp	RMSProp
Gradient clipping	0.5, 1., 10.	10.	1.0	10.
Learning rate	3.e-5, 1.e-4, 3.e-4, 5.e-4, 1.e-3	5.e-4	5.e-4	1.e-4

Table 3: DVRL hyperparameters

	Range	**RepeatPrevious**	**StatelessCartpole**	**NoisyStatelessCartpole**
Optimizer	Adam, RMSProp	RMSProp	Adam	Adam
Gradient clipping	0.5, 1., 10.	1.	0.5	1.
Learning rate	3.e-5, 1.e-4, 3.e-4, 5.e-4, 1.e-3	3.e-5	5.e-4	1.e-3
Encoding loss factor	1,.1,.5,.05	1.	0.5	1.
Number of particles	5,10,15	10	10	5

Additional References

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

[44] 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, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5767–5777, 2017.

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