The Wasserstein Believer: Learning Belief Updates for Partially Observable Environments through Reliable Latent Space Models

Raphael Avalos*
AI Lab, Vrije Universiteit Brussel
Belgium
raphael.avalos@vub.be

Florent Delgrange*
AI Lab, Vrije Universiteit Brussel
University of Antwerp
Belgium
florent.delgrange@ai.vub.ac.be

Ann Nowé
AI Lab, Vrije Universiteit Brussel
Belgium

Guillermo A. Pérez
University of Antwerp
Flanders Make
Belgium

Diederik M. Roijers
AI Lab, Vrije Universiteit Brussel (BE)
Urban Innovation and R&D, City of Amsterdam (NL)

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. Current state-of-the-art algorithms use Recurrent Neural Networks (RNNs) 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 (WBU), 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.

KEYWORDS

Reinforcement Learning, Partial Observability, Representation Learning, Model Based

1 INTRODUCTION

The *Partially Observable Markov Decision Process* (POMDP) [39] is 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 [28], recommendation systems [42], and autonomous vehicles [30]. In contrast to in a *Markov Decision Process* (MDP), in a POMDP, the agent observes a noisy function 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 to histories 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 and, when used as states, define a *belief MDP* equivalent to the original POMDP [3]. While two closed-form expressions to compute the belief exist — using the full history or through the recursive belief update — they both require 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, current state-of-the-art algorithms focus on compressing the observation-action history with the help of *Recurrent Neural Networks* (RNNs) [21] in the hope of learning a sufficient statistic. 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 different techniques such as variational inference, particle filtering, and regularization through the ability to predict future observations [7, 17, 24]. However, *none of these techniques guarantee that the representation of histories induced by RNNs is suitable to optimize the return.*

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 action-observation 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 or medical trials, where additional sensors can be used during training in a laboratory setting. In *multi-agent* RL, using additional information during training is known as the *Centralized Training with a Decentralized Execution paradigm* [35] from which we draw inspiration.

We learn the latent model of the POMDP via a *Wasserstein auto-encoded MDP* (WAE-MDP) [12]. We then learn the *belief update*

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

Proc. of the Adaptive and Learning Agents Workshop (ALA 2023), Cruz, Hayes, Wang, Yates (eds.), May 29-30, 2023, London, UK, https://alaworkshop2023.github.io/. 2023.

network (BUN) by minimizing the Wasserstein distance with the exact belief update rule in the latent POMDP. To allow for complex belief distributions we use *Normalizing Flows* [26]. Unlike the current state-of-the-art 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*).

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.

This paper presents the theoretical results of our algorithm and the preliminary results on small environments with a modified belief update network that uses KL-divergence as a proxy for the Wasserstein distance. In future work we plan to implement the theoretical losses for the belief update network, test our algorithm on a larger set of POMDP environments and carry out ablative studies on the different components of our algorithm.

2 BACKGROUND

2.1 Probability Distributions

Notations. Let $\mathcal{X}$ be a complete and separable space and $\Sigma(\mathcal{X})$ be the set of all Borel subsets of $\mathcal{X}$. We write $\Delta(\mathcal{X})$ for the set of measures P defined on $\mathcal{X}$ and $\delta_a \in \Delta(\mathcal{X})$ for the *Dirac measure* with impulse $a \in \mathcal{X}$, having the following properties: $\delta_a = \lim_{\sigma\to 0} \mathcal{N}(a, \sigma^2)$ where $\mathcal{N}(a, \sigma^2)$ is the normal distribution with mean a and variance σ^2; $\delta_a(A) = 1$ if $a \in A$ and $\delta_a(A) = 0$ otherwise, for $A \in \Sigma(\mathcal{X})$; and $\int_{\mathcal{X}} \delta_a(x) f(x)\, dx = f(a)$ for any compactly supported function f. Given another space $\mathcal{Y}$, we use the standard conditional probability notation $P(\cdot \mid y)$ for $P\colon \mathcal{Y} \to \Delta(\mathcal{X})$.

Discrepancy measures. Let $P, Q \in \Delta(\mathcal{X})$, the divergence between P and Q can be measured according to the following discrepancies:

- the *Kullback-Leibler* (KL) divergence, defined as
$$D_{\mathrm{KL}}(P, Q) = \mathop{\mathbb{E}}_{x\sim P}\left[\log(P(x)/Q(x))\right].$$
- the solution of the *optimal transport problem* (OT), given by $\mathcal{W}_c(P, Q) = \inf_{\lambda\in\Lambda(P,Q)} \mathbb{E}_{x,y\sim\lambda}\, c(x, y)$, which is the *minimum cost of changing P into Q* [40], where $c\colon \mathcal{X}\times\mathcal{X} \to [0, \infty)$ is a cost function and $\Lambda(P, Q)$ is the set of all *couplings* of P and Q. If c is equal to a distance metric d over $\mathcal{X}$, then $\mathcal{W}_d$ is the *Wasserstein distance* between the two distributions.
- 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.

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 [37]. 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 drop the superscript when the context is clear.

Partial Observability. A *partially observable Markov decision process* (POMDP) [39] is a tuple $\mathcal{P} = \langle \mathcal{M}, \Omega, O\rangle$ where $\mathcal{M}$ is an MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$; Ω is a set of *observations*; and $O\colon \mathcal{S}\times\mathcal{A} \to \Delta(\Omega)$ is an *observation function* that defines the distribution of possible observations that may occur when the MDP $\mathcal{M}$ transitions to a state upon the execution of a particular action. An agent interacting 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 $o_{t+1} \sim O(\cdot \mid s_{t+1}, a_t)$ 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 $t \geq 0$, the 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 $\mathcal{M}$ [43]. 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

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

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 [38]. We write $\tau^*(a_{0:t}, o_{1:t+1}) = \tau(\cdot, a_t, o_{t+1}) \circ \cdots \circ \tau\big(\delta_{s_I}, a_0, o_1\big) = b_{t+1}$ for the recursive application of τ along the history. The belief update rule derived from τ allows to formulate $\mathcal{P}$ as a continuous[1] *belief MDP* $\mathcal{M}_{\mathcal{B}} = \langle \mathcal{B}, \mathcal{A}, \mathbf{P}_{\mathcal{B}}, \mathcal{R}_{\mathcal{B}}, b_I, \gamma\rangle$, where

$$\mathbf{P}_{\mathcal{B}}\big(b' \mid b, a\big) = \mathop{\mathbb{E}}_{s\sim b}\ \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)}\ \mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)} \delta_{\tau(b,a,o')}\big(b'\big)$$

is the transition function in the belief space, $\mathcal{R}_{\mathcal{B}}(b, a) = \mathbb{E}_{s\sim b}\mathcal{R}(s, a)$ is the reward function based on the current belief; and $b_I = \delta_{s_I}$ the initial belief state. As for all MDPs, $\mathcal{M}_{\mathcal{B}}$ as well as 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 enable the optimization of the expected return in $\mathcal{P}$ [3].

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

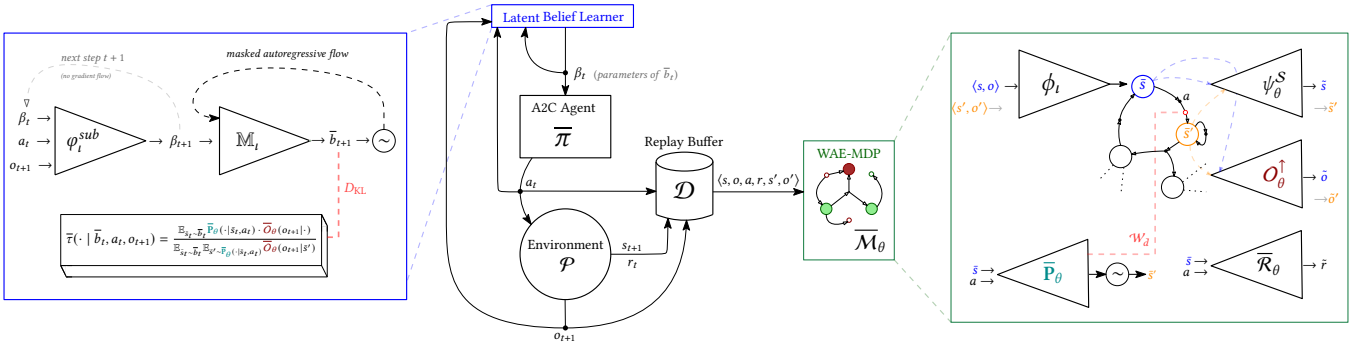

Figure 1: High-level picture of our Wasserstein Belief Updater framework. The *WAE-MDP* component is presented in Sect. 3, and the *Latent Belief Learner* is presented in Sect. 4. The learning of the different components happens in a round-robin fashion. The WAE-MDP learns from data collected by the RL agent and stored in a Replay Buffer. The Latent Belief Learner uses the latent transition function $\bar{\mathrm{P}}_\theta$ and observation decoder $\overline{O}_\theta$ 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 $\bar{b}_t$.

2.3 Latent Space Modeling

In the following, we write $f_\theta \colon \mathcal{X} \to \mathcal{Y}$ to denote a neural network f parameterized by θ, mapping inputs from $\mathcal{X}$ to outputs in $\mathcal{Y}$.

Latent MDPs. Given the original (continuous or very large, possibly unknown) environment $\mathcal{M}$, a *latent space model* is another (tractable, explicit) MDP $\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \mathcal{A}, \bar{\mathrm{P}}, \overline{\mathcal{R}}, \bar{s}_I, \gamma \rangle$ 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) [11] are latent space models that are trained based on the optimal transport from trajectory distributions, resulting from the execution of the RL agent policy in the real environment $\mathcal{M}$, to that reconstructed from the latent model $\overline{\mathcal{M}}_\theta$. 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 $\overline{\mathcal{M}}_\theta$ to be probably approximately *bisimilarly close* [12, 16, 27] 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 [13, 14].

Specifically, a WAE-MDP learns the following components:

$$\begin{aligned}
&\text{a } \textit{state embedding function} && \phi_\iota \colon \mathcal{S} \to \overline{\mathcal{S}},\\
&\text{a } \textit{latent transition function} && \bar{\mathrm{P}}_\theta \colon \overline{\mathcal{S}} \times \mathcal{A} \to \Delta(\overline{\mathcal{S}}),\\
&\text{a } \textit{latent reward function} && \overline{\mathcal{R}}_\theta \colon \overline{\mathcal{S}} \times \mathcal{A} \to \mathbb{R}, \text{ and}\\
&\text{a } \textit{state decoder} && \psi_\theta \colon \overline{\mathcal{S}} \to \mathcal{S};
\end{aligned}$$

the latter allowing to reconstruct original states from latent states. The objective function of WAE-MDPs — derived from the OT — incorporates *local losses* [15] that minimize the expected distance between the original and latent reward and transition functions:

$$\begin{aligned}
L_{\mathcal{R}} &= \mathop{\mathbb{E}}_{s,a\sim\mathcal{D}} \left| \mathcal{R}(s,a) - \overline{\mathcal{R}}_\theta(\phi_\iota(s), a) \right| \\
L_{\mathrm{P}} &= \mathop{\mathbb{E}}_{s,a\sim\mathcal{D}} \mathcal{W}_{\bar{d}} \left(\phi_\iota \mathbf{P}(\cdot \mid s,a), \bar{\mathrm{P}}_\theta(\cdot \mid \phi_\iota(s), a) \right) \qquad (2)
\end{aligned}$$

where $\mathcal{D} \in \Delta(\mathcal{S} \times \mathcal{A})$ is the distribution of experiences gathered by the RL agent when it interacts with $\mathcal{M}$, $\phi_\iota \mathbf{P}(\cdot \mid s, a)$ is the distribution of transitioning to $s' \sim \mathbf{P}(\cdot \mid s, a)$, then embedding it to the latent space $\bar{s}' = \phi_\iota(s')$, and $\bar{d}$ is a metric on $\overline{\mathcal{S}}$.

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 $\mathcal{P} = \langle \mathcal{M}, \Omega, O \rangle$ 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 O. To do so, we make the agent learn an internal, explicit representation of the experiences gathered, through a latent space model. The trick to getting the agent 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 $\mathcal{P}$ through a WAE-MDP by considering the POMDP $\mathcal{P}^\uparrow = \left\langle \mathcal{M}_\Omega, \Omega, O^\uparrow \right\rangle$, where

(1) the underlying MDP is refined to encode the observations in its state space, defined as $\mathcal{M}_\Omega = \langle \mathcal{S}_\Omega, \mathcal{A}, \mathbf{P}_\Omega, \mathcal{R}_\Omega, s_I^\star, \gamma \rangle$ so that $\mathcal{S}_\Omega = \mathcal{S} \times (\Omega \cup \{\star\})$; $\star$ is a special observation symbol indicating that no observation has been perceived yet; $\mathbf{P}_\Omega(s', o' \mid s, o, a) = \mathbf{P}(s' \mid s, a) \cdot O(o' \mid s', a)$; $\mathcal{R}_\Omega(\langle s, o \rangle, a) = \mathcal{R}(s, a)$; and $s_I^\star = \langle s_I, \star \rangle$.

(2) the observation function $O^{\uparrow}\colon \mathcal{S}_\Omega \to \Omega$ is now deterministic and defined as the projection of the refined state on the observation space, with $O^{\uparrow}(\langle s, o\rangle) = o$.

The POMDPs $\mathcal{P}$ and $\mathcal{P}^{\uparrow}$ are equivalent [6]: $\mathcal{P}^{\uparrow}$ captures the stochastic dynamics of the observations 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 $\overline{M}_\theta = \langle \overline{\mathcal{S}}, \mathcal{A}, \overline{\mathbf{P}}_\theta, \overline{\mathcal{R}}_\theta, \bar{s}_I\rangle$ linked to $\mathcal{M}_\Omega$ via the embedding $\phi_\iota\colon \mathcal{S}_\Omega \to \overline{\mathcal{S}}$, and we achieve this via the WAE-MDP framework. Not only does the latter allow for the learning of the observation dynamics through $\overline{\mathbf{P}}_\theta$, but it also enables the learning of the deterministic observation function $O^{\uparrow}$ through the use of the state decoder ψ_θ, by decomposing the latter in two networks $\psi^{\mathcal{S}}_\theta\colon \overline{\mathcal{S}} \to \mathcal{S}$ and $O^{\uparrow}_\theta\colon \overline{\mathcal{S}} \to \Omega$, which yield $\psi_\theta(\bar{s}) = \langle \psi^{\mathcal{S}}_\theta(\bar{s}), O^{\uparrow}_\theta(\bar{s})\rangle$. This way, the WAE-MDP procedure learns all the components of $\mathcal{P}^{\uparrow}$, being equivalent to $\mathcal{P}$. With this model, we construct a *latent POMDP* $\overline{\mathcal{P}}_\theta = \langle \overline{\mathcal{M}}_\theta, \Omega, \overline{O}_\theta\rangle$, where the observation function outputs a normal distribution centered in $O^{\uparrow}_\theta$: $\overline{O}_\theta(\cdot \mid \bar{s}) = \mathcal{N}(O^{\uparrow}_\theta(\bar{s}), \sigma^2)$. Note that the deterministic function is retrieved as the variance approaches zero. However, it is worth mentioning that the smoothness of $\overline{O}_\theta$ 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 $\overline{\mathcal{P}}_\theta$ allows to reason on the belief space to optimize the expected return. Formally, at any time step $t \geq 0$ 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 $\bar{s}_{t+1} \in \overline{\mathcal{S}}$,

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

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 enabled by processing h through the latent space, to obtain a belief $\bar{\tau}^*(h) = \bar{b}$ over $\overline{\mathcal{S}}$ and execute the action prescribed by $\bar{\pi}(\cdot \mid \bar{b})$. Training $\overline{M}_\theta$ 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 $\overline{M}_\theta$ to learn a *latent belief encoder* $\varphi_\iota\colon \overline{\mathcal{B}} \times \mathcal{A} \times \overline{\mathcal{S}} \to \overline{\mathcal{B}}$ that *approximates the belief update function* via *some* discrepancy D:

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

for $h \in (\mathcal{A} \cdot \Omega)^*$ drawn from *some* distribution. The belief encoder φ_ι 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 $a \sim \bar{\pi}\big(\cdot \mid \varphi^*_\iota(h)\big)$.*

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_ι suitable to optimize the expected return in $\mathcal{P}$?"*. Clearly, the guarantees that can be obtained through this approach depend on the history distribution and the discrepancy chosen. We dedicate the rest of this section to answering those questions through a rigorous theoretical discussion on the distribution and losses required to obtain such learning guarantees.

3.2 Losses and Theoretical Guarantees

We start by formally defining the process allowing to draw experiences from the interaction and the related history distribution.

Episodic RL process. The RL procedure is *episodic* if the environment $\mathcal{P}$ embeds a special *reset state* $s_{\text{reset}} \in \mathcal{S}$ so that (i) under any policy π, the environment is almost surely eventually reset: $\mathbb{P}^{\mathcal{M}}_\pi(\{\, s_{0:\infty}, a_{0:\infty} \mid \exists t > 0, s_t = s_{\text{reset}} \,\}) = 1$; (ii) when reset, the environment transitions to the initial state: $\mathbf{P}(s_I \mid s_{\text{reset}}, a) > 0$ and $\mathbf{P}(\mathcal{S} \setminus \{\, s_I, s_{\text{reset}} \,\} \mid s_{\text{reset}}, a) = 0$ for all $a \in \mathcal{A}$; and (iii) *the reset state is observable*: there is an observation $o^\star \in \Omega$ so that $O\big(o^\star \mid s', a\big) = 0$ when $s' \neq s_{\text{reset}}$, and $O(\cdot \mid s_{\text{reset}}, a) = \delta_{o^\star}$ for $a \in \mathcal{A}$. An *episode* is a history $\langle a_{0:T-1}, o_{1:T}\rangle$ where $O(o_1 \mid s_I, a_0) > 0$ and $o_T = o^\star$.

Assumption 3.1. *The environment $\mathcal{P}$ is an episodic process.*

Lemma 3.2. *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 interacts with $\mathcal{P}$, by executing $\bar{\pi}$.*

Proof sketch. Build a *history unfolding* as the MDP whose state space consists of all histories, and keeps track of the current history of $\mathcal{P}$ at any time of the interaction. The resulting MDP remains episodic since it is equivalent to $\mathcal{P}$: the former mimics the behaviors of the latter under $\bar{\pi}$. All episodic processes are *ergodic* [23], which guarantees the existence of such a distribution. The complete proof including the details of this construction and the equivalence relation between the unfolding and $\mathcal{P}$ is in Appendix A. □

Local losses. For the sake of clarity, we reformulate the local losses (Eq. 2) to align with the fact that the experiences come from a distribution that generates histories instead of state-action pairs, and the states processed by the WAE-MDP are those of $\mathcal{M}_\Omega$:

$$\begin{aligned} L_{\mathcal{R}} &= \mathop{\mathbb{E}}_{s,o,a \sim \mathcal{H}_{\bar{\pi}}} \Big| \mathcal{R}(s,a) - \overline{\mathcal{R}}_\theta(\phi_\iota(s,o), a) \Big| \\ L_{\mathbf{P}} &= \mathop{\mathbb{E}}_{s,o,a \sim \mathcal{H}_{\bar{\pi}}} \mathcal{W}_{\tilde{d}}\Big(\phi_\iota \mathbf{P}_\Omega(\cdot \mid s,o,a), \overline{\mathbf{P}}_\theta(\cdot \mid \phi_\iota(s,o), a)\Big) \end{aligned} \tag{5}$$

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) $a \sim \bar{\pi}\big(\cdot \mid \varphi^*_\iota(h)\big)$.

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 transitions of $\mathcal{M}_\Omega$ encountered while executing $\bar{\pi}$. We also introduce an *observation loss* in addition to the reconstruction loss that allows learning $\overline{O}_\theta$:

$$\begin{aligned} L_O = \mathop{\mathbb{E}}_{s,o,a \sim \mathcal{H}_{\bar{\pi}}} \mathop{\mathbb{E}}_{s' \sim \mathbf{P}(\cdot \mid s,a)} \\ d_{TV}\Big(O(\cdot \mid s', a), \mathop{\mathbb{E}}_{o' \sim O(\cdot \mid s', a)} \overline{O}_\theta\big(\cdot \mid \phi_\iota(s', o')\big)\Big) \end{aligned} \tag{6}$$

Intuitively, L_O provides a way to gauge the variation between the latent state reconstruction using $O^{\uparrow}_\theta$ and the actual observation generation process, allowing us to set the variance of $\overline{O}_\theta$.

Belief Losses. Setting D as the Wasserstein between the true latent belief update and our belief encoder leads to the following loss:

$$L_{\bar{\tau}} = \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\bar{\pi}}} \mathcal{W}_{\bar{d}}\left(\bar{\tau}^*(h), \varphi_\iota^*(h)\right) \tag{7}$$

In addition, we argue that the following reward and transition regularizers are required to bound the gap between the fully observable model $\overline{\mathcal{M}}_\theta$ and the partially observable one $\overline{\mathcal{P}}_\theta$:

$$L^{\varphi}_{\overline{\mathcal{R}}} = \mathop{\mathbb{E}}_{h,s,o,a\sim\mathcal{H}_{\bar{\pi}}} \mathop{\mathbb{E}}_{\bar{s}\sim\varphi_\iota^*(h)} \left|\overline{\mathcal{R}}_\theta(\phi_\iota(s,o),a) - \overline{\mathcal{R}}_\theta(\bar{s},a)\right|$$
$$L^{\varphi}_{\overline{\mathbf{P}}} = \mathop{\mathbb{E}}_{h,s,o,a\sim\mathcal{H}_{\bar{\pi}}} \mathop{\mathbb{E}}_{\bar{s}\sim\varphi_\iota^*(h)} \mathcal{W}_{\bar{d}}\left(\bar{\mathbf{P}}_\theta(\cdot \mid \phi_\iota(s,o),a), \bar{\mathbf{P}}_\theta(\cdot \mid \bar{s},a)\right) \tag{8}$$

The two aim at regularizing φ_ι 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 allows the optimization of those losses. Intuitively, the belief loss and the related two regularizers can be optimized *on-policy*, i.e., coupled with the optimization of $\bar{\pi}$ that is used to generate the episodes.

Value difference bounds. We provide guarantees suited for partial observability, 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: $\mathrm{V}_\pi(h) = \mathbb{E}_\pi^{\mathcal{M}_\mathcal{B}[\tau^*(h)]}\left[\sum_{t=0}^{\infty}\gamma^t\cdot\mathcal{R}_\mathcal{B}(b_t,a_t)\right]$, where $\mathcal{M}_\mathcal{B}[\tau^*(h)]$ is the belief MDP of $\mathcal{P}$ obtained by changing the initial belief state by $\tau^*(h)$. Similarly, we write $\overline{\mathrm{V}}_{\bar{\pi}}$ for the values of the latent policy $\bar{\pi}$ in $\overline{\mathcal{P}}_\theta$. The following two Theorems state that, when the local and belief losses are all minimized and go to zero,

(1) the agent behaviors are the same in the original and latent POMDPs when the former executes a latent policy by using our belief encoder φ_ι to maintain a latent belief from the interaction; and
(2) any pair of histories whose belief representation induced by φ_ι are close also have close optimal behaviors.

While (1) guarantees the "average equivalence" of the two models and justifies the usage of $\overline{\mathcal{P}}$ as model of the environment, (2) shows that φ_ι induces a suitable representation of the history to optimize the expected return.

Theorem 3.3. *Assume that the WAE-MDP is at the zero-temperature limit (i.e., $\lambda\to 0$) and let $\overline{\mathcal{R}}^\star = \sup_{\bar{s},a}|\overline{\mathcal{R}}(\bar{s},a)|$, $K_{\overline{\mathrm{V}}} = \overline{\mathcal{R}}^\star/_{1-\gamma}$, then for any latent policy $\bar{\pi}\colon\overline{\mathcal{B}}\to\Delta(\mathcal{A})$, the values of $\mathcal{P}$ and $\overline{\mathcal{P}}_\theta$ are guaranteed to be bounded by those losses in average:*

$$\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\bar{\pi}}}\left|\mathrm{V}_{\bar{\pi}}(h) - \overline{\mathrm{V}}_{\bar{\pi}}(h)\right| \le \frac{L_\mathcal{R} + L^{\varphi}_{\overline{\mathcal{R}}} + \overline{\mathcal{R}}^\star L_{\bar{\tau}} + \gamma K_{\overline{\mathrm{V}}}\cdot\left(L_\mathbf{P} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\bar{\tau}} + L_O\right)}{1-\gamma}. \tag{9}$$

Theorem 3.4. *Let $\bar{\pi}^*$ be the optimal policy of the POMDP $\overline{\mathcal{P}}_\theta$, then for any couple of histories $h_1, h_2 \in (\mathcal{A}\cdot\Omega)^*$ mapped to latent beliefs through $\varphi_\iota^*(h_1) = \bar{b}_1$ and $\varphi_\iota^*(h_2) = \bar{b}_2$, the belief representation induced by φ_ι yields:*

$$\left|\mathrm{V}_{\bar{\pi}^*}(h_1) - \mathrm{V}_{\bar{\pi}^*}(h_2)\right| \le K_{\overline{\mathrm{V}}}\mathcal{W}_{\bar{d}}\left(\bar{b}_1,\bar{b}_2\right) + \frac{L_\mathcal{R} + L^{\varphi}_{\overline{\mathcal{R}}} + \left(K_{\overline{\mathrm{V}}} + \overline{\mathcal{R}}^\star\right)L_{\bar{\tau}} + \gamma K_{\overline{\mathrm{V}}}\cdot\left(L_\mathbf{P} + L^{\varphi}_{\overline{\mathbf{P}}} + L_O\right)}{1-\gamma} \cdot\left(\mathcal{H}^{-1}_{\bar{\pi}^*}(h_1) + \mathcal{H}^{-1}_{\bar{\pi}^*}(h_2)\right) \tag{10}$$

when the WAE-MDP temperature λ goes to zero.

Notice that the right-hand side of Eq. 9 and Eq. 10 are both only composed of constants multiplied by the losses, which are null as the losses go to zero. We prove those claims in Appendix B.

4 LEARNING TO BELIEVE

In this section, we assume that we have access to the latent model learned by the WAE-MDP. The goal of the belief updater is to compute the belief states of the "behaviorally equivalent" (Thm. 3.3) latent POMDP $\overline{\mathcal{P}}_\theta$ so that an RL agent can learn to optimize a latent policy based on those latent beliefs, guaranteed to be a suitable representation of the histories for optimizing the return (Thm. 3.4). The rest of this section is composed as follows: first we analyze the belief update rule $\bar{\tau}$, second we explain our belief encoder architecture and our design choices, third we describe our training procedure, and last we explain how the agent learns its policy.

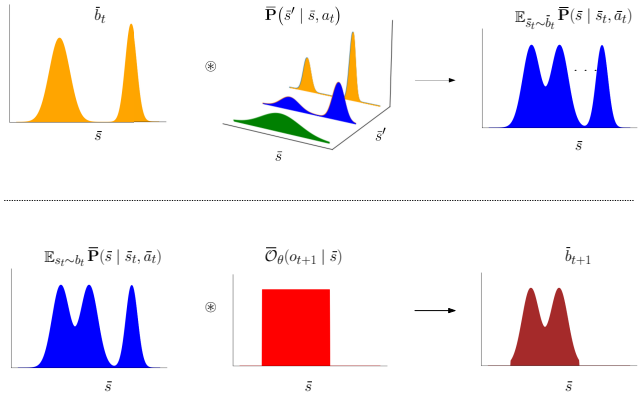

Figure 2: The Belief Update rule: a) transformation of the current belief $\bar{b}_t$ with the transition probability function $\bar{\mathbf{P}}$, evaluated on the current action a_t, into the next state probability density; b) filtering out the next states that could not have produced the next observation 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. As shown in Fig. 2, the update rule is divided into two steps. First, the current belief distribution $\bar{b}_t$ is used to marginalise the latent transition function $\bar{\mathbf{P}}_\theta$ 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

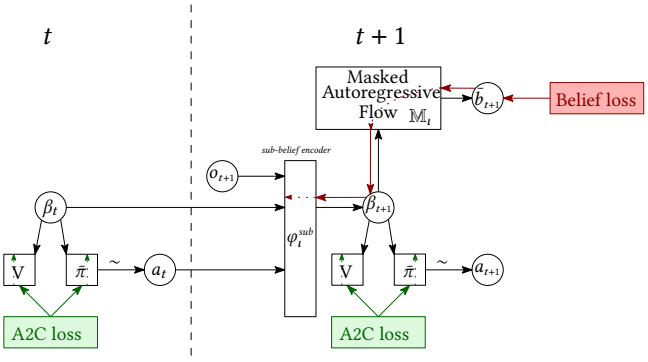

Figure 3: The Belief A2C agent; the trick to make the agent learn a policy conditioned on latent belief distributions is to condition them on *sub-beliefs* in place, i.e., the parameters β_t of each belief distribution. We therefore define a *sub-belief encoder* $\varphi_t^{\text{sub}}(\beta_t, a_t, o_{t+1}) = \beta_{t+1}$ which generates the next *sub-belief*, and a MAF $\mathbb{M}_t(\beta_{t+1}) = \bar{b}_{t+1}$ is then used to retrieve the latent belief $\bar{b}_{t+1}$ linked to the parameters β_{t+1}. The red arrows represent the flow of the belief loss gradients, and the green arrows the flow of the A2C loss gradients. Both gradients do not back-propagate through time.

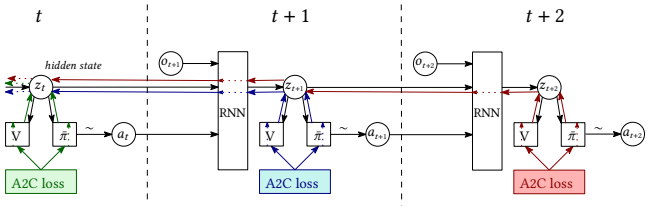

Figure 4: The RNN A2C agent uses back-propagation through time (BPTT): the RNN leverages gradients from future time steps to improve its compression of the history for learning a policy and value function. The colored arrows represent the flow of the A2C loss across time.

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 $\overline{O}_\theta$ 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) [36], 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 [11] that learns multivariate, latent, relaxed distributions which become discrete (binary) in the zero-temperature limit of the WAE-MDP.

We define the *sub-belief* as the vector that embed the parameters of the belief distribution, the MAF allowing the transformation of sub-beliefs into beliefs. The sub-belief functions similarly to the hidden states in an RNN as it is updated recursively. However, as we do not allow gradients to back-propagate through time (BPTT), we use a simple feed-forward network instead of a GRU or LSTM [10, 22]. This choice is motivated by the difference between the nature of the RNN hidden states in the partially observable version of A2C [32] (R-A2C), and sub-beliefs.

On the one hand, the goal of the RNN hidden states is to compress the history into a fixed sized 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. Therefore, using the gradients of future time steps with BPTT 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 regularized to follow the belief update rule. Since beliefs of earlier time steps are easier to compute as the history is smaller, gradients from future time steps tend to be of worse quality compared to the current one. Therefore, disabling BPTT effectively improves the quality of the update to learn the sub-beliefs. Allowing gradients from the belief loss of future time steps to flow would negatively impact learning, as it would lead to an accumulation of errors in the update of the sub-belief. Fig. 3 and 4 outline the differences between both methods and the way the RL gradients flow.

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_t (Eq. 7) to leverage the theoretical learning guarantees of Thm. 3.3 and 3.4. However, Wasserstein optimization is known to be challenging, often requiring the use of 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 [9] 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, 12].

On-policy KL divergence. Using D_{KL} as a proxy for the Wassertein distance allows to close the gap between $\bar{\tau}$ and φ_t while optimizing the policy; at any time step $t \geq 0$, 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 given by

$$D_{\mathrm{KL}}\Big(\varphi_t\Big(\bar{b}_t, a_t, o_{t+1}\Big) \parallel \bar{\tau}\Big(\bar{b}_t, a_t, o_{t+1}\Big)\Big) =$$

$$\mathop{\mathbb{E}}_{\bar{s}_{t+1}\sim\varphi_t\left(\bar{b}_t,a_t,o_{t+1}\right)}\Bigg[\log\varphi_t\Big(\bar{s}_{t+1} \mid \bar{b}_t, a_t, o_{t+1}\Big) - \log \mathop{\mathbb{E}}_{\bar{s}\sim\bar{b}_t} \bar{\mathbf{P}}_\theta(\bar{s}_{t+1} \mid \bar{s}, a_t)$$

$$- \log \overline{O}_\theta(o_{t+1} \mid \bar{s}_{t+1})\Bigg] + \log\left(\mathop{\mathbb{E}}_{\bar{s}\sim\bar{b}_t} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot\mid\bar{s},a_t)} \overline{O}_\theta\big(o_{t+1} \mid \bar{s}'\big)\right)$$

The divergence is composed of 4 terms, the first one corresponds to the negative entropy of φ_t, the second term ensures that the belief update follows the state transition function of the latent MDP, the third term filters out the latent states which are not associated with the observation o_{t+1}, and the fourth one is a normalizing factor. We minimize the KL divergence D_{KL} by gradient descent on the Monte-Carlo estimate of the divergence:

$$\nabla_t D_{\mathrm{KL}}\Big(\varphi_t\Big(\bar{b}_t, a_t, o_{t+1}\Big) \parallel \bar{\tau}\Big(\bar{b}_t, a_t, o_{t+1}\Big)\Big) =$$

$$\nabla_t \mathop{\mathbb{E}}_{\bar{s}_{t+1}\sim\varphi_t\left(\bar{b}_t,a_t,o_{t+1}\right)}\Bigg[\log\varphi_t\Big(\bar{s}_{t+1} \mid \bar{b}_t, a_t, o_{t+1}\Big)$$

$$- \log \mathop{\mathbb{E}}_{\bar{s}\sim\bar{b}_t} \bar{\mathbf{P}}_\theta(\bar{s}_{t+1} \mid \bar{s}, a_t) - \log \overline{O}_\theta(o_{t+1} \mid \bar{s}_{t+1})\Bigg]$$

Notice that the fourth term of this KL divergence does not depends on φ_t and therefore yields a gradient of zero.

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.

Learning the policy is enabled by feeding the sub-belief vector as input of the former. As mentioned earlier, we do not allow the RL agent to optimize the parameters of the belief encoder. In our experiments, we use A2C as on-policy algorithm but our method can be applied to any on-policy algorithm.

5 EXPERIMENTS

We evaluate the performance of our algorithm on a subset of the POPGym environments [33], which were designed to test different features required for generalization to various POMDPs. These features include short-term memory for control and long-term memory capacity. Our algorithm is tested on two distinct environments:

- *Repeat Previous*: the first environment tests the agent's ability to maintain and retrieve long-term memory. At the start of each episode, two decks of cards are shuffled and the agent is presented 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 there are no more cards in the deck. The agent receives a positive for every correct card and a negative reward otherwise. The rewards are scaled so that the maximum return is 1.
- *Stateless Cart-Pole*: the second environment challenges the agent to control a cart that can move left or right on a rail, while maintaining the attached pole within a specific angle range. The system state is comprised of the cart position and velocity, as well as the pole angular position and velocity. In the partially observable version, the velocity components are hidden, requiring the agent to rely on its short-term memory to estimate them. The agent receives a positive reward for every time step, and the maximum return is 1.

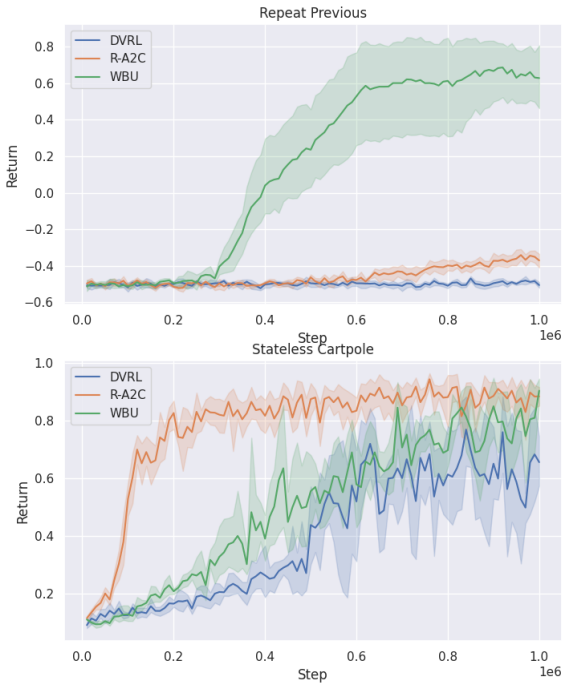

Figure 5: Comparison of the evolution of the un-discounted cumulative return between WBU, R-A2C and DVRL. For both environment the maximum return is 1

We compare the performance of WBU in those two environments with R-A2C and DVRL [24], an algorithm based on R-A2C that uses a Variational Auto-Encoder and particle filtering to maintain some belief distribution. We train 5 instances of each algorithm for 1 million time steps using 16 parallel environments.

In the Repeat Previous environment (Fig. 5 top), WBU is the only algorithm out of the three that can memorize almost perfectly the 8 last cards and output the suit of the 8th card. While after 400k environmental steps, WBU already obtains positive return, which corresponds to getting half of the cards correct, R-A2C only starts to slowly improve over its initial return of -0.5 after 600k timestep to reach less than -0.3 at the end of the learning. DVRL on the other hand does not seem to learn anything as its returns do not evolve. This first experiment showcases the capacity of our algorithm to remember previous observations and retrieve them when needed.

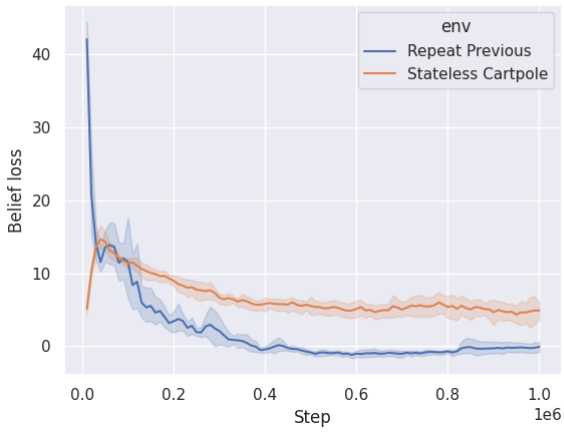

Figure 6: Evolution of the belief loss during learning in the two environments. While the belief loss is a divergence and should therefore be positive its Monte Carlo approximation can be negative as it is the case in the Repeat Previous experiment.

In the stateless version of Cart-Pole, the RNN of R-A2C can learn relatively fast how to estimate the velocities of both the cart and the pole as it manages to score more than 0.5 in returns after only 100k steps. WBU and R-A2C present a similar learning curve, with the exception that at the end of the learning, WBU is able to catch up with R-A2C performance, while DVRL appears to plateau. This experiment shows that WBU can leverage short-term memory for control.

The evolution of the belief loss of WBU for the two environments is presented in Figure 6. For both environments, the belief loss is decreasing and converges. We note an increase in belief loss at the beginning of the learning. This is because we learn the model in parallel.

To summarize, these environments provide a rigorous test of WBU's ability to generalize to various partially observable problems and demonstrate its capability to learn and effectively utilize a belief state representation.

6 RELATED WORK

POMDPs pose a significant challenge to RL due to the loss of the Markovian property — the next observation distribution do not only depend on the current state, but on the entire action-observation history. As a consequence, this history need be considered to derive an optimal policy, making it essential to obtain a sufficient statistic such as belief states [25]. Most deep-RL methods use RNNs to compress the action-observation history into a fixed-size vector [21] and apply regularization techniques to improve the likelihood of obtaining a belief. These techniques include generative models [8, 19, 20], particle filtering [24, 31], and predicting distant observations [17, 18]. However, most algorithms assume the beliefs to be simple distributions of states, such as Gaussian distributions, limiting their applicability [17, 20, 29]. We note that using Normalizing Flows for the belief distribution as been experienced in FORBES [8]. However, FORBES does not condition its policy on the beliefs but rather on sample latent states (which is known to be sub-optimal). Some works also focus on specific types of POMDPs, such as building compact latent representation of images for visual motor tasks [29], or environment where the observation are masked states with Gaussian noise [41]. Compared to those works, WBU's beliefs are not trained to help the learning but to follow the belief update rule.

While accessing the state is common in partially observable deep multi-agent RL, and known as the centralized training decentralized execution paradigm [4, 35], it is not a common practice in single-agent RL with the exception of kernel-POMDPs [34] that uses the states to build models based on RKHSs.

Finally, the works of [12, 15] study similar value difference bounds to ours in the context of fully observable environments, guaranteeing the quality of (i) the latent space model learned, and (ii) the representation induced by ϕ_t for optimizing the returns. They further link their bounds with bisimulation theory (e.g., [16, 27]): in a nutshell, bisimulation define an equivalence relation between models in terms of the behaviors induced when a a policy is executed (in particular, the optimal policy). We defer as future work the study of bisimulation metrics [13, 14] in the context of POMDPs.

7 CONCLUSION

Wasserstein Belief Updater provides a novel approach that approximates directly belief update for POMDPs, in contrast to state-of-the-art 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 can 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 on two environments demonstrate the potential of our approach. While our current implementation uses the Kullback-Leibler divergence as a proxy for the Wasserstein distance, we plan to improve our algorithm in future work by incorporating the true Wasserstein distance and additional theoretical losses (Eq. 6 and 8). We also aim to explore the use of simulated trajectories for policy learning, which is theoretically enabled through the model and representation quality guarantees yielded by Thm 3.3 and 3.4, to evaluate our approach on a broader range of environments, and carry ablative studies. Overall, our proposed Wasserstein Belief-Updater 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.

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. 2003. An Introduction to MCMC for Machine Learning. *Machine Learning* 50, 1-2 (2003), 5–43. https://doi.org/10.1023/A:1020281327116

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

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

[4] Raphael Avalos, Mathieu Reymond, Ann Nowé, and Diederik M. Roijers. 2022. 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)*.

[5] J.M. Borwein and A.S. Lewis. 2005. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer New York. https://books.google.be/books?id=TXWzqEkAa7IC

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

[7] Xiaoyu Chen, Yao Mu, Ping Luo, Shengbo Li, and Jianyu Chen. 2022. Flow-based Recurrent Belief State Learning for POMDPs. (5 2022). https://doi.org/10.48550/arxiv.2205.11051

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

[9] Siddhartha Chib and Edward Greenberg. 1995. Understanding the Metropolis-Hastings Algorithm. *The American Statistician* 49, 4 (1995), 327–335. http://www.jstor.org/stable/2684568

[10] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

[11] Florent Delgrange, Ann Nowe, and Guillermo Perez. 2023. Wasserstein Auto-encoded MDPs: Formal Verification of Efficiently Distilled RL Policies with Many-sided Guarantees. In *International Conference on Learning Representations*. https://openreview.net/forum?id=JLLTtEdh1ZY

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

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

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

[15] Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. 2019. DeepMDP: Learning Continuous Latent Space Models for Representation Learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2170–2179. http://proceedings.mlr.press/v97/gelada19a.html

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

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

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

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

[20] Danijar Hafner, Timothy P Lillicrap, Mohammad Norouzi, and Jimmy Ba. 2021. Mastering Atari with Discrete World Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=0oabwyZbOu

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

[22] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (11 1997), 1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735

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

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

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

[26] Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. 2021. Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 11 (nov 2021), 3964–3979. https://doi.org/10.1109/tpami.2020.2992934

[27] Kim Guldstrand Larsen and Arne Skou. 1989. 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*. ACM Press, 344–352. https://doi.org/10.1145/75277.75307

[28] Mikko Lauri, David Hsu, and Joni Pajarinen. 2023. Partially Observable Markov Decision Processes in Robotics: A Survey. *IEEE Transactions on Robotics* 39, 1 (2023), 21–40. https://doi.org/10.1109/TRO.2022.3200138

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

[30] Wei Liu, Seong-Woo Kim, Scott Pendleton, and Marcelo H Ang. 2015. Situation-aware decision making for autonomous driving on urban road using online POMDP. In *2015 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 1126–1133.

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

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

[33] Steven Morad, Ryan Kortvelesy, Matteo Bettini, Stephan Liwicki, and Amanda Prorok. 2023. POPGym: Benchmarking Partially Observable Reinforcement Learning. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=chDrutUTs0K

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

[35] Frans A. Oliehoek, Matthijs T.J. Spaan, and Nikos Vlassis. 2008. Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research* 32 (10 2008), 289–353. https://doi.org/10.1613/jair.2447

[36] George Papamakarios, Theo Pavlakou, and Iain Murray. 2017. Masked Autoregressive Flow for Density Estimation. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/6c1da886822c67822bcf3679d04369fa-Paper.pdf

[37] Martin L. Puterman. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley. https://doi.org/10.1002/9780470316887

[38] Richard D. Smallwood and Edward J. Sondik. 1973. The Optimal Control of Partially Observable Markov Processes over a Finite Horizon. *Oper. Res.* 21, 5 (1973), 1071–1088. https://doi.org/10.1287/opre.21.5.1071

[39] Matthijs TJ Spaan. 2012. Partially observable Markov decision processes. *Reinforcement learning: State-of-the-art* (2012), 387–414.

[40] Cédric Villani. 2009. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, Berlin, Heidelberg. 93–111 pages. https://doi.org/10.1007/978-3-540-71050-9_6

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

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

[43] K.J Åström. 1965. Optimal control of Markov processes with incomplete state information. *J. Math. Anal. Appl.* 10, 1 (1965), 174–205. https://doi.org/10.1016/0022-247X(65)90154-X

A PROOF OF LEMMA 3.2: STATIONARITY OVER HISTORIES

Let us restate the Lemma:

Lemma A.1. *Let $\mathcal{P}$ be an episodic POMDP with action space $\mathcal{A}$ and observation space Ω (Assumption 3.1). 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 $\bar{\pi}$ conditioned over the beliefs of a latent POMDP $\overline{\mathcal{P}}$, the latter sharing the action and observation spaces of $\mathcal{P}$.*

Therefore, we want to show the existence of a *limiting distribution* over histories, when a latent policy is executed. Before going further, we formally introduce the notions of *memory-based policies*, *Markov Chains*, and *limiting distributions in Markov Chains*.

A.1 Preliminaries

Definition A.2 (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* $\pi = \big\langle Q, \pi_\alpha, \pi_\mu, q_I \big\rangle$, 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 $\mathcal{P} = \langle \mathcal{M}, \Omega, O \rangle$ 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 A.1. Then, any latent (stationary) policy $\bar{\pi} \colon \overline{\mathcal{B}} \to \Delta(\mathcal{A})$ 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 $\bar{\pi}' = \langle \overline{\mathcal{B}}, \bar{\pi}_\alpha, \bar{\pi}_\mu, \delta_{\bar{s}_I} \rangle$, keeping track in its memory of the current latent belief $\bar{b} \in \overline{\mathcal{B}}$. This enables the agent to take its decisions solely based on the latter: $\bar{\pi}_\alpha(\cdot \mid b, \bar{b}) = \bar{\pi}(\cdot \mid \bar{b})$. When the belief MDP transitions to the next belief b', the memory is then updated according to the observation dynamics:

$$\bar{\pi}_\mu(\bar{b}' \mid b, \bar{b}, a, b') = \frac{\mathbb{E}_{s\sim b}\, \mathbb{E}_{s'\sim \mathbf{P}(\cdot|s,a)}\, \mathbb{E}_{o'\sim O(\cdot|s',a)}\, \delta_{\varphi_\iota(\bar{b},a,o')}\Big(\bar{b}'\Big) \cdot \delta_{\tau(b,a,o')}(b')}{\mathbb{E}_{s\sim b}\, \mathbb{E}_{s'\sim \mathbf{P}(\cdot|s,a)}\, \mathbb{E}_{o'\sim O(\cdot|s',a)}\, \delta_{\tau(b,a,o')}(b')} \qquad \text{if } b' \neq \delta_{s_{\text{reset}}}, \text{ and}$$

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

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

Definition A.3 (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 $\pi = \big\langle Q, \pi_\alpha, \pi_\mu, q_I \big\rangle$ 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.,

$$\mathbf{P}_\pi\big(\langle s', q' \rangle \mid \langle s, q \rangle\big) = \mathop{\mathbb{E}}_{a\sim\pi_\alpha(\cdot|s,q)} \pi_\mu\big(q' \mid s, q, a, s'\big) \cdot \mathbf{P}\big(s' \mid s, a\big), \text{ and}$$

- the rewards are averaged over the possible actions produced by the next action function, i.e., $\mathcal{R}_\pi(\langle s, q \rangle) = \mathbb{E}_{a\sim\pi_\alpha(\cdot|s,q)}\, \mathcal{R}(s, a)$.

Furthermore, the probability measure $\mathbb{P}_\pi^{\mathcal{M}}$ is actually the unique probability measure defined over the measurable infinite trajectories of the MC $\mathcal{M}^\pi$ [37].

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 existence conditions of such a distribution.

Definition A.4 (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 distribution $\xi_\pi^t : \mathcal{S} \to \Delta(\mathcal{S})$ 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 $\bar{\pi}$.

A.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 all the possible histories (i.e., sequence of actions and observations) that can be encountered in $\mathcal{P}$, i.e., $\mathcal{S}_{\mathcal{H}} = (\mathcal{A} \cdot \Omega)^* \cup \{ \star, h_{\text{reset}} \}$, which additionally embeds a special symbol $\star$ indicating that no observation has been perceived yet (cf. the definition of $\mathcal{M}_{\Omega}$ in Sect. 3.1) 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.,

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

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

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

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

Definition A.5 (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 [16, 27]. The relation can be extended to compare two MDPs by considering the disjoint union of their state space.

Lemma A.6. *Let $\mathcal{P}$ be the POMDP of Lemma A.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 $\overline{\pi}^{\clubsuit}(\cdot \mid h) = \overline{\pi}\big(\cdot \mid \varphi_{\iota}^*(h)\big)$, and the memory-based policy $\overline{\pi}^{\diamond}$ for $\mathcal{M}_{\mathcal{B}}$ encoded by the Mealy machine detailed in Example 2. Then, $\mathcal{M}_{\mathcal{H}}^{\overline{\pi}^{\clubsuit}}$ and $\mathcal{M}_{\mathcal{B}}^{\overline{\pi}^{\diamond}}$ are in stochastic bisimulation.*

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

$$\begin{aligned}
\mathbf{P}_{\overline{\pi}^{\diamond}}\left(b', \bar{b}' \mid b, \bar{b}\right) &= \mathop{\mathbb{E}}_{a \sim \overline{\pi}\left(\cdot \mid \bar{b}\right)} \overline{\pi}_{\mu}\left(\bar{b}' \mid b, \bar{b}, a, b'\right) \cdot \mathbf{P}_{\mathcal{B}}\left(b' \mid b, a\right) \\
&= \mathop{\mathbb{E}}_{a \sim \overline{\pi}\left(\cdot \mid \bar{b}\right)} \mathop{\mathbb{E}}_{s \sim b} \mathop{\mathbb{E}}_{s' \sim \mathbf{P}(\cdot \mid s,a)} \mathop{\mathbb{E}}_{o' \sim O(\cdot \mid s,a)} \delta_{\varphi_{\iota}\left(\bar{b},o,a\right)}\left(\bar{b}'\right) \cdot \delta_{\tau(b,o,a)}\left(b'\right), \text{and} \\
\mathcal{R}_{\overline{\pi}^{\diamond}}(b, \bar{b}) &= \mathop{\mathbb{E}}_{a \sim \overline{\pi}\left(\cdot \mid \bar{b}\right)} \mathop{\mathbb{E}}_{s \sim b} \mathcal{R}(s, a). \qquad \text{(cf. Definition A.3)}
\end{aligned}$$

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

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

(2) Each equivalence class $T \in \left(\mathcal{S}_{\mathcal{H}} \times \mathcal{B} \times \overline{\mathcal{B}}\right) / \rightrightarrows_{\varphi_{\iota}}^{\tau}$ consists of histories sharing the same belief and latent beliefs. Since $\tau^* \colon \mathcal{S}_{\mathcal{H}} \to \mathcal{B}$ and $\varphi_{\iota}^* \colon \mathcal{S}_{\mathcal{H}} \to \overline{\mathcal{B}}$ 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 $\rightrightarrows_{\varphi_{\iota}}^{\tau}$ has the form $T = \left[\left\langle b', \bar{b}' \right\rangle\right]_{\rightrightarrows_{\varphi_{\iota}}^{\tau}}$ so that
   (a) the projection of $\left[\left\langle b', \bar{b}' \right\rangle\right]_{\rightrightarrows_{\varphi_{\iota}}^{\tau}}$ on $\mathcal{S}_{\mathcal{H}}$ is the set $\{ h \in \mathcal{S}_{\mathcal{H}} \mid \tau^*(h) = b \text{ and } \varphi_{\iota}^*(h) = \bar{b} \}$, and
   (b) the projection of $\left[\left\langle b', \bar{b}' \right\rangle\right]_{\rightrightarrows_{\varphi_{\iota}}^{\tau}}$ on the state space of $\mathcal{M}_{\mathcal{B}}^{\overline{\pi}^{\diamond}}$ is merely the pair $\langle b', \bar{b}' \rangle$.

Therefore,

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

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

Corollary A.7. *The agent behaviors, formulated through the expected return, that are obtained by executing the policies respectively in the two models are the same:* $\mathbb{E}^{\mathcal{M}_{\mathcal{H}}}_{\overline{\pi}^{\clubsuit}}\left[\sum_{t=0}^{\infty}\gamma^t\cdot\mathcal{R}_{\mathcal{H}}(a_{0:t}, o_{1:t})\right] = \mathbb{E}^{\mathcal{M}_{\mathcal{B}}}_{\overline{\pi}^{\diamondsuit}}\left[\sum_{t=0}^{\infty}\gamma^t\cdot\mathcal{R}_{\mathcal{B}}(b_t, a_t)\right]$.

Proof. Follows directly from [16, 27]: 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.

A.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 A.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 [23]), 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 $\mathcal{M}^{\overline{\pi}^{\diamondsuit}}_{\mathcal{B}}$. □

B VALUE DIFFERENCE BOUND

Let us restate Theorem 3.3:

Theorem B.1. *Let $\mathcal{P}$, $\overline{\mathcal{P}}_\theta$, 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 A.1, so that the latent POMDP is learned through a WAE-MDP, via the minimization of the local losses $L_{\mathcal{R}}, L_{\mathbf{P}}$ of Eq. 5. 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 $\overline{\mathcal{P}}_\theta$ are guaranteed to be bounded by the local losses in average:*

$$
\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}}\left|\mathrm{V}_{\overline{\pi}}(h) - \overline{\mathrm{V}}_{\overline{\pi}}(h)\right| \leq \frac{L_{\mathcal{R}} + L^{\varphi}_{\overline{\mathcal{R}}} + \overline{\mathcal{R}}^{\star}L_{\bar{\tau}} + \gamma K_{\overline{\mathrm{V}}}\cdot\left(L_{\mathbf{P}} + L^{\varphi}_{\overline{\mathbf{P}}} + L_{\bar{\tau}} + L_O\right)}{1-\gamma}. \tag{11}
$$

Before going further, let us formally define the *value function* of any POMDP.

B.1 Value Functions

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

Definition B.2 (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., $\mathrm{V}_\pi(s) = \mathbb{E}^{\mathcal{M}[s]}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t\cdot\mathcal{R}(s_t, a_t)\right]$. 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 A.3). Then, the value function can be defined as the unique fixed point of the Bellman's equations [37]: $\mathrm{V}_\pi(s) = \mathcal{R}_\pi(s) + \mathbb{E}_{s'\sim\mathbf{P}_\pi(s)}\left[\gamma\cdot\mathrm{V}_\pi(s')\right]$. The typical goal of an RL agent is to learn a policy $\pi^\star$ that maximizes the value of the initial state of $\mathcal{M}$: $\max_{\pi^\star}\mathrm{V}_{\pi^\star}(s_I)$.

Property B.3 (POMDP values). *We obtain the value function of any POMDP $\mathcal{P} = \langle \mathcal{M}, \Omega, O \rangle$ 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 $\mathrm{V}_\pi(h)$ for $\mathrm{V}_\pi(\tau^*(h))$. Therefore, we have in particular for any latent policy $\overline{\pi} \colon \overline{\mathcal{B}} \to \Delta(\mathcal{A})$:*

$$
\begin{aligned}
\mathrm{V}_{\overline{\pi}}(h) &= \mathbb{E}_{\overline{\pi}^{\diamondsuit}}^{\mathcal{M}_{\mathcal{B}}[\tau^*(h)]} \left[\sum_{t=0}^{\infty} \gamma^t \cdot \mathcal{R}_{\mathcal{B}}(b_t, a_t) \right] && \text{(cf. Lemma A.6 for } \overline{\pi}^{\diamondsuit} \text{ and } \overline{\pi}^{\clubsuit}) \\
&= \mathbb{E}_{\overline{\pi}^{\clubsuit}}^{\mathcal{M}_{\mathcal{H}}[h]} \left[\sum_{t=0}^{\infty} \gamma^t \cdot \mathcal{R}_{\mathcal{H}}(h_t, a_t) \right] && \text{(cf. Corollary A.7)} \\
&= \mathop{\mathbb{E}}_{a \sim \overline{\pi}^{\clubsuit}(\cdot \mid h)} \left[\mathcal{R}_{\mathcal{H}}(h, a) + \mathop{\mathbb{E}}_{h' \sim \mathbf{P}_{\mathcal{H}}(\cdot \mid h, a)} \left[\gamma \cdot \mathrm{V}_{\overline{\pi}}(h') \right] \right] && \text{(by Definition B.2)} \\
&= \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi_t^*(h))} \left[\mathcal{R}_{\mathcal{H}}(h, a) + \mathop{\mathbb{E}}_{h' \sim \mathbf{P}_{\mathcal{H}}(\cdot \mid h, a)} \left[\gamma \cdot \mathrm{V}_{\overline{\pi}}(h') \right] \right] && \text{(by definition of } \overline{\pi}^{\clubsuit}) \\
&= \mathop{\mathbb{E}}_{a \sim \overline{\pi}(\cdot \mid \varphi_t^*(h))} \mathop{\mathbb{E}}_{s \sim \tau^*(h)} \left[\mathcal{R}(s, a) + \mathop{\mathbb{E}}_{s' \sim \mathbf{P}(\cdot \mid s, a)} \mathop{\mathbb{E}}_{o' \sim O(\cdot \mid s', a)} \left[\gamma \cdot \mathrm{V}_{\overline{\pi}}(h \cdot a \cdot o') \right] \right]. && \text{(by definition of } \mathcal{M}_{\mathcal{H}})
\end{aligned}
$$

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

B.2 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 claim. 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 $\mathbf{1}_{\neq} \colon \mathcal{X} \to \{1, 0\}$ [11], formally defined as $\mathbf{1}_{\neq}(x1, x2) = 1$ iff $x_1 \neq x_2$.

Definition B.4 (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 the metric $d \colon \mathcal{X} \to [0, \infty)$. Then, given a constant $K \geq 0$, we say that f is K-Lipschitz iff, for any $x_1, x_2 \in \mathcal{X}$, $|f(x_1) - f(x_2)| \leq K \cdot d(x_1, x_2)$. We write $\mathcal{F}_d^K$ for the set of K-Lipschitz functions.

Definition B.5 (Wasserstein dual). The Kantorovich-Rubinstein duality [40] allows formulating the Wasserstein distance between P and Q as $\mathcal{W}_d(P, Q) = \sup_{f \in \mathcal{F}_d^1} \left| \mathbb{E}_{x \sim P} f(x) - \mathbb{E}_{y \sim Q} f(y) \right|$.

Property B.6 (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 $P, Q \in \Delta(\mathcal{X})$, $\left| \mathbb{E}_{x_1 \sim P} f(x_1) - \mathbb{E}_{x_2 \sim Q} f(x_2) \right| \leq K \cdot \mathcal{W}_d(P, Q)$.*

*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 $\left| \mathbb{E}_{x_1 \sim P} g(x_1) - \mathbb{E}_{x_2 \sim Q} g(x_2) \right| \leq K_Y \cdot \mathcal{W}_{\mathbf{1}_{\neq}}(P, Q) = K_Y \cdot d_{TV}(P, Q)$, where $K_Y \geq \sup_{x \in \mathcal{X}} |g(x)|$ (see, e.g., [15, Sect. 6] for a discussion).*

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

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

Lemma B.7 (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 $\mathbb{E}_{y \sim \mathcal{D}} \left| \mathbb{E}_{x_1 \sim P} f(y, x_1) - \mathbb{E}_{x_2 \sim Q} f(y, x_2) \right| \leq K \cdot \mathcal{W}_d(P, Q)$.*

Proof. The proof is straightforward by construction of g_y:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{y \sim \mathcal{D}} \left| \mathop{\mathbb{E}}_{x_1 \sim P} f(y, x_1) - \mathop{\mathbb{E}}_{x_2 \sim Q} f(y, x_2) \right| \\
=&\mathop{\mathbb{E}}_{y \sim \mathcal{D}} \left| \mathop{\mathbb{E}}_{x_1 \sim P} g_y(x_1) - \mathop{\mathbb{E}}_{x_2 \sim Q} g_y(x_2) \right| \\
\leq&\mathop{\mathbb{E}}_{y \sim \mathcal{D}} \left[K \cdot \mathcal{W}_d(P, Q) \right] && \text{(by Property B.6, since } g_y \text{ is } K\text{-Lipschitz)} \\
=&K \cdot \mathcal{W}_d(P, Q)
\end{aligned}
$$

□

B.3 Value Difference Bounds: Time to Raise your Expectations

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 highlight 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 B.6 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:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}}\left|\mathrm{V}_{\overline{\pi}}(h)-\overline{\mathrm{V}}_{\overline{\pi}}(h)\right| \\
&= \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}}\Bigg|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_i^*(h))}\left[\mathcal{R}(s,a)+\gamma\mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)}\mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)}\mathrm{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\right] && \text{(see Property B.3)}\\
&\qquad -\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_i^*(h))}\left[\overline{\mathcal{R}}_\theta(\bar{s},a)+\gamma\mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)}\mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')}\overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\right]\Bigg| \\
&\leq \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_i^*(h))}\Bigg[\left|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathcal{R}(s,a)-\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\overline{\mathcal{R}}_\theta(\bar{s},a)\right| \\
&\qquad +\gamma\left|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)}\mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)}\mathrm{V}_{\pi}\big(h\cdot a\cdot o'\big)-\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)}\mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')}\overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot o'\big)\right|\Bigg] && \text{(Triangular inequality)}
\end{aligned}
$$

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}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*_\iota(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}}_\theta(\bar{s},a)\right|\\
=&\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*_\iota(h))}\left|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\left[\mathcal{R}(s,a)-\overline{\mathcal{R}}_\theta(\phi_\iota(s,o),a)\right]+\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\left[\overline{\mathcal{R}}_\theta(\phi_\iota(s,o),a)-\overline{\mathcal{R}}_\theta(\bar{s},a)\right]\right|\\
&\text{(o is the last observation of h; the state embedding function ϕ_ι that links the original and latent state spaces comes into play)}\\
\leq&\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*_\iota(h))}\left[\left|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\left[\mathcal{R}(s,a)-\overline{\mathcal{R}}_\theta(\phi_\iota(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}}_\theta(\phi_\iota(s,o),a)-\overline{\mathcal{R}}_\theta(\bar{s},a)\right]\right|\right] && \text{(Triangular inequality)}\\
\leq&\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*_\iota(h))}\left[\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\left|\mathcal{R}(s,a)-\overline{\mathcal{R}}_\theta(\phi_\iota(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}}_\theta(\phi_\iota(s,o),a)-\overline{\mathcal{R}}_\theta(\bar{s},a)\right]\right|\right] && \text{(Jensen's inequality)}\\
=&\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*_\iota(h))}\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\left|\mathcal{R}(s,a)-\overline{\mathcal{R}}_\theta(\phi_\iota(s,o),a)\right|+\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*_\iota(h))}\left|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\left[\overline{\mathcal{R}}_\theta(\phi_\iota(s,o),a)-\overline{\mathcal{R}}_\theta(\bar{s},a)\right]\right|\\
=&L_{\mathcal{R}}+\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*_\iota(h))}\left|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\left[\overline{\mathcal{R}}_\theta(\phi_\iota(s,o),a)-\overline{\mathcal{R}}_\theta(\bar{s},a)\right]\right| && \text{(by definition of $L_{\mathcal{R}}$, Eq. 5)}\\
=&L_{\mathcal{R}}+\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*_\iota(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^*_\iota(h)}\left[\left[\overline{\mathcal{R}}_\theta(\phi_\iota(s,o),a)-\overline{\mathcal{R}}_\theta(\bar{s}_\perp,a)\right]+\left[\overline{\mathcal{R}}_\theta(\bar{s}_\perp,a)-\overline{\mathcal{R}}_\theta(\bar{s},a)\right]\right]\right|\\
&\text{(the belief encoder φ_ι comes into play)}\\
=&L_{\mathcal{R}}+\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*_\iota(h))}\left|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar{s}\sim\varphi^*_\iota(h)}\left[\overline{\mathcal{R}}_\theta(\phi_\iota(s,o),a)-\overline{\mathcal{R}}_\theta(\bar{s},a)\right]+\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\mathop{\mathbb{E}}_{\bar{s}_\perp\sim\varphi^*_\iota(h)}\left[\overline{\mathcal{R}}_\theta(\bar{s}_\perp,a)-\overline{\mathcal{R}}_\theta(\bar{s},a)\right]\right|\\
\leq&L_{\mathcal{R}}+\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*_\iota(h))}\left[\left|\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar{s}\sim\varphi^*_\iota(h)}\left[\overline{\mathcal{R}}_\theta(\phi_\iota(s,o),a)-\overline{\mathcal{R}}_\theta(\bar{s},a)\right]\right|+\left|\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\mathop{\mathbb{E}}_{\bar{s}_\perp\sim\varphi^*_\iota(h)}\left[\overline{\mathcal{R}}_\theta(\bar{s}_\perp,a)-\overline{\mathcal{R}}_\theta(\bar{s},a)\right]\right|\right]\\
& && \text{(Triangular inequality)}\\
\leq&L_{\mathcal{R}}+\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*_\iota(h))}\left[\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar{s}\sim\varphi^*_\iota(h)}\left|\overline{\mathcal{R}}_\theta(\phi_\iota(s,o),a)-\overline{\mathcal{R}}_\theta(\bar{s},a)\right|+\left|\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\mathop{\mathbb{E}}_{\bar{s}_\perp\sim\varphi^*_\iota(h)}\left[\overline{\mathcal{R}}_\theta(\bar{s}_\perp,a)-\overline{\mathcal{R}}_\theta(\bar{s},a)\right]\right|\right] && \text{(Jensen's inequality)}\\
=&L_{\mathcal{R}}+\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*_\iota(h))}\mathop{\mathbb{E}}_{s\sim\tau^*(h)}\mathop{\mathbb{E}}_{\bar{s}\sim\varphi^*_\iota(h)}\left|\overline{\mathcal{R}}_\theta(\phi_\iota(s,o),a)-\overline{\mathcal{R}}_\theta(\bar{s},a)\right|+\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*_\iota(h))}\left|\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\mathop{\mathbb{E}}_{\bar{s}_\perp\sim\varphi^*_\iota(h)}\overline{\mathcal{R}}_\theta(\bar{s}_\perp,a)-\overline{\mathcal{R}}_\theta(\bar{s},a)\right|\\
=&L_{\mathcal{R}}+L^{\varphi}_{\overline{\mathcal{R}}}+\mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}}\mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi^*_\iota(h))}\left|\mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)}\mathop{\mathbb{E}}_{\bar{s}_\perp\sim\varphi^*_\iota(h)}\overline{\mathcal{R}}_\theta(\bar{s}_\perp,a)-\overline{\mathcal{R}}_\theta(\bar{s},a)\right| && \text{(by definition of $L^{\varphi}_{\overline{\mathcal{R}}}$, Eq. 8)}\\
\leq&L_{\mathcal{R}}+L^{\varphi}_{\overline{\mathcal{R}}}+\mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}}\overline{\mathcal{R}}^{\star}\mathcal{W}_{\bar{d}}\left(\bar{\tau}^*(h),\varphi^*_\iota(h)\right) && \text{(as $\lambda\to 0$, by Lem. B.7 and Prop. B.6)}\\
=&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 $\left\|\overline{\mathcal{R}}_\theta\right\|_\infty=\sup_{\bar{s},a\in\overline{\mathcal{S}}\times\mathcal{A}}\left|\overline{\mathcal{R}}_\theta(\bar{s},a)\right|$.

Part 2: Next value bounds

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

$$= \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)} \left[\mathrm{V}_{\overline{\pi}}(h\cdot a\cdot o') - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') \right] \right.$$
$$\left. + \left[\mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)} \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot \hat{o}') - \mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') \right] \right|$$

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

$$\leq \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)} \left[\mathrm{V}_{\overline{\pi}}(h\cdot a\cdot o') - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot \hat{o}') \right] \right|$$
$$+ \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)} \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot \hat{o}') - \mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') \right|$$

(Triangular inequality)

$$= \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)} \left[\mathrm{V}_{\overline{\pi}}(h\cdot a\cdot o') - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot \hat{o}') \right] \right|$$
$$+ \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s',o'\sim\mathbf{P}_\Omega(\cdot|s,o,a)} \left[\mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot \hat{o}') \right] - \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\phi_\iota(s,o),a)} \left[\mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot \hat{o}') \right] \right.$$
$$\left. + \left[\mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\phi_\iota(s,o),a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') - \mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') \right] \right|$$

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

$$\leq \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)} \left[\mathrm{V}_{\overline{\pi}}(h\cdot a\cdot o') - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot \hat{o}') \right] \right|$$
$$+ \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s',o'\sim\mathbf{P}_\Omega(\cdot|s,o,a)} \left[\mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot \hat{o}') \right] - \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\phi_\iota(s,o),a)} \left[\mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot \hat{o}') \right] \right|$$
$$+ \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\phi_\iota(s,o),a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') - \mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') \right|$$

(Triangular inequality)

$$\leq \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)} \left[\mathrm{V}_{\overline{\pi}}(h\cdot a\cdot o') - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot \hat{o}') \right] \right|$$
$$+ \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \left| \mathop{\mathbb{E}}_{\bar{s}'\sim\phi_\iota\mathbf{P}_\Omega(\cdot|s,o,a)} \left[\mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') \right] - \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\phi_\iota(s,o),a)} \left[\mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') \right] \right|$$
$$+ \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\phi_\iota(s,o),a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') - \mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') \right|$$

(Jensen's inequality)

$$\leq \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)} \left[\mathrm{V}_{\overline{\pi}}(h\cdot a\cdot o') - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot \hat{o}') \right] \right|$$
$$+ \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \mathop{\mathbb{E}}_{s\sim\tau^*(h)} K_{\overline{\mathrm{V}}} \cdot \mathcal{W}_{\bar{d}} \left(\phi_\iota\mathbf{P}_\Omega(\cdot \mid s,a), \bar{\mathbf{P}}_\theta(\cdot \mid \phi_\iota(s,o),a) \right)$$
$$+ \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\phi_\iota(s,o),a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') - \mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') \right|$$

(as $\lambda \to 0$, by Lem. B.7)

$$
\begin{aligned}
\leq\ & \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)} \left[\mathrm{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot \hat{o}'\big) \right] \right| \\
& + \gamma K_{\overline{\mathrm{V}}} L_{\mathbf{P}} \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\phi_\iota(s,o),a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) \right|
\end{aligned}
$$

(by definition of $L_{\mathbf{P}}$, Eq. 5)

$$
\begin{aligned}
=\ & \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)} \left[\mathrm{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot \hat{o}'\big) \right] \right| \\
& + \gamma K_{\overline{\mathrm{V}}} L_{\mathbf{P}} \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \left[\mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\phi_\iota(s,o),a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \mathop{\mathbb{E}}_{\bar{s}\sim\varphi_\iota^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) \right] \right. \\
& \qquad \left. + \left[\mathop{\mathbb{E}}_{\bar{s}\sim\varphi_\iota^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) \right] \right|
\end{aligned}
$$

(the belief encoder φ_ι comes into play)

$$
\begin{aligned}
\leq\ & \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)} \left[\mathrm{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot \hat{o}'\big) \right] \right| \\
& + \gamma K_{\overline{\mathrm{V}}} L_{\mathbf{P}} \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\phi_\iota(s,o),a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \mathop{\mathbb{E}}_{\bar{s}\sim\varphi_\iota^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) \right| \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{\bar{s}\sim\varphi_\iota^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) \right|
\end{aligned}
$$

(triangular inequality)

$$
\begin{aligned}
\leq\ & \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)} \left[\mathrm{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot \hat{o}'\big) \right] \right| \\
& + \gamma K_{\overline{\mathrm{V}}} L_{\mathbf{P}} \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{\bar{s}\sim\varphi_\iota^*(h)} \left| \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\phi_\iota(s,o),a)} \left[\mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) \right] - \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \left[\mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) \right] \right| \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{\bar{s}\sim\varphi_\iota^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) \right|
\end{aligned}
$$

(Jensen's inequality)

$$
\begin{aligned}
\leq\ & \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)} \left[\mathrm{V}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot \hat{o}'\big) \right] \right| \\
& + \gamma K_{\overline{\mathrm{V}}} L_{\mathbf{P}} \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{\bar{s}\sim\varphi_\iota^*(h)} K_{\overline{\mathrm{V}}} \mathcal{W}_{\bar{d}} \left(\bar{\mathbf{P}}_\theta(\cdot \mid \phi_\iota(s,o),a), \bar{\mathbf{P}}_\theta(\cdot \mid \bar{s},a) \right) \\
& + \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{\bar{s}\sim\varphi_\iota^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) - \mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}\big(h\cdot a\cdot o'\big) \right|
\end{aligned}
$$

(as $\lambda \to 0$, by Lem. B.7)

$$
\begin{aligned}
&= \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)} \left[\mathrm{V}_{\overline{\pi}}(h\cdot a\cdot o') - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot \hat{o}') \right] \right| \\
&\quad + \gamma K_{\overline{\mathrm{V}}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\widetilde{\mathbf{P}}} \right) \\
&\quad + \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{\bar{s}\sim\varphi_\iota^*(h)} \left[\mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') \right] - \mathop{\mathbb{E}}_{\bar{s}\sim\bar{\tau}^*(h)} \left[\mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') \right] \right| \\
&\qquad\qquad \text{(by definition of } L^{\varphi}_{\widetilde{\mathbf{P}}}\text{, Eq. 8)} \\
&\leq \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)} \left[\mathrm{V}_{\overline{\pi}}(h\cdot a\cdot o') - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot \hat{o}') \right] \right| \\
&\quad + \gamma K_{\overline{\mathrm{V}}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\widetilde{\mathbf{P}}} \right) \\
&\quad + \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} K_{\overline{\mathrm{V}}} \mathcal{W}_{\bar{d}} \left(\bar{\tau}^*(h), \varphi_\iota^*(h) \right) \\
&\qquad\qquad \text{(as } \lambda \to 0\text{, by Lem. B.7; note that Wasserstein is symmetric since it is a distance metric [40])} \\
&\leq \gamma \cdot \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)} \left[\mathrm{V}_{\overline{\pi}}(h\cdot a\cdot o') - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot \hat{o}') \right] \right| \\
&\quad + \gamma K_{\overline{\mathrm{V}}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\widetilde{\mathbf{P}}} + L_{\bar{\tau}} \right) \qquad \text{(by definition of } L_{\bar{\tau}}\text{, Eq. 7)} \\
&= \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s',o'\sim\mathbf{P}_\Omega(\cdot|s,o,a)} \left[\left(\mathrm{V}_{\overline{\pi}}(h\cdot a\cdot o') - \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') \right) + \left(\overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot \hat{o}') \right) \right] \right| \\
&\quad + \gamma K_{\overline{\mathrm{V}}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\widetilde{\mathbf{P}}} + L_{\bar{\tau}} \right) \\
&\leq \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s',o'\sim\mathbf{P}_\Omega(\cdot|s,o,a)} \left[\mathrm{V}_{\overline{\pi}}(h\cdot a\cdot o') - \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') \right] \right| \\
&\quad + \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s',o'\sim\mathbf{P}_\Omega(\cdot|s,o,a)} \left[\overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot \hat{o}') \right] \right| \\
&\quad + \gamma K_{\overline{\mathrm{V}}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\widetilde{\mathbf{P}}} + L_{\bar{\tau}} \right) \qquad \text{(triangular inequality)} \\
&\leq \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s',o'\sim\mathbf{P}_\Omega(\cdot|s,o,a)} \left[\mathrm{V}_{\overline{\pi}}(h\cdot a\cdot o') - \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') \right] \right| \\
&\quad + \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} \left| \mathop{\mathbb{E}}_{o'\sim O(\cdot|s',a)} \left[\overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') - \mathop{\mathbb{E}}_{\hat{o}'\sim\overline{O}_\theta(\cdot|\phi_\iota(s',o'))} \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot \hat{o}') \right] \right| \\
&\quad + \gamma K_{\overline{\mathrm{V}}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\widetilde{\mathbf{P}}} + L_{\bar{\tau}} \right) \qquad \text{(Jensen's inequality)} \\
&\leq \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s',o'\sim\mathbf{P}_\Omega(\cdot|s,o,a)} \left[\mathrm{V}_{\overline{\pi}}(h\cdot a\cdot o') - \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') \right] \right| \\
&\quad + \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s'\sim\mathbf{P}(\cdot|s,a)} K_{\overline{\mathrm{V}}} d_{TV} \left(O(\cdot \mid s',a), \mathop{\mathbb{E}}_{o'\sim s',a} \overline{O}_\theta(\cdot \mid \phi_\iota(s',o')) \right) \\
&\quad + \gamma K_{\overline{\mathrm{V}}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\widetilde{\mathbf{P}}} + L_{\bar{\tau}} \right) \qquad \text{(cf. Prop. B.6 and Lem B.7)} \\
&= \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \left| \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s',o'\sim\mathbf{P}_\Omega(\cdot|s,o,a)} \left[\mathrm{V}_{\overline{\pi}}(h\cdot a\cdot o') - \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') \right] \right| \\
&\quad + \gamma K_{\overline{\mathrm{V}}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\widetilde{\mathbf{P}}} + L_{\bar{\tau}} + L_O \right) \qquad \text{(by definition of } L_O\text{, Eq. 6)} \\
&\leq \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \mathop{\mathbb{E}}_{a\sim\overline{\pi}(\cdot|\varphi_\iota^*(h))} \mathop{\mathbb{E}}_{s\sim\tau^*(h)} \mathop{\mathbb{E}}_{s',o'\sim\mathbf{P}_\Omega(\cdot|s,o,a)} \left| \mathrm{V}_{\overline{\pi}}(h\cdot a\cdot o') - \overline{\mathrm{V}}_{\overline{\pi}}(h\cdot a\cdot o') \right| + \gamma K_{\overline{\mathrm{V}}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\widetilde{\mathbf{P}}} + L_{\bar{\tau}} + L_O \right) \qquad \text{(Jensen's inequality)} \\
&= \gamma \cdot \mathop{\mathbb{E}}_{h,o\sim\mathcal{H}_{\overline{\pi}}} \left| \mathrm{V}_{\overline{\pi}}(h) - \overline{\mathrm{V}}_{\overline{\pi}}(h) \right| + \gamma K_{\overline{\mathrm{V}}} \cdot \left(L_{\mathbf{P}} + L^{\varphi}_{\widetilde{\mathbf{P}}} + L_{\bar{\tau}} + L_O \right) \\
&\qquad\qquad \text{(}\mathcal{H}_{\overline{\pi}} \text{ is a stationary distribution (Lem. A.1) which allows us to apply the stationary property (Def. A.4))}
\end{aligned}
$$

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

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

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

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

which finally concludes the proof. □

B.4 Representation Quality Bound

We start by showing that the optimal latent value function is Lipschitz continuous in the latent belief space. Coupled with Theorem B.1, 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.

Lemma B.8. *The optimal latent value function, given by* $\max_{\overline{\pi}^*}\overline{\mathrm{V}}_{\overline{\pi}^*}\left(\bar{b}\right)=\overline{\mathrm{V}}^{*}\left(\bar{b}\right)$ *for any belief* $\bar{b}\in\overline{\mathcal{B}}$*, is* $\overline{\mathcal{R}}^{\star}/_{1-\gamma}$*-Lipschitz as the temperature parameter of the WAE-MDP* λ *goes to* 0.

Proof. To prove this claim, we consider the dynamic programming value sequence $\mathrm{V}_n(s)=\max_{a\in\mathcal{A}}Q_n(s,a)$ so that $Q_{n+1}(s,a)=\mathcal{R}(s,a)+\mathbb{E}_{s'\sim\mathbf{P}(\cdot|s,a)}\left[\gamma\max_{a'\in\mathcal{A}}Q_n(s',a')\right]$, proven to converge to $\mathrm{V}^*(s)$ as $n\to\infty$ (e.g., [37]), for any initial q-value $Q_0^*(s,a)\in\mathbb{R}$, and MDP $\mathcal{M}$ with state and action spaces $\mathcal{S}$ and $\mathcal{A}$. We show by recursion that all sequences of such latent values for the latent belief MDP $\overline{\mathcal{M}}_{\overline{\mathcal{B}}}$ are all K_n-Lipchitz, for any $n\geq 0$ and some $K_n\geq 0$. We further prove that $\lim_{n\to\infty}K_n=\frac{\overline{\mathcal{R}}^{\star}}{1-\gamma}$. Set $\overline{Q}_0\left(\bar{b},a\right)=0$ for all $\bar{b}\in\overline{\mathcal{B}}$, clearly $K_0=0$. Assume now that $K_n=\sup_{a\in\mathcal{A},\bar{b}_1\neq\bar{b}_2}\frac{\left|\overline{Q}_n\left(\bar{b}_1,a\right)-\overline{Q}_n\left(\bar{b}_2,a\right)\right|}{\mathcal{W}_{\bar{d}}\left(\bar{b}_1,\bar{b}_2\right)}$. Then, we have

$$\begin{aligned}
K_{n+1} &= \sup_{a\in\mathcal{A},\bar{b}_1\neq\bar{b}_2}\frac{\left|\overline{Q}_{n+1}\left(\bar{b}_1,a\right)-\overline{Q}_{n+1}\left(\bar{b}_2,a\right)\right|}{\mathcal{W}_{\bar{d}}\left(\bar{b}_1,\bar{b}_2\right)}\\
&\leq \sup_{a\in\mathcal{A},\bar{b}_1\neq\bar{b}_2}\frac{\left|\overline{\mathcal{R}}_{\overline{\mathcal{B}}}\left(\bar{b}_1,a\right)-\overline{\mathcal{R}}_{\overline{\mathcal{B}}}\left(\bar{b}_2,a\right)\right|}{\mathcal{W}_{\bar{d}}\left(\bar{b}_1,\bar{b}_2\right)}+\gamma\cdot\sup_{a\in\mathcal{A},\bar{b}_1\neq\bar{b}_2}\frac{\left|\mathbb{E}_{\bar{b}'\sim\bar{\mathbf{P}}_{\overline{\mathcal{B}}}\left(\cdot|\bar{b}_1,a\right)}\overline{Q}_n\left(\bar{b}',a\right)-\mathbb{E}_{\bar{b}'\sim\bar{\mathbf{P}}_{\overline{\mathcal{B}}}\left(\cdot|\bar{b}_2,a\right)}\overline{Q}_n\left(\bar{b}',a\right)\right|}{\mathcal{W}_{\bar{d}}\left(\bar{b}_1,\bar{b}_2\right)} \quad \text{(triangular inequality)}
\end{aligned}$$

Now, observe that $\left|\overline{\mathcal{R}}_{\overline{\mathcal{B}}}\left(\bar{b}_1,a\right)-\overline{\mathcal{R}}_{\overline{\mathcal{B}}}\left(\bar{b}_2,a\right)\right|=\left|\mathbb{E}_{\bar{s}\sim\bar{b}_1}\overline{\mathcal{R}}_{\theta}(\bar{s},a)-\mathbb{E}_{\bar{s}\sim\bar{b}_2}\overline{\mathcal{R}}_{\theta}(\bar{s},a)\right|\leq\overline{\mathcal{R}}^{\star}\mathcal{W}_{\bar{d}}\left(\bar{b}_1,\bar{b}_2\right)$ as $\lambda\to 0$ (see Property B.6). Therefore,

$$\begin{aligned}
&\sup_{a\in\mathcal{A},\bar{b}_1\neq\bar{b}_2}\frac{\left|\overline{\mathcal{R}}_{\overline{\mathcal{B}}}\left(\bar{b}_1,a\right)-\overline{\mathcal{R}}_{\overline{\mathcal{B}}}\left(\bar{b}_2,a\right)\right|}{\mathcal{W}_{\bar{d}}\left(\bar{b}_1,\bar{b}_2\right)}+\gamma\cdot\sup_{a\in\mathcal{A},\bar{b}_1\neq\bar{b}_2}\frac{\left|\mathbb{E}_{\bar{b}'\sim\bar{\mathbf{P}}_{\overline{\mathcal{B}}}\left(\cdot|\bar{b}_1,a\right)}\overline{Q}_n\left(\bar{b}',a\right)-\mathbb{E}_{\bar{b}'\sim\bar{\mathbf{P}}_{\overline{\mathcal{B}}}\left(\cdot|\bar{b}_2,a\right)}\overline{Q}_n\left(\bar{b}',a\right)\right|}{\mathcal{W}_{\bar{d}}\left(\bar{b}_1,\bar{b}_2\right)}\\
\leq&\overline{\mathcal{R}}^{\star}+\gamma\cdot\sup_{a\in\mathcal{A},\bar{b}_1\neq\bar{b}_2}\frac{\left|\mathbb{E}_{\bar{b}'\sim\bar{\mathbf{P}}_{\overline{\mathcal{B}}}\left(\cdot|\bar{b}_1,a\right)}\overline{Q}_n\left(\bar{b}',a\right)-\mathbb{E}_{\bar{b}'\sim\bar{\mathbf{P}}_{\overline{\mathcal{B}}}\left(\cdot|\bar{b}_2,a\right)}\overline{Q}_n\left(\bar{b}',a\right)\right|}{\mathcal{W}_{\bar{d}}\left(\bar{b}_1,\bar{b}_2\right)}\\
\leq&\overline{\mathcal{R}}^{\star}+\gamma K_n\cdot\sup_{a\in\mathcal{A},\bar{b}_1\neq\bar{b}_2}\frac{\mathcal{W}_{\bar{d}}\left(\bar{\mathbf{P}}_{\overline{\mathcal{B}}}\left(\cdot\mid\bar{b}_1,a\right),\bar{\mathbf{P}}_{\overline{\mathcal{B}}}\left(\cdot\mid\bar{b}_2,a\right)\right)}{\mathcal{W}_{\bar{d}}\left(\bar{b}_1,\bar{b}_2\right)} \quad \text{(by induction, } \overline{Q}_n \text{ is } K_n\text{-Lipschitz and by Prop. B.6)}
\end{aligned}$$

By definition,

$$\mathcal{W}_{\bar{d}}\left(\bar{\mathbf{P}}_{\overline{\mathcal{B}}}\left(\cdot\mid\bar{b}_1,a\right),\bar{\mathbf{P}}_{\overline{\mathcal{B}}}\left(\cdot\mid\bar{b}_2,a\right)\right)=\sup_{f\in\mathcal{F}^1_{\bar{d}}}\left|\mathop{\mathbb{E}}_{\bar{s}\sim\bar{b}_1}\mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_{\theta}(\cdot|\bar{s},a)}\mathop{\mathbb{E}}_{o'\sim\overline{O}_{\theta}(\cdot|\bar{s}')}\mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\tau}\left(\bar{b}_1,a,o'\right)}f\left(\bar{s}'\right)-\mathop{\mathbb{E}}_{\bar{s}\sim\bar{b}_2}\mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_{\theta}(\cdot|\bar{s},a)}\mathop{\mathbb{E}}_{o'\sim\overline{O}_{\theta}(\cdot|\bar{s}')}\mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\tau}\left(\bar{b}_2,a,o'\right)}f\left(\bar{s}'\right)\right| \quad (12)$$

Let $f^\star$ be the 1-Lipschitz function that meets the supremum of Eq. 12 and define the next 1-Lipschitz belief operator $\mathcal{T} : \overline{\mathcal{B}} \to \mathcal{F}^1_{\bar{d}}$ as

$$\mathcal{T}\left(\bar{b}\right) = \mathop{\mathbb{E}}_{\bar{s}\sim\bar{b}} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\mathbf{P}}_\theta(\cdot|\bar{s},a)} \mathop{\mathbb{E}}_{o'\sim\overline{O}_\theta(\cdot|\bar{s}')} \mathop{\mathbb{E}}_{\bar{s}'\sim\bar{\tau}\left(\bar{b},a,o'\right)} f^\star(\bar{s}').$$

Therefore, the Wasserstein distance between the two latent belief transition functions can be rewritten as:

$$\begin{aligned}
&\mathcal{W}_{\bar{d}}\left(\bar{\mathbf{P}}_{\overline{\mathcal{B}}}\left(\cdot \mid \bar{b}_1, a\right), \bar{\mathbf{P}}_{\overline{\mathcal{B}}}\left(\cdot \mid \bar{b}_2, a\right)\right) \\
=&\left| \mathop{\mathbb{E}}_{\bar{b}\sim\delta_{\bar{b}_1}} \mathcal{T}\left(\bar{b}\right) - \mathop{\mathbb{E}}_{\bar{b}\sim\delta_{\bar{b}_2}} \mathcal{T}\left(\bar{b}\right) \right| \\
\leq&\mathcal{W}_{\mathcal{W}_{\bar{d}}}\left(\delta_{\bar{b}_1}, \delta_{\bar{b}_2}\right) && \text{(since } \mathcal{T} \text{ is 1-Lipschitz)} \\
=&\mathcal{W}_{\bar{d}}\left(\bar{b}_1, \bar{b}_2\right).
\end{aligned}$$

Back to our induction proof, we have

$$\begin{aligned}
K_{n+1} \leq&\overline{\mathcal{R}}^\star + \gamma K_n \cdot \sup_{a\in\mathcal{A},\bar{b}_1\neq\bar{b}_2} \frac{\mathcal{W}_{\bar{d}}\left(\bar{\mathbf{P}}_{\overline{\mathcal{B}}}\left(\cdot \mid \bar{b}_1, a\right), \bar{\mathbf{P}}_{\overline{\mathcal{B}}}\left(\cdot \mid \bar{b}_2, a\right)\right)}{\mathcal{W}_{\bar{d}}\left(\bar{b}_1, \bar{b}_2\right)} \\
\leq&\overline{\mathcal{R}}^\star + \gamma K_n \cdot \sup_{a\in\mathcal{A},\bar{b}_1\neq\bar{b}_2} \frac{\mathcal{W}_{\bar{d}}\left(\bar{b}_1, \bar{b}_2\right)}{\mathcal{W}_{\bar{d}}\left(\bar{b}_1, \bar{b}_2\right)} \\
\leq&\overline{\mathcal{R}}^\star + \gamma K_n \\
\leq&\sum_{i=0}^{n-1} (\gamma)^i \overline{\mathcal{R}}^\star + \gamma^n K_0 \\
=&\sum_{i=0}^{n-1} \gamma^i \overline{\mathcal{R}}^\star
\end{aligned}$$

We are thus left with a geometric serie that converges to $\frac{\overline{\mathcal{R}}^\star}{1-\gamma} = K_{\bar{\mathrm{V}}}$ as n goes to ∞. To conclude, we thus have $\left|\bar{\mathrm{V}}^*\left(\bar{b}_1\right) - \bar{\mathrm{V}}^*\left(\bar{b}_2\right)\right| \leq K_{\bar{\mathrm{V}}}\mathcal{W}_{\bar{d}}\left(\bar{b}_1, \bar{b}_2\right)$ for any $\bar{b}_1, \bar{b}_2 \in \overline{\mathcal{B}}$. □

Theorem B.9. *Let $\overline{\pi}^*$ be the optimal policy of the POMDP $\overline{\mathcal{P}}_\theta$, then for any couple of histories $h_1, h_2 \in (\mathcal{A}\cdot\Omega)^*$ mapped to latent beliefs through $\varphi^*_\iota(h_1) = \bar{b}_1$ and $\varphi^*_\iota(h_2) = \bar{b}_2$, the belief representation induced by φ_ι yields:*

$$\left|\mathrm{V}_{\overline{\pi}^*}(h_1) - \mathrm{V}_{\overline{\pi}^*}(h_2)\right| \leq K_{\bar{\mathrm{V}}}\mathcal{W}_{\bar{d}}\left(\bar{b}_1, \bar{b}_2\right) + \frac{L_{\mathcal{R}} + L^\varphi_{\overline{\mathcal{R}}} + \left(K_{\bar{\mathrm{V}}} + \overline{\mathcal{R}}^\star\right)L_{\bar{\tau}} + \gamma K_{\bar{\mathrm{V}}}\cdot\left(L_{\mathbf{P}} + L^\varphi_{\bar{\mathbf{P}}} + L_O\right)}{1-\gamma}\left(\mathcal{H}^{-1}_{\overline{\pi}^*}(h_1) + \mathcal{H}^{-1}_{\overline{\pi}^*}(h_2)\right)$$

when the WAE-MDP temperature λ goes to 0.

Proof. First, observe that for any history $h \in (\mathcal{A}\cdot\Omega)^*$, $\left|\mathrm{V}_{\overline{\pi}^*}(h) - \bar{\mathrm{V}}_{\overline{\pi}^*}(h)\right| \leq \mathcal{H}^{-1}_{\overline{\pi}}(h)\cdot\mathbb{E}_{h'\sim\mathcal{H}_{\overline{\pi}}}\left|\mathrm{V}_{\overline{\pi}^*}(h') - \bar{\mathrm{V}}_{\overline{\pi}^*}(h')\right|$ (cf. [15]). Therefore, we have:

$$\begin{aligned}
&\left|\mathrm{V}_{\overline{\pi}^*}(h_1) - \mathrm{V}_{\overline{\pi}^*}(h_2)\right| \\
=&\left|\mathrm{V}_{\overline{\pi}^*}(h_1) - \bar{\mathrm{V}}_{\overline{\pi}^*}(h_1) + \bar{\mathrm{V}}_{\overline{\pi}^*}(h_1) - \bar{\mathrm{V}}_{\overline{\pi}^*}(h_2) + \bar{\mathrm{V}}_{\overline{\pi}^*}(h_2) - \mathrm{V}_{\overline{\pi}^*}(h_2)\right| \\
\leq&\left|\mathrm{V}_{\overline{\pi}^*}(h_1) - \bar{\mathrm{V}}_{\overline{\pi}^*}(h_1)\right| + \left|\bar{\mathrm{V}}_{\overline{\pi}^*}(h_1) - \bar{\mathrm{V}}_{\overline{\pi}^*}(h_2)\right| + \left|\bar{\mathrm{V}}_{\overline{\pi}^*}(h_2) - \mathrm{V}_{\overline{\pi}^*}(h_2)\right| && \text{(triangular inequality)} \\
\leq&\mathcal{H}^{-1}_{\overline{\pi}}(h_1) \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}}\left|\mathrm{V}_{\overline{\pi}^*}(h) - \bar{\mathrm{V}}_{\overline{\pi}^*}(h)\right| + \left|\bar{\mathrm{V}}_{\overline{\pi}^*}(h_1) - \bar{\mathrm{V}}_{\overline{\pi}^*}(h_2)\right| + \mathcal{H}^{-1}_{\overline{\pi}}(h_2) \mathop{\mathbb{E}}_{h\sim\mathcal{H}_{\overline{\pi}}}\left|\mathrm{V}_{\overline{\pi}^*}(h) - \bar{\mathrm{V}}_{\overline{\pi}^*}(h)\right| \\
\leq&\left|\bar{\mathrm{V}}_{\overline{\pi}^*}(h_1) - \bar{\mathrm{V}}_{\overline{\pi}^*}(h_2)\right| + \frac{L_{\mathcal{R}} + L^\varphi_{\overline{\mathcal{R}}} + \overline{\mathcal{R}}^\star L_{\bar{\tau}} + \gamma K_{\bar{\mathrm{V}}}\cdot\left(L_{\mathbf{P}} + L^\varphi_{\bar{\mathbf{P}}} + L_{\bar{\tau}} + L_O\right)}{1-\gamma}\left(\mathcal{H}^{-1}_{\overline{\pi}^*}(h_1) + \mathcal{H}^{-1}_{\overline{\pi}^*}(h_2)\right) && \text{(Thm 3.3)}
\end{aligned}$$

Define the Bellman operator $\mathbb{B} \colon (\mathcal{A} \cdot \Omega)^* \times \overline{\mathcal{S}}, \langle h, \bar{s} \rangle \mapsto \mathbb{E}_{a \sim \bar{\pi}^*(\cdot \mid \varphi_\iota^*(h))} \left[\overline{\mathcal{R}}(\bar{s}, a) + \gamma \cdot \mathbb{E}_{\bar{s}' \sim \bar{\mathbf{P}}(\cdot \mid \bar{s}, a)} \mathbb{E}_{o' \sim \overline{\mathcal{O}}(\cdot \mid \bar{s}')} \overline{\mathrm{V}}_{\bar{\pi}^*}(h \cdot a \cdot o') \right]$, then

$$
\begin{aligned}
&\left| \overline{\mathrm{V}}_{\bar{\pi}^*}(h_1) - \overline{\mathrm{V}}_{\bar{\pi}^*}(h_2) \right| \\
=& \left| \mathop{\mathbb{E}}_{\bar{s} \sim \bar{\tau}^*(h_1)} \mathbb{B}(h_1, \bar{s}) - \mathop{\mathbb{E}}_{\bar{s} \sim \bar{\tau}^*(h_2)} \mathbb{B}(h_2, \bar{s}) \right| \\
\leq& \left| \mathop{\mathbb{E}}_{\bar{s} \sim \bar{\tau}^*(h_1)} \mathbb{B}(h_1, \bar{s}) - \mathop{\mathbb{E}}_{\bar{s} \sim \varphi_\iota^*(h_1)} \mathbb{B}(h_1, \bar{s}) \right| + \left| \mathop{\mathbb{E}}_{\bar{s} \sim \varphi_\iota^*(h_1)} \mathbb{B}(h_1, \bar{s}) - \mathop{\mathbb{E}}_{\bar{s} \sim \varphi_\iota^*(h_2)} \mathbb{B}(h_2, \bar{s}) \right| + \left| \mathop{\mathbb{E}}_{\bar{s} \sim \bar{\tau}^*(h_2)} \mathbb{B}(h_2, \bar{s}) - \mathop{\mathbb{E}}_{\bar{s} \sim \varphi_\iota^*(h_2)} \mathbb{B}(h_2, \bar{s}) \right| \\
& \hspace{20em} \text{(triangular inequality)} \\
\leq& \left| \mathop{\mathbb{E}}_{\bar{s} \sim \bar{\tau}^*(h_1)} \mathbb{B}(h_1, \bar{s}) - \mathop{\mathbb{E}}_{\bar{s} \sim \varphi_\iota^*(h_1)} \mathbb{B}(h_1, \bar{s}) \right| + \left| \overline{\mathrm{V}}^*\left(\bar{b}_1\right) - \overline{\mathrm{V}}^*\left(\bar{b}_2\right) \right| + \left| \mathop{\mathbb{E}}_{\bar{s} \sim \bar{\tau}^*(h_2)} \mathbb{B}(h_2, \bar{s}) - \mathop{\mathbb{E}}_{\bar{s} \sim \varphi_\iota^*(h_2)} \mathbb{B}(h_2, \bar{s}) \right| \\
\leq& K_{\overline{\mathrm{V}}} \left(\mathcal{W}_{\bar{d}} \left(\bar{\tau}^*(h_1), \varphi_\iota^*(h_1) \right) + \mathcal{W}_{\bar{d}} \left(\bar{\tau}^*(h_2), \varphi_\iota^*(h_2) \right) \right) + \left| \overline{\mathrm{V}}^*\left(\bar{b}_1\right) - \overline{\mathrm{V}}^*\left(\bar{b}_2\right) \right| && \text{(as } \lambda \to 0\text{, by Lem. B.7)} \\
\leq& K_{\overline{\mathrm{V}}} \left(\mathcal{H}_{\bar{\pi}^*}^{-1}(h_1) + \mathcal{H}_{\bar{\pi}^*}^{-1}(h_2) \right) \mathop{\mathbb{E}}_{h \sim \mathcal{H}_{\bar{\pi}^*}} \mathcal{W}_{\bar{d}} \left(\bar{\tau}^*(h), \varphi_\iota^*(h) \right) + \left| \overline{\mathrm{V}}^*\left(\bar{b}_1\right) - \overline{\mathrm{V}}^*\left(\bar{b}_2\right) \right| \\
\leq& K_{\overline{\mathrm{V}}} \left(\mathcal{H}_{\bar{\pi}^*}^{-1}(h_1) + \mathcal{H}_{\bar{\pi}^*}^{-1}(h_2) \right) L_{\bar{\tau}} + \left| \overline{\mathrm{V}}^*\left(\bar{b}_1\right) - \overline{\mathrm{V}}^*\left(\bar{b}_2\right) \right| && \text{(by definition of } L_{\bar{\tau}}\text{, Eq. 7)} \\
\leq& K_{\overline{\mathrm{V}}} \left(\mathcal{H}_{\bar{\pi}^*}^{-1}(h_1) + \mathcal{H}_{\bar{\pi}^*}^{-1}(h_2) \right) L_{\bar{\tau}} + K_{\overline{\mathrm{V}}} \mathcal{W}_{\bar{d}} \left(\bar{b}_1, \bar{b}_2 \right) && \text{(Lem. B.8)}
\end{aligned}
$$

Putting all together, we have:

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

□