A Wasserstein Distances

Definition 5 (Wasserstein metric [50]). *Let $d : X \times X \to [0, \infty)$ be a distance function and Ω the set of all joint distributions with marginals μ and λ over the space X;*

$$W_p(d)(\mu, \lambda) = \left(\inf_{\omega \in \Omega} \mathbb{E}_{(x_1, x_2) \sim \omega}[d(x_1, x_2)^p] \right)^{\frac{1}{p}}. \tag{18}$$

Definition 6 (Dual formulation of the Wasserstein metric [50]). *Let $d : X \times X \to [0, \infty)$ be a distance function, and μ and λ marginals over the space X;*

$$W_p(d)(\mu, \lambda) = \left(\sup_{\phi \oplus \psi \leq d^p} \mathbb{E}_{x_1 \sim \mu}[\phi(x_1)] + \mathbb{E}_{x_2 \sim \lambda}[\psi(x_2)] \right)^{\frac{1}{p}}, \tag{19}$$

where $\phi \oplus \psi \leq d^p \iff \phi(x) + \psi(y) \leq d(x, y)^p, \ \forall (x, y) \in X \times X$.

This dual formulation takes a simple form for $p = 1$:

$$W_1(d)(\mu, \lambda) = \sup_{f \in \mathrm{Lip}_{1,d}(X)} \mathbb{E}_{x_1 \sim \mu}[f(x_1)] - \mathbb{E}_{x_2 \sim \lambda}[f(x_2)], \tag{20}$$

where $\mathrm{Lip}_{1,d}(X)$ denotes 1-Lipschitz functions $f : X \to \mathbb{R}$ such that $|f(x_1) - f(x_2)| \leq d(x_1, x_2)$. Note that the 2-Wasserstein metric $W_2(\|\cdot\|_2)$ (or simply W_2) has a closed-form for Gaussians [35]:

$$W_2(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}(\mu_j, \Sigma_j))^2 = \|\mu_i - \mu_j\|_2^2 + \|\Sigma_i - \Sigma_j\|_{\mathcal{F}}^2, \tag{21}$$

where $\|\cdot\|_{\mathcal{F}}$ denotes the Frobenius norm. We can observe that for point masses (i.e., $\Sigma_i, \Sigma_j \to 0$), the 2-Wasserstein metric is equivalent to the Euclidean distance between the two points.

Lemma 4 (p-Wasserstein Inequality [50]). *For any two distributions μ, λ, if $p \leq q$:*

$$W_p(\mu, \lambda) \leq W_q(\mu, \lambda). \tag{22}$$

Lemma 5 (Bounds on Wasserstein distances [41]). *For any two distributions μ, λ over a space X, for all $p \geq 1$:*

$$W_1(\mu, \lambda) \leq W_p(\mu, \lambda) \leq \mathrm{diam}(X)^{\frac{p-1}{p}} W_1(\mu, \lambda)^{\frac{1}{p}}. \tag{23}$$

B Proofs

Remark 1. *If $p = 1$, or both the environment and policy are deterministic, A1 holds.*

Proof. The existence proof is virtually identical to the proof of Thm. 3.12 of [14], except it discards $\max_{\mathbf{a} \in \mathcal{A}}$ operations in favor of expectations under a policy π. We need to show that the following fixed-point update is a contraction:

$$\mathcal{F}(d_\pi)(\mathbf{s}_i, \mathbf{s}_j) \coloneqq c_R |r_i^\pi - r_j^\pi| \ + \ c_T W_p(d_\pi)(\mathcal{P}^\pi(\cdot|\mathbf{s}_i), \mathcal{P}^\pi(\cdot|\mathbf{s}_j)),$$

and invoke the Banach fixed-point theorem to show the existence of a unique metric.

First, consider the case where $p = 1$:

$$\begin{aligned}
&\mathcal{F}(d_\pi)(\mathbf{s}_i, \mathbf{s}_j) - \mathcal{F}(d'_\pi)(\mathbf{s}_i, \mathbf{s}_j) \\
&= c_T \left(W_1(d_\pi)(\mathcal{P}^\pi(\cdot|\mathbf{s}_i), \mathcal{P}^\pi(\cdot|\mathbf{s}_j)) - W_1(d'_\pi)(\mathcal{P}^\pi(\cdot|\mathbf{s}_i), \mathcal{P}^\pi(\cdot|\mathbf{s}_j)) \right) \\
&= c_T \left(W_1(d_\pi - d'_\pi + d'_\pi)(\mathcal{P}^\pi(\cdot|\mathbf{s}_i), \mathcal{P}^\pi(\cdot|\mathbf{s}_j)) - W_1(d'_\pi)(\mathcal{P}^\pi(\cdot|\mathbf{s}_i), \mathcal{P}^\pi(\cdot|\mathbf{s}_j)) \right) \\
&\leq c_T \left(W_1(\|d_\pi - d'_\pi\|_\infty + d'_\pi)(\mathcal{P}^\pi(\cdot|\mathbf{s}_i), \mathcal{P}^\pi(\cdot|\mathbf{s}_j)) - W_1(d'_\pi)(\mathcal{P}^\pi(\cdot|\mathbf{s}_i), \mathcal{P}^\pi(\cdot|\mathbf{s}_j)) \right) \\
&\leq c_T \left(\|d_\pi - d'_\pi\|_\infty + W_1(d'_\pi)(\mathcal{P}^\pi(\cdot|\mathbf{s}_i), \mathcal{P}^\pi(\cdot|\mathbf{s}_j)) - W_1(d'_\pi)(\mathcal{P}^\pi(\cdot|\mathbf{s}_i), \mathcal{P}^\pi(\cdot|\mathbf{s}_j)) \right) \\
&= c_T \|d_\pi - d'_\pi\|_\infty, \ \forall (\mathbf{s}_i, \mathbf{s}_j) \in \mathcal{S} \times \mathcal{S}.
\end{aligned}$$

For $c_T \in [0, 1)$, there exists a unique fixed-point due to the Banach fixed-point theorem.

Next, we consider the case where both $\mathcal{P}$ and π are deterministic, such that $\mathcal{P}^\pi$ is a delta distribution. Observe that for point masses, $W_p(d)(\delta(\mathbf{s}_i), \delta(\mathbf{s}_j)) = d(\mathbf{s}_i, \mathbf{s}_j)$, due to Definition 5 of the Wasserstein metric. Then:

$$\begin{aligned}\mathcal{F}(d_\pi)(\mathbf{s}_i, \mathbf{s}_j) - \mathcal{F}(d'_\pi)(\mathbf{s}_i, \mathbf{s}_j) &= c_T \left(W_p(d_\pi)(\mathcal{P}^\pi(\cdot|\mathbf{s}_i), \mathcal{P}^\pi(\cdot|\mathbf{s}_j)) - W_p(d'_\pi)(\mathcal{P}^\pi(\cdot|\mathbf{s}_i), \mathcal{P}^\pi(\cdot|\mathbf{s}_j))\right) \\ &= c_T(d_\pi(\mathbf{s}'_i, \mathbf{s}'_j) - d'_\pi(\mathbf{s}'_i, \mathbf{s}'_j)) \\ &\leq c_T \|d_\pi - d'_\pi\|_\infty, \ \forall (\mathbf{s}_i, \mathbf{s}_j) \in \mathcal{S} \times \mathcal{S}.\end{aligned}$$

Then, fixed point iterations that update the metric as $d^{(n+1)}(\mathbf{s}_i, \mathbf{s}_j) \leftarrow \mathcal{F}(d^{(n)})(\mathbf{s}_i, \mathbf{s}_j)$ will converge for finite MDPs. ■

Lemma 6 (p-Wasserstein value difference bound)**.** *For an on-policy bisimulation metric given by Eq. (6), for any $c_T \in [\gamma, 1)$ and $p \geq 1$, the bisimulation distance between a pair of states upper-bounds the difference in their values:*

$$c_R |V^\pi(\mathbf{s}_i) - V^\pi(\mathbf{s}_j)| \ \leq \ d_\pi(\mathbf{s}_i, \mathbf{s}_j), \ \forall (\mathbf{s}_i, \mathbf{s}_j) \in \mathcal{S} \times \mathcal{S}. \tag{24}$$

Proof. The proof follows similarly to the proofs of Thm. 5.1 of [13] and Thm. 3 of [10]. We will prove by induction. Consider the following updates:

$$\begin{aligned}V^{(n+1)}(\mathbf{s}_i) &= r_i^\pi + \gamma \int_{\mathbf{s}' \in \mathcal{S}} \mathcal{P}^\pi(\mathbf{s}'|\mathbf{s}_i) V^{(n)}(\mathbf{s}') d\mathbf{s}' \\ d_\pi^{(n+1)}(\mathbf{s}_i, \mathbf{s}_j) &= c_R |r_i^\pi - r_j^\pi| \ + \ c_T W_p(d_\pi^{(n)})(\mathcal{P}^\pi(\cdot|\mathbf{s}_i), \mathcal{P}^\pi(\cdot|\mathbf{s}_j)).\end{aligned}$$

We need to show that the following holds for all $n \in \mathbb{N}$:

$$c_R \left|V^{(n)}(\mathbf{s}_i) - V^{(n)}(\mathbf{s}_j)\right| \ \leq \ d_\pi^{(n)}(\mathbf{s}_i, \mathbf{s}_j), \ \forall (\mathbf{s}_i, \mathbf{s}_j) \in \mathcal{S} \times \mathcal{S}.$$

Then, Eq. (24) holds by taking a limit $n \to \infty$. The base case holds since:

$$\left|V^{(0)}(\mathbf{s}_i) - V^{(0)}(\mathbf{s}_j)\right| = d_\pi^{(0)}(\mathbf{s}_i, \mathbf{s}_j) = 0, \ \forall (\mathbf{s}_i, \mathbf{s}_j) \in \mathcal{S} \times \mathcal{S}.$$

In the general case:

$$\begin{aligned}c_R |V^{(n+1)}(\mathbf{s}_i) - V^{(n+1)}(\mathbf{s}_j)| &= c_R \left| r_i^\pi - r_j^\pi + \gamma \int_{\mathbf{s}' \in \mathcal{S}} \left(\mathcal{P}^\pi(\mathbf{s}'|\mathbf{s}_i) - \mathcal{P}^\pi(\mathbf{s}'|\mathbf{s}_j)\right) V^{(n)}(\mathbf{s}') d\mathbf{s}' \right| \\ &\leq c_R \left| r_i^\pi - r_j^\pi \right| + c_R \gamma \left| \int_{\mathbf{s}' \in \mathcal{S}} \left(\mathcal{P}^\pi(\mathbf{s}'|\mathbf{s}_i) - \mathcal{P}^\pi(\mathbf{s}'|\mathbf{s}_j)\right) V^{(n)}(\mathbf{s}') d\mathbf{s}' \right| \\ &= c_R \left| r_i^\pi - r_j^\pi \right| + c_T \left| \int_{\mathbf{s}' \in \mathcal{S}} \left(\mathcal{P}^\pi(\mathbf{s}'|\mathbf{s}_i) - \mathcal{P}^\pi(\mathbf{s}'|\mathbf{s}_j)\right) \frac{c_R \gamma}{c_T} V^{(n)}(\mathbf{s}') d\mathbf{s}' \right|.\end{aligned}$$

Notice that by the induction hypothesis, $c_R V^{(n)}(\mathbf{s})$ is a 1-Lipschitz function with respect to the distance function $d_\pi^{(n)}$, i.e., $c_R V^{(n)}(\mathbf{s}) \in \text{Lip}_{1, d_\pi^{(n)}}$. Then, since $\gamma \leq c_T$ by assumption, $\frac{c_R \gamma}{c_T} V^{(n)}(\mathbf{s})$ is also 1-Lipschitz. Using the dual form of the W_1 metric in Eq. (20):

$$\begin{aligned}c_R |V^{(n+1)}(\mathbf{s}_i) - V^{(n+1)}(\mathbf{s}_j)| &\leq c_R \left| r_i^\pi - r_j^\pi \right| + c_T W_1(d_\pi^{(n)})(\mathcal{P}^\pi(\cdot|\mathbf{s}_i), \mathcal{P}^\pi(\cdot|\mathbf{s}_j)) \\ &\leq c_R \left| r_i^\pi - r_j^\pi \right| + c_T W_p(d_\pi^{(n)})(\mathcal{P}^\pi(\cdot|\mathbf{s}_i), \mathcal{P}^\pi(\cdot|\mathbf{s}_j)) \\ &= d_\pi^{(n+1)},\end{aligned}$$

where the last inequality is due to Lemma 4. ■

Lemma 7 (Value function difference bound for different discount factors [37])**.** *Consider two otherwise identical MDPs with different discount factors $\gamma_1 \leq \gamma_2$, and a bounded reward function $R \in [0, 1]$. Let V_γ^π denote the value function for policy π given discount factor γ.*

$$\left|V_{\gamma_1}^\pi(\mathbf{s}) - V_{\gamma_2}^\pi(\mathbf{s})\right| \leq \frac{\gamma_2 - \gamma_1}{(1 - \gamma_1)(1 - \gamma_2)}, \forall \mathbf{s} \in \mathcal{S}. \tag{25}$$

Proof. See Thm. 2 of [37]. ■

Theorem 1 (Generalized value difference bound)**.** *Let the reward function be bounded as $R \in [0,1]$. For an on-policy bisimulation metric given by Eq. (6), for any $c_T \in [0,1)$ and $p \geq 1$, define $\overline{\gamma} = \min(c_T, \gamma)$. Given A1, the bisimulation distance between a pair of states upper-bounds the difference in their values:*

$$c_R|V^\pi(\mathbf{s}_i) - V^\pi(\mathbf{s}_j)| \;\leq\; d_\pi(\mathbf{s}_i, \mathbf{s}_j) + \frac{2c_R(\gamma - \overline{\gamma})}{(1-\gamma)(1-c_T)},\; \forall (\mathbf{s}_i, \mathbf{s}_j) \in \mathcal{S} \times \mathcal{S}. \tag{7}$$

Proof. A bisimulation metric with $c_T \leq \gamma$ can be viewed as approximating a value function for another MDP with $\gamma' = c_T$. We will make use of this view and apply Lemma 6 with Lemma 7 to derive the above relation.

First, note that by Lemma 6:

$$c_R|V^\pi_{c_T}(\mathbf{s}_i) - V^\pi_{c_T}(\mathbf{s}_j)| \;\leq\; d_\pi(\mathbf{s}_i, \mathbf{s}_j),\; \forall (\mathbf{s}_i, \mathbf{s}_j) \in \mathcal{S} \times \mathcal{S}.$$

Then;

$$\begin{aligned}
c_R|V^\pi(\mathbf{s}_i) - V^\pi(\mathbf{s}_j)| &= c_R\left|V^\pi(\mathbf{s}_i) - V^\pi_{c_T}(\mathbf{s}_i) + V^\pi_{c_T}(\mathbf{s}_i) - V^\pi(\mathbf{s}_j) + V^\pi_{c_T}(\mathbf{s}_j) - V^\pi_{c_T}(\mathbf{s}_j)\right| \\
&\leq c_R\left(|V^\pi(\mathbf{s}_i) - V^\pi_{c_T}(\mathbf{s}_i)| + |V^\pi(\mathbf{s}_j) - V^\pi_{c_T}(\mathbf{s}_j)| + |V^\pi_{c_T}(\mathbf{s}_i) - V^\pi_{c_T}(\mathbf{s}_j)|\right) \\
&\leq d_\pi(\mathbf{s}_i, \mathbf{s}_j) + c_R\left(|V^\pi(\mathbf{s}_i) - V^\pi_{c_T}(\mathbf{s}_i)| + |V^\pi(\mathbf{s}_j) - V^\pi_{c_T}(\mathbf{s}_j)|\right) \\
&\leq d_\pi(\mathbf{s}_i, \mathbf{s}_j) + \frac{2c_R(\gamma - c_T)}{(1-\gamma)(1-c_T)},
\end{aligned}$$

where the last inequality is due to Lemma 7. Due to Lemma 6:

$$c_R|V^\pi(\mathbf{s}_i) - V^\pi(\mathbf{s}_j)| \;\leq\; d_\pi(\mathbf{s}_i, \mathbf{s}_j) + \frac{2c_R(\gamma - \min(c_T, \gamma))}{(1-\gamma)(1-c_T)}. \tag{26}$$

■

Lemma 8 (On-policy VFA bound)**.** *Let the reward function be bounded as $R \in [0,1]$ and $\Phi : \mathcal{S} \to \widetilde{\mathcal{S}}$ a function mapping states to a finite partitioning $\widetilde{\mathcal{S}}$ such that $\Phi(\mathbf{s}_i) = \Phi(\mathbf{s}_j) \Rightarrow d_\pi(\mathbf{s}_i, \mathbf{s}_j) \leq 2\epsilon$, which produces an aggregated MDP $\langle \widetilde{\mathcal{S}}, \mathcal{A}, \widetilde{\mathcal{P}}, \widetilde{R}, \widetilde{\rho}_0 \rangle$. For $c_T \in [\gamma, 1)$:*

$$|V^\pi(\mathbf{s}_i) - \widetilde{V}^\pi(\Phi(\mathbf{s}_i))| \;\leq\; \frac{2\epsilon}{c_R(1-\gamma)},\; \forall \mathbf{s}_i \in \mathcal{S}. \tag{27}$$

Proof. Let ξ be a measure on $\mathcal{S}$. Given a partition $\Phi(\mathbf{s}) \in \widetilde{\mathcal{S}}$, i.e., a set of points in $\mathcal{S}$ clustered in an ϵ-neighborhood such that $\xi(\Phi(\mathbf{s})) > 0$, we can define the reward function and transition probabilities of a ξ-average finite MDP as in Thm. 3.21 of [14]:

$$\begin{aligned}
\tilde{r}^\pi(\Phi(\mathbf{s})) &= \frac{1}{\xi(\Phi(\mathbf{s}))} \int\limits_{\mathbf{z} \in \Phi(\mathbf{s})} r^\pi(\mathbf{z}) d\xi(\mathbf{z}), \\
\widetilde{\mathcal{P}}^\pi(\Phi(\mathbf{s}')|\Phi(\mathbf{s})) &= \frac{1}{\xi(\Phi(\mathbf{s}))} \int\limits_{\mathbf{z} \in \Phi(\mathbf{s})} \mathcal{P}^\pi(\Phi(\mathbf{s}')|\mathbf{z}) d\xi(\mathbf{z}).
\end{aligned}$$

Then,

$$
\begin{aligned}
&|V^\pi(\mathbf{s}) - \widetilde{V}^\pi(\Phi(\mathbf{s}))| \\
&= \left| r^\pi(\mathbf{s}) - \tilde{r}^\pi(\Phi(\mathbf{s})) + \gamma \int_{\mathbf{s}'\in\mathcal{S}} \mathcal{P}^\pi(\mathbf{s}'|\mathbf{s}) V^\pi(\mathbf{s}') d\mathbf{s}' - \gamma \int_{\Phi(\mathbf{s}')\in\widetilde{\mathcal{S}}} \widetilde{\mathcal{P}}^\pi(\Phi(\mathbf{s}')|\Phi(\mathbf{s}))\widetilde{V}^\pi(\Phi(\mathbf{s}'))d\Phi(\mathbf{s}') \right| \\
&\leq \frac{1}{\xi(\Phi(\mathbf{s}))} \int_{\mathbf{z}\in\Phi(\mathbf{s})} |r^\pi(\mathbf{s}) - r^\pi(\mathbf{z})| + \gamma \left| \int_{\mathbf{s}'\in\mathcal{S}} \mathcal{P}^\pi(\mathbf{s}'|\mathbf{s}) V^\pi(\mathbf{s}') d\mathbf{s}' - \int_{\Phi(\mathbf{s}')\in\widetilde{\mathcal{S}}} \widetilde{\mathcal{P}}^\pi(\Phi(\mathbf{s}')|\mathbf{z})\widetilde{V}^\pi(\Phi(\mathbf{s}'))d\Phi(\mathbf{s}') \right| d\xi(\mathbf{z}) \\
&\leq \frac{1}{\xi(\Phi(\mathbf{s}))} \int_{\mathbf{z}\in\Phi(\mathbf{s})} |r^\pi(\mathbf{s}) - r^\pi(\mathbf{z})| + \gamma \left| \int_{\mathbf{s}'\in\mathcal{S}} \mathcal{P}^\pi(\mathbf{s}'|\mathbf{s}) V^\pi(\mathbf{s}') - \mathcal{P}^\pi(\mathbf{s}'|\mathbf{z})\widetilde{V}^\pi(\Phi(\mathbf{s}'))d\mathbf{s}' \right| d\xi(\mathbf{z}) \\
&\leq \frac{1}{\xi(\Phi(\mathbf{s}))} \int_{\mathbf{z}\in\Phi(\mathbf{s})} |r^\pi(\mathbf{s}) - r^\pi(\mathbf{z})| + \gamma \left| \int_{\mathbf{s}'\in\mathcal{S}} \mathcal{P}^\pi(\mathbf{s}'|\mathbf{s}) V^\pi(\mathbf{s}') - \mathcal{P}^\pi(\mathbf{s}'|\mathbf{z}) V^\pi(\mathbf{s}') d\mathbf{s}' \right| d\xi(\mathbf{z}) \\
&\quad + \gamma \frac{1}{\xi(\Phi(\mathbf{s}))} \int_{\mathbf{z}\in\Phi(\mathbf{s})} \left| \int_{\mathbf{s}'\in\mathcal{S}} \mathcal{P}^\pi(\mathbf{s}'|\mathbf{z}) \left(V^\pi(\mathbf{s}') - \widetilde{V}^\pi(\Phi(\mathbf{s}')) \right) d\mathbf{s}' \right| d\xi(\mathbf{z}) \\
&\leq \frac{1}{\xi(\Phi(\mathbf{s}))} \int_{\mathbf{z}\in\Phi(\mathbf{s})} |r^\pi(\mathbf{s}) - r^\pi(\mathbf{z})| + \gamma \left| \int_{\mathbf{s}'\in\mathcal{S}} \left(\mathcal{P}^\pi(\mathbf{s}'|\mathbf{s}) - \mathcal{P}^\pi(\mathbf{s}'|\mathbf{z}) \right) V^\pi(\mathbf{s}') d\mathbf{s}' \right| d\xi(\mathbf{z}) \\
&\quad + \gamma \|V^\pi - \widetilde{V}^\pi\|_\infty \\
&\leq \frac{c_R^{-1}}{\xi(\Phi(\mathbf{s}))} \int_{\mathbf{z}\in\Phi(\mathbf{s})} c_R \left| r^\pi(\mathbf{s}) - r^\pi(\mathbf{z}) \right| + c_T \left| \int_{\mathbf{s}'\in\mathcal{S}} \left(\mathcal{P}^\pi(\mathbf{s}'|\mathbf{s}) - \mathcal{P}^\pi(\mathbf{s}'|\mathbf{z}) \right) \frac{c_R \gamma}{c_T} V^\pi(\mathbf{s}') d\mathbf{s}' \right| d\xi(\mathbf{z}) \\
&\quad + \gamma \|V^\pi - \widetilde{V}^\pi\|_\infty,
\end{aligned}
$$

where $\|\cdot\|_\infty$ is the supremum norm over $\mathcal{S}$. Due to Lemma 6, $c_R V(\mathbf{s})$ is a 1-Lipschitz function with respect to the distance function d_π. Then, since $\gamma \leq c_T$ by assumption, $\frac{c_R\gamma}{c_T} V(\mathbf{s})$ is also 1-Lipschitz. Using the dual form of the W_1 metric:

$$
\begin{aligned}
|V^\pi(\mathbf{s}) - \widetilde{V}^\pi(\Phi(\mathbf{s}))| &\leq \frac{c_R^{-1}}{\xi(\Phi(\mathbf{s}))} \int_{\mathbf{z}\in\Phi(\mathbf{s})} c_R \left| r^\pi(\mathbf{s}) - r^\pi(\mathbf{z}) \right| + c_T W_1(d_\pi)(\mathcal{P}^\pi(\mathbf{s}'|\mathbf{s}), \mathcal{P}^\pi(\mathbf{s}'|\mathbf{z}))d\xi(\mathbf{z}) \\
&\quad + \gamma \|V^\pi - \widetilde{V}^\pi\|_\infty \\
&\leq \frac{c_R^{-1}}{\xi(\Phi(\mathbf{s}))} \int_{\mathbf{z}\in\Phi(\mathbf{s})} d_\pi(\mathbf{s}, \mathbf{z}) d\xi(\mathbf{z}) + \gamma \|V^\pi - \widetilde{V}^\pi\|_\infty \\
&\leq c_R^{-1} 2\epsilon + \gamma \|V^\pi - \widetilde{V}^\pi\|_\infty.
\end{aligned}
$$

Thus, taking the supremum on the LHS over the state space $\mathcal{S}$:

$$
|V^\pi(\mathbf{s}) - \widetilde{V}^\pi(\Phi(\mathbf{s}))| \leq \frac{2\epsilon}{c_R(1-\gamma)}, \ \forall \mathbf{s} \in \mathcal{S}.
$$

■

Theorem 2 (Generalized VFA bound)**.** *Let rewards be bounded as $R \in [0,1]$ and $\Phi : \mathcal{S} \to \widetilde{\mathcal{S}}$ be a function mapping states to a finite partitioning $\widetilde{\mathcal{S}}$ such that $\Phi(\mathbf{s}_i) = \Phi(\mathbf{s}_j) \Rightarrow d_\pi(\mathbf{s}_i, \mathbf{s}_j) \leq 2\epsilon$, which produces an aggregated MDP $\langle \widetilde{\mathcal{S}}, \mathcal{A}, \widetilde{\mathcal{P}}, \widetilde{R}, \widetilde{\rho}_0 \rangle$. For any $c_T \in [0,1)$, let $\overline{\gamma} = \min(c_T, \gamma)$. Given A1,*

$$
|V^\pi(\mathbf{s}) - \widetilde{V}^\pi(\Phi(\mathbf{s}))| \ \leq \ \frac{2\epsilon}{c_R(1-\overline{\gamma})} + \frac{2(\gamma - \overline{\gamma})}{(1-\gamma)(1-c_T)}, \ \forall \mathbf{s} \in \mathcal{S}. \tag{8}
$$

Proof. Due to Lemma 8,

$$|V^\pi_{\gamma'}(\mathbf{s}) - \widetilde{V}^\pi_{\gamma'}(\Phi(\mathbf{s}))| \leq \frac{2\epsilon}{c_R(1-\gamma')}, \tag{28}$$

where $V^\pi_{\gamma'}$ denotes the expected value under a policy π for a discount factor γ'. Then, for $c_T < \gamma$:

$$\begin{aligned}
|V^\pi(\mathbf{s}) - \widetilde{V}^\pi(\Phi(\mathbf{s}))| &= |V^\pi(\mathbf{s}) - \widetilde{V}^\pi(\Phi(\mathbf{s})) + \widetilde{V}^\pi_{c_T}(\Phi(\mathbf{s})) - \widetilde{V}^\pi_{c_T}(\Phi(\mathbf{s})) + V^\pi_{c_T}(\mathbf{s}) - V^\pi_{c_T}(\mathbf{s})| \\
&\leq |V^\pi_{c_T}(\mathbf{s}) - \widetilde{V}^\pi_{c_T}(\Phi(\mathbf{s}))| \ + \ |V^\pi(\mathbf{s}) - V^\pi_{c_T}(\mathbf{s})| \ + \ |\widetilde{V}^\pi(\Phi(\mathbf{s})) - \widetilde{V}^\pi_{c_T}(\Phi(\mathbf{s}))| \\
&\leq \frac{2\epsilon}{c_R(1-c_T)} \ + \ |V^\pi(\mathbf{s}) - V^\pi_{c_T}(\mathbf{s})| \ + \ |\widetilde{V}^\pi(\Phi(\mathbf{s})) - \widetilde{V}^\pi_{c_T}(\Phi(\mathbf{s}))| \\
&\leq \frac{2\epsilon}{c_R(1-c_T)} + \frac{2(\gamma - c_T)}{(1-\gamma)(1-c_T)},
\end{aligned}$$

where the second and third inequalities are due to Eq. (28) and Lemma 7 respectively. For $\gamma \leq c_T$, we recover Lemma 8, hence, for all $c_T \in [0, 1)$:

$$|V^\pi(\mathbf{s}) - \widetilde{V}^\pi(\Phi(\mathbf{s}))| \ \leq \ \frac{2\epsilon}{c_R(1-\min(c_T,\gamma))} + \frac{2(\gamma - \min(c_T,\gamma))}{(1-\gamma)(1-c_T)}, \ \forall \mathbf{s} \in \mathcal{S}.$$

■

Lemma 1 (Diameter of $\mathcal{S}$ is bounded)**.** *Let $d : \mathcal{S} \times \mathcal{S} \to [0, \infty)$ be any bisimulation metric:*

$$\text{diam}(\mathcal{S}; d) := \sup_{\mathbf{s}_i, \mathbf{s}_j \in \mathcal{S} \times \mathcal{S}} d(\mathbf{s}_i, \mathbf{s}_j) \leq \frac{c_R}{1 - c_T}(R_{\max} - R_{\min}). \tag{9}$$

Proof. This lemma is a slight generalization of the distance bounds given in Thm. 3.12 of [14], and the proof follows similarly:

$$\begin{aligned}
d(\mathbf{s}_i, \mathbf{s}_j) &= \max_{\mathbf{a} \in \mathcal{A}} (c_R|R(\mathbf{s}_i, \mathbf{a}) - R(\mathbf{s}_j, \mathbf{a})| \ + \ c_T W_1(d)(\mathcal{P}(\cdot|\mathbf{s}_i, \mathbf{a}), \mathcal{P}(\cdot|\mathbf{s}_j, \mathbf{a}))) \\
&\leq c_R(R_{\max} - R_{\min}) + c_T \text{diam}(\mathcal{S}; d), \ \forall (\mathbf{s}_i, \mathbf{s}_j) \in \mathcal{S} \times \mathcal{S},
\end{aligned}$$

due to Lemma 5 (upper bound as $p \to \infty$). Then,

$$\begin{aligned}
\text{diam}(\mathcal{S}; d) &\leq c_R(R_{\max} - R_{\min}) + c_T \text{diam}(\mathcal{S}; d) \\
&\leq \frac{c_R}{1 - c_T}(R_{\max} - R_{\min}).
\end{aligned}$$

■

Theorem 3 (Boundedness condition for convergence)**.** *Assume $\mathcal{S}$ is compact. If the support of an approximate dynamics model $\widehat{\mathcal{P}}$, i.e., $\mathcal{S}' = \text{supp}(\widehat{\mathcal{P}})$, is a closed subset of $\mathcal{S}$, then there exists a unique on-policy bisimulation metric $\widehat{d}_\pi$ of the form Eq. (10), and this metric is bounded:*

$$\text{supp}(\widehat{\mathcal{P}}) \subseteq \mathcal{S} \Rightarrow \text{diam}(\mathcal{S}; \widehat{d}_\pi) \leq \frac{c_R}{1 - c_T}(R_{\max} - R_{\min}). \tag{11}$$

Proof. The existence proof is virtually identical to the proof of Remark 1, except it replaces $\mathcal{P}$ with an approximate dynamics model $\widehat{\mathcal{P}}$. This is possible since $\mathcal{S}$ is compact by assumption such that $\text{supp}(\widehat{\mathcal{P}}) \subseteq \mathcal{S}$ is also compact:

$$\begin{aligned}
&\mathcal{F}(d_\pi)(\mathbf{s}_i, \mathbf{s}_j) - \mathcal{F}(d'_\pi)(\mathbf{s}_i, \mathbf{s}_j) \\
&= c_T \left(W_1(d_\pi)(\widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_i), \widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_j)) - W_1(d'_\pi)(\widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_i), \widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_j)) \right) \\
&= c_T \left(W_1(d_\pi - d'_\pi + d'_\pi)(\widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_i), \widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_j)) - W_1(d'_\pi)(\widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_i), \widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_j)) \right) \\
&\leq c_T \left(W_1(\|d_\pi - d'_\pi\|_\infty + d'_\pi)(\widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_i), \widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_j)) - W_1(d'_\pi)(\widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_i), \widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_j)) \right) \\
&\leq c_T \left(\|d_\pi - d'_\pi\|_\infty + W_1(d'_\pi)(\widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_i), \widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_j)) - W_1(d'_\pi)(\widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_i), \widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_j)) \right) \\
&= c_T \|d_\pi - d'_\pi\|_\infty, \ \forall (\mathbf{s}_i, \mathbf{s}_j) \in \mathcal{S} \times \mathcal{S},
\end{aligned}$$

which implies $\mathcal{F}$ is a c_T-contraction. It remains to prove that the distance is bounded. First, note that due to Lemma 5:

$$\text{supp}(\widehat{\mathcal{P}}) \subseteq \mathcal{S} \Rightarrow \sup_{\mathbf{s}_i,\mathbf{s}_j \in \mathcal{S}\times\mathcal{S}} W_p(\widehat{d}_\pi)(\widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_i), \widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_j)) \leq \text{diam}(\mathcal{S}; \widehat{d}_\pi),\ \forall p \geq 1. \tag{29}$$

Then, similarly to Lemma 1,

$$\begin{aligned}\widehat{d}_\pi(\mathbf{s}_i, \mathbf{s}_j) &= c_R|r_i^\pi - r_j^\pi| \ + \ c_T W_p(\widehat{d}_\pi)(\widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_i), \widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}_j)) \\ &\leq c_R(R_{\max} - R_{\min}) + c_T \text{diam}(\mathcal{S}; \widehat{d}_\pi),\ \forall (\mathbf{s}_i, \mathbf{s}_j) \in \mathcal{S} \times \mathcal{S},\end{aligned}$$

which implies,

$$\begin{aligned}\text{diam}(\mathcal{S}; \widehat{d}_\pi) &\leq c_R(R_{\max} - R_{\min}) + c_T \text{diam}(\mathcal{S}; \widehat{d}_\pi) \\ &\leq \frac{c_R}{1 - c_T}(R_{\max} - R_{\min}).\end{aligned}$$

■

Lemma 2 (A reason for caution in on-policy bisimulation)**.** *On-policy bisimulation metrics of the form Eq. (6) have an upper bound determined by their policy:*

$$\text{diam}(\mathcal{S}; d_\pi) \leq \frac{c_R}{1 - c_T} \sup_{i,j} |r_i^\pi - r_j^\pi|. \tag{15}$$

Proof. In the on-policy case, the bound in Lemma 1 can be much tighter depending on the policy:

$$\begin{aligned}d_\pi(\mathbf{s}_i, \mathbf{s}_j) &= c_R|r_i^\pi - r_j^\pi| \ + \ c_T W_1(d_\pi)(\mathcal{P}^\pi(\cdot|\mathbf{s}_i), \mathcal{P}^\pi(\cdot|\mathbf{s}_j)) \\ &\leq c_R \sup|r_i^\pi - r_j^\pi| \ + \ c_T \text{diam}(\mathcal{S}; d_\pi),\ \forall (\mathbf{s}_i, \mathbf{s}_j) \in \mathcal{S} \times \mathcal{S}.\end{aligned}$$

As before,

$$\text{diam}(\mathcal{S}; d_\pi) \leq \frac{c_R}{1 - c_T} \sup|r_i^\pi - r_j^\pi|.$$

■

Lemma 3 (Relating collapse and low-dispersion rewards)**.** *Assume deterministic transitions and the existence of a stationary distribution ρ_π over states. Given a bisimulation metric of the form Eq. (6):*

$$\mu_{bd}^\pi = \frac{c_R}{1 - c_T} \mu_{rd}^\pi. \tag{17}$$

Proof. For point masses, $W_p(d)(\delta(\mathbf{s}_i), \delta(\mathbf{s}_j)) = d(\mathbf{s}_i, \mathbf{s}_j)$:

$$d_\pi(\mathbf{s}_i, \mathbf{s}_j) = c_R|r_i^\pi - r_j^\pi| \ + \ c_T d_\pi(\mathbf{s}_i', \mathbf{s}_j').$$

Simply taking an expectation under ν^π, due to the stationarity assumption:

$$\begin{aligned}\mu_{bd}^\pi &= c_R \mu_{rd}^\pi + c_T \mu_{bd}^\pi \\ &= \frac{c_R}{1 - c_T} \mu_{rd}^\pi.\end{aligned}$$

■

C Notes on Reward Scale, c_R and c_T

In recent work [10, 17, 53], various forms of bisimulation metrics have been presented with different scaling constants; $(c_R = 1 - c, c_T = c)$ as in Definition 1, and $(c_R = 1, c_T = \gamma)$ as in Definition 2. Here, we aim to add clarity to the effect of these choices and how they relate to the reward scale. First, note that due to Lemma 1, setting $c_R = 1 - c_T$ serves to ensure that $d \in [0, 1]$ when the reward range is specified as $(R_{\max} = 1, R_{\min} = 0)$ as in [13, 14].

Corollary 1 (Policy-independent mean and variance bounds)**.** *Due to Lemma 1, the first two moments of the random variable given by bisimulation distance have the following bounds independently of any policy π:*

$$\mu_{bd} \leq \frac{c_R(R_{\max} - R_{\min})}{2(1-c_T)}, \tag{30}$$

$$\sigma^2_{bd} \leq \frac{c_R^2(R_{\max} - R_{\min})^2}{4(1-c_T)^2}. \tag{31}$$

Conversely, if $(c_R = 1, c_T = \gamma)$ as in Zhang et al. [53], the formulation allows large distances between embeddings since γ is commonly set to a value close to 1. Large pairwise distances imply large norms and high variance (see Corollary 1), which may cause instabilities in optimization (especially in the absence of norm constraints), considering the compactness conditions discussed in Sec. 3.2.2.

Definition 7 (Variance of distances and reward differences)**.** *Given a stationary distribution ρ^π over states, and ν^π the distribution over pairs of states, $(\mathbf{s}_i, \mathbf{s}_j)$ sampled independently from ρ^π:*

$$(\sigma^\pi_{bd})^2 \coloneqq \mathbb{V}_{(\mathbf{s}_i,\mathbf{s}_j)\sim\nu^\pi}[d^\pi(\mathbf{s}_i,\mathbf{s}_j)] \qquad (\sigma^\pi_{rd})^2 \coloneqq \mathbb{V}_{(\mathbf{s}_i,\mathbf{s}_j)\sim\nu^\pi}[|r^\pi_i - r^\pi_j|]. \tag{32}$$

Proposition 1 (On-policy variance bound)**.** *Assume a deterministic environment. Given $c_T \in [0, \sqrt{0.5})$, the variance of the optimal on-policy bisimulation distance for an objective of the form Eq. (3) can be upper-bounded as follows:*

$$(\sigma^\pi_{bd})^2 \leq \frac{2c_R^2}{1-2c_T^2}(\sigma^\pi_{rd})^2 + \frac{c_R^2\,(1-2c_T)^2}{(1-2c_T^2)\,(1-c_T)^2}(\mu^\pi_{rd})^2, \tag{33}$$

while for all $c_T \in [0, 1)$, the bound in Eq. (31) applies.

Proof. Given a deterministic environment:

$$d_\pi(\mathbf{s}_i,\mathbf{s}_j) = c_R|r^\pi_i - r^\pi_j| \;+\; c_T d_\pi(\mathbf{s}'_i,\mathbf{s}'_j).$$

Then,

$$\begin{aligned}(\sigma^\pi_{bd})^2 &= \mathbb{E}_{(\mathbf{s}_i,\mathbf{s}_j)\sim\nu^\pi}[d_\pi(\mathbf{s}_i,\mathbf{s}_j)^2] - (\mu^\pi_{bd})^2 \\ &\leq 2c_R^2((\sigma^\pi_{rd})^2 + (\mu^\pi_{rd})^2) + 2c_T^2((\sigma^\pi_{bd})^2 + (\mu^\pi_{bd})^2) - (\mu^\pi_{bd})^2.\end{aligned}$$

When $c_T \geq \sqrt{0.5}$ (as in Zhang et al. [53]), the above bound is loose. However, $c_T < \sqrt{0.5}$ provides a convenient upper bound:

$$\begin{aligned}(\sigma^\pi_{bd})^2 &\leq \frac{2c_R^2}{1-2c_T^2}((\sigma^\pi_{rd})^2 + (\mu^\pi_{rd})^2) - (\mu^\pi_{bd})^2 \\ &= \frac{2c_R^2}{1-2c_T^2}(\sigma^\pi_{rd})^2 + \frac{c_R^2\,(1-2c_T)^2}{(1-2c_T^2)\,(1-c_T)^2}(\mu^\pi_{rd})^2,\end{aligned}$$

where the equality is due to Lemma 3. ■

Tighter bounds on the variance of the on-policy bisimulation metric may be important, since the statistics of r^π undergo change throughout training between policy updates. Hence, it is desirable to remove the dependence of the bound from the μ^π_{rd} term with a choice of $c_T = 0.5$, such that tighter bounds can be obtained. This choice renders the formulation more robust to changes in the scale of the expected rewards. The resulting bound for $c_T = 0.5$ has a simpler form:

$$(\sigma^\pi_{bd})^2 \leq 4c_R^2(\sigma^\pi_{rd})^2 \leq c_R^2(R_{\max} - R_{\min})^2. \tag{34}$$

Indeed, in Figure 4, we show that such a choice can stabilize the DBC [53] algorithm significantly, resulting in higher overall performance.

D Value Bounds with Model Error

Our goal in this section is to characterize the errors induced by using approximate dynamics and an imperfect encoder, with respect to estimating both the ground-truth on-policy bisimulation metric, as well as preserving the value function with the encoded state.

For this section, we first remark on the difference between three forms of the PBSM, for some fixed policy π:

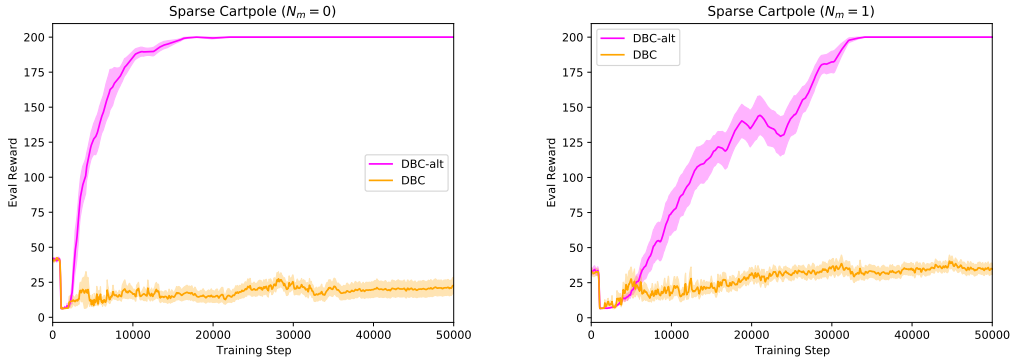

Figure 4: Performance on the Sparse Cartpole task, comparing the standard DBC algorithm (using $c_R = 1.0$ and $c_T = \gamma$) with an alternative weighting formulation (denoted "DBC-alt"), where $c_R = 0.5$ and $c_T = 0.5$. For this task, the alternative weighting formulation is much more robust to reward sparsity. Shaded bars are standard errors over 10 seeds.

- $d_\pi(\mathbf{s}_i, \mathbf{s}_j)$ is the *ground-truth* bisimulation metric, defined as the fixed point of the operator defined in Assumption 1 (when it exists uniquely).
- $\widehat{d}_\pi(\mathbf{s}_i, \mathbf{s}_j; \widehat{\mathcal{P}})$ is the fixed point of the same operator but with *approximate dynamics* (i.e., using $\widehat{\mathcal{P}}$ and $\widehat{r}^\pi$ instead of the true $\mathcal{P}$ and r^π; we leave out conditioning on $\widehat{r}^\pi$ for brevity).
- $\widehat{d}_{\pi,\phi}(\mathbf{s}_i, \mathbf{s}_j) := ||\phi(\mathbf{s}_i) - \phi(\mathbf{s}_j)||_q$ is a non-negative function of states, dependent on an encoder function ϕ. For example, ϕ may be produced by a metric learning process, such as by stochastic minimization of the objective in Eq. 3.

Notice that, if we attempt to learn ϕ with a metric learning process based on approximate dynamics, then the best we can do is obtain $\widehat{d}_{\pi,\phi} \to \widehat{d}_\pi$, which can still have some irreducible error. On the other hand, even with perfect approximate dynamics, the metric learning process may be incomplete, meaning $\widehat{d}_{\pi,\phi} \neq \widehat{d}_\pi$. We therefore hope to characterize the error into two types: dynamics approximation error and metric learning error.

Next, we define three types of model errors, relating to the quality of encoding and forward dynamics prediction (in terms of state and reward).

Definition 8 (Model and Encoder Errors)**.** *The bisimulation distance approximation error $\mathcal{E}_\phi$, transition probability error $\mathcal{E}_\mathcal{P}$, and reward prediction error $\mathcal{E}_r$ are given by*

$$\mathcal{E}_\phi := \|\widehat{d}_{\pi,\phi} - \widehat{d}_\pi\|_\infty \tag{35}$$

$$\mathcal{E}_\mathcal{P} := \sup_{\mathbf{s}\in\mathcal{S}} W_p(d_\pi)(\mathcal{P}^\pi(\cdot|\mathbf{s}), \widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s})) \tag{36}$$

$$\mathcal{E}_r := \|\widehat{r}^\pi - r^\pi\|_\infty \tag{37}$$

where $||\cdot||_\infty$ is the supremum (or uniform) norm over states and $\widehat{d}_\pi$ is the fixed-point bisimulation metric with the W_p distance defined by using $\widehat{\mathcal{P}}^\pi$ and $\widehat{r}^\pi$, instead of the true dynamics $\mathcal{P}^\pi$ and r^π.[7]

Note that we have defined our forward dynamics model errors here irrespective of ϕ (i.e., it may be used by $\widehat{\mathcal{P}}$ and/or $\widehat{r}^\pi$, or not). Our goal is to use these errors to bound the difference in the MDP value function induced by the approximate nature of the environmental model and state encoder. First, we consider how the model errors affect the optimal policy-dependent bisimulation distance that we can obtain given our approximate forward dynamics.

[7] Firstly, note that for stochastic reward signals and/or policies, this is a difference between expectations, meaning that there could be sampling noise. Otherwise, for the DBC use case, the observed reward collected by the agent is used for training (meaning there will be zero modelling error for the reward term when computing $\widehat{d}_\pi$ between observed states). Nevertheless, given a reward model, $\widehat{d}_{\pi,\phi}$ and $\widehat{d}_\pi$ can still be queried for states where the ground truth reward (or reward distribution) may not be known.

Lemma 9 (Bisimulation distance error)**.** *Let* $c_T \in [0,1)$ *and* $c_R \geq 0$*. Assume* $\text{supp}(\widehat{\mathcal{P}}) \subseteq \mathcal{S}$ *and* $1 - c_T a_p > 0$*. Then*

$$\|d_\pi - \widehat{d}_\pi\|_\infty \leq \frac{2c_R}{1-c_T a_p}\mathcal{E}_r + \frac{2c_T}{1-c_T a_p}\mathcal{E}_\mathcal{P} + \frac{c_T[a_p-1]}{1-c_T a_p}\text{diam}(\mathcal{S}; d_\pi) \tag{38}$$

where $a_p = 2^{(p-1)/p}$ *and* $\text{diam}(\mathcal{S}; d_\pi) \leq \frac{c_R}{1-c_T}(R_{\max} - R_{\min})$ *by Theorem 3.*

Proof. We can first use the triangle inequality to bound the difference between reward distances:

$$|r_i^\pi - r_j^\pi| \leq |r_i^\pi - \widehat{r}_i^\pi| + |r_j^\pi - \widehat{r}_i^\pi| \leq \mathcal{E}_r + \mathcal{E}_r + |\widehat{r}_i^\pi - \widehat{r}_j^\pi|,$$

so that $|r_i^\pi - r_j^\pi| - |\widehat{r}_i^\pi - \widehat{r}_j^\pi| \leq \mathcal{E}_r + \mathcal{E}_r$. Symmetrically, we can show that $|\widehat{r}_i^\pi - \widehat{r}_j^\pi| - |r_i^\pi - r_j^\pi| \leq 2\mathcal{E}_r$ as well. For notational clarity, let $a_p = 2^{(p-1)/p}$ and $W_p(d_\pi, \mathcal{P}^\pi) := W_p(d_\pi)(\mathcal{P}^\pi(\cdot|\mathbf{s}_i), \mathcal{P}^\pi(\cdot|\mathbf{s}_j))$, as well as similarly for $W_p(d_\pi, \widehat{\mathcal{P}}^\pi)$ and $W_p(\widehat{d}_\pi, \widehat{\mathcal{P}}^\pi)$.

First, by the Wasserstein triangle inequality [11], as for the reward difference:

$$|W_p(d_\pi, \widehat{\mathcal{P}}^\pi) - W_p(d_\pi, \mathcal{P}^\pi)| \leq 2\mathcal{E}_\mathcal{P}. \tag{39}$$

Second, the convexity of d^p implies that,

$$\begin{aligned}
W_p(||d_\pi - \widehat{d}_\pi||_\infty + d_\pi, \widehat{\mathcal{P}}^\pi) &= \left(\inf_{\omega \in \Omega} \mathbb{E}_{(\mathbf{s}_i, \mathbf{s}_j)\sim\omega}[(||d_\pi - \widehat{d}_\pi||_\infty + d_\pi(\mathbf{s}_i, \mathbf{s}_j))^p]\right)^{\frac{1}{p}} \\
&\leq \left(\inf_{\omega \in \Omega} 2^{p-1}\mathbb{E}_{(\mathbf{s}_i, \mathbf{s}_j)\sim\omega}[(||d_\pi - \widehat{d}_\pi||_\infty^p + d_\pi(\mathbf{s}_i, \mathbf{s}_j)^p]\right)^{\frac{1}{p}} \\
&\leq a_p \left(||d_\pi - \widehat{d}_\pi||_\infty^p + W_p^p(d_\pi, \widehat{\mathcal{P}}^\pi)\right)^{\frac{1}{p}} \\
&\leq a_p \left(\left[||d_\pi - \widehat{d}_\pi||_\infty + W_p(d_\pi, \widehat{\mathcal{P}}^\pi)\right]^p\right)^{1/p} \\
&= a_p \left(||d_\pi - \widehat{d}_\pi||_\infty + W_p(d_\pi, \widehat{\mathcal{P}}^\pi)\right).
\end{aligned} \tag{40}$$

Third, recall that when $\text{supp}(\widehat{\mathcal{P}}) \subseteq \mathcal{S}$, due to Lemma 5, we have:

$$W_p(d_\pi, \widehat{\mathcal{P}}^\pi) \leq \text{diam}(\mathcal{S}; d_\pi). \tag{41}$$

Then, the difference in distances can be bounded by:

$$\begin{aligned}
&|W_p(d_\pi, \mathcal{P}^\pi) - W_p(\widehat{d}_\pi, \widehat{\mathcal{P}}^\pi)| \\
&\leq |W_p(\widehat{d}_\pi, \widehat{\mathcal{P}}^\pi) - W_p(d_\pi, \widehat{\mathcal{P}}^\pi)| + |W_p(d_\pi, \mathcal{P}^\pi) - W_p(d_\pi, \widehat{\mathcal{P}}^\pi)| \\
&\leq |W_p(\widehat{d}_\pi, \widehat{\mathcal{P}}^\pi) - W_p(d_\pi, \widehat{\mathcal{P}}^\pi)| + 2\mathcal{E}_\mathcal{P} && \text{By Eq. 39} \\
&= |W_p(\widehat{d}_\pi - d_\pi + d_\pi, \widehat{\mathcal{P}}^\pi) - W_p(d_\pi, \widehat{\mathcal{P}}^\pi)| + 2\mathcal{E}_\mathcal{P} \\
&\leq |W_p(||\widehat{d}_\pi - d_\pi||_\infty + d_\pi, \widehat{\mathcal{P}}^\pi) - W_p(d_\pi, \widehat{\mathcal{P}}^\pi)| + 2\mathcal{E}_\mathcal{P} \\
&= |W_p(||d_\pi - \widehat{d}_\pi||_\infty + d_\pi, \widehat{\mathcal{P}}^\pi) - W_p(d_\pi, \widehat{\mathcal{P}}^\pi)| + 2\mathcal{E}_\mathcal{P} \\
&\leq |a_p||d_\pi - \widehat{d}_\pi||_\infty + a_p W_p(d_\pi, \widehat{\mathcal{P}}^\pi) - W_p(d_\pi, \widehat{\mathcal{P}}^\pi)| + 2\mathcal{E}_\mathcal{P} && \text{By Eq. 40} \\
&\leq a_p||d_\pi - \widehat{d}_\pi||_\infty + [a_p - 1]\text{diam}(\mathcal{S}; d_\pi) + 2\mathcal{E}_\mathcal{P}. && \text{By Eq. 41}
\end{aligned}$$

We can then plug these into the difference between the true and approximate policy-dependent bisimulation distances:

$$\begin{aligned}
|d_\pi(\mathbf{s}_i, \mathbf{s}_j) - \widehat{d}_\pi(\mathbf{s}_i, \mathbf{s}_j)| &\leq c_R \left||r_i^\pi - r_j^\pi| - |\widehat{r}_i^\pi - \widehat{r}_j^\pi|\right| + c_T \left|W_p(d_\pi, \mathcal{P}^\pi) - W_p(\widehat{d}_\pi, \widehat{\mathcal{P}}^\pi)\right| \\
&\leq 2c_R\mathcal{E}_r + c_T \left|a_p||d_\pi - \widehat{d}_\pi||_\infty + [a_p - 1]\text{diam}(\mathcal{S}; d_\pi) + 2\mathcal{E}_\mathcal{P}\right| \\
||d_\pi - \widehat{d}_\pi||_\infty &\leq 2c_R\mathcal{E}_r + 2c_T\mathcal{E}_\mathcal{P} + c_T a_p||d_\pi - \widehat{d}_\pi||_\infty + c_T[a_p - 1]\text{diam}(\mathcal{S}; d_\pi) \\
||d_\pi - \widehat{d}_\pi||_\infty &\leq \frac{2c_R}{1-c_T a_p}\mathcal{E}_r + \frac{2c_T}{1-c_T a_p}\mathcal{E}_\mathcal{P} + \frac{c_T[a_p-1]}{1-c_T a_p}\text{diam}(\mathcal{S}; d_\pi)
\end{aligned}$$

where the second-last inequality follows by taking the supremum over states for both sides. ■

For the remainder of this section, we assume $p = 1$.

Corollary 2 (Bisimulation distance error with $p = 1$)**.** *Let $p = 1$, with the remaining conditions as in Lemma 9. Then*

$$\|d_\pi - \widehat{d}_\pi\|_\infty \leq \frac{2c_R}{1 - c_T}\mathcal{E}_r + \frac{2c_T}{1 - c_T}\mathcal{E}_\mathcal{P}. \tag{42}$$

Proof. When $p = 1$, we have $a_p = a_1 = 1$, giving the expression above. ■

This bounds the error between the true on-policy bisimulation distance and the optimal *approximate* bisimulation distance (i.e., the best distance function we could hope to achieve with our encoder, given the error in our forward dynamics model). However, ultimately, we wish to bound the error in the value function in terms of $\widehat{d}_{\pi,\phi}$, not just $\widehat{d}_\pi$ (to take the error of the encoder ϕ into account, as well as that of the dynamics model). First, we can bound the true bisimulation distance in terms of the encoder and model error as follows:

Lemma 10 (Bound on bisimulation distance with encoder error)**.** *Consider the same conditions as Corollary 2. Then*

$$\|d_\pi - \widehat{d}_{\pi,\phi}\|_\infty \leq \mathcal{E}_\phi + \frac{2c_R}{1 - c_T}\mathcal{E}_r + \frac{2c_T}{1 - c_T}\mathcal{E}_\mathcal{P}. \tag{43}$$

Proof.

$$\begin{aligned}
\|d_\pi - \widehat{d}_{\pi,\phi}\|_\infty &= \|d_\pi - \widehat{d}_{\pi,\phi} - \widehat{d}_\pi + \widehat{d}_\pi\|_\infty \\
&\leq \|d_\pi - \widehat{d}_\pi\|_\infty + \|\widehat{d}_{\pi,\phi} - \widehat{d}_\pi\|_\infty \\
&\leq \frac{2c_R}{1 - c_T}\mathcal{E}_r + \frac{2c_T}{1 - c_T}\mathcal{E}_\mathcal{P} + \mathcal{E}_\phi
\end{aligned}$$

using Corollary 2 and Equation 35. ■

Thus, if we can relate d_π to the value function, we can also do so for $\widehat{d}_{\pi,\phi}$, as a function of model error.

Finally, we look at bounding the difference in the state value function, using the *approximate* bisimulation distance defined through the learned encoder (i.e., our partitioning Z is defined via $\widehat{d}_{\pi,\phi}$). Let $\widehat{\epsilon}$ be the aggregation radius in ϕ-space (meaning the maximum diameter with respect to $\widehat{d}_{\pi,\phi}$ per partition subset, or equivalence class, is at most $2\widehat{\epsilon}$):

$$\sup_{\mathbf{z} \in Z} \sup_{\mathbf{s}_i, \mathbf{s}_j \in \mathbf{z}} ||\phi(\mathbf{s}_i) - \phi(\mathbf{s}_j)||_q \ \leq \ 2\widehat{\epsilon}.$$

Notice that $\widehat{\epsilon}$ bounds the maximal diameter of the partition cells with respect to the *learned* metric, using ϕ, rather than the ground truth bisimulation distance.

Theorem 4 (VFA bound in terms of model error)**.** *Consider the same conditions as in Theorem 2, except that $c_T \in [\gamma, 1)$, $p = 1$, and $\Phi(\mathbf{s}_i) = \Phi(\mathbf{s}_j) \Rightarrow \widehat{d}_{\pi,\phi}(\mathbf{s}_i, \mathbf{s}_j) = \|\phi(\mathbf{s}_i) - \phi(\mathbf{s}_j)\|_q \leq 2\widehat{\epsilon}$. Then:*

$$|V^\pi(\mathbf{s}) - \widetilde{V}^\pi(\Phi(\mathbf{s}))| \leq \frac{1}{c_R(1 - \gamma)}\left(2\widehat{\epsilon} + \mathcal{E}_\phi + \frac{2c_R}{1 - c_T}\mathcal{E}_r + \frac{2c_T}{1 - c_T}\mathcal{E}_\mathcal{P}\right), \forall \mathbf{s} \in \mathcal{S}. \tag{14}$$

where $\mathcal{E}_\phi := \|\widehat{d}_{\pi,\phi} - \widehat{d}_\pi\|_\infty$ is the metric learning error, $\mathcal{E}_r := \|\widehat{r}^\pi - r^\pi\|_\infty$ is the reward approximation error, and $\mathcal{E}_\mathcal{P} := \sup_{\mathbf{s} \in \mathcal{S}} W_1(d_\pi)(\mathcal{P}^\pi(\cdot|\mathbf{s}), \widehat{\mathcal{P}}^\pi(\cdot|\mathbf{s}))$ is the state transition model error.

Proof. Following the proof of Lemma 8, we have that [8]

$$
\begin{aligned}
(1-\gamma)|V^{\pi}(\mathbf{s}) - \widetilde{V}^{\pi}(\Phi(\mathbf{s}))| &\leq \frac{c_R^{-1}}{\xi(\Phi(\mathbf{s}))} \int\limits_{\mathbf{z}\in\Phi(\mathbf{s})} d_{\pi}(\mathbf{s},\mathbf{z}) d\xi(\mathbf{z}) \\
&\leq \frac{c_R^{-1}}{\xi(\Phi(\mathbf{s}))} \int\limits_{\mathbf{z}\in\Phi(\mathbf{s})} \widehat{d}_{\pi,\phi}(\mathbf{s},\mathbf{z}) + \underbrace{|d_{\pi}(\mathbf{s},\mathbf{z}) - \widehat{d}_{\pi,\phi}(\mathbf{s},\mathbf{z})|_{\infty}}_{\mathcal{L}} d\xi(\mathbf{z}) \\
&\leq \frac{c_R^{-1}}{\xi(\Phi(\mathbf{s}))} \int\limits_{\mathbf{z}\in\Phi(\mathbf{s})} 2\widehat{\epsilon} + \mathcal{L}\, d\xi(\mathbf{z}) \\
&= c_R^{-1}(2\widehat{\epsilon} + \mathcal{L}) \qquad (\dagger) \\
&\leq \frac{1}{c_R}\left(2\widehat{\epsilon} + \mathcal{E}_{\phi} + \frac{2c_R}{1-c_T}\mathcal{E}_r + \frac{2c_T}{1-c_T}\mathcal{E}_{\mathcal{P}}\right)
\end{aligned}
$$

where the last line used Lemma 10.

■

Rather than bound the ground-truth on-policy bisimulation distance d_{π}, we have instead bound the *estimated* distance $\widehat{d}_{\pi,\phi}$, using both approximate predictive dynamics and error in the metric learning process.

Notice that as we shrink the size of the equivalence classes in the partition, so $\widehat{\epsilon} \to 0$, we get $\Phi \to \phi$. This tells us that information about the value of a state is preserved by the encoder ϕ, as long as the error in the forward dynamics model and metric learning process is small. Further, in the low error case, if the vectors $\phi(\mathbf{s}_i)$ and $\phi(\mathbf{s}_j)$ are close, we are guaranteed they have similar value under π, with the difference growing only linearly with the error.

Corollary 3 (VFA bound in terms of model error for arbitrary c_T)**.** *Consider the same conditions and definitions as Thm. 4, except $c_T \in [0,1)$. Let $\overline{\gamma} = \min(c_T, \gamma)$:*

$$
|V^{\pi}(\mathbf{s}) - \widetilde{V}^{\pi}(\Phi(\mathbf{s}))| \leq \frac{1}{c_R(1-\overline{\gamma})}\left(2\widehat{\epsilon} + \mathcal{E}_{\phi} + \frac{2c_R}{1-c_T}\mathcal{E}_r + \frac{2c_T}{1-c_T}\mathcal{E}_{\mathcal{P}}\right) + \frac{2(\gamma-\overline{\gamma})}{(1-\gamma)(1-c_T)},\ \forall \mathbf{s} \in \mathcal{S}. \tag{44}
$$

Proof. The proof follows similarly to the proof of Thm. 2. Suppose $c_T < \gamma$:

$$
\begin{aligned}
|V^{\pi}(\mathbf{s}) - \widetilde{V}^{\pi}(\Phi(\mathbf{s}))| &= |V^{\pi}(\mathbf{s}) - \widetilde{V}^{\pi}(\Phi(\mathbf{s})) + \widetilde{V}^{\pi}_{c_T}(\Phi(\mathbf{s})) - \widetilde{V}^{\pi}_{c_T}(\Phi(\mathbf{s})) + V^{\pi}_{c_T}(\mathbf{s}) - V^{\pi}_{c_T}(\mathbf{s})| \\
&\leq |V^{\pi}_{c_T}(\mathbf{s}) - \widetilde{V}^{\pi}_{c_T}(\Phi(\mathbf{s}))| \ + \ |V^{\pi}(\mathbf{s}) - V^{\pi}_{c_T}(\mathbf{s})| \ + \ |\widetilde{V}^{\pi}(\Phi(\mathbf{s})) - \widetilde{V}^{\pi}_{c_T}(\Phi(\mathbf{s}))| \\
&\leq \frac{1}{c_R(1-c_T)}\left(2\widehat{\epsilon} + \mathcal{E}_{\phi} + \frac{2c_R}{1-c_T}\mathcal{E}_r + \frac{2c_T}{1-c_T}\mathcal{E}_{\mathcal{P}}\right) \\
&\quad + \ |V^{\pi}(\mathbf{s}) - V^{\pi}_{c_T}(\mathbf{s})| \ + \ |\widetilde{V}^{\pi}(\Phi(\mathbf{s})) - \widetilde{V}^{\pi}_{c_T}(\Phi(\mathbf{s}))| \\
&\leq \frac{1}{c_R(1-c_T)}\left(2\widehat{\epsilon} + \mathcal{E}_{\phi} + \frac{2c_R}{1-c_T}\mathcal{E}_r + \frac{2c_T}{1-c_T}\mathcal{E}_{\mathcal{P}}\right) + \frac{2(\gamma-c_T)}{(1-\gamma)(1-c_T)},
\end{aligned}
$$

where the second and third inequalities are due to Thm. 4 and Lemma 7 respectively. For $\gamma \leq c_T$, we recover Thm. 4, hence, for all $c_T \in [0,1)$ and $\mathbf{s} \in \mathcal{S}$:

$$
|V^{\pi}(\mathbf{s}) - \widetilde{V}^{\pi}(\Phi(\mathbf{s}))| \ \leq \ \frac{1}{c_R(1-\min(c_T,\gamma))}\left(2\widehat{\epsilon} + \mathcal{E}_{\phi} + \frac{2c_R}{1-c_T}\mathcal{E}_r + \frac{2c_T}{1-c_T}\mathcal{E}_{\mathcal{P}}\right) + \frac{2(\gamma-\min(c_T,\gamma))}{(1-\gamma)(1-c_T)}.
$$

■

[8] Notice that using the Equation (†) allows us to recover the value bound with model error from [53]: $|V^{\pi}(\mathbf{s}) - \widetilde{V}^{\pi}(\Phi(\mathbf{s}))| \leq \frac{2\widehat{\varepsilon}+\mathcal{L}}{(1-\gamma)(1-c)}$, where $c_R = 1-c$.

E Experimental Details and Additional Results

E.1 OpenAI Gym

E.1.1 Sparse Noisy Cartpole Additional Details

First, we consider modifications to the Cartpole-v0 task. In the standard version, the agent receives a constant reward of $+1$ at each time-step for as long as the cart-pole system is upright between $[-\theta_{\text{term}}, \theta_{\text{term}}]$ degrees. If the pole falls below this range, the episode terminates. To make the task more challenging, we introduce a second parameter $\theta_{\text{rew}} \ll \theta_{\text{term}}$, such that the agent receives a reward only if the pole is between $[-\theta_{\text{rew}}, \theta_{\text{rew}}]$ degrees. We refer to this task as Sparse Cartpole and set $\theta_{\text{term}} = 12^\circ$ with $\theta_{\text{rew}} = 0.01\,\theta_{\text{term}}$. Additionally, we consider a noisy version of Sparse Cartpole to mimic distractors. In particular, we concatenate an $N_m \dim(\mathcal{S})$-dimensional vector sampled from an isotropic Gaussian to the state vector. The resulting task is referred to as Noisy Sparse Cartpole. Thus, the encoder ϕ must learn to embed functionally similar states in close proximity, despite the distractions, and maintain a well-behaved embedding, despite reward sparsity. While we provide sparse reward signals at training time to make the learning problem more difficult, we evaluate the resulting models in the standard environment based on θ_{term}, since this provides a lower variance return signal and still allows us determine whether the task has been solved.

E.1.2 Noisy Mountain Car Additional Details

We next tested on the Noisy Mountain Car task [32], implemented as MountainCarContinuous-v0 in the OpenAI Gym [7], and modified to concatenate $N_m \dim(\mathcal{S})$ noise dimensions to the observed state to simulate distraction. Briefly, the agent controls a car that should reach the top of a mountain, but has an engine of insufficient power to attain it directly. It must therefore learn a sequence of actions that build enough momentum to complete the task. The reward signal is highly uninformative: a small negative reward is given at every step, unless the task is solved, in which case a large positive reward is given. For $N_m > 0$, this task has both noisy distractors and high sparsity, making it rather challenging.

Note that only methods with intrinsic reward were able to solve the task (see Fig. 2 and Fig. 5), and that DBC without normalization was also unable to complete it. We remark that all methods rely on the maximum policy entropy RL formulation, which is known to improve exploration [23, 24, 55]; nevertheless, our results suggest that curiosity-driven exploration, induced by intrinsic rewards based on predictive error, is at least complementary to such techniques.

E.1.3 Sparse Pendulum Task Details

For the Pendulum-v0 task, we implement similar modifications to those in SparseCartpole. The standard Pendulum task starts with a pole in downright position, and provides negative rewards proportional to $|\theta_{\text{pend}}|$, degrees away from upright position. Our SparsePendulum instead provides a reward of $+1$ only when the pendulum is between $[-\theta_{\text{rew}}, \theta_{\text{rew}}]$ degrees, where θ_{rew}, and does not provide a reward otherwise. NoisySparsePendulum similarly concatenates to state vectors a noise vector of N_m times the original dimensionality. Note that we evaluate with an environment with $\theta_{\text{rew}} = 1^\circ$, to reduce variance in the evaluation reward.

Results (see Fig. 6) show that our method performs comparably to DBC. In the presence of high distraction ($N_m = 10$), DBC-based approaches can do much better than SAC, which is no longer able to solve the task.

E.1.4 Plots with Additional Distraction

In Fig. 5, we show results for Sparse Cartpole (left) and Mountain Car (right), with an even higher level of distraction ($N_m = 3$). On Sparse Cartpole, only DBC-normed-IR and DBC-normed-IR-ID were able to solve the task, with the latter being slightly more stable, while DBC-normed performed significantly better than any methods without normalization. On Mountain Car, all methods struggle to solve the task at this level of distraction; however, DBC-normed-IR-ID performs significantly better than the others, showing the utility of the inverse dynamics regularization.

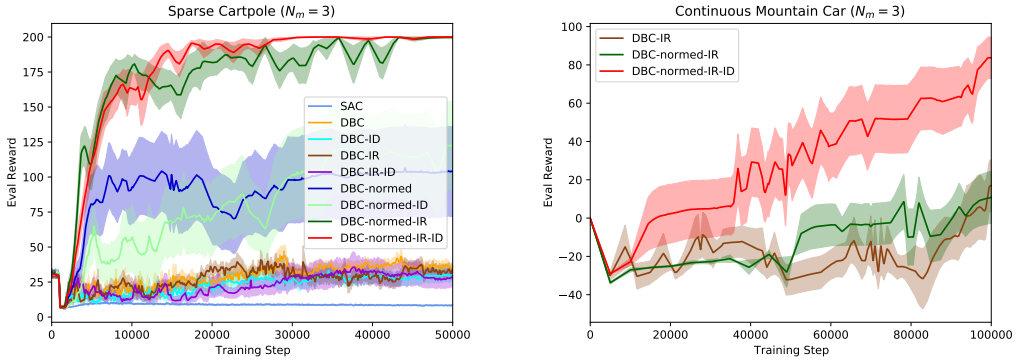

Figure 5: Additional plots on Sparse Noisy Cartpole (left) and Mountain Car (right) at high distraction ($N_m = 3$). Shaded bars are standard errors over 10 seeds.

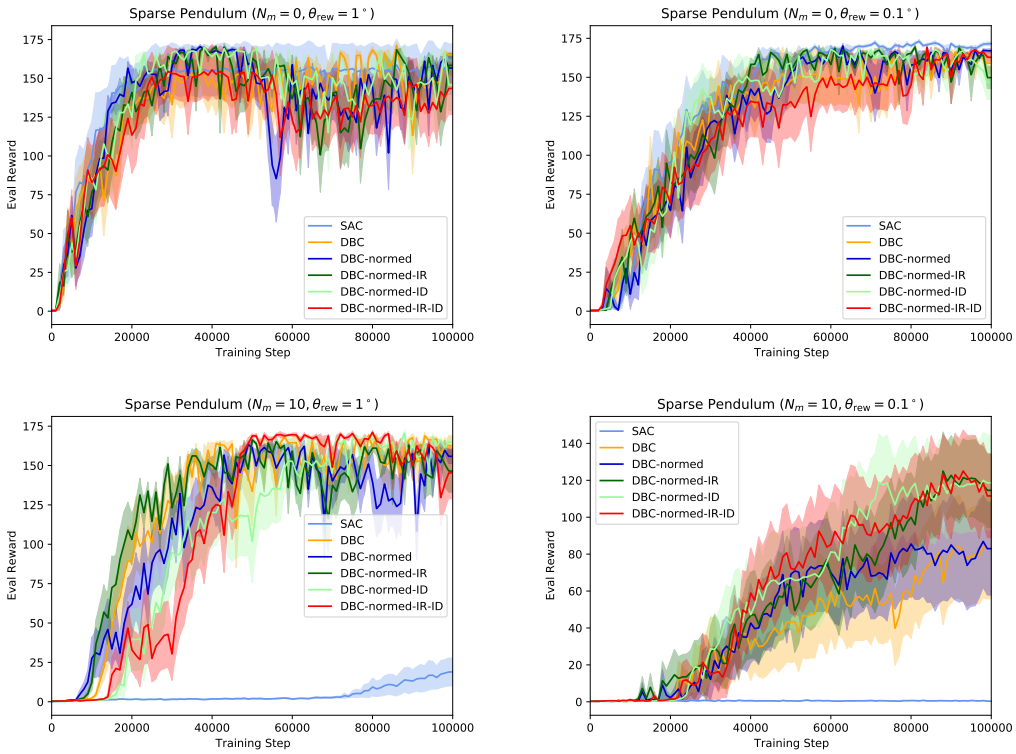

Figure 6: Results on the Sparse Pendulum task with differing levels of distraction and sparsity (shaded bars are standard errors over 10 seeds).

E.1.5 Hyper-parameters and Architectures for OpenAI Gym Tasks

All models use the code generously released alongside the DBC paper [53] (CC-BY-NC 4.0 licensed), with default hyper-parameters and architectures according to the released code, unless otherwise specified. DBC and our modifications of it are all built on top of the Soft Actor-Critic implementation therein. Actor, critic, and encoder learning rates were all 0.001, using the Adam optimizer. A replay buffer of 50K was used, with a batch size for all training steps of 512. All encoders were implemented as multilayer perceptrons (MLPs) with four layers, except for the SparsePendulum task where two layers were used for all normalized approaches (unnormalized DBC still performed better with four). Encoder feature dimensionality was set to $\dim(\phi(\mathbf{s})) = 50$. When inverse dynamics prediction was needed, we implemented g_I as an MLP with two hidden layers (of size 256 and 128), with ELU activations. Note that for the predictive approximate transition model $\widehat{\mathcal{P}}$, we use a deterministic predictor. Following the implementation of [53], distances in reward and predicted encoded state, for the bisimulation metric, were computed with a Huber loss, defining the value of q as a function of the distance. However, embedding normalization was computed with the L_2 norm. A discount of $\gamma = 0.99$ was always used. In all tasks, rewards are bounded as $R \in [0, 1]$.

We remark that all evaluation rewards in plots (per seed) are computed as averages over 10 episodes.

We used a maximum intrinsic reward clamping value of $R_{\max,I} = 0.1$ for all Gym tasks, and set η_r and η_d per task as in Table 1 below. Hyper-parameters not left at default ($R_{\max,I}$, η_r, η_d, and number of encoder layers and latent dimensionality) were set by searching over a small, manually defined set of values.

E.2 DeepMind Control Suite

For the DeepMind Control Suite [45], our encoder model architecture is identical to the open-source code repository released by [53]. Namely, a 3x3 convolutional layer with stride 2 is followed by another 3x3 convolution with stride 1 (both with 32 channels), before a fully-connected layer with 50-dimensional output and layer normalization. `ReLU` activations are used between neural layers. When an inverse dynamics model is used, we use the same architecture as those used for the OpenAI experiments (described in Appendix E.1.5), namely, a two-layer MLP. Differently from [53], to speed up training, we run 16 environments in parallel, all of which add experience tuples to a shared replay buffer. After every 16 environment steps (i.e., each parallel step), we apply 2 gradient updates. Our hyperparameters for inverse dynamics and intrinsic motivation are given in the last column of Table 1, and were selected by searching over a small, manually defined set of values.

	Cartpole	Pendulum	Mountain Car	DMC
η_r	2	0.1	20	1
η_d	1	0.1	20	10

Table 1: Hyper-parameters used per task for intrinsic reward and inverse dynamics.

E.3 Computational Resources and Timing

All training and evaluation was done on a small set of NVIDIA GPUs (GTX 1080 TI, Titan X, or RTX 2080 TI), less than 10 in total and shared with other users.

OpenAI Gym tasks were run with multiple seeds per GPU (up to the GPU memory limit) during training. In this parallel training context, which allowed us to complete a Gym task for all methods and seeds (for a single distraction level) within roughly a day on 2-4 GPUs, we obtain the following approximate timings (in seconds per training iteration). Cartpole: $\sim$0.1 for DBC-based methods and $\sim$0.07 for SAC. Pendulum: $\sim$0.09 for DBC-based methods ($\sim$0.1 with IR+ID present) and $\sim$0.06 for SAC. Mountain Car: $\sim$0.07 for DBC-based methods and $\sim$0.04 for SAC.

For the Deepmind Control (DMC) tasks, experiments were performed on 4 GPUs over the course of a week. Our 16-process parallelization for running MuJoCo [47] simulations greatly sped up training, resulting in approximately 0.03, 0.05 and 0.07 seconds per environment step (with 2 gradient updates for every 16 environment step) for SAC, DBC and DBC+IR+ID respectively.

References

[1] Joshua Achiam and Shankar Sastry. Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv preprint arXiv:1703.01732*, 2017.

[2] Rishabh Agarwal, Marlos C. Machado, Pablo Samuel Castro, and Marc G. Bellemare. Contrastive behavioral similarity embeddings for generalization in reinforcement learning. In *International Conference on Learning Representations*, 2021.

[3] Pulkit Agrawal, Ashvin Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Learning to poke by poking: experiential learning of intuitive physics. *arXiv preprint arXiv:1606.07419*, 2016.

[4] Arthur Aubret, Laetitia Matignon, and Salima Hassas. A survey on intrinsic motivation in reinforcement learning. *arXiv preprint arXiv:1908.06976*, 2019.

[5] Marc G Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *arXiv preprint arXiv:1606.01868*, 2016.

[6] Ondrej Biza and Robert Platt. Online abstraction with MDP homomorphisms for deep learning. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2019.

[7] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI gym. *arXiv preprint arXiv:1606.01540*, 2016.

[8] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.

[9] Pablo Castro and Doina Precup. Using bisimulation for policy transfer in MDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, 2010.

[10] Pablo Samuel Castro. Scalable methods for computing state similarity in deterministic Markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10069–10076, 2020.

[11] Philippe Clement and Wolfgang Desch. An elementary proof of the triangle inequality for the Wasserstein metric. *Proceedings of the American Mathematical Society*, 136(1):333–339, 2008.

[12] Dane Corneil, Wulfram Gerstner, and Johanni Brea. Efficient model-based deep reinforcement learning with variational state tabulation. In *International Conference on Machine Learning*, pages 1049–1058. PMLR, 2018.

[13] Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, page 162–169, Arlington, Virginia, USA, 2004. AUAI Press.

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

[15] Norm Ferns and Doina Precup. Bisimulation metrics are optimal value functions. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, page 210–219, Arlington, Virginia, USA, 2014. AUAI Press.

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

[17] Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning*, pages 2170–2179. PMLR, 2019.

[18] Dibya Ghosh, Abhishek Gupta, and Sergey Levine. Learning actionable representations with goal-conditioned policies. *arXiv preprint arXiv:1811.07819*, 2018.

[19] Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1):163–223, 2003. Planning with Uncertainty and Incomplete Information.

[20] Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. *arXiv preprint arXiv:1611.07507*, 2016.

[21] Christopher Grimm, Andre Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[22] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[23] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870. PMLR, 2018.

[24] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

[25] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pages 2555–2565. PMLR, 2019.

[26] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[27] Thanard Kurutach, Aviv Tamar, Ge Yang, Stuart Russell, and Pieter Abbeel. Learning plannable representations with causal InfoGAN. *arXiv preprint arXiv:1807.09341*, 2018.

[28] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pages 5639–5650. PMLR, 2020.

[29] Charline Le Lan, Marc G Bellemare, and Pablo Samuel Castro. Metrics and continuity in reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8261–8269, 2021.

[30] Alex X Lee, Anusha Nagabandi, Pieter Abbeel, and Sergey Levine. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. *arXiv preprint arXiv:1907.00953*, 2019.

[31] Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for MDPs. In *In Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics*, pages 531–539, 2006.

[32] Andrew William Moore. Efficient memory-based learning for robot control. 1990.

[33] Ashvin Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. *arXiv preprint arXiv:1807.04742*, 2018.

[34] Masashi Okada and Tadahiro Taniguchi. Dreaming: Model-based reinforcement learning by latent imagination without reconstruction. *arXiv preprint arXiv:2007.14535*, 2020.

[35] Ingram Olkin and Friedrich Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

[36] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, pages 2778–2787. PMLR, 2017.

[37] Marek Petrik and Bruno Scherrer. Biasing approximate dynamic programming with a lower discount factor. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2009.

[38] Balaraman Ravindran. *An algebraic approach to abstraction in reinforcement learning*. PhD thesis, 2004.

[39] Balaraman Ravindran and Andrew G Barto. Approximate homomorphisms: A framework for non-exact minimization in Markov decision processes. 2004.

[40] Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment–an introduction. In *Guided Self-Organization: Inception*, pages 67–114. Springer, 2014.

[41] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[42] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[43] Bradly C Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*, 2015.

[44] Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling representation learning from reinforcement learning. *arXiv preprint arXiv:2009.08319*, 2020.

[45] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. DeepMind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

[46] Jonathan Taylor, Doina Precup, and Prakash Panagaden. Bounding performance loss in approximate MDP homomorphisms. *Advances in Neural Information Processing Systems*, 21:1649–1656, 2008.

[47] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[48] Elise van der Pol, Thomas Kipf, Frans A Oliehoek, and Max Welling. Plannable approximations to MDP homomorphisms: Equivariance under actions. *arXiv preprint arXiv:2002.11963*, 2020.

[49] Elise van der Pol, Daniel Worrall, Herke van Hoof, Frans Oliehoek, and Max Welling. MDP homomorphic networks: Group symmetries in reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4199–4210. Curran Associates, Inc., 2020.

[50] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[51] Manuel Watter, Jost Tobias Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: a locally linear latent dynamics model for control from raw images. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015.

[52] Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2859–2867, 2017.

[53] Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021.

[54] Marvin Zhang, Sharad Vikram, Laura Smith, Pieter Abbeel, Matthew Johnson, and Sergey Levine. Solar: Deep structured representations for model-based reinforcement learning. In *International Conference on Machine Learning*, pages 7444–7453. PMLR, 2019.

[55] Brian D Ziebart. Modeling purposeful adaptive behavior with the principle of maximum causal entropy. 2010.