ACCV2022 #585

A Proofs

We first prove the inequality of Theorem 1.

Theorem 1 restated. *(Divergence on belief state representation) Consider a POMDP $\mathcal{M}$, and let b_t be a latent representation of the belief state, such that $P\left(s_t|x_{\leq t}, a_{<t}\right) = P\left(s_t|b_t\right)$. Let the policy $a_t \sim \pi(\cdot|b_t)$, so that $P(s_t|b_t, a_t) = P(s_t|b_t)$ Let D_f be a generic f-divergence. Then the following inequalities hold:*

$$D_f\left(\rho_{\mathcal{M}}^{\pi}(s,a) \parallel \rho_{\mathcal{M}}^{\pi_E}(s,a)\right) \leq D_f\left(\rho_{\mathcal{M}}^{\pi}(b,a) \parallel \rho_{\mathcal{M}}^{\pi_E}(b,a)\right). \tag{16}$$

Proof. With the condition $P\left(s_t|x_{\leq t}, a_{<t}\right) = P\left(s_t|b_t\right)$, there is:

$$\begin{aligned}
D_f\left(\rho_{\mathcal{M}}^{\pi}(b,a) \parallel \rho_{\mathcal{M}}^{\pi_E}(b,a)\right) &= \mathbb{E}_{b,a\sim\rho_{\mathcal{M}}^{\pi_E}(b,a)}\left[f\left(\frac{\rho_{\mathcal{M}}^{\pi}(b,a)}{\rho_{\mathcal{M}}^{\pi_E}(b,a)}\right)\right] &(17)\\
&= \mathbb{E}_{s,b,a\sim\rho_{\mathcal{M}}^{\pi_E}(s,b,a)}\left[f\left(\frac{\rho_{\mathcal{M}}^{\pi}(s,b,a)}{\rho_{\mathcal{M}}^{\pi_E}(s,b,a)}\right)\right] &(18)\\
&= \mathbb{E}_{s,a\sim\rho_{\mathcal{M}}^{\pi_E}(s,a)}\left[\mathbb{E}_{b\sim\rho_{\mathcal{M}}^{\pi_E}(b|s,a)}\left[f\left(\frac{\rho_{\mathcal{M}}^{\pi}(s,b,a)}{\rho_{\mathcal{M}}^{\pi_E}(s,b,a)}\right)\right]\right] &(19)\\
&\geq \mathbb{E}_{s,a\sim\rho_{\mathcal{M}}^{\pi_E}(s,a)}\left[f\left(\mathbb{E}_{b\sim\rho_{\mathcal{M}}^{\pi_E}(b|s,a)}\left[\frac{\rho_{\mathcal{M}}^{\pi}(s,b,a)}{\rho_{\mathcal{M}}^{\pi_E}(s,b,a)}\right]\right)\right] &(20)\\
&= \mathbb{E}_{s,a\sim\rho_{\mathcal{M}}^{\pi_E}(s,a)}\left[f\left(\mathbb{E}_{b\sim\rho_{\mathcal{M}}^{\pi_E}(b|s,a)}\left[\frac{\rho_{\mathcal{M}}^{\pi}(s,a)\rho_{\mathcal{M}}^{\pi}(b|s,a)}{\rho_{\mathcal{M}}^{\pi_E}(s,a)\rho_{\mathcal{M}}^{\pi_E}(b|s,a)}\right]\right)\right] &(21)\\
&= \mathbb{E}_{s,a\sim\rho_{\mathcal{M}}^{\pi_E}(s,a)}\left[f\left(\mathbb{E}_{b\sim\rho_{\mathcal{M}}^{\pi}(b|s,a)}\left[\frac{\rho_{\mathcal{M}}^{\pi}(s,a)}{\rho_{\mathcal{M}}^{\pi_E}(s,a)}\right]\right)\right] &(22)\\
&= \mathbb{E}_{s,a\sim\rho_{\mathcal{M}}^{\pi_E}(s,a)}\left[f\left(\frac{\rho_{\mathcal{M}}^{\pi}(s,a)}{\rho_{\mathcal{M}}^{\pi_E}(s,a)}\right)\right] &(23)\\
&= D_f\left(\rho_{\mathcal{M}}^{\pi}(s,a) \parallel \rho_{\mathcal{M}}^{\pi_E}(s,a)\right). &(24)
\end{aligned}$$

Lemma 1. *(Simultaneous policy and model deviation) Let $\hat{M}$ be approximate MDP constructed using approximate dynamics model $\hat{\mathcal{T}}$. $\hat{M}$ differs $\mathcal{M}$ only in transition dynamics - $\hat{\mathcal{T}}$ and $\mathcal{T}$. Let $R_{max} = max_{(s,a)\mathcal{R}(s,a)}$ be the upper-bound of rewards. Let $\mu_{\mathcal{M}}^{\pi}(s,a) = (1-\gamma)\rho_{\mathcal{M}}^{\pi}(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\left(\gamma^t P\left(s_t = s, a_t = a|\pi\right)\right)$ be the normalized γ-discounted state-visitation distribution of a policy π in MDP $\mathcal{M}$. If the dynamics of both $\mathcal{M}$ and $\hat{\mathcal{M}}$ are such that*

$$\mathbb{E}_{(s,a)\sim\mu_{\mathcal{M}}^{\pi,t}}\left[D_{TV}(\mathcal{T}(\cdot|s,a), \hat{\mathcal{T}(\cdot|}s,a))\right] \leq \epsilon \quad \forall t. \tag{25}$$

then, the performance difference between policy π and expert policy π_E can be bounded as:

$$\left|J(\pi,\mathcal{M}) - J(\pi_E,\mathcal{M})\right| \leq \frac{R_{max}}{1-\gamma}D_{TV}(\mu_{\mathcal{M}}^{\pi}, \mu_{\hat{\mathcal{M}}}^{\pi_E}) + \frac{\epsilon\gamma R_{max}}{(1-\gamma)^2} \tag{26}$$

Proof. Depending on the definition of $J(\pi, \mathcal{M})$ and the triangle inequality on D_{TV}, we have:

$$\begin{aligned}
\Big| J(\pi, \mathcal{M}) - J(\pi_E, \mathcal{M}) \Big| &= \left| \frac{1}{1-\gamma} \mathbb{E}_{\mu^{\pi}_{\mathcal{M}}} \left[\mathcal{R}(s,a) \right] - \frac{1}{1-\gamma} \mathbb{E}_{\mu^{\pi_E}_{\mathcal{M}}} \left[\mathcal{R}(s,a) \right] \right| && (27) \\
&\leq \frac{R_{max}}{1-\gamma} D_{TV}(\mu^{\pi}_{\mathcal{M}}, \mu^{\pi_E}_{\mathcal{M}}) && (28) \\
&\leq \frac{R_{max}}{1-\gamma} \left(D_{TV}(\mu^{\pi}_{\mathcal{M}}, \mu^{\pi}_{\hat{\mathcal{M}}}) + D_{TV}(\mu^{\pi}_{\hat{\mathcal{M}}}, \mu^{\pi_E}_{\mathcal{M}}) \right) && (29) \\
&\leq \frac{R_{max}}{1-\gamma} D_{TV}(\mu^{\pi}_{\mathcal{M}}, \mu^{\pi_E}_{\hat{\mathcal{M}}}) + \frac{\epsilon \gamma R_{max}}{(1-\gamma)^2} && (30)
\end{aligned}$$

Thus, the divergence minimization in model-based IL guarantees the suboptimality with a bias that is proportional to the model error. It allows us to collect on-policy rollouts to update policies without interaction with the environment.

B Environments

To evaluate the performance of IL algorithms, we chose three classic control tasks from the Distracting DeepMind Control Suite. In each task, IL agents are provided with RGB images whose height and width are set to 64 pixels. To evaluate the generalization further, the dynamic backgrounds of testing environments are chosen at random from the DAVIS 2017 dataset [58], as illustrated in Fig. 5. When the start or end frame is reached, we reverse the order of playing videos. In this way, the background changes are constantly continuous without abrupt alterations. As mentioned before, we use DrQ to train experts to generate demonstrations. The action repeat for the three tasks is set to 2, so that the episode length is 500. We construct the observation input as a 3-stack of consecutive frames for experts. As a result, experts make decisions depending on the observation with $64 \times 64 \times 3$ dimensions.

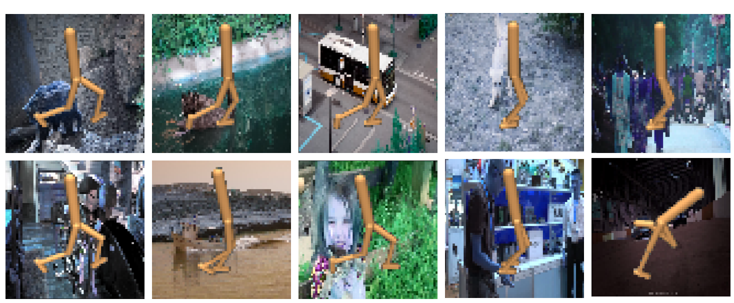

Fig. 5. Snapshots of testing environments used for evaluating generalization. The dynamic backgrounds are randomly selected from scenes of 10 videos of the DAVIS 2017 dataset.

C Hyperparameters

We provide a list of hyperparameters in Table 1.

Table 1. Hyperparameters of IMAIL in experiments

Hyperparameters	Value
Environment parameters	
Image size	$64 \times 64 \times 3$
Action repeat	2
Exploration steps	1000
Common parameters	
Batch size	64
Optimizer	Adam
Discount	0.99
Invariant encoder feature dim	256
Noise encoder feature dim	30
Belief encoder feature dim	30
Expert demonstrations length	50
Rollout length	15
Dynamics model learning rate	6×10^{-4}
Mutual information learning rate	6×10^{-4}
Statistics network parameters	
MLP network	$2 \times \{256\text{-}FC, Tanh\}$
Learning rate	1×10^{-3}
SAC parameters	
Action MLP network	$4 \times \{256\text{-}FC, Elu\}$
Value MLP network	$3 \times \{256\text{-}FC, Elu\}$
Actor learning rate	8×10^{-5}
Value learning rate	8×10^{-5}
Discriminator parameters	
MLP network	$2 \times \{256\text{-}FC, Elu\}$
Learning rate	8×10^{-5}