Supplementary Material: Learning Belief Representations for Imitation Learning in POMDPs

7.1 Proof of Inequalities in Section 3.2

We first prove the inequality connecting D_{JS} between the state-visitation distribution and belief-visitation distribution of the agent and the expert:

$$D_{JS}[\rho_\pi(s) \mid\mid \rho_E(s)] \leq D_{JS}[\rho_\pi(b) \mid\mid \rho_E(b)]$$

Proof. The proof is a simple application of the data-processing inequality for f-divergences (Ali & Silvey, 1966), of which D_{JS} is a type.

We denote the filtering posterior distribution over states, given the belief, by $p(s|b)$. Note that $p(s|b)$ is characterized by the environment, and does not depend on the policy (agent or expert). The posterior over belief, given the state, however, is policy-dependent and obtained using Bayes rule as: $p_\pi(b|s) = \frac{p(s|b)\rho_\pi(b)}{\rho_\pi(s)}$. Also, $\rho_\pi(s,b) = \rho_\pi(s)p_\pi(b|s) = \rho_\pi(b)p(s|b)$. Analogously definitions exist for expert E.

We write $D_{JS}[\rho_\pi(b) \mid\mid \rho_E(b)]$ in terms of the template used for f-divergences. Let $f : (0,\infty) \mapsto \mathbb{R}$ be the following convex function with the property $f(1) = 0$: $f(u) = -(u+1)\log\frac{1+u}{2} + u\log u$. Then,

$$\begin{aligned}
&D_{JS}[\rho_\pi(b) \mid\mid \rho_E(b)] \\
&\qquad = \mathbb{E}_{b\sim\rho_E(b)}\Big[f(\frac{\rho_\pi(b)}{\rho_E(b)})\Big] \\
&\qquad = \mathbb{E}_{s,b\sim\rho_E(s,b)}\Big[f(\frac{\rho_\pi(s,b)}{\rho_E(s,b)})\Big] \\
&\qquad = \mathbb{E}_{s\sim\rho_E(s)}\Big[\mathbb{E}_{b\sim\rho_E(b|s)}f(\frac{\rho_\pi(s,b)}{\rho_E(s,b)})\Big] \\
\textit{(Jensen's)}\ &\qquad \geq \mathbb{E}_{s\sim\rho_E(s)}\Big[f\Big(\mathbb{E}_{b\sim\rho_E(b|s)}\frac{\rho_\pi(s,b)}{\rho_E(s,b)}\Big)\Big] \\
&\qquad = \mathbb{E}_{s\sim\rho_E(s)}\Big[f\Big(\mathbb{E}_{b\sim\rho_\pi(b|s)}\frac{\rho_\pi(s,b)\rho_E(b|s)}{\rho_E(s,b)\rho_\pi(b|s)}\Big)\Big] \\
&\qquad = \mathbb{E}_{s\sim\rho_E(s)}\Big[f\Big(\mathbb{E}_{b\sim\rho_\pi(b|s)}\frac{\rho_\pi(s)}{\rho_E(s)}\Big)\Big] \\
&\qquad = \mathbb{E}_{s\sim\rho_E(s)}\Big[f(\frac{\rho_\pi(s)}{\rho_E(s)})\Big] \\
&\qquad = D_{JS}[\rho_\pi(s) \mid\mid \rho_E(s)]
\end{aligned}$$

□

Similarity, we can prove the inequality connecting D_{JS} between belief-visitation distribution and belief-action-visitation distribution of the agent and the expert:

$$D_{JS}[\rho_\pi(b) \mid\mid \rho_E(b)] \leq D_{JS}[\rho_\pi(b,a) \mid\mid \rho_E(b,a)]$$

Proof. Replace $s \mapsto b'$ and $b \mapsto (b,a)$ in the previous proof. The only *required* condition for that result to hold is the non-dependence of the distribution $p(s|b)$ on the policy. Therefore, if it holds that $p(b'|b,a)$ is independent of the policy, then we have,

$$D_{JS}[\rho_\pi(b') \mid\mid \rho_E(b')] \leq D_{JS}[\rho_\pi(b,a) \mid\mid \rho_E(b,a)]$$

The independence holds under the trivial case of a deterministic mapping $b'{=}b$. This gives us the desired inequality. □

7.2 MDP and POMDP Sensors

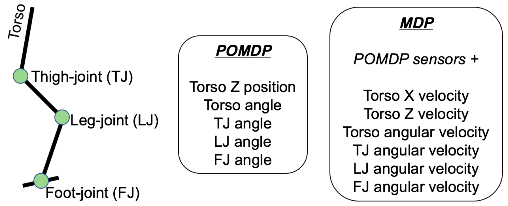

Figure 5: Comparison of sensor information available to the agent in the MDP (original) and the POMDP (modified) settings for Hopper-v2 from the Gym MuJoCo suite.

Description of the sensor measurements given to the agent in the MDP and POMDP environments is provided in Table 3. As an example, for the Hopper agent composed of 4 links connected via actuated joints (Figure 5), the MDP space is 11-dimensional, which includes 6 velocity sensors and 5 position sensors, whereas the POMDP space is 5-dimensional, comprising of 5 position sensors. Amongst sensor categories, *velocity* includes translation and angular velocities of the torso, and also the velocities for all the joints; *position* includes torso position and orientation (quaternion), and the angle of the joints. The sensors in the MDP column marked in **bold** are removed in the POMDP setting.

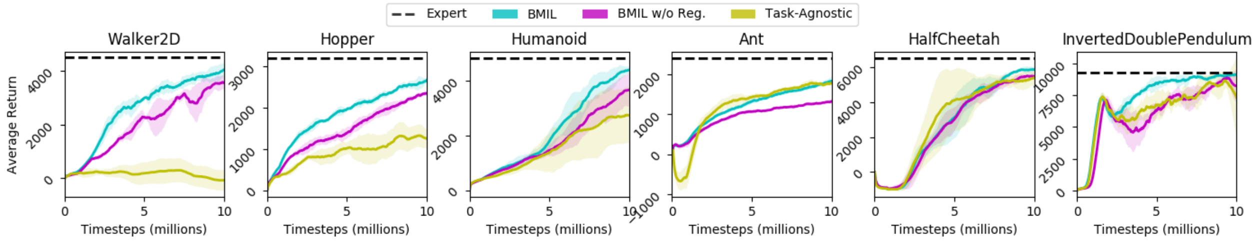

Figure 3: Mean episode-returns vs. timesteps of environment interaction. BMIL is our proposed architecture (Figure 2); BMIL w/o Reg excludes the various regularization terms (Section 3.4) from this design; *Task-Agnostic* learns the belief module separately from the policy using a task-agnostic loss ($\mathcal{L}^{AR}$, Section 3.3). We plot the average and standard-deviation over 5 random seeds.

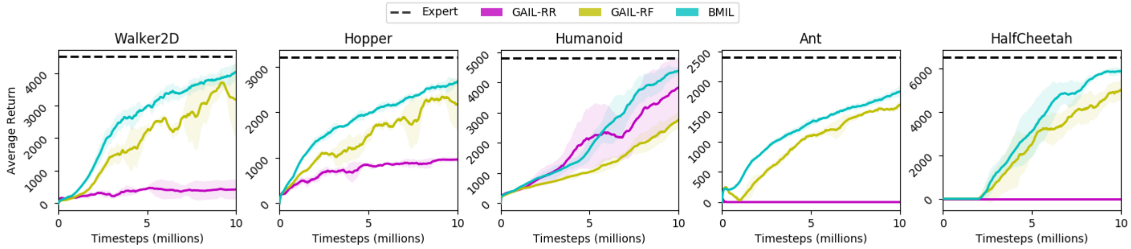

Figure 4: Mean episode-returns vs. timesteps of environment interaction. BMIL is our proposed architecture (Figure 2); GAIL-RF uses a recurrent policy and a feed-forward discriminator, while in GAIL-RR, both the policy and the discriminator are recurrent. We plot the average and standard-deviation over 5 random seeds.

Environment	MDP sensors ($s \in \mathcal{S}$)	POMDP sensors ($o \in \mathcal{O}$)
Hopper	($\lvert\mathcal{S}\rvert$=11) **velocity(6)** + position(5)	($\lvert\mathcal{O}\rvert$=5) position(5)
Half-Cheetah	($\lvert\mathcal{S}\rvert$=17) **velocity(9)** + position(8)	($\lvert\mathcal{O}\rvert$=8) position(8)
Walker2d	($\lvert\mathcal{S}\rvert$=17) **velocity(9)** + position(8)	($\lvert\mathcal{O}\rvert$=8) position(8)
Inv.DoublePend.	($\lvert\mathcal{S}\rvert$=11) **velocity(3)** + position(5) + actuator forces(3)	($\lvert\mathcal{O}\rvert$=8) position(5) + actuator forces(3)
Ant	($\lvert\mathcal{S}\rvert$=111) **velocity(14)** + position(13) + external forces(84)	($\lvert\mathcal{O}\rvert$=97) position(13) + external forces(84)
Humanoid	($\lvert\mathcal{S}\rvert$=376) **velocity(23) + center-of-mass based velocity(84)** + position(22) + center-of-mass based inertia(140) + actuator forces(23) + external forces(84)	($\lvert\mathcal{O}\rvert$=269) position(22) + center-of-mass based inertia(140) + actuator forces(23) + external forces(84)

Table 3: MDP and POMDP sensors (MuJoCo). The sensors in the MDP column marked in **bold** are removed in the POMDP setting.

7.3 Hyperparameters

Hyper-parameter	Value
Parameters for Convolution networks (encoding past & future action-sequences)	Layers=2, Stride=1, Padding=1, Kernel_size=3, Num_filters = {5,5}
Belief Regularization Coefficients	λ_1=λ_2=λ_3=0.2
Rollout length (c) in Algorithm 1	5
Size of expert demonstrations $\mathcal{M}_E$	50 (trajectories)
Size of replay buffer $\mathcal{R}$	1000 (trajectories)
Optimizer, Learning Rate	RMSProp, 3e-4 (linear-decay)
γ, λ (GAE)	0.99, 0.95

7.4 Ablation Plots

Learning curves for our ablation on the components of the belief regularization loss, and the effect of multi-step predictions are presented in Figure 6 and Figure 7, respectively. Please see Section 5.3 for the analysis.

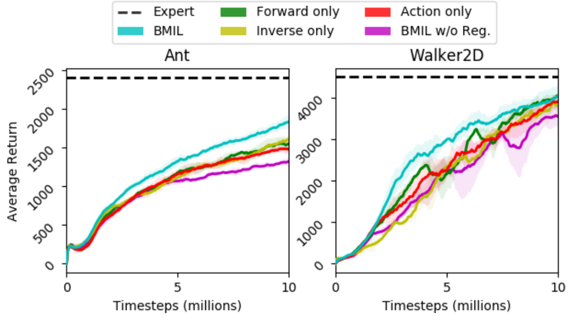

Figure 6: Ablation on components of belief regularization. Forward-, Inverse-, Action-only correspond to using $\mathcal{L}^f$, $\mathcal{L}^i$, $\mathcal{L}^a$, respectively, in isolation, without the other two.

7.5 Predictions in Encoding-space

In our approach, we regularize with single- and multi-step predictions in the space of *raw observations*. For many high-dimensional, complex spaces (e.g. visual inputs), it may be more efficient to operate in a lower-dimensional, compressed encoding-space, either pre-trained, or learnt online (Cuccu et al., 2019).

The encoder in our architecture (yellow box in Figure 2) pre-processes the raw observations before passing them to the RNN for temporal integration. We now evalu-

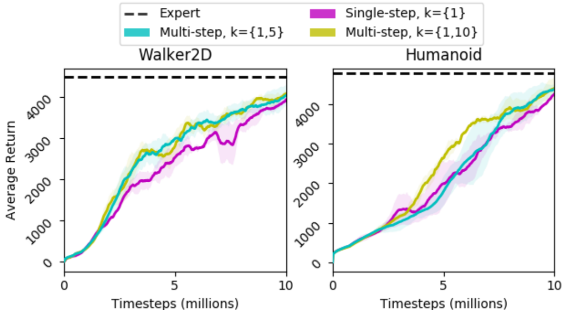

Figure 7: Ablation on hyperparameter k in the regularization terms. Multi-step design builds over single-step by adding predictions at different temporal offsets, k=5 and k=10.

ate BMIL with single- and multi-step predictions in the space of this encoder output. For instance, the forward regularization loss function is:

$$\mathcal{L}^f(\phi) = \mathbb{E}_{\mathcal{R}} || Enc(o_{t+k}) - g(b_t^\phi, a_{t:t+k-1}) ||_2^2$$

We do not pass the gradient through the target value $Enc(o_{t+k})$. The encoder is trained online as part of the belief module. Table 4 indicates that, for the tasks considered, BMIL performance is fairly similar when predicting in observation-space vs. encoding-space.

	BMIL: Predictions in observation-space	BMIL: Predictions in encoding-space
Invd.DoublePend.	9104 ± 134	8883 ± 448
Hopper	2665 ± 70	2700 ± 116
Ant	1832 ± 92	1784 ± 44
Walker	4038 ± 259	4043 ± 113
Humanoid	4382 ± 117	4322 ± 263
Half-cheetah	5860 ± 171	5912 ± 128

Table 4: Mean and std. of episode-returns, averaged over 5 random seeds, after 10M timesteps in POMDP MuJoCo.

7.6 Experiments in environment variants that accentuate partial observability

To test robustness of BMIL, we evaluate it on two new POMDP environment variants designed to make inferring the true state from given sensors more challenging. These new environments are:

- **Inv.DoublePend. from velocities only -** The partially observable *Inverted-Double-Pendulum* used in Section 5 removes the velocity sensors to achieve partial observability, and provides as sensors only the cart-position and sin/cos of link angles. In this new environment, we remove the previously shown sensors (cart-position and link angles), and instead provide only the velocity sensors (which were removed in our original environment). Note that the motivation is to exacerbate partial observability by restricting sensors such that inferring the true state is more challenging (i.e. it is easier to infer velocity from subsequent positions than to integrate position over time from only velocity information). Indeed, our experiments indicate this is a harder imitation learning scenario.
- **Walker from velocities only -** In the same spirit as above. We remove all position sensors and instead provide only the velocity sensors to the agent.

We compare BMIL to GAIL-RF (the strongest baseline).

	GAIL-RF	BMIL
Inv.DoublePend. (velocity only)	4988	6578
Walker (velocity only)	1539	4199

Table 5: Mean episode-returns, averaged over 5 runs with random seeds, after 10M timesteps.