A EXAMPLE FOR FORWARD AND BACKWARD K-OBSERVABILITY

Example (Refinement HMM) RHMMs [Stratos et al., 2013] can be considered as a special case of latent chain-CRFs (e.g., Figure 1a) with directed edges (e.g., all edges in Figure 1a point from left to right). Let a_t and x_t be vectors with one-hot representation, and z a new variable such that $z = a \times s$, where $\times$ stands for tensor product. Then, we can define the transition matrix for z as A_z, where $A_z[i, j]$ stands for the probability of transiting from $z_t = j$ to $z_{t+1} = i$. Additionally, we can define a new observation matrix C_z, where $C_z[i, j]$ stands for the probability of observing $x_t = i$ given $z_t = j$. With A_z and C_z, we reduce the RHMM model to a HMM model, where A_z and C_z are the transition and observation matrices, while z and x are, respectively, the latent state and the observation. The observability condition for HMM is well defined. Let us assume the HMM model is k observable with some constant k. This means that there exists a bijective map between $P(a_t, s_t \mid x_{1:t-1})$ and $P(x_{t:t+k-1} \mid x_{1:t-1})$. Let us define such map as M here. First of all, given $P(s_t, a_t \mid x_{1:t-1})$, it is straightforward to compute $P(a_{t:t+k}, x_{t:t+k-1} \mid x_{1:t-1})$ from the RHMM's graphical model itself. Secondly, given any distribution $P(a_{t:t+k}, x_{t:t+k-1} \mid x_{1:t-1})$ resulting from some $P(a_t, s_t \mid x_{1:t-1})$, to recover $P(a_t, s_t \mid x_{1:t-1})$, we can first marginalize $P(a_{t:t+k}, x_{t:t+k-1} \mid x_{1:t-1})$ over $a_{t:t+k}$ to get $P(x_{t:t+k-1} \mid x_{1:t-1})$, and then apply the HMM's bijective map M to get $P(z_t \mid x_{1:t-1})$, which is $P(a_t, s_t \mid x_{1:t-1})$ based on our definition.

Following the example of RHMM for forward k-observability, let us define the reversed transition matrix $\bar{A}_z$, where $\bar{A}_z[i, j]$ stands for $P(z_t = i \mid z_{t+1} = j)$, which always exists and can be computed by using the Bayes rule with A_z. With $\bar{A}_z$ and C_z, we have a new HMM, which runs in a backward fashion: from time step $t + 1$ to step t. Assuming that this new HMM has k-observability, then we have a bijective map between $P(a_t, s_t \mid x_{t:T})$ and $P(x_{t-k:t-1} \mid x_{t:T})$. Given $P(a_{t-k:t}, x_{t-k:t-1} \mid x_{t:T})$ resulting from some $P(a_t, s_t \mid x_{t:T})$, we can reveal $P(a_t, s_t \mid x_{t:T})$ by first marginalizing $a_{t-k:t}$ and, then, applying the bijective map of HMM to $P(x_{t-k:t-1} \mid x_{t:T})$ to get $P(a_t, s_t \mid x_{t:T})$. With $P(a_t, s_t \mid x_{t:T})$, it is straightforward to compute $P(a_{t-k:t}, x_{t-k:t-1} \mid x_{t:T})$ from the RHMM's graphical model itself.

B PROOF FOR LEMMA. 5.1

Proof. Define $\tau_{t-1} = \{a_1, x_1, ..., a_{t-1}, x_{t-1}\}$. For $\sum_{t=1}^{T} \mathbb{E}_\tau \|\Delta_{m_t^\tau}\|^2$, we have:

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_\tau[\|\hat{m}_t^\tau - m_t^\tau\|^2] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_\tau[\|\hat{m}_t^\tau - \phi(f_t) + \phi(f_t) - m_t^\tau\|^2] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau_{t-1}}\big[\mathbb{E}_{f_t}[\|\hat{m}_t^\tau - \phi(f_t) + \phi(f_t) - m_t^\tau\|^2 \mid \tau_{t-1}]\big] \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau_{t-1}}\big[\mathbb{E}_{f_t}[2\|\hat{m}_t^\tau - \phi(f_t)\|^2 + 2\|\phi(f_t) - m_t^\tau\|^2 \mid \tau_{t-1}]\big] \\
&= \frac{2}{T}\sum_{t=1}^{T}\mathbb{E}_\tau[\|\hat{m}_t^\tau - \phi(f_t)\|^2] + \frac{2}{T}\sum_{t=1}^{T}\mathbb{E}_\tau[\|\phi(f_t) - m_t^\tau\|^2] \\
&\leq 2\epsilon_m + 2\delta_m \\
&= O(\epsilon_m + \delta_m),
\end{aligned} \tag{15}
$$

where the first inequality comes from the fact that $\|a + b\|^2 = \|a\|^2 + 2a^T b + \|b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for any vector a and b. Following similar derivation, we can prove the conclusion for Δ_{v_t}. □

Getting rid of Bayes error δ_m (δ_v) in general is difficult. Define the risk in Eq. 5 as $l(F)$ with respect to any hypothesis F in $\mathcal{F}_1$. In the worst case, it is possible that there exist two different hypothesis F_1 and F_2 in $\mathcal{F}_1$ both globally minimize the risk $l(F)$ due to the non-convexity of $l(F)$. Assuming in realizable case, F_1 is the true underlying filter that generates m_t^τ for time step t and trajectory τ. Define $\delta_t^\tau = m_t^\tau - \phi(f_t^\tau)$, we have $l(F_1) = \mathbb{E}_\tau \sum_{t=1}^{T} \|\delta_t^\tau\|^2/T = \delta_m$ by definition. Assume that F_2 generates predictive state $\hat{m}_t^\tau$ that are always equal to $m_t^\tau - 2\delta_t^\tau$, it is not hard to verify that the F_2 is risk consistent: $l(F_2) = \delta_m$. However, when we compute the difference between m_t from F_1 and $\hat{m}_t$ from F_2, we see that the gap $\mathbb{E}_\tau \sum_{t=1}^{T} \|\hat{m}_t^\tau - m_t^\tau\|^2/T$ is $2\delta_m$. Hence without any further assumptions on the loss $l(F)$ and hypothesis F, in the worst-case, the bound can be tight.

C PROOF FOR THEOREM. 5.2

Proof. Given τ, let us define z_t^τ as $(m_t^\tau, v_{t+1}^\tau, x_t^\tau)$, the concatenation of the forward true message m_t^τ, the true backward message v_{t+1}^τ, and the local feature x_t^τ. Remind that G^* is defined as:

$$\arg\min_{G\in\mathcal{F}_3} \frac{1}{T}\mathbb{E}_{\tau\sim\mathcal{D}}\sum_{t=1}^{T}[\ell(G(z_t^\tau), a_t^\tau))], \tag{16}$$

which is the optimal hypothesis from $\mathcal{F}_3$ that minimizes the prediction error given the exact messages m_t^τ and v_t^τ on τ sampled from $\mathcal{D}$ (messages m_t^τ, v_t^τ is fully determined by τ). The learned hypothesis $\hat{G}$ (Eq. 11) is

$$\arg\min_{G\in\mathcal{F}_3} \frac{1}{T}\mathbb{E}_{\tau\sim\mathcal{D}}\sum_{t=1}^{T}[\ell(G(\hat{z}_t^\tau), a_t^\tau))]. \tag{17}$$

The difference between z_t and $\hat{z}_t$ is determined by $\delta_{m_t^\tau}$ and $\delta_{v_t^\tau}$. Since $G(\cdot)$ is L_2-Lipschitz continuous, we have that, given a τ and a time step t:

$$\|G^*(\hat{z}_t^\tau) - G^*(z_t^\tau)\| \le L_2\|\hat{z}_t^\tau - z_t^\tau\| = L_2\|\Delta_{m_t^\tau}\| + L_2\|\Delta_{v_t^\tau}\|.$$

since loss function ℓ is L_1-Lipschitz continuous, we have that for a given τ and a time step t:

$$|\ell(G^*(\hat{z}_t^\tau), a_t^\tau) - \ell(G^*(z_t^\tau, a_t^\tau))| \le L_1\|G^*(\hat{z}_t^\tau) - G^*(z_t^\tau)\| \le L_1L_2\|\Delta_{m_t^\tau}\| + L_1L_2\|\Delta_{v_t^\tau}\|.$$

Now let us put expectation with respect to τ back on both sides of the above inequality and use Jensen inequality, we have:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}|\mathbb{E}_{\tau\sim\mathcal{D}}[\ell(G^*(\hat{z}_t^\tau), a_t^\tau)] - \mathbb{E}_{\tau\sim\mathcal{D}}[\ell(G^*(z_t^\tau), a_t^\tau)]| \le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau\sim\mathcal{D}}|\ell(G^*(\hat{z}_t^\tau), a_t^\tau) - \ell(G^*(z_t^\tau), a_t^\tau)| \\
&\le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau\sim\mathcal{D}}[L_1L_2\|\Delta_{m_t^\tau}\| + L_1L_2\|\Delta_{v_t^\tau}\|] \le 2L_1L_2(\epsilon_m + \epsilon_v + \delta_m + \delta_v),
\end{aligned} \tag{18}$$

where the first inequality comes from Jensen inequality. Since $\hat{G}$ is the minimizer of $\mathbb{E}_{\tau\sim\mathcal{D}}\sum_{t=1}^{T}[\ell(G(\hat{z}_t^\tau), a_t^\tau)]$, we must have:

$$\mathbb{E}_{\tau\sim\mathcal{D}}\sum_{t=1}^{T}[\ell(\hat{G}(\hat{z}_t^\tau), a_t^\tau)] \le \mathbb{E}_{\tau\sim\mathcal{D}}\sum_{t=1}^{T}[\ell(G^*(\hat{z}_t^\tau), a_t^\tau)].$$

Combine the above two inequality together, we have:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau\sim\mathcal{D}}[\ell(\hat{G}(\hat{z}_t^\tau), a_t^\tau)] \le \frac{1}{T}\mathbb{E}_{\tau\sim\mathcal{D}}\sum_{t=1}^{T}[\ell(G^*(\hat{z}_t^\tau), a_t^\tau)] \\
&\le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau\sim\mathcal{D}}[\ell(G^*(z_t^\tau), a_t^\tau)] + 2L_1L_2(\epsilon_m + \epsilon_v + \delta_m + \delta_v)
\end{aligned}$$

Hence we prove the above theorem. □

D PROOF FOR LEMMA. 5.3

Proof. The finite sample analysis for PSIM with Data Aggregation shows that given M training sequences, with probability $1-\delta$, for F_f and F_b we have:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau\sim\mathcal{D}}(\hat{m}_t^\tau - \phi(f_t^\tau))^2 \le \hat{\gamma}_m + \hat{\epsilon}_m + O(\sqrt{\frac{\ln(1/\delta)}{MN}}); \tag{19}$$

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau\sim\mathcal{D}}(\hat{v}_t^\tau - \xi(h_t^\tau))^2 \le \hat{\gamma}_v + \hat{\epsilon}_v + O(\sqrt{\frac{\ln(1/\delta)}{MN}}); \tag{20}$$

where N is the number of iterations the PSIM used for learning F_f and F_b, $\hat{\gamma}_m$ and $\hat{\gamma}_n$ is the average of regret which converges to zero as $N \to \infty$, and $\hat{\epsilon}_m$ and $\hat{\epsilon}_v$ are the minimum regression error or classification error in hindsight on all collected data with respect to the best hypothesis in the hypothesis class.

Now using the similar derivation as in the proof for lemma. 5.1, it is easy to show that with probability $1-\delta$:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau}[(\Delta_{m_t}^{\tau})] \leq 2\hat{\gamma}_m + 2\hat{\epsilon}_m + 2\delta_m + O(\sqrt{\frac{\ln(1/\delta)}{MN}});\\
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau}[(\Delta_{v_t}^{\tau})] \leq 2\hat{\gamma}_v + 2\hat{\epsilon}_v + 2\delta_v + O(\sqrt{\frac{\ln(1/\delta)}{MN}});
\end{aligned} \tag{21}$$

□

E PROOF FOR THEOREM. 5.4

Proof. Remind that $\hat{d}_t$ represents the joint distribution of $\mu(\hat{m}_t^{\tau}, \hat{v}_{t+1}^{\tau}, x_t^{\tau})$ and a_t^{τ} at time step t. With fixed F and F', $\hat{d}_t$ is fully determined by the underlying Since we use the first half M training sequences for PSIM to learn F and F', the left half M training sequences for the learned F and F' to generate $\hat{m}_t$ and $\hat{v}_t$, the corresponding sample set of $\hat{d}_t$ is $\{(\hat{z}_t^{\tau_{M+1}}, a_t^{\tau_{M+1}}), ..., (\hat{z}_t^{\tau_{2M}}, a_t^{\tau_{2M}})\}$. These samples are i.i.d, since $\tau_{M+1}, ..., \tau_{2M}$ are i.i.d samples from $\mathcal{D}$ (F and F' are trained on the first half dataset).

Due to the uniform bound from Rademacher theorem, we known that for the learned hypothesis $\hat{G}$, we have with probability $1-\delta'$:

$$\mathbb{E}_{(\hat{z},a)\sim\hat{d}_t}[\ell(\hat{G}(\hat{z}),a)] - \frac{1}{M}\sum_{i=M+1}^{2M}\ell(\hat{G}(\hat{z}_t^{\tau_i}), a_t^{\tau_i}) \leq 2\mathcal{R}_t(\mathcal{L}) + \sqrt{\frac{\ln(1/\delta')}{2M}} \tag{22}$$

where $\mathcal{R}_t(\mathcal{L})$ is the Rademacher number for the function class $\mathcal{L} = \{(z,a) \to \ell(g(z),a); g \in \mathcal{F}_3\}$ under distribution $\hat{d}_t$. Let us define the above inequality as event A_t being true. The probability that all A_t being true for for $1 \leq t \leq T$ is less than $(1-\delta')^T$. Hence, we have with with probability at least $(1-\delta')^T$:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{(\hat{z},a)\sim\hat{d}_t}[\ell(\hat{G}(\hat{z}),a)] \leq \frac{1}{T}\sum_{t=1}^{T}\frac{1}{M}\sum_{i=M+1}^{2M}\ell(\hat{G}(\hat{z}_t^{\tau_i}), a_t^{\tau_i}) + \frac{2}{T}\sum_{t=1}^{T}\mathcal{R}_t(\mathcal{L}) + \sqrt{\frac{\ln(1/\delta')}{2M}}. \tag{23}$$

Note that $\hat{G}^*$ is the minimizer of the average risk :

$$\hat{G}^* = \arg\min_{g\in\mathcal{F}_3}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\hat{z},a\sim\hat{d}_t}[\ell(g(\hat{z}),a)] = \arg\min_{g\in\mathcal{F}_3}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau\sim\mathcal{D}}[\ell(g(\hat{z}_t^{\tau}),a_t^{\tau})], \tag{24}$$

where $\hat{G}$ is the corresponding empirical risk minimizer, we must have:

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{M}\sum_{i=M+1}^{2M}\ell(\hat{G}(\hat{z}_t^{\tau_i}), a_t^{\tau_i}) \leq \frac{1}{T}\sum_{t=1}^{T}\frac{1}{M}\sum_{i=M+1}^{2M}\ell(\hat{G}^*(\hat{z}_t^{\tau_i}), a_t^{\tau_i}). \tag{25}$$

Substitute the above inequality into the RHS of Inequality. 23, we have with probability $(1-\delta')^T$:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{(\hat{z},a)\sim\hat{d}_t}[\ell(\hat{G}(\hat{z}),a)] \leq \frac{1}{T}\sum_{t=1}^{T}\frac{1}{M}\sum_{i=M+1}^{2M}\ell(\hat{G}^*(\hat{z}_t^{\tau_i}), a_t^{\tau_i}) + \frac{2}{T}\sum_{t=1}^{T}\mathcal{R}_t(\mathcal{L}) + \sqrt{\frac{\ln(1/\delta')}{2M}}. \tag{26}$$

Now for every time step t, let us re-apply the uniform bound from Rademacher analysis to $\hat{G}^*$, we have with probability $(1-\delta')$:

$$\frac{1}{M}\sum_{i=M+1}^{M}\ell(\hat{G}^*(\hat{z}_t^{\tau_i}), a_t^{\tau_i}) \leq \mathbb{E}_{(\hat{z},a)\sim\hat{d}_t}[\ell(\hat{G}^*(\hat{z}),a)] + 2\mathcal{R}_t(\mathcal{L}) + \sqrt{\frac{\ln(1/\delta')}{M}}. \tag{27}$$

Sum both LHS and RHS of the above inequality from $t = 1$ to T, we have with probability at least $(1 - \delta')^T$:

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{M}\sum_{i=M+1}^{2M}\ell(\hat{G}^*(\hat{z}_t^{\tau_i}), a_t^{\tau_i}) \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{(\hat{z},a)\sim\hat{d}_t}[\ell(\hat{G}^*(\hat{z}), a)] + \frac{2}{T}\sum_{t=1}^{T}\mathcal{R}_t(\mathcal{L}) + \sqrt{\frac{\ln(1/\delta')}{M}}. \tag{28}$$

Now let us combine Inequality. 23 and 28 together, we have with probability at least $(1 - \delta')^{2T}$:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{(\hat{z},a)\sim\hat{d}_t}[\ell(\hat{G}(\hat{z}), a)] \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{(\hat{z},a)\sim\hat{d}_t}[\ell(\hat{G}^*(\hat{z}), a)] + \frac{4}{T}\sum_{t=1}^{T}\mathcal{R}_t(\mathcal{L}) + 2\sqrt{\frac{\ln(1/\delta')}{M}}. \tag{29}$$

Now using the similar derivation for Theorem. 5.2 with the finite sample bounds for Δ_m and Δ_v as shown in Lemma. 5.3, we will have with probability $1 - \delta'$:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau\sim\mathcal{D}}[\ell(\hat{G}^*(\hat{z}_t^\tau), a_t^\tau)] \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau\sim\mathcal{D}}[\ell(G^*(z_t^\tau), a_t^\tau)] + O(\hat{\gamma}_m + \hat{\gamma}_v + \hat{\epsilon}_m + \hat{\epsilon}_v + \delta_m + \delta_v) + O(\sqrt{\frac{\ln(1/\delta')}{MN}}). \tag{30}$$

Now let us combine the above results together, we have with probability at least $(1 - \delta')^{2T+1}$

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau\sim\mathcal{D}}[\ell(\hat{G}(\hat{z}_t^\tau), a_t^\tau)] \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau\sim\mathcal{D}}[\ell(\hat{G}^*(\hat{z}_t^\tau), a_t^\tau)] + 4\bar{\mathcal{R}}(\mathcal{L}) + 2\sqrt{\frac{\ln(1/\delta')}{2M}} \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau\sim\mathcal{D}}[\ell(G^*(z_t^\tau), a_t^\tau)] + 4\bar{\mathcal{R}}(\mathcal{L}) + 2\sqrt{\frac{\ln(1/\delta')}{2M}} + O(\sqrt{\frac{\ln(1/\delta')}{MN}}) + O(\hat{\gamma}_m + \hat{\gamma}_v + \hat{\epsilon}_m + \hat{\epsilon}_v + \delta_m + \delta_v).
\end{aligned} \tag{31}$$

Since $(1 - \delta')^{2T+1} \geq 1 - (2T + 1)\delta'$, let $\delta' = \delta/(2T + 1)$ and substitute it back, we have with probability $1 - \delta$:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau\sim\mathcal{D}}[\ell(\hat{G}(\hat{z}_t^\tau), a_t^\tau)] &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau\sim\mathcal{D}}[\ell(G^*(z_t^\tau), a_t^\tau)] + 4\bar{\mathcal{R}}(\mathcal{L}) + O(\sqrt{\frac{\ln((2T+1)/\delta)}{2M}} + \sqrt{\frac{\ln((2T+1)/\delta)}{MN}}) \\
&\quad + O(\hat{\gamma}_m + \hat{\gamma}_v + \hat{\epsilon}_m + \hat{\epsilon}_v + \delta_v + \delta_m) \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau\sim\mathcal{D}}[\ell(G^*(z_t^\tau), a_t^\tau)] + 4\bar{\mathcal{R}}(\mathcal{L}) + \tilde{O}(\sqrt{\frac{\ln(1/\delta)}{M}} + \sqrt{\frac{\ln(1/\delta)}{MN}}) \\
&\quad + O(\hat{\gamma}_m + \hat{\gamma}_v + \hat{\epsilon}_m + \hat{\epsilon}_v + \delta_v + \delta_m)
\end{aligned} \tag{32}$$

Here we assumed that T is a bounded constant. When $N \to \infty$, we have $\hat{\gamma}_m \to 0$ and $\hat{\gamma}_v \to 0$, hence, when the number of iterations of DAgger approaches to infinity, we have with probability at least $1 - \delta$:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau\sim\mathcal{D}}[\ell(\hat{G}(\hat{z}_t^\tau), a_t^\tau)] &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tau\sim\mathcal{D}}[\ell(G^*(z_t^\tau), a_t^\tau)] + 4\bar{\mathcal{R}}(\mathcal{L}) + \tilde{O}(\sqrt{\frac{\ln(1/\delta)}{M}}) \\
&\quad + O(\hat{\epsilon}_m + \hat{\epsilon}_v + \delta_v + \delta_m).
\end{aligned} \tag{33}$$

Since we assume that the loss function ℓ is L_1-Lipschitz continuous with respect to the first term, following the composition property of Rademacher number, we will have that $\mathcal{R}_t(\mathcal{L}) = L_1\mathcal{R}_t(\mathcal{F}_3)$. Substitute $\mathcal{R}_t(\mathcal{L}) = L_1\mathcal{R}_t(\mathcal{F}_3)$ into the above inequality, we prove the theorem. □

F DETAILS OF THE ROBOTICS DATASETS

F.1 CART-POLE

The 4-d state q of cart-pole consists of the angular position and angular velocity of the pendulum, as well as the position and velocity of the cart. The 1-d action is the force applied on the cart. The stochastic observation model returns the relative position of the tip with respect to the cart.

F.2 BICYCLE BALANCING

The 7-d state q of the bicycle consists of ω: angle from the vertical to the bicycle, $\dot{\omega}$: angular velocity, θ: angle of the handlebars displacement, and $\dot{\theta}$, and ψ: angle between the bicycle frame and the x-axis. The 2-d action consists of the torque applied to the handlebar and the displacement of the rider. The stochastic observation model returns θ and $\dot{\theta}$, subject to Gaussian noise.

For the experimental setting where we have the latent state s_t, we move $\dot{\theta}$ and $\dot{\omega}$ to s_t, and leave θ, ω and ψ for a_t. It is straightforward, in this setting, to see that using multiple steps of future a_t is helpful, since we can estimate the velocity information s_t from a_{t+1} and a_t. The inference task (smoothing), here, is to predict a_t given all observations $x_{1:T}$, without any access to latent states s_t.

F.3 HELICOPTER HOVER

The 19-d state q of the helicopter consists of the 3-d position, relative to the desired hover position, 3-d velocity, 3-d angular velocity in the helicopter's coordinate, 4-d quaternion in world's coordinate, and an additional 5-d vector modelling the gusts. The observation model returns the 3-d position and the corresponding velocity in the world coordinate, subject to Gaussian noise.

For the experimental setting where we have the latent state s_t, we partition the full state q into two sets: a_t consists of 3-d position and 4-d quaternion; s_t consists of 3-d velocity, 3-d angular velocity, and the 5-d vector modelling the gusts.

F.4 SWIMMER

The swimmer has 3-links, and its state q consists of the 2-d position of the nose, relative to the goal, 2-d angles, 2-d velocity of the nose, and 3-d angular velocities. The 3-d action consists of torques applied on the 3 links. The observation model returns the position and velocity of the nose in the swimmer's body coordinate, subject to Gaussian noise.

For the experimental setting where we have the latent state s_t, we partition the full state q into two sets: a_t consists of 2-d position of the nose and 2-d angles, while the latent state s_t consists of the 3-d velocity of the nose and 3-d angular velocities.