Informed POMDP: Leveraging Additional Information in Model-Based RL

Gaspard Lambrechts, Adrien Bolland and Damien Ernst

1. Informed POMDP

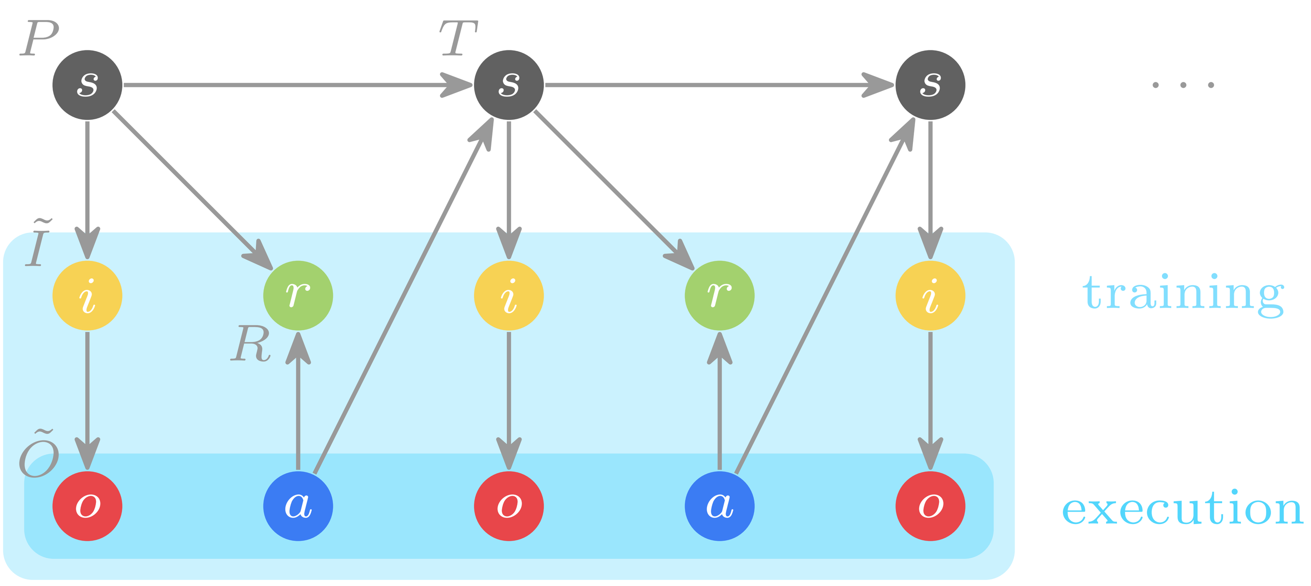
While partial observability at execution time is a realistic assumption, assuming the same partial observability at training time is too pessimistic.

Informed POMDP

Formally, an informed POMDP $\widetilde{\mathcal{P}}$ is defined as $\widetilde{\mathcal{P}} = (\mathcal{S}, \mathcal{A}, \mathcal{I}, \mathcal{O}, T, R, \widetilde{I}, \widetilde{O}, P, \gamma)$,

- State $s \in \mathcal{S}$,
- Action $a \in \mathcal{A}$,
- Information $i \in \mathcal{I}$,
- Observation $o \in \mathcal{O}$,
- Transition distribution $T(s' \mid s, a)$,
- Reward function $r = R(s, a)$,
- Information distribution $\widetilde{I}(i \mid s)$,
- Observation distribution $\widetilde{O}(o \mid i)$,
- Initialization distribution $P(s_0)$,
- Discount factor $\gamma \in [0, 1[$.

NB: o is conditionally independent of s given i.



Execution POMDP

The underlying execution POMDP $\mathcal{P}$ of the informed POMDP $\widetilde{\mathcal{P}}$ is defined as $\mathcal{P} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, T, R, O, P, \gamma)$, where $O(o|s) = \int_I \widetilde{O}(o|i)\widetilde{I}(i|s)\,\mathrm{d}i$.

The history at time t is defined as $h_t = (o_0, a_0, \ldots, o_t) \in \mathcal{H}$, where $\mathcal{H}$ is the set of histories of arbitrary length.
A history-dependent policy $\eta \colon \mathcal{H} \to \Delta(\mathcal{A})$ is a mapping from histories to probability measures over the action space, and is optimal when it maximizes the return,

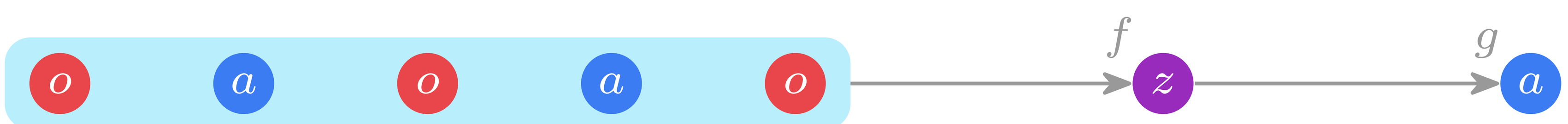
$$J(\eta) = \mathop{\mathbb{E}}_{\mathcal{P},\eta}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]. \tag{1}$$

The RL objective is to find an optimal policy for the execution POMDP using interaction samples $(i_0, o_0, a_0, r_0, \ldots, i_t, o_t)$ from the informed POMDP.

2. Learning Sufficient Statistics

If a statistic from the history is recurrent and predictive of the reward and information given the action, it is sufficient for the optimal control.

We consider policies that compute a statistic from the history $z = f(h)$, before outputting the action distribution $\eta(a|h) = g(a|f(h))$, denoted $\eta = g \circ f$.



This statistic needs to contain all relevant information from the history to act optimally.

Theorem (Sufficiency of Recurrent Predictive Sufficient Statistics)

In an informed POMDP $\widetilde{\mathcal{P}}$, a statistic $f \colon \mathcal{H} \to \mathcal{Z}$ is sufficient for the optimal control, i.e., $\max_g J(g \circ f) = \max_\eta J(\eta)$, if it is,

(i) recurrent $f(h') = u(f(h), a, o')$, $\forall h' = (h, a, o')$,

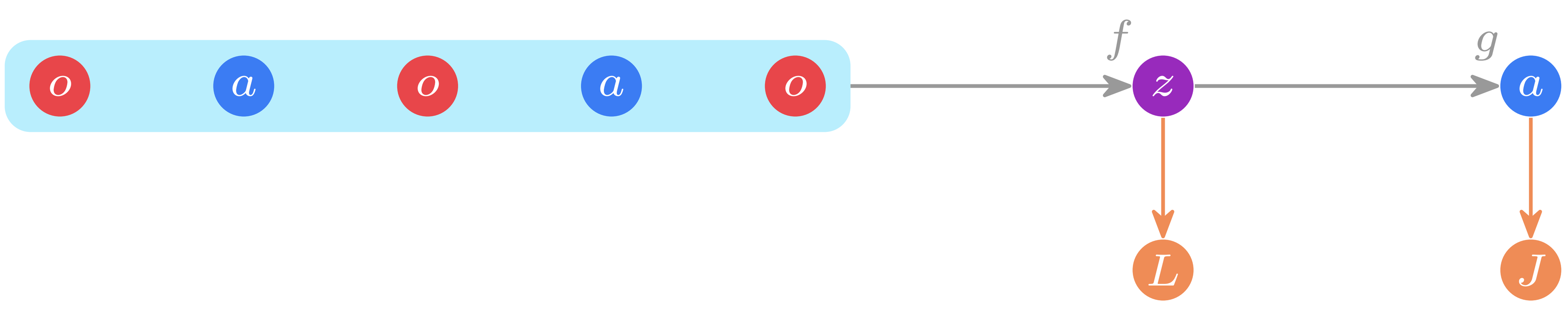
(ii) predictive sufficient $p(r, i'|h, a) = p(r, i'|f(h), a)$, $\forall (h, a, r, i')$.

Under mild assumptions, those sufficiency conditions can be satisfied (i) by design (e.g., using an RNN f_θ) and (ii) by maximising the following variational objective,

$$\max_\theta \underbrace{\mathop{\mathbb{E}}_{p(h,a,r,i')} \log q_\theta(r, i'|f_\theta(h), a)}_{L(f_\theta)}. \tag{2}$$

In practice, we jointly maximize the sufficiency objective and the RL objective, using a parametrized history-dependent policy $\eta_{\theta,\phi} = g_\phi \circ f_\theta$,

$$\max_{\theta,\phi} J(g_\phi \circ f_\theta) + L(f_\theta). \tag{3}$$



3. Informed Dreamer

Learning a sufficient statistic using the reward and information still provides a world model from which latent trajectories can be sampled.

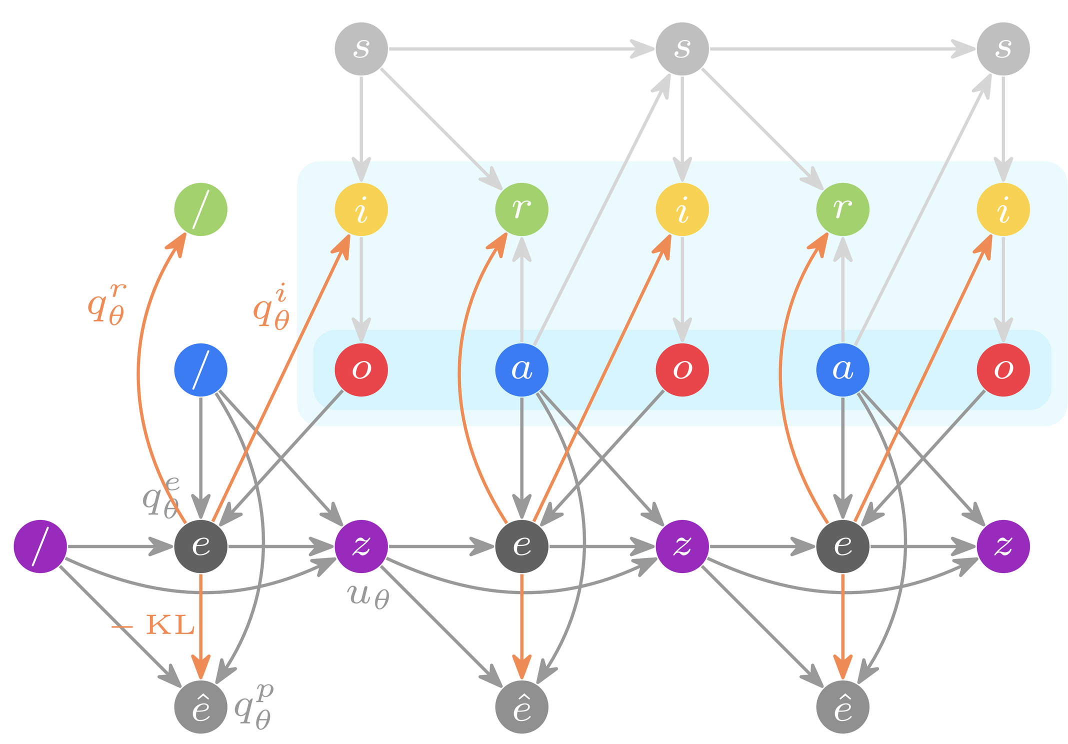
Informed World Model

The informed world model writes,

$$\begin{aligned}
\hat{e} &\sim q^p_\theta(\cdot|z, a), && \text{(prior, 4)}\\
\hat{r} &\sim q^r_\theta(\cdot|z, \hat{e}), && \text{(reward decoder, 5)}\\
\hat{i}' &\sim q^i_\theta(\cdot|z, \hat{e}), && \text{(information decoder, 6)}\\
e &\sim q^e_\theta(\cdot|z, a, o'), && \text{(encoder, 7)}\\
z' &= u_\theta(z, a, e). && \text{(recurrence, 8)}
\end{aligned}$$

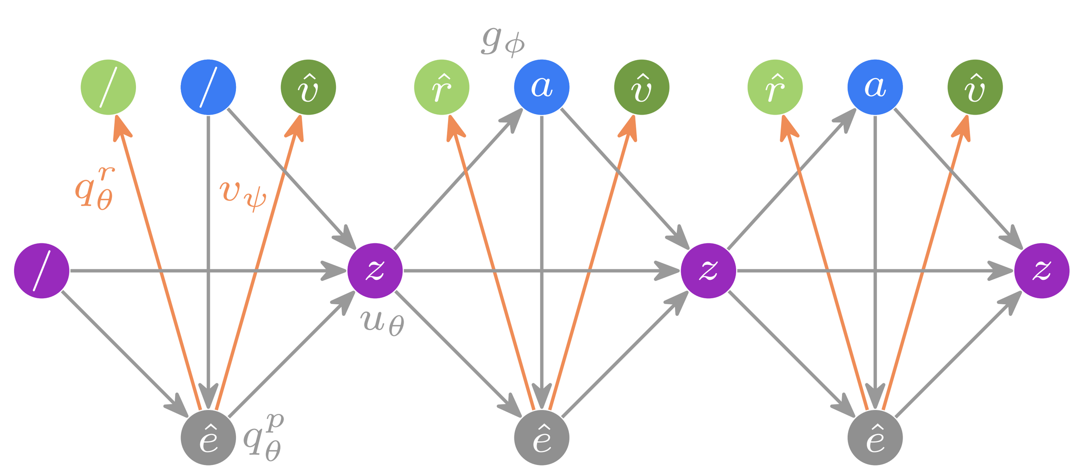
where $\hat{e}$ is the latent variable. The prior q^p_θ and the decoders q^i_θ and q^r_θ are jointly trained with the encoder to maximize the likelihood (2) using the ELBO.

Note that the statistic z is no longer deterministically updated to z' given a and o', instead we have $z \sim f_\theta(\cdot|h)$, which is induced by u_θ and q^e_θ.

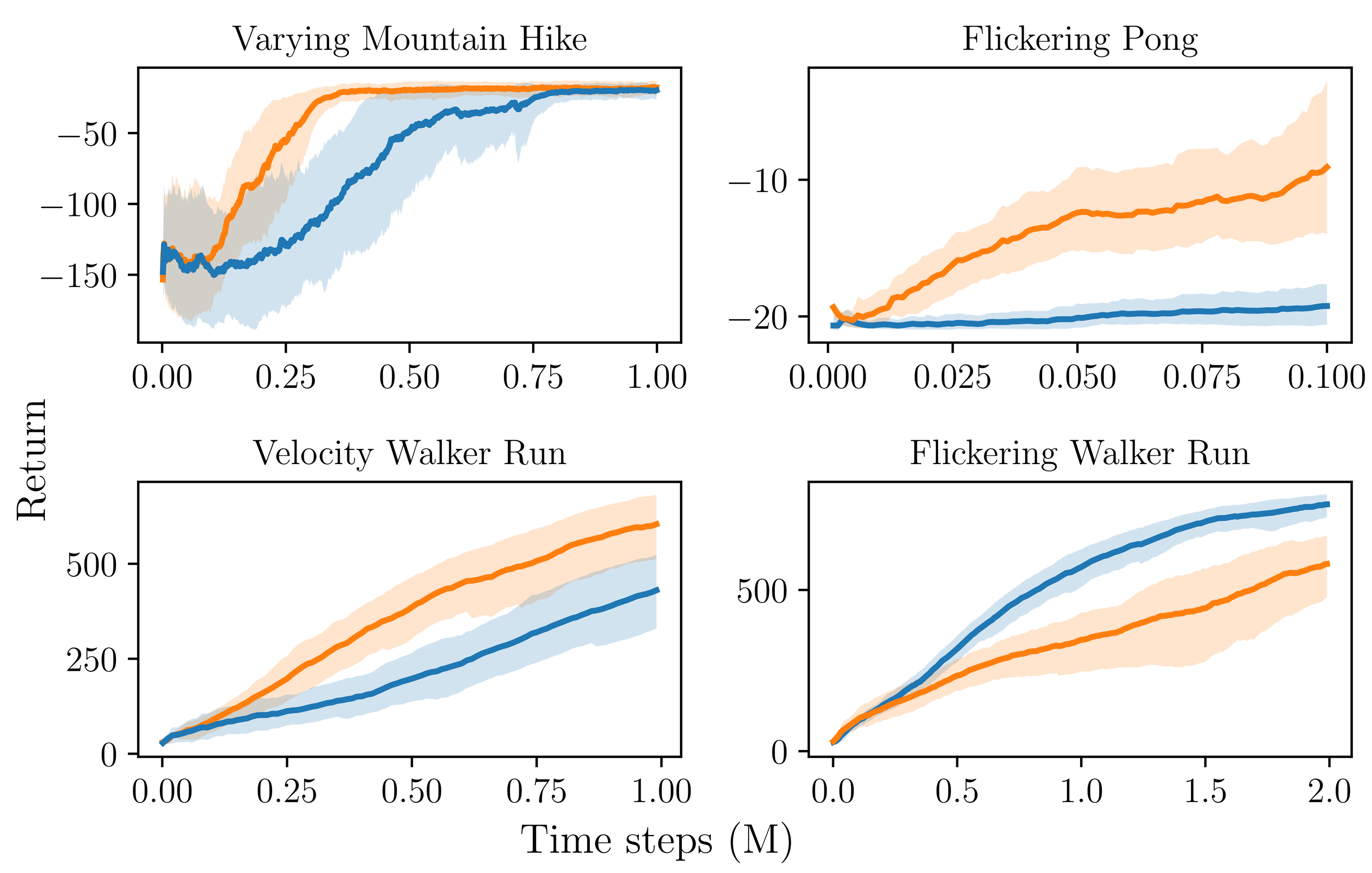


The latent representation $\hat{e}$, trained to minimize KL divergence in to e in expectation, encodes the whole dependence of r, i' (and thus o') on the history.

$\Rightarrow$ It allows sampling latent trajectoires without needing an observation decoder, but using its latent representation $\hat{e} \sim q^p_\theta(\cdot|z, a)$ in update (8).



The learning curves of the Uninformed Dreamer and the Informed Dreamer are given below for some illustrative (cherry-picked) environments.



Conclusion

Take-Home Message

- It is easy and useful to exploit additional information when available at training.
- If i is designed carefully, recurrently learning $p(r, i'|h, a)$ provides a sufficient statistic.
- It also provides an informed world model.

Future Works

- Generalize theorem to stochastic $z \sim f_\theta(\cdot|h)$ to better support the world model.
- Study conditions on the information i for the convergence speed to improve.
- Study robustness and generalization of the informed world model.