Informed POMDP: Leveraging Additional Information in MBRL

Reinforcement Learning Conference - August 12th, 2024

Gaspard Lambrechts, Adrien Bolland and Damien Ernst

Informed POMDP

A story of partial observability

Decision process	Execution	Training	Generality
MDP	s	s	**Too optimistic.**
POMDP	o	o	**Too pessimistic.**
Asymmetric POMDP	o	s	**Too optimistic.**
Informed POMDP	o	i	**Just right?**

Classical POMDP

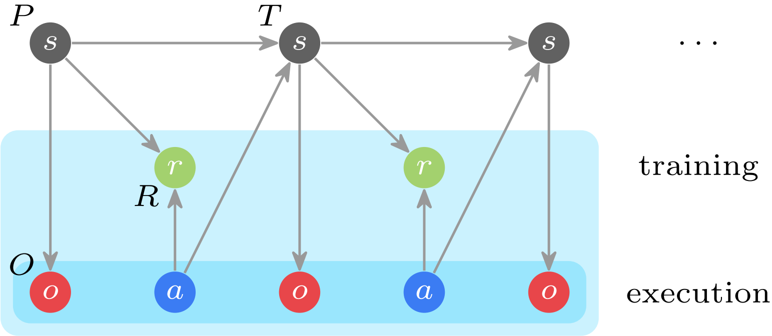

Fig. 1: Bayesian graph of a POMDP.

Informed POMDP

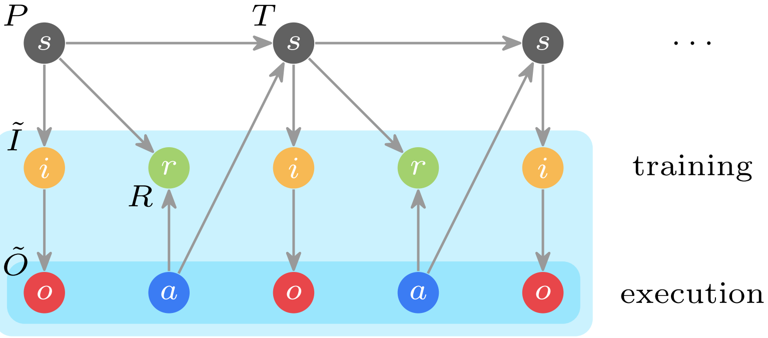

Fig. 2: Bayesian graph of an informed POMDP.

Informed Dreamer

Sufficiency for optimal control

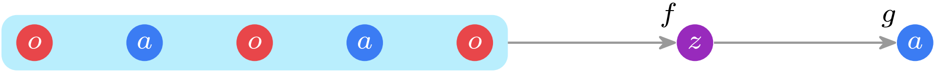

Fig. 3: Statistic $z = f(h)$ of the history h.

- The history h is compressed into a statistic z by a function f.
  - RNN, Transformer, SSM, etc.
- It should summarize all relevant information to act optimally.

Definition 1: Sufficiency for optimal control.

A statistic $f : \mathcal{H} \to \mathcal{Z}$ is **sufficient for optimal control** if, and only if,

$$\max_g J(g \circ f) = \max_\eta J(\eta).$$

Sufficiency in an informed POMDP

Theorem 1: Sufficiency of recurrent predictive statistics.

In an **informed POMDP**, a statistic $f : \mathcal{H} \to \mathcal{Z}$ is **sufficient** for optimal control if it is,

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

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

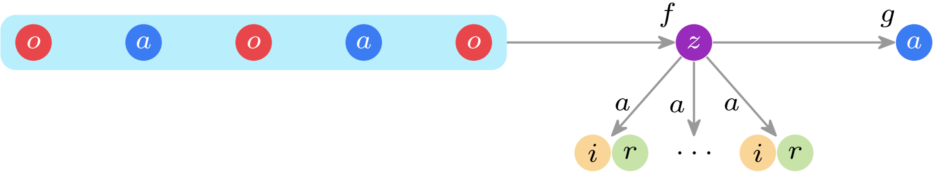

Fig. 4: Statistic $z = f(h)$ of the history h encoding the transition distribution.

A simple view of the Informed Dreamer

The **informed world model** $q(r, i' | f(h), a)$ is learned through likelihood maximization:

$$\max \underbrace{\mathop{\mathbb{E}}_{p(r,i'|h,a)} q(r, i' | f(h), a)}_{L}.$$

- The statistic $z = f(h)$ is **recurrent**.
- At optimum, the statistic is **predictive**.

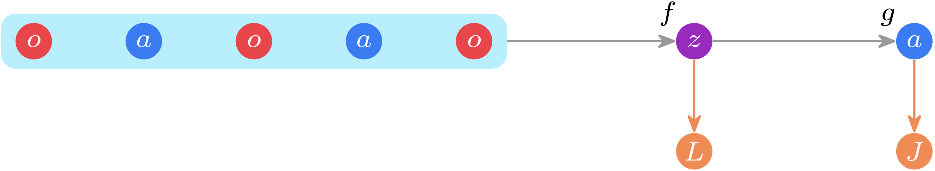

Fig. 5: Sufficiency objective L and reinforcement objective J.

Informed Dreamer

- Prior $\hat{e} \sim q^e(\cdot\,|z, a)$
- **Information** $\hat{i} \sim q^i(\cdot\,|z, \hat{e})$
  - Instead of observation $\hat{o} \sim q^o(\cdot\,|z, \hat{e})$
- Reward $\hat{r} \sim q^r(\cdot\,|z, \hat{e})$
- **Encoder** $e \sim q^e(\cdot\,|z, a, o')$
- Update $z' = u(z, a, e)$.

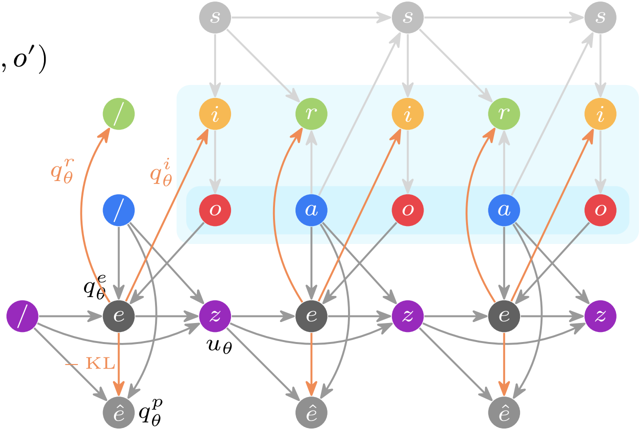

Fig. 6: Informed Dreamer

Results

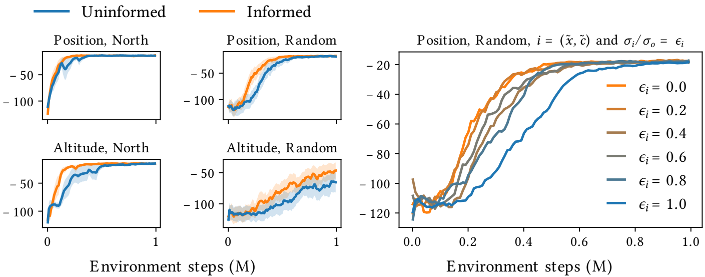

Fig. 7: Varying Mountain Hike

Results (ii)

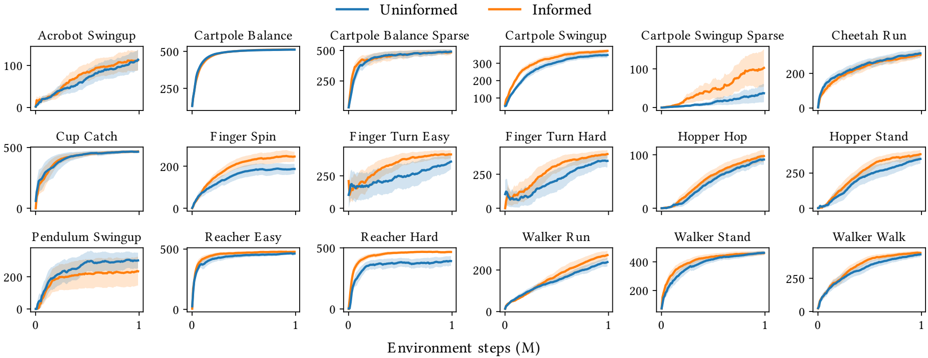

Fig. 8: Velocity DeepMind Control

Results (iii)

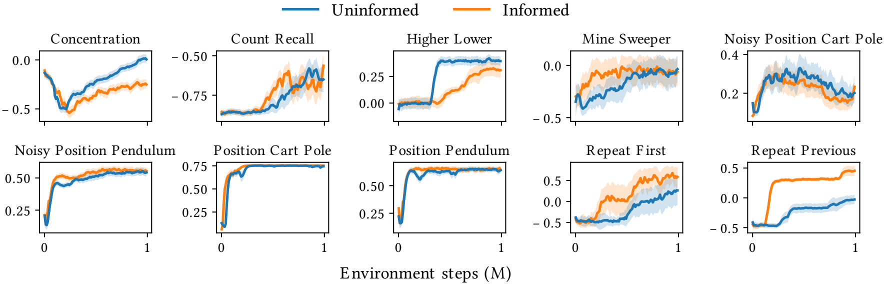

Fig. 9: Pop Gym

Take-home message

Don’t make the problem harder than it is.

Consider all available information at training.