# Learning to Remember the Past by Learning to Predict the Future

**VUB Reinforcement Learning Talks** - November 17th, 2023

Gaspard Lambrechts

# Outline

# Partial observability

# Partial observability

A **POMDP** is described by a model $\mathcal{P} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, P, O, T, R, \gamma)$.

- States $s_t \in \mathcal{S}$,
- Actions $a_t \in \mathcal{A}$,
- **Observations** $o_t \in \mathcal{O}$,
- Initialisation $P(s_0)$,
- **Perception** $O(o_t \mid s_t)$,
- Transition $T(s_{t+1} \mid s_t, a_t)$,
- Reward $r_t = R(s_t, a_t)$,
- Discount $\gamma \in [0, 1[$.

**States** satisfy the **Markov property** but are **not available**,

$$p(s_{t+1} \mid s_0, a_0, ..., s_t, a_t) = p(s_{t+1} \mid s_t, a_t) = T(s_{t+1} \mid s_t, a_t).$$

**Observations** do not satisfy the **Markov property**,

$$p(o_{t+1} \mid o_0, a_0, ..., o_t, a_t) \neq p(o_{t+1} \mid o_t, a_t).$$

$\Rightarrow$ Contrary to MDP, selecting $a_t$ based on $o_t$ only is **suboptimal**.

# Partial observability (ii)

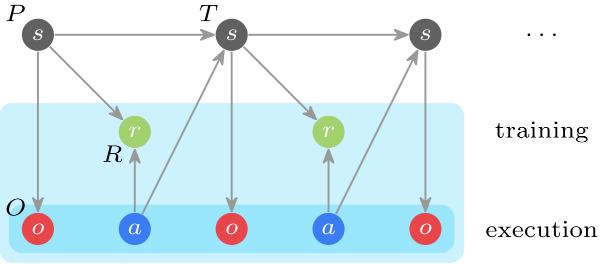


**Fig. 1**: Bayesian graph of a POMDP execution.

The **history** at time $t$ is $h_t = (o_0, a_0, \ldots, o_t) \in \mathcal{H}$, with $\mathcal{H}$ the set of histories. Includes all information available to select action $a_t$.

**Notation:** $(h, a, r, o') \equiv (h_t, a_t, r_t, o_{t+1})$ for arbitrary $t$.

# History-dependent policies

**Definition 1:** History-dependent stochastic policy.

A history-dependent stochastic policy $\eta \in \mathrm{H} = \mathcal{H} \to \Delta(\mathcal{A})$ is a mapping from histories to distributions over the actions, whose density writes $\eta(a \mid h)$.

**Definition 2:** Value function of a policy.

The value function of a history-dependent stochastic policy gives the expected return of the policy starting from a given history,

$$V^\eta(h) = \mathop{\mathbb{E}}_{\mathcal{P},\eta}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid h_0 = h\right], \ s_0 \sim p(S \mid h).$$

**Definition 3:** Q-function of a policy.

The Q-function of a history-dependent stochastic policy gives the expected return of the policy starting from a given history and a given action,

$$Q^\eta(h, a) = \mathop{\mathbb{E}}_{\mathcal{P},\eta}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid h_0 = h, a_0 = a\right], \ s_0 \sim p(S \mid h).$$

# Optimal history-dependent policies

**Definition 4:** Optimal value function.

The optimal Q-function gives the optimal expected return starting from a given history,

$$V(h) = \max_{\eta \in \mathrm{H}} V^{\eta}(h).$$

**Definition 5:** Optimal Q-function.

The optimal Q-function gives the optimal expected return starting from a given history and a given action,

$$Q(h, a) = \max_{\eta \in \mathrm{H}} Q^{\eta}(h, a).$$

**Definition 6:** Optimal policy.

A policy $\eta^*$ is optimal when its value function is maximised in every history,

$$V^{\eta^*} = \max_{\eta \in \mathrm{H}} V^{\eta}(h) = V(h).$$

# History-dependent RL

# History-dependent RL

**Substitute the history $h$ to the state $s$** in the value function or policy.

Requires **function approximators** able to process **variable-size histories**:
- Sliding window (window size)
- Recurrent neural network (truncated BPTT)
- Transformer (window size)

They all suffer from the **unbounded growth of the history**, but RNN are able to process histories indefinitely and efficiently.

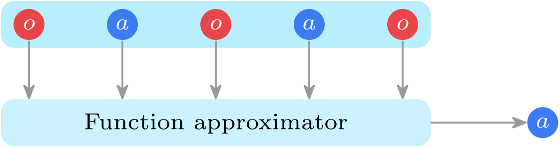


**Fig. 2**: Sequence approximator for history-dependent policy or Q-function.

# Belief-dependent RL

**Definition 7:** Belief of a history.

The belief $b = f(h) \in \Delta(\mathcal{S})$ of a history $h \in \mathcal{H}$ is defined as the posterior probability distribution over the states given the history: $b(s) = p(s \mid h)$.

**Theorem 1:** Sufficiency of the belief.

The Q-function can be written as a function of the belief,

$$Q(h, a) = \mathcal{Q}(b, a),\ \ b = f(h).$$

Moreover, the belief is **recurrent**: $f(h') = u(f(h), a, o'),\ \ h' = (h, a, o')$.

$$b_0(s_0) = \frac{P(s_0)O(o_0 \mid s_0)}{\int_{\mathcal{S}} P(s_0')O(o_0 \mid s_0')\, \mathrm{d}s_0'},$$

$$b_t(s_t) = \frac{O(o_t \mid s_t) \int_{\mathcal{S}} T(s_t \mid s_{t-1}, a_{t-1}) b_{t-1}(s_{t-1})\, \mathrm{d}s_{t-1}}{\int_{\mathcal{S}} O(o_t \mid s_t') \int_{\mathcal{S}} T(s_t' \mid s_{t-1}, a_{t-1}) b_{t-1}(s_{t-1})\, \mathrm{d}s_{t-1}\, \mathrm{d}s_t'}.$$

$\Rightarrow$ If the belief is known, we can discard the history.

# Recurrent RL

**Recurrent Q-learning** learns $Q_\theta(h, a) = g_\theta(f_\theta(h), a)$ where $f_\theta$ is an RNN,

$$f_\theta(h') = u_\theta(f_\theta(h), a, o'), \ \forall h' = (h, a, o').$$

**Reminder:** the belief filter is recurrent,

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

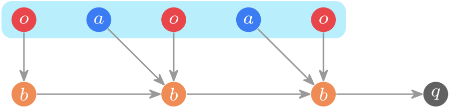


**Fig. 3**: Belief and Q-function.

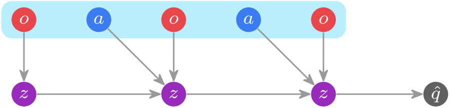


**Fig. 4**: RNN state and Q-function.

$\Rightarrow$ Should RNN states encode the belief?

# Recurrent RL (ii)

Lambrechts, Bolland, and Ernst (2022) highlights that **beliefs emerge** in RNN states during model-free recurrent Q-learning (LSTM, GRU, BRC, NBRC, MGU).

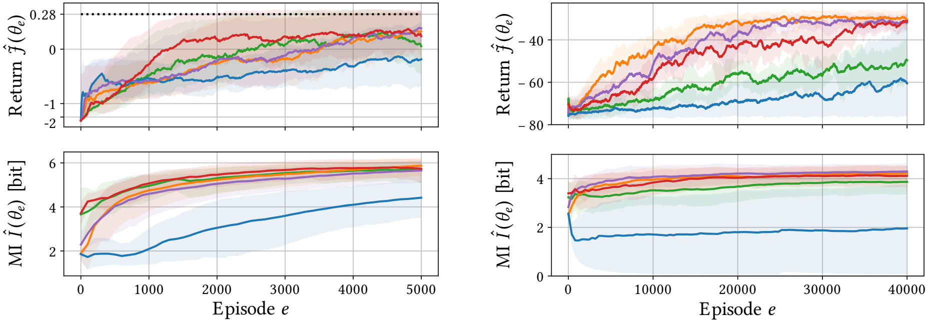


**Fig. 5**: Return and mutual information throughout training.

$\hat{I}(\theta) \approx I(z, b)$ under stationary distribution $p^{\eta_\theta}(h)$.

# Recurrent RL (iii)

The **informativeness** of states about the belief is **correlated** with the **performance** (LSTM, GRU, BRC, NBRC, MGU).

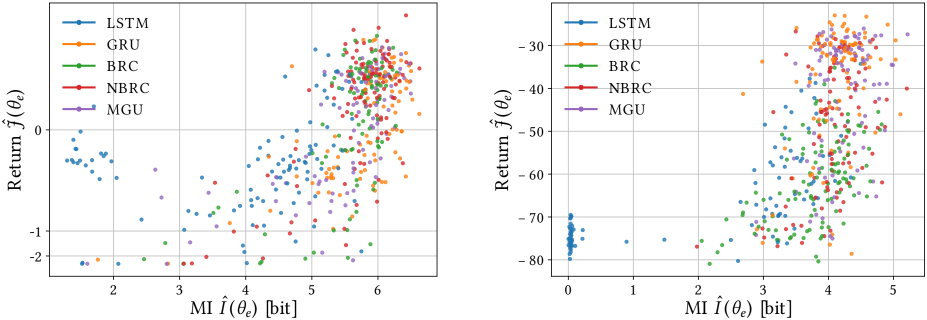


**Fig. 6**: Correlation between return and mutual information.

$\hat{I}(\theta) \approx I(z, b)$ under stationary distribution $p^{\eta_\theta}(h)$.

# Recurrent RL (iv)

The belief of **irrelevant state variable** is not encoded in RNN states (GRU).

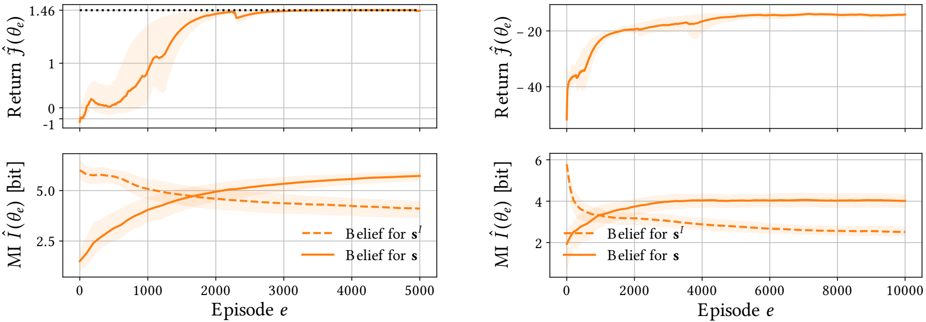


**Fig. 7**: Return and mutual information (belief of relevant and irrelevant state variables) throughout training.

$\hat{I}(\theta) \approx I(z, b)$ under stationary distribution $p^{\eta_\theta}(h)$.

# Sufficient statistic

# Sufficient statistic

**Notation:** $g \circ f$ is the policy $\eta(a \mid h) = g(a \mid f(h))$.

**Definition 8:** Sufficient statistic.

A statistic $f : \mathcal{H} \to \mathcal{Z}$ of the history is sufficient for the optimal control iff,

$$\max_{g:\mathcal{Z}\to\Delta(\mathcal{A})} J(g \circ f) = \max_{\eta\in\mathcal{H}\to\Delta(\mathcal{A})} J(\eta).$$

**Corollary 1:** Sufficiency of optimal policies.

If a policy $\eta = g \circ f$ is optimal, then the statistic $f : \mathcal{H} \to \mathcal{Z}$ is sufficient for the optimal control.

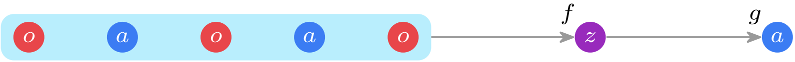


**Fig. 8**: Statistic and policy.

**NB:** The **belief** is a sufficient statistic of the history for the optimal control.

# Sufficiency of recurrent predictive statistics

**Theorem 2:** Sufficiency of recurrent predictive statistics (Subramanian et al. 2022).

A statistic of the history $f : \mathcal{H} \to \mathcal{Z}$ is **sufficient for the optimal control** if it is (i) **recurrent** and (ii) **predictive** of the reward and next observation given the action,

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

(ii) $p(r, o' \mid h, a) = p(r, o' \mid f(h), a), \ \forall (h, a, r, o').$

Intuitively, if a statistic encodes the distribution of the reward and next observation given an action, and can be updated using this observation, then it is **virtually able to simulate all future execution** of the POMDP.

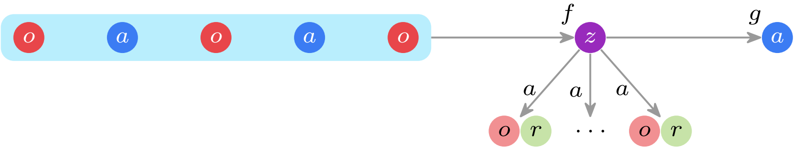


**Fig. 9**: Sufficiency of recurrent and predictive statistics.

# Learning recurrent predictive statistics

Under mild assumptions (e.g., $p(h,a) > 0$), any statistic $f : \mathcal{H} \to \mathcal{Z}$ satisfying

$$\max_{\substack{f:\mathcal{H}\to\mathcal{Z} \\ q:\mathcal{Z}\times\mathcal{A}\to\Delta(\mathbb{R}\times\mathcal{O})}} \mathop{\mathbb{E}}_{p(h,a,r,o')} \log q(r, o' \mid f(h), a), \tag{1}$$

is **predictive** of the reward and next observation given the action (ii). If in addition, the statistic is **recurrent** (i), then it is **sufficient for the optimal control**.

**Algorithm 1:** Sufficient statistic learning.

1. Select a recurrent universal dynamical system approximator $f_\theta$ (e.g., RNN).
2. Select a universal density approximator $q_\theta$ (e.g., GM).
3. Repeat:
    1. Sample trajectories and store transitions $(h, a, r, o')$.
    2. Maximize the log likelihood of (1) using SGD.

**NB:** If $f_\theta$ and $q_\theta$ are expressive enough and satisfy objective (1), then

$$q_\theta(r, o' \mid f_\theta(h), a) = p(r, o' \mid h, a).$$

# Jointly learning statistics and policies

The statistic and the policy can be learned **jointly**,

$$\max_{f,g} J(g \circ f) + L(f). \tag{2}$$

where $J(\eta) = \mathbb{E}_{s_0 \sim P}[V^\eta(s_0)]$ and $L(f) = \max_q \mathbb{E}_{p(h,a,r,o')} \log q(r, o' \mid f(h), a)$.

**Choices** for this objective and its optimization:

- The distribution $p(h, a)$ in $L(f)$ may be that of $g \circ f$ or another policy $\eta$.
- The RL algorithm maximizing $J$ may optimize $f$ or not.

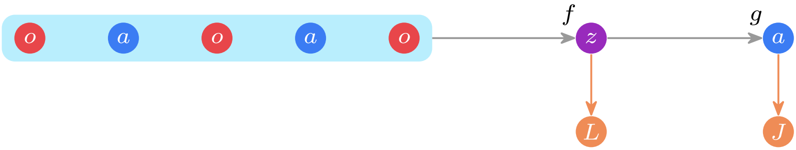


**Fig. 10**: Joint optimization of the statistic and policy.

# Jointly learning statistics and policies (ii)

We have a generic algorithm for optimizing both **sufficiency** and **optimality**.

**Algorithm 2:** Sufficient statistic and policy learning.

1. Select a recurrent universal dynamical system approximator $f_\theta$ (e.g., RNN).
2. Select universal density approximators $q_\theta$ and $g_\varphi$ (e.g., GM).
3. Repeat
   1. Interact (policy $\eta$ or $\eta_{\theta,\varphi} = g_\varphi \circ f_\theta$) and store transitions $(h, a, r, o')$.
   2. Maximize objective (2) using SGD (off- or on-policy RL algorithm).

# Sufficiency of recurrent world models

The model $q_\theta(r, o' \mid f_\theta(h), a) \approx p(r, o' \mid h, a)$ is a **world model**.
$\Rightarrow$ Trajectories can be sampled for free.

It can be exploited in a Dyna / Dreamer algorithm adapted to POMDP.

**Algorithm 3:** Dyna with sufficient statistic for POMDP.

1. Select a recurrent universal dynamical system approximator $f_\theta$ (e.g., RNN).
2. Select universal density approximators $q_\theta$ and $g_\varphi$ (e.g., GM).
3. Repeat
   1. Interact (policy $\eta$ or $\eta_{\theta,\varphi} = g_\varphi \circ f_\theta$) and store transitions $(h, a, r, o')$.
   2. Maximize the log likelihood of objective $(1)$ using SGD.
   3. Repeat:
      1. Imagine trajectories using policy $\eta_{\theta,\varphi} = g_\varphi \circ f_\theta$.
      2. Optimize policy to maximize imagined rewards. **NB:** $\nabla_\varphi \sum_{t=0}^{\infty} \gamma^t \hat{r}_t$ is computable since the world model $q_\theta \circ f_\theta$ is differentiable.

# Asymmetric learning

# Asymmetric learning

Asymmetric learning consists of **exploiting state information** at training.
**Motivation:** assuming the same partial observability at training is restrictive.

We generalise asymmetric learning to **non Markovian** additional information.

The **informed POMDP** is described by $\tilde{\mathcal{P}} = \left(\mathcal{S}, \mathcal{A}, \mathcal{I}, \mathcal{O}, P, \tilde{I}, \tilde{O}, T, R, \gamma\right)$,

- States $s_t \in \mathcal{S}$,
- Actions $a_t \in \mathcal{A}$,
- **Information** $i_t \in \mathcal{I}$,
- Observations $o_t \in \mathcal{O}$,
- Initialisation $P(s_0)$,
- **Supervision** $\tilde{I}(i_t \mid s_t)$,
- **Perception** $\tilde{O}(o_t \mid i_t)$,
- Transition $T(s_{t+1} \mid s_t, a_t)$,
- Reward $r_t = R(s_t, a_t)$,
- Discount $\gamma \in [0, 1[$.

During execution, the information is unavailable and we obtain the POMDP $\mathcal{P} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, P, O, T, R, \gamma)$, where $O(o_t \mid s_t) = \int_{\mathcal{I}} \tilde{O}(o_t \mid i) \tilde{I}(i \mid s_t) \, \mathrm{d}i$.

# Asymmetric learning (ii)

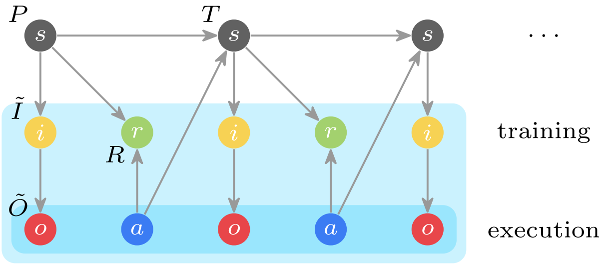


**Fig. 11**: Bayesian graph of an informed POMDP execution.

**NB:** The information is designed such that $o_t$ is independent of $s_t$ given $i_t$.

# Asymmetric learning of sufficient statistics

Usually, the state information is exploited through either
- (constrained) **imitation learning**,
- (unbiased) **asymmetric actor-critic** approaches.

Lambrechts, Bolland, and Ernst (2023) propose to leverage the additional information in the learning of sufficient statistic.
- Exploits additional information **only through the objective**.
- Handles **partial additional information** about the state.

**Theorem 3:** Sufficiency of recurrent informed predictive statistics.

A statistic of the history $f : \mathcal{H} \to \mathcal{Z}$ is **sufficient for the optimal control** if it is (i) **recurrent** and (ii) **predictive** of the reward and next **information** given the action,

$$\begin{aligned} &\text{(i)} \quad f(h') = u(f(h), a, o'), \ \forall h' = (h, a, o'), \\ &\text{(ii)} \quad p(r, \boldsymbol{i}' \mid h, a) = p(r, \boldsymbol{i}' \mid f(h), a), \ \forall (h, a, r, \boldsymbol{i}'). \end{aligned}$$

# Asymmetric learning of sufficient statistics (ii)

The resulting **informed** learning objective is

$$\max_{\substack{f:\mathcal{H}\to\mathcal{Z}\\ q:\mathcal{Z}\times\mathcal{A}\to\Delta(\mathbb{R}\times\mathcal{I})}} \mathop{\mathbb{E}}_{p(h,a,r,\boldsymbol{i}')} \log q(r,\boldsymbol{i}' \mid f(h),a). \tag{3}$$

**Motivation:** $i$ is more informative than $o$: $I(s', i' \mid h, a) \geq I(s', o' \mid h, a)$.

**Algorithm 4:** Informed sufficient statistic learning.

1. Select a recurrent universal dynamical system approximator $f_\theta$ (e.g., RNN).
2. Select a universal density approximator $q_\theta$ (e.g., GM).
3. Repeat
    1. Sample trajectories and store transitions $(h, a, r, \boldsymbol{i}')$.
    2. Maximize the log likelihood of (3) using SGD.

**NB:** If $f_\theta$ and $q_\theta$ are expressive enough and satisfy objective (3), then

$$q_\theta(r, \boldsymbol{i}' \mid f_\theta(h), a) = p(r, \boldsymbol{i}' \mid h, a).$$

# Informed world-model

We use a Dyna / Dreamer algorithm with an **informed world model** using a variational RNN (VRNN or RSSM). Formally, we have,

$$\hat{e} \sim q_\theta^p(\cdot \mid z, a),$$
$$\hat{r} \sim q_\theta^r(\cdot \mid z, \hat{e}),$$
$$\hat{i}' \sim q_\theta^i(\cdot \mid z, \hat{e}),$$

where $\hat{e}$ is the latent variable of the VRNN when generating trajectories. The prior $q_\theta^p$ and decoders $q_\theta^r$ and $q_\theta^i$ are jointly trained with the encoder,

$$e \sim q_\theta^e(\cdot \mid z, a, o'),$$

to maximise likelihood of $(r, i')$. The latent representation $e \sim q_\theta^e(\cdot \mid z, a, o')$ of the next observation $o'$ can be used to update the statistic to $z'$,

$$z' = u_\theta(z, a, e).$$

⚠ The statistic is no longer deterministic, instead we have $z \sim f(\cdot \mid h)$.

# Informed world-model (ii)

In practice, we maximize the **evidence lower bound** (ELBO), a variational lower bound on the likelihood,

$$\mathop{\mathbb{E}}_{\substack{p(h,a,r,i')\\ f_\theta(z\mid h)}} \log q_\theta(r,i' \mid z,a) \geq \mathop{\mathbb{E}}_{\substack{p(h,a,r,i',o')\\ f_\theta(z\mid h)}} \Bigg[ \underbrace{\mathop{\mathbb{E}}_{q^e_\theta(e\mid z,a,o')}\left[\log q^i_\theta(i'\mid z,e) + \log q^r_\theta(r\mid z,e)\right]}_{\text{reconstruction}} \underbrace{- \mathrm{KL}\left(q^e_\theta(\cdot\mid z,a,o') \,\|\, q^p_\theta(\cdot\mid z,a)\right)}_{\text{regularization}} \Bigg].$$

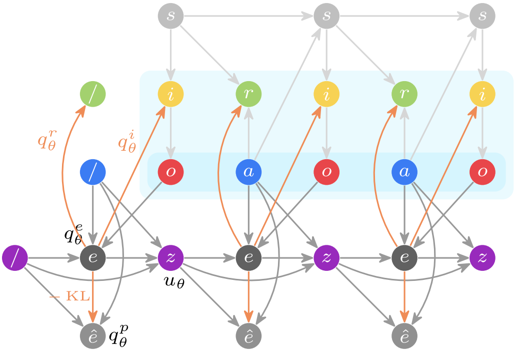


**Fig. 12**: Informed world model training.

# Informed Dreamer

The informed world model allows **sampling in latent space**.

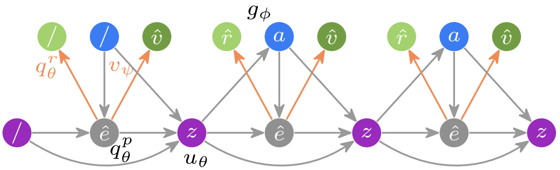


**Fig. 13**: Informed world model imagination.

The policy is executed on the **latent representations** of the observations.

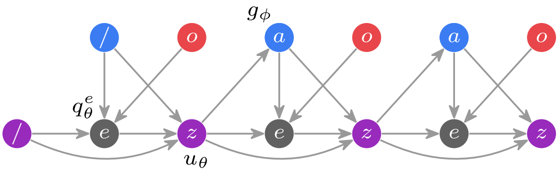


**Fig. 14**: Informed world model imagination

# Informed Dreamer (ii)

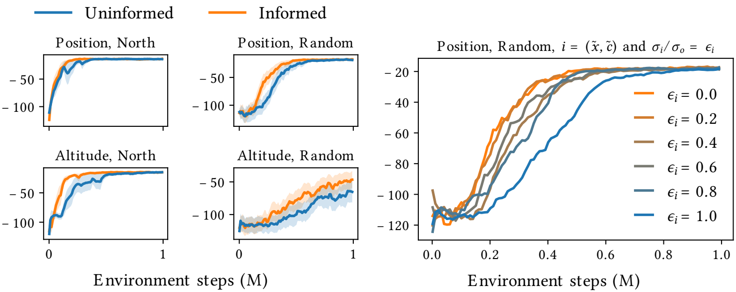


**Fig. 15**: Informed and Uninformed Dreamer in Mountain Hike.

# Informed Dreamer (iii)

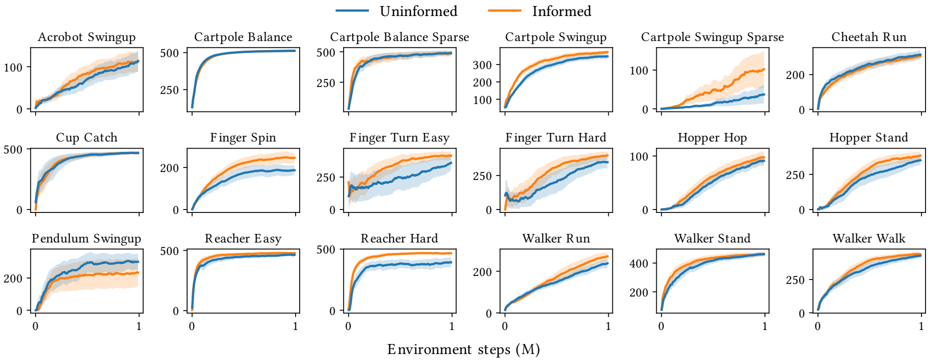


**Fig. 16**: Informed and Uninformed Dreamer in Velocity Control.

# Informed Dreamer (iv)

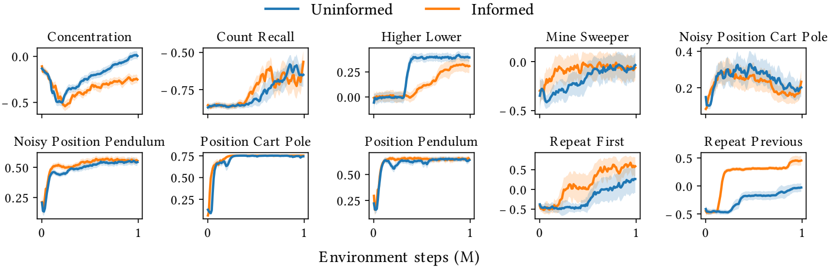


**Fig. 17**: Informed and Uninformed Dreamer in Pop Gym.

# Limitations

- In **theory**:
  - No theoretical support for **stochastic statistics**.
  - Guarantees are for the maximisers only, **bounds are missing**.

- In **practice**:
  - Approximating the conditional information distribution **sometimes hurts performance**.

- Others:
  - **Ill-posed ELBO objective**: the encoder is not conditioned on $i'$ (only the distribution of $o'$, encoded in the distribution of $i'$, and the informational content of $i'$ that is encoded in $h$ are approximated).

# Future works

# Future works

- In **model-based RL**:
  - Fixing the ELBO learning objective.
  - Generalizing the theory to stochastic statistic.
  - Proposing an efficient deterministic (and latent) world model.
- In **model-free RL**:
  - Comparing asymmetric actor-critic to the statistic learning approach.
- In **multi-agent RL**:
  - Considering model-free statistic learning from the local histories of agents.
- In **theory**:
  - Studying generalization when using state supervision.

# Bibliography

Lambrechts, G., Bolland, A., & Ernst, D. (2022, August). *Recurrent networks, hidden states and beliefs in partially observable environments.*

Lambrechts, G., Bolland, A., & Ernst, D. (2023, July). *Informed POMDP: Leveraging additional information in model-based RL.*

Subramanian, J., Sinha, A., Seraj, R., & Mahajan, A. (2022, January). *Approximate information state for approximate planning and reinforcement learning in partially observed systems.*

For other related works, see Section 2 of (Lambrechts, Bolland, and Ernst 2023).

*A warm thank to my coauthors Adrien Bolland and Damien Ernst.*