Outcome-Driven Reinforcement Learning via Variational Inference





Tim G. J. Rudner* **Vitchyr H. Pong*** Rowan McAllister Yarin Gal Sergey Levine

Neural Information Processing Systems 2021

Correspondence to

tim.rudner@cs.ox.ac.uk

vitchyr@berkeley.edu

Can we derive RL from probabilistic inference?

Can we derive RL from probabilistic inference

without access to a reward function?

Yes!

Yes!

A reward function

Yes!

A reward function emerges naturally from inference,

Yes!

A reward function emerges naturally from inference,

can be learned from data, and

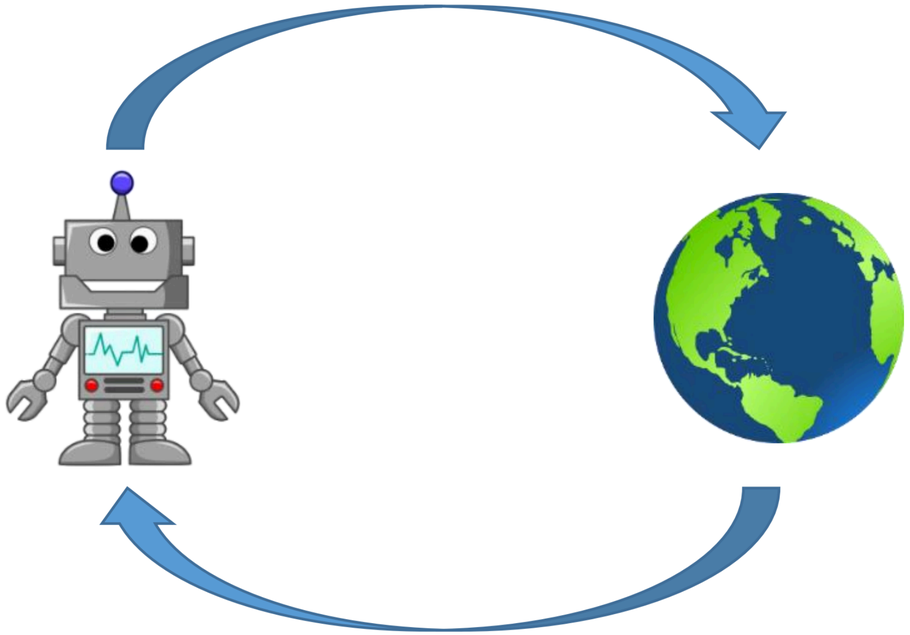
Yes!

A reward function emerges naturally from inference,

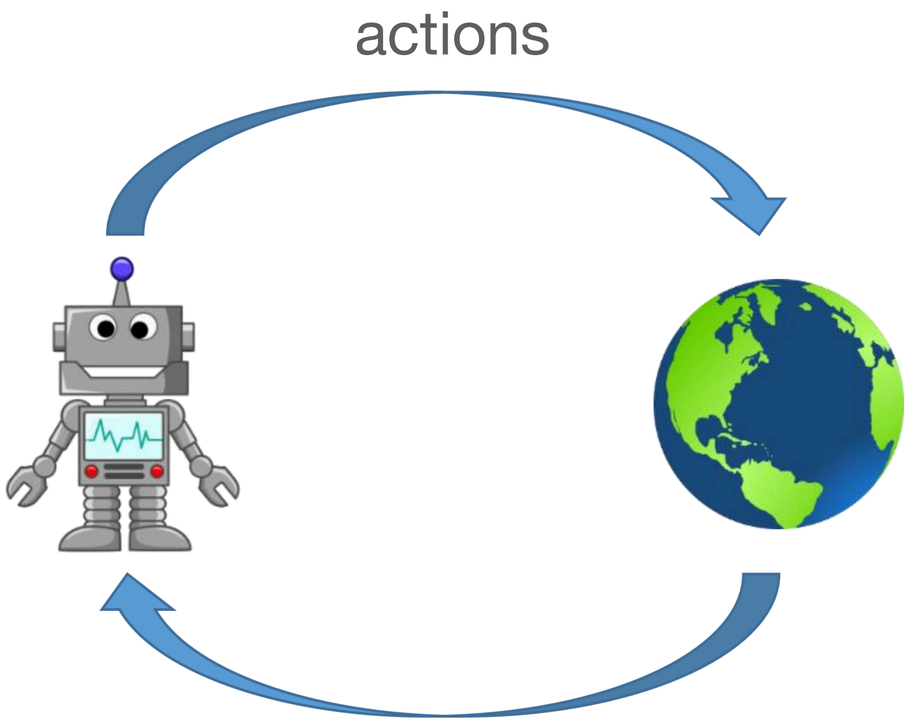
can be learned from data, and

is well-shaped.

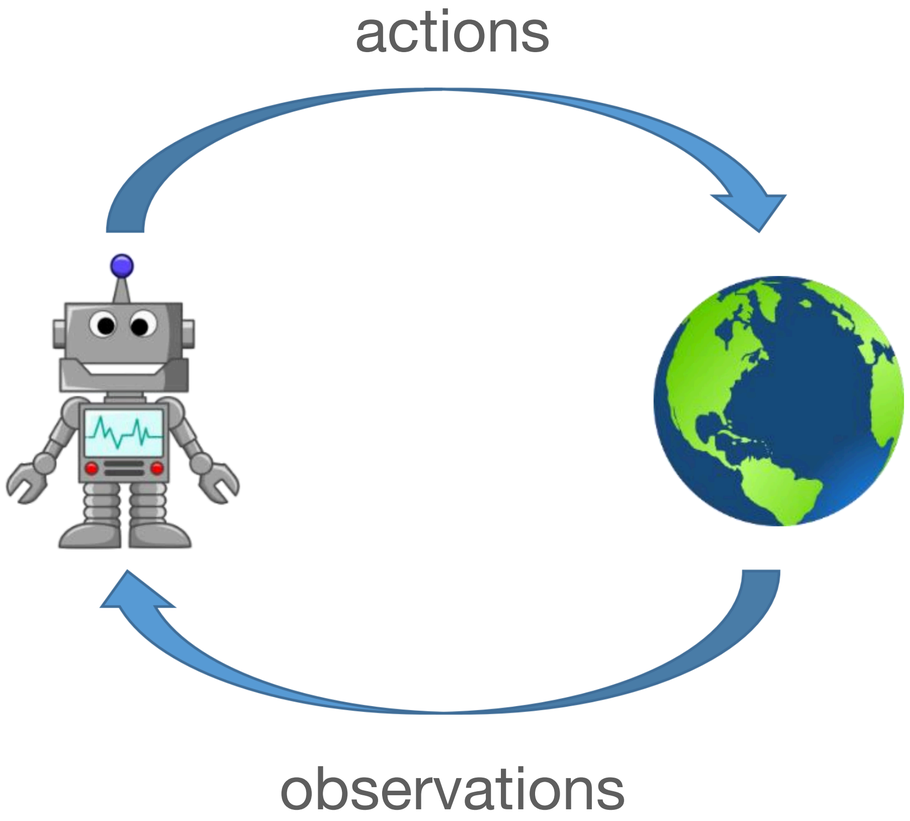
The Reinforcement Learning Loop



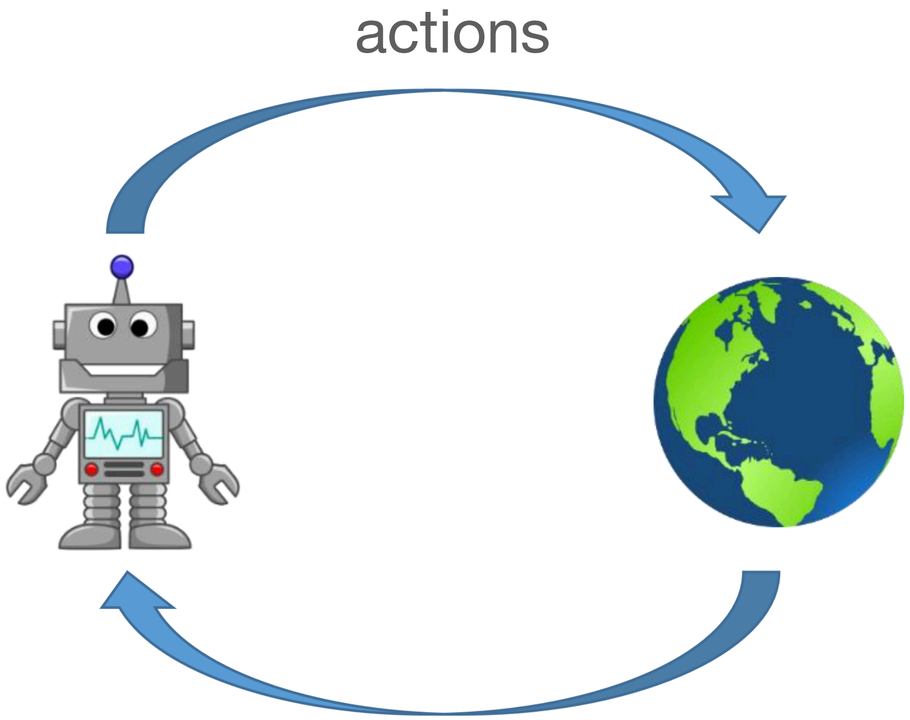
The Reinforcement Learning Loop



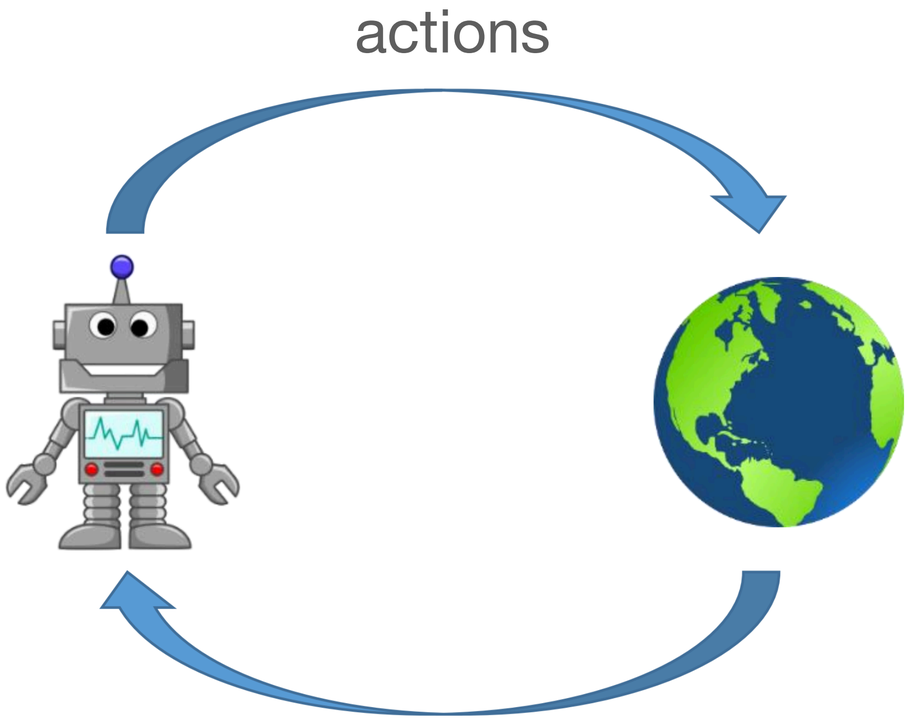
The Reinforcement Learning Loop



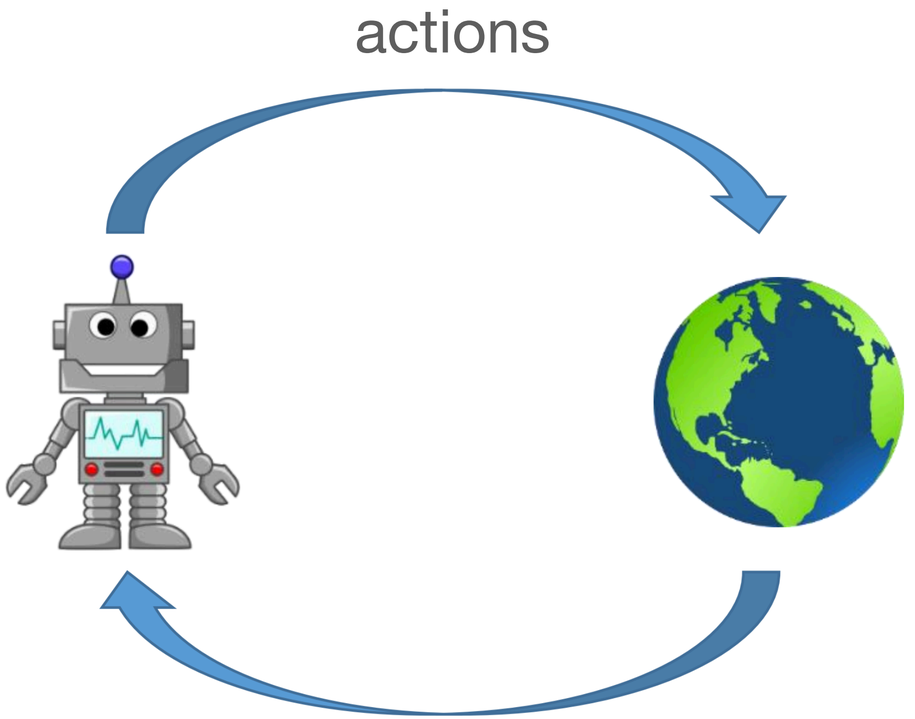
The Reinforcement Learning Loop



The Reinforcement Learning Loop



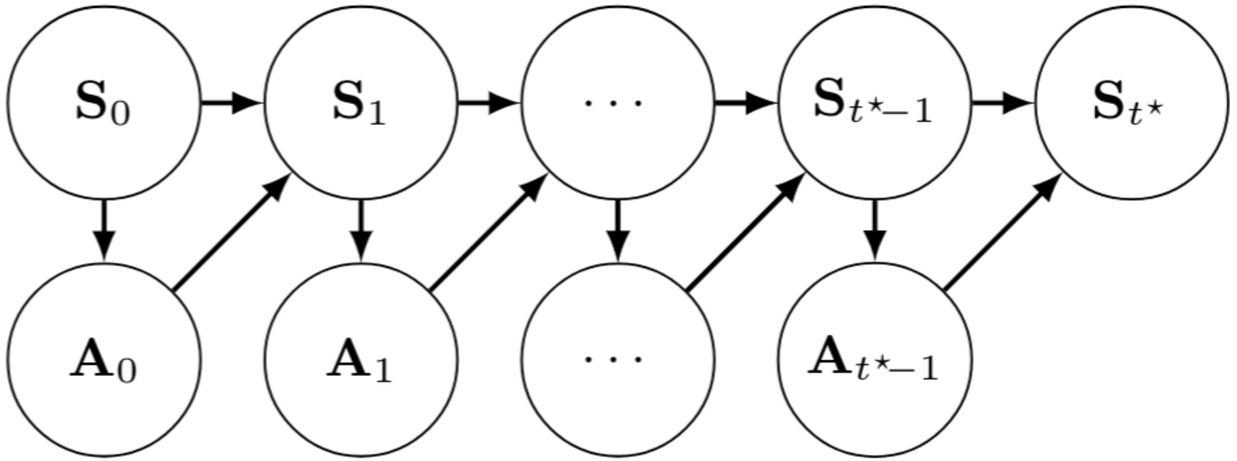
The Reinforcement Learning Loop



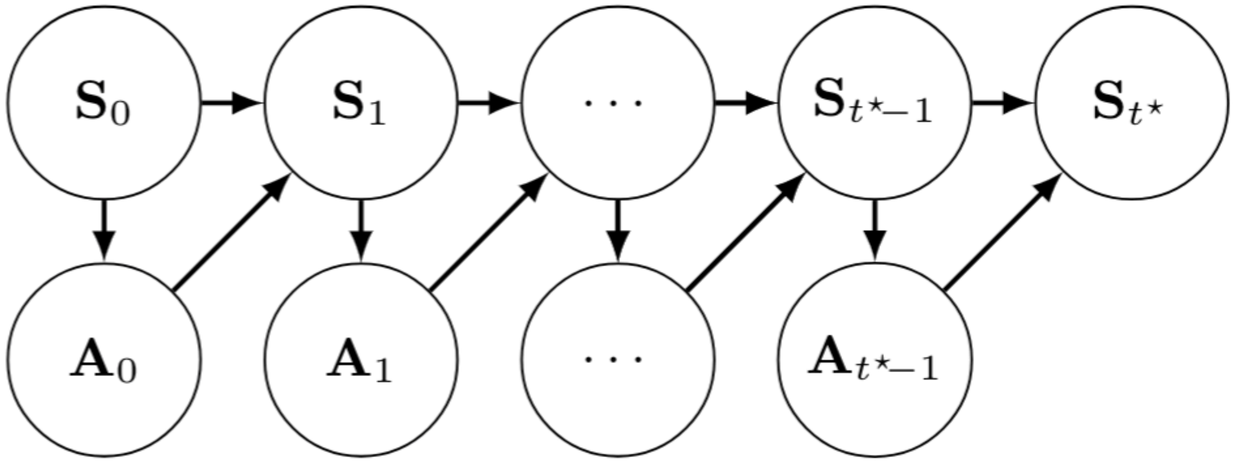
desired outcome

$$\mathbf{S} = \mathbf{g}$$

Inferring Outcome-Driven Trajectory Distributions



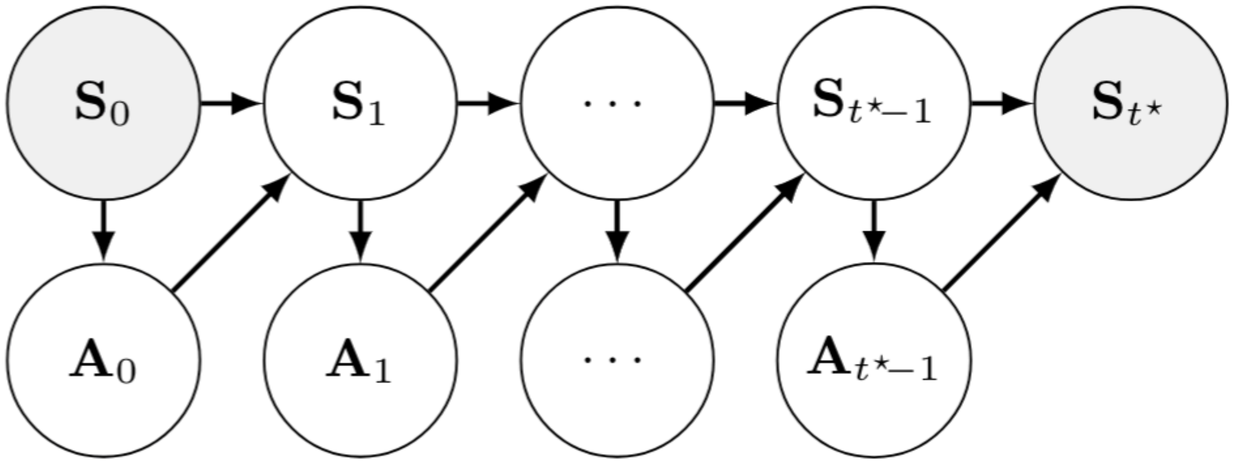
Inferring Outcome-Driven Trajectory Distributions



$$\mathbf{S}_{t^\star} = \mathbf{g}$$

$t^\star$ is known

Inferring Outcome-Driven Trajectory Distributions

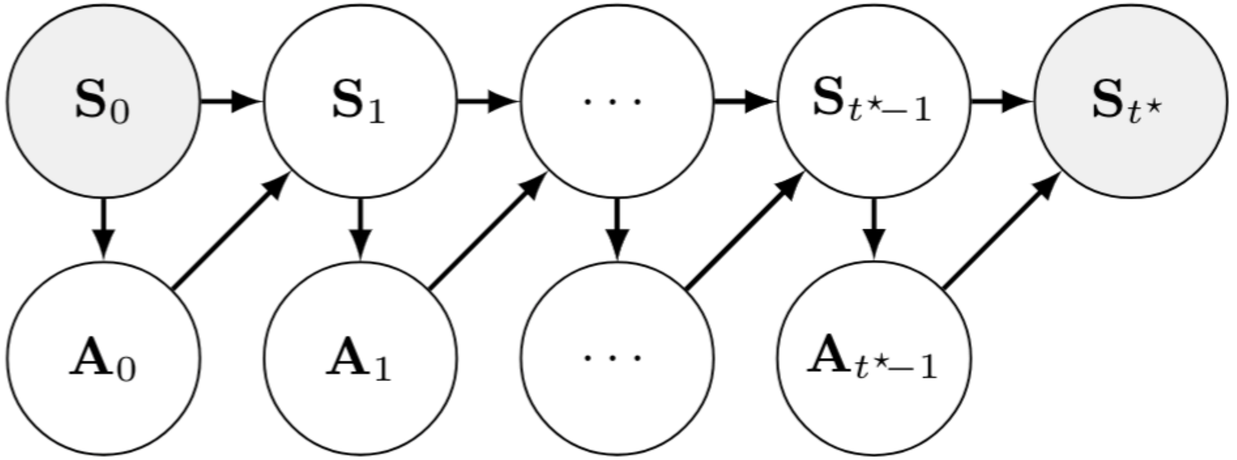


$$\mathbf{S}_{t^\star} = \mathbf{g}$$

$t^\star$ is known

Inferring Outcome-Driven Trajectory Distributions

$$\tilde{\boldsymbol{\tau}}_{0:t} \doteq \{\mathbf{a}_0, \mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_t, \mathbf{a}_t\} \quad t \doteq t^\star - 1$$

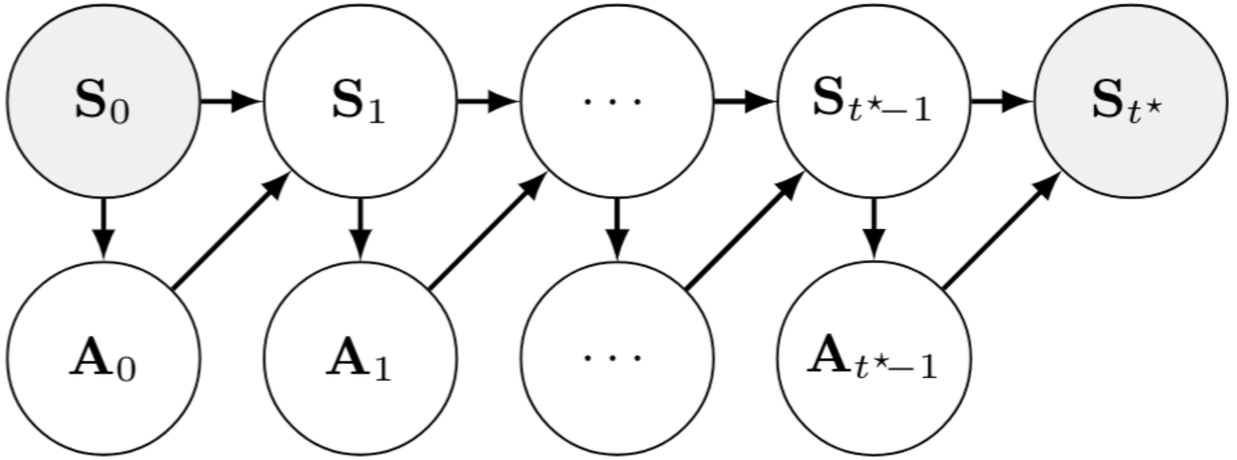


$$\mathbf{S}_{t^\star} = \mathbf{g}$$

$t^\star$ is known

Inferring Outcome-Driven Trajectory Distributions

$$\tilde{\boldsymbol{\tau}}_{0:t} \doteq \{\mathbf{a}_0, \mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_t, \mathbf{a}_t\} \quad t \doteq t^\star - 1$$



$\mathbf{S}_{t^\star} = \mathbf{g}$

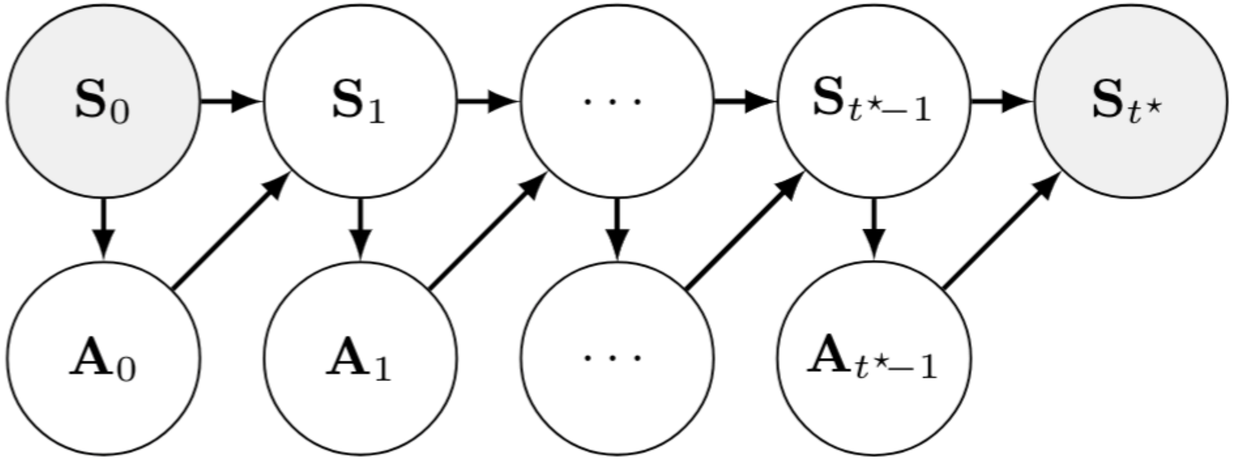
$t^\star$ is known

Goal

- Find a state-action trajectory distribution leading to $\mathbf{S}_{t^\star} = \mathbf{g}$
- Infer conditional distribution $p_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{S}_0,\mathbf{S}_{t^\star}}(\cdot \,|\, \mathbf{s}_0, \mathbf{g})$

Inferring Outcome-Driven Trajectory Distributions

$$\tilde{\boldsymbol{\tau}}_{0:t} \doteq \{\mathbf{a}_0, \mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_t, \mathbf{a}_t\} \quad t \doteq t^\star - 1$$



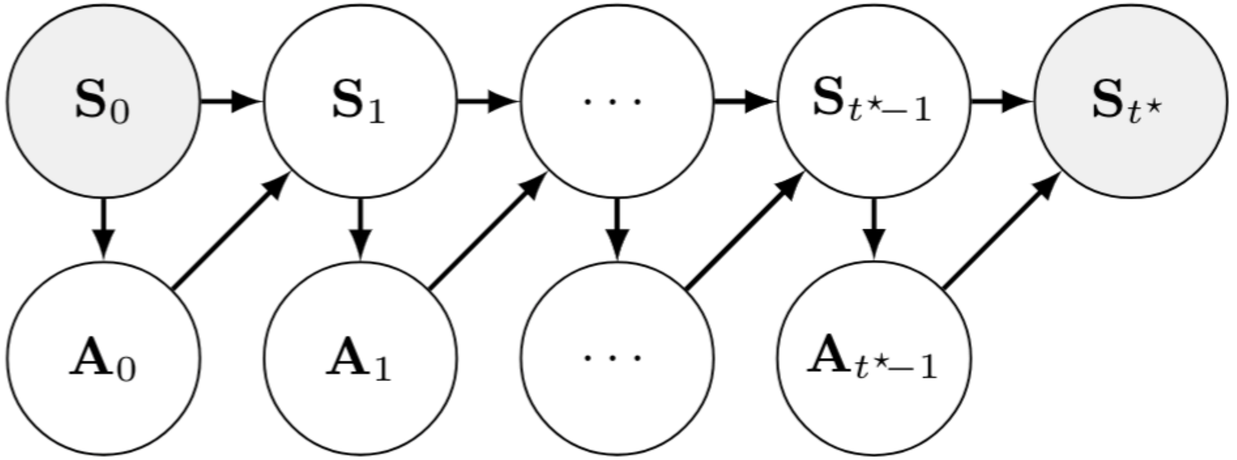
$$\mathbf{S}_{t^\star} = \mathbf{g}$$

$t^\star$ is known

How?

Inferring Outcome-Driven Trajectory Distributions

$$\tilde{\boldsymbol{\tau}}_{0:t} \doteq \{\mathbf{a}_0, \mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_t, \mathbf{a}_t\} \quad t \doteq t^\star - 1$$



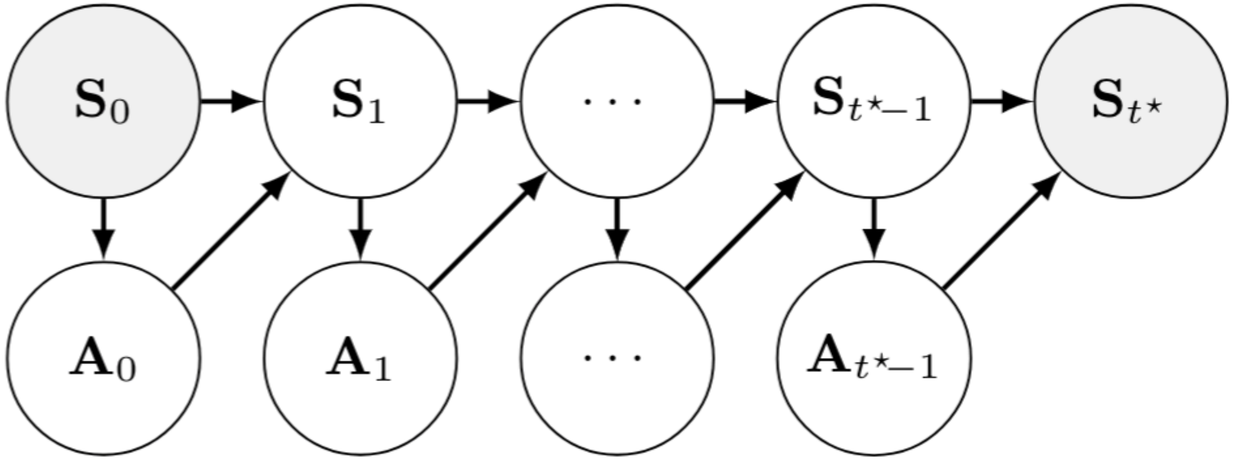
$\mathbf{S}_{t^\star} = \mathbf{g}$

$t^\star$ is known

How? $\mathbb{I}\{\mathbf{S}_{t^\star} = \mathbf{g}\}$ **?**

Inferring Outcome-Driven Trajectory Distributions

$$\tilde{\boldsymbol{\tau}}_{0:t} \doteq \{\mathbf{a}_0, \mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_t, \mathbf{a}_t\} \quad t \doteq t^\star - 1$$



$$\mathbf{S}_{t^\star} = \mathbf{g}$$

$t^\star$ is known

How? ~~$\mathbb{I}\{\mathbf{S}_{t^\star} = \mathbf{g}\}$~~**?**

- Solve variational problem

$$\min_{q_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0} \in \hat{\mathcal{Q}}} D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0}(\cdot \,|\, \mathbf{s}_0) \,\|\, p_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0, \mathbf{s}_{t^\star}}(\cdot \,|\, \mathbf{s}_0, \mathbf{g}))$$

Inferring Outcome-Driven Trajectory Distributions

Problem:

$$\min_{q_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0} \in \hat{\mathcal{Q}}} D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0}(\cdot \mid \mathbf{s}_0) \parallel p_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0, \mathbf{s}_{t^\star}}(\cdot \mid \mathbf{s}_0, \mathbf{g}))$$

Inferring Outcome-Driven Trajectory Distributions

Problem:

$$\min_{q_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0} \in \hat{\mathcal{Q}}} D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0}(\cdot \mid \mathbf{s}_0) \parallel p_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0, \mathbf{s}_{t^\star}}(\cdot \mid \mathbf{s}_0, \mathbf{g}))$$

- Intractable!

Inferring Outcome-Driven Trajectory Distributions

Problem:

$$\min_{q_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0} \in \hat{\mathcal{Q}}} D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0}(\cdot \mid \mathbf{s}_0) \parallel p_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0, \mathbf{s}_{t^\star}}(\cdot \mid \mathbf{s}_0, \mathbf{g}))$$

- Intractable!

Instead:

- Derive an **equivalent tractable variational objective**

Outcome-Driven Variational Objective (Finite Horizon)

Deriving a tractable variational lower bound

Outcome-Driven Variational Objective (Finite Horizon)

Deriving a tractable variational lower bound

- Define variational distribution

$$q_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t} \mid \mathbf{s}_0) \doteq \pi(\mathbf{a}_t \mid \mathbf{s}_t) \prod_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1} \mid \mathbf{s}_{t'}, \mathbf{a}_{t'}) \pi(\mathbf{a}_{t'} \mid \mathbf{s}_{t'})$$

Outcome-Driven Variational Objective (Finite Horizon)

Deriving a tractable variational lower bound

- Define variational distribution

$$q_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t} \mid \mathbf{s}_0) \doteq \pi(\mathbf{a}_t \mid \mathbf{s}_t) \prod_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1} \mid \mathbf{s}_{t'}, \mathbf{a}_{t'}) \pi(\mathbf{a}_{t'} \mid \mathbf{s}_{t'})$$

- Then:

$$D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0}(\cdot \mid \mathbf{s}_0) \parallel p_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0,\mathbf{s}_{t^\star}}(\cdot \mid \mathbf{s}_0, \mathbf{g})) = -\bar{\mathcal{F}}(\pi, \mathbf{s}_0, \mathbf{g}) + \log p(\mathbf{g}|\mathbf{s}_0)$$

Outcome-Driven Variational Objective (Finite Horizon)

Deriving a tractable variational lower bound

- Define variational distribution

$$q_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t} \mid \mathbf{s}_0) \doteq \pi(\mathbf{a}_t \mid \mathbf{s}_t) \prod_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1} \mid \mathbf{s}_{t'}, \mathbf{a}_{t'}) \pi(\mathbf{a}_{t'} \mid \mathbf{s}_{t'})$$

- Then:

$$D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0}(\cdot \mid \mathbf{s}_0) \parallel p_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0,\mathbf{s}_{t^\star}}(\cdot \mid \mathbf{s}_0, \mathbf{g})) = -\bar{\mathcal{F}}(\pi, \mathbf{s}_0, \mathbf{g}) + \log p(\mathbf{g}|\mathbf{s}_0)$$

with

$$\bar{\mathcal{F}}(\pi, \mathbf{s}_0, \mathbf{g}) \doteq \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t} \mid \mathbf{s}_0)} \left[\log p_d(\mathbf{g} \mid \mathbf{s}_t, \mathbf{a}_t) + \sum_{t'=0}^{t-1} D_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{s}_{t'}) \,||\, p(\cdot \mid \mathbf{s}_{t'})) \right]$$

Outcome-Driven Variational Objective (Finite Horizon)

Deriving a tractable variational lower bound

- Define variational distribution

$$q_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t} \mid \mathbf{s}_0) \doteq \pi(\mathbf{a}_t \mid \mathbf{s}_t) \prod_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1} \mid \mathbf{s}_{t'}, \mathbf{a}_{t'}) \pi(\mathbf{a}_{t'} \mid \mathbf{s}_{t'})$$

- **Hence**:

$$\arg\min_{\pi \in \Pi} D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0}(\cdot \mid \mathbf{s}_0) \,\|\, p_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0, \mathbf{s}_{t^\star}}(\cdot \mid \mathbf{s}_0, \mathbf{g})) = \arg\max_{\pi \in \Pi} \bar{\mathcal{F}}(\pi, \mathbf{s}_0, \mathbf{g}).$$

with

$$\bar{\mathcal{F}}(\pi, \mathbf{s}_0, \mathbf{g}) \doteq \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_{0:t}|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t} \mid \mathbf{s}_0)} \left[\log p_d(\mathbf{g} \mid \mathbf{s}_t, \mathbf{a}_t) + \sum_{t'=0}^{t-1} D_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{s}_{t'}) \,||\, p(\cdot \mid \mathbf{s}_{t'})) \right]$$

Outcome-Driven Variational Objective (Finite Horizon)

What reward function is this objective optimizing?

- Variational objective

$$\bar{\mathcal{F}}(\pi, \mathbf{s}_0, \mathbf{g}) \doteq \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_{0:t} | \mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t} \mid \mathbf{s}_0)} \left[\log p_d(\mathbf{g} \mid \mathbf{s}_t, \mathbf{a}_t) + \sum_{t'=0}^{t-1} D_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{s}_{t'}) \,||\, p(\cdot \mid \mathbf{s}_{t'})) \right]$$

Outcome-Driven Variational Objective (Finite Horizon)

What reward function is this objective optimizing?

- Variational objective

$$\bar{\mathcal{F}}(\pi, \mathbf{s}_0, \mathbf{g}) \doteq \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_{0:t} \mid \mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t} \mid \mathbf{s}_0)} \left[\log p_d(\mathbf{g} \mid \mathbf{s}_t, \mathbf{a}_t) + \sum_{t'=0}^{t-1} D_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{s}_{t'}) \,||\, p(\cdot \mid \mathbf{s}_{t'})) \right]$$

- Equivalent to KL-regularized RL with rewards

$$r(\mathbf{s}_{t'}, \mathbf{a}_{t'}, \mathbf{g}, t') \doteq \mathbb{I}\{t' = t\} \log p_d(\mathbf{g} \mid \mathbf{s}_{t'}, \mathbf{a}_{t'})$$

Outcome-Driven Variational Objective (Finite Horizon)

What reward function is this objective optimizing?

- Variational objective

$$\bar{\mathcal{F}}(\pi, \mathbf{s}_0, \mathbf{g}) \doteq \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_{0:t} \mid \mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t} \mid \mathbf{s}_0)} \left[\log p_d(\mathbf{g} \mid \mathbf{s}_t, \mathbf{a}_t) + \sum_{t'=0}^{t-1} D_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{s}_{t'}) \,||\, p(\cdot \mid \mathbf{s}_{t'})) \right]$$

- Equivalent to KL-regularized RL with rewards

$$r(\mathbf{s}_{t'}, \mathbf{a}_{t'}, \mathbf{g}, t') \doteq \mathbb{I}\{t' = t\} \log p_d(\mathbf{g} \mid \mathbf{s}_{t'}, \mathbf{a}_{t'})$$

- Sparse rewards given by transition dynamics

Infinite-Horizon Outcome-Driven MDP

***But*: We don't care *when* the outcome is achieved**

$$p_{\tilde{\boldsymbol{\tau}}_{0:T},\mathbf{s}_{T^\star},T|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t},\mathbf{g},t\,|\,\mathbf{s}_0) = p_T(t)p_d(\mathbf{g}\,|\,\mathbf{s}_t,\mathbf{a}_t)p(\mathbf{a}_t\,|\,\mathbf{s}_t)\prod_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1}\,|\,\mathbf{s}_{t'},\mathbf{a}_{t'})p(\mathbf{a}_{t'}\,|\,\mathbf{s}_{t'})$$

Infinite-Horizon Outcome-Driven MDP

<u>*But*</u>: We don't care *when* the outcome is achieved

- Treat **termination time** as a **random variable**:

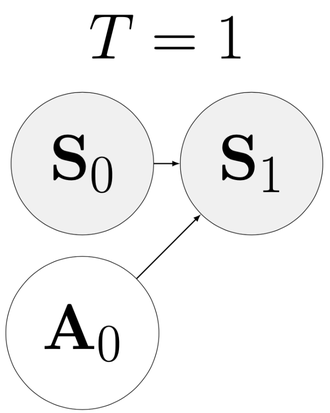
$$p_{\tilde{\boldsymbol{\tau}}_{0:T},\mathbf{s}_{T^\star},T|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t},\mathbf{g},t\,|\,\mathbf{s}_0) = p_T(t)p_d(\mathbf{g}\,|\,\mathbf{s}_t,\mathbf{a}_t)p(\mathbf{a}_t\,|\,\mathbf{s}_t)\prod_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1}\,|\,\mathbf{s}_{t'},\mathbf{a}_{t'})p(\mathbf{a}_{t'}\,|\,\mathbf{s}_{t'})$$

Infinite-Horizon Outcome-Driven MDP

<u>*But*</u>: We don't care *when* the outcome is achieved

- Treat **termination time** as a **random variable**:

$$p_{\tilde{\boldsymbol{\tau}}_{0:T},\mathbf{s}_{T^\star},T|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t},\mathbf{g},t\,|\,\mathbf{s}_0) = p_T(t)p_d(\mathbf{g}\,|\,\mathbf{s}_t,\mathbf{a}_t)p(\mathbf{a}_t\,|\,\mathbf{s}_t)\prod_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1}\,|\,\mathbf{s}_{t'},\mathbf{a}_{t'})p(\mathbf{a}_{t'}\,|\,\mathbf{s}_{t'})$$

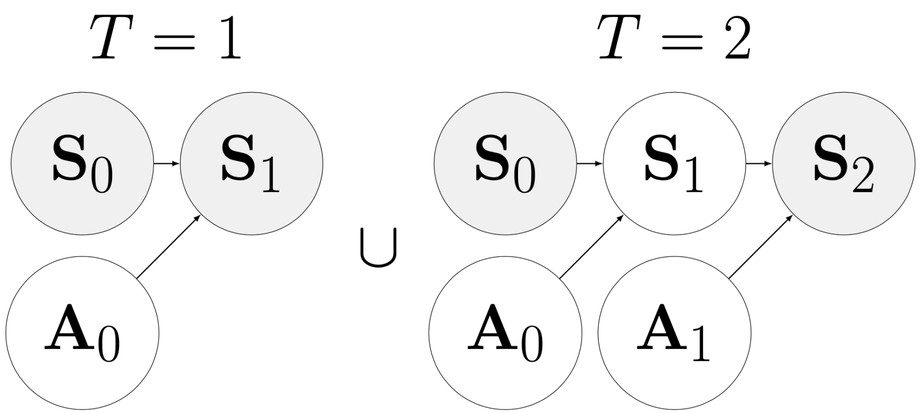


Infinite-Horizon Outcome-Driven MDP

But: We don't care _when_ the outcome is achieved

- Treat **termination time** as a **random variable**:

$$p_{\tilde{\boldsymbol{\tau}}_{0:T},\mathbf{s}_{T^\star},T|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t},\mathbf{g},t\,|\,\mathbf{s}_0) = p_T(t)p_d(\mathbf{g}\,|\,\mathbf{s}_t,\mathbf{a}_t)p(\mathbf{a}_t\,|\,\mathbf{s}_t)\prod_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1}\,|\,\mathbf{s}_{t'},\mathbf{a}_{t'})p(\mathbf{a}_{t'}\,|\,\mathbf{s}_{t'})$$

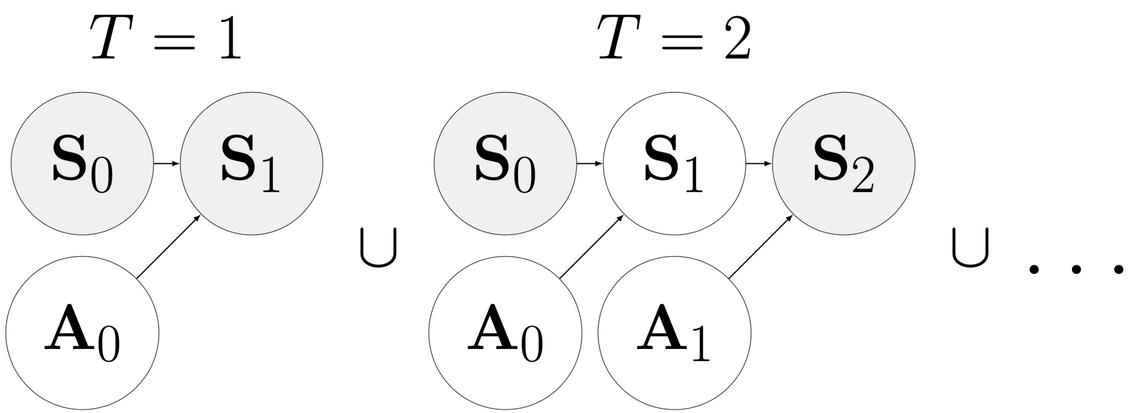


Infinite-Horizon Outcome-Driven MDP

But: We don't care _when_ the outcome is achieved

- Treat **termination time** as a **random variable**:

$$p_{\tilde{\boldsymbol{\tau}}_{0:T},\mathbf{s}_{T^\star},T|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t},\mathbf{g},t\,|\,\mathbf{s}_0) = p_T(t)p_d(\mathbf{g}\,|\,\mathbf{s}_t,\mathbf{a}_t)p(\mathbf{a}_t\,|\,\mathbf{s}_t)\prod_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1}\,|\,\mathbf{s}_{t'},\mathbf{a}_{t'})p(\mathbf{a}_{t'}\,|\,\mathbf{s}_{t'})$$



Inference in an Infinite-Horizon Outcome-Driven MDP

But: We don't care *when* the outcome is achieved

- Treat **termination time** as a **random variable**:

$$p_{\tilde{\boldsymbol{\tau}}_{0:T},\mathbf{s}_{T^\star},T|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t},\mathbf{g},t\,|\,\mathbf{s}_0) = p_T(t)p_d(\mathbf{g}\,|\,\mathbf{s}_t,\mathbf{a}_t)p(\mathbf{a}_t\,|\,\mathbf{s}_t)\prod_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1}\,|\,\mathbf{s}_{t'},\mathbf{a}_{t'})p(\mathbf{a}_{t'}\,|\,\mathbf{s}_{t'})$$

- Perform inference over **states**, **actions**, and the **termination time**

$$\min_{q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}\in\mathcal{Q}} D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\cdot\,|\,\mathbf{s}_0)\,\|\,p_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0,\mathbf{s}_{T^\star}}(\cdot\,|\,\mathbf{s}_0,\mathbf{g}))$$

Outcome-Driven Variational Objective

Infinite-horizon variational objective

- Define variational distribution

Outcome-Driven Variational Objective

Infinite-horizon variational objective

- Define variational distribution

$$q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t},t\,|\,\mathbf{s}_0) = q_{\tilde{\boldsymbol{\tau}}_{0:T}|T,\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t,\mathbf{s}_0)q_T(t)$$

Outcome-Driven Variational Objective

Infinite-horizon variational objective

- Define variational distribution

$$q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t},t\,|\,\mathbf{s}_0)=q_{\tilde{\boldsymbol{\tau}}_{0:T}|T,\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t,\mathbf{s}_0)q_T(t)$$

- Then:

$$D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\cdot\,|\,\mathbf{s}_0)\,\|\,p_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0,\mathbf{s}_{T^\star}}(\cdot\,|\,\mathbf{s}_0,\mathbf{g}))=\log p(\mathbf{g}|\mathbf{s}_0)-\mathcal{F}(\pi,q_T,\mathbf{s}_0,\mathbf{g})$$

with

$$\mathcal{F}(\pi,q_T,\mathbf{s}_0,\mathbf{g})$$

$$\doteq\sum_{t=0}^{\infty}q_T(t)\,\mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_{0:T}|T,\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t}\,|\,t,\mathbf{s}_0)}\Big[\log p_d(\mathbf{g}\,|\,\mathbf{s}_t,\mathbf{a}_t)-D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\cdot\,|\,\mathbf{s}_0)\,||\,p_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\cdot\,|\,\mathbf{s}_0))\Big]$$

Outcome-Driven Variational Objective

Infinite-horizon variational objective

- Define variational distribution

$$q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t}, t \,|\, \mathbf{s}_0) = q_{\tilde{\boldsymbol{\tau}}_{0:T}|T,\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t} \,|\, t, \mathbf{s}_0) q_T(t)$$

- Then:

$$D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\cdot \,|\, \mathbf{s}_0) \,\|\, p_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0,\mathbf{s}_{T^\star}}(\cdot \,|\, \mathbf{s}_0, \mathbf{g})) = \log p(\mathbf{g}|\mathbf{s}_0) - \mathcal{F}(\pi, q_T, \mathbf{s}_0, \mathbf{g})$$

with

$$\mathcal{F}(\pi, q_T, \mathbf{s}_0, \mathbf{g})$$

$$\doteq \sum_{t=0}^{\infty} q_T(t)\, \mathbb{E}_{q_{\tilde{\boldsymbol{\tau}}_{0:T}|T,\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t} \,|\, t,\mathbf{s}_0)} \Big[\log p_d(\mathbf{g} \,|\, \mathbf{s}_t, \mathbf{a}_t) - D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\cdot \,|\, \mathbf{s}_0) \,||\, p_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\cdot \,|\, \mathbf{s}_0)) \Big]$$

not factorized

Outcome-Driven Variational Distribution

Infinite-horizon variational objective

- Define variational distribution

$$q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t}, t \,|\, \mathbf{s}_0) = q_{\tilde{\boldsymbol{\tau}}_{0:T}|T,\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t} \,|\, t, \mathbf{s}_0) q_T(t)$$

$$q_T(t) = q_{\Delta_{t+1}}(\Delta_{t+1} = 1) \prod_{t'=1}^{t} q_{\Delta_{t'}}(\Delta_{t'} = 0)$$

- Goal: Derive recursive variational objective

Outcome-Driven Variational Inference As TD Learning

Theorem 1. (informal)

Outcome-Driven Variational Inference As TD Learning

Theorem 1. (informal)

- Define variational distributions $q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t}, t \,|\, \mathbf{s}_0)$ and $q_T(t)$

Outcome-Driven Variational Inference As TD Learning

Theorem 1. (informal)

- Define variational distributions $q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t},t\,|\,\mathbf{s}_0)$ and $q_T(t)$

- Then

$$V^\pi(\mathbf{s}_t,\mathbf{g};q_T) \doteq \mathbb{E}_{\pi(\mathbf{a}_t\,|\,\mathbf{s}_t)}[Q^\pi(\mathbf{s}_t,\mathbf{a}_t,\mathbf{g};q_T)] - D_{\mathrm{KL}}(\pi(\cdot\,|\,\mathbf{s}_t)\,\|\,p(\cdot\,|\,\mathbf{s}_t))$$

Outcome-Driven Variational Inference As TD Learning

Theorem 1. (informal)

- Define variational distributions $q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t}, t \,|\, \mathbf{s}_0)$ and $q_T(t)$

- Then

$$V^\pi(\mathbf{s}_t, \mathbf{g}; q_T) \doteq \mathbb{E}_{\pi(\mathbf{a}_t \,|\, \mathbf{s}_t)}[Q^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_T)] - D_{\mathrm{KL}}(\pi(\cdot \,|\, \mathbf{s}_t) \,\|\, p(\cdot \,|\, \mathbf{s}_t))$$

$$Q^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_T) \doteq r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_\Delta) + q_{\Delta_{t+1}}(\Delta_{t+1} = 0)\, \mathbb{E}_{p_d(\mathbf{s}_{t+1} \,|\, \mathbf{s}_t, \mathbf{a}_t)}[V^\pi(\mathbf{s}_{t+1}, \mathbf{g}; \pi, q_T)]$$

Outcome-Driven Variational Inference As TD Learning

Theorem 1. (informal)

- Define variational distributions $q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t}, t \,|\, \mathbf{s}_0)$ and $q_T(t)$

- Then

$$V^\pi(\mathbf{s}_t, \mathbf{g}; q_T) \doteq \mathbb{E}_{\pi(\mathbf{a}_t \,|\, \mathbf{s}_t)}[Q^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_T)] - D_{\mathrm{KL}}(\pi(\cdot \,|\, \mathbf{s}_t) \,\|\, p(\cdot \,|\, \mathbf{s}_t))$$

$$Q^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_T) \doteq r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_\Delta) + q_{\Delta_{t+1}}(\Delta_{t+1} = 0)\, \mathbb{E}_{p_d(\mathbf{s}_{t+1} \,|\, \mathbf{s}_t, \mathbf{a}_t)}[V^\pi(\mathbf{s}_{t+1}, \mathbf{g}; \pi, q_T)]$$

$$r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_\Delta) \doteq q_{\Delta_{t+1}}(\Delta_{t+1} = 1) \log p_d(\mathbf{g} \,|\, \mathbf{s}_t, \mathbf{a}_t) - D_{\mathrm{KL}}(q_{\Delta_{t+1}} \,\|\, p_{\Delta_{t+1}})$$

Outcome-Driven Variational Inference As TD Learning

Theorem 1. (informal)

- Define variational distributions $q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\tilde{\boldsymbol{\tau}}_{0:t}, t \,|\, \mathbf{s}_0)$ and $q_T(t)$

- Then

$$V^\pi(\mathbf{s}_t, \mathbf{g}; q_T) \doteq \mathbb{E}_{\pi(\mathbf{a}_t \,|\, \mathbf{s}_t)}[Q^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_T)] - D_{\mathrm{KL}}(\pi(\cdot \,|\, \mathbf{s}_t) \,\|\, p(\cdot \,|\, \mathbf{s}_t))$$

$$Q^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_T) \doteq r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_\Delta) + q_{\Delta_{t+1}}(\Delta_{t+1} = 0)\, \mathbb{E}_{p_d(\mathbf{s}_{t+1} \,|\, \mathbf{s}_t, \mathbf{a}_t)}[V^\pi(\mathbf{s}_{t+1}, \mathbf{g}; \pi, q_T)]$$

$$r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_\Delta) \doteq q_{\Delta_{t+1}}(\Delta_{t+1} = 1) \log p_d(\mathbf{g} \,|\, \mathbf{s}_t, \mathbf{a}_t) - D_{\mathrm{KL}}(q_{\Delta_{t+1}} \,\|\, p_{\Delta_{t+1}})$$

with

$$D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\cdot \,|\, \mathbf{s}_0) \,\|\, p_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0,\mathbf{s}_{T^*}}(\cdot \,|\, \mathbf{s}_0, \mathbf{g})) = -V^\pi(\mathbf{s}_0, \mathbf{g}; q_T) + \log p(\mathbf{g} \,|\, \mathbf{s}_0)$$

A Lower Bound on the Log Marginal Likelihood

What does this mean?

A Lower Bound on the Log Marginal Likelihood

What does this mean?

- Since

$$D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\cdot\,|\,\mathbf{s}_0)\,\|\,p_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0,\mathbf{s}_{T^*}}(\cdot\,|\,\mathbf{s}_0,\mathbf{g})) = -V^{\pi}(\mathbf{s}_0,\mathbf{g};q_T) + \log p(\mathbf{g}\,|\,\mathbf{s}_0),$$

the **value function** is a **variational lower bound** on $\log p(\mathbf{g}\,|\,\mathbf{s}_0)$

A Lower Bound on the Log Marginal Likelihood

What does this mean?

- Since

$$D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\cdot\,|\,\mathbf{s}_0)\,\|\,p_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0,\mathbf{s}_{T^*}}(\cdot\,|\,\mathbf{s}_0,\mathbf{g})) = -V^{\pi}(\mathbf{s}_0,\mathbf{g};q_T) + \log p(\mathbf{g}\,|\,\mathbf{s}_0),$$

the **value function** is a **variational lower bound** on $\log p(\mathbf{g}\,|\,\mathbf{s}_0)$ and

$$\underset{\pi\in\Pi,q_T\in\mathcal{Q}_T}{\arg\min}\ \{D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\cdot\,|\,\mathbf{s}_0)\,\|\,p_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0,\mathbf{s}_{T^\star}}(\cdot\,|\,\mathbf{s}_0,\mathbf{g}))\}$$

A Lower Bound on the Log Marginal Likelihood

What does this mean?

- Since

$$D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\cdot\,|\,\mathbf{s}_0)\,\|\,p_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0,\mathbf{s}_{T^*}}(\cdot\,|\,\mathbf{s}_0,\mathbf{g})) = -V^{\pi}(\mathbf{s}_0,\mathbf{g};q_T) + \log p(\mathbf{g}\,|\,\mathbf{s}_0),$$

the **value function** is a **variational lower bound** on $\log p(\mathbf{g}\,|\,\mathbf{s}_0)$ and

$$\begin{aligned} &\underset{\pi\in\Pi,q_T\in\mathcal{Q}_T}{\arg\min}\ \{D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\cdot\,|\,\mathbf{s}_0)\,\|\,p_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0,\mathbf{s}_{T^\star}}(\cdot\,|\,\mathbf{s}_0,\mathbf{g}))\} \\ =\ &\underset{\pi\in\Pi,q_T\in\mathcal{Q}_T}{\arg\max}\ \mathcal{F}(\pi,q_T,\mathbf{s}_0,\mathbf{g}) \end{aligned}$$

A Lower Bound on the Log Marginal Likelihood

What does this mean?

- Since

$$D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\cdot\,|\,\mathbf{s}_0)\,\|\,p_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0,\mathbf{s}_{T^*}}(\cdot\,|\,\mathbf{s}_0,\mathbf{g})) = -V^{\pi}(\mathbf{s}_0,\mathbf{g};q_T) + \log p(\mathbf{g}\,|\,\mathbf{s}_0),$$

the **value function** is a **variational lower bound** on $\log p(\mathbf{g}\,|\,\mathbf{s}_0)$ and

$$\begin{aligned}
&\underset{\pi\in\Pi,q_T\in\mathcal{Q}_T}{\arg\min}\ \{D_{\mathrm{KL}}(q_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0}(\cdot\,|\,\mathbf{s}_0)\,\|\,p_{\tilde{\boldsymbol{\tau}}_{0:T},T|\mathbf{s}_0,\mathbf{s}_{T^\star}}(\cdot\,|\,\mathbf{s}_0,\mathbf{g}))\}\\
=&\underset{\pi\in\Pi,q_T\in\mathcal{Q}_T}{\arg\max}\ \mathcal{F}(\pi,q_T,\mathbf{s}_0,\mathbf{g})\\
=&\underset{\pi\in\Pi,q_T\in\mathcal{Q}_T}{\arg\max}\ V^{\pi}(\mathbf{s}_0,\mathbf{g};q_T).
\end{aligned}$$

Outcome-Driven Bellman Operator

$$\mathcal{T}^{\pi} Q(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_T) \doteq r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_\Delta) + q_{\Delta_{t+1}}(\Delta_{t+1} = 0)\, \mathbb{E}_{p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)} \left[V(\mathbf{s}_{t+1}, \mathbf{g}; q_T) \right]$$

with

$$V(\mathbf{s}_t, \mathbf{g}; q_T) \doteq \mathbb{E}_{\pi(\mathbf{a}_t \mid \mathbf{s}_t)} \left[Q(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_T) \right] + D_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{s}_t) \,\|\, p(\cdot \mid \mathbf{s}_t))$$

$$r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_\Delta) \doteq q_{\Delta_{t+1}}(\Delta_{t+1} = 1) \log p_d(\mathbf{g} \mid \mathbf{s}_t, \mathbf{a}_t) - D_{\mathrm{KL}}(q_{\Delta_{t+1}} \,\|\, p_{\Delta_{t+1}})$$

Outcome-Driven Bellman Operator

$$\mathcal{T}^{\pi} Q(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_T) \doteq r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_\Delta) + q_{\Delta_{t+1}}(\Delta_{t+1} = 0)\, \mathbb{E}_{p_d(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)} \left[V(\mathbf{s}_{t+1}, \mathbf{g}; q_T) \right]$$

with

$$V(\mathbf{s}_t, \mathbf{g}; q_T) \doteq \mathbb{E}_{\pi(\mathbf{a}_t \mid \mathbf{s}_t)} \left[Q(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_T) \right] + D_{\mathrm{KL}}(\pi(\cdot \mid \mathbf{s}_t) \,\|\, p(\cdot \mid \mathbf{s}_t))$$

$$r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_\Delta) \doteq q_{\Delta_{t+1}}(\Delta_{t+1} = 1) \log p_d(\mathbf{g} \mid \mathbf{s}_t, \mathbf{a}_t) - D_{\mathrm{KL}}(q_{\Delta_{t+1}} \,\|\, p_{\Delta_{t+1}})$$

- **Dense, shaped reward function + dynamic discount factor**

Outcome-Driven Policy Iteration

Theorem 2. (informal)

- Assume that the MDP is ergodic and set of available action is **finite**.

Outcome-Driven Policy Iteration

Theorem 2. (informal)

- Assume that the MDP is ergodic and set of available action is **finite**.
- Consider the **outcome-driven Bellman operator**.

Outcome-Driven Policy Iteration

Theorem 2. (informal)

- Assume that the MDP is ergodic and set of available action is **finite**.
- Consider the **outcome-driven Bellman operator**.
- Consider the **optimal variational distribution** over the termination time

Outcome-Driven Policy Iteration

Theorem 2. (informal)

- Assume that the MDP is ergodic and set of available action is **finite**.
- Consider the **outcome-driven Bellman operator**.
- Consider the **optimal variational distribution** over the termination time
- Alternating between **policy evaluation** and **policy improvement** will result in an **optimal policy**

Outcome-Driven Actor Critic

Algorithm 1 ODAC: Outcome-Driven Actor–Critic

Initialize policy π_θ, replay buffer $\mathcal{R}$, Q-function Q_ϕ, and dynamics model p_ψ.
for iteration $i = 1, 2, \ldots$ **do**
  Collect on-policy samples to add to $\mathcal{R}$ by sampling $\mathbf{g}$ from environment and executing π.
  Sample batch $(\mathbf{s}, \mathbf{a}, \mathbf{s}', \mathbf{g})$ from $\mathcal{R}$.
  Compute approximate reward and optimal weights
  Update Q_ϕ, π_θ, and p_ψ
end for

with $$\hat{r}(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_\Delta) \doteq \hat{q}_{\Delta_{t+1}}(\Delta_{t+1} = 1) \log p_\psi(\mathbf{g} \mid \mathbf{s}_t, \mathbf{a}_t) - D_{\mathrm{KL}}(q_{\Delta_t} \parallel p_{\Delta_t})$$

$$\mathcal{F}_Q(\phi) = \mathbb{E}\left[\left(Q_\phi(\mathbf{s}, \mathbf{a}, \mathbf{g}) - (\hat{r}(\mathbf{s}, \mathbf{a}, \mathbf{g}; q_\Delta) + q_{\Delta_t}(\Delta_t = 0)\, \hat{V}(\mathbf{s}', \mathbf{g}))\right)^2\right]$$

$$\mathcal{F}_\pi(\theta) = -\,\mathbb{E}_{\mathbf{s}\sim\mathcal{D}, \mathbf{a}\sim\pi_\theta(\cdot \mid \mathbf{s};\mathbf{g})}\left[Q_\phi(\mathbf{s}, \mathbf{a}, \mathbf{g}) - \log \pi_\theta(\mathbf{a} \mid \mathbf{s}; \mathbf{g})\right]$$

$$\mathcal{F}_p(\psi) = \mathbb{E}_{(\mathbf{s},\mathbf{a},\mathbf{s}')\sim\mathcal{D}}[\log p_\psi(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})]$$

Outcome-Driven Reward Estimation

Estimation of transition dynamics

- Objective is defined in terms of **likelihood of achieving the outcome**

Outcome-Driven Reward Estimation

Estimation of transition dynamics

- Objective is defined in terms of **likelihood of achieving the outcome**
- Need to know transition dynamics

Outcome-Driven Reward Estimation

Estimation of transition dynamics

- Objective is defined in terms of **likelihood of achieving the outcome**
- Need to know transition dynamics
- If transition dynamics are unknown, estimation is needed:

Outcome-Driven Reward Estimation

Estimation of transition dynamics

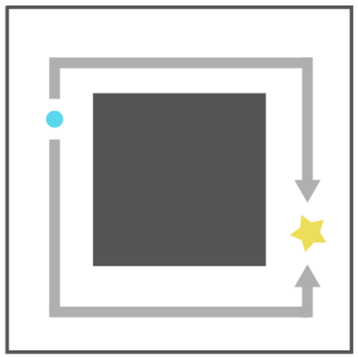
- Objective is defined in terms of **likelihood of achieving the outcome**
- Need to know transition dynamics
- If transition dynamics are unknown, estimation is needed:
  - Fix parametric model

Outcome-Driven Reward Estimation

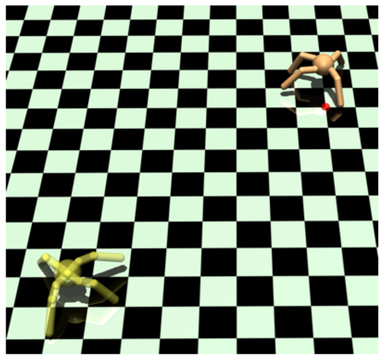
Estimation of transition dynamics

- Objective is defined in terms of **likelihood of achieving the outcome**
- Need to know transition dynamics
- If transition dynamics are unknown, estimation is needed:
  - Fix parametric model
  - Learn parametric model via $\mathcal{F}_p(\psi) = \mathbb{E}_{(\mathbf{s},\mathbf{a},\mathbf{s}')\sim\mathcal{D}}[\log p_\psi(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})]$

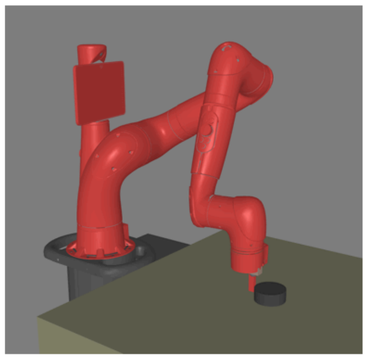
Experiments



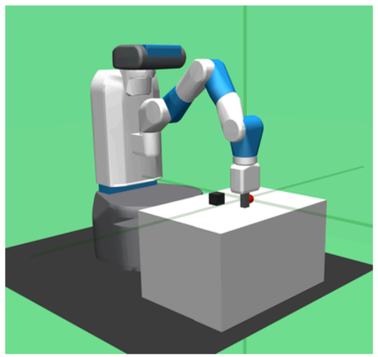
(a) Box 2D



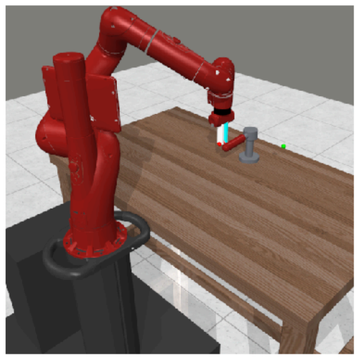
(b) Ant



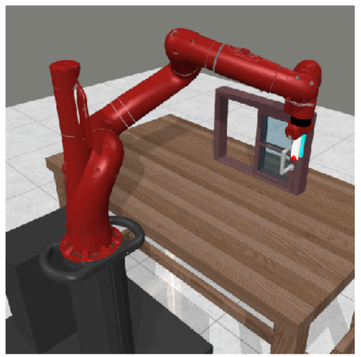
(c) Sawyer Push



(d) Fetch Push



(e) Faucet

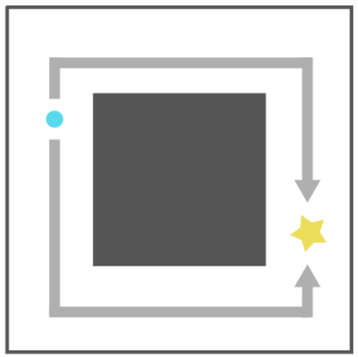


(f) Window

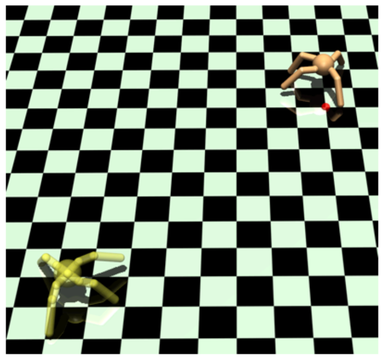
Setting

- No oracle goal sampling

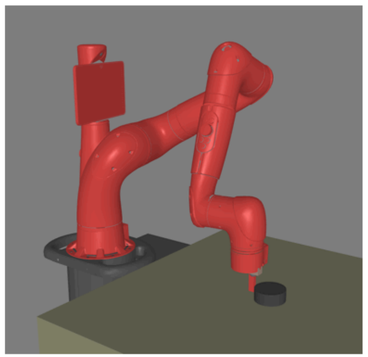
Experiments



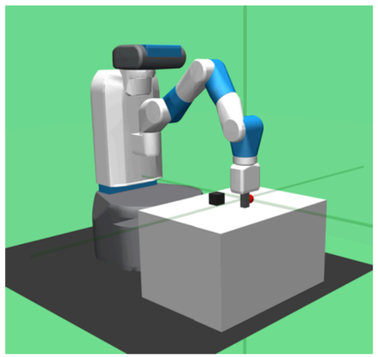
(a) Box 2D



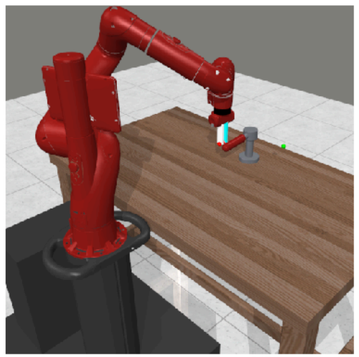
(b) Ant



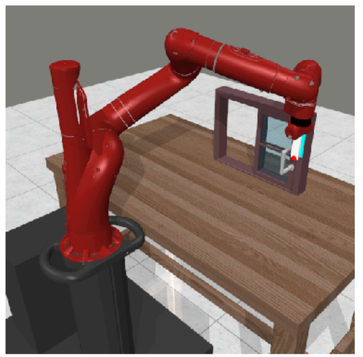
(c) Sawyer Push



(d) Fetch Push



(e) Faucet

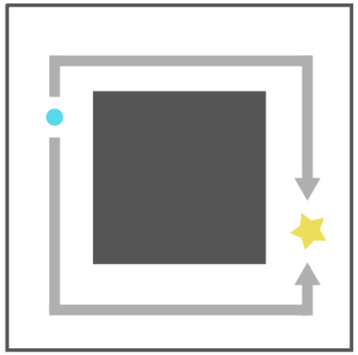


(f) Window

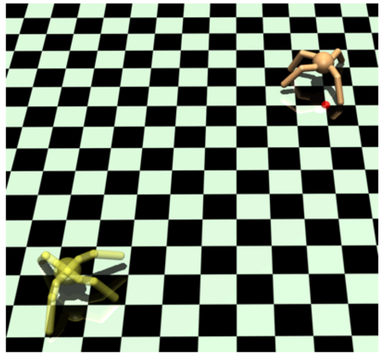
Setting

- No oracle goal sampling
- Future-style relabeling

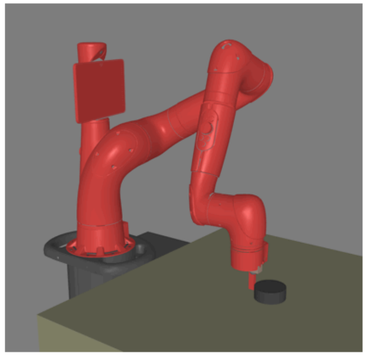
EXPERIMENTS



(a) Box 2D



(b) Ant

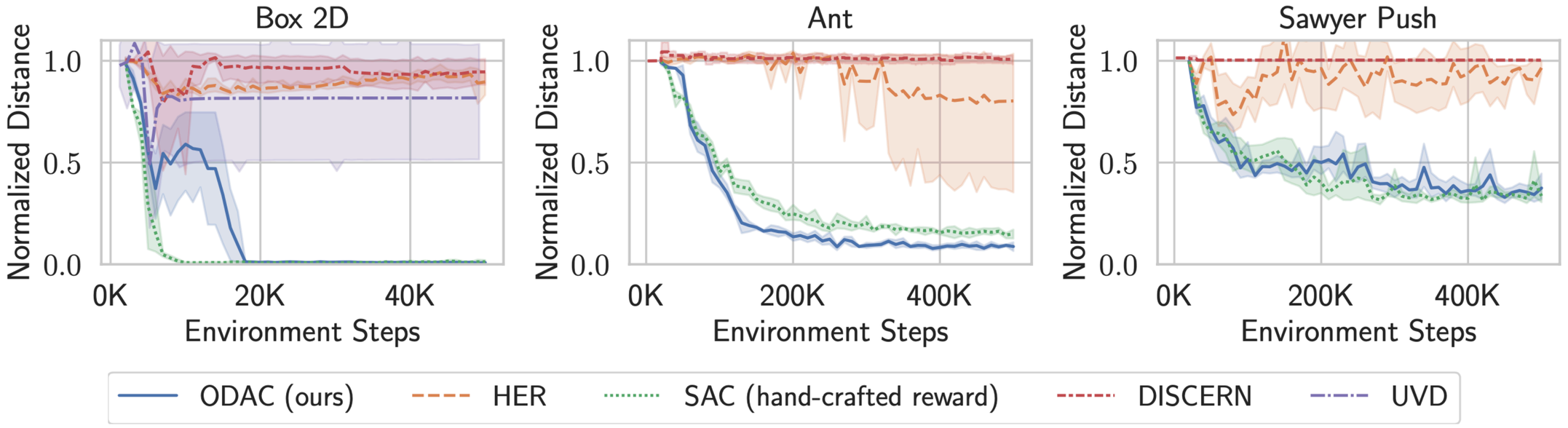


(c) Sawyer Push

(d) Fetch Push

(e) Faucet

(f) Window

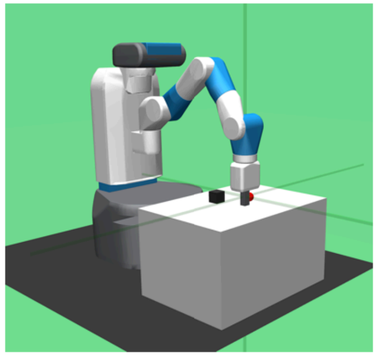


EXPERIMENTS

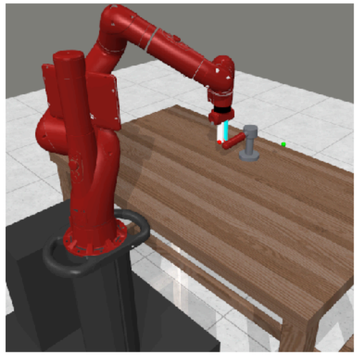
(a) Box 2D

(b) Ant

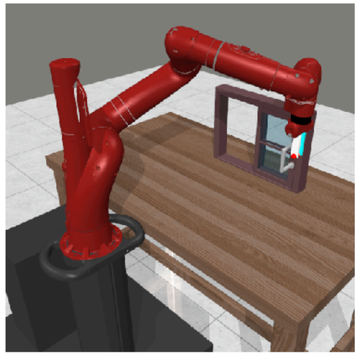
(c) Sawyer Push



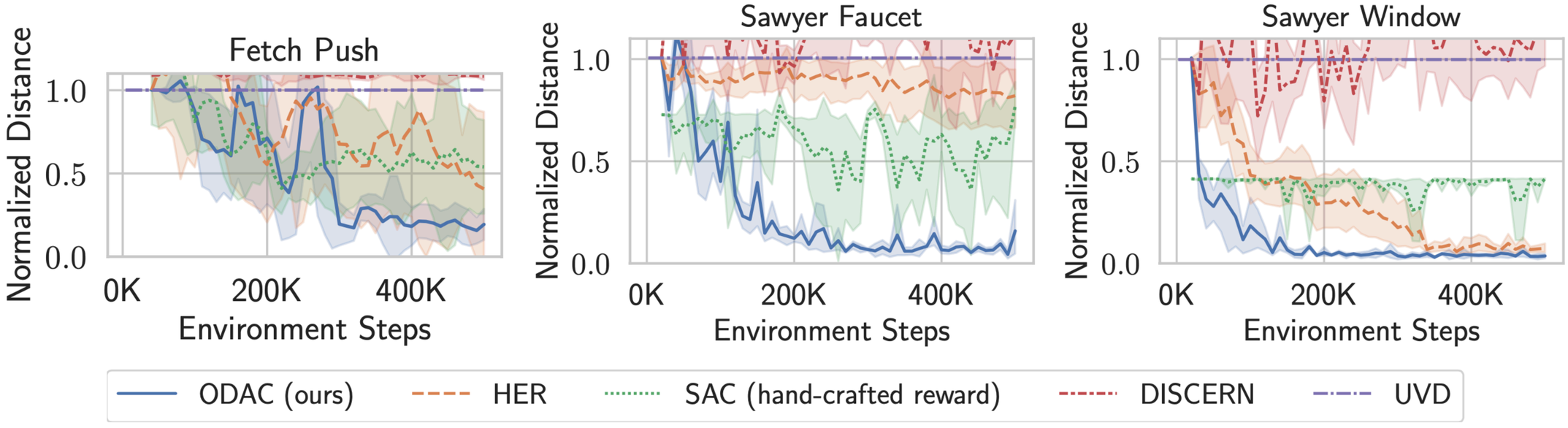
(d) Fetch Push



(e) Faucet



(f) Window



Ablation Studies

Effect of **dynamic discount factor** and **learned dynamics**

Env	ODAC	fixed $\hat{p}_d$	fixed q_T	fixed q_T, $\hat{p}_d$
2D	1.7 (1.20)	1.2 (0.14)	1.0 (0.24)	1.3 (0.29)
Ant	9 (0.48)	11 (0.57)	12 (0.41)	13 (0.20)
Push	35 (2.7)	34 (1.5)	37 (1.5)	38 (3.1)
Fetch	19 (6)	15 (3)	53 (13)	66 (15)
Window	5.4 (0.62)	5.0 (0.62)	7.9 (0.71)	6.0 (0.12)
Faucet	13 (4.2)	15 (3.3)	37 (8.3)	38 (7.2)

Ablation Studies

Effect of **dynamic discount factor** and **learned dynamics**

- Fixing the dynamics model works well

Env	ODAC	fixed $\hat{p}_d$	fixed q_T	fixed q_T, $\hat{p}_d$
2D	1.7 (1.20)	1.2 (0.14)	1.0 (0.24)	1.3 (0.29)
Ant	9 (0.48)	11 (0.57)	12 (0.41)	13 (0.20)
Push	35 (2.7)	34 (1.5)	37 (1.5)	38 (3.1)
Fetch	19 (6)	15 (3)	53 (13)	66 (15)
Window	5.4 (0.62)	5.0 (0.62)	7.9 (0.71)	6.0 (0.12)
Faucet	13 (4.2)	15 (3.3)	37 (8.3)	38 (7.2)

Ablation Studies

Effect of **dynamic discount factor** and **learned dynamics**

- Fixing the dynamics model works well
- Fixing the discount factor deteriorates performance

Env	ODAC	fixed $\hat{p}_d$	fixed q_T	fixed q_T, $\hat{p}_d$
2D	1.7 (1.20)	1.2 (0.14)	1.0 (0.24)	1.3 (0.29)
Ant	9 (0.48)	11 (0.57)	12 (0.41)	13 (0.20)
Push	35 (2.7)	34 (1.5)	37 (1.5)	38 (3.1)
Fetch	19 (6)	15 (3)	53 (13)	66 (15)
Window	5.4 (0.62)	5.0 (0.62)	7.9 (0.71)	6.0 (0.12)
Faucet	13 (4.2)	15 (3.3)	37 (8.3)	38 (7.2)

Main Takeaways

1. RL can be derived from **probabilistic inference** **without** access to a pre-specified **reward function**.

Main Takeaways

1. RL can be derived from **probabilistic inference** **without** access to a pre-specified **reward function**.

2. **Outcome-driven** variational inference can be formulated as a **temporal-difference algorithm**,

Main Takeaways

1. RL can be derived from **probabilistic inference** **without** access to a pre-specified **reward function**.

2. **Outcome-driven** variational inference can be formulated as a **temporal-difference algorithm**,

3. and yields a **dense reward function** that can be estimated from environment interactions.

Thank You!





Tim G. J. Rudner* **Vitchyr H. Pong*** Rowan McAllister Yarin Gal Sergey Levine

Correspondence: tim.rudner@cs.ox.ac.uk & vitchyr@berkeley.edu

Project Website: https://sites.google.com/view/od-ac