Outcome-Driven Reinforcement Learning via Variational Inference

Tim G. J. Rudner[†] [*], Vitchyr H. Pong[‡] [*], Rowan McAllister[‡], Yarin Gal[†], Sergey Levine[‡]

[†] University of Oxford [‡] University of California, Berkeley [*] Equal contribution. Correspondence to: {tim.rudner@cs.ox.ac.uk, vitchyr@berkeley.edu}. @timrudner @vitchyr

Overview

We present a unified probabilistic framework for learning behaviors that allow agents to achieve desired outcomes. This way, we get:

- User-specified reward → **derived reward function**
- Fixed discount factor → **dynamic discount factor**
- **Unification of discrete- and continuous- state space formulations**

Problem Statement & Model

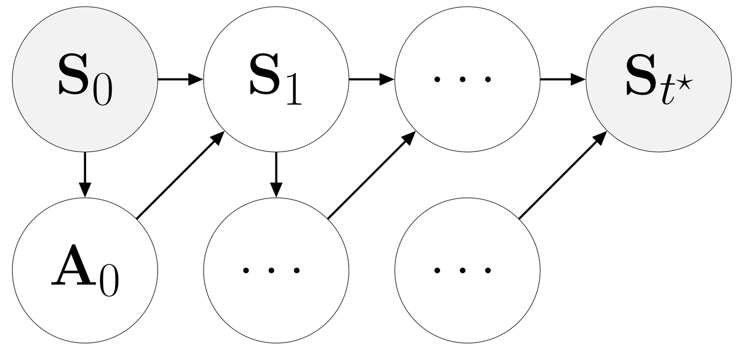
How can we find policies that lead to desired outcomes?

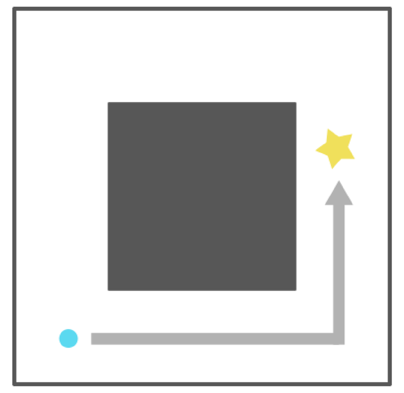
→ Frame the policy search problem probabilistically.

→ Treat the desired outcome as a state realization $\mathbf{S} = \mathbf{g}$.

Known Time of Outcome: $\mathbf{S}_{t^\star} = \mathbf{g}$

Infer a policy $\pi(\mathbf{a}_t|\mathbf{s}_t, \mathbf{S}_{t^\star} = \mathbf{g})$ that induces $p(\boldsymbol{\tau}|\mathbf{s}_0, \mathbf{S}_{t^\star} = \mathbf{g})$.

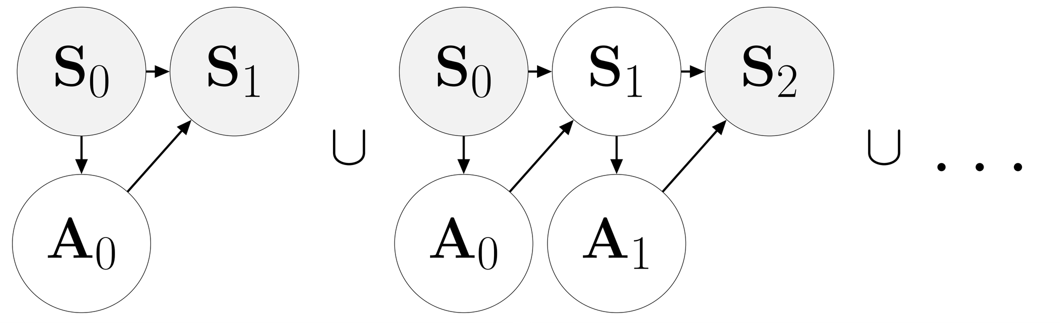




$$p(\boldsymbol{\tau}_{0:t}, \mathbf{g}|\mathbf{s}_0) = p_d(\mathbf{g}|\mathbf{s}_t, \mathbf{a}_t)\pi_0(\mathbf{a}_t|\mathbf{s}_t)\prod_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1}|\mathbf{s}_{t'}, \mathbf{a}_{t'})\pi_0(\mathbf{a}_{t'}|\mathbf{s}_{t'})$$

Unknown Time of Outcome: $\mathbf{S}_{T^\star} = \mathbf{g}$

Infer a policy $\pi(\mathbf{a}_t|\mathbf{s}_t, \mathbf{S}_{T^\star} = \mathbf{g})$ that induces $p(\boldsymbol{\tau}|\mathbf{s}_0, \mathbf{S}_{T^\star} = \mathbf{g})$.



transdimensional distribution over finite trajectories:

$$p(\boldsymbol{\tau}_{0:t}, \mathbf{g}, t|\mathbf{s}_0) = p_T(t)p_d(\mathbf{g}|\mathbf{s}_t, \mathbf{a}_t)\pi_0(\mathbf{a}_t|\mathbf{s}_t)\prod_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1}|\mathbf{s}_{t'}, \mathbf{a}_{t'})\pi_0(\mathbf{a}_{t'}|\mathbf{s}_{t'}) \quad (1)$$

Outcome-Driven Variational Inference

Solve the variational inference problem

$$\min_{\pi, q_T} D_{\mathrm{KL}}(q(\boldsymbol{\tau}, t|\mathbf{s}_0) \,\|\, p(\boldsymbol{\tau}, t|\mathbf{s}_0, \mathbf{S}_{T^\star} = \mathbf{g})), \quad (2)$$

where $q(\boldsymbol{\tau}, t) = q(\boldsymbol{\tau}|t)q_T(t)$ with

$$q_T(t|\mathbf{s}_0) = q(\Delta_{t+1} = 1|\mathbf{s}_0)\prod_{t'=1}^{t} q(\Delta_{t'} = 0|\mathbf{s}_0) \quad (3)$$

$$q(\boldsymbol{\tau}|t, \mathbf{s}_0) = \pi(\mathbf{a}_t|\mathbf{s}_t)\prod_{t'=0}^{t-1} p_d(\mathbf{s}_{t'+1}|\mathbf{s}_{t'}, \mathbf{a}_{t'})\pi(\mathbf{a}_{t'}|\mathbf{s}_{t'}). \quad (4)$$

Theorem 1 (Outcome-Driven Variational Inference).
Solving Equation (2) is equivalent to maximizing

$$V^\pi(\mathbf{s}_0, \mathbf{g}; q_T) = \sum_{t=0}^{\infty} q_T(t)\, \mathbb{E}_{q(\boldsymbol{\tau}_{0:t}|t, \mathbf{s}_0)}\Big[\log p_d(\mathbf{g}|\mathbf{s}_t, \mathbf{a}_t) - D_{\mathrm{KL}}(q(\boldsymbol{\tau}_{0:t}, t|\mathbf{s}_0) \,\|\, p(\boldsymbol{\tau}_{0:t}, t|\mathbf{s}_0))\Big] \quad (5)$$

which can be expressed recursively as

$$V^\pi(\mathbf{s}_0, \mathbf{g}; q_T) = \mathbb{E}_\pi[Q^\pi(\mathbf{s}_0, \mathbf{a}_0, \mathbf{g}; q_T)] - D_{\mathrm{KL}}(\pi(\cdot|\mathbf{s}_0) \,\|\, \pi_0(\cdot|\mathbf{s}_0)) \quad (6)$$

with a **novel Bellman backup**

$$Q^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_T) = r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_T) + \underbrace{q(\Delta_{t+1} = 0)}_{\text{dynamic discount}}\, \mathbb{E}_{p_d}[V^\pi(\mathbf{s}_{t+1}, \mathbf{g}; q_T)], \quad (7)$$

a **derived reward function**

$$r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_T) = \underbrace{(1 - q(\Delta_{t+1} = 0))}_{\text{reward weight}}\underbrace{\log p_d(\mathbf{g}|\mathbf{s}_t, \mathbf{a}_t)}_{\text{learnable from data}} - D_{\mathrm{KL}}(q_{\Delta_t} \,\|\, p_{\Delta_t}), \quad (8)$$

and an optimal **"dynamic" discount function**

$$q(\Delta_{t+1} = 0) = \sigma\left(\log \frac{e^{Q^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_T)}}{p_d(\mathbf{g}|\mathbf{s}_t, \mathbf{a}_t)} + \sigma^{-1}(\gamma)\right). \quad (9)$$

Full paper: timrudner.com/odrl

Theorem 2 (Outcome-Driven Policy Iteration).

A. Outcome-Driven Policy Evaluation: Given π and a Q^0, the sequence $Q^{i+1} = \mathcal{T}^\pi Q^i$ converges to Q^π.

B. The policy π^+ that solves

$$\max_{\pi' \in \Pi}\{\mathbb{E}_{\pi'(\mathbf{a}_t|\mathbf{s}_t)}[Q^\pi(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g}; q_T)] - D_{\mathrm{KL}}(\pi'(\cdot|\mathbf{s}_t) \,\|\, \pi_0(\cdot|\mathbf{s}_t))\}$$

and the variational distribution over T defined in Equation (9) improve the variational objective, that is, for all $\mathbf{s}_0, \pi, q_T$: $V^\pi(\pi^+, q_T, \mathbf{s}_0) \geq V^\pi(\pi, q_T, \mathbf{s}_0)$ and $V^\pi(\pi, q_T^+, \mathbf{s}_0) \geq V^\pi(\pi, q_T, \mathbf{s}_0)$.

C. Alternating between (A) and (B) converges to optimal π and q_T in the variational family.

Empirical Evaluation

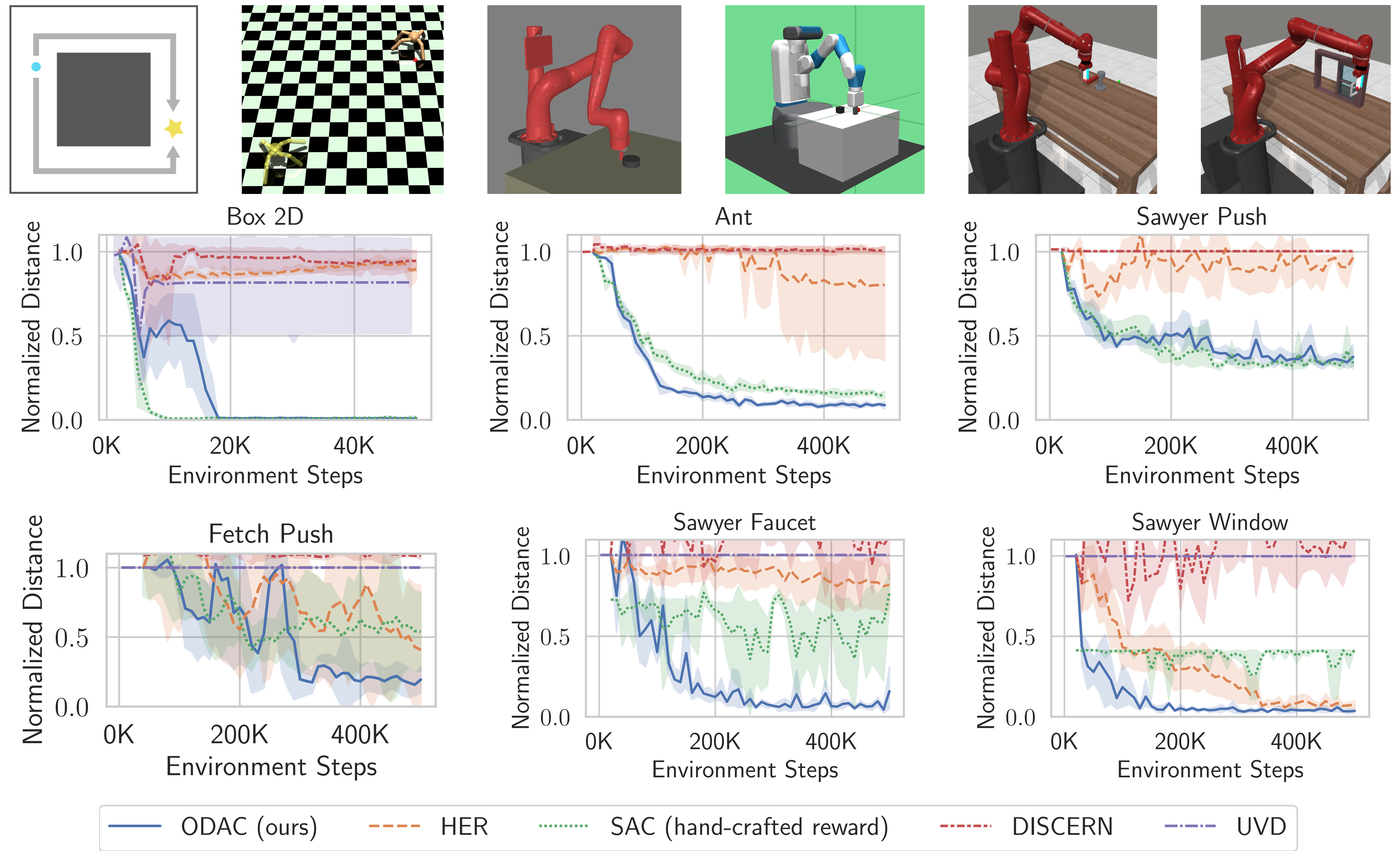


Figure 1: Only ODAC consistently performs well on all six tasks.

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

Figure 2: Ablation results, showing mean final normalized distance ($\times 100$) at the end of training across 4 seeds. ODAC is not sensitive to the dynamics models $\hat{p}_d$ but benefits from the dynamic q_T variant.