Whiteson Research Lab
WhiRL
UNIVERSITY OF OXFORD

VIREL: A Variational Inference Framework for Reinforcement Learning

Matthew Fellows

Talk Structure

- Background in Reinforcement Learning
- Existing RL as Inference Methods
- VIREL: a new framework

Reinforcement Learning
A Primer

$r_t := r(a_t, s_t)$

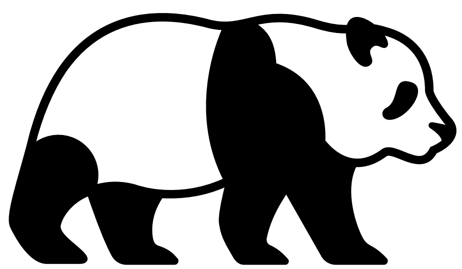
$s_{t+1} \sim p(\,\cdot\,|\, s_t, a_t)$

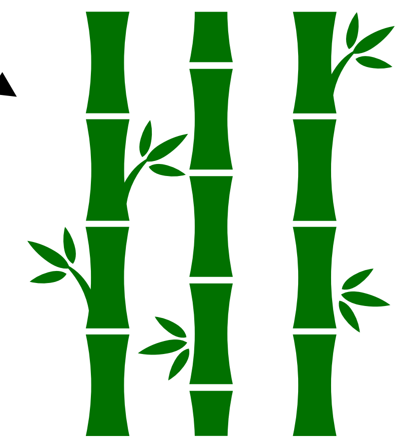
Sample action from a policy

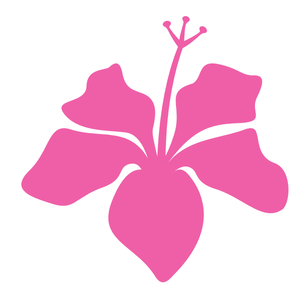
$a_t \sim \pi(\,\cdot\,|\, s_t)$

?

a_t







Currently in state

$s_t \in \mathcal{S}$

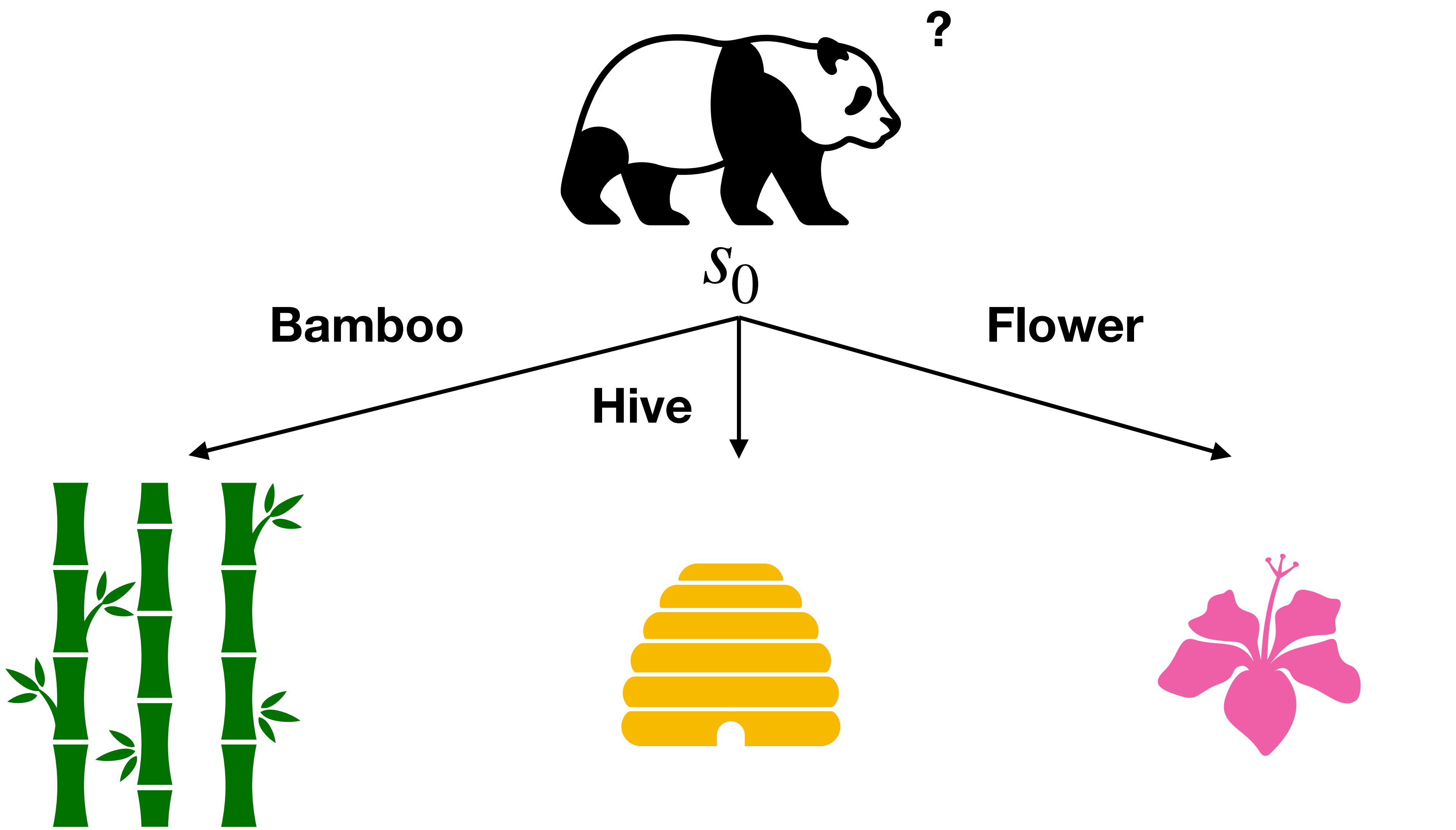
Agent

r_t

s_{t+1}

Environment

Reinforcement Learning
A Primer



$r_0 = 1$ | $r_0 = -5$ | $r_0 = 5$

$r_{10} = 1$ | $r_{10} = 100$ | $r_{10} = -100$

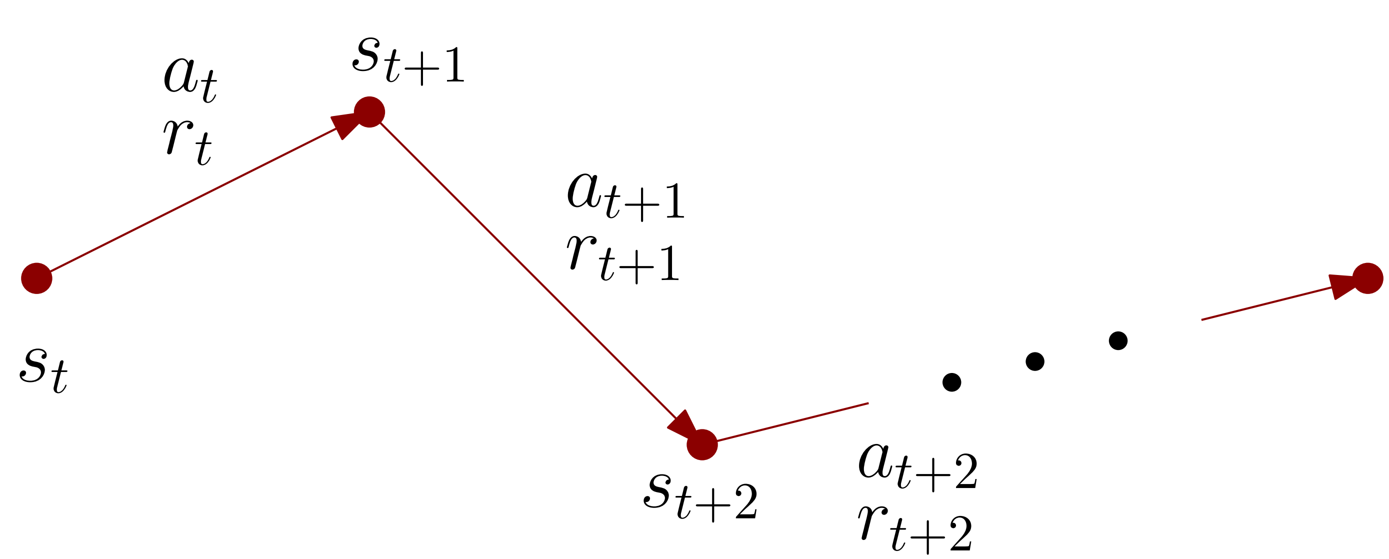
Reinforcement Learning
A Primer

Define the return as $R_{t,N} := \sum_{i=t}^{N-1} \gamma^{i-t} r_t$

Discount factor $\gamma \in [0,1)$

Returns are specific to a particular trajectory

$$\tau_{t,N} := \{s_t, a_t, r_t, s_{t+1}, a_{t+1}, r_{t+1}, \ldots s_{t+N-1}, a_{t+N-1}, r_{t+N-1}\}$$

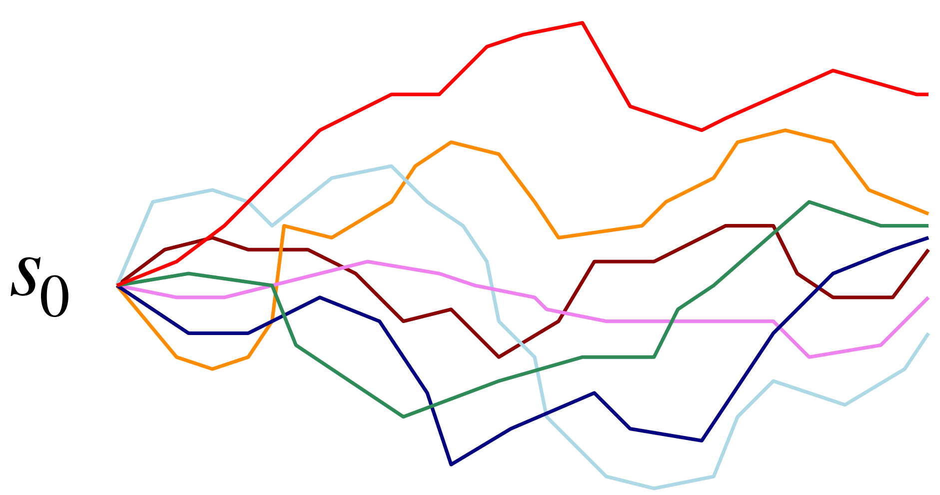


Reinforcement Learning Objective

Denote the probability of a trajectory starting from s_0 :

$$p^{\pi}(\tau_N) = p_0(s_0)\pi(a_0 \mid s_0)\prod_{i=1}^{N-1} p(s_i \mid s_{i-1}, a_{i-1})\pi(a_i \mid s_i)$$

GOAL: Find an optimal policy that maximises the overall *expected* return over all trajectories:



RL OBJECTIVE: $\pi^* \in \arg_{\pi} \max J_N^{\pi} := \arg_{\pi} \max \mathbb{E}_{p^{\pi}(\tau_N)} \left[R_{N,0}\right]$

Reinforcement Learning Objective

More general to work with *infinite horizon* problems:

$$J^\pi := \lim_{N \to \infty} J_N^\pi = \lim_{N \to \infty} \mathbb{E}_{p^\pi(\tau_N)} \left[R_{N,0} \right]$$

Proof of existence, see, for example,
Reinforcement Learning and Optimal Control, Bertsekas

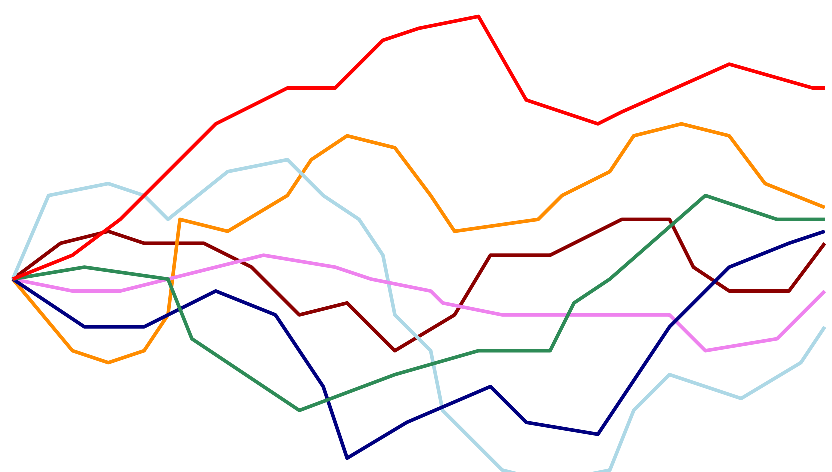
Action-Value Functions

Denote the probability of a trajectory given:

Starting state-action pair, s_t, a_t $\quad p^\pi(\tau_t \mid s_t, a_t) = \prod_{i=1}^{\infty} p(s_{t+i} \mid s_{t+i-1}, a_{t+i-1}) \pi(a_{t+i} \mid s_{t+i})$

Averaging return over all possible trajectories starting in s_t taking action a_t under π

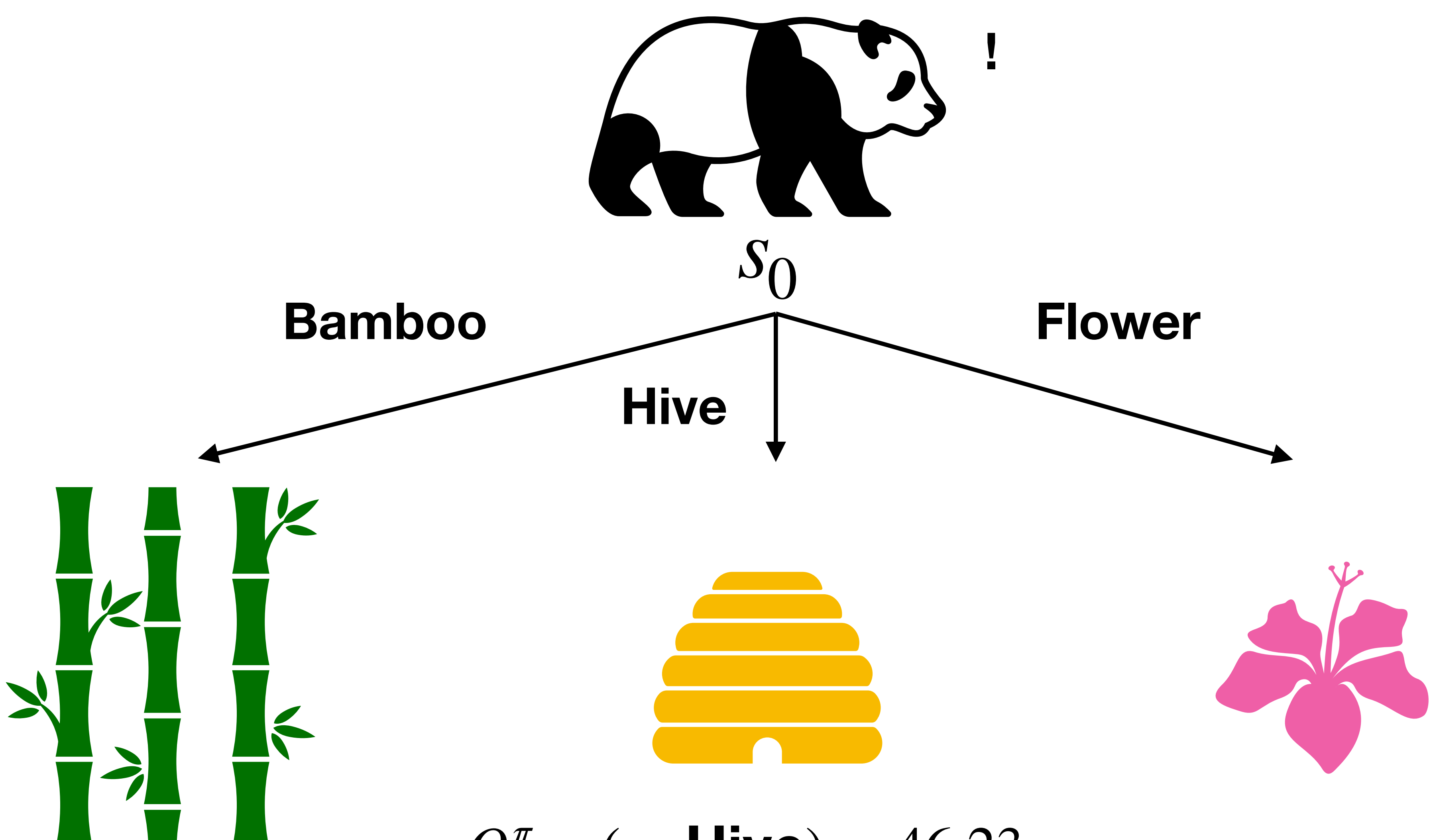
s_t, a_t



Action-value (Q) function: $Q^\pi(s_t, a_t) := \mathbb{E}_{p^\pi(\tau_t \mid s_t, a_t)}\left[R_t\right]$

Re-write RL objective for Q: $J^\pi = \mathbb{E}_{p_0(s)\pi_\theta(a \mid s)}\left[Q^\pi(a, s)\right]$

Q-Functions as ‘Quality’ Functions



$Q^{\pi}_{N=11}(s_0, \textbf{Hive}) = 46.23$

$Q^{\pi}_{N=11}(s_0, \textbf{Bamboo}) = 10.21$

$Q^{\pi}_{N=11}(s_0, \textbf{Flower}) = -25.97$

Bellman Equations and Function Approximators:

Consider the Bellman operator:

$$\mathcal{T}^\pi Q^\pi(a,s) := r(a,s) + \gamma \mathbb{E}_{p(s'|s,a)\pi(a'|s')}\left[Q^\pi(s',a')\right]$$

Any Q-function will satisfy a Bellman equation:

$$\mathcal{T}^\pi Q^\pi(a,s) - Q^\pi(a,s) = 0 \quad \forall\ s,a \in S \times A$$

For any approximate $\hat{Q}_\omega(a,s)$ Q-function parametrised by $\omega \in \Omega$
we define the residual error as:

$$\|\mathcal{T}^\pi \hat{Q}_\omega(a,s) - \hat{Q}_\omega(a,s)\|_p^p$$

$$\|\mathcal{T}^\pi \hat{Q}_\omega(a,s) - \hat{Q}_\omega(a,s)\|_p^p = 0 \implies \hat{Q}_\omega(\,\cdot\,) = Q^\pi(\,\cdot\,)$$

Conditions for Optimality

Definite the optimal Q-function as: $Q^*(\cdot) = Q^{\pi^*}(\cdot)$

Howard (1960): For infinite horizon MDPS, there always exists at least one stationary, deterministic policy:

$$\pi^*(a \mid s) = \delta\left(a \in \arg_{a'} \max Q^*(a', s)\right)$$

Consider the optimal Bellman operator:

$$\mathcal{T}^* Q^\pi(a, s) := r(h) + \gamma \mathbb{E}_{p(s'|s,a)}\left[\max_{a'} Q^\pi(a', s')\right]$$

Any *optimal* Q-Function satisfies the *optimal* Bellman equation:

$$\mathcal{T}^* Q^*(s, a) - Q^*(s, a) = 0 \quad \forall\ s, a \in S \times A$$

Actor-Critic

Probably the most successful class of RL algorithms

Parametrise policy $\pi_\theta(a \mid s)$ with $\theta \in \Theta$ and use function approximator $\hat{Q}_\omega(\cdot) \approx Q^\pi(\cdot)$

ACTOR: $\theta \leftarrow \theta + \alpha_{ac} \nabla_\theta J(\theta)$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\rho^\pi(s)\pi(a|s)} \left[\hat{Q}_\omega(a, s) \nabla_\theta \log \pi_\theta(a \mid s) \right]$$

Like policy improvement, updates θ in direction of increasing rewards

CRITIC: $\omega \leftarrow \omega - \frac{1}{2} \alpha_{cr} \nabla_\omega \mathbb{E}_{d(s)} \left[\left(\mathcal{T}^\pi \hat{Q}_\omega(a, s) - \hat{Q}_\omega(a, s) \right)^2 \right]$

$$\frac{1}{2} \nabla_\omega \mathbb{E}_{d(s)} \left[\left(\mathcal{T}^\pi \hat{Q}_\omega(a, s) - \hat{Q}_\omega(a, s) \right)^2 \right] \approx - \mathbb{E}_{d(s)} \left[\left(\mathcal{T}^\pi \hat{Q}_\omega(a, s) - \hat{Q}_\omega(a, s) \right) \nabla_\omega \hat{Q}_\omega(h) \right]$$

Like policy evaluation, updates ω to minimiser error between $\hat{Q}_\omega(\cdot)$ and $Q^{\pi_{new}}(\cdot)$

Reinforcement Learning as Inference-Motivation

- Powerful methods from variational inference literature can be applied to RL
- Bayesian interpretation of RL problem can be exploited for uncertainty driven exploration
- Deeper theoretical understanding of RL can highlight key problems in existing algorithms

Reinforcement Learning as Inference-A Brief Review

Introduce a binary variable $\mathcal{O}_t \in \{0,1\}$

$\mathcal{O}_t = 1$ is the event that agent is behaving 'optimally'

However, semantics of $\mathcal{O}_t$ are not formally defined

We write $\mathcal{O}_t$ for $\mathcal{O}_t = 1$ and introduce a new restriction, $r(\cdot) \leq 0$

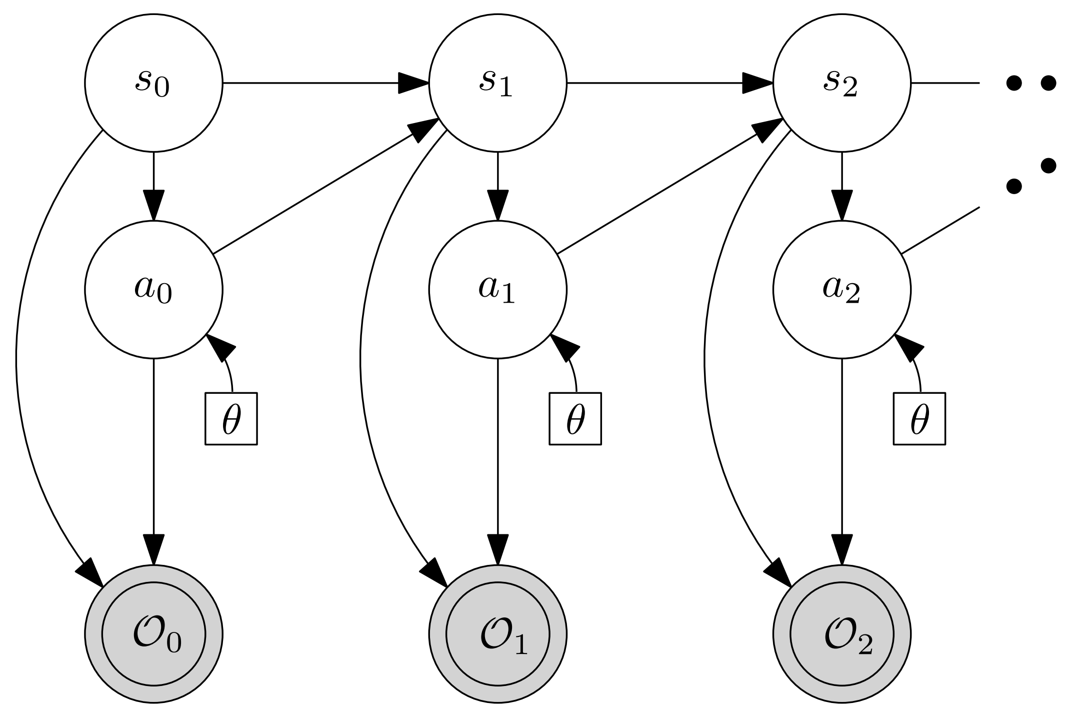
The distribution over $\mathcal{O}_t$ is defined as: $p(\mathcal{O}_t \mid s_t, a_t) := \exp\left(r_t\right)$

Likelihood is defined as: $p(\mathcal{O} \mid \tau) = \prod_{t=0}^{N-1} p(\mathcal{O}_t \mid s_t, a_t) = \exp\left(\sum_{t=0}^{N-1} r_t\right)$

Two approaches follow:

θ **= Model parameters** **Maximum Likelihood Problem**	θ **= Variational parameters** **Inference Problem**

Approach i: Pseudo-Likelihood Methods: θ as model parameters



Introducing a prior over trajectories: $p_\theta(\tau) := p_0(s_0)\pi_\theta(a_0 \mid s_0)\prod_{i=1}^{N-1} p(s_i \mid s_{i-1}, a_{i-1})\pi_\theta(a_i \mid s_i)$

The joint follows as: $p_\theta(\tau, \mathcal{O}) = P(\mathcal{O} \mid \tau)p_\theta(\tau) = \exp\left(\sum_{i=0}^{N-1} r_i\right)p_\theta(\tau)$

Approach i: Pseudo-Likelihood Methods

The marginal-likelihood is thus the expected *exponential* return:

$$p_\theta(\mathcal{O}) = \int P(\mathcal{O} \mid \tau) p_\theta(\tau) d\tau = \mathbb{E}_{p_\theta(\tau)} \left[\exp \left(\sum_{i=0}^{N-1} r_i \right) \right]$$

Compare to the (episodic, undiscounted) reinforcement learning objective:

$$J(\theta) = \mathbb{E}_{p_\theta(\tau)} \left[\sum_{i=0}^{N-1} r_i \right]$$

Finding maximum marginal likelihood equivalent to solving MDP with transformed rewards-solved using (V)EM!

State of the art: MPO (ish!) [Abdolmaleki et al 18]

Critical Problem with Pseudo-Likelihood

The (V) E-step infers posterior $q(\tau) \approx p_\theta(\tau \mid \mathcal{O})$ which characterises return in MDP

The M-step minimises the *forward* (mass-covering) KL divergence for θ:

Pseudo-likelihood: $KL(q(\tau) \| p_\theta(\tau \mid \mathcal{O}))$

Target distribution, proportional to exponential return

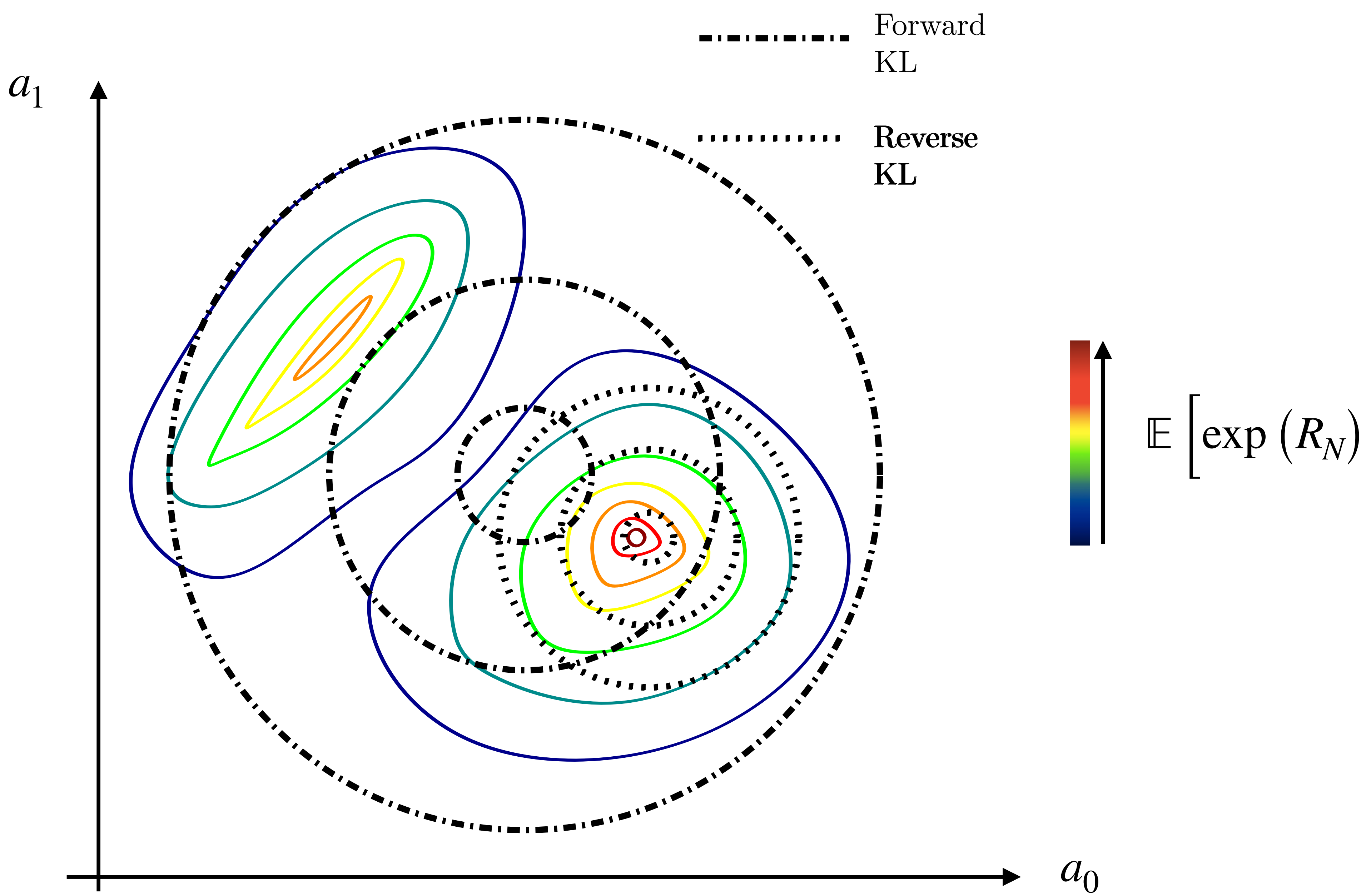
Distribution containing policy to be improved

Classic RL optimises the *reverse* (mode-seeking) form of KL divergence:

Classic RL: $KL(p_\theta(\tau \mid \mathcal{O}) \| q(\tau))$

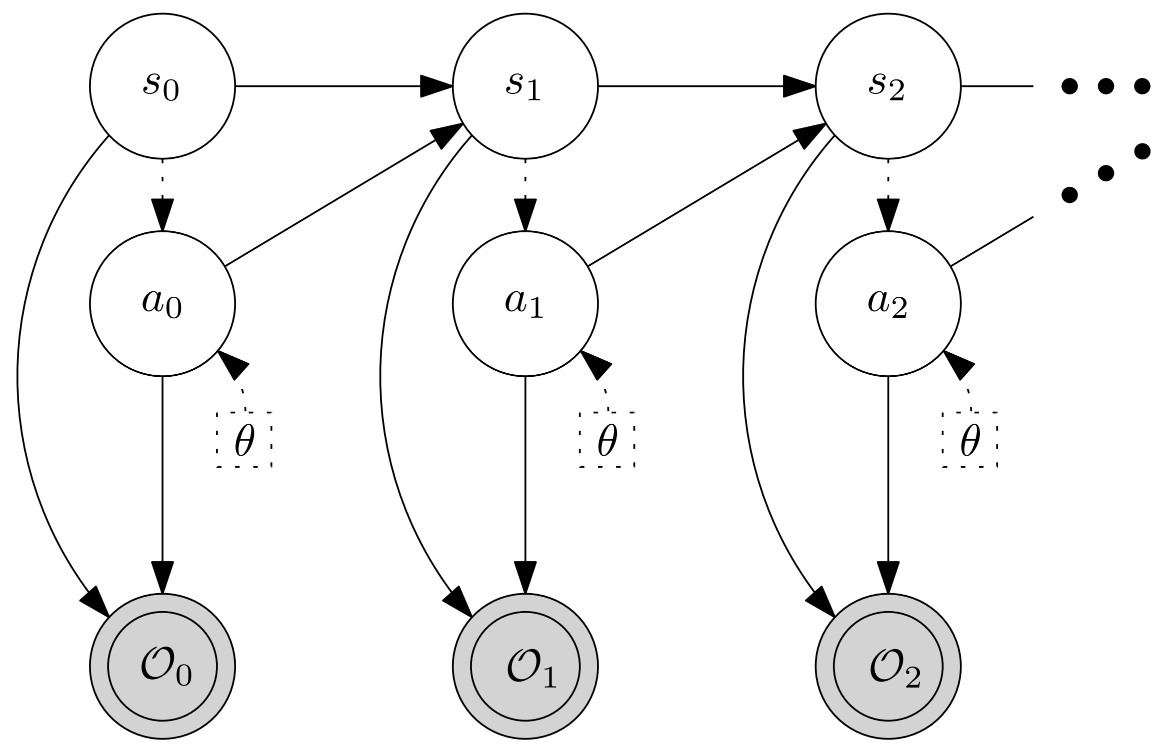
Pseudo-likelihood promotes risk-seeking behaviour

Critical Problem with Pseudo-Likelihood



See [Neumann 11] for examples of this in practice

Approach ii: Maximum Entropy RL (MERL)

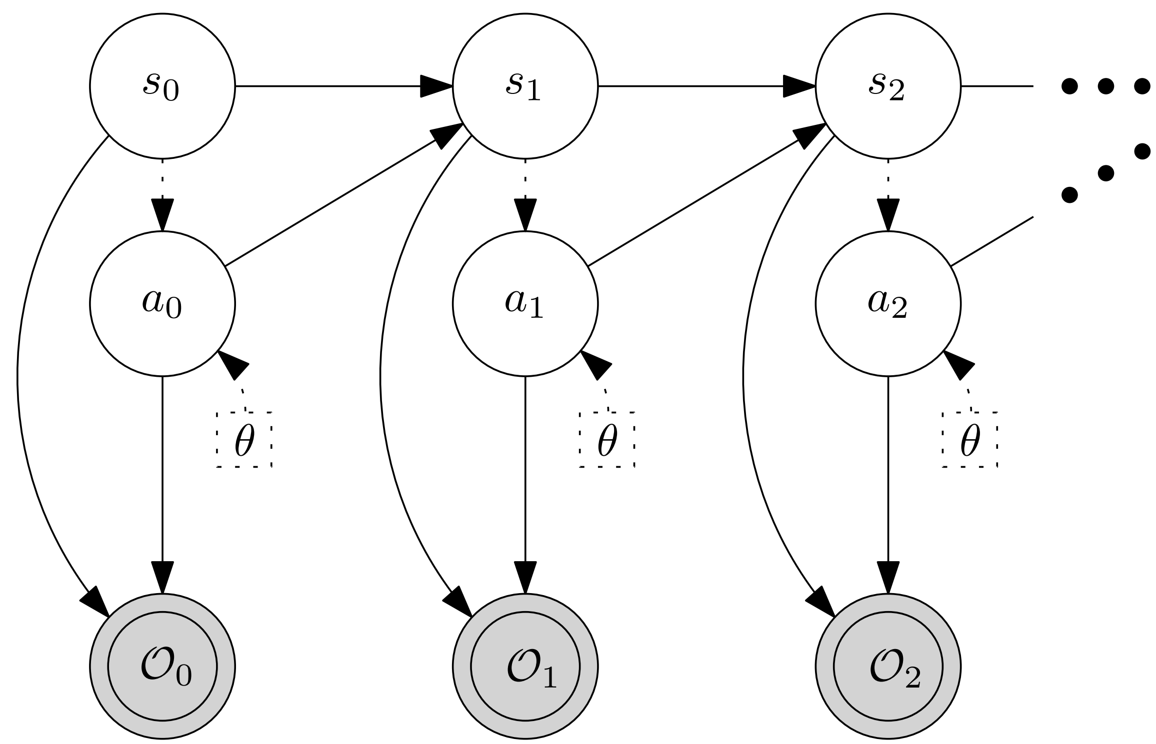


For MERL, the prior is independent of θ: $p(\tau) := p_0(s_0)\prod_{i=1}^{N-1} p(s_i \mid s_{i-1}, a_{i-1})\mathcal{U}(a_i)$

The joint follows as: $p(\tau, \mathcal{O}) = P(\mathcal{O} \mid \tau)p_\theta(\tau) = \exp\left(\sum_{i=0}^{N-1} r_i\right)p(\tau)$

See [Levine 18] for a full overview

Approach ii: Maximum Entropy RL (MERL)



The posterior distribution is derived as: $$p(\tau \mid \mathcal{O}) = \frac{\exp\left(R_N\right) p(\tau)}{\int \exp\left(R_N\right) p(\tau) d\tau}$$

The variational distribution is defined as: $$q_\theta(\tau) := p_0(s_0) \prod_{i=1}^{N-1} p(s_i \mid s_{i-1}, a_{i-1}) \pi_\theta(a_i \mid s_i)$$

Maximum Entropy RL Objective

Optimising the *reverse* KL divergence:

$$\arg_\theta \min KL\left(q_\theta(\tau) \| p(\tau \mid \mathcal{O})\right) = \arg_\theta \max \mathcal{L}(\theta)$$

The ELBO can be derived as:

Temperature parameter

$$\mathcal{L}(\theta) = \mathbb{E}_{q_\theta(\tau)}\left[\sum_{i=0}^{N-1}\left(r_i - \log \pi_\theta(a_i \mid s_i)\right)\right] = \mathbb{E}_{q_\theta(\tau)}\left[\sum_{i=0}^{N-1} r_i\right] + c\sum_{i=0}^{N-1}\mathbb{E}_{p(s_i|s_{i-1},a_{i-1})}\left[\mathcal{H}\left(\pi_\theta(\cdot \mid s_i)\right)\right]$$

Again, compare to the (episodic, undiscounted) reinforcement learning objective:

$$J(\theta) = \mathbb{E}_{p_\theta(\tau)}\left[\sum_{i=0}^{N-1} r_i\right]$$

Inferring $q_\theta(\tau)$ closest in KL Divergence to the posterior is equivalent to solving the maximum entropy RL objective

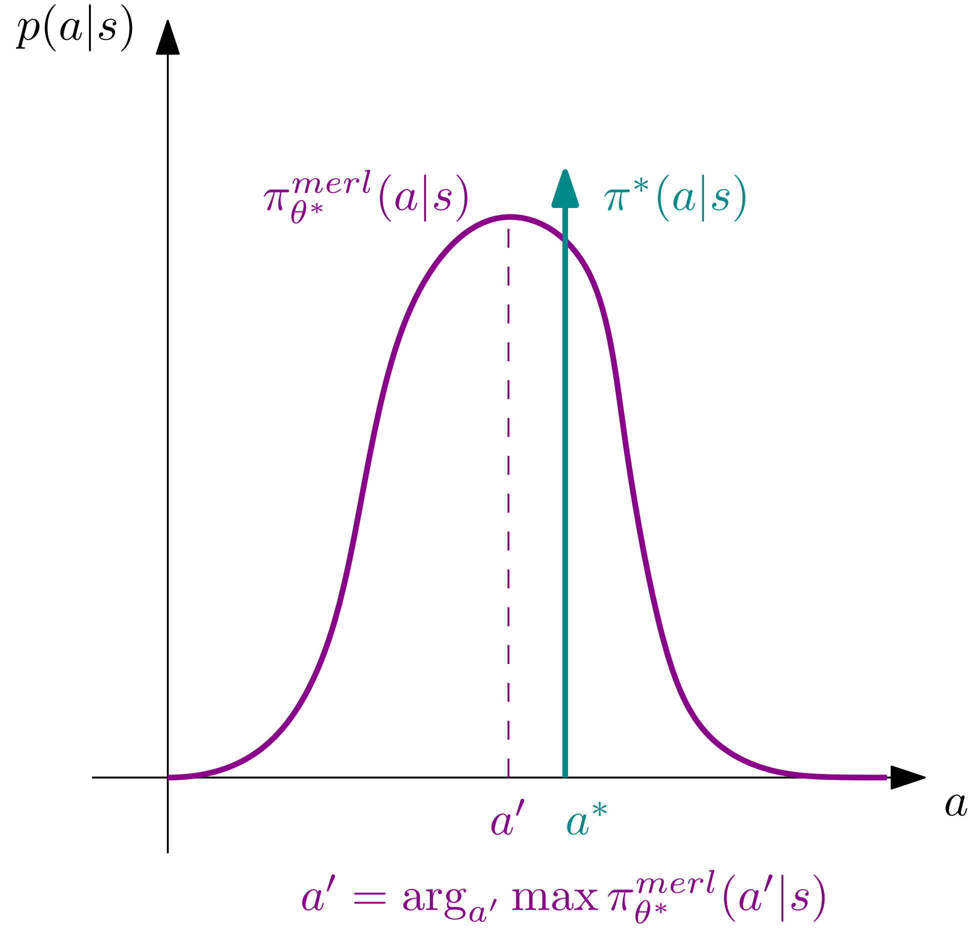
State of the art: Soft Actor Critic [Haarnoja et al 18]

Problems with MERL

Discounting and infinite horizon MDPs are complicated (see [Thompson 14])

Defining $\pi_{\theta^*}^{merl}(a \,|\, s)$ as the optimal policy under $\mathscr{L}(\theta)$

$\pi_{\theta^*}^{merl}(a \,|\, s)$ is not deterministic and in general, $\arg_{a'} \max \pi_{\theta^*}^{merl}(a' \,|\, s) \neq \arg_{a'} \max Q^*(a', s)$



Restricting to deterministic policies renders inference intractable [Rawlik 10]

Optimal deterministic policies are not learned

Simple Counterexample

$r(s_0, a_2) = 1$

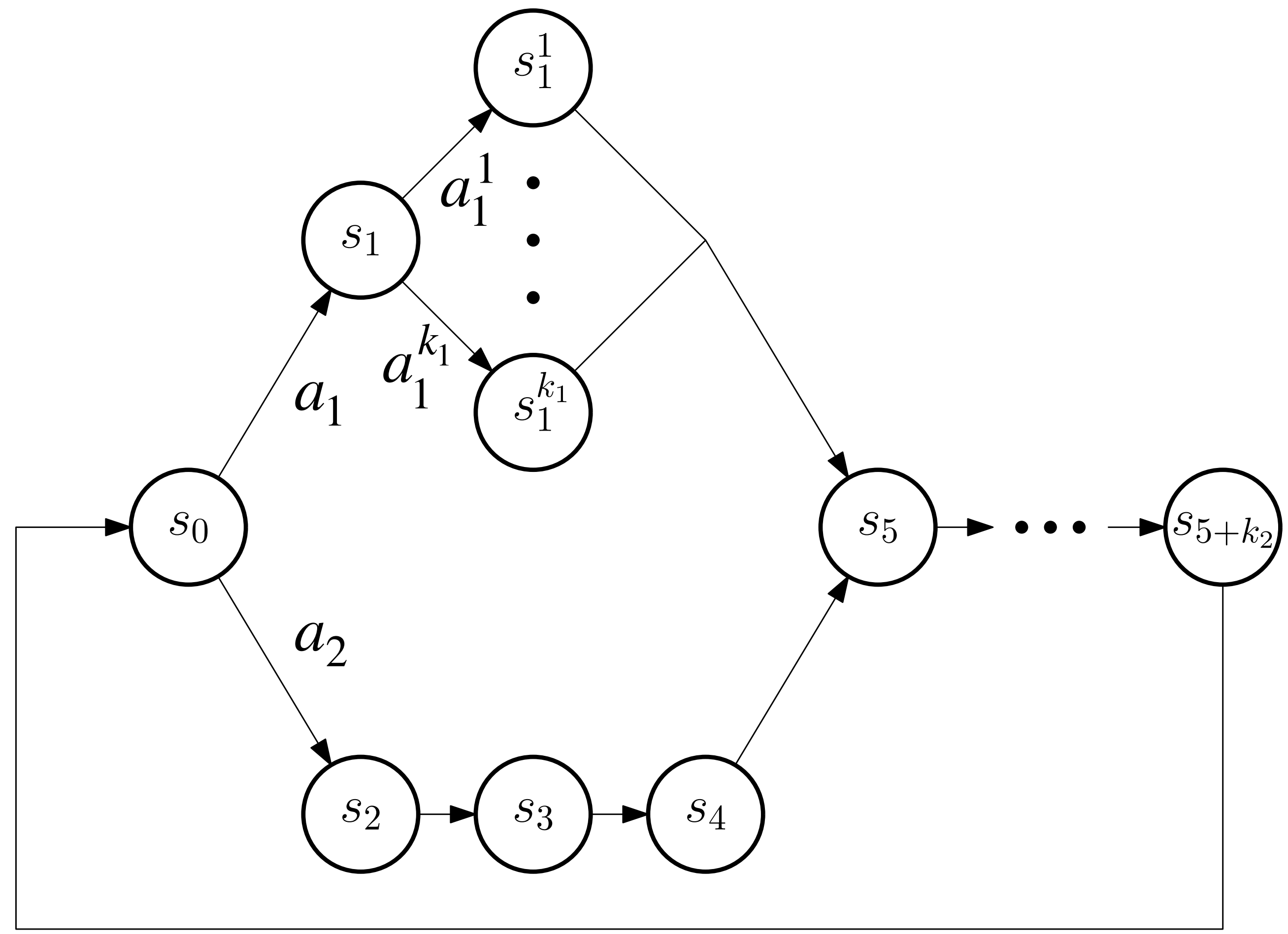
$r = 0$ everywhere else

Deterministic state transitions

Optimal policy has $\pi^*(a_2 \mid s_0) = 1$

Optimal MERL policy has

$$\pi^*_{merl}(a_1 \mid s_0) = \frac{1}{k_1^{-\gamma} \exp(\frac{1}{c}) + 1}$$



For any $k_1^{-\gamma} \exp\left(\frac{1}{c}\right) < 1 \implies \pi^*_{merl}(a_1 \mid s_0) > \frac{1}{2}$

π^* cannot be recovered from π^*_{merl}

Optimality of Soltuion Highly Sensitive to Temperature Parameter

Goals for a General RL Inference Framework

Naturally learns optimal deterministic policies

Variational distribution is a policy, not trajectory

Optimises the reverse form of KL divergence

VIREL

Temperature not a hyperparameter

Discounting easily incorporated

Function approximators explicitly used

Stochastic policies used for learning

VIREL

A Variational Inference Framework for Reinforcement Learning

M Fellows   A Mahajan   T G J Rudner   S Whiteson

For simplicity of notation, define hidden variables: $h := \langle a, s \rangle$

ASSUMPTION I: Optimal Q-function is finite and positive $0 < Q^*(\cdot) < \infty$

Introducing an approximate Q-function: $\hat{Q}_\omega(\cdot), \quad \omega \in \Omega$

ASSUMPTION II: $\exists\ \omega^* \in \Omega\ s.t.\ \hat{Q}_{\omega^*}(\cdot) = Q^*(\cdot)$ i.e. Optimal Q-function can be represented by an approximator

(**A II** relaxed using projected Bellman errors, extending [Bhatnagar et al 09])

ASSUMPTION III: $\hat{Q}_{\omega^*}(\cdot)$ has unique maximum and is a locally $\mathbb{C}^2$ smooth about that maximum.

Model Specification

Define the residual error as $\varepsilon_\omega := \frac{1}{p|H|} \|\mathscr{T}_\omega \hat{Q}_\omega(h) - \hat{Q}_\omega(h)\|_p^p$

Which is the temperature of a Boltzmann policy:

$$\pi_\omega(a \mid s) = \frac{\exp\left(\frac{\hat{Q}_\omega(h)}{\varepsilon_\omega}\right)}{\int \exp\left(\frac{\hat{Q}_\omega(h)}{\varepsilon_\omega}\right) da}$$

Function approximators explicitly used

Temperature defined explicitly

Theorem 1: In the limit $\varepsilon_\omega \to 0,\ \pi_\omega(a \mid s)$ tends towards a Dirac-delta distribution centred on $\arg_{a'} \max \hat{Q}_\omega(a', s)$, that is:

$$\lim_{\varepsilon_\omega \to 0} \int \varphi(a) \pi_\omega(a \mid s) da = \varphi(a = \arg_{a'} \max \hat{Q}_\omega(a', s)) \quad \forall\ \varphi(\,\cdot\,) \in \mathbb{C}_0^\infty(A)$$

Model Specification

Define the residual error as $\varepsilon_\omega := \frac{1}{p|H|} \|\mathscr{T}_\omega \hat{Q}_\omega(h) - \hat{Q}_\omega(h)\|_p^p$

$\mathscr{T}_\omega \cdot$ any operator which recovers the optimal Bellman operator when $\varepsilon_\omega \to 0$

$$\mathscr{T}_\omega \cdot \in \mathbb{T} := \left\{ \mathscr{T}_\omega \cdot \;\middle|\; \lim_{\varepsilon_\omega \to 0} \mathscr{T}_\omega \hat{Q}_\omega(\,\cdot\,) = \mathscr{T}^* \hat{Q}_\omega(\,\cdot\,) \right\}$$

e.g. $\mathscr{T}_\omega \hat{Q}_\omega(\,\cdot\,) := r(\,\cdot\,) + \gamma \mathbb{E}_{p(s'|\cdot)\pi_\omega(a'|s')} \left[\hat{Q}_\omega(h') \right] \in \mathbb{T}$ (note: constrains Ω)

e.g. $\mathscr{T}^* \hat{Q}_\omega(\,\cdot\,) := r(\,\cdot\,) + \gamma \mathbb{E}_{p(s'|\cdot)} \left[\max_{a'} \hat{Q}_\omega(a', s') \right] \in \mathbb{T}$

Discounting easily incorporated

Main Theoretical Result

Theorem 2: For any $\omega^* \; s.t. \; \varepsilon_{\omega^*} = 0$, it follows that:
i) the corresponding approximator is optimal, i.e. $\hat{Q}_{\omega^*}(\cdot) = Q^*(\cdot)$
ii) the corresponding Boltzmann policy is optimal, i.e.

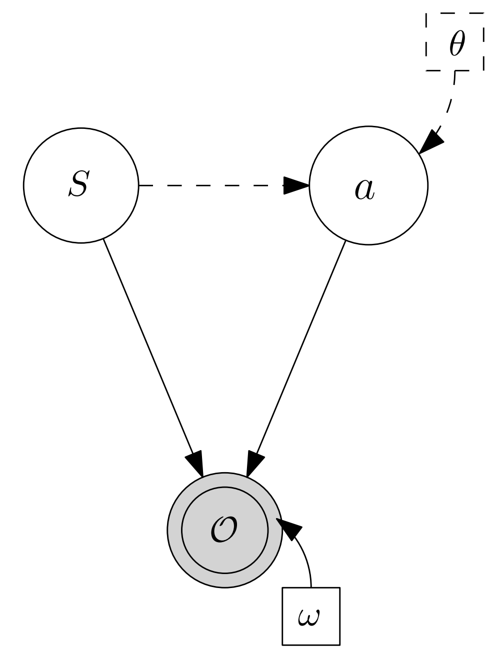
$$\pi_{\omega^*}(a \mid \cdot) = \delta(a = \arg_{a'} \max Q^*(a', \cdot))$$

OBJECTIVE: $\arg_\omega \min \varepsilon_\omega$ Naturally learns optimal deterministic policies

Corollary 1: $\varepsilon_\omega = 0$ is also a necessary condition for i) and ii), hence $\varepsilon_\omega > 0$ for any non-optimal $\hat{Q}_\omega(\cdot)$ and π_ω

IMPLIES: π_ω stochastic whenever $\varepsilon_\omega > 0$ Stochastic policies used for learning

Probabilistic Interpretation



Introduce binary variable $\mathscr{O} \in \{0,1\}$

$$p_\omega(\mathscr{O} \,|\, h) := y_\omega(h)^{\mathscr{O}}(1 - y_\omega(h))^{1-\mathscr{O}}$$

$$y_\omega(h) := \exp\left(\frac{\hat{Q}_\omega(h) - \max_{a'} \hat{Q}_\omega(a', s)}{\varepsilon_\omega}\right)$$

(well defined for $\varepsilon_\omega > 0$)

$\mathscr{O} = 1$ event that samples are optimal under $\hat{Q}_\omega(\,\cdot\,)$ i.e. greedy under $\hat{Q}_\omega(\,\cdot\,)$:

Given $s \in S$ and $a^\star \in \arg_{a'} \max \hat{Q}_\omega(a', s)$, $\mathscr{O} = 1$ with complete certainty,

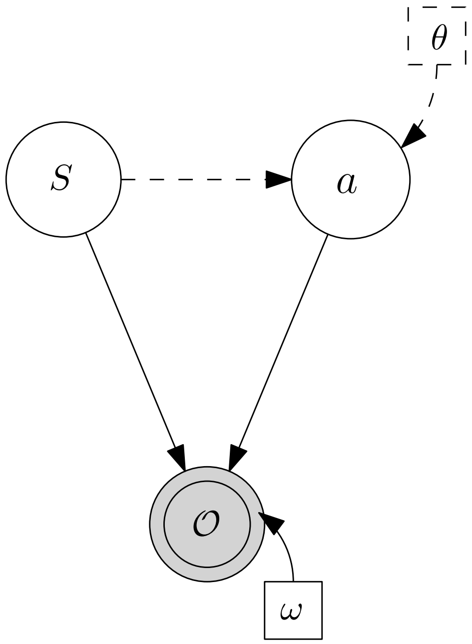
$$p_\omega(\mathscr{O} = 1 \,|\, a^\star, s) = \exp\left(\frac{\hat{Q}_\omega(a^\star, s) - \max_{a'} \hat{Q}_\omega(a', s)}{\varepsilon_\omega}\right) = 1$$

Probabilistic Interpretation

Writing $\mathcal{O}$ for $\mathcal{O} = 1$ and defining:

$$y_\omega(s) := \exp\left(\frac{-\max_{a'} \hat{Q}_\omega(a', s)}{\varepsilon_\omega}\right)$$

we have $p_\omega(\mathcal{O} \mid h) = \exp\left(\frac{\hat{Q}_\omega(h)}{\varepsilon_\omega}\right) y_\omega(s)$



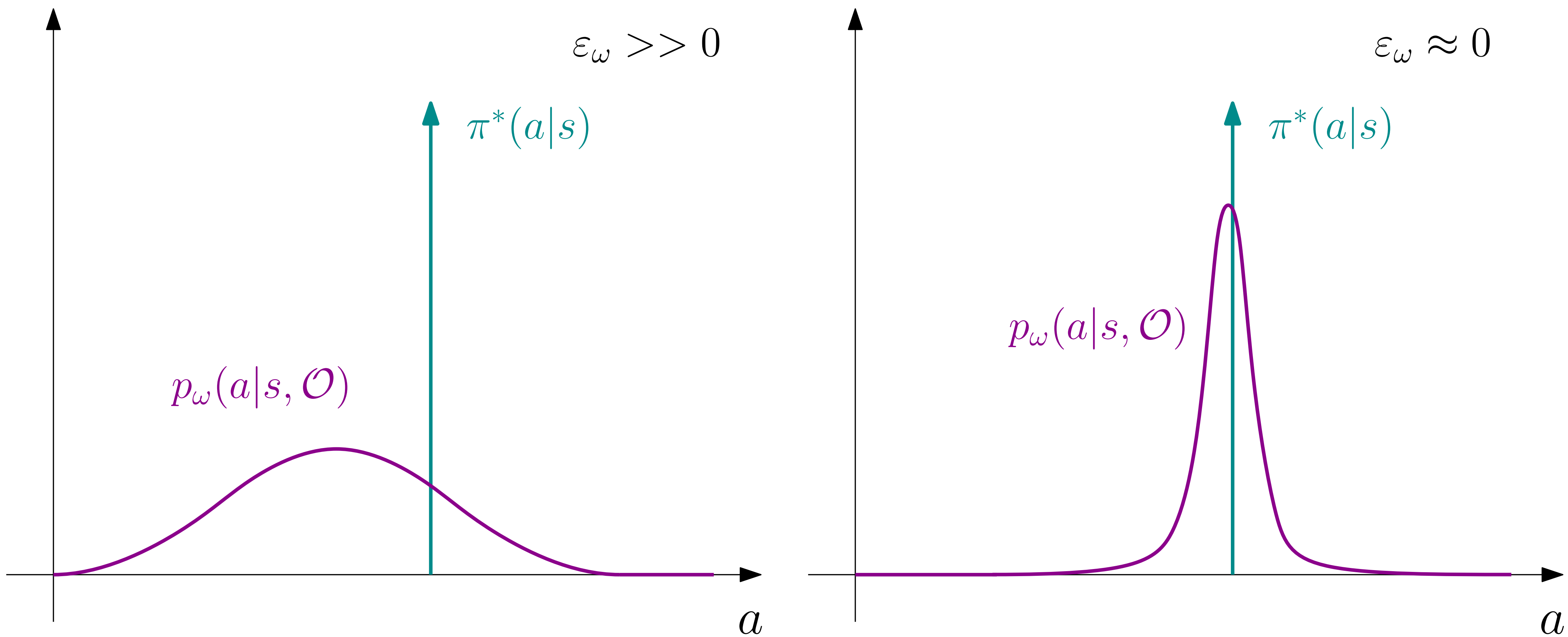
(well defined for $\varepsilon_\omega > 0$)

Defining a prior to be uniform $p(h) = \mathcal{U}(h)$ the state-conditional *action posterior* is:

$$p_\omega(a \mid s, \mathcal{O}) = \frac{\exp\left(\frac{\hat{Q}_\omega(h)}{\varepsilon_\omega}\right)}{\int \exp\left(\frac{\hat{Q}_\omega(h)}{\varepsilon_\omega}\right) da} = \pi_\omega(a \mid s)$$

We recover our Boltzmann distribution!

Probabilistic Interpretation



Model not confident about optimal policy

Model confident about optimal policy

Sampling from $p_\omega(a \mid s, \mathcal{O}) = \pi_\omega(a \mid s)$ affords uncertainty driven exploration

Inferring the Action Posterior

Sampling directly from the action posterior is not possible in general

Introduce variational distribution:

Variational distribution is a policy, not trajectory

$$q_\theta(h) := d(s)\pi_\theta(a \mid s)$$

Arbitrary sampling distribution with support over $\mathcal{S}$

Variational policy $\pi_\theta \approx \pi_\omega$

Optimises the *reverse* form of KL divergence

Full posterior is $p_\omega(h \mid \mathcal{O}) = \dfrac{\exp\left(\frac{\hat{Q}_\omega(h)}{\varepsilon_\omega}\right) y_\omega(s)}{\int \exp\left(\frac{\hat{Q}_\omega(h)}{\varepsilon_\omega}\right) y_\omega(s) dh}$

Objective: $\arg_\theta \min KL(q_\theta(h) \| p_\omega(h \mid \mathcal{O})) = \arg_\theta \max \mathcal{L}_\omega(\theta)$

$$\mathcal{L}_\omega(\theta) = \mathbb{E}_{d(s)}\left[\mathbb{E}_{\pi_\theta(a|s)}\left[\frac{\hat{Q}_\omega(h)}{\varepsilon_\omega}\right] + \mathcal{H}(\pi_\theta(\cdot \mid s))\right] + \mathbb{E}_{d(s)}\left[\log\left(\frac{y_\omega(s)}{d(s)}\right)\right]$$

Inferring the Action Posterior

Theorem 3: For any $\varepsilon_\omega > 0,\ \max_\theta \mathscr{L}_\omega(\theta) = \min_\theta \mathbb{E}_{d(s)}\left[KL(\pi_\theta(\cdot\,|\,s) \| \pi_\omega(\cdot\,|\,s))\right]$

What about when $\varepsilon_\omega = 0$?

To prevent ill conditioning, we maximise $\varepsilon_\omega \mathscr{L}_\omega(\theta)$ anyway:

$$\varepsilon_\omega \mathscr{L}_\omega(\theta) = \mathbb{E}_{d(s)}\left[\mathbb{E}_{\pi_\theta(a|s)}\left[\hat{Q}_\omega(h)\right] + \varepsilon_\omega \mathscr{H}(\pi_\theta(\cdot\,|\,s))\right]$$

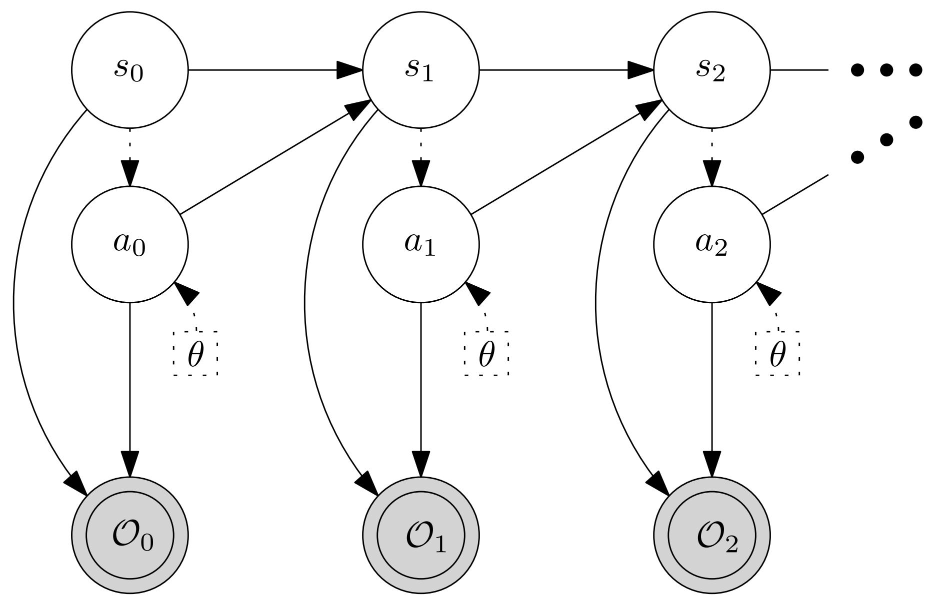
Reduces influence of entropy

$$\varepsilon_\omega = 0 \implies \hat{Q}_\omega(\cdot) = Q^*(\cdot)$$ **(Theorem 2)**

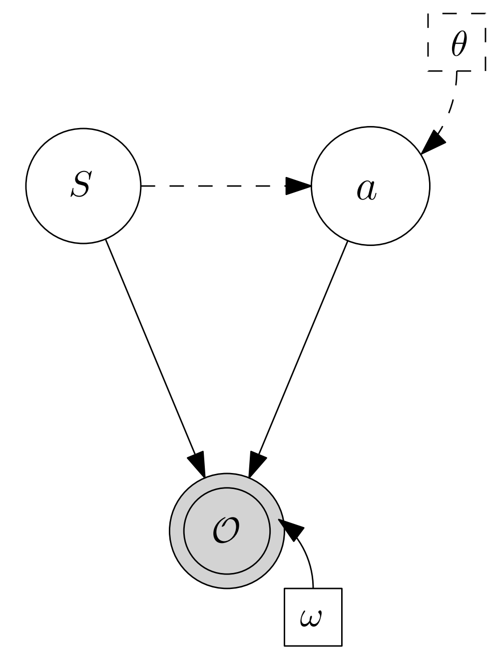
$$\lim_{\varepsilon_\omega \to 0} \varepsilon_\omega \mathscr{L}_\omega(\theta) = \mathbb{E}_{d(s)\pi_\theta(a|s)}\left[Q^*(h)\right] = J(\theta)$$

Hence $\pi^*(a\,|\,s)$ can still be found using e.g. classic policy gradient updates

Comparing MERL and VIREL



MERL



VIREL

In VIREL, $q_\theta(h)$ approximates the posterior for a single interaction, $\hat{Q}_\omega(h)$ models all future interactions

In MERL, $q_\theta(\tau)$ needs to model underlying long-term dynamics of the MDP

For high dimensional MDPs, expressiveness of $q_\theta(\tau)$ could be a bottleneck to performance (see experiments…)

A Simple Algorithm:

ELBO Objective: $\mathscr{L}_\omega(\theta) = \mathbb{E}_{d(s)}\left[\mathbb{E}_{\pi_\theta(a|s)}\left[\frac{\hat{Q}_\omega(h)}{\varepsilon_\omega}\right] + \mathscr{H}(\pi_\theta(\cdot\,|\,s))\right]$

$\mathscr{L}_\omega(\theta) \to \infty$ whenever $\varepsilon_\omega \to 0$ therefore treat $\mathscr{L}_\omega(\theta)$ as overall objective or, even simpler:

VEM/AC-style algorithm:

E-step (actor): $\theta_{k+1} \leftarrow \arg_\theta \max \mathscr{L}_{\omega_k}(\theta)$

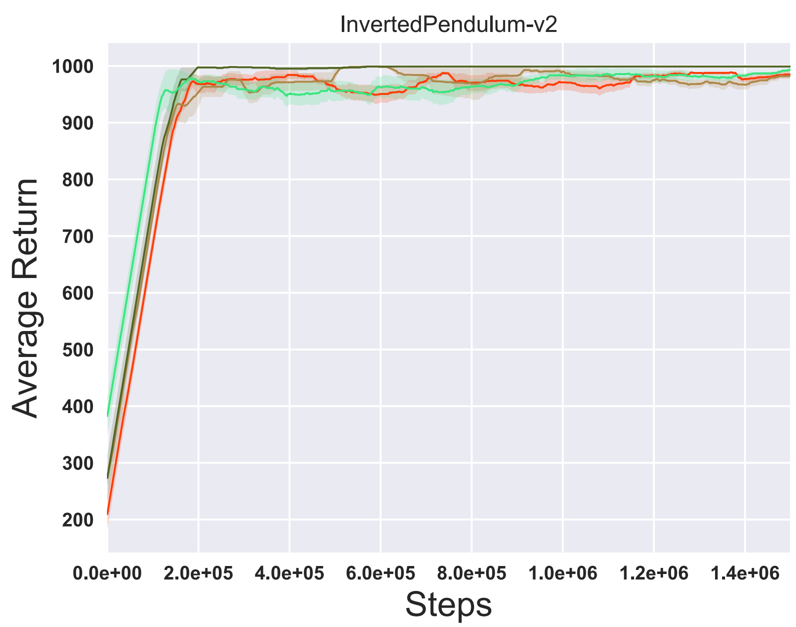
Using gradient based optimisation: $\theta_{i+1} \leftarrow \theta_i + \alpha_{ac}\left(\varepsilon_{\omega_k}\nabla_\theta \mathscr{L}_{\omega_k}(\theta)\,|_{\theta=\theta_i}\right)$

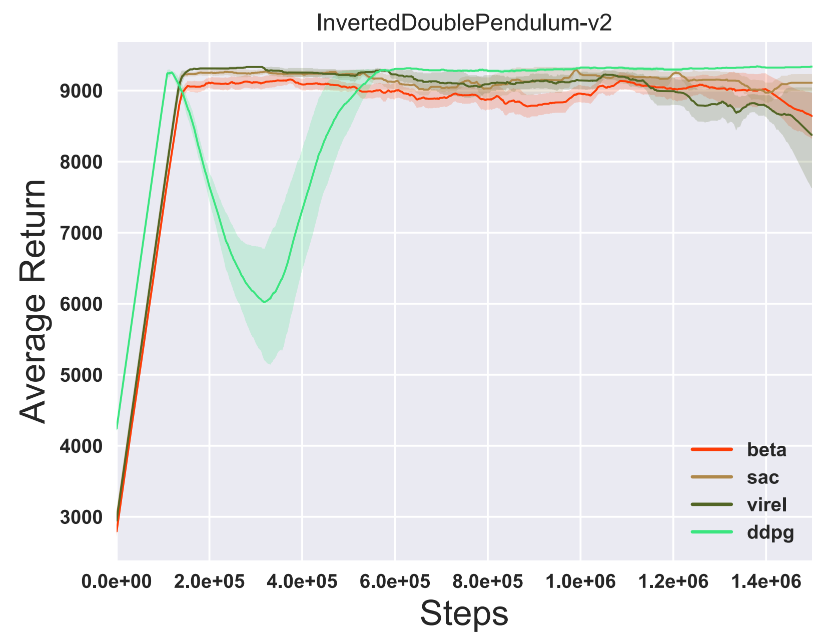
$$\varepsilon_{\omega_k}\nabla_\theta \mathscr{L}(\omega_k, \theta)\,|_{\theta=\theta_i} = \mathbb{E}_{d(s)}\left[\mathbb{E}_{\pi_\theta(a|s)}\left[\hat{Q}_{\omega_k}(h)\,\nabla_\theta \log \pi_\theta(a\,|\,s)\right] + \varepsilon_{\omega_k}\nabla_\theta \mathscr{H}(\pi_\theta(\cdot\,|\,s))\right]$$

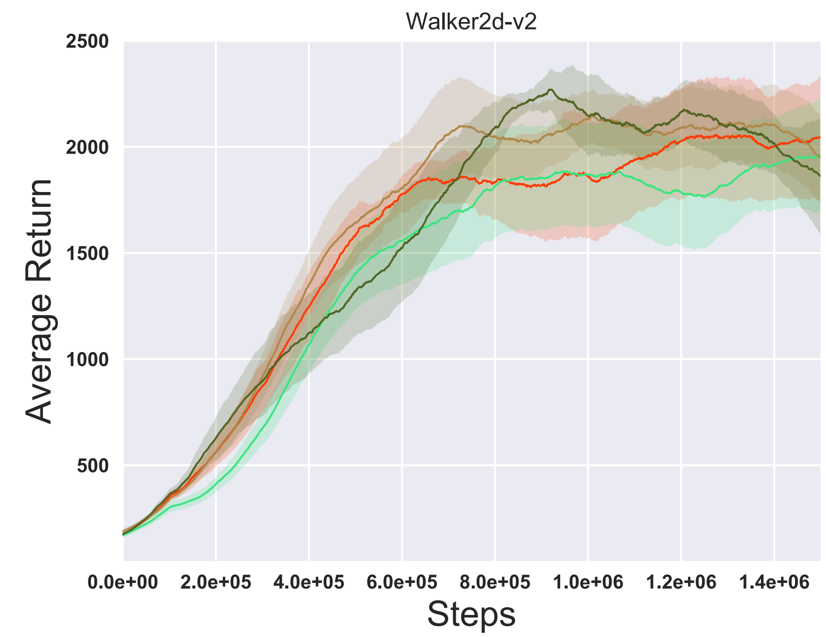
M-step (critic): Sample $\pi_{\theta_{k+1}}$ and update ω_{k+1} using gradient based optimisation:

$$\omega_{k+1} \leftarrow \omega_k - \alpha_{cr}\nabla_\omega \varepsilon_\omega\,|_{\omega=\omega_k}$$

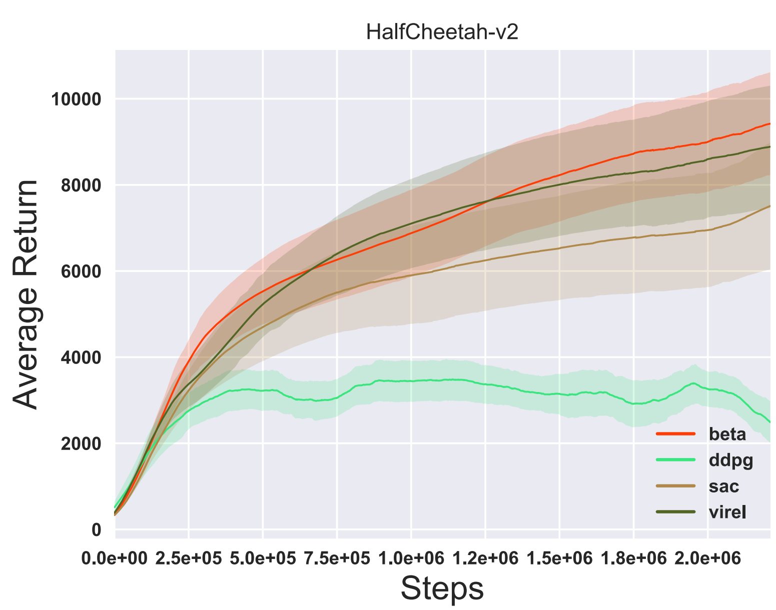
Results

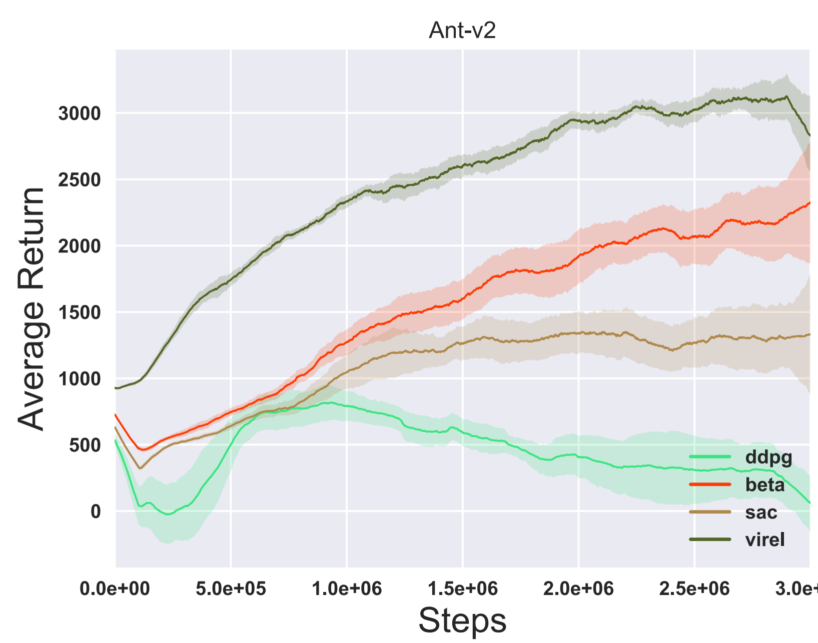


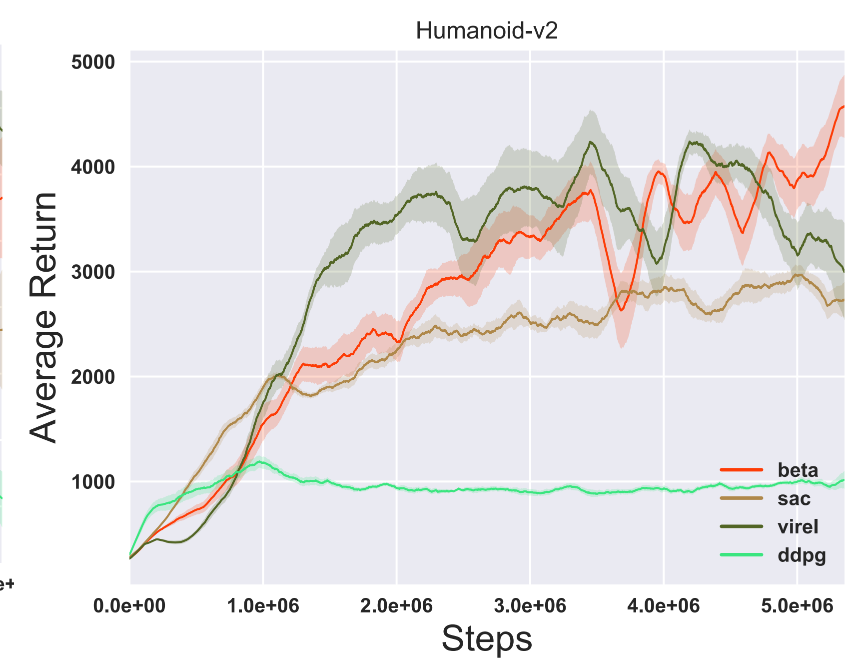




Lowest dimensional task → Higher dimensional tasks







Higher dimensional task → Highest dimensional tasks