VIREL: A Variational Inference Framework for Reinforcement Learning

Matthew Fellows*, Anuj Mahajan*, Tim G. J. Rudner, Shimon Whiteson (NeurIPS 2019)

Presented by: **Cem Anil & Jenny Bao**

University of Toronto | Vector Institute

Presentation outline

- Background: control as inference
- Problems with current approaches in probabilistic RL
- VIREL
  - Key ideas & key properties
  - Derived actor-critic algorithm
- Colab presentation
- Conclusion

Background: control as inference

This section is directly adapted from Sergey Levine's slides in CS285 lecture 19 (fall 2019)

`http://rail.eecs.berkeley.edu/deeprlcourse/static/slides/lec-19.pdf`

Background: control as inference

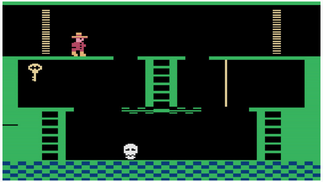
Conventional RL / optimal control:

$$a_1, \ldots, a_T = \arg\max_{a_1,\ldots,a_T} \sum_{t=1}^{T} r(s_t, a_t)$$

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

arg max finds *one* optimal sequence.

Does not model suboptimal, but reasonable behaviour. ✗



Background: control as inference

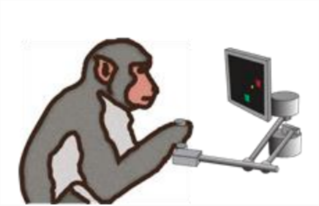
Conventional RL / optimal control:

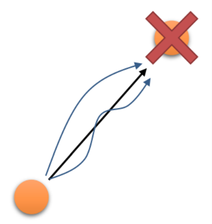
$$a_1, \ldots, a_T = \arg\max_{a_1,\ldots,a_T} \sum_{t=1}^{T} r(s_t, a_t)$$

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

arg max finds *one* optimal sequence.

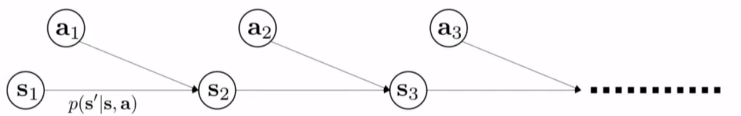
Does not model suboptimal, but reasonable behaviour. ✗





Background: control as inference

Model the decision making as a probabilistic graphical model (PGM)



Inference problem:

$$p(s_{1:T}, a_{1:T}) = ???$$

What's the probability of a trajectory?

But, no assumption of optimal behaviour!

Background: control as inference

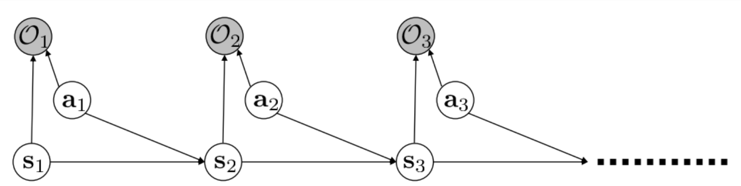
Introduce an optimality random variable $\mathcal{O}_t$.

- $\mathcal{O}_t = 1$ — optimal at time t.
- $\mathcal{O}_t = 0$ — not optimal at time t.

New inference problem:

$$p(\tau | \mathcal{O}_{1:T}) = ??? \qquad \text{where } \tau = (s_{1:T}, a_{1:T})$$

What's the probability of a trajectory, given that it is optimal at all timesteps?



Background: control as inference

Important assumption:

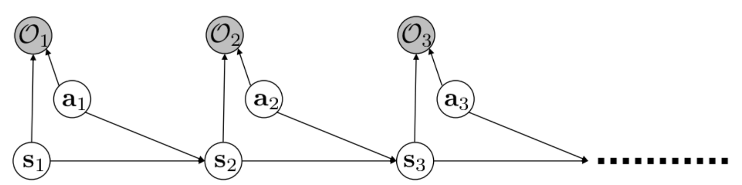
$$p(\mathcal{O}_t | s_t, a_t) \propto \exp(r(s_t, a_t))$$

This gives us a convenient form for $p(\tau | \mathcal{O}_{1:T})$:

$$p(\tau | \mathcal{O}_{1:T}) \propto p(\tau) \exp\left(\sum_t r(s_t, a_t)\right)$$

Higher-reward trajectories are exponentially more likely.

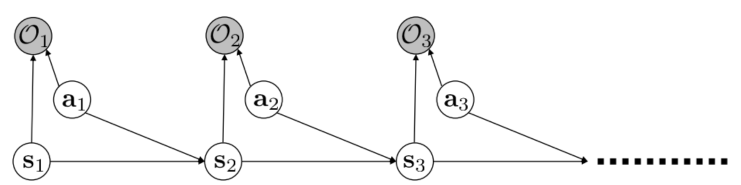
Background: control as inference



Benefits of modeling control as an inference problem:

- Can model suboptimal behaviour (important for inverse RL)
- Provides an explanation for why stochastic behaviour might be preferred.
- Can apply inference algorithms to solve control and planning problems

Background: control as inference



How to do inference?

1. compute backward messages $\beta_t(s_t, a_t) = p(\mathcal{O}_{t:T} | s_t, a_t)$ (Q-function)
2. compute policy $\pi(a_t | s_t) = p(a_t | s_t, \mathcal{O}_{t:T})$
3. compute forward messages $\alpha_t(s_t) = p(s_t | \mathcal{O}_{1:t-1})$

Background: control as variational inference

let $q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = p(\mathbf{s}_1) \prod_t p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) q(\mathbf{a}_t|\mathbf{s}_t)$

same dynamics and initial state as p

only new thing

Variational inference: match $p(\tau|\mathcal{O}_{1:T})$ with $q(s_{1:T}, a_{1:T})$:

$$\text{minimize} \;\; \mathrm{KL}\Big(q(\tau) \;\|\; p(\tau|\mathcal{O}_{1:T})\Big)$$

$\Leftrightarrow$ Maximize the variational lower bound on $\log p(\mathcal{O}_{1:T})$:

$$\log p(\mathcal{O}_{1:T}) \geq \cdots = \sum_t \mathbb{E}_{(s_t, a_t)\sim q}[\underbrace{r(s_t, a_t)}_{\text{reward}} + \underbrace{\mathcal{H}(\pi(a_t|s_t))}_{\text{action entropy}}]$$

Motivates maximum entropy RL

Background: control as variational inference

$$\mathrm{ELBO} = \mathbb{E}_{(s_t,a_t)\sim q}[\underbrace{r(s_t,a_t)}_{\text{reward}} + \underbrace{\mathcal{H}(\pi(a_t|s_t))}_{\text{action entropy}}]$$

The ELBO is maximized with the Boltzmann policy

$$\pi(a_t|s_t) \propto \exp(Q(s_t,a_t))$$

Problems with existing methods

Maximum Entropy RL methods (MERLIN)

- Recall: optimal policy $\pi(a_t|s_t) \propto \exp(Q(s_t, a_t))$
  - Sensitive to temperature (reward magnitude) ✗
  - Cannot learn deterministic policy (temperature is fixed and non-zero) ✗

Pseudo likelihood methods

- Minimizes $\mathrm{KL}\big(p(\tau|\mathcal{O}_{1:T})\|q(\tau)\big)$ instead of $\mathrm{KL}\big(q(\tau)\|p(\tau|\mathcal{O}_{1:T})\big)$
  - Favours risk-seeking policy ✗

VIREL — desiderata

Desired properties of the VIREL objective:

1. When objective is maximized, policy should be deterministic
2. When objective is not maximized, policy should be stochastic
3. Should minimize the "correct" (risk-neutral) KL — $\mathrm{KL}\big(q(\tau) \| p(\tau|\mathcal{O}_{1:T})\big)$

VIREL — key idea

Key idea: Boltzmann policy with adaptive temperature

$$\pi_\omega(a|s) := \frac{\exp(\frac{\hat{Q}_\omega(s,a)}{\epsilon_\omega})}{\int_{\mathcal{A}} \exp(\frac{\hat{Q}_\omega(s,a)}{\epsilon_\omega}) da}$$

- $\hat{Q}_\omega(s, a)$ is the approximate Q-function parameterized by ω.
- ϵ_ω is the adaptive temperature, defined as the Bellman error:

$$\epsilon_\omega := \frac{c}{p} \| \mathcal{T}_\omega \hat{Q}_\omega(s, a) - \hat{Q}_\omega(s, a) \|_p^p$$

where $\mathcal{T}_\omega(\cdot) = r(s, a) + \gamma \mathbb{E}_{(s',a') \sim p(s'|s,a)\pi_\omega(a'|s')}[\cdot]$ is the Bellman operator, $c > 0$ is an arbitrary constant, and assume $p = 2$ WLOG.

VIREL — objective

First try on the objective:

$$\arg\min \mathcal{L}(\omega) := \arg\min \underbrace{\frac{c}{p}\|\mathcal{T}_\omega \hat{Q}_\omega(s,a) - \hat{Q}_\omega(s,a)\|_p^p}_{\epsilon_\omega}$$

Check the desiderata:

1. Main result: finding optimal ω^* such that $\epsilon_{\omega^*} = 0 \implies$ Q function is optimal $\hat{Q}_\omega(s,a) = Q^*(s,a)$.
   $\implies \pi_{\omega^*}(a|s) = \delta(a = \arg\max_{a'} \hat{Q}_{\omega^*}(a', s))$ is the deterministic optimal policy ✓
2. When the objective is not optimized ($\epsilon_\omega > 0$), the temperature is positive, and $\pi_\omega(a|s)$ is stochastic ✓

VIREL — objective

However, it's intractable to compute the normalization constant of the Boltzmann policy

$$\pi_\omega(a|s) := \frac{\exp(\frac{\hat{Q}_\omega(s,a)}{\epsilon_\omega})}{\int_{\mathcal{A}} \exp(\frac{\hat{Q}_\omega(s,a)}{\epsilon_\omega}) da}$$

Solution: learn a variational policy

$$\pi_\theta(a|s) \approx \pi_\omega(a|s)$$

VIREL — objective

New objective:

$$\mathcal{L}(\omega,\theta) = \mathbb{E}_{s\sim d(s)}\left[\mathbb{E}_{a\sim\pi_\theta(a|s)}\left[\frac{\hat{Q}_\omega(s,a)}{\epsilon_\omega}\right] + \mathcal{H}\left(\pi_\theta(a|s)\right)\right]$$

where $d(s)$ is an arbitrary sampling distribution for the state.

- Can check that it still satisfies desiderata 1 and 2 ✓
- Can show that it minimizes the risk-neutral KL:

$$\mathcal{L}(\omega,\theta) = \log\int_{\mathcal{S}x\mathcal{A}} \exp\left(\frac{\hat{Q}_\omega(s,a)}{\epsilon_\omega}\right) - \mathrm{KL}(q_\theta(s,a)\|p_\omega(s,a)) - \mathcal{H}(d(s))$$

It satisfies desiderata 3 as well ✓

VIREL — derived actor-critic

VIREL provides a variational framework for probabilistic RL, from which we can derive specific algorithms.

A natural derivation: EM $\longleftrightarrow$ variational actor-critic

VIREL — derived actor-critic

Use EM to optimize the objective $\longleftrightarrow$ actor-critic

$$\mathcal{L}(\omega, \theta) = \mathbb{E}_{s \sim d(s)}\left[\mathbb{E}_{a \sim \pi_\theta(a|s)}\left[\frac{\hat{Q}_\omega(s, a)}{\epsilon_\omega}\right] + \mathcal{H}\big(\pi_\theta(a|s)\big)\right]$$

E-step (actor)

$$\theta_{i+1} = \theta_i + \alpha_{\mathrm{actor}} \epsilon_{\omega_k} \nabla_\theta \mathcal{L}(\omega_k, \theta_i)$$

M-step (critic)

$$\omega_{i+1} = \omega_i + \alpha_{\mathrm{critic}} \epsilon^2_{\omega_i} \nabla_\omega \mathcal{L}(\omega_i, \theta_{k+1})$$

A lot of techniques in advanced actor-critic methods naturally apply here (e.g. control variates, baselines, ...)

VIREL — experiments

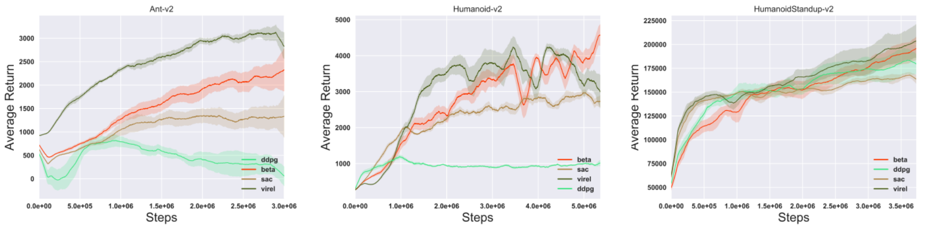


Figure 3: Training curves on continuous control benchmarks gym-Mujoco-v2 : High-dimensional domains.

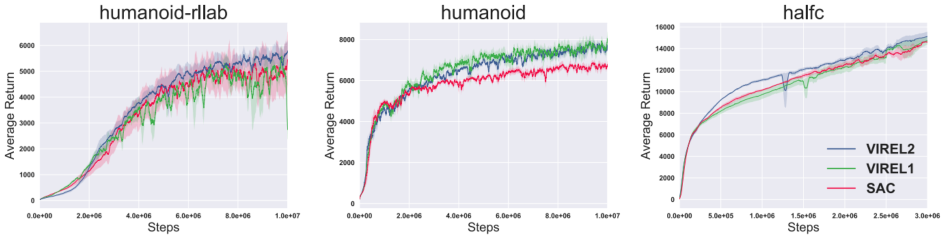


Figure 4: Training curves on continuous control benchmarks gym-Mujoco-v1.

Colab presentation

Link to Colab Notebook

Summary

- RL as an inference problem
- Existing methods in probabilistic RL suffer from various issues
- VIREL
  - Key idea: Boltzmann policy with adaptive temperature
  - Variational objective that satisfies all the desiderata
  - Naturally derived actor-critic algorithm

Scope & limitations

The experiments are on the Mujoco continuous control tasks. It's unclear how VIREL performs on tasks with discrete state / action spaces, or tasks with higher dimensional inputs (e.g. pixel input).

To accurately estimate ϵ_ω can be costly. Current implementation uses a buffer to reduce sample complexity, but may introduce complications to the learning dynamics. An immediate future work is to find better estimates of ϵ_ω.

Another direction for future work: extend the framework to multi-agent settings.

Thank you for listening!