# VIREL: A Variational Inference Framework for Reinforcement Learning

Matthew Fellows, Anuj Mahajan, Tim G. J. Rudner, Shimon Whiteson

{matthew.fellows,anuj.mahajan,tim.rudner,shimon.whiteson}@cs.ox.ac.uk

UNIVERSITY OF OXFORD

AIMS Autonomous Intelligent Machines & Systems

THE RHODES TRUST

## Abstract


- VIREL is a novel, theoretically grounded probabilistic inference framework for reinforcement learning (RL) that utilises the action-value function in a parametrised form to capture future dynamics of the underlying Markov decision process
- Applying the variational expectation-maximisation algorithm to our framework, we show that the actor-critic algorithm can be reduced to expectation-maximization
- VIREL is a more flexible and mathematically grounded alternative to existing RL-as-inference frameworks such as the maximum entropy or pseudo-likelihood approaches


## Reinforcement Learning as Inference

- The RL problem is to find an optimal policy $\pi^*(a|s) \in \Pi^* \triangleq \arg\max_\pi J^\pi$, where $J^\pi \triangleq \int Q^\pi(h)p_0(s)\pi(a|s)dh$
- RL as inference approaches recast the RL problem as an inference problem in which maximizing a marginal likelihood is equivalent to maximizing the reward function $J^\pi$
- **Existing RL-as-inference frameworks:**
  - Maximum Entropy RL: no closed-form updates for the parameters of value functions without using approximations
  - Pseudo-Likelihood RL: promotes risk-seeking policies

## Variational Expectation-Maximization

- In variational inference, we seek to maximize the marginal likelihood, $p(x)$
- For any valid probability distribution $q(h)$ over $h$ we can rewrite the log-marginal likelihood objective as a difference of two KL divergences,

$$\begin{aligned}\mathcal{L}(x;\omega) &= \int q(h)\log\left(\frac{p(x,h;\omega)}{q(h)}\right)dh - \int q(h)\log\left(\frac{p(h|x;\omega)}{q(h)}\right)dh,\\ &=\mathrm{ELBO}\left(q(h);\omega\right) + \mathbf{KL}(q(h)\parallel p(h|x;\omega)),\end{aligned}$$

where $\mathrm{ELBO}\left(q(h)\right) \triangleq \int q(h)\log\left(\frac{p(x,h)}{q(h)}\right)dh$ is known as the evidence lower bound

- **Variational expectation-maximization:**

$$\text{Variational E-Step: } \theta_{n+1} \leftarrow \arg\max_\theta \mathrm{ELBO}\left(q(h;\theta);\omega_n\right)$$

$$\text{Variational M-Step: } \omega_{n+1} \leftarrow \arg\max_\omega \mathrm{ELBO}\left(q(h;\theta_{n+1});\omega\right)$$

## VIREL Framework

- Optimality of reward:

$$p(\mathcal{O}|h;\omega) = \exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right)^{\mathcal{O}}\left(1-\exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right)\right)^{(1-\mathcal{O})}$$

- Mean squared Bellman error (MSBE):

$$\beta(\omega) = \mathbb{E}_{h\sim p(h|\mathcal{O};\omega)}\left[\left(Q^{p_\omega}(h)-\hat{Q}(h;\omega)\right)^2\right]$$

- $Q^{p_\omega}(h)$ is the target $Q$-function: the action-value of the policy corresponding to the action-posterior distribution, $p(a|s,\mathcal{O};\omega)$
- Inference objective:

$$\mathrm{ELBO}(q,\omega) = \frac{\int Q^{p_\omega}(h)p_0(s)\pi^q(a|s)dh}{\beta(\omega)} + \mathcal{H}(q(h)) + \mathbb{E}_{h\sim q(h)}[\log(p(h))]$$

| DISTRIBUTION/ FUNCTION | DEFINITION |
|---|---|
| Conditional Likelihood | $p(\mathcal{O}|h;\omega) = \exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right)$ |
| Posterior $Q$-function | $Q^{p_\omega}(h) = \int\left(\sum_{i=0}^{\infty}\gamma^i r_i\right)p^{p(a|s,\mathcal{O};\omega)}(\tau|h)d\tau$ |
| Mean Squared Bellman Error | $\beta(\omega) = \mathop{\mathbb{E}}_{h\sim p(h|\mathcal{O};\omega)}[(Q^{p_\omega}(h)-\hat{Q}(h;\omega))^2]$ |
| Prior | $p(h) = \mathcal{U}(h)$ |
| Joint | $p(\mathcal{O},h;\omega) = \exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right)p(h)$ |
| Posterior | $p(h|\mathcal{O};\omega) = \frac{\exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right)p(h)}{\int\exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right)dh}$ |
| Action-posterior | $p(a|s,\mathcal{O};\omega) = \frac{\exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right)}{\int\exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right)da}$ |
| Variational Posterior | $q(h) = p_0(s)\pi^q(a|s)$ |
| Log-likelihood | $\mathcal{L}(\mathcal{O};\omega) = \mathrm{ELBO}(q,\omega) + \mathbf{KL}(q(h)\parallel p(h|\mathcal{O};\omega))$ |
| Evidence Lower Bound | $\mathrm{ELBO}(q,\omega) = \int q(h)\log\left(\frac{p(\mathcal{O},h;\omega)}{q(h)}\right)dh$ |

## Main Results

**Lemma 1** (Characterisation of posterior)**.** *If all optimal policies and corresponding optimal $Q$-functions can be represented exactly by distributions parametrised by $\omega$, then the action-posterior $p(a|s,\mathcal{O};\omega)$ defines a soft policy with respect to $Q^{p_\omega}(h)$ with the temperature given by the residual error $\beta(\omega)$. In the limit $\lim_{\beta(\omega)\to 0}p(a|s,\mathcal{O};\omega)$ is greedy with respect to $Q^{p_\omega}(h)$.*

**Theorem 1** (Optimal Posterior Distributions as Optimal Policies)**.** *For any $\omega$ that maximizes $\mathcal{L}(\omega)$, the corresponding policy induced must be optimal, i.e.,*

$$\omega^* \in \arg\max_\omega \mathcal{L}(\omega) \implies p(a|s,\mathcal{O};\omega^*) \in \arg\max_\pi J^\pi.$$

## Variational Actor-Critic Algorithm

**Variational E-Step (Actor):**

$$\theta_{k+1} \leftarrow \theta_k + \alpha_{\text{actor}}\nabla_\theta\mathrm{ELBO}\left(\omega_k,\theta\right),$$

with

$$\nabla_\theta\mathrm{ELBO}\left(\omega_k,\theta\right) = \nabla_\theta\sum_{t=1}^{T-1}\int\hat{Q}(s_t,a;\omega)\pi^q(a|s_t;\theta)da + \beta(\omega)\nabla_\theta\sum_{t=1}^{T-1}\mathcal{H}(\pi^q(a|s_t;\theta)),$$

where we have used a $T$ time step Monte Carlo estimation of the outer expectation with respect to $s$

**Variational M-Step (Critic):**

$$\omega_{k+1} \leftarrow \omega_k + \alpha_{\text{critic}}\nabla_\omega\mathrm{ELBO}\left(\omega,\theta_{k+1}\right),$$

with

$$\nabla_\omega\mathrm{ELBO}\left(\omega,\theta_{k+1}\right) = \mathbb{E}_{h\sim q(h;\theta_{k+1})}\left[\nabla_\omega\hat{Q}(h;\omega)\left(\psi(h)-\hat{Q}(h;\omega)\right)\right].$$

Our choice of estimate $\psi(h_0)$ thus determines the form of policy evaluation. We can recover, for example, recover $Q$-learning by letting $\psi(h) = r(h) + \gamma\max_a'\hat{Q}(h';\omega_k)$

## Conclusions

- Owing to its generality, our framework is amenable by a wide range of variational inference methods
- Our framework does not suffer from the same shortcomings as existing RL-as-inference methods
- An empirical evaluation showed that VIREL outperforms or performs on par with current state-of-the-art RL models, performing particularly well in difficult high-dimensional domains (such as MuJoCo humanoid)