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

Abstract

• VIREL is a novel, theoretically grounded probabilistic inference framework for reinforcement learning (RL) that utilizes 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-maximisation

• 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 \( \pi^* (s) \in \Pi \) \( = \arg \max_s J^* \), where \( J^* \triangleq \int Q^*(b) \pi(s) \, db \)

• RL as inference problems can be formulated as inference problems in which maximizing a marginal likelihood is equivalent to maximizing the reward function \( J^* \)

• 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,

\[
\mathcal{L}(x; \omega) = \int q(h) \log \left( \frac{p(x, h; \omega)}{q(h)} \right) \, dh \quad \text{and} \quad \mathcal{L}(x; \omega) = \text{ELBO}(q(h)) + \text{KL}(q(h) \| p(h; x, \omega)),
\]

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

• Variational expectation-maximization:

  Variational E-Step: \( \theta_{i+1} \leftarrow \arg \max_\theta \text{ELBO}(q(h; \theta), \omega_i) \)

  Variational M-Step: \( \omega_{i+1} \leftarrow \arg \max_\omega \text{ELBO}(q(h; \theta_{i+1}), \omega) \)

VIREL Framework

• Optimality of reward:

\[
p(O|h, \omega) = \exp \left( \frac{Q^*(b)}{\beta(h)} \right)^\omega \left( 1 - \exp \left( \frac{Q^*(b)}{\beta(h)} \right)^{1-\omega} \right)
\]

• Mean squared Bellman error (MSBE):

\[
\beta(h) = \mathbb{E}_{b \sim p(h; \omega)} \left( Q^*(h) - Q(h; \omega) \right)^2
\]

• Inference objective:

\[
\text{ELBO}(q_h; \omega) = \int Q^*(b) p(h) q(h) \, db + \mathcal{H}(p(h)) + \mathbb{E}_{b \sim p(h; \omega)} \left( \log p(h) \right)
\]

Main Results

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

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.,

\[
\omega^* \in \arg \max_\omega \mathcal{L}(\omega) \Rightarrow p(a|s, \omega^*) \in \arg \max_a J^*
\]

Variational Actor-Critic Algorithm

Variational E-Step (Actor):

\[
\theta_{i+1} \leftarrow \theta_i + \alpha_{\text{act}} \nabla_{\theta_i} \mathbb{E}_{b \sim p(h; \omega_i)} \left[ Q^*(h; \theta) \right]
\]

with

\[
\nabla_{\theta_i} \mathbb{E}_{b \sim p(h; \omega_i)} = \mathbb{E}_{b \sim p(h; \omega_i)} \left[ \int Q^*(h; \theta) \, db + \beta(h) \nabla_{\theta_i} \mathcal{H}(p(h)) \right],
\]

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

Variational M-Step (Critic):

\[
\omega_{i+1} \leftarrow \omega_i + \alpha_{\text{crit}} \nabla_{\omega} \mathbb{E}_{b \sim p(h; \omega_i)} \left[ \mathcal{H}(Q^*(h; \theta_i)) \right]
\]

with

\[
\nabla_{\omega} \mathbb{E}_{b \sim p(h; \omega_i)} = \mathbb{E}_{b \sim p(h; \omega_i)} \left[ \int \left( \psi(h) - Q(h; \omega_i) \right) \, db \right]
\]

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

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)