VIREL:
A Variational Inference Framework for Reinforcement Learning

Anuj Mahajan

WhiRL, University of Oxford

Joint work with Matt and Shimon, NeurIPS 2019

Outline

Reinforcement Learning(RL) problem

Markov decision process (MDP) defined by the tuple $\langle S, A, r, p, p_0, \gamma \rangle$

- S is the set of states
- $A \subseteq \mathbb{R}^n$ the set of available actions
- $h \triangleq \langle s, a \rangle$
- $r(h_t) \in \mathbb{R}^+$ is the reward at time t
- p is the transition kernel of MDP
- p_0 is the initial state distribution
- γ is discount factor
- $\pi(a|s)$ is the policy

RL problem continued

$$Q^{\pi}(h) \triangleq \int \left(\sum_{t=0}^{\infty} \gamma^t r_t \right) p^{\pi}(\tau|h) d\tau, \tag{1}$$

Here, $\tau = (h_0, r_0, h_1, r_1, ...)$ is the trajectory under policy π.
The objective is:

$$\arg\max_{\pi} J^{\pi} \triangleq \int Q^{\pi}(h) p_0(s) \pi(a|s) dh. \tag{2}$$

Casting RL as Inference

- Use a "optimality" variable $\mathcal{O}$ that relates to returns of the MDP. $\mathcal{O} = 1$ for the rest of discussion.
- Maximum entropy reinforcement learning(MERL) : augment rewards with action entropy

$$J^{\pi}_{\text{merl}} \triangleq \mathbb{E}_{\tau \sim p(\tau)} \left[\sum_{t=0}^{T-1} \left(r_t - c \log(\pi(a_t|s_t)) \right) \right], \tag{3}$$

- Pseudo Likelihood approaches: Define a trajectory distribution by multiplying actual probabilities with returns.

$$J_{PL} = \int R(\tau) p(\tau) d\tau, \tag{4}$$

- Don't capture mode as min $KL\langle p_{\mathcal{O}}, p_{\pi} \rangle$

Main contributions: VIREL

- Exact reduction of entropy regularised RL to inference.
- Reconciling empirically successful methods with accurate theoretical motivation.
- Framework for developing inference style algorithms for RL.
- Adaptive weighing of "exploration driving" entropy and RL objective.

VIREL

- $p(\mathcal{O} = 1|h;\omega) \propto \exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right)$
- Mean Squared Bellman Error
  $\beta(\omega) = \mathop{\mathbb{E}}_{h\sim p(h|\mathcal{O};\omega)}[(Q^{p_\omega}(h) - \hat{Q}(h;\omega))^2]$
- Joint $p(\mathcal{O}, h;\omega) = \exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right) p(h)$
- $p(h)$ is assumed to be uniform.
- $Q^{p_\omega}(h)$ is the action-value function of action posterior $p(a|s,\mathcal{O};\omega)$ of the joint distribution.
- $\hat{Q}(h;\omega)$ is the approximator of $Q^{p_\omega}(h)$ parametrized by ω

Action posterior

Marginalising out a from the joint we obtain:

$$p(s|\mathcal{O};\omega) = \int p(h|\mathcal{O};\omega)da, \tag{5}$$

$$= \frac{\int \exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right) p(h)da}{\int \exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right) p(h)dh}. \tag{6}$$

Action posterior

$$p(a|s,\mathcal{O};\omega) = \frac{p(h|\mathcal{O};\omega)}{p(s|\mathcal{O};\omega)}, \tag{7}$$

$$= \frac{\exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right) p(h)}{\int \exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right) p(h)dh} \cdot \frac{\int \exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right) p(h)dh}{\int \exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right) p(h)da}, \tag{8}$$

$$= \frac{\exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right) \mathcal{U}(h)}{\int \exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right) \mathcal{U}(h)da}, \tag{9}$$

$$= \frac{\exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right)}{\int \exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right) da}, \tag{10}$$

A closer look at action posterior

$$p(a|s,\mathcal{O};\omega) = \frac{\exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right)}{\int \exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right) da} \tag{11}$$

- Action posterior "encodes" for the optimal policy

Theorem

The action-posterior from defines a soft policy with respect to $Q^{p_\omega}(h)$ *with the temperature given by the residual error* $\beta(\omega)$*, under realisation assumption, in the limit* $\lim_{\beta(\omega)\to 0} p(a|s,\mathcal{O};\omega)$ *is greedy with respect to* $Q^{p_\omega}(h)$*.*

Inference problem and approximation

- The ML problem is typically intractable

$$\arg\max_{\omega} p(\mathcal{O};\omega) = \arg\max_{\omega} \int \exp\left(\frac{Q^{p_\omega}(h)}{\beta(\omega)}\right) p(h)dh \quad (12)$$

- Introduce variational distribution $q(h) = p_0(s)\pi^q(a|s)$ to get lower bound.(here on we assume π^q is parameterised by θ)
- Here $\mathcal{L}(\mathcal{O};\omega) = log[p(\mathcal{O};\omega)]$,

$$\mathcal{L}(\mathcal{O};\omega) = \mathrm{ELBO}(q,\omega) + KL\langle q(h), p(h|\mathcal{O};\omega)\rangle \quad (13)$$

More results

$$\mathrm{ELBO}(\omega, \theta) = \frac{\int Q^{p_\omega}(h) p_0(s) \pi^q(a|s;\theta) dh}{\beta(\omega)} \tag{14}$$

$$+ \mathcal{H}(q(h;\theta)) + \mathbb{E}_{h\sim q(h;\theta)}[\log(p(h))]. \tag{15}$$

Theorem

For any pair $\{\omega^*, \theta^*\}$ that maximises $\mathrm{ELBO}(\omega, \theta)$, the corresponding variational policy induced must be optimal, i.e.,

$$\{\omega^*, \theta^*\} \in \arg\max_{\omega,\theta} \mathrm{ELBO}(\omega, \theta) \tag{16}$$

$$\implies \pi^q(a|s;\theta^*) = p(a|s, \mathcal{O}; \omega^*) \in \arg\max_{\pi} J^\pi. \tag{17}$$

More results

Actor-Critic algorithms can be recovered from VIREL using Expectation-Maximisation on the $ELBO(\theta, \omega)$ objective.

- Variational E-Step (Actor):

$$\theta_{k+1} \leftarrow \theta_k + \alpha_{\text{actor}} \nabla_\theta \text{ELBO}\left(\omega_k, \theta\right), \tag{18}$$

- Variational M-Step (Critic):

$$\omega_{k+1} \leftarrow \omega_k + \alpha_{\text{critic}} \nabla_\omega \hat{\beta}(\omega), \tag{19}$$

Actor-Critic correspondence: E-step

$$\nabla_\theta \mathrm{ELBO}\,(\omega_k, \theta) = \nabla_\theta \int \frac{\hat{Q}(h;\omega)\pi^q(a|s;\theta)p_0(s)}{\beta(\omega)} dh + \tag{20}$$

$$\nabla_\theta \int \mathcal{H}(\pi^q(a|s;\theta))p_0(s)ds, \tag{21}$$

$$\propto \nabla_\theta \int \hat{Q}(h;\omega)\pi^q(a|s;\theta)p_0(s)dh + \tag{22}$$

$$\beta(\omega)\nabla_\theta \int \mathcal{H}(\pi^q(a|s;\theta))p_0(s)ds. \tag{23}$$

Actor-Critic correspondence: M-step

$$\arg\max_{\omega} \mathrm{ELBO}\,(\omega, \theta_{k+1}) = \tag{24}$$

$$\arg\max_{\omega} \int q(h;\theta_{k+1}) \log\left(\frac{p(\mathcal{O}, h;\omega)}{q(h;\theta_{k+1})}\right) dh, \tag{25}$$

$$= \arg\max_{\omega} \left(\int \frac{Q^{p_\omega}(h)}{\beta(\omega)} q(h;\theta_{k+1}) dh + \right. \tag{26}$$

$$\left. \mathcal{H}(q(h;\theta_{k+1})) + \int q(h;\theta_{k+1}) \log p(h;\theta_{k+1}) dh \right), \tag{27}$$

$$= \arg\max_{\omega} \frac{\int Q^{p_\omega}(h) q(h;\theta_{k+1}) dh}{\beta(\omega)}, \tag{28}$$

$$= \arg\min_{\omega} \frac{\beta(\omega)}{\int Q^{p_\omega}(h) q(h;\theta_{k+1}) dh}, \tag{29}$$

$$= \arg\min_{\omega} \beta(\omega). \tag{30}$$

Actor-Critic correspondence

- E-step :

$$\nabla_\theta \mathrm{ELBO}\,(\omega_k, \theta) = \nabla_\theta \sum_{t=1}^{T-1} \int \hat{Q}(s_t, a; \omega) \pi^q(a|s_t; \theta) da \quad (31)$$

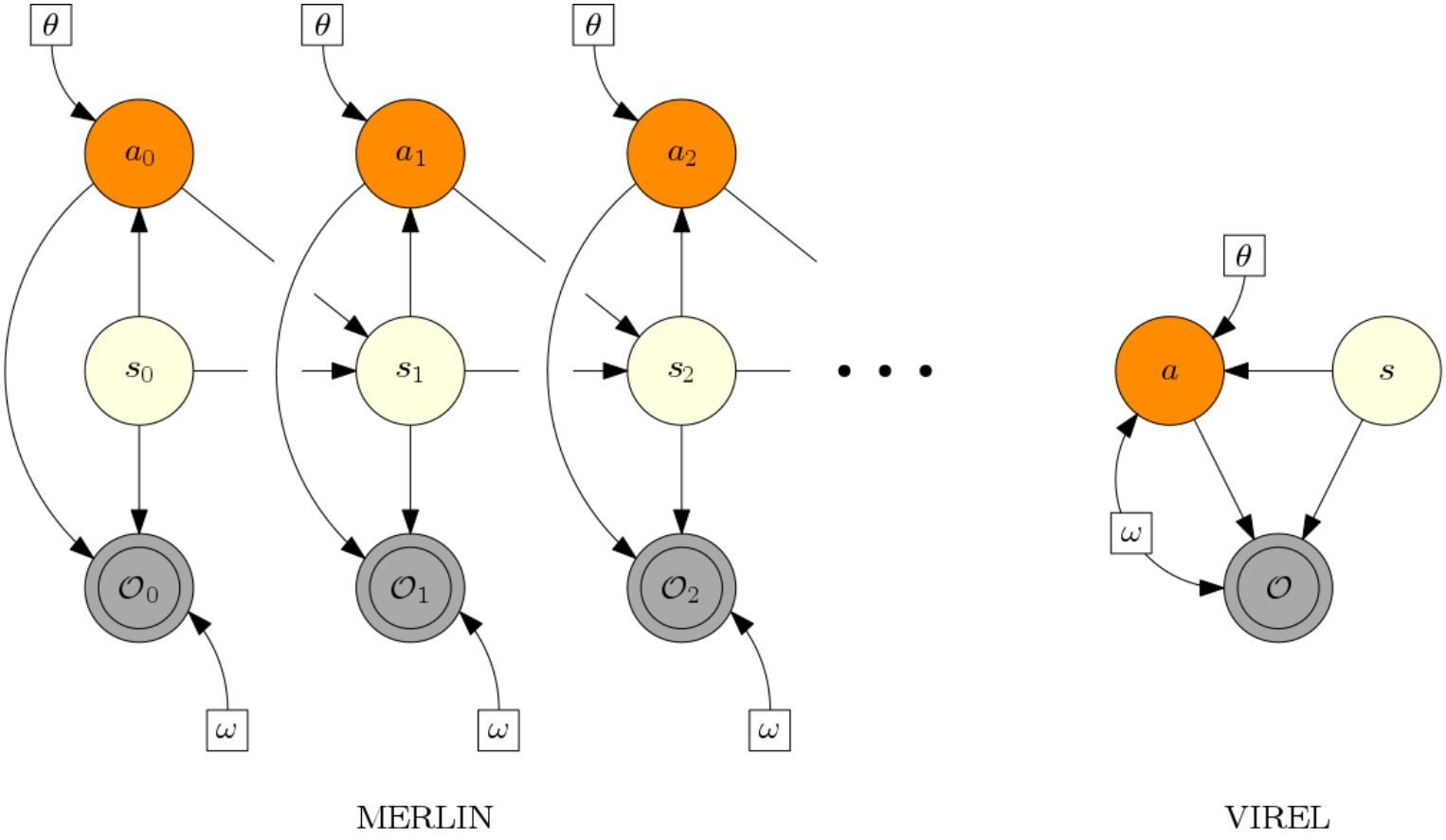
$$+ \beta(\omega) \nabla_\theta \sum_{t=1}^{T-1} \mathcal{H}(\pi^q(a|s_t; \theta)), \quad (32)$$

- M-step:

$$\nabla_\omega \hat{\beta}(\omega) = \mathbb{E}_{h \sim \mathcal{D}(h)} \left[\nabla_\omega \hat{Q}(h; \omega) \left(\psi(h) - \hat{Q}(h; \omega) \right) \right] \quad (33)$$

- Where $\mathcal{D}(h)$ ensures visitation to all state-actions and $\psi(h)$ is the target. ex. $r(h) + \gamma \max_{a'} \hat{Q}(a', s; \omega_k)$ for Q-learning.

Graphical models



- MERL inference based approaches typically model entire trajectories.
- VIREL summarizes future dynamics of the MDP using an action-value function.

Experiments

- Comparison against SAC and DDPG
- Two versions of actor critic algorithms derived from VIREL were compared:
  - Initial version using estimate for $\beta(\omega)$ for scaling entropy gradient (*beta*).
  - Version using fixed scaling (*virel*).

Experiments: Mujoco:v2

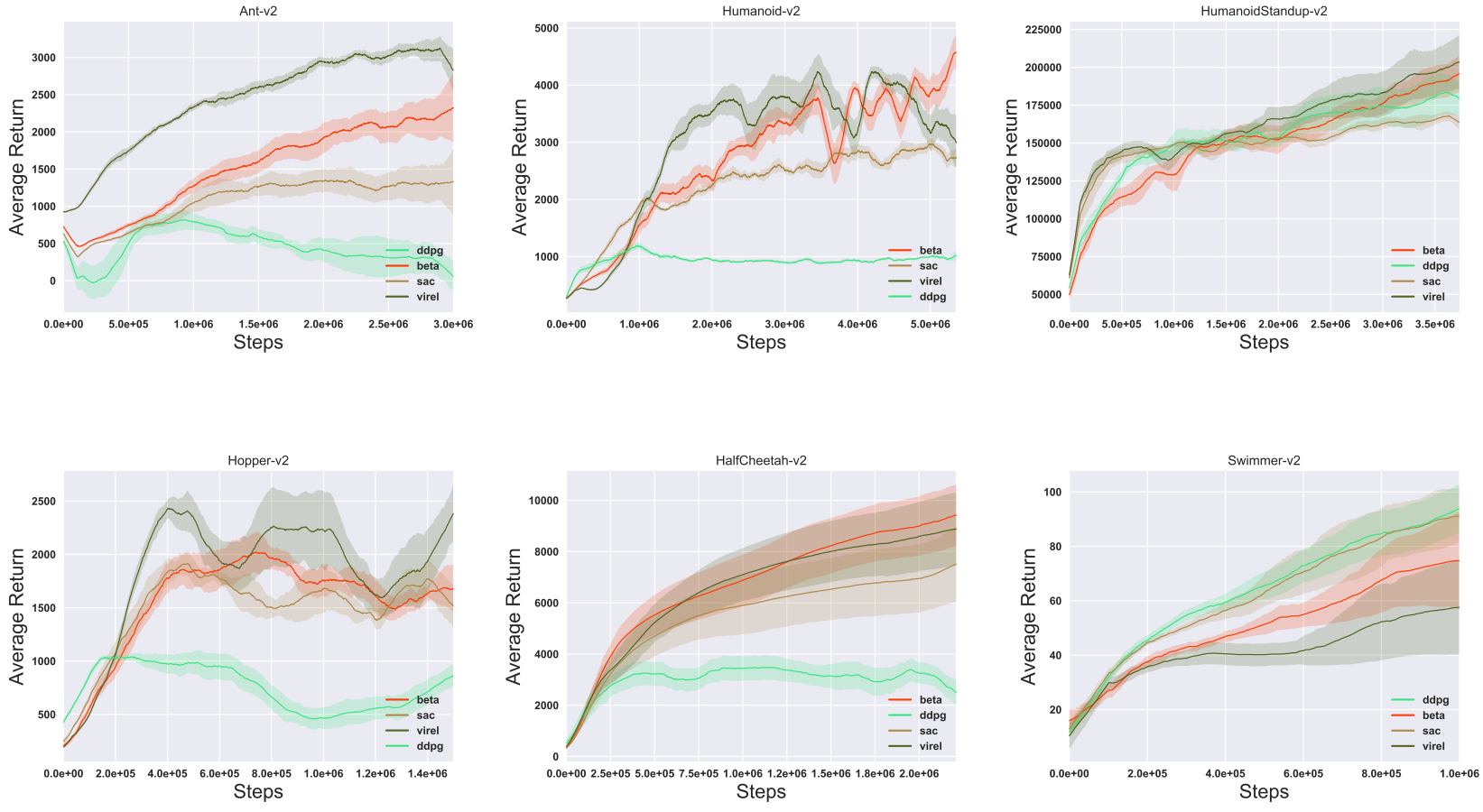


Figure 1: Training curves gym-Mujoco-v2

Experiments

- Comparison against SAC baseline using TD3 based bias minimization.
- Two versions of actor critic algorithms derived from VIREL were compared:
  - Twin critics for reducing bias
  - VIREL1 uses regular Q functions
  - VIREL2 uses a regular and a soft Q function.

Experiments: Mujoco:v1

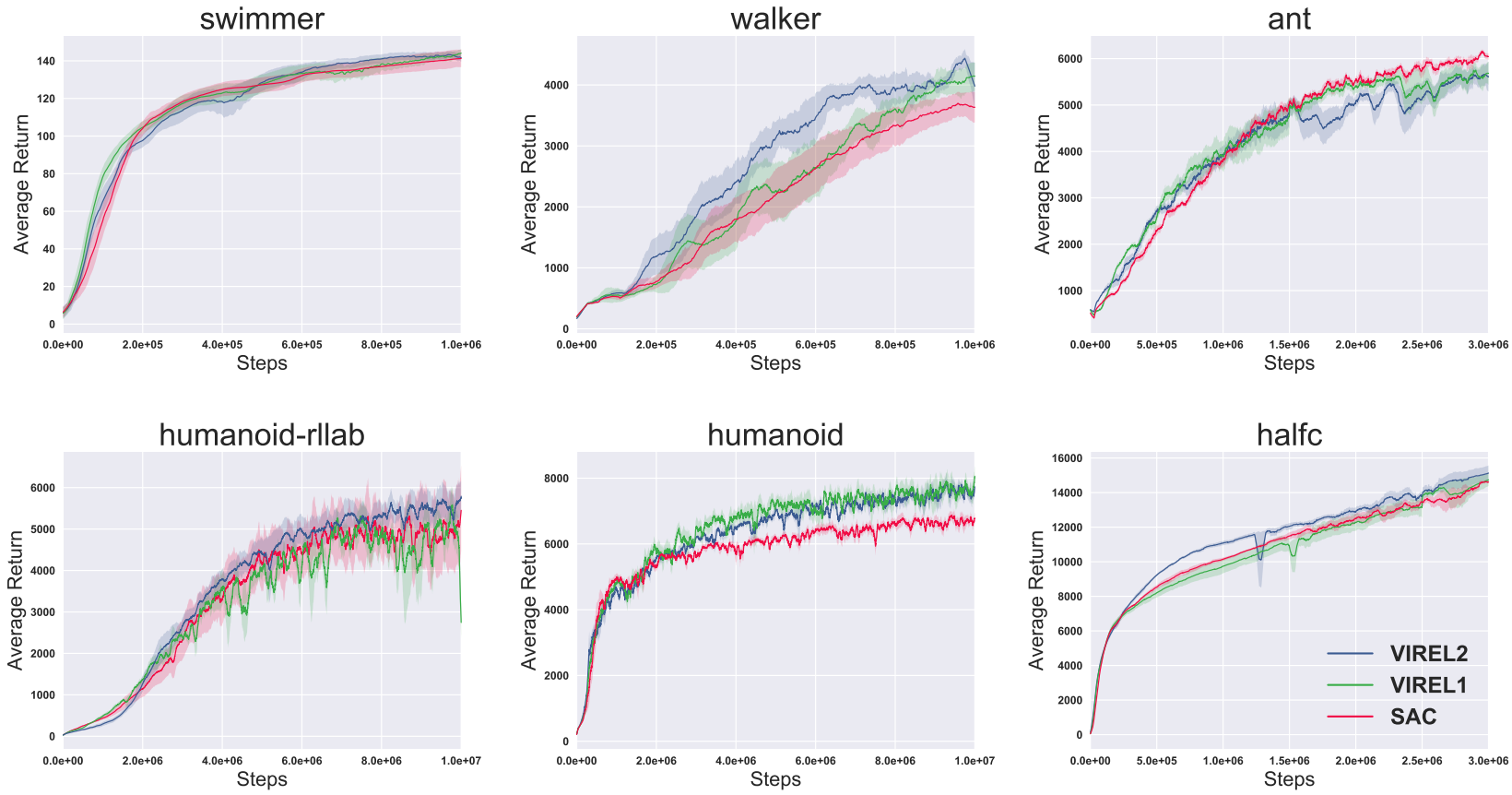


Figure 2: Training curves gym-Mujoco-v1.