Policy Gradient Methods

Milan Straka

April 02, 2025

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research,
Development and Education

Policy Gradient Methods

Policy Gradient Methods

Instead of predicting expected returns, we could train the method to directly predict the policy

$$\pi(a|s;\boldsymbol{\theta}).$$

Obtaining the full distribution over all actions would also allow us to sample the actions according to the distribution π instead of just ε-greedy sampling.

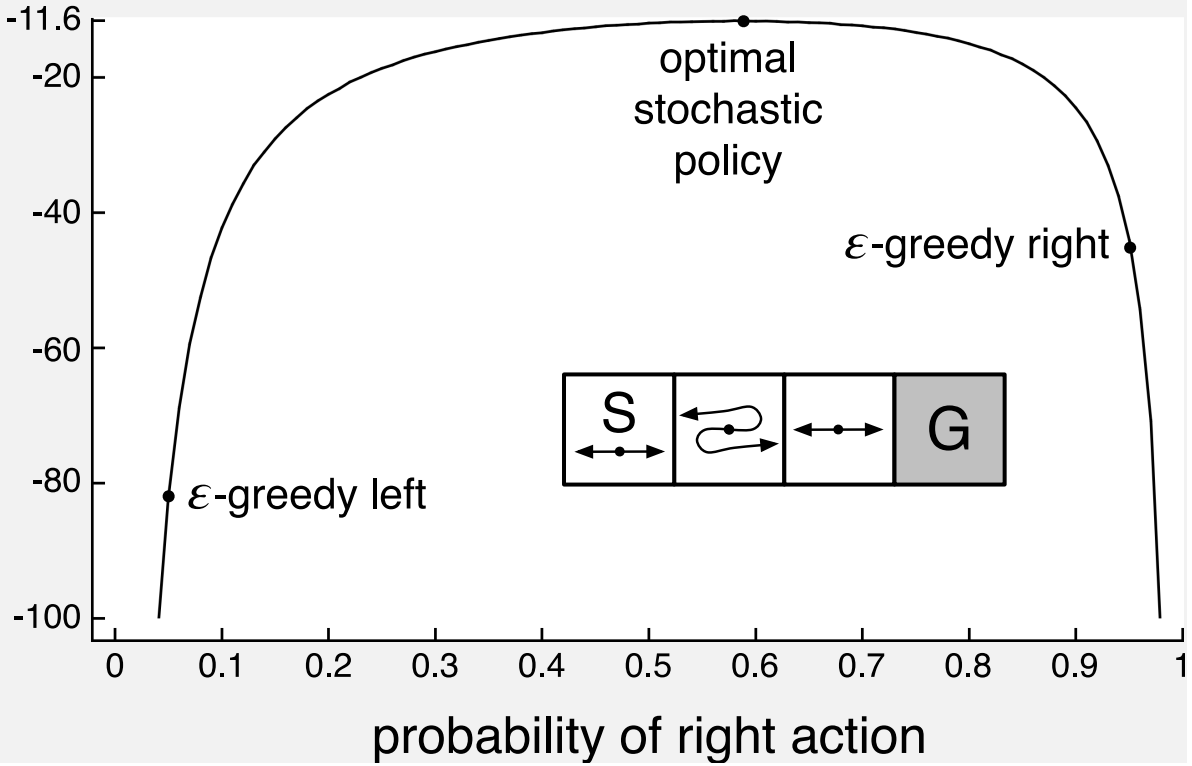
However, to train the network, we maximize the expected return $v_\pi(s)$ and to that account we need to compute its *gradient* $\nabla_{\boldsymbol{\theta}} v_\pi(s)$.

Policy Gradient Methods

In addition to discarding ε-greedy action selection, policy gradient methods allow producing policies which are by nature stochastic, as in card games with imperfect information, while the action-value methods have no natural way of finding stochastic policies (distributional RL might be of some use though).

In the example, the reward is -1 per step, and we assume the three states appear identical under the function approximation.

$$J(\boldsymbol{\theta}) = v_{\pi_{\boldsymbol{\theta}}}(\mathrm{S})$$



Example 13.1 of "Reinforcement Learning: An Introduction, Second Edition".

Policy Gradient Theorem

Let $\pi(a|s;\boldsymbol{\theta})$ be a parametrized policy. We denote the initial state distribution as $h(s)$ and the on-policy distribution under π as $\mu(s)$. Let also $J(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathbb{E}_{s\sim h} v_\pi(s)$.

Then

$$\nabla_{\boldsymbol{\theta}} v_\pi(s) \propto \sum_{s'\in\mathcal{S}} P(s\to\ldots\to s'|\pi) \sum_{a\in\mathcal{A}} q_\pi(s',a)\nabla_{\boldsymbol{\theta}}\pi(a|s';\boldsymbol{\theta})$$

and

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \propto \sum_{s\in\mathcal{S}} \mu(s) \sum_{a\in\mathcal{A}} q_\pi(s,a)\nabla_{\boldsymbol{\theta}}\pi(a|s;\boldsymbol{\theta}),$$

where $P(s\to\ldots\to s'|\pi)$ is the probability of getting to state s' when starting from state s, after any number of 0, 1, ... steps. The γ parameter should be treated as a form of termination, i.e., $P(s\to\ldots\to s'|\pi) \propto \sum_{k=0}^{\infty} \gamma^k P(s\to s' \text{ in } k \text{ steps }|\pi)$.

Proof of Policy Gradient Theorem

$$\begin{aligned}
\nabla v_\pi(s) &= \nabla\Big[\sum_a \pi(a|s;\boldsymbol{\theta}) q_\pi(s,a)\Big] \\
&= \sum_a \Big[q_\pi(s,a)\nabla\pi(a|s;\boldsymbol{\theta}) + \pi(a|s;\boldsymbol{\theta})\nabla q_\pi(s,a)\Big] \\
&= \sum_a \Big[q_\pi(s,a)\nabla\pi(a|s;\boldsymbol{\theta}) + \pi(a|s;\boldsymbol{\theta})\nabla\big(\sum_{s',r} p(s',r|s,a)(r+\gamma v_\pi(s'))\big)\Big] \\
&= \sum_a \Big[q_\pi(s,a)\nabla\pi(a|s;\boldsymbol{\theta}) + \gamma\pi(a|s;\boldsymbol{\theta})\big(\sum_{s'} p(s'|s,a)\nabla v_\pi(s')\big)\Big]
\end{aligned}$$

We now expand $v_\pi(s')$.

$$\begin{aligned}
&= \sum_a \Big[q_\pi(s,a)\nabla\pi(a|s;\boldsymbol{\theta}) + \gamma\pi(a|s;\boldsymbol{\theta})\Big(\sum_{s'} p(s'|s,a)\Big(\\
&\quad \sum_{a'}\Big[q_\pi(s',a')\nabla\pi(a'|s';\boldsymbol{\theta}) + \gamma\pi(a'|s';\boldsymbol{\theta})\big(\sum_{s''} p(s''|s',a')\nabla v_\pi(s'')\big)\Big]\Big)\Big)\Big]
\end{aligned}$$

Continuing to expand all $v_\pi(s'')$, *we obtain the following:*

$$\nabla v_\pi(s) = \sum_{s'\in\mathcal{S}}\sum_{k=0}^{\infty}\gamma^k P(s\to s' \text{ in } k \text{ steps}\,|\pi)\sum_{a\in\mathcal{A}} q_\pi(s',a)\nabla_{\boldsymbol{\theta}}\pi(a|s';\boldsymbol{\theta}).$$

Proof of Policy Gradient Theorem

To finish the proof of the first part, recall that

$$\sum_{k=0}^{\infty} \gamma^k P(s \to s' \text{ in } k \text{ steps} \,|\pi) \propto P(s \to \ldots \to s'|\pi).$$

For the second part, we know that

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{s\sim h} \nabla_{\boldsymbol{\theta}} v_\pi(s) \propto \mathbb{E}_{s\sim h} \sum_{s'\in\mathcal{S}} P(s \to \ldots \to s'|\pi) \sum_{a\in\mathcal{A}} q_\pi(s', a) \nabla_{\boldsymbol{\theta}} \pi(a|s'; \boldsymbol{\theta}),$$

therefore using the fact that $\mu(s') = \mathbb{E}_{s\sim h} P(s \to \ldots \to s'|\pi)$ we get

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \propto \sum_{s\in\mathcal{S}} \mu(s) \sum_{a\in\mathcal{A}} q_\pi(s, a) \nabla_{\boldsymbol{\theta}} \pi(a|s; \boldsymbol{\theta}).$$

The REINFORCE Algorithm

The REINFORCE Algorithm

The REINFORCE algorithm (Williams, 1992) directly uses the policy gradient theorem, minimizing $-J(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\mathbb{E}_{s\sim h} v_\pi(s)$. The loss gradient is then

$$\nabla_{\boldsymbol{\theta}} - J(\boldsymbol{\theta}) \propto -\sum_{s\in\mathcal{S}} \mu(s) \sum_{a\in\mathcal{A}} q_\pi(s,a) \nabla_{\boldsymbol{\theta}} \pi(a|s;\boldsymbol{\theta}) = -\mathbb{E}_{s\sim\mu} \sum_{a\in\mathcal{A}} q_\pi(s,a) \nabla_{\boldsymbol{\theta}} \pi(a|s;\boldsymbol{\theta}).$$

However, the sum over all actions is problematic. Instead, we rewrite it to an expectation which we can estimate by sampling:

$$\nabla_{\boldsymbol{\theta}} - J(\boldsymbol{\theta}) \propto \mathbb{E}_{s\sim\mu} \mathbb{E}_{a\sim\pi} q_\pi(s,a) \nabla_{\boldsymbol{\theta}} - \ln \pi(a|s;\boldsymbol{\theta}),$$

where we used the fact that

$$\nabla_{\boldsymbol{\theta}} \ln \pi(a|s;\boldsymbol{\theta}) = \frac{1}{\pi(a|s;\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} \pi(a|s;\boldsymbol{\theta}).$$

The REINFORCE Algorithm

REINFORCE therefore minimizes the loss $-J(\boldsymbol{\theta})$ with gradient

$$\mathbb{E}_{s\sim\mu}\mathbb{E}_{a\sim\pi} q_\pi(s,a)\nabla_{\boldsymbol{\theta}} - \ln \pi(a|s;\boldsymbol{\theta}),$$

where we estimate the $q_\pi(s,a)$ by a single sample.

Note that the loss is just a weighted variant of negative log-likelihood (NLL), where the sampled actions play a role of gold labels and are weighted according to their return.

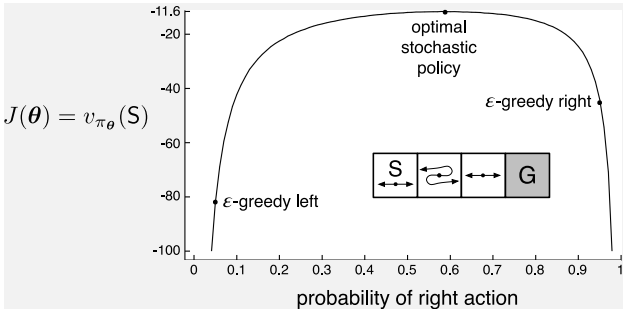
REINFORCE: Monte-Carlo Policy-Gradient Control (episodic) for π_*

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Algorithm parameter: step size $\alpha > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ (e.g., to $\mathbf{0}$)

Loop forever (for each episode):
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot,\boldsymbol{\theta})$
  Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
    $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ $\qquad (G_t)$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \quad \alpha\, G \nabla \ln \pi(A_t|S_t,\boldsymbol{\theta})$

Modified from Algorithm 13.3 of "Reinforcement Learning: An Introduction, Second Edition" by removing γ^t from the update of θ.

The REINFORCE Algorithm Example Performance



Example 13.1 of "Reinforcement Learning: An Introduction, Second Edition".

G_0
Total reward
on episode
averaged over 100 runs

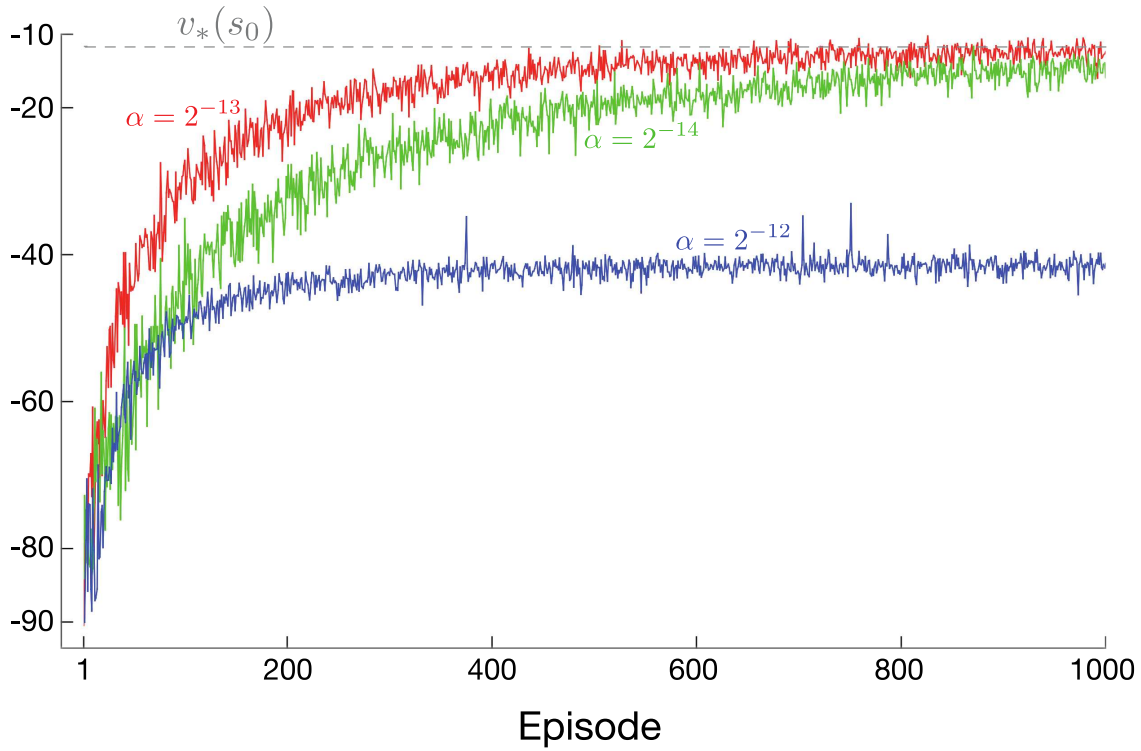


Figure 13.1 of "Reinforcement Learning: An Introduction, Second Edition".

On-policy Distribution in REINFORCE

In the proof, we assumed γ is used as a form of termination in the definition of the on-policy distribution.

However, even when discounting is used during training (to guarantee convergence even for very long episodes), evaluation is often performed without discounting.

Consequently, the distribution μ used in the REINFORCE algorithm is almost always the unterminated (undiscounted) on-policy distribution (I am not aware of any implementation or paper that would use the discounted one), so that we learn even in states that are far from the beginning of an episode.

Note that this is actually true even for DQN and its variants. Therefore, the discounting parameter γ is used mostly as a variance-reduction technique.

REINFORCE with Baseline

REINFORCE with Baseline

The returns can be arbitrary: better-than-average and worse-than-average returns cannot be recognized from the absolute value of the return.

Hopefully, we can generalize the policy gradient theorem using a baseline $b(s)$ to

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \propto \sum_{s \in \mathcal{S}} \mu(s) \sum_{a \in \mathcal{A}} \big(q_\pi(s,a) - \boldsymbol{b(s)}\big) \nabla_{\boldsymbol{\theta}} \pi(a|s;\boldsymbol{\theta}).$$

The baseline $b(s)$ can be a function or even a random variable, as long as it does not depend on a, because

$$\sum_a b(s) \nabla_{\boldsymbol{\theta}} \pi(a|s;\boldsymbol{\theta}) = b(s) \sum_a \nabla_{\boldsymbol{\theta}} \pi(a|s;\boldsymbol{\theta}) = b(s) \nabla_{\boldsymbol{\theta}} \sum_a \pi(a|s;\boldsymbol{\theta}) = b(s) \nabla_{\boldsymbol{\theta}} 1 = 0.$$

REINFORCE with Baseline

A good choice for $b(s)$ is $v_\pi(s)$, which can be shown to minimize the variance of the gradient estimate (in the limit $\gamma \to 1$; see L. Weaver and N. Tao, The Optimal Reward Baseline for Gradient-Based Reinforcement Learning, https://arxiv.org/abs/1301.2315, for the proof). Such baseline reminds centering of returns, given that

$$v_\pi(s) = \mathbb{E}_{a\sim\pi} q_\pi(s,a).$$

Then, better-than-average returns are positive and worse-than-average returns are negative. The resulting $q_\pi(s,a) - v_\pi(s)$ function is also called the **advantage** function

$$a_\pi(s,a) \stackrel{\text{def}}{=} q_\pi(s,a) - v_\pi(s).$$

Of course, the $v_\pi(s)$ baseline can be only approximated. If neural networks are used to estimate $\pi(a|s;\boldsymbol{\theta})$, then some part of the network is usually shared between the policy and value function estimation, which is trained using mean square error of the predicted and observed return.

REINFORCE with Baseline

In REINFORCE with baseline, we train:

1. the *policy network* using the REINFORCE algorithm, and
2. the *value network* by minimizing the mean squared value error $\overline{VE}$.

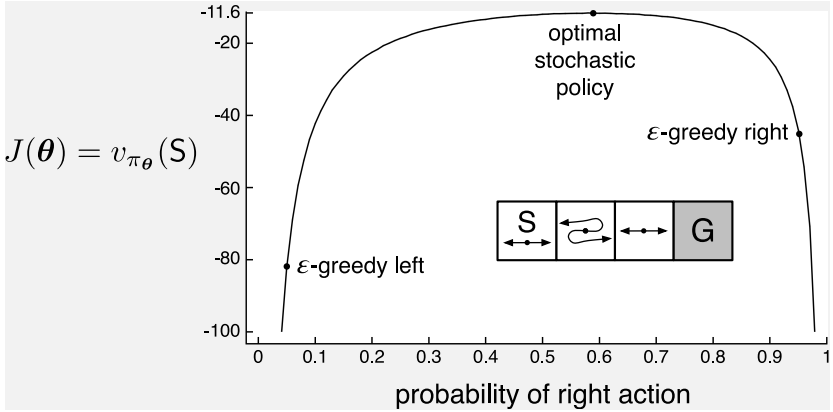
REINFORCE with Baseline (episodic), for estimating $\pi_{\boldsymbol{\theta}} \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Algorithm parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to $\mathbf{0}$)

Loop forever (for each episode):
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot,\boldsymbol{\theta})$
  Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
    $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ $\qquad (G_t)$
    $\delta \leftarrow G - \hat{v}(S_t,\mathbf{w})$
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S_t,\mathbf{w})$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \quad \alpha^{\boldsymbol{\theta}} \delta \nabla \ln \pi(A_t|S_t,\boldsymbol{\theta})$

Modified from Algorithm 13.4 of "Reinforcement Learning: An Introduction, Second Edition" by removing γ^t from the update of θ.

REINFORCE with Baseline Example Performance



Example 13.1 of "Reinforcement Learning: An Introduction, Second Edition".

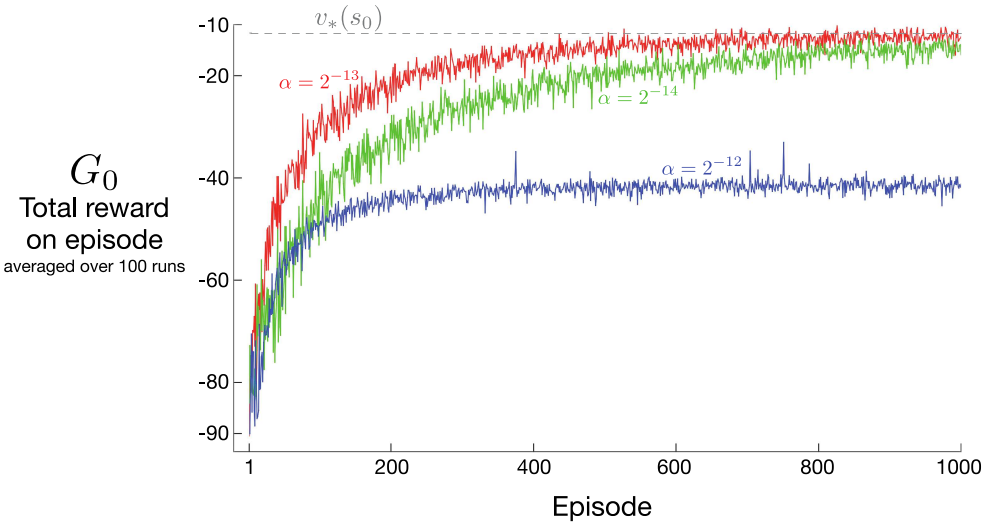


Figure 13.1 of "Reinforcement Learning: An Introduction, Second Edition".

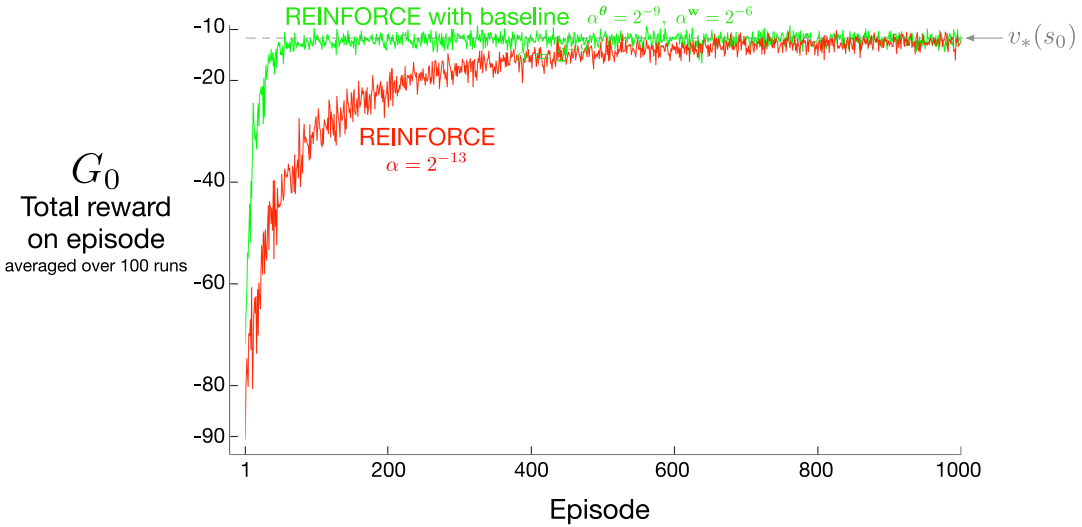


Figure 13.2 of "Reinforcement Learning: An Introduction, Second Edition".

Operator View of REINFORCE

In the middle of 2020, *Dibya Ghosh et al.* introduced the operator view of policy gradient methods in their paper An operator view of policy gradient methods, https://arxiv.org/abs/2006.11266.

Trajectory Formulation

Let $\tau = (S_0, A_0, S_1, A_1, \ldots)$ be a specific trajectory with return $G(\tau) = \sum_{k=0}^{\infty} \gamma^k R_{k+1}(\tau)$.
The probability of τ under a policy π is $\pi(\tau) = h(S_0) \prod_i \pi(A_i|S_i) p(S_{i+1}|S_i, A_i)$.

Our goal is then to find

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\tau \sim \pi_{\boldsymbol{\theta}}} \big[G(\tau)\big] = \arg\max_{\boldsymbol{\theta}} \int_{\tau} \pi(\tau) G(\tau) \,\mathrm{d}\tau,$$

and the REINFORCE algorithm at each step sets the weights $\boldsymbol{\theta}_{t+1}$ to

$$\boldsymbol{\theta}_t + \alpha \mathbb{E}_{\tau \sim \pi_{\boldsymbol{\theta}_t}} \left[G(\tau) \frac{\partial \log \pi_{\boldsymbol{\theta}}(\tau)}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t} \right] = \boldsymbol{\theta}_t + \alpha \int_{\tau} \pi_{\boldsymbol{\theta}_t}(\tau) G(\tau) \frac{\partial \log \pi_{\boldsymbol{\theta}}(\tau)}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t} \mathrm{d}\tau.$$

Trajector Formulation of OP-REINFORCE

In the operator view, policy improvement is achieved by a successive application of a **policy improvement operator** $\mathcal{I}$ and a **projection operator** $\mathcal{P}$. For tabular methods, the projection operator is identity, but it is needed for functional approximation methods.

The operator version of REINFORCE is then the iterative application of $\mathcal{P} \circ \mathcal{I}$ with

$$(\mathcal{I}\pi)(\tau) \overset{\text{def}}{\propto} G(\tau)\pi(\tau),$$

$$\mathcal{P}\nu \overset{\text{def}}{=} \arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}\left(\nu \| \pi_{\boldsymbol{\theta}}\right).$$

As formulated, the operator version of REINFORCE computes the projection perfectly in each step, while the REINFORCE performs just one step of gradient descent in the direction of $\mathcal{P}$. However, it is easy to show that the fixed points of both algorithms are the same.

Trajector Formulation of OP-REINFORCE

The proposition is actually not difficult to prove, we just need to expand the definitions.

Denoting ν the distribution over trajectories such that $\nu(\tau) \propto G(\tau)\pi(\tau)$, we get

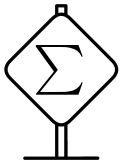
$$D_{\mathrm{KL}}\left(\nu \| \pi_{\boldsymbol{\theta}}\right) = \int_{\tau} \nu(\tau) \log \frac{\nu(\tau)}{\pi_{\boldsymbol{\theta}}(\tau)} \,\mathrm{d}\tau.$$

Therefore, the gradient is

$$\frac{\partial D_{\mathrm{KL}}\left(\nu \| \pi_{\boldsymbol{\theta}}\right)}{\partial \boldsymbol{\theta}} = -\int_{\tau} \nu(\tau) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\tau) \,\mathrm{d}\tau \propto -\int_{\tau} \pi_{\boldsymbol{\theta}}(\tau) G(\tau) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\tau) \,\mathrm{d}\tau.$$

For optimal policy $\pi_{\boldsymbol{\theta}^*}$, we therefore get $\frac{\partial D_{\mathrm{KL}}(\nu \| \pi_{\boldsymbol{\theta}^*})}{\partial \boldsymbol{\theta}^*} \propto -\int_{\tau} \pi_{\boldsymbol{\theta}^*}(\tau) G(\tau) \nabla_{\boldsymbol{\theta}^*} \log \pi_{\boldsymbol{\theta}^*}(\tau) \,\mathrm{d}\tau$, but the latter is zero because of the optimality of $\pi_{\boldsymbol{\theta}^*}$ according to the policy gradient theorem; therefore, $\pi_{\boldsymbol{\theta}^*}$ is also the fixed point of $\mathcal{P} \circ \mathcal{I}$.

State-Action Formulation of OP-REINFORCE



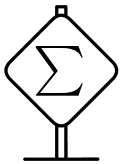
We can formulate the operator view also employing the action-value function q and the on-policy distribution μ_π; however, the policy improvement operator needs to return not just a policy, but a joint distribution over the states and actions.

The REINFORCE algorithm can be seen as performing one gradient step to minimize the composition $\mathcal{P} \circ \mathcal{I}$, where

$$(\mathcal{I}\pi)(s,a) \overset{\text{def}}{\propto} \mu_\pi(s) q_\pi(s,a) \pi(a|s),$$

$$\mathcal{P}\nu \overset{\text{def}}{=} \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{s\sim\nu(s)} \Big[D_{\mathrm{KL}}\big(\nu(\cdot|s) \| \pi_{\boldsymbol{\theta}}(\cdot|s)\big) \Big].$$

State-Action Formulation of OP-REINFORCE



For completeness, we can explicitly express the joint distribution $(\mathcal{I}\pi)(s,a)$ as a product of $(\mathcal{I}\pi)(s) \cdot (\mathcal{I}\pi)(a|s)$, where

- the distribution over the states is

$$(\mathcal{I}\pi)(s) \stackrel{\text{def}}{=} \frac{\mu_\pi(s) v_\pi(s)}{\sum_{s'} \mu_\pi(s') v_\pi(s')} = \frac{\mu_\pi(s) v_\pi(s)}{\mathbb{E}_{s' \sim \mu_\pi}[v_\pi(s')]},$$

- the conditional distribution over the actions is

$$(\mathcal{I}\pi)(a|s) \stackrel{\text{def}}{=} \frac{q_\pi(s,a)\pi(a|s)}{\sum_{a'} q_\pi(s,a')\pi(a'|s)} = \frac{q_\pi(s,a)\pi(a|s)}{\mathbb{E}_{a' \sim \pi(s)}[q_\pi(s,a')]} = \frac{q_\pi(s,a)\pi(a|s)}{v_\pi(s)}.$$

Higher Powers of the Returns

Instead of $\nu(\tau) \propto G(\tau)\pi(\tau)$, we now for $k \geq 1$ consider

$$(\mathcal{I}^k \pi)(\tau) \overset{\text{def}}{\propto} G(\tau)^k \pi(\tau).$$

However, it is not obvious if $\pi_{\boldsymbol{\theta}^*}$ is still a fixed point of $\mathcal{P} \circ \mathcal{I}^k$. In fact, it is not:

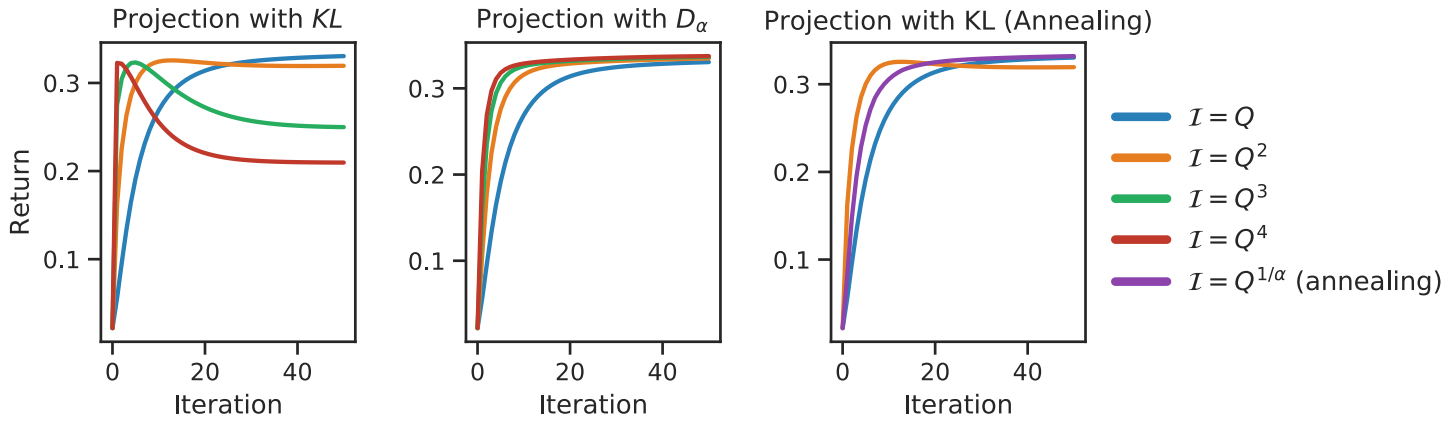
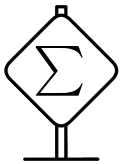


Figure 1: Evaluation of polynomial reward improvement operators $\mathcal{I}_V^{1/\alpha}$ paired with different projection steps in the four-room domain. The operator $\mathcal{I}_V^{1/\alpha}$ generally speeds up learning, but if paired with the KL projection (left), it can converge to a sub-optimal policy. If the improvement operator is paired with an α-divergence (middle) or the value of α is annealed to 1 (right), learning is fast and converges to the optimal policy. Figure best seen in color.

Figure 1 of "An operator view of policy gradient methods", https://arxiv.org/abs/2006.11266

Higher Powers of the Returns



Let reformulate the $\mathcal{I}^k$ to $\mathcal{I}^{\frac{1}{\alpha}}$ for $\alpha \leq 1$:

$$(\mathcal{I}^{\frac{1}{\alpha}}\pi)(\tau) \overset{\text{def}}{\propto} G(\tau)^{\frac{1}{\alpha}}\pi(\tau).$$

We then define a projection operator $\mathcal{P}^\alpha$ using α-divergence (also known as Rényi divergence of order α) instead of the KL divergence:

$$\mathcal{P}^\alpha \nu \overset{\text{def}}{=} \arg\min_{\boldsymbol{\theta}} D^\alpha\left(\nu \| \pi_{\boldsymbol{\theta}}\right),$$

$$D^\alpha(p\|q) \overset{\text{def}}{=} \frac{1}{1-\alpha}\log \mathbb{E}_{x\sim p}\left[(p(x)/q(x))^{\alpha-1}\right].$$

For $\alpha = 1$, D^α is not defined, but its limit in $\alpha \to 1$ is D_{KL}.

Proposition 8 of the OP-REINFORCE paper proves that for $\alpha \in (0,1)$, $\pi_{\boldsymbol{\theta}^*}$ is the fixed point of $\mathcal{P}^\alpha \circ \mathcal{I}^{\frac{1}{\alpha}}$.

Higher Powers of the Returns

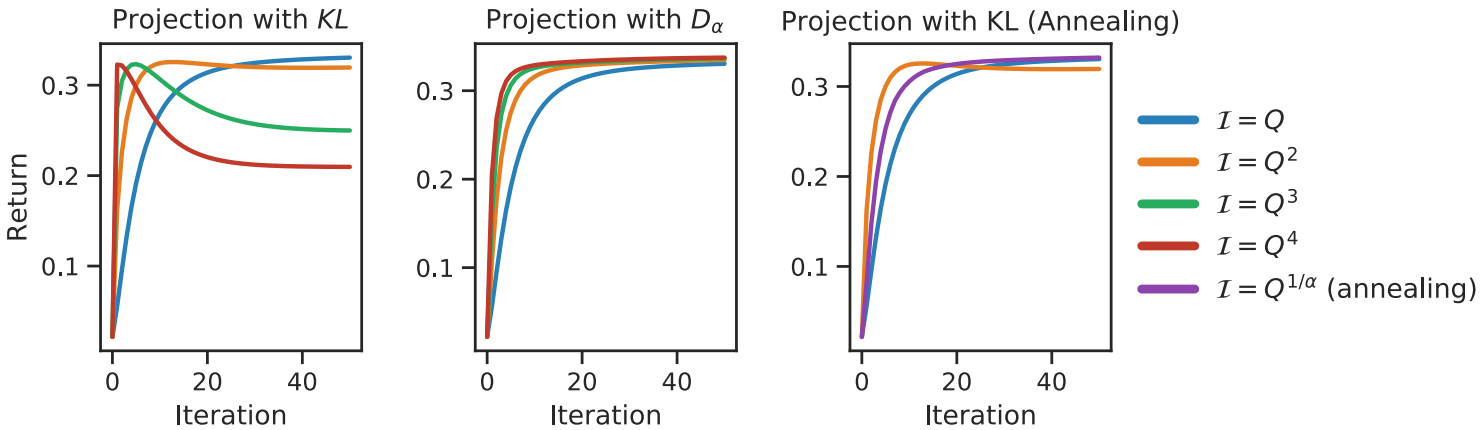


Figure 1: Evaluation of polynomial reward improvement operators $\mathcal{I}_V^{1/\alpha}$ paired with different projection steps in the four-room domain. The operator $\mathcal{I}_V^{1/\alpha}$ generally speeds up learning, but if paired with the KL projection (left), it can converge to a sub-optimal policy. If the improvement operator is paired with an α-divergence (middle) or the value of α is annealed to 1 (right), learning is fast and converges to the optimal policy. Figure best seen in color.

Figure 1 of "An operator view of policy gradient methods", https://arxiv.org/abs/2006.11266

Note that $\mathcal{I}^\alpha$ in the limit $\alpha \to 0$ assigns probability of 1 to the greedy action, so it becomes the *greedy policy improvement operator*.

Actor-Critic Methods

Actor-Critic Methods

It is possible to combine the policy gradient methods and temporal difference methods, creating a family of algorithms usually called the **actor-critic** methods.

The idea is straightforward – similarly to the REINFORCE with baseline, we train the policy network together with the value network. However, instead of estimating the episode return using the whole episode rewards, we use n-step return TD estimate in both the policy gradient and the mean squared value error $\overline{VE}$.

One-step Actor–Critic (episodic), for estimating $\pi_{\boldsymbol{\theta}} \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to $\mathbf{0}$)
Loop forever (for each episode):
  Initialize S (first state of episode)

  Loop while S is not terminal (for each time step):
    $A \sim \pi(\cdot|S,\boldsymbol{\theta})$
    Take action A, observe S', R
    $\delta \leftarrow R + \gamma \hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})$ (if S' is terminal, then $\hat{v}(S',\mathbf{w}) \doteq 0$)
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S,\mathbf{w})$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} \delta \nabla \ln \pi(A|S,\boldsymbol{\theta})$
    $S \leftarrow S'$

Modified from Algorithm 13.5 of "Reinforcement Learning: An Introduction, Second Edition" by removing I.

Asynchronous Advantage Actor-Critic (A3C)

Asynchronous Methods for Deep RL

The A3C was introduced in a 2016 paper from Volodymyr Mnih et al. (the same group as DQN) Asynchronous Methods for Deep Reinforcement Learning, https://arxiv.org/abs/1602.01783.

The authors propose an asynchronous framework, where multiple workers share one neural network, each training using either an off-line or on-line RL algorithm.

They compare 1-step Q-learning, 1-step Sarsa, n-step Q-learning and A3C (an *asynchronous advantage actor-critic* method). For A3C, they compare a version with and without LSTM.

The authors also introduce *entropy regularization term* $-\beta H(\pi(s;\boldsymbol{\theta}))$ to the loss to support exploration and discourage premature convergence (they use $\beta = 0.01$).

- The entropy regularization has since become the standard way of encouraging exploration with a policy network.

  The entropy regularization keeps a controllable level of surprise (i.e., exploration) in the distribution. Compared to ε-greedy approach, the exploration actions are sampled proportionally to their expected utility, not randomly.

Algorithm 1 Asynchronous one-step Q-learning - pseudocode for each actor-learner thread.

// Assume global shared θ, θ^-, and counter $T = 0$.
Initialize thread step counter $t \leftarrow 0$
Initialize target network weights $\theta^- \leftarrow \theta$
Initialize network gradients $d\theta \leftarrow 0$
Get initial state s
repeat
  Take action a with ϵ-greedy policy based on $Q(s, a; \theta)$
  Receive new state s' and reward r
  $y = \begin{cases} r & \text{for terminal } s' \\ r + \gamma \max_{a'} Q(s', a'; \theta^-) & \text{for non-terminal } s' \end{cases}$
  Accumulate gradients wrt θ: $d\theta \leftarrow d\theta + \frac{\partial (y - Q(s,a;\theta))^2}{\partial \theta}$
  $s = s'$
  $T \leftarrow T + 1$ and $t \leftarrow t + 1$
  if $T \mod I_{target} == 0$ **then**
    Update the target network $\theta^- \leftarrow \theta$
  end if
  if $t \mod I_{AsyncUpdate} == 0$ or s is terminal **then**
    Perform asynchronous update of θ using $d\theta$.
    Clear gradients $d\theta \leftarrow 0$.
  end if
until $T > T_{max}$

Algorithm 1 of "Asynchronous Methods for Deep Reinforcement Learning" by Volodymyr Mnih et al.

Algorithm S2 Asynchronous n-step Q-learning - pseudocode for each actor-learner thread.

// Assume global shared parameter vector θ.
// Assume global shared target parameter vector θ^-.
// Assume global shared counter $T = 0$.
Initialize thread step counter $t \leftarrow 1$
Initialize target network parameters $\theta^- \leftarrow \theta$
Initialize thread-specific parameters $\theta' = \theta$
Initialize network gradients $d\theta \leftarrow 0$
repeat
  Clear gradients $d\theta \leftarrow 0$
  Synchronize thread-specific parameters $\theta' = \theta$
  $t_{start} = t$
  Get state s_t
  repeat
    Take action a_t according to the ϵ-greedy policy based on $Q(s_t, a; \theta')$
    Receive reward r_t and new state s_{t+1}
    $t \leftarrow t + 1$
    $T \leftarrow T + 1$
  until terminal s_t **or** $t - t_{start} == t_{max}$
  $R = \begin{cases} 0 & \text{for terminal } s_t \\ \max_a Q(s_t, a; \theta^-) & \text{for non-terminal } s_t \end{cases}$
  for $i \in \{t-1, \ldots, t_{start}\}$ **do**
    $R \leftarrow r_i + \gamma R$
    Accumulate gradients wrt θ': $d\theta \leftarrow d\theta + \frac{\partial \left(R - Q(s_i, a_i; \theta')\right)^2}{\partial \theta'}$
  end for
  Perform asynchronous update of θ using $d\theta$.
  if $T \mod I_{target} == 0$ **then**
    $\theta^- \leftarrow \theta$
  end if
until $T > T_{max}$

Algorithm S2 of "Asynchronous Methods for Deep Reinforcement Learning" by Volodymyr Mnih et al.

Algorithm S3 Asynchronous advantage actor-critic - pseudocode for each actor-learner thread.

// Assume global shared parameter vectors θ and θ_v and global shared counter $T = 0$
// Assume thread-specific parameter vectors θ' and θ'_v
Initialize thread step counter $t \leftarrow 1$
repeat
  Reset gradients: $d\theta \leftarrow 0$ and $d\theta_v \leftarrow 0$.
  Synchronize thread-specific parameters $\theta' = \theta$ and $\theta'_v = \theta_v$
  $t_{start} = t$
  Get state s_t
  repeat
    Perform a_t according to policy $\pi(a_t|s_t;\theta')$
    Receive reward r_t and new state s_{t+1}
    $t \leftarrow t + 1$
    $T \leftarrow T + 1$
  until terminal s_t **or** $t - t_{start} == t_{max}$
  $R = \begin{cases} 0 & \text{for terminal } s_t \\ V(s_t, \theta'_v) & \text{for non-terminal } s_t \text{// Bootstrap from last state} \end{cases}$
  for $i \in \{t-1, \ldots, t_{start}\}$ **do**
    $R \leftarrow r_i + \gamma R$
    Accumulate gradients wrt θ': $d\theta \leftarrow d\theta + \nabla_{\theta'} \log \pi(a_i|s_i;\theta')(R - V(s_i;\theta'_v))$
    Accumulate gradients wrt θ'_v: $d\theta_v \leftarrow d\theta_v + \partial \left(R - V(s_i;\theta'_v)\right)^2 / \partial \theta'_v$
  end for
  Perform asynchronous update of θ using $d\theta$ and of θ_v using $d\theta_v$.
until $T > T_{max}$

Algorithm S3 of "Asynchronous Methods for Deep Reinforcement Learning", https://arxiv.org/abs/1602.01783

Asynchronous Methods for Deep RL

All methods performed updates every 5 actions ($t_{\max} = I_{\text{AsyncUpdate}} = 5$), updating the target network each $40\,000$ frames.

The Atari inputs were processed as in DQN, using also action repeat 4.

The network architecture is: 16 filters 8×8 stride 4, 32 filters 4×4 stride 2, followed by a fully connected layer with 256 units. All hidden layers apply a ReLU nonlinearity. Values and/or action values were then generated from the (same) last hidden layer.

The LSTM methods utilized a 256-unit LSTM cell after the dense hidden layer.

All experiments used a discount factor of $\gamma = 0.99$ and used RMSProp with momentum decay factor of 0.99.

Asynchronous Methods for Deep RL

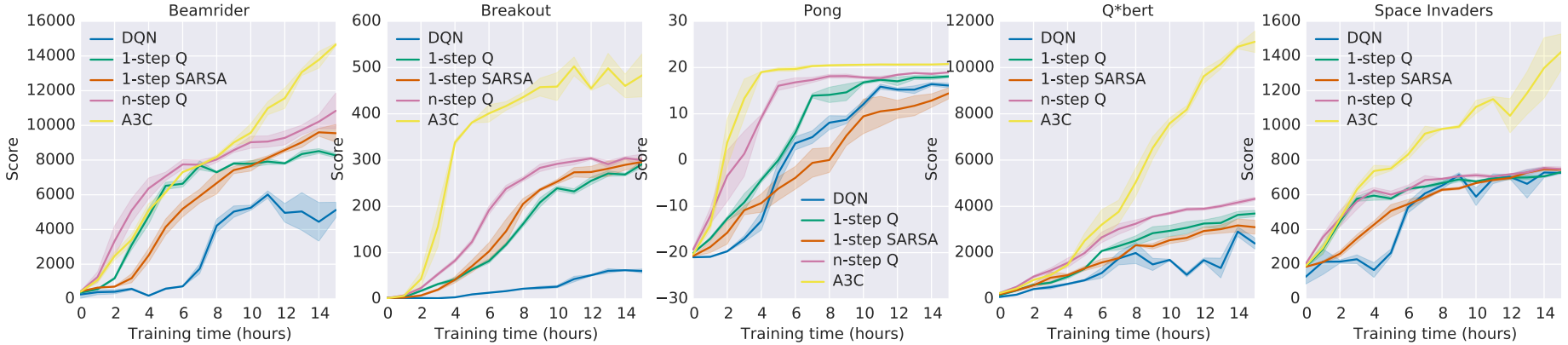


Figure 1. Learning speed comparison for DQN and the new asynchronous algorithms on five Atari 2600 games. DQN was trained on a single Nvidia K40 GPU while the asynchronous methods were trained using 16 CPU cores. The plots are averaged over 5 runs. In the case of DQN the runs were for different seeds with fixed hyperparameters. For asynchronous methods we average over the best 5 models from 50 experiments with learning rates sampled from $LogUniform(10^{-4}, 10^{-2})$ and all other hyperparameters fixed.

Figure 1 of "Asynchronous Methods for Deep Reinforcement Learning", https://arxiv.org/abs/1602.01783

Method	Training Time	Mean	Median
DQN	8 days on GPU	121.9%	47.5%
Gorila	4 days, 100 machines	215.2%	71.3%
D-DQN	8 days on GPU	332.9%	110.9%
Dueling D-DQN	8 days on GPU	343.8%	117.1%
Prioritized DQN	8 days on GPU	463.6%	127.6%
A3C, FF	1 day on CPU	344.1%	68.2%
A3C, FF	4 days on CPU	496.8%	116.6%
A3C, LSTM	4 days on CPU	623.0%	112.6%

Table 1 of "Asynchronous Methods for Deep Reinforcement Learning", https://arxiv.org/abs/1602.01783

	Number of threads				
Method	1	2	4	8	16
1-step Q	1.0	**3.0**	**6.3**	**13.3**	**24.1**
1-step SARSA	1.0	**2.8**	**5.9**	**13.1**	**22.1**
n-step Q	1.0	**2.7**	**5.9**	**10.7**	**17.2**
A3C	1.0	2.1	3.7	6.9	12.5

Table 2 of "Asynchronous Methods for Deep Reinforcement Learning", https://arxiv.org/abs/1602.01783

Asynchronous Methods for Deep RL

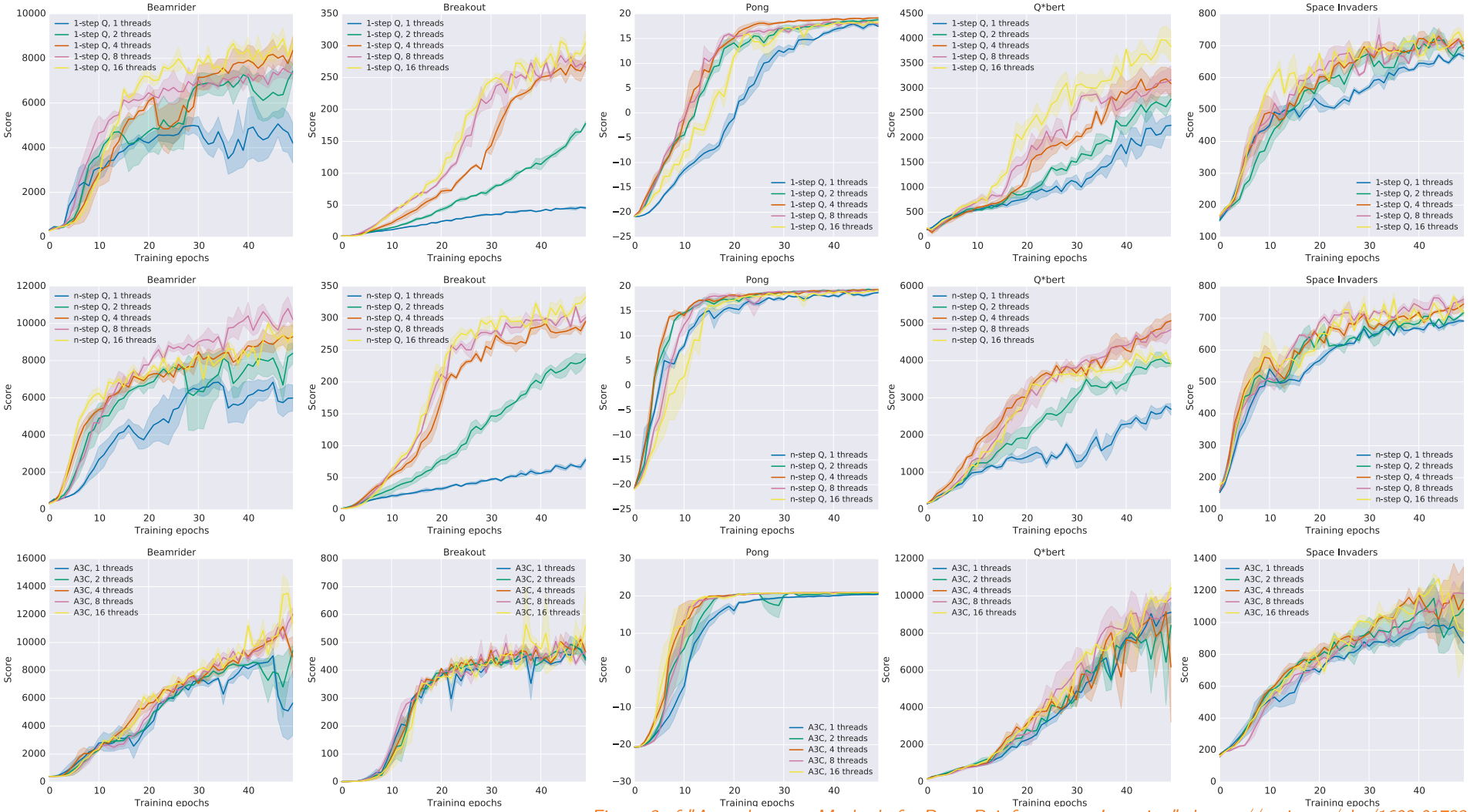


Figure 3 of "Asynchronous Methods for Deep Reinforcement Learning", https://arxiv.org/abs/1602.01783

Asynchronous Methods for Deep RL

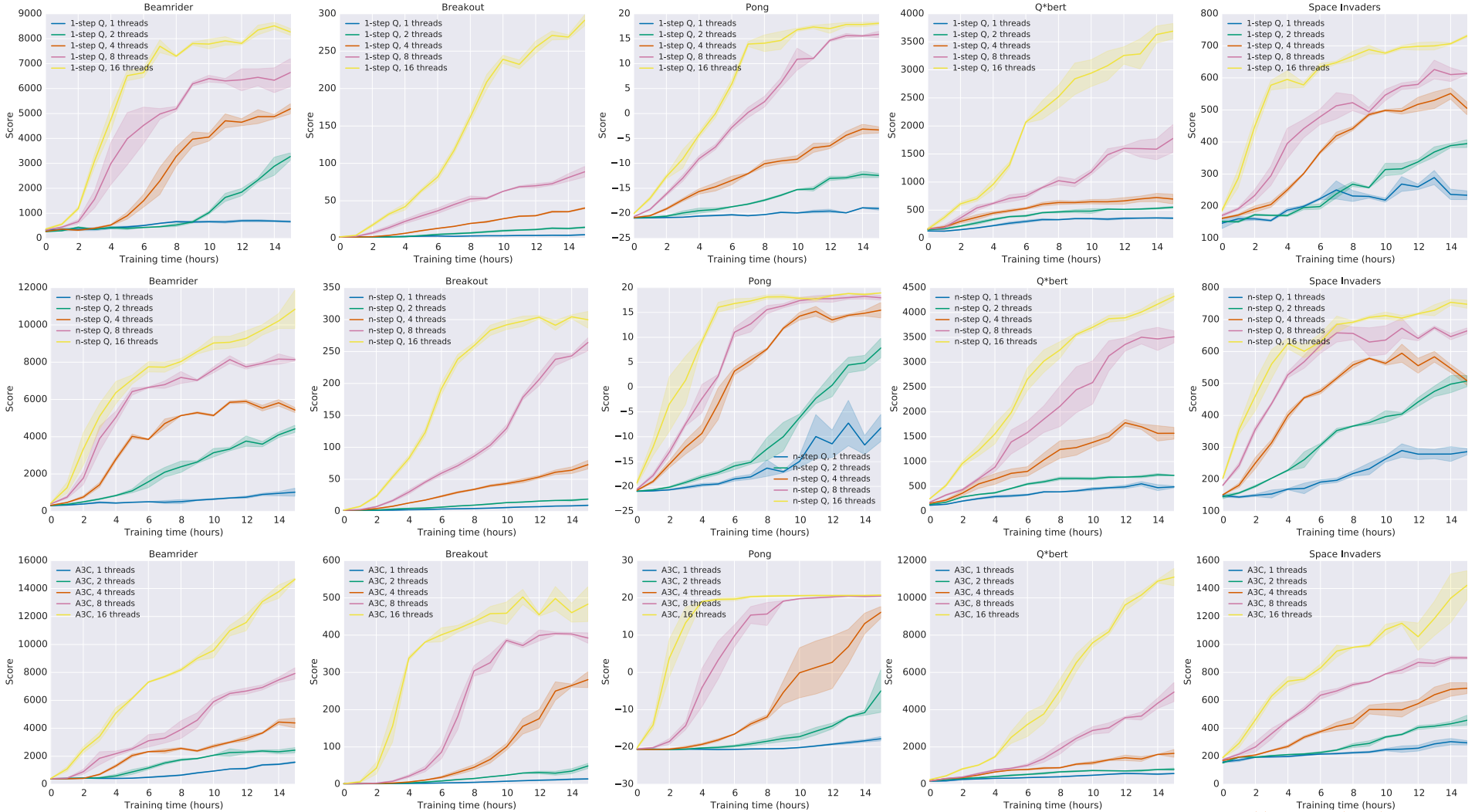


Figure 4 of "Asynchronous Methods for Deep Reinforcement Learning", https://arxiv.org/abs/1602.01783

Asynchronous Methods for Deep RL

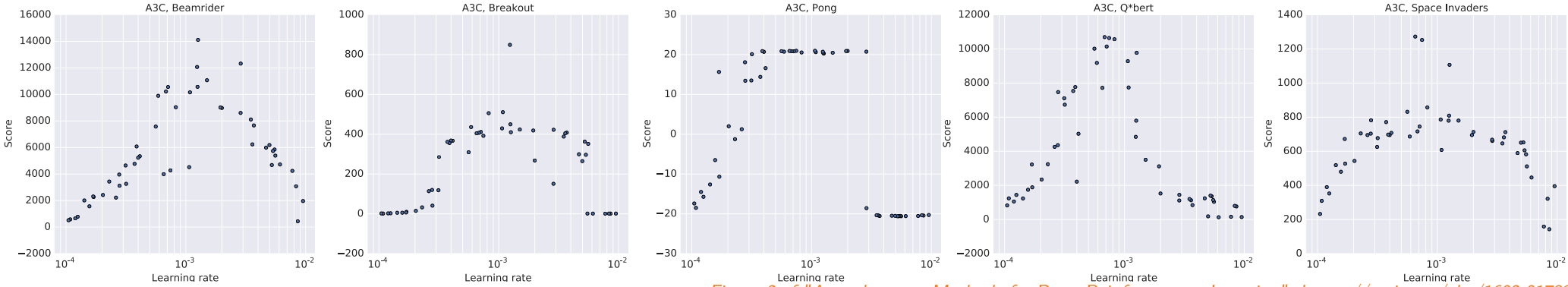


Figure 2 of "Asynchronous Methods for Deep Reinforcement Learning", https://arxiv.org/abs/1602.01783

Parallel Advantage Actor Critic (PAAC)

Parallel Advantage Actor Critic

An alternative to independent workers is to train in a synchronous and centralized way by having the workers to only generate episodes. Such approach was described in 2017 as **parallel advantage actor-critic** (PAAC) by Clemente et al., https://arxiv.org/abs/1705.04862.

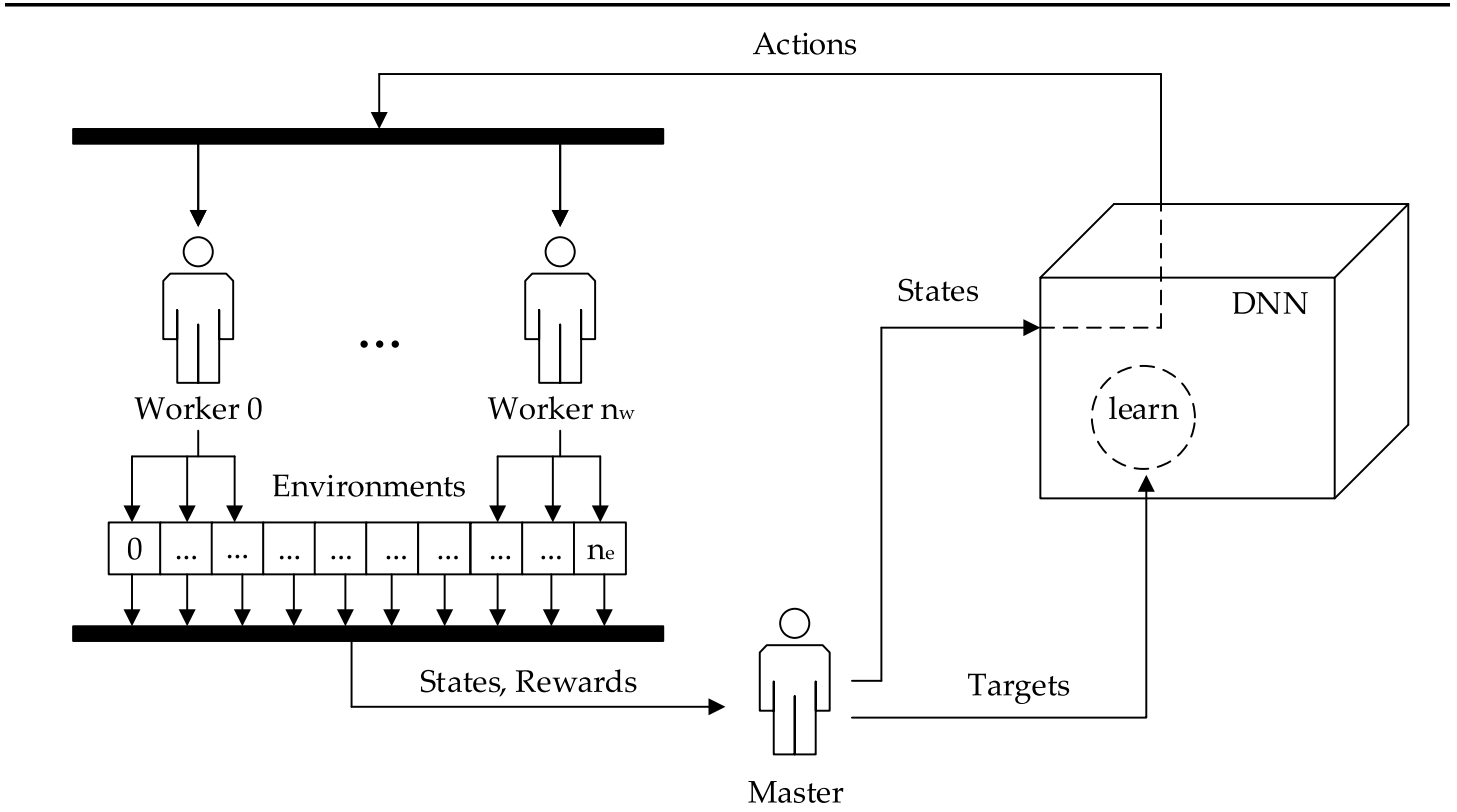


Figure 1 of "Efficient Parallel Methods for Deep Reinforcement Learning" by Alfredo V. Clemente et al.

Parallel Advantage Actor Critic

Algorithm 1 Parallel advantage actor-critic

Initialize timestep counter $N = 0$ and network weights θ, θ_v
Instantiate set $\boldsymbol{e}$ of n_e environments
repeat
  for $t = 1$ to t_{max} **do**
    Sample $\boldsymbol{a}_t$ from $\pi(\boldsymbol{a}_t|\boldsymbol{s}_t;\theta)$
    Calculate $\boldsymbol{v}_t$ from $V(\boldsymbol{s}_t;\theta_v)$
    parallel for $i = 1$ to n_e **do**
      Perform action $a_{t,i}$ in environment e_i
      Observe new state $s_{t+1,i}$ and reward $r_{t+1,i}$
    end parallel for
  end for
  $\boldsymbol{R}_{t_{\max}+1} = \begin{cases} 0 & \text{for terminal } \boldsymbol{s}_t \\ V(s_{t_{\max}+1};\theta) & \text{for non-terminal } \boldsymbol{s}_t \end{cases}$
  for $t = t_{\max}$ down to 1 **do**
    $\boldsymbol{R}_t = \boldsymbol{r}_t + \gamma \boldsymbol{R}_{t+1}$
  end for
  $d\theta = \frac{1}{n_e \cdot t_{max}} \sum_{i=1}^{n_e} \sum_{t=1}^{t_{max}} (R_{t,i} - v_{t,i}) \nabla_\theta - \log \pi(a_{t,i}|s_{t,i};\theta) - \beta \nabla_\theta H(\pi(s_{e,t};\theta))$
  $d\theta_v = \frac{1}{n_e \cdot t_{max}} \sum_{i=1}^{n_e} \sum_{t=1}^{t_{max}} \nabla_{\theta_v} (R_{t,i} - V(s_{t,i};\theta_v))^2$
  Update θ using $d\theta$ and θ_v using $d\theta_v$.
  $N \leftarrow N + n_e \cdot t_{\max}$
until $N \geq N_{max}$

Modification of Algorithm 1 of "Efficient Parallel Methods for Deep Reinforcement Learning" by Alfredo V. Clemente et al.

Parallel Advantage Actor Critic

Game	Gorila	A3C FF	GA3C	PAAC $\text{arch}_{\text{nips}}$	PAAC $\text{arch}_{\text{nature}}$
Amidar	1189.70	263.9	218	701.8	1348.3
Centipede	8432.30	3755.8	7386	5747.32	7368.1
Beam Rider	3302.9	22707.9	N/A	4062.0	6844.0
Boxing	94.9	59.8	92	99.6	99.8
Breakout	402.2	681.9	N/A	470.1	565.3
Ms. Pacman	3233.50	653.7	1978	2194.7	1976.0
Name This Game	6182.16	10476.1	5643	9743.7	14068.0
Pong	18.3	5.6	18	20.6	20.9
Qbert	10815.6	15148.8	14966.0	16561.7	17249.2
Seaquest	13169.06	2355.4	1706	1754.0	1755.3
Space Invaders	1883.4	15730.5	N/A	1077.3	1427.8
Up n Down	12561.58	74705.7	8623	88105.3	100523.3
Training	4d CPU cluster	4d CPU	1d GPU	12h GPU	15h GPU

Table 1 of "Efficient Parallel Methods for Deep Reinforcement Learning" by Alfredo V. Clemente et al.

The authors use 8 workers, $n_e = 32$ parallel environments, 5-step returns, $\gamma = 0.99$, $\varepsilon = 0.1$, $\beta = 0.01$, and a learning rate of $\alpha = 0.0007 \cdot n_e = 0.0224$.

The $\text{arch}_{\text{nips}}$ is from A3C: 16 filters 8×8 stride 4, 32 filters 4×4 stride 2, a dense layer with 256 units. The $\text{arch}_{\text{nature}}$ is from DQN: 32 filters 8×8 stride 4, 64 filters 4×4 stride 2, 64 filters 3×3 stride 1 and 512-unit fully connected layer. All nonlinearities are ReLU.

Parallel Advantage Actor Critic

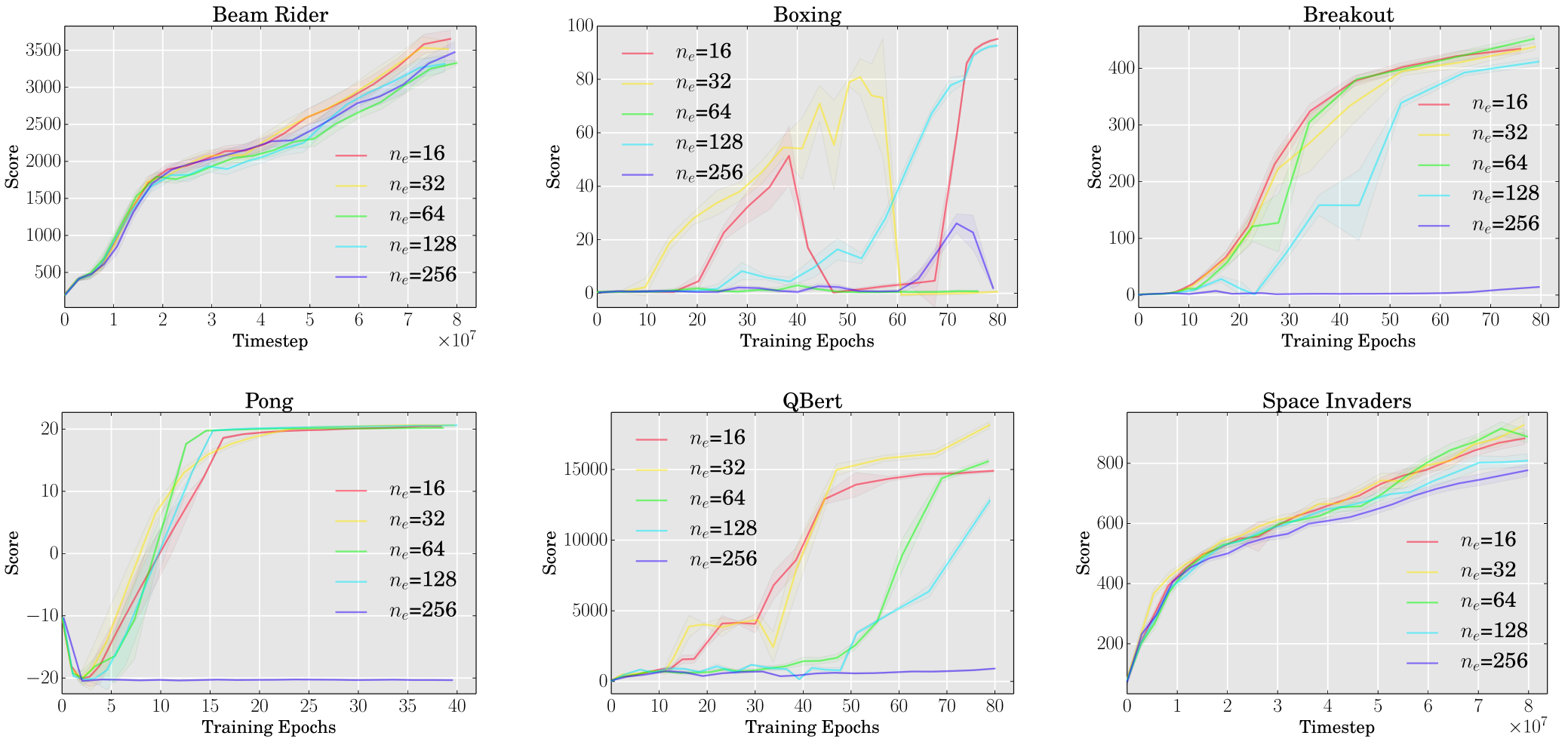


Figure 3 of "Efficient Parallel Methods for Deep Reinforcement Learning" by Alfredo V. Clemente et al.

Parallel Advantage Actor Critic

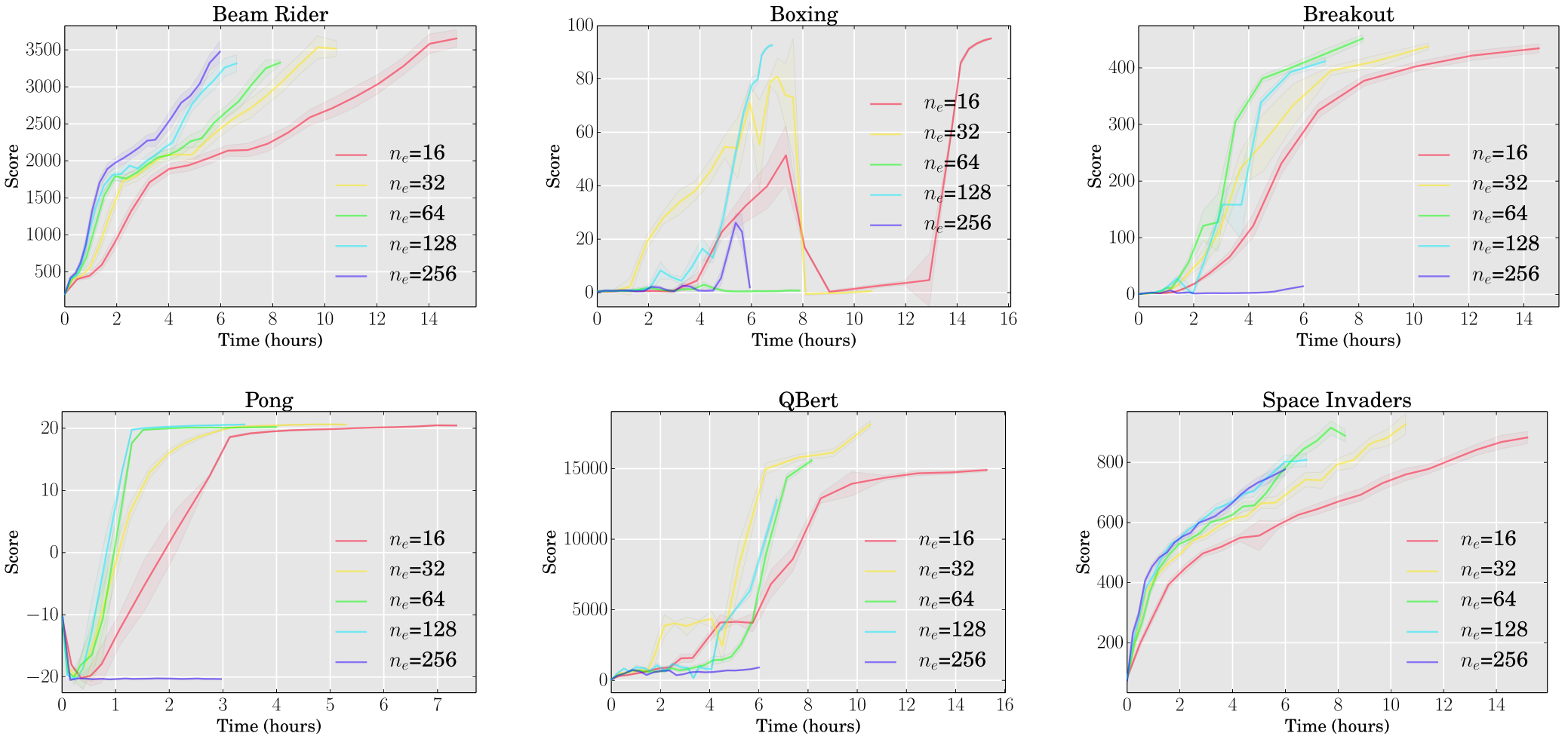
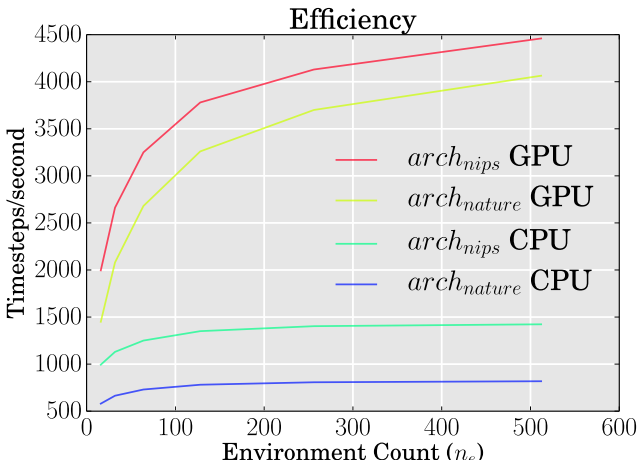
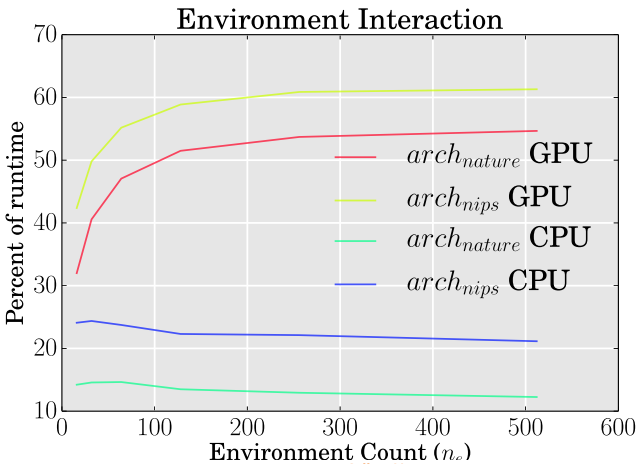


Figure 4 of "Efficient Parallel Methods for Deep Reinforcement Learning" by Alfredo V. Clemente et al.

Parallel Advantage Actor Critic





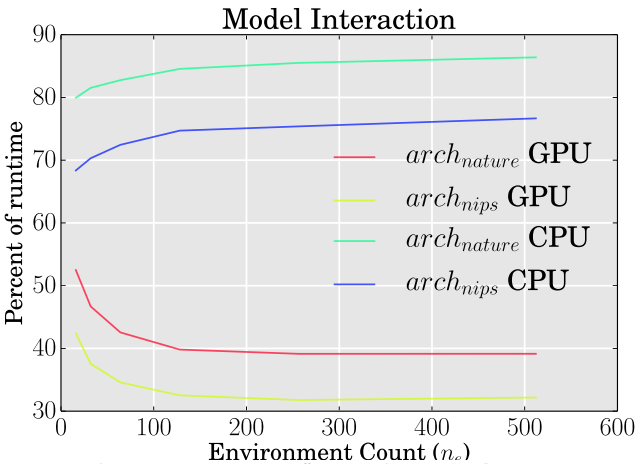


Figure 2 of "Efficient Parallel Methods for Deep Reinforcement Learning" by Alfredo V. Clemente et al.