SAMPLE EFFICIENT ACTOR-CRITIC WITH EXPERIENCE REPLAY

Ziyu Wang
DeepMind
ziyu@google.com

Victor Bapst
DeepMind
vbapst@google.com

Nicolas Heess
DeepMind
heess@google.com

Volodymyr Mnih
DeepMind
vmnih@google.com

Remi Munos
DeepMind
Munos@google.com

Koray Kavukcuoglu
DeepMind
korayk@google.com

Nando de Freitas
DeepMind, CIFAR, Oxford University
nandodefreitas@google.com

ABSTRACT

This paper presents an actor-critic deep reinforcement learning agent with experience replay that is stable, sample efficient, and performs remarkably well on challenging environments, including the discrete 57-game Atari domain and several continuous control problems. To achieve this, the paper introduces several innovations, including truncated importance sampling with bias correction, stochastic dueling network architectures, and a new trust region policy optimization method.

SAMPLE EFFICIENT ACTOR-CRITIC WITH EXPERIENCE REPLAY

RIKEN BSI-TOYOTA Collaboration Center
Behavioral and Cognitive Neuroscience Unit
Yoshito OGAWA

Outline

- INTRODUCTION
- DISCRETE ACTOR CRITIC WITH EXPERIENCE REPLAY
- RESULTS ON ATARI
- CONTINUOUS ACTOR CRITIC WITH EXPERIENCE REPLAY
- RESULTS ON MUJOCO
- CONCLUSION

INTRODUCTION

Costs of Simulation

- With richer realistic environments, costs of simulation increases
- We need to reduce the number of simulation steps
- Sample efficiency
  - In particular, when agents are deployed in the real world

Experience Replay [Lin, 1992] and Deep Q-Learning

- A valuable tool for improving sample efficiency
- Popular in deep Q-learning
- However, deep Q-learning has two important limitations
  - In indeterministic domain
  - Cost for large action space

Actor Critic Method

- Sample efficient actor critic methods that apply to both continuous and discrete action spaces has been a long-standing hurdle of reinforcement learning

- This paper introduces an actor critic with experience replay (ACER)

DISCRETE ACTOR CRITIC WITH EXPERIENCE REPLAY

Problem Setup

$$Q^{\pi}(x_t, a_t) = \mathbb{E}_{x_{t+1:\infty}, a_{t+1:\infty}}\left[R_t \middle| x_t, a_t\right]$$

$$V^{\pi}(x_t) = \mathbb{E}_{a_t}\left[Q^{\pi}(x_t, a_t) | x_t\right]$$

$$A^{\pi}(x_t, a_t) = Q^{\pi}(x_t, a_t) - V^{\pi}(x_t)$$

Policy Gradient

$$g = \mathbb{E}_{x_{0:\infty}, a_{0:\infty}} \left[\sum_{t \geq 0} A^{\pi}(x_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | x_t) \right]$$

Replacing $A^{\pi}(x_t, a_t)$

with

- the state action value $Q^{\pi}(x_t, a_t)$
- the discounted return R_t
- the temporal difference residual $r_t + \gamma V^{\pi}(x_{t+1}) - V^{\pi}(x_t)$

These have different variance

- estimator using R_t have higher variance and lower bias
- estimator using function approximation have lower variance and higher bias

Combining R_t with the current value function approximation to minimize bias while maintaining bounded variance is one of the central design principles behind ACER

Policy Gradient of A3C [Mnih, 2016]

$$\widehat{g}^{\mathrm{a3c}} = \sum_{t\geq 0}\left(\left(\sum_{i=0}^{k-1}\gamma^{i} r_{t+i}\right) + \gamma^{k} V^{\pi}_{\theta_v}(x_{t+k}) - V^{\pi}_{\theta_v}(x_t)\right)\nabla_{\theta}\log \pi_{\theta}(a_t|x_t)$$

Importance weighted Policy Gradient (Importance Sampling)

$$\widehat{g}^{\text{imp}} = \left(\prod_{t=0}^{k} \rho_t\right) \sum_{t=0}^{k} \left(\sum_{i=0}^{k} \gamma^i r_{t+i}\right) \nabla_\theta \log \pi_\theta(a_t|x_t)$$

$$\rho_t = \frac{\pi(a_t|x_t)}{\mu(a_t|x_t)}$$

Policy Gradient using Marginal Value Function [Degris, 2012]

$$g^{\text{marg}} = \mathbb{E}_{x_t \sim \beta, a_t \sim \mu} \left[\rho_t \nabla_\theta \log \pi_\theta(a_t|x_t) Q^\pi(x_t, a_t) \right]$$

$$R_t^\lambda = r_t + (1 - \lambda)\gamma V(x_{t+1}) + \overline{\lambda\gamma\breve{\rho}_{t+1} R_{t+1}^\lambda}$$

Estimation of Q^{π} Using Retrace [Munos, 2016]

$$Q^{\text{ret}}(x_t, a_t) = r_t + \gamma \bar{\rho}_{t+1}[Q^{\text{ret}}(x_{t+1}, a_{t+1}) - Q(x_{t+1}, a_{t+1})] + \gamma V(x_{t+1})$$

$$\bar{\rho}_t = \min\{c, \rho_t\} \text{ with } \rho_t = \frac{\pi(a_t|x_t)}{\mu(a_t|x_t)}$$

$\lambda = 1$ setting of original retrace

Estimation the critic $Q_{\theta_v}(x_t, a_t)$ Using Retrace

$$(Q^{\text{ret}}(x_t, a_t) - Q_{\theta_v}(x_t, a_t))\nabla_{\theta_v} Q_{\theta_v}(x_t, a_t))$$

Importance Weight Truncation with Bias Correction

$$
\begin{aligned}
g^{\text{marg}} &= \mathbb{E}_{x_t a_t}\left[\rho_t \nabla_\theta \log \pi_\theta(a_t|x_t) Q^\pi(x_t, a_t)\right] \\
&= \mathbb{E}_{x_t}\left[\mathbb{E}_{a_t}[\bar{\rho}_t \nabla_\theta \log \pi_\theta(a_t|x_t) Q^\pi(x_t, a_t)] + \mathop{\mathbb{E}}_{a\sim\pi}\left(\left[\frac{\rho_t(a) - c}{\rho_t(a)}\right]_+ \nabla_\theta \log \pi_\theta(a|x_t) Q^\pi(x_t, a)\right)\right]
\end{aligned}
$$

$$
\widehat{g}^{\text{marg}} = \mathbb{E}_{x_t}\left[\mathbb{E}_{a_t}[\bar{\rho}_t \nabla_\theta \log \pi_\theta(a_t|x_t) Q^{ret}(x_t, a_t)] + \mathop{\mathbb{E}}_{a\sim\pi}\left(\left[\frac{\rho_t(a) - c}{\rho_t(a)}\right]_+ \nabla_\theta \log \pi_\theta(a|x_t) Q_{\theta_v}(x_t, a)\right)\right]
$$

$$
\bar{\rho}_t = \min\{c, \rho_t\} \text{ with } \rho_t = \frac{\pi(a_t|x_t)}{\mu(a_t|x_t)} \;, \quad \rho_t(a) = \frac{\pi(a|x_t)}{\mu(a|x_t)}
$$

$[x]_+ = x$ if $x > 0$ and it is zero otherwise

Off-Policy ACER gradient

$$\begin{aligned}
\widehat{g}^{\text{acer}} \quad &= \quad \bar{\rho}_t \nabla_\theta \log \pi_\theta(a_t|x_t)[Q^{\text{ret}}(x_t, a_t) - V_{\theta_v}(x_t)] \\
&\quad\quad + \mathop{\mathbb{E}}_{a\sim\pi}\left(\left[\frac{\rho_t(a) - c}{\rho_t(a)}\right]_+ \nabla_\theta \log \pi_\theta(a|x_t)[Q_{\theta_v}(x_t, a) - V_{\theta_v}(x_t)]\right)
\end{aligned}$$

Efficient Trust Region Policy Optimization [Schulman, 2015a]

$$\begin{aligned}\widehat{g}^{\text{acer}} &= \bar{\rho}_t \nabla_\theta \log \pi_\theta(a_t|x_t)[Q^{\text{ret}}(x_t, a_t) - V_{\theta_v}(x_t)] \\ &\quad + \mathop{\mathbb{E}}_{a\sim\pi}\left(\left[\frac{\rho_t(a) - c}{\rho_t(a)}\right]_+ \nabla_\theta \log \pi_\theta(a|x_t)[Q_{\theta_v}(x_t, a) - V_{\theta_v}(x_t)]\right)\end{aligned}$$

$$\begin{aligned}\widehat{g}_t^{\text{acer}} &= \bar{\rho}_t \nabla_{\phi_\theta(x_t)} \log f(a_t|\phi_\theta(x))[Q^{\text{ret}}(x_t, a_t) - V_{\theta_v}(x_t)] \\ &\quad + \mathop{\mathbb{E}}_{a\sim\pi}\left(\left[\frac{\rho_t(a) - c}{\rho_t(a)}\right]_+ \nabla_{\phi_\theta(x_t)} \log f(a_t|\phi_\theta(x))[Q_{\theta_v}(x_t, a) - V_{\theta_v}(x_t)]\right)\end{aligned}$$

$$\phi_\theta\text{: } \pi(\cdot|x) = f(\cdot|\phi_\theta(x))$$

$$z^* = \hat{g}_t^{\text{acer}} - \max\left\{0, \frac{k^T \hat{g}_t^{\text{acer}} - \delta}{\|k\|_2^2}\right\} k$$

$$k = \nabla_{\phi_\theta(x_t)} D_{KL}\left[f(\cdot|\phi_{\theta_a}(x_t)\|f(\cdot|\phi_\theta(x_t))\right]$$

RESULTS ON ATARI

Results on Atari

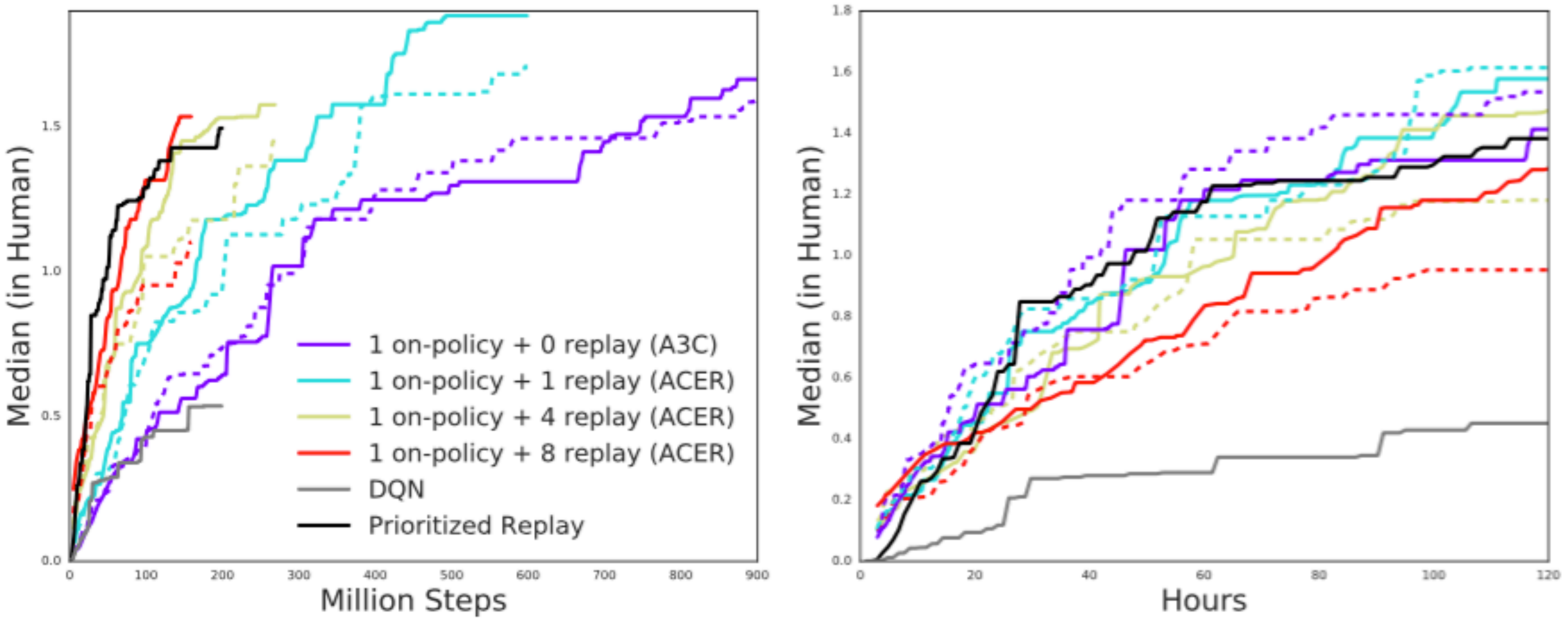

Figure 1: ACER improvements in sample (**LEFT**) and computation (**RIGHT**) complexity on Atari. On each plot, the median of the human-normalized score across all 57 Atari games is presented for 4 ratios of replay with 0 replay corresponding to on-policy A3C. The colored solid and dashed lines represent ACER with and without trust region updating respectively. The environment steps are counted over all threads. The gray curve is the original DQN agent (Mnih et al., 2015) and the black curve is one of the Prioritized Double DQN agents from Schaul et al. (2016).

CONTINUOUS ACTOR CRITIC WITH EXPERIENCE REPLAY

Stochastic Dueling Networks

- Stochastic Dueling Networks estimates both V^{π} and Q^{π}

$$\widetilde{Q}_{\theta_v}(x_t, a_t) \sim V_{\theta_v}(x_t) + A_{\theta_v}(x_t, a_t) - \frac{1}{n}\sum_{i=1}^{n} A_{\theta_v}(x_t, u_i), \text{ and } u_i \sim \pi_\theta(\cdot|x_t)$$

$$V^{target}(x_t) = \min\left\{1, \frac{\pi(a_t|x_t)}{\mu(a_t|x_t)}\right\}\Big(Q^{\text{ret}}(x_t, a_t) - Q_{\theta_v}(x_t, a_t)\Big) + V_{\theta_v}(x_t)$$

Trust Region Updating

$$\begin{aligned} g_t^{\text{acer}} &= \mathbb{E}_{x_t}\Bigg[\mathbb{E}_{a_t}\Big[\bar{\rho}_t \nabla_{\phi_\theta(x_t)} \log f(a_t|\phi_\theta(x_t))(Q^{\text{opc}}(x_t,a_t) - V_{\theta_v}(x_t))\Big] \\ &\quad + \mathop{\mathbb{E}}_{a\sim\pi}\left(\left[\frac{\rho_t(a)-c}{\rho_t(a)}\right]_+ (\widetilde{Q}_{\theta_v}(x_t,a) - V_{\theta_v}(x_t))\nabla_{\phi_\theta(x_t)} \log f(a|\phi_\theta(x_t))\right)\Bigg] \end{aligned}$$

$$\begin{aligned} \hat{g}_t^{\text{acer}} &= \bar{\rho}_t \nabla_{\phi_\theta(x_t)} \log f(a_t|\phi_\theta(x_t))(Q^{\text{opc}}(x_t,a_t) - V_{\theta_v}(x_t)) \\ &\quad + \left[\frac{\rho_t(a'_t)-c}{\rho_t(a'_t)}\right]_+ (\widetilde{Q}_{\theta_v}(x_t,a'_t) - V_{\theta_v}(x_t))\nabla_{\phi_\theta(x_t)} \log f(a'_t|\phi_\theta(x_t)) \end{aligned}$$

RESULTS ON MUJOCO

Results on Muioco

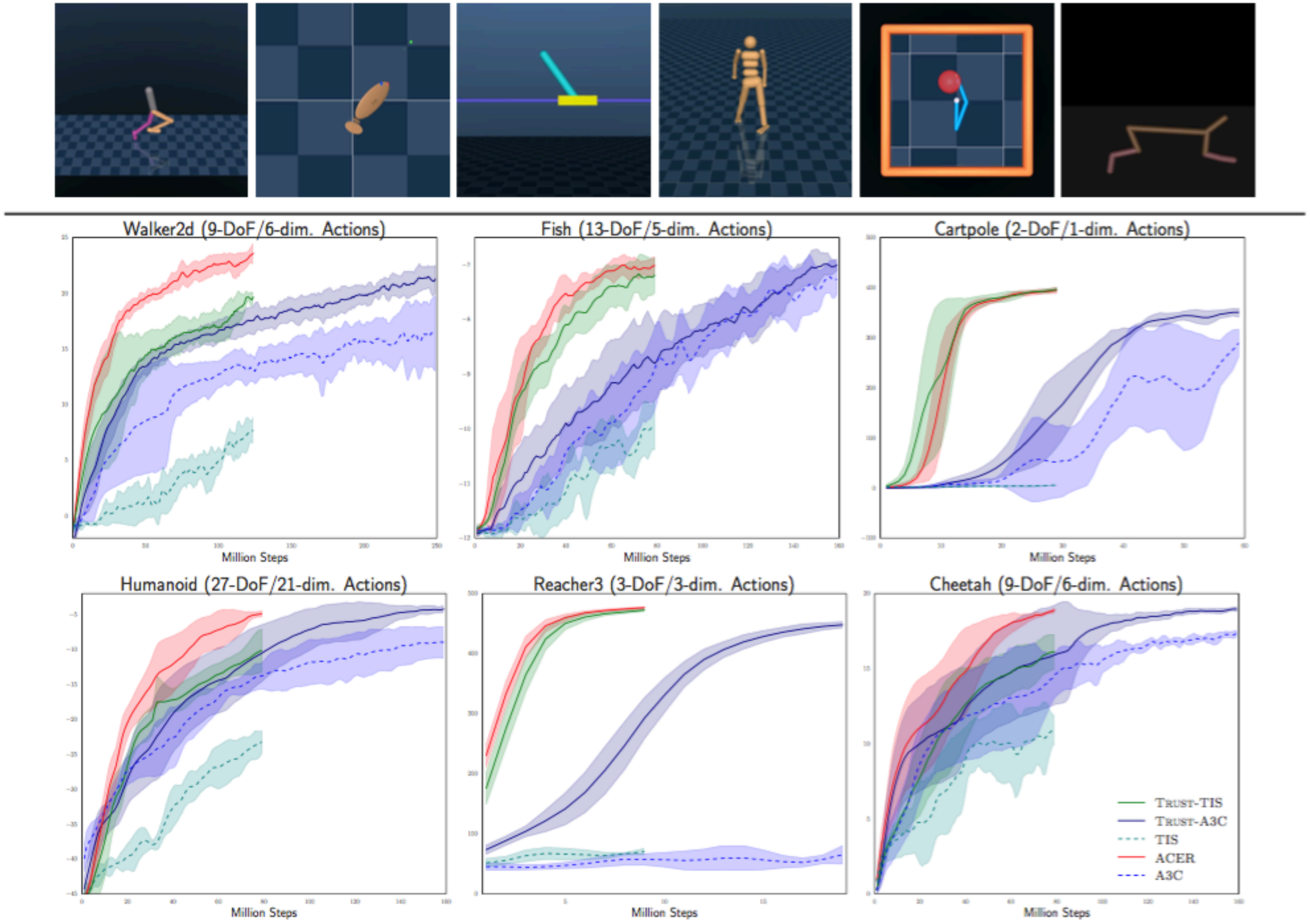

Figure 2: [Top] Screen shots of the continuous control tasks. [Bottom] Performance of different methods on these tasks. ACER outperforms all other methods and shows clear gains for the higher-dimensionality tasks (humanoid, cheetah, walker and fish). The proposed trust region method by itself improves the two baselines (truncated importance sampling and A3C) significantly.

CONCLUSION

Conclusion

- Add off-policy experience replay to A3C
- Use some technique for bound variance
  - Marginal Importance weight
  - Retrace
  - Importance weight truncation with bias correction
  - Efficient trust region policy optimization
- The proposal method improves performance in discrete and continuous condition