AIL 722: Reinforcement Learning

Lecture 37: A2C & A3C

Raunak Bhattacharyya

YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Recap

- Policy gradient
- Reinforce
- Variance Reduction

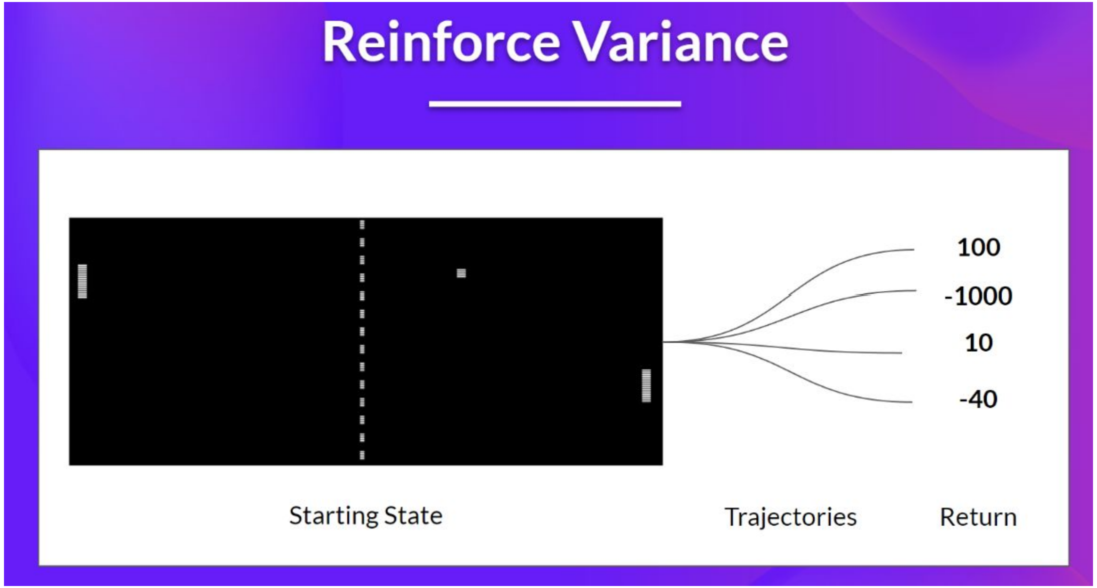
Why Policy Optimisation instead of Value Function

- Might be simpler than V or Q
  - Robotic grasping

- V does not prescribe what action to pick
  - Compute 1 Bellman backup
  - Needs dynamics model

- Q also does not directly prescribe the action
  - Need to compute argmax efficiently

Variance



Let's see a toy example

Hugging Face Deep RL Blog, A2C

Variance Reduction

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \right) \left(\sum_{t=1}^{T} r(s_t, a_t) \right) \right]$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\mathbf{a}_{i,t} | \mathbf{s}_{i,t}) \left(\sum_{t'=t}^{T} r(\mathbf{s}_{i,t'}, \mathbf{a}_{i,t'}) \right)$$

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_{i,t} | s_{i,t}) \left(Q(s_{i,t}, a_{i,t}) - V(s_{i,t}) \right)$$

Recap & Today's Outline

- Policy gradient
- Reinforce
- Variance Reduction
- Actor-critic
- Performance and Examples
- DQN with continuous actions

Actor-Critic

Sample $\{s_i, a_i\}$ from $\pi_\theta(a|s)$

Fit $\hat{V}_\phi(s)$ to sampled reward sums

How to do this?

Evaluate $\hat{A}^\pi(s_i, a_i) = r(s_i, a_i) + \hat{V}_\phi(s_i') - \hat{V}_\phi(s_i)$

Approximation here

$$\nabla_\theta J(\theta) \simeq \sum_i \nabla_\theta \log \pi_\theta(a_i|s_i) \hat{A}^\pi(s_i, a_i)$$

Improve policy by $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$

Advantage computation

$$A(s_t, a_t) = Q_w(s_t, a_t) - V_v(s_t)$$

$$Q(s_t, a_t) = \mathbb{E}[r_{t+1} + \gamma V(s_{t+1})]$$

$$A(s_t, a_t) = r_{t+1} + \gamma V_v(s_{t+1}) - V_v(s_t)$$

Different Approaches

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a)\, G_t\right] && \text{REINFORCE} \\
&= \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a)\; Q^w(s,a)\right] && \text{Q Actor-Critic} \\
&= \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a)\; A^w(s,a)\right] && \text{Advantage Actor-Critic} \\
&= \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a)\; \delta\right] && \text{TD Actor-Critic}
\end{aligned}
$$

Image taken from CMU CS10703 lecture slides

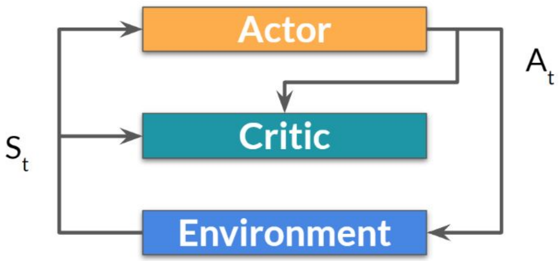
From Sutton and Barto

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 $\mathbf{0}$)

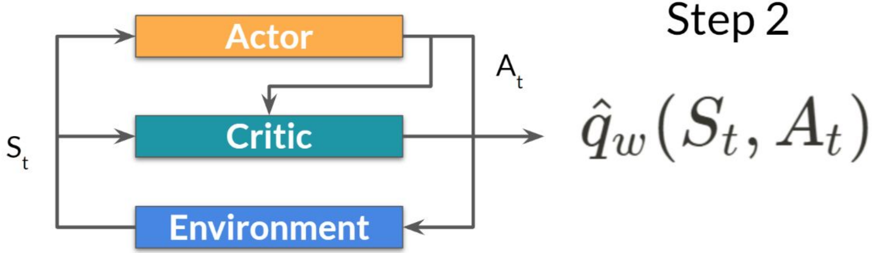
Loop forever (for each episode):
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot, \boldsymbol{\theta})$
  Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
    $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ (G_t)
    $\delta \leftarrow G -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

Actor-Critic Process: Visuals



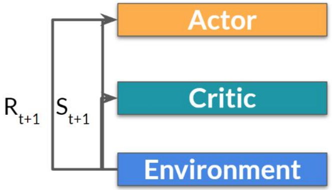
Step 1

Step 2



Hugging Face Deep RL Blog, A2C

Actor-Critic Process: Visuals

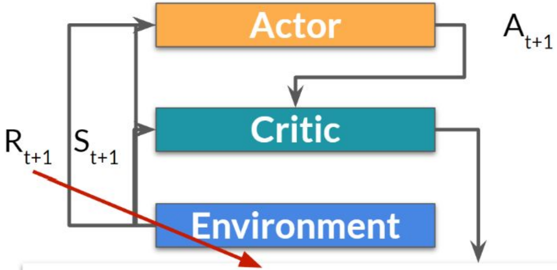


Step 3

Step 4

$$\underbrace{\Delta\theta}_{\text{Change in policy parameters (weights)}} = \alpha\nabla_\theta(log\pi_\theta(s,a))\underbrace{\hat{q}_w(s,a)}_{\text{Action value estimate}}$$

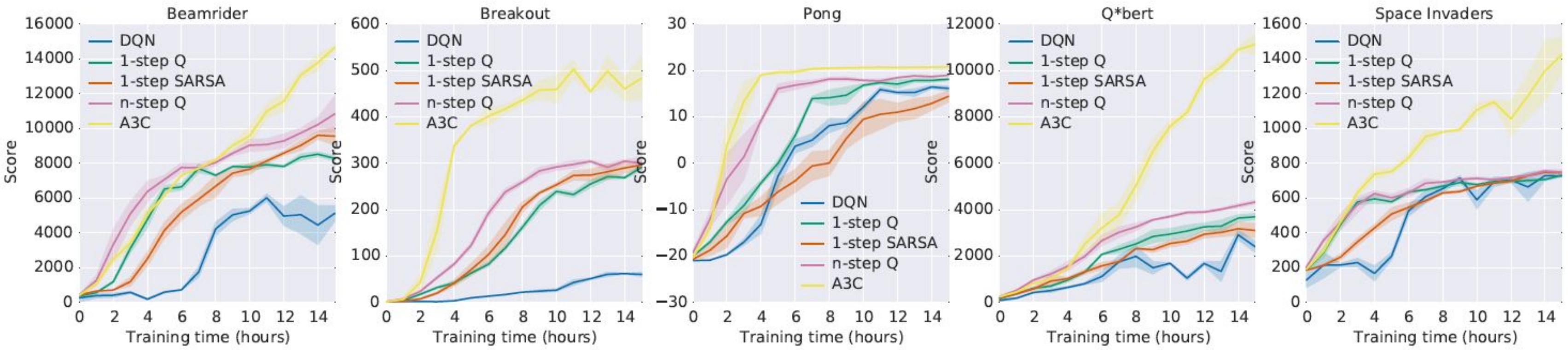
Step 5



$$\Delta w = \beta\,\underbrace{(R(s,a) + \gamma\hat{q}_w(s_{t+1},a_{t+1}) - \hat{q}_w(s_t,a_t))}_{\text{TD error}}\underbrace{\nabla_w\hat{q}_w(s_t,a_t)}_{\text{Gradient of our value function}}$$

Policy and value have different learning rates

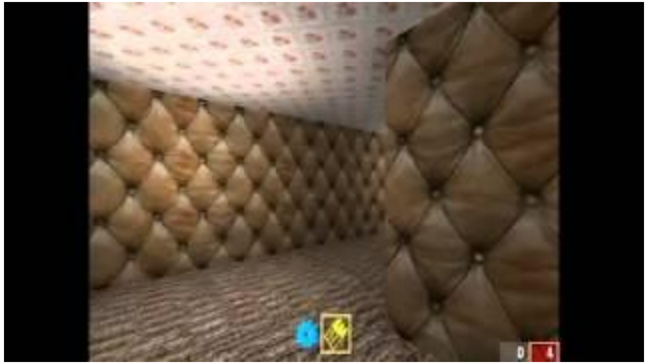
Actor-Critic Performance



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%

Asynchronous Advantage Actor-Critic, Mnih, ICML 2016

Example



Labyrinth, A3C paper, Source: Youtube

Source: Youtube

Paper Presentation

- What are they trying to do:
  - Title
  - Abstract
  - Background
  - Conclusion
- Why are they doing this
  - Title
  - Abstract
  - Background
- How do they do it
  - Methods
  - Appendix
- How do they justify it works: results

- Answer the questions
  - What is the problem and why important
  - Why is it challenging
  - What was done before
  - What has the paper done
- Demonstrate that you understand the math and algorithm
- Rehearse
  - We will impose a hard stop at 10 mins