Move 37
Actor-Critic Methods - Study Guide

Basics:

- Alternates between taking an action and criticising the policy
- **Actor:** decides the policy (which action to take) based on direct rewards
- **Critic:** tells us how good our policy is relative to the environment state value estimates
- Works with continuous action spaces

Relation to other methods:

- Estimates both policy and value function (combines policy and value based methods)
- While policy gradients require waiting till the end of an episode, actor-critic updates at each step, greatly improving efficiency (source)
- ‘Gradient version of policy iteration’ (source)
- Similarities between Actor-Critic and GANs present an opportunity for cross-pollination of methods (source)

Main Actor-Critic methods:

- **A2C, Advantage Actor-Critic** uses the equation A(S, A) = Q(S, A) - V(S) to get the advantage (extra reward) of an action over the estimated value of a state. (source)
- **DDPG, Deep Deterministic Policy Gradient** based on A2C (Maxim Lapan: Deep learning hands on); combines DPG and DQN (2015) (source) (paper)
- **TRPO, Trust Region Policy Optimization** alters the parameter update for actors (source) (2015) (paper)
- **A3C, Asynchronous Advantage Actor-Critic** similar to A2C; uses many actors trained asynchronously (Deepmind, 2016) (paper)
- **PPO, Proximal Policy Optimization** is the OpenAI algorithm of choice for RL as of 2017 (source) (paper)
  OpenAI Five won a best of 3 against a team of top players at Dota using a scaled up version of PPO this August (2018) (source)

A2C vs A3C:

- Maxim Lapan describes A3C as prefered over A2C (see chapter 11 of Deep Reinforcement Learning, Hands On)
- OpenAI lists A2C as prefered to A3C (source)
- A2C is simpler

For a great explanation of A2C see this comic

Recent Advances in Actor-Critic:

- **BAC, Bayesian Actor-Critic** models the policy gradient as a gaussian process (2016) (paper)
- **ACER, Actor-Critic with Experience Replay** improves TRPO (2016) (paper)
- **ACKTR, Actor-Critic using Kronecker-factored Trust Region,** developed by OpenAI (2017) (source) (paper)
- **GAC, Guide Actor-Critic** (2017) (paper)
- **SAC, Soft Actor-Critic** (2018) (paper)
- **TD3, Twin Delayed Deep Deterministic** works to reduce variance; improves DDPG (2018) (paper)
- **D4PG, Deep Distributional DDPG** improves DDPG by allowing critic to use probability distributions (2018) (paper)
- **SPU, Supervised Policy Update** improves both PPO and TRPO (2018) (paper)
- **POP3D, Policy Optimization with Penalized Point Probability Distance** improves TRPO; competitive with PPO (2018) (paper)
- **SIL, Self Imitation Learning** (2018) (paper)

Variations on PPO:

- **AMBER, Adaptive Multi-Batch Experience Replay for Continuous Action Control** (2017) (paper)
- **PPO-CMA, Proximal Policy Optimization with Covariance Matrix Adaptation** (2018) (paper)
- **MPPO, Memory Proximal Policy Optimization** (2018) (paper)
- **PPO-λ** (2018) (paper)