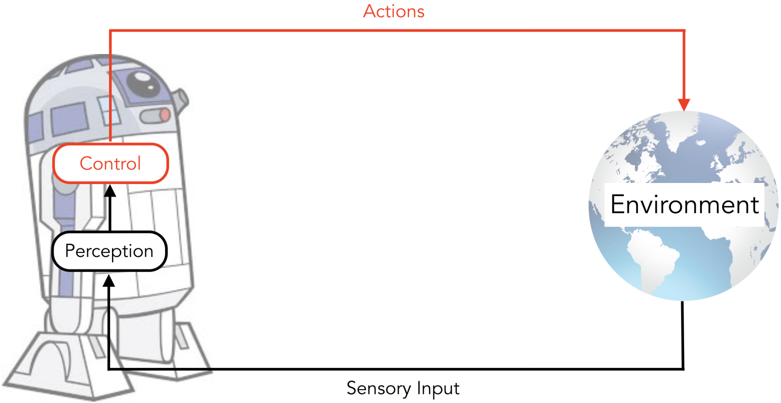
Cooperative Multi-Agent Reinforcement Learning

Shimon Whiteson
Dept. of Computer Science
University of Oxford

joint work with Jakob Foerster, Gregory Farquhar,
Triantafyllos Afouras, and Nantas Nardelli

July 4, 2018

Single-Agent Paradigm



Markov Decision Process

- Agent observes the *state* s
- Selects an *action*: $u \in U$
- State transitions: $P(s'|s,u) : S \times U \times S \rightarrow [0,1]$
- Receives *reward*: $r(s,u) : S \times U \rightarrow \mathbb{R}$
- Goal: maximise expected cumulative discounted *return*:

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$$

Value Functions

- Given a policy $\pi(s, a)$, the *value function* is:

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[R_t | s_t = s\right]$$

- The *action-value function* is:

$$Q^{\pi}(s, a) = \mathbb{E}_{\pi}\left[R_t | s_t = s, a_t = a\right]$$

- Estimate Q-values using a *temporal difference* update rule:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$

- Act (soft) *greedily* wrt to Q-values:

$$a_t = \arg\max_{a} Q(s_t, a)$$

Policy Gradient Methods

- What about when *greedification* is hard, e.g., continuous actions?
- Optimise π_θ with gradient ascent on expected return:

$$J_\theta = \mathbb{E}_{s \sim \rho^\pi(s), u \sim \pi_\theta(s,\cdot)} \left[r(s, u) \right]$$

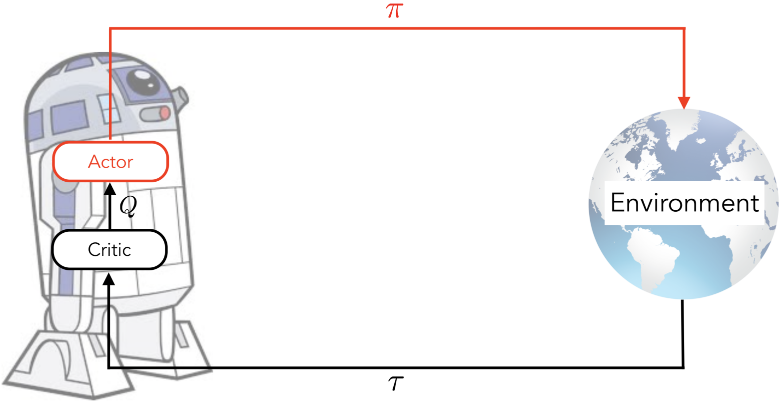
- Policy gradient theorem [Sutton et al. 2000]:

$$\nabla_\theta J_\theta = \mathbb{E}_{s \sim \rho^\pi(s), u \sim \pi_\theta(s,\cdot)} \left[\nabla_\theta \log \pi_\theta(u|s) Q^\pi(s, u) \right]$$

Actor-Critic Methods [Sutton et al. 00]

- Estimate gradient with trajectory τ and learned *critic* $Q(s, u)$:

$$g(\tau) = \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(u_t|s_t) Q(s_t, u_t)$$



Baselines

- Reduce variance with a *baseline* $b(s)$:

$$g(\tau) = \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(u_t|s_t)(Q(s_t, u_t) - b(s_t))$$

- $b(s) = V(s) \implies Q(s,u) - b(s) = A(s,u)$, the *advantage function*:

$$g(\tau) = \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(u_t|s_t) A(s_t, u_t)$$

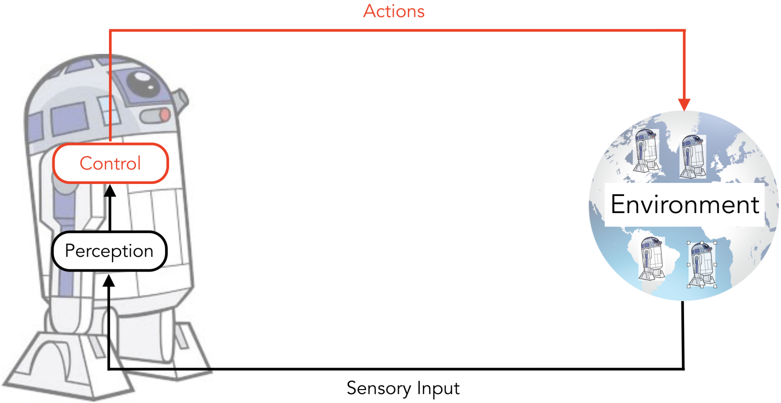
- *TD-error* $r_t + \gamma V(s_{t+1}) - V(s)$ is an unbiased estimate of $A(s_t, u_t)$:

$$g(\tau) = \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(u_t|s_t)(r_t + \gamma V(s_{t+1}) - V(s_t))$$

Deep Actor-Critic Methods

- Actor and critic are both deep neural networks
  - Convolutional and recurrent layers
  - Actor and critic share layers
- Both trained with stochastic gradient descent
  - Actor trained on policy gradient
  - Critic trained on TD(λ) or Sarsa(λ)

Multi-Agent Paradigm



Types of Multi-Agent Systems

- *Cooperative:*
  - Shared team reward
  - Coordination problem
- *Competitive:*
  - Zero-sum games
  - Individual opposing rewards
  - Minimax equilibria
- *Mixed:*
  - General-sum games
  - Nash equilibria
  - What is the question?

Coordination Problems are Everywhere

Multi-Agent MDP

- All agents see the global state s
- Individual actions: $u^a \in U$
- State transitions: $P(s'|s, \mathbf{u}) : S \times \mathbf{U} \times S \to [0, 1]$
- Shared team reward: $r(s, \mathbf{u}) : S \times \mathbf{U} \to \mathbb{R}$
- Equivalent to an MDP with a factored action space

Dec-POMDP

- Observation function: $O(s, a) : S \times A \rightarrow Z$
- Action-observation history: $\tau^a \in T \equiv (Z \times U)^*$
- Decentralised policies: $\pi^a(u^a|\tau^a) : T \times U \rightarrow [0, 1]$
- Centralised learning of decentralised policies

Independent Actor-Critic

- Inspired by *independent Q-learning* [Tan 1993]
  - Each agent learns independently with its own actor and critic
  - Treats other agents as part of the environment
- Speed learning with *parameter sharing*
  - Different inputs, including a, induce different behaviour
  - Still independent: critics condition only on τ^a and u^a
- Limitations:
  - Nonstationary learning
  - Hard to learn to coordinate
  - Multi-agent credit assignment

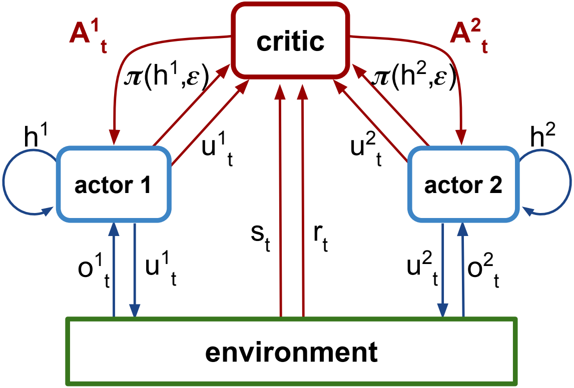
Counterfactual Multi-Agent Policy Gradients

- Centralised critic: stabilise learning to coordinate
- Counterfactual baseline: tackle multi-agent credit assignment
- Efficient critic representation: scale to large NNs

Centralised Critic

Centralisation $\rightarrow$ Hard Greedification $\rightarrow$ Actor-Critic

$$g_a(\tau) = \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(u_t^a | \tau_t^a)(r_t + \gamma V(s_{t+1}) - V(s_t))$$



Wonderful Life Utility [Wolpert & Tumer 2000]

Difference Rewards [Tumer & Agogino 2007]

- Per-agent shaped reward:

$$D^a(s, \mathbf{u}) = r(s, \mathbf{u}) - r(s, (\mathbf{u}^{-a}, c^a))$$

  where c^a is a *default action*

- Limitations:
  - Need extra simulation to estimate counterfactual $r(s, (\mathbf{u}^{-a}, c^a))$
  - Need domain knowledge to choose c^a

Counterfactual Baseline

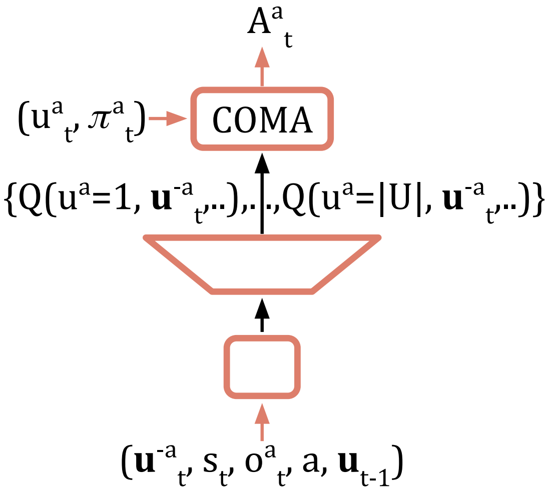
- Use $Q(s, \mathbf{u})$ to estimate difference rewards:

$$g_a(\tau) = \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(u_t^a | \tau_t^a) A^a(s_t, \mathbf{u}_t)$$

$$A^a(s, \mathbf{u}) = Q(s, \mathbf{u}) - \sum_{u^a} \pi^a(u^a | \tau^a) Q(s, (\mathbf{u}^{-a}, u^a))$$

- Baseline computes expectation wrt u^a
- Critic obviates need for extra simulations
- Expectation obviates need for default action

Efficient Critic Representation



Starcraft

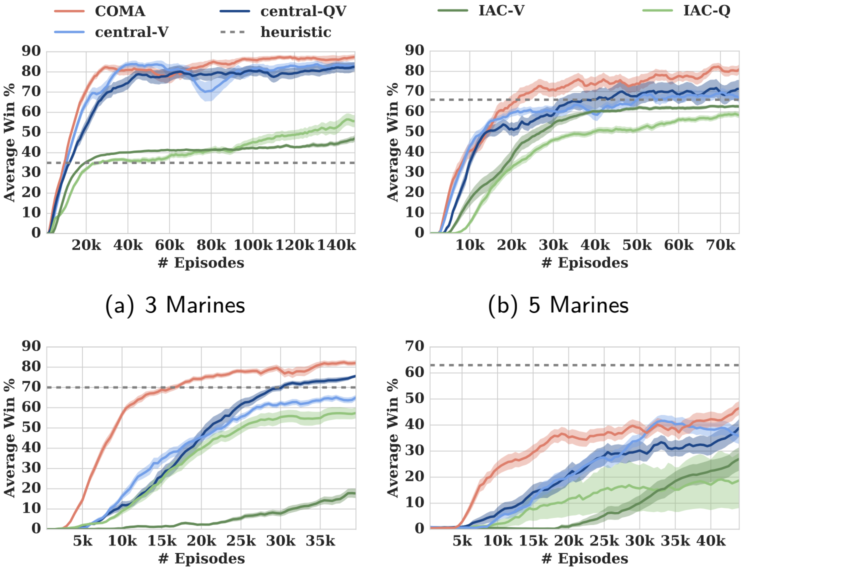
Starcraft Micromanagement [Synnaeve et al. 2016]

Decentralised Starcraft Micromanagement

Baseline Algorithms

- *IAC-V:* independent actor-critic with $V(\tau^a)$ (TD error)
- *IAC-Q:* independent actor-critic with $A(\tau^a, u^a) = Q(\tau^a, u^a) - V(\tau^a)$
- *Central-V:* centralised critic $V(s)$ (TD error)
- *Central-QV:*
  - Centralised critics $Q(s, \mathbf{u})$ and $V(s)$
  - Advantage gradient $A(s, \mathbf{u}) = Q(s, \mathbf{u}) - V(s)$
  - COMA but with $b(s) = V(s)$

COMA Results vs. Baselines (Average Performance)



(a) 3 Marines

(b) 5 Marines

(c) 5 Wraiths

(d) 2 Dragoons & 3 Zealots

COMA Results vs. Centralised (Best Agents)

Map	COMA	Heuristic	DQN	GMEZO
3 Marines	98	74	-	-
5 Marines	95	98	99	100
5 Wraiths*	98	82	70	74
2 Dragoons & 3 Zealots	65	68	61	90

Paper

Counterfactual Multi-Agent Policy Gradients

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras,
Nantas Nardelli, and Shimon Whiteson

The Outstanding Student Paper of AAAI-18