CS4400
DEEP REINFORCEMENT LEARNING

Lecture 10: Deep Multi-agent RL

Wendelin Böhmer

<j.w.bohmer@tudelft.nl>

9th of January 2024

Content of this lecture

10.1

Deep Multi-agent RL

Multi-agent games

10.1 Single- vs. Multi-agent environments

- Single agents assume *stationary* environment
  - all other objects are *passive*
  - always react the same to player actions
- Real world has many different actors
  - actors have *intentions* (reward functions)
  - actors can change behavior (in response)
  - environment no longer stationary
- Multi-agent RL formulates environment as a *game*

10.1 Terminology of games

- single/two/multi-player game
  - Solitaire/Chess/Settlers-of-Catan

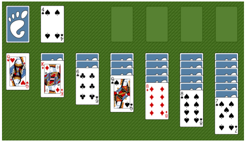


Image sources: wikimedia.org

10.1 Terminology of games

- single/two/multi-player game
  - Solitaire/Chess/Settlers-of-Catan
- simultaneous/sequential moves
  - Rock-paper-scissors/Go

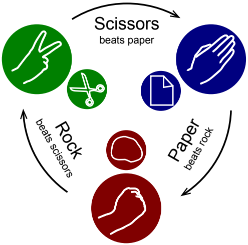


Image sources: wikimedia.org

10.1 Terminology of games

- single/two/multi-player game
  - Solitaire/Chess/Settlers-of-Catan
- simultaneous/sequential moves
  - Rock-paper-scissors/Go
- stochastic/deterministic moves
  - Poker/Chess

Image sources: wikimedia.org

Terminology of games

- single/two/multi-player game
  - Solitaire/Chess/Settlers-of-Catan
- simultaneous/sequential moves
  - Rock-paper-scissors/Go
- stochastic/deterministic moves
  - Poker/Chess
- partial/perfect information
  - Card-games/Connect-four

Image sources: wikimedia.org

10.1 Terminology of games

- single/two/multi-player game
  - Solitaire/Chess/Settlers-of-Catan
- simultaneous/sequential moves
  - Rock-paper-scissors/Go
- stochastic/deterministic moves
  - Poker/Chess
- partial/perfect information
  - Card-games/Connect-four
- discrete/continual state/actions/time
  - turn-based/real-time strategy

Image sources: wikimedia.org, Samvelyan et al. (2019)

10.1 Terminology of games

- single/two/multi-player game
  - Solitaire/Chess/Settlers-of-Catan
- simultaneous/sequential moves
  - Rock-paper-scissors/Go
- stochastic/deterministic moves
  - Poker/Chess
- partial/perfect information
  - Card-games/Connect-four
- discrete/continual state/actions/time
  - turn-based/real-time strategy
- zero-sum/general-sum/cooperative game
  - Chess/Settlers-of-Catan/Hanabi

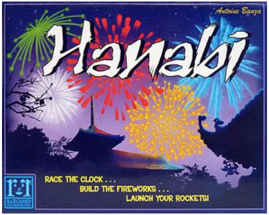


Image sources: wikimedia.org

10.1 Partially observable stochastic games

- POSG $\langle \mathcal{S}, \{\mathcal{A}^i\}_{i=1}^N, \{\mathcal{O}^i\}_{i=1}^N, \rho, P, \{R^i\}_{i=1}^N, \{O^i\}_{i=1}^N \rangle$
  - state space $s \in \mathcal{S}$ of the game
  - action space $a^i \in \mathcal{A}^i$ for each agent $1 \leq i \leq N$
    - joint actions denoted as $\boldsymbol{a} \in \mathcal{A} := \mathcal{A}^1 \times \ldots \times \mathcal{A}^N$
  - observation space $o^i \in \mathcal{O}^i$ for each agent $1 \leq i \leq N$
  - initial state distribution $s_0 \sim \rho(\cdot)$
  - transition probability $s_{t+1} \sim P(\cdot|s_t, \boldsymbol{a}_t)$
  - reward probability $r_t^i \sim R^i(\cdot|s_t, \boldsymbol{a}_t, s_{t+1})$ for each agent $1 \leq i \leq N$
  - observation function $o_t^i \sim O^i(\cdot|s_t)$ for each agent $1 \leq i \leq N$
    - action-observation histories $\tau_t^i := [o_0^i, a_0^i, \ldots, o_t^i] \in (\mathcal{O}^i \times \mathcal{A}^i)^t \times \mathcal{O}^i$
    - joint action-observation history $\boldsymbol{\tau}_t := \{\tau_t^i\}_{i=1}^N$
    - joint history is sufficient belief: $r(\boldsymbol{\tau}_t, \boldsymbol{a})$ and $P(\boldsymbol{\tau}_{t+1}|\boldsymbol{\tau}_t, \boldsymbol{a}_t)$ exist
- Decentralized policy $\boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{\tau}_t) = \prod_{i=1}^N \pi^i(a^i|\tau_t^i)$

see Oliehoek and Amato (2016) for a sound introduction

core concept: Independent Q-learning (IQL)

- Simply ignore non-stationarity induced by other agents
  - estimate decentralized Q-values $\{q^i_{\theta_i}(\tau^i_t, a^i)\}_{i=1}^N$

$$\mathcal{L}^{\text{IQL}} := \sum_{i=1}^{N} \mathbb{E}\Big[\sum_{t=0}^{n-1} \Big(r^i_t + \gamma \max_{a^i} q^i_{\theta'_i}(\tau^i_{t+1}, a^i) - q^i_{\theta_i}(\tau^i_t, a^i_t)\Big)^2 \Big| \langle \tau^i_t, a^i_t, r^i_t, \tau^i_{t+1}\rangle \in \mathcal{D}^i\Big]$$

- Convergence for *joint training* with small learning rates
  - other agents' policies almost stationary

- Similar for all flavors of independent policy-gradient methods

IQL first by Tan (1993); see Schröder de Witt et al. (2020) for an example of independent PPO

10.1 Summary

- Multiple non-stationary agents are formalized by games
- POSG formalize discrete time games:
  - multi-agent
  - simultaneous-move
  - stochastic
  - partial-information
- IQL learns POSG by assuming stationary players

Learning Objectives

LO10.1: Classify a game in the given terminology
LO10.2: Explain and implement IQL

10.2

Deep Multi-agent RL
Game theory

core concept: Nash equilibria

- How *should* rational agents/players behave?
- **Rational**: players **maximize** *only* their *own outcome*
  - choosing (silent, silent) appears optimal

Prisoner's dilemma

A \ B	confess	silent
confess	A: -5, B: -5	A: 0, B: -10
silent	A: -10, B: 0	A: -1, B: -1

core concept: Nash equilibria

- How *should* rational agents/players behave?
- **Rational**: players **maximize** *only* their *own outcome*
  - choosing (silent, silent) appears optimal
  - assume A remains silent and B remains silent
  - if A remains silent, B should confess

Prisoner's dilemma

A \ B	confess	silent
confess	A: -5, B: -5	A: 0, B: -10
silent	A: -10, B: 0	A: -1, B: -1

core concept: Nash equilibria

- How *should* rational agents/players behave?
- **Rational**: players **maximize** *only* their *own outcome*
  - choosing (silent, silent) appears optimal
  - assume A remains silent and B remains silent
  - if A remains silent, B should confess
  - if B remains confess, A should confess
  - if A confess and B confess,
    neither should change

Prisoner's dilemma

A \ B	confess	silent
confess	A: -5, B: -5	A: 0, B: -10
silent	A: -10, B: 0	A: -1, B: -1

en.wikipedia.org/wiki/Nash_equilibrium

core concept: Nash equilibria

- How *should* rational agents/players behave?
- **Rational**: players **maximize** *only* their *own outcome*
  - choosing (silent, silent) appears optimal
  - assume A remains silent and B remains silent
  - if A remains silent, B should confess
  - if B remains confess, A should confess
  - if A confess and B confess,
    neither should change
- Definition **Nash equilibrium** (a^*, b^*):
  - $r^A(a^*, b^*) \geq r^A(a, b^*), \quad \forall a$
  - $r^B(a^*, b^*) \geq r^B(a^*, b), \quad \forall b$
- (silent, silent) is not a Nash equilibrium!

Prisoner's dilemma

A \ B	confess	silent
confess	-5 / -5	0 / -10
silent	-10 / 0	-1 / -1

en.wikipedia.org/wiki/Nash_equilibrium

10.2 General-sum multi-player games

- Every agent has its own centralized value function $Q_i^{\boldsymbol{\pi}}(\boldsymbol{\tau}_t, \boldsymbol{a})$
  - depends on other agents' actions $\boldsymbol{a}^{-i} \in \mathcal{A}^{-i}, \boldsymbol{a} = \{a^i\} \cup \boldsymbol{a}^{-i}$
- Differently valued *Nash equilibria* $\boldsymbol{a}_* \in \mathcal{A}$ (or none) can exist
  - no other action a^i of agent i is better if all other actions $\boldsymbol{a}_*^{-i}$ remain
  - $Q_i^{\boldsymbol{\pi}}(\boldsymbol{\tau}_t, \boldsymbol{a}_*) \geq Q_i^{\boldsymbol{\pi}}(\boldsymbol{\tau}_t, \{a^i\} \cup \boldsymbol{a}_*^{-i}), \forall a^i \in \mathcal{A}^i \setminus \{a_*^i\}, \forall i \in \{1, \dots, N\}$

A \ B	stag	hare
stag	2 / 2	1 / 0
hare	0 / 1	1 / 1

the stag-hunt game

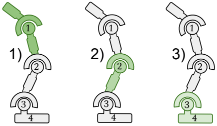
A \ B	head	tail
head	0 / 1	1 / 0
tail	1 / 0	0 / 1

matching pennies

10.2 General-sum multi-player games

- Every agent has its own centralized value function $Q_i^{\boldsymbol{\pi}}(\boldsymbol{\tau}_t, \boldsymbol{a})$
  - depends on other agents' actions $\boldsymbol{a}^{-i} \in \mathcal{A}^{-i}, \boldsymbol{a} = \{a^i\} \cup \boldsymbol{a}^{-i}$
- Differently valued *Nash equilibria* $\boldsymbol{a}_* \in \mathcal{A}$ (or none) can exist
  - no other action a^i of agent i is better if all other actions $\boldsymbol{a}_*^{-i}$ remain
  - $Q_i^{\boldsymbol{\pi}}(\boldsymbol{\tau}_t, \boldsymbol{a}_*) \geq Q_i^{\boldsymbol{\pi}}(\boldsymbol{\tau}_t, \{a^i\} \cup \boldsymbol{a}_*^{-i}), \forall a^i \in \mathcal{A}^i \setminus \{a_*^i\}, \forall i \in \{1, \dots, N\}$
- Requires extensive search for all Nash equilibria (NE)
  - there might not be a NE
  - not all NE are equally good for everyone
  - not all information is available to decentralized agents
  - agents might play different NE (bad for everyone)

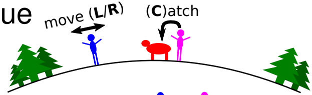
Cooperative multi-player games



- Very common engineering setup
  - independent components need to work together
- Reward functions r^i of all agents are identical

$$r^i(\boldsymbol{\tau}_t, \boldsymbol{a}) \; := \; r^j(\boldsymbol{\tau}_t, \boldsymbol{a}), \quad \forall i, j, \boldsymbol{\tau}_t, \boldsymbol{a}$$

- No conflict of interest: unique centralized value



$$Q_i^{\boldsymbol{\pi}}(\boldsymbol{\tau}_t, \boldsymbol{a}) \; = \; Q_j^{\boldsymbol{\pi}}(\boldsymbol{\tau}_t, \boldsymbol{a}), \quad \forall i, j, \boldsymbol{\tau}_t, \boldsymbol{a}$$

- Independent value is decentralized expectation
  - $q^i(\tau_t^i, a^i; \boldsymbol{\pi}) \; := \; \mathbb{E}_{\boldsymbol{\pi}}\Big[Q_i^{\boldsymbol{\pi}}(\boldsymbol{\tau}_t, \boldsymbol{a}) \,\Big|\, \tau_t^i, a^i \Big], \quad \forall i, \tau_t^i, a^i$
  - $\pi^i(a^i | \tau_t^i) := 1$ iff $a^i = \arg\max_{a^i \in \mathcal{A}^i} q^i(\tau_t^i, a^i; \boldsymbol{\pi})$

	C	L	R
C	+r	-p	-p
L	-p	0	0
R	-p	0	0

- Optimal centralized $\neq$ optimal decentralized policy
  - decentralization may require more *information gathering actions*

assignment sheet 4

called Decentralized POMDP, see Oliehoek and Amato (2016) for a sound derivation

10.2 Zero-sum two-player games

- Very common board game setup
- Reward functions always add to zero (or a constant)

$$r^1(\boldsymbol{\tau}_t, \boldsymbol{a}) \; := \; -r^2(\boldsymbol{\tau}_t, \boldsymbol{a}), \quad \forall \boldsymbol{\tau}_t, \boldsymbol{a}$$

- Which previously discussed games are zero-sum?

assignment sheet 4

for a more formal definition see e.g. Raghavan (1994)

10.2 Zero-sum two-player games

- Very common board game setup
- Reward functions always add to zero (or a constant)

$$r^1(\boldsymbol{\tau}_t, \boldsymbol{a}) \;:=\; -r^2(\boldsymbol{\tau}_t, \boldsymbol{a}), \quad \forall \boldsymbol{\tau}_t, \boldsymbol{a}$$

- Which previously discussed games are zero-sum?

A \ B	head	tail
head	1 / 0	0 / 1
tail	0 / 1	1 / 0

matching pennies

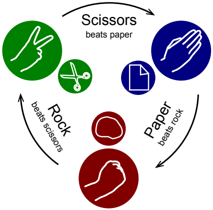
10.2 Zero-sum two-player games

- Very common board game setup
- Reward functions always add to zero (or a constant)

$$r^1(\boldsymbol{\tau}_t, \boldsymbol{a}) \;:=\; -r^2(\boldsymbol{\tau}_t, \boldsymbol{a}), \quad \forall \boldsymbol{\tau}_t, \boldsymbol{a}$$

- Unique value for *equal information* games
  - mirrored value function: $Q_1^{\boldsymbol{\pi}}(\boldsymbol{\tau}_t, \boldsymbol{a}) = -Q_2^{\boldsymbol{\pi}}(\boldsymbol{\tau}_t, \boldsymbol{a}),\ \forall \boldsymbol{\tau}_t, \boldsymbol{a}$
  - all Nash equilibria have the same value

$$V(\boldsymbol{\tau}_t) \;=\; \max_{a^1 \in \mathcal{A}^1} \min_{a^2 \in \mathcal{A}^2} Q_1^{\boldsymbol{\pi}}(\boldsymbol{\tau}_t, \boldsymbol{a}), \quad \forall \boldsymbol{\tau}_t$$

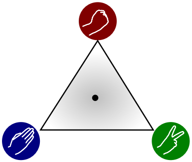


- Can be learned by *self-play*
  - agent plays against mirrored self ($\max \to \min$)
  - unequal information $\tau_t^1 \neq \tau_t^2$ can lead to cycles

A4.2 self-play became popular with AlphaGo (Silver et al., 2016, 2017, 2018); see e.g. Vinyals et al. (2019) for cyclic games

10.2 Cyclic games

- Optimal move can depend on other player's *policy/strategy*
  - sometimes no Nash equilibrium exists
  - simultaneous moves or unequal information

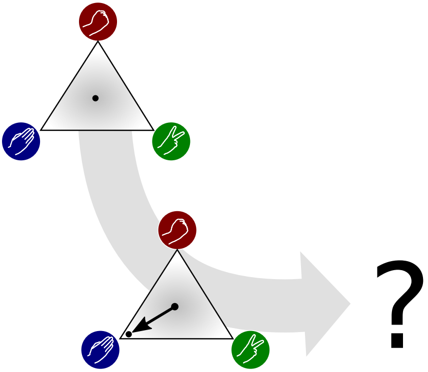


A \ B	rock	paper	scissors
rock	0 / 0	+1 / -1	-1 / +1
paper	-1 / +1	0 / 0	+1 / -1
scissors	+1 / -1	-1 / +1	0 / 0

images modified from wikimedia.org

10.2 Cyclic games

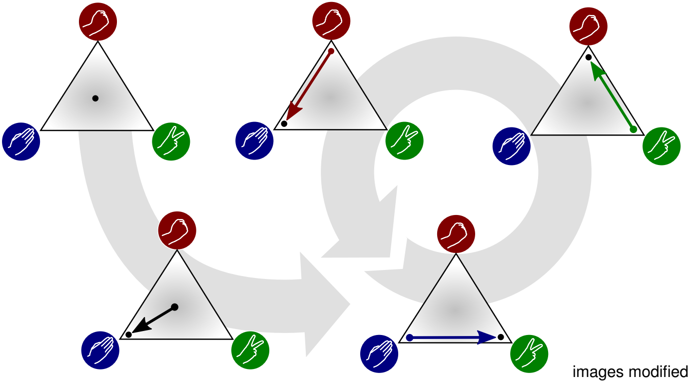
- Optimal move can depend on other player's *policy/strategy*
  - sometimes no Nash equilibrium exists
  - simultaneous moves or unequal information



images modified from wikimedia.org

10.2 Cyclic games

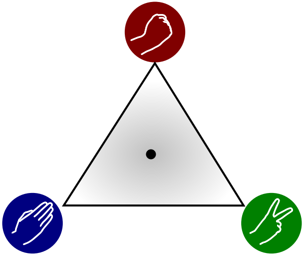
- Optimal move can depend on other player's *policy/strategy*
  - sometimes no Nash equilibrium exists
  - simultaneous moves or unequal information
- Self-play assumes opponent uses the same policy!



images modified from wikimedia.org

10.2 Leagues

- What is the optimal response in a zero sum cyclic game?
  - mixed Nash equilibrium $\max_{\pi^1} \min_{\pi^2} \mathbb{E}\Big[Q(\boldsymbol{\tau}_t, \boldsymbol{a}) \,\Big|\, \substack{a^1 \sim \pi^1(\cdot|\tau_t^1) \\ a^2 \sim \pi^2(\cdot|\tau_t^2)} \Big]$
  - average case response $\max_{\pi^1} \mathbb{E}\Big[Q(\boldsymbol{\tau}_t, \boldsymbol{a}) \,\Big|\, \substack{a^1 \sim \pi^1(\cdot|\tau_t^1) \\ a^2 \sim \pi'(\cdot|\tau_t^2)}, \pi' \sim \Pi \Big]$



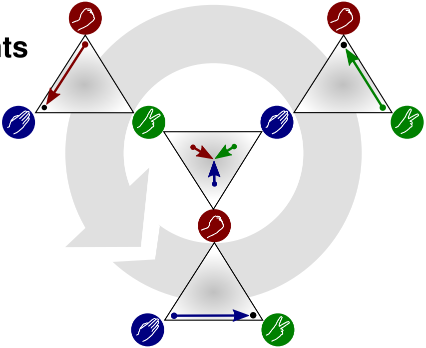
A4.3

images modified from wikimedia.org

Leagues

- What is the optimal response in a zero sum cyclic game?
  - mixed Nash equilibrium $\max\limits_{\pi^1} \min\limits_{\pi^2} \mathbb{E}\Big[Q(\boldsymbol{\tau}_t, \boldsymbol{a}) \,\Big|\, \substack{a^1 \sim \pi^1(\cdot|\tau_t^1) \\ a^2 \sim \pi^2(\cdot|\tau_t^2)} \Big]$
  - average case response $\max\limits_{\pi^1} \mathbb{E}\Big[Q(\boldsymbol{\tau}_t, \boldsymbol{a}) \,\Big|\, \substack{a^1 \sim \pi^1(\cdot|\tau_t^1) \\ a^2 \sim \pi'(\cdot|\tau_t^2)}, \pi' \sim \Pi \Big]$

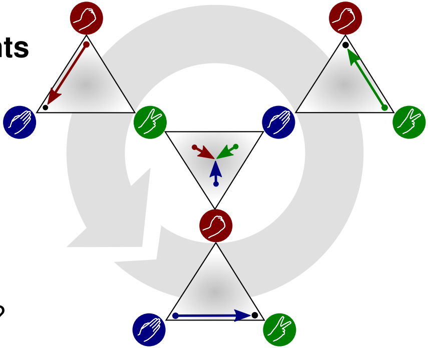
- Can be trained in a **league of agents**
  - keep old policies around
  - play against all of them
  - use either worst or average loss



A4.3 this idea has been recently popularized by AlphaStar (Vinyals et al., 2019); images modified from `wikimedia.org`

10.2 Leagues

- What is the optimal response in a zero sum cyclic game?
  - mixed Nash equilibrium $\max_{\pi^1} \min_{\pi^2} \mathbb{E}\Big[Q(\boldsymbol{\tau}_t, \boldsymbol{a}) \,\Big|\, \substack{a^1 \sim \pi^1(\cdot|\tau_t^1) \\ a^2 \sim \pi^2(\cdot|\tau_t^2)}\Big]$
  - average case response $\max_{\pi^1} \mathbb{E}\Big[Q(\boldsymbol{\tau}_t, \boldsymbol{a}) \,\Big|\, \substack{a^1 \sim \pi^1(\cdot|\tau_t^1) \\ a^2 \sim \pi'(\cdot|\tau_t^2)}, \pi' \sim \Pi\Big]$
- Can be trained in a **league of agents**
  - keep old policies around
  - play against all of them
  - use either worst or average loss
- Open research questions:
  - *which* policies should be kept?
  - let *adversary* choose who to play?



A4.3 this idea has been recently popularized by AlphaStar (Vinyals et al., 2019); images modified from wikimedia.org

10.2 Summary

- Nash equilibria (NE) are stable for rational agents
- General-sum games can have many or no NE
- Cooperative and zero-sum games have unique centralized values
- Without NE, cyclic games must be solved by leagues

Learning Objectives

LO10.3: Define, explain and find Nash equilibria
LO10.4: Explain general-sum, zero-sum and collaborative games

10.3

Deep Multi-agent RL
Centralized training

10.3 Centralized training and decentralized execution

- More information available during *centralized training*
  - other agents' actions $\boldsymbol{a}_t^{-i}$ and histories $\boldsymbol{\tau}_t^{-i}$
  - sometimes true state s_t, e.g. for value functions $V^\pi(s_t, \boldsymbol{\tau}_t)$
- Centralized training allows *parameter sharing*
  - all agents have the same architecture and parameters θ
  - extending input with class/role/id differentiates agents
- Effectively reuses training data of all agents
  - enforces *permutation invariance* between agents
  - can dramatically improve sample efficiency
- Example: IQL with centralized training and parameter sharing
  - DRQN implementation where `dim=-2` stacks agents

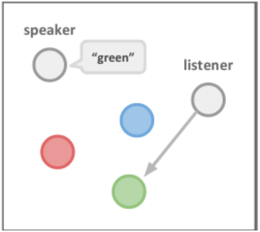
see e.g. `pymarl` for a collaborative IQL implementation (Samvelyan et al., 2019)

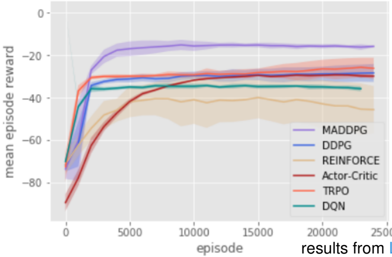
Multi-agent DDPG (MADDPG)

- Independent DDPG with centralized Q-value functions $Q^{\boldsymbol{\pi}}_{\phi_i}$
  - parameter sharing only for collaborative games
  - fixes all other agents' behavior to $\boldsymbol{a}_t^{-i}$ from replay buffer

$$\mathcal{L}^{\text{MADDPG}}_{\mu[\boldsymbol{\theta}]} := -\sum_{i=1}^{N} \mathbb{E}_\mu\Big[\sum_{t=0}^{n-1} Q^{\boldsymbol{\pi}}_{\phi_i}\big(s_t, \{\pi^i_{\theta_i}(\tau^i_t)\} \cup \boldsymbol{a}_t^{-i}\big)\Big]$$

$$\mathcal{L}^{\text{MADDPG}}_{Q[\boldsymbol{\phi}]} := \sum_{i=1}^{N} \mathbb{E}_\mu\Big[\sum_{t=0}^{n-1}\Big(r^i_t + \gamma Q^{\boldsymbol{\pi}}_{\phi'_i}\big(s_{t+1}, \{\pi^j_{\theta'_j}(\tau^j_{t+1})\}_{j=1}^N\big) - Q^{\boldsymbol{\pi}}_{\phi_i}(s_t, \boldsymbol{a}_t)\Big)^2\Big]$$





results from Lowe et al. (2017)

Counterfactual multi-agent learning (COMA)

- Stochastic policy-gradients in cooperative games
  - centralized training with parameter sharing
  - centralized value $Q^{\boldsymbol{\pi}}_{\phi}(s, \boldsymbol{a})$, decentr. policy $\boldsymbol{\pi}_\theta(\boldsymbol{a}|\boldsymbol{\tau}_t) = \prod_{i=1}^{N} \pi^i_\theta(a^i|\tau^i_t)$

$$\mathcal{L}^{\text{C-QV}}_{\pi[\theta]} := -\sum_{i=1}^{N} \mathbb{E}_{\pi_\theta}\Big[A_t \ln \pi^i_\theta(a^i_t|\tau^i_t)\Big], \quad A_t := Q^{\boldsymbol{\pi}}_{\phi}\big(s_t, \boldsymbol{a}_t\big) - V^{\boldsymbol{\pi}}(s_t)$$

$$= -\mathbb{E}_{\pi_\theta}\Big[A_t \ln \boldsymbol{\pi}_\theta(\boldsymbol{a}_t|\boldsymbol{\tau}_t)\Big],$$ centralized = sum of independent

StarCraft 2 results of COMA from Förster et al. (2018)

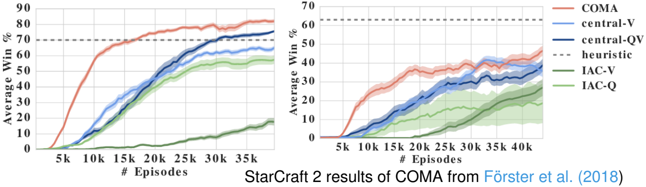
Counterfactual multi-agent learning (COMA)

- Stochastic policy-gradients in cooperative games
  - centralized training with parameter sharing
  - centralized value $Q^{\boldsymbol{\pi}}_{\phi}(s, \boldsymbol{a})$, decentr. policy $\boldsymbol{\pi}_\theta(\boldsymbol{a}|\boldsymbol{\tau}_t) = \prod_{i=1}^{N} \pi^i_\theta(a^i|\tau^i_t)$

$$\mathcal{L}^{\text{COMA}}_{\pi[\theta]} \;:=\; -\sum_{i=1}^{N} \mathbb{E}_{\pi_\theta}\Big[A^i_t \ln \pi^i_\theta(a^i_t|\tau^i_t)\Big], \qquad A_t \;:=\; Q^{\boldsymbol{\pi}}_{\phi}\big(s_t, \boldsymbol{a}_t\big) - V^{\boldsymbol{\pi}}(s_t)$$

- Same baseline $V^{\boldsymbol{\pi}}(s_t)$ for all joint actions $\boldsymbol{a}_t$ has high variance
  - different counterfactual baseline in advantage A^i_t for each $\boldsymbol{a}^{-i}_t$

$$A^i_t \;:=\; Q^{\boldsymbol{\pi}}_{\phi}(s_t, \boldsymbol{a}_t) - \sum_{a'^i \in \mathcal{A}^i} \pi^i_\theta(a'^i|\tau^i_t)\, Q^{\boldsymbol{\pi}}_{\phi}\big(s_t, \{a'^i\} \cup \boldsymbol{a}^{-i}_t\big)$$



StarCraft 2 results of COMA from Förster et al. (2018)

10.3 Summary

- During training we might have centralized information
- MADDPG extends DDPG with centralized Q-values
- COMA extends AC additionally with low-variance bias

Learning Objectives

LO10.5: Explain centralized training and decentralized execution
LO10.6: Explain how MADDPG and COMA exploit centralized training

10.3 Next lecture

- Next lecture: **advanced MARL**!
- Remember assignment sheet 4 (and exercise sheet 4)!
- Questions? Ask them here: answers.ewi.tudelft.nl

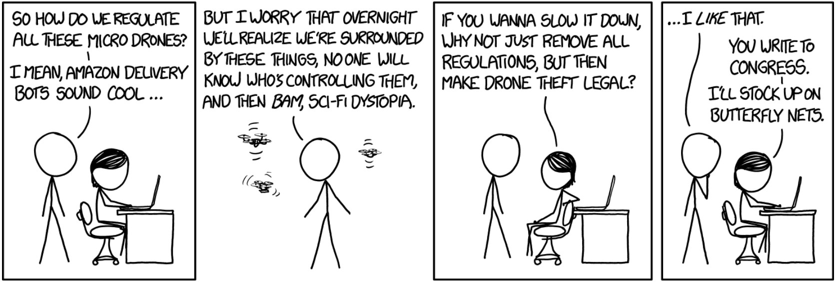


image source: xkcd.com

References I

Jakob Förster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 2974–2982. AAAI Press, 2018. URL https://arxiv.org/abs/1705.08926.

Ryan Lowe, YI WU, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems 30*, pages 6379–6390. 2017. URL https://arxiv.org/pdf/1706.02275.pdf.

Frans A. Oliehoek and Christopher Amato. *A concise introduction to decentralized POMDPs*. Springer Publishing Company, Incorporated, 1st edition, 2016. ISBN 3319289276, 9783319289274. URL https://www.fransoliehoek.net/docs/OliehoekAmato16book.pdf.

T.E.S. Raghavan. Chapter 20 zero-sum two-person games. volume 2 of *Handbook of Game Theory with Economic Applications*, pages 735–768. Elsevier, 1994. doi: https://doi.org/10.1016/S1574-0005(05)80052-9. URL https://www.sciencedirect.com/science/article/pii/S1574000505800529.

Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philiph H. S. Torr, Jakob Förster, and Shimon Whiteson. The StarCraft Multi-Agent Challenge. *CoRR*, abs/1902.04043, 2019.

Christian Schröder de Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip H. S. Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the starcraft multi-agent challenge?, 2020. URL https://arxiv.org/abs/2011.09533.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, pages 484–489, 2016. doi: 10.1038/nature16961.

References II

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550:354–359, October 2017. URL http://dx.doi.org/10.1038/nature24270.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018. URL https://science.sciencemag.org/content/362/6419/1140.

Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pages 330–337, 1993.

Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575:350–354, 2019.