# Background

## Markov Games

* https://scholar.google.co.uk/scholar?q=Markov+games+as+a+framework+for+multi-agent+reinforcement+learning&hl=zh-CN&as\_sdt=0&as\_vis=1&oi=scholart

## RL and DRL

* V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. Nature, 518(7540):529–533, 2015.
  + DRL
* Mastering the game of Go with deep neural networks and tree search.
  + DRL playing game of GO

# Related work

## General MARL

* Multiagent bidirectionally coordinated nets for learning to play starcraft combat games.
  + Multiagent game playing

Independent Q Learning

* Independent reinforcement learners in cooperative Markov games: a survey regarding coordination problems
  + Independent Q-Learning agents, shown in paper that they don’t perform well in multi-agent settings.
  + Policy of each agent changes during training, causing non-stationarity in the perspective of a single agent, preventing naïve approach of experience replay
    - <http://proceedings.mlr.press/v70/foerster17b/foerster17b.pdf>
      * introduces nonstationarity that makes it incompati-  
        ble with the experience replay memory on which  
        deep Q-learning relies
  + Violates Markov assumption for convergence of Q Learning
* <https://proceedings.neurips.cc/paper/2003/hash/e71e5cd119bbc5797164fb0cd7fd94a4-Abstract.html>
  + Attempts in inputting other agents’ policy parameters to Q function to overcome such non-stationarity
* <https://arxiv.org/abs/1702.08887>
  + using a multi-agent variant of importance sampling to naturally decay obsolete data
  + and conditioning each agent's value function on a fingerprint that disambiguates the age of the data sampled from the replay memory
    - Essentially indexing the experiences

Cooperative settings

* <https://ieeexplore.ieee.org/abstract/document/4399095>
  + Hysteric Q Learning

Policy Gradient

* General Policy Gradient
  + <https://proceedings.neurips.cc/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html>
  + Known to exhibit high variance gradient estimations
    - More so in multiagent context as agent’s reward depends on action of many agents, therefore reward conditioned only on agent’s own actions exhibits higher variance
* Actor Critic
  + Rather similar to GAN
  + Temporal Difference version of policy gradient
  + Contains an actor and critic
    - Actor decides which action should be taken
    - Critic inform the actor how good was the action and how it should adjust
  + Actor learns through policy gradient approach, critics evaluate the action produced by actor by computing the value function
* Extend policy gradient framework to Deterministic Policy Gradient (DPG) algorithms
  + Deep DPG is a variant of DPG where the policy and critic are approximated with deep neural networks
    - <https://arxiv.org/pdf/1509.02971.pdf?source=post_page--------------------------->
    - Also make use of experience replay as in DQN to stabilize the neural network
* <https://ojs.aaai.org/index.php/AAAI/article/view/11794> (COMA)
  + Counterfactual multi-agent policy gradients, uses centralised critic to estimate Q function and decentralised actors to optimise agents’ policies
  + Address the challenges of multi-agent credit assignment, uses a counterfactual baseline that marginalised out a single agent’s action while keeping the other agents’ actions fixed
  + Learns a single centralized critic for all agents
* <https://arxiv.org/abs/1706.02275> (MADDPG)
  + Actor critic policy gradient where the critic is augmented with extra information about the policies of other agents, while the actor only has access to local information. After training is completed, only the local actors are used at execution phase
  + Since the centralized critic function explicitly uses the decision-making policies of other agents, we additionally show that agents can learn approximate models of other agents online and effectively use them in their own policy learning procedure.
  + acting in a decentralized manner and equally applicable in cooperative and competitive settings.
  + learn a centralized critic for each agent, allowing for agents with differing reward functions including competitive scenarios
  + consider environments with explicit communication between agents
  + learns continuous policies

Learning grounded cooperative communication protocols between agents

* <https://proceedings.neurips.cc/paper/2016/file/55b1927fdafef39c48e5b73b5d61ea60-Paper.pdf>
* <https://proceedings.neurips.cc/paper/2016/file/c7635bfd99248a2cdef8249ef7bfbef4-Paper.pdf>
* usually only applicable when the communication between agents is carried out over a dedicated, differentiable communication channel.

## Communication

Communication treated as differentiable process (continuous) optimized through backprop

* Tend to converge quickly to higher-quality policies compared to traditional RL framework
* <https://proceedings.neurips.cc/paper/2016/file/c7635bfd99248a2cdef8249ef7bfbef4-Paper.pdf>
  + Centralised training, decentralised execution
  + DIAL approach, allows real valued messages to pass between agents during centralised learning allowing gradients to be pushed through the communication channel, and allows real-valued messages to pass between agents during centralised learning, thereby treating communication actions as bottleneck connections between agents. As a result, gradients can be pushed through the communication channel, yielding a system that is end-to-end trainable even across agents. During decentralised execution, real-valued messages are discretised and mapped to the discrete set of communication actions allowed by the task. Because DIAL passes gradients from agent to agent, it is an inherently deep learning approach.
* <https://arxiv.org/pdf/1605.07736.pdf>
  + CommNet
  + Learns a shared Deep Neural Net that is shared across agents

Message treated as an extension to action space

* Communication behaviour learned and optimized via standard reinforcement learning
* <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5045926>
  + Make use of tabular Q Learning to solve the predator-prey task with communication
  + Similarly, in <https://scholar.google.co.uk/scholar?q=Efficient+Distributed+Reinforcement+Learning+through+Agreement,&hl=zh-CN&as_sdt=0&as_vis=1&oi=scholart>
* <https://proceedings.neurips.cc/paper/2016/file/c7635bfd99248a2cdef8249ef7bfbef4-Paper.pdf>
  + Same paper as DIAL
  + In case that communicated messages are discrete values and gradient are not able to be calculated
  + RIAL method makes use of deep Q Learning for better scalability to learn content of the message
* <https://arxiv.org/pdf/1602.02672.pdf>
  + DDRQN
  + Discover communication protocols to solve multiagent learning problems based on well known riddles (partially observable tasks)
  + Not really solving coordination problems
* <https://link.springer.com/chapter/10.1007/978-3-540-45173-0_29>
  + Genetic algorithm to learn the languages for the predator agents in the predator-prey problem
  + Not scalable for larger problems
* <https://ojs.aaai.org/index.php/AAAI/article/view/6205>
  + Communication Learning via Backpropagation in Discrete Channels with Unknown Noise
    - Stochastic message encoding/decoding procedure that makes a discrete communication channel mathematically equivalent to an analog channel with additive noise
    - Then which gradients can be backpropagated

Surveys

* <http://proceedings.mlr.press/v70/omidshafiei17a/omidshafiei17a.pdf>
* <https://d1wqtxts1xzle7.cloudfront.net/50476164/A_Comprehensive_Survey_of_Multiagent_Rei20161122-3056-9gsdvn-with-cover-page-v2.pdf?Expires=1668710668&Signature=Ggqn3hV3pNw~mFUk8cniapX0P1yQlKCjb~iUtFd21YA-e6QLomogR5htK9vGefiYzFkx6pCytRyaoquO2klRiGFQbnjjwPjmt~JVuUr8u9FMOQJ~bDvY7ZhaL~fFfjibXnWPqMsYyQbBd1jzVU3YKNEjPqXE1n6UUJj9MAg6ZGKCvzpy838TDJJLnLJrFSMt7ptzM3w51u2XEx9FmfzK~nXXjtzGt-VguNQ4MrQyJa2x1Lzot2HBYYHKzC7fxXLlEizHovHtwA39~2twVP3bYL-qK0eYibF5vHqq5jLXhPUP0GHom~ZqKI5AoiSzijAmV3873Olbo3OFSAUdg0O-xg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA>