**9.27.2018 Paper reading**

1. More Robust Doubly Robust Off-policy Evaluation, Mehrdad Farajtabar, Yinlam Chow, Mohammad Ghavamzadeh, ICML 2018.

Overview: In many real-world decision-making problems (healthcare, marketing,…) deploying a policy without an accurate estimate of its performance is unethical. This paper proposes a variation of DR with a DM loss derived from minimizing variance to estimate the policy..

* Problem
  + Estimate the performance of a policy from the data generated by another policy(ies). For example, infering the causal effect of a new treatment from historical data.
* Related work
  + **Direct method (DM)**: learn a model of the system to estimate the policy. (low variance high bias )
  + **Importance sampling (IS)**: correct the mismatch between distributions of evaluation and behavior policies.(low bias high variance)
  + **Doubly Robust (DR)**: combine DM and IS
* Method
  + Doubly robust estimators: learn a DM estimator by minimizing the variance of DR.

1. Automatic Goal Generation for Reinforcement Learning Agents, Carlos Florensa, David Held, Xinyang Geng, Pieter Abbeel, ICML 2018.

Overview: RL is powerful in solving a single task. It does not scale well in diverse tasks. For example, moving objects to varying locations. This paper uses a generator network to propose a set of tasks for the agent to accomplish, where the generator is optimized using adversarial training.

* Method
  + Goal GAN: Generator: trained to generate goals that is at the appropriate level of difficulty for the current policy. Discriminator is trained to evaluate whether a goal is at this criteria.

1. On Reinforcement Learning for Full-length Game of StarCraft, Zhen-Jia Pang, Ruo-Ze Liu, Zhou-Yu Meng, Yi Zhang, Yang Yu , Tong Lu, arXiv 2018.

Overview: The main difficulties of StarCraft II are large state space and action space and long-time horizon. This paper proposes a HRL algorithm to address these difficulties. One is macro-action extracted from expert’s trajectories (large action space). The other is a two-layer hierarchical architecture to decompose a long time horizon into several sub-horizon (sub-question) and solving each sub-question in turn.

1. Latent Space Policies for Hierarchical Reinforcement Learning, Tuomas Haarnoja, Kristian Hartikainen, Pieter Abbeel, Sergey Levine, ICML 2018

Overview: Prior HRL methods explicitly cripple lower layers of a hierarchy to force them to use higher-level modulating signals. This paper imagines a hierarchical framework in which each layer directly attempts to solve the task and, if it is not fully successful, makes the job easier for the layer above it.