

CSC2621 Topics in Robotics

Reinforcement Learning in Robotics

Week 2: Supervised & Imitation Learning

Instructor: Animesh Garg

TA: Dylan Turpin & Tingwu Wang

VARIANCE REDUCTION FOR POLICY GRADIENT WITH ACTION-DEPENDENT FACTORIZED BASELINES

Wu, Rajeswaran, Duan, Kumar, Bayen, Kakade, Mordatch, Abbeel

Topic: Policy Gradients

Presenter: Animesh Garg


Motivation

- Policy Gradient:
Improve the log-prob of actions weighted by expected future returns
- The gradient estimator has **high variance!**
 - Credit assignment to actions (especially in long horizon tasks)
 - High Dimensional action spaces
- Variance of estimator matters because of stability and convergence properties.

Key Insights

- Use a baseline! But, what does a baseline do?
Intuition: Removes the effect of future actions from total reward!

$$A = Q(s, a) - V(s) \quad \begin{array}{l} \text{State-dependent} \\ \text{Baseline} \end{array}$$

- But do we have to be limited to state-only?
 If/when the **individual actions** produced by the policy can be **decomposed** into multiple factors, we can incorporate this additional information into the baseline to **further reduce variance**.
Information about the other factors can provide a better evaluation of how well a specific factor performs

Key Insights

When do actions decouple?

- Different Actions dimensions act of independent components of Observation Space.
- Action space composed of multiple independent Function Approximators (no weight sharing)
- Multivariate Gaussian policies with a diagonal covariance
- Multi-agent & Distributed RL
Centralized Learning + Decentralized Execution

Contributions

- An action-dependent baseline enables using additional signals beyond the state to achieve bias-free variance reduction.
- Derive an Optimal Action-Dependent Baseline
- Analysis of improvement in variance reduction
- Empirical results to show the effects of the proposal baselines and comparison of several choices of baselines

Background

- Value-Function Based Gradient through Q(or V)
- Actor-Critic Gradient through critic

Low-Variance
Biased (often)
Sample Efficient
Can be unstable

- Policy Gradients Gradient through rollouts

High-Variance
Unbiased
Less Sample Efficient
More Stable

Background: Variance Reduction

- MDP Objective

$$\eta(\pi_\theta) = \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]$$

- Score Function Estimator (Williams 1992)

$$\begin{aligned} \nabla_\theta \mathbb{E}_x [f(x)] &= \nabla_\theta \int p_\theta(x) f(x) dx = \int p_\theta(x) \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)} f(x) dx \\ &= \int p_\theta(x) \nabla_\theta \log p_\theta(x) f(x) dx = \mathbb{E}_x [\nabla_\theta \log p_\theta(x) f(x)] \end{aligned}$$

Background: Variance Reduction

- Policy Gradient

$$\nabla_{\theta} \eta(\pi_{\theta}) = \mathbb{E}_{\tau} \left[\sum_{t=0}^{\infty} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \sum_{t'=t}^{\infty} \gamma^{t'-t} r_{t'} \right].$$

$$\nabla_{\theta} \eta(\pi_{\theta}) = \mathbb{E}_{\rho_{\pi}, \pi} \left[\nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \hat{Q}(s_t, a_t) \right]$$

- State-Dependent Baseline

$$\nabla_{\theta} \eta(\pi_{\theta}) = \mathbb{E}_{\rho_{\pi}, \pi} \left[\nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \left(\hat{Q}(s_t, a_t) - b(s_t) \right) \right]$$

$$\mathbb{E}_{a_t} [\nabla_{\theta} \log \pi_{\theta}(a_t | s_t) b(s_t)] = \nabla_{\theta} \mathbb{E}_{a_t} [b(s_t)] = 0$$

Action-Dependent Baseline

- Assume m-dimensional action space

$$\pi_{\theta}(a_t|s_t) = \prod_{i=1}^m \pi_{\theta}(a_t^i|s_t)$$

- So the MDP objective becomes

$$\nabla_{\theta} \eta(\pi_{\theta}) = \mathbb{E}_{\rho_{\pi}, \pi} \left[\nabla_{\theta} \log \pi_{\theta}(a_t|s_t) \hat{Q}(s_t, a_t) \right] = \mathbb{E}_{\rho_{\pi}, \pi} \left[\sum_{i=1}^m \nabla_{\theta} \log \pi_{\theta}(a_t^i|s_t) \hat{Q}(s_t, a_t) \right]$$

- Baseline is independent of other action components, hence:

$$\mathbb{E}_{a_t} \left[\nabla_{\theta} \log \pi_{\theta}(a_t^i|s_t) b_i(s_t, a_t^{-i}) \right] = \mathbb{E}_{a_t^{-i}} \left[\nabla_{\theta} \mathbb{E}_{a_t^i} \left[b_i(s_t, a_t^{-i}) \right] \right] = 0$$

Action-Dependent Baseline

- New Gradient Estimator

$$\nabla_{\theta} \eta(\pi_{\theta}) = \mathbb{E}_{\rho_{\pi}, \pi} \left[\sum_{i=1}^m \nabla_{\theta} \log \pi_{\theta}(a_t^i | s_t) \left(\hat{Q}(s_t, a_t) - b_i(s_t, a_t^{-i}) \right) \right]$$

- Notice that this is similar to Advantage Function

$$\hat{A}_i(s_t, a_t) = Q(s_t, a_t) - b_i(s_t, a_t^{-i})$$

Optimal State-dependent Baseline

- Reformulate the objective

$$\nabla_{\theta} \eta(\pi_{\theta}) := \mathbb{E}_{\rho_{\pi}, \pi} \left[\nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \left(\hat{Q}(s_t, a_t) - b(s_t) \right) \right]$$

$$g := \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \left(\hat{Q}(s_t, a_t) - b(s_t) \right), \quad a_t \sim \pi_{\theta}(a_t | s_t), s_t \sim \rho_{\pi}(s_t)$$

- Optimal State-dep Baseline

$$\frac{\partial}{\partial b} [\text{Var}(g)] = 0$$

$$b^*(s_t) = \frac{\mathbb{E}_{\rho_{\pi}, \pi} \left[\nabla_{\theta} \log \pi_{\theta}(a_t | s_t)^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \hat{Q}(s_t, a_t) \right]}{\mathbb{E}_{\rho_{\pi}, \pi} \left[\nabla_{\theta} \log \pi_{\theta}(a_t | s_t)^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right]}$$

Optimal Action-Dependent Baseline

- Reformulate the objective

$$\begin{aligned}\nabla \eta_i(\pi_\theta) &:= \mathbb{E}_{\rho_{\pi, \pi}} \left[\nabla_\theta \log \pi_\theta(a_t^i | s_t) \left(\hat{Q}(s_t, a_t) - b_i(s_t, a_t^{-i}) \right) \right] \\ z_i &:= \nabla_\theta \log \pi_\theta(a_t^i | s_t)\end{aligned}$$

- Independent action dimensions

$$\nabla_\theta \log \pi_\theta(a_t^i | s_t)^T \nabla_\theta \log \pi_\theta(a_t^j | s_t) \equiv z_i^T z_j = 0, \quad \forall i \neq j$$

- Optimal Action Baseline

$$b_i^*(s_t, a_t^{-i}) = \frac{\mathbb{E}_{a_t^i} \left[\nabla_\theta \log \pi_\theta(a_t^i | s_t)^T \nabla_\theta \log \pi_\theta(a_t^i | s_t) \hat{Q}(s_t, a_t) \right]}{\mathbb{E}_{a_t^i} \left[\nabla_\theta \log \pi_\theta(a_t^i | s_t)^T \nabla_\theta \log \pi_\theta(a_t^i | s_t) \right]}.$$

How are they related

- Action-dep. Baseline doesn't degenerate to State-dep. Baseline

$$Z_i := Z_i(s_t, a_t^{-i}) = \mathbb{E}_{a_t^i} [\nabla_{\theta} \log \pi_{\theta}(a_t^i | s_t)^T \nabla_{\theta} \log \pi_{\theta}(a_t^i | s_t)]$$

$$Y_i := Y_i(s_t, a_t^{-i}) = \mathbb{E}_{a_t^i} [\nabla_{\theta} \log \pi_{\theta}(a_t^i | s_t)^T \nabla_{\theta} \log \pi_{\theta}(a_t^i | s_t) \hat{Q}(s_t, a_t)]$$

$$I_{b=b^*(s)} = \sum_i \mathbb{E}_{\rho_{\pi, a_t^{-i}}} \left[\frac{1}{Z_i} \left(\frac{Z_i}{\sum_j Z_j} \sum_j Y_j - Y_i \right)^2 \right]$$

weighted sum of the deviation of the per-component score-weighted marginalized Q from the component weight (based on score only, not Q) of the overall aggregated marginalized Q values

- The difference is particularly large when the Q function is highly sensitive to the actions, esp. along directions that influence the gradient the most.

Potential Choices of Action-Dep Baselines

- Marginalized Q-Baseline $b = \mathbb{E}_{a_t^i} [\hat{Q}(a_t, s_t)]$

nearly optimal if $Corr(\log \pi, Q) \approx 0$

- Monte Carlo marginalized Q baseline

$$b_i(s_t, a_t^{-i}) = \frac{1}{M} \sum_{j=0}^M Q_{\pi_\theta}(s_t, (a_t^{-i}, \alpha_j))$$

- Mean marginalized Q baseline

$$b_i(s_t, a_t^{-i}) = Q_{\pi_\theta}(s_t, (a_t^{-i}, \bar{a}_t^i)) \quad \bar{a}_t^i = \mathbb{E}_{\pi_\theta} [a_t^i]$$

Experiments



Variance Reduction for Policy
Gradient with Action-Dependent
Factorized Baselines

Experiments: Does it help v/s b(s)

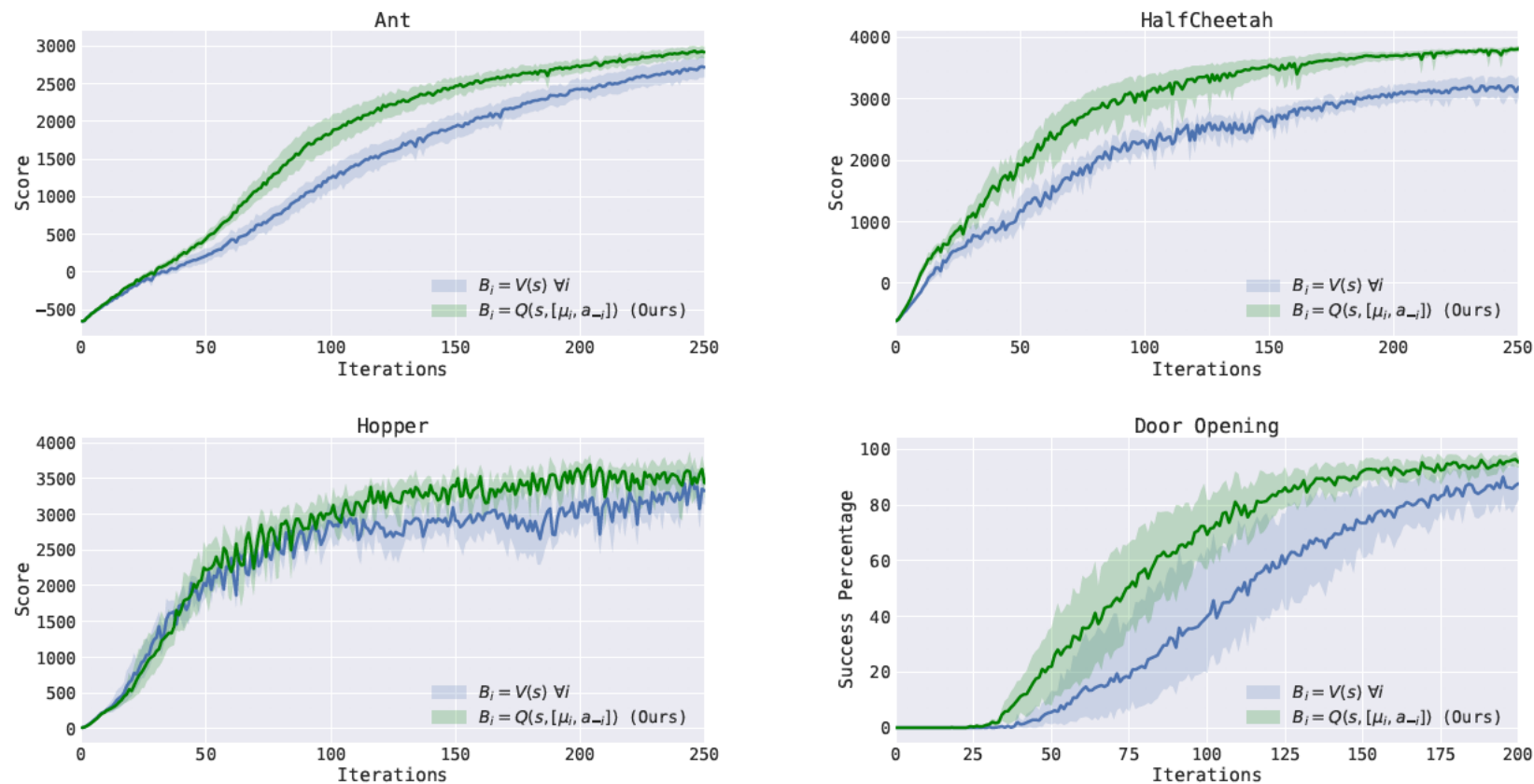
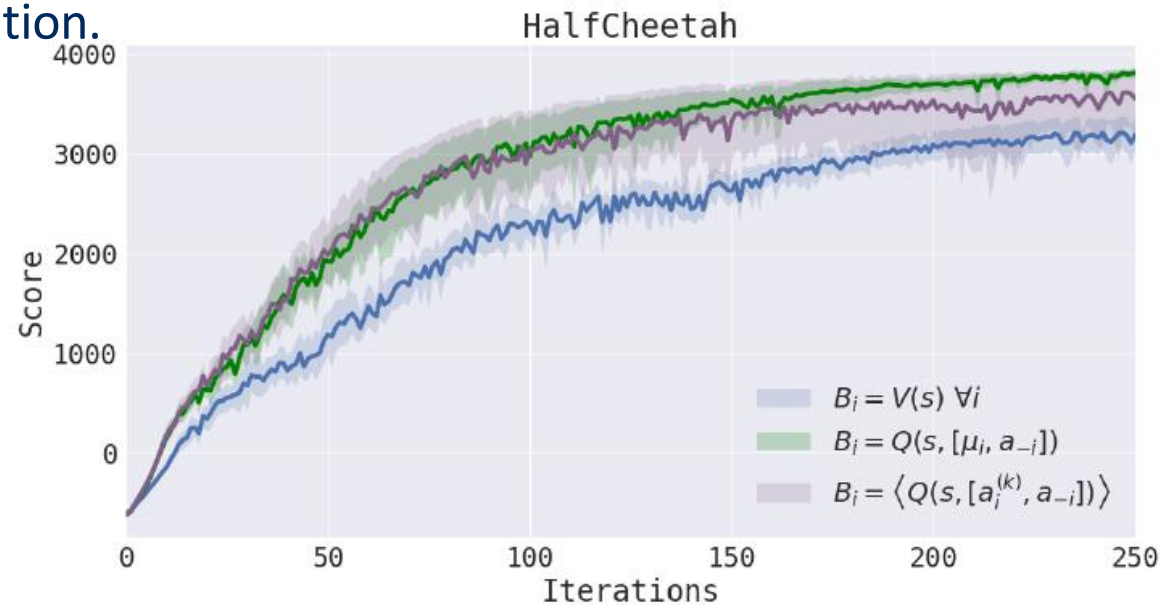


Figure 1: Comparison between value function baseline and action-conditioned baseline on various continuous control tasks. Action-dependent baseline performs consistently better across all the tasks.

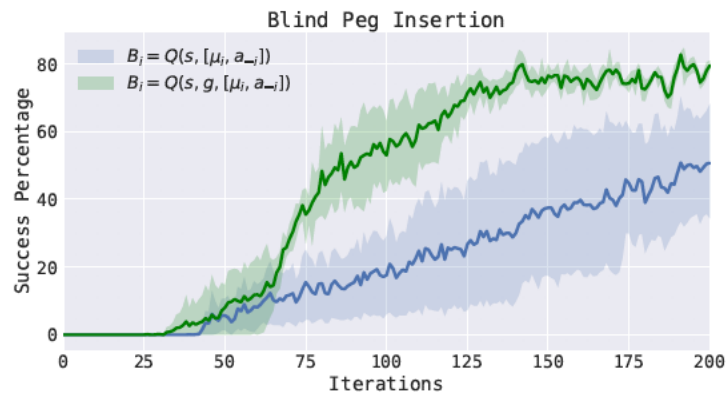
Experiments: Different Baseline Choices

- Variants of the action-dependent baseline that use: (i) sampling from the Q-function to estimate the conditional expectation; (ii) Using the mean action to form a linear approximation to the conditional expectation.

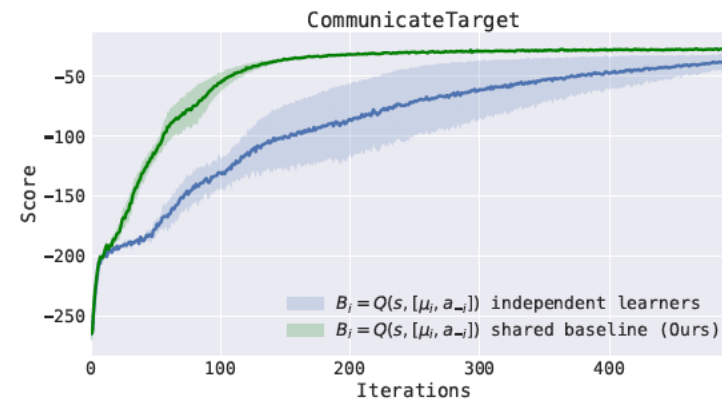


- Both variants are comparable with the latter being more comp. efficient

Experiments: Effect of global information



(a) Success percentage on the blind peg insertion task. The policy still acts on the observations and does not know the hole location. However, the baseline has access to this goal information, in addition to the observations and action, and helps to speed up the learning. By comparison, in blue, the baseline has access only to the observations and actions.



(b) Training curve for multi-agent communication task with two agents. Two policies are simultaneously trained, one for each agent. Each policy acts on the observations of its respective agent only. However, the shared baseline has access to the other agent's state and action, in addition to its own state and action, and results in considerably faster training. By comparison, in blue, the independent learners baseline has access to only a single agent's state and action.

Contributions (Recap)

- An action-dependent baseline enables using additional signals beyond the state to achieve bias-free variance reduction.
- Derive an Optimal Action-Dependent Baseline
- Analysis of improvement in variance reduction
- Empirical results to show the effects of the proposal baselines and comparison of several choices of baselines

Is this it?

- Some people don't believe this works at all!

The Mirage of Action-Dependent Baselines in Reinforcement Learning

George Tucker¹ **Surya Bhupatiraju**^{1,2} **Shixiang Gu**^{1,3,4} **Richard E. Turner**³ **Zoubin Ghahramani**^{3,5}
Sergey Levine^{1,6}

References

- Variance Reduction for Policy Gradient with Action-Dependent Factorized Baselines
<https://arxiv.org/pdf/1803.07246.pdf>
- Q-Prop: Sample-Efficient Policy Gradient with An Off-Policy Critic
<https://arxiv.org/abs/1611.02247>
- The Mirage of Action-Dependent Baselines in Reinforcement Learning
<https://arxiv.org/pdf/1802.10031.pdf>
- Trajectory-wise Control Variates for Variance Reduction in Policy Gradient Methods
<https://arxiv.org/abs/1908.03263>

Presentations

Jan 28

- Need 4 students
- Presentation Review Thurs/Fri (sign up)

Feb 5

- Need a minimum of 4 students
- Presentation Review Tues Jan 28 and Wed Jan 29

Projects

Jan 28

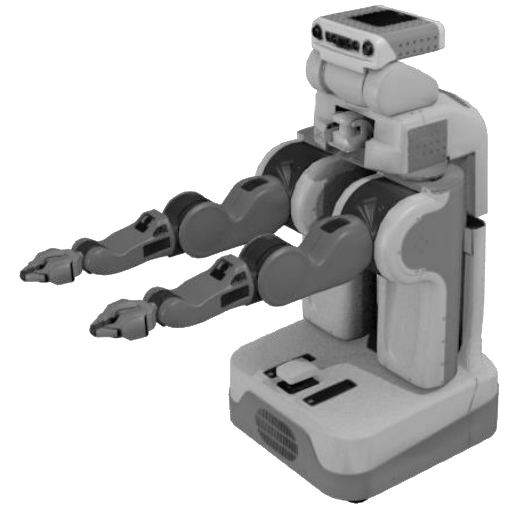
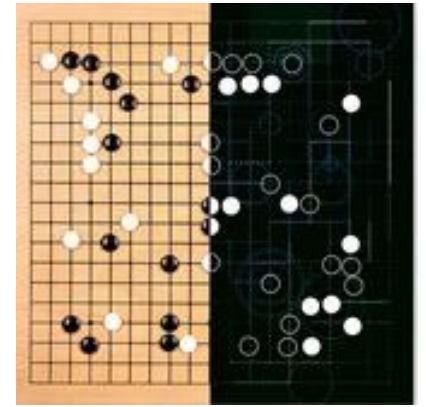
- Proposal: **Due**
- 2 pages
- Latex Template provided. (CoRL)
- Should include introduction with motivation

Intuition & description of the aimed contribution.

- If empirical, state broad experimental plan and condition of success
- If analysis, then state what property you are analysing.

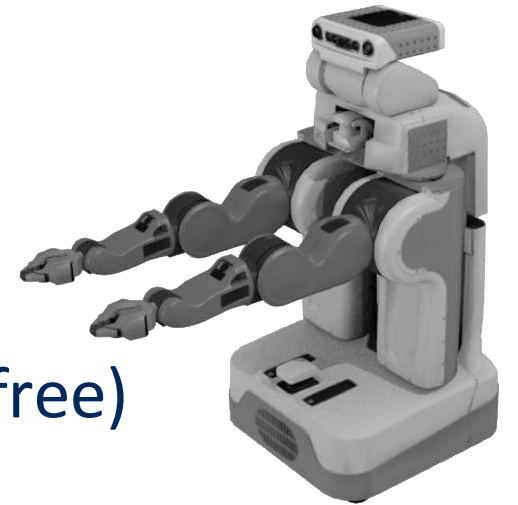
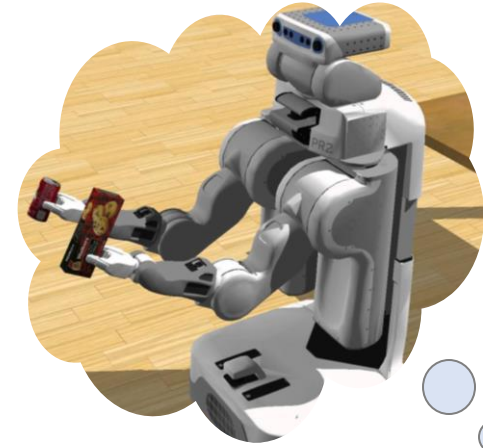
Going from Go to Robot/Control

- Known Environment vs Unstructured/Open World
- Need for Behavior Transfer
- Discrete vs Continuous States-Actions
- Single vs Variable Goals
- Reward Oracle vs Reward Inference



Other Open Problems

- Single algorithm for multiple tasks
- Learn new tasks very quickly
- Reuse past information about related problems
- Reward modelling in open environment
- How and what to build a model of?
- How much to rely on the model vs direct reflex (model-free)
- Learn without interaction if seen a lot of data



What this course plans to cover

- Imitation Learning: Supervised
- Policy Gradient Algorithms
- Actor-Critic Methods
- Value Based Methods
- Distributional RL
- Model-Based Methods
- Imitation Learning: Inverse RL
- Exploration Methods
- Bayesian RL
- Hierarchical RL