ENGINEERING
TEXAS A&M UNIVERSITY

Inverse Reinforcement Learning Applied to Guidance and Control of Small Unmanned Aerial Vehicles

Akshay Sarvesh, Kishan P. Badrinath, Vinicius G. Goecks

ECEN 689-602, Dr. Dileep Kalathil

December 5, 2018

Overview

- Introduction
  - Robotics task based on human demonstration
- IRL, and why IRL?
- Why Previous Approaches Fail
- Guided Cost Learning (GCL)
  - Study case: Inverted Pendulum
- Generative Adversarial Imitation Learning (GAIL)
  - Study case: Lunar Lander Continuous
  - Study case: Unmanned Aerial Vehicle
- Future work
- Conclusions

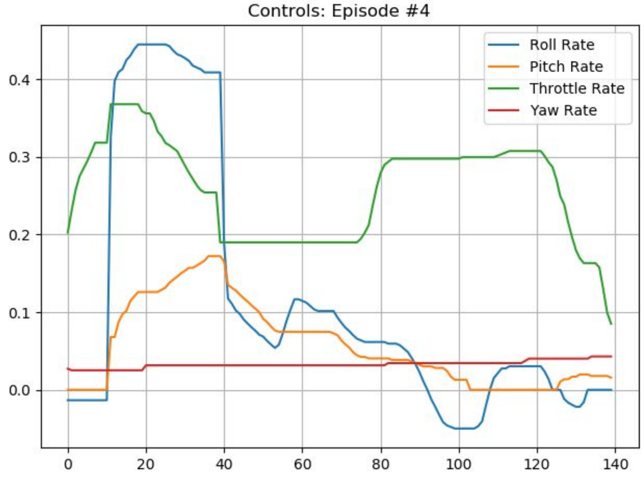
Introduction

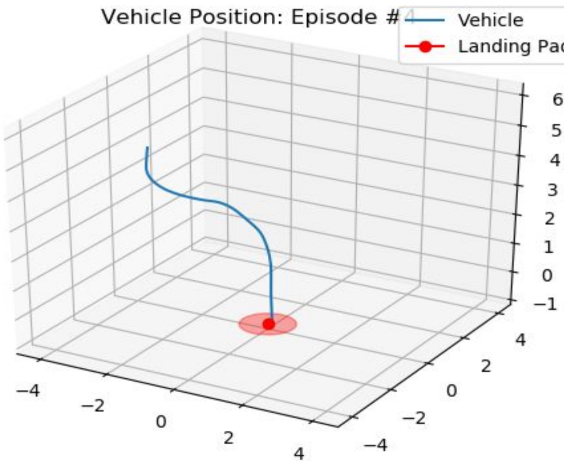
- **Goal:**
  - Train a policy initially based on human demonstration which is able to match or surpass human performance.

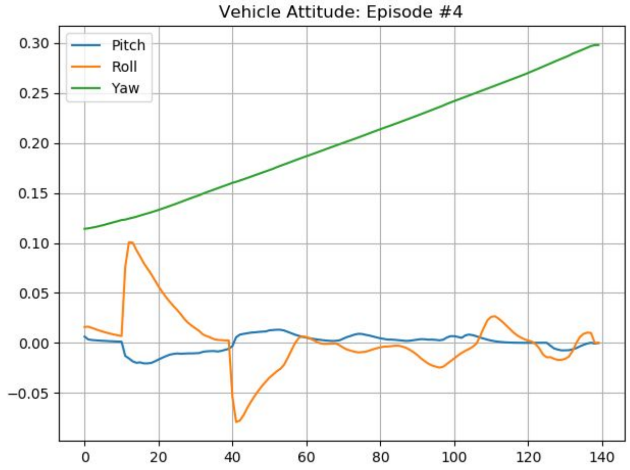
- **Applications:**
  - Robotics applications where machines are currently being teleoperated or entirely controlled by humans.
  - UAV Landing task:
    - Continuous states and actions
    - Unknown dynamics
    - Safety is critical

Introduction

- Example of UAV data:







Overview

- Introduction
  - Robotics task based on human demonstration
- IRL, and why IRL?
- Why Previous Approaches Fail
- Guided Cost Learning (GCL)
  - Study case: Inverted Pendulum
- Generative Adversarial Imitation Learning (GAIL)
  - Study case: Lunar Lander Continuous
  - Study case: Unmanned Aerial Vehicle
- Future work
- Conclusions

IRL and Why IRL?

ENGINEERING
TEXAS A&M UNIVERSITY

- Determine the optimal reward function with:
  - Measurements (both inputs and outputs) of expert's behaviour
    - Should adapt even if expert is suboptimal
  - Model of environment
    - Unknown dynamics

- Why?
  - Handcrafted reward functions in RL might not be the best
    - Eg: Driving in Weather Conditions, driving in complex environments, etc
  - Moreover, RL is built upon deriving a policy from a reward function. So mimicking expert behaviour is better.

Overview

- Introduction
  - Robotics task based on human demonstration
- IRL, and why IRL?
- Previous Approaches and why they fail?
- Guided Cost Learning (GCL)
  - Study case: Inverted Pendulum
- Generative Adversarial Imitation Learning (GAIL)
  - Study case: Lunar Lander Continuous
  - Study case: Unmanned Aerial Vehicle
- Future work
- Conclusions

Why Previous Approaches Fail

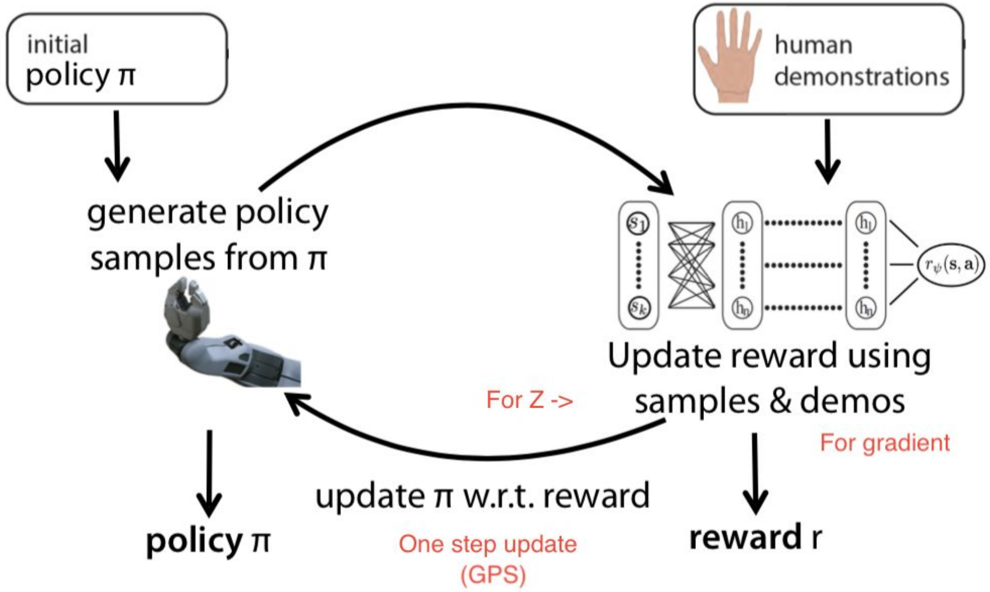
- (Abbeel and Ng ’04)
  - Non-linear space of reward’s exploration
  - Ill-posed problem to find exact reward function
- (Ziebart et al. ’08)
  - Takes care of above challenges
  - Unknown dynamics in the MDP
  - Solving MDP in each loop

Overview

- Introduction
  - Robotics task based on human demonstration
- IRL, and why IRL?
- Previous Approaches and why they fail?
- Guided Cost Learning (GCL)
  - Study case: Inverted Pendulum
- Generative Adversarial Imitation Learning (GAIL)
  - Study case: Lunar Lander Continuous
  - Study case: Unmanned Aerial Vehicle
- Future work
- Conclusions

Guided Cost Learning (GCL) (Finn et al. ICML '16)

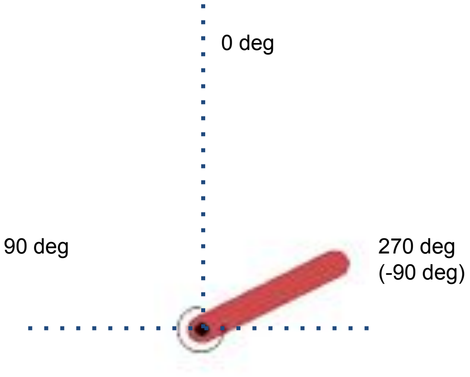
- Handling unknown dynamics - Using sample distribution to estimate Z
- Avoid solving MDP in inner loop - One step (lazy) update of the policy



Slides taken from Deep RL Bootcamp at UCB

Inverted Pendulum (GCL)

- Average Return using GCL for different number “n” of expert trajectories



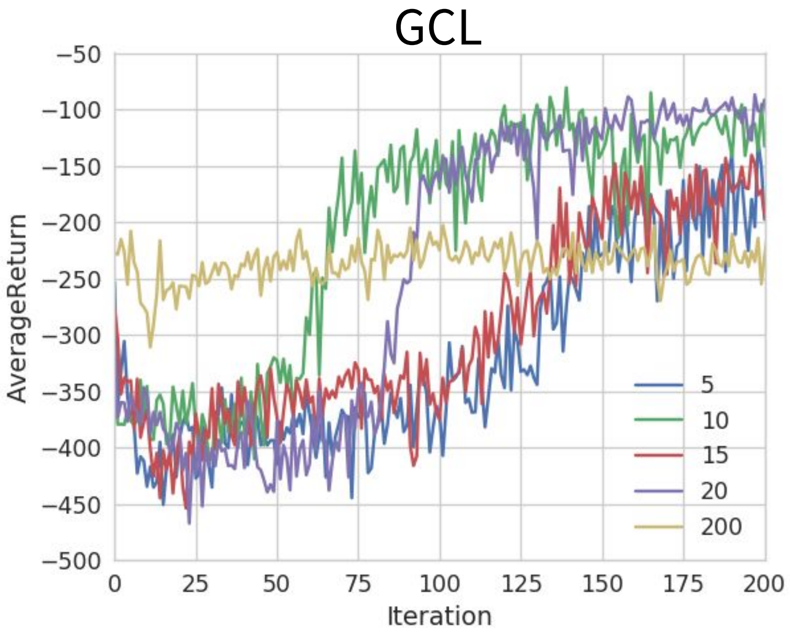
States (continuous, dim = 3)**:**

cos(θ), sin(θ), θ_dot

Actions (continuous, dim = 1)**:**

torque on joint

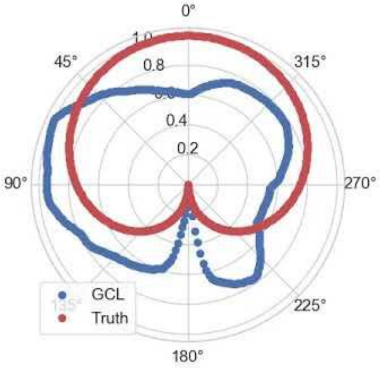
Reward(Cost): minimize θ, θ_dot, and torque



Guided Cost Learning

- Reward model for “n” = 10, evaluated for 1000 iterations (theta_dot = 0, torque = 0)

GCL Policy
(200 iterations)



GCL Reward/Energy Model
(normalized)

Overview

- Introduction
  - Robotics task based on human demonstration
- IRL, and why IRL?
- Previous Approaches and why they fail?
- Guided Cost Learning (GCL)
  - Study case: Inverted Pendulum
- Generative Adversarial Imitation Learning (GAIL)
  - Study case: Lunar Lander Continuous
  - Study case: Unmanned Aerial Vehicle
- Future work
- Conclusions

Generative Adversarial Imitation Learning (GAIL)

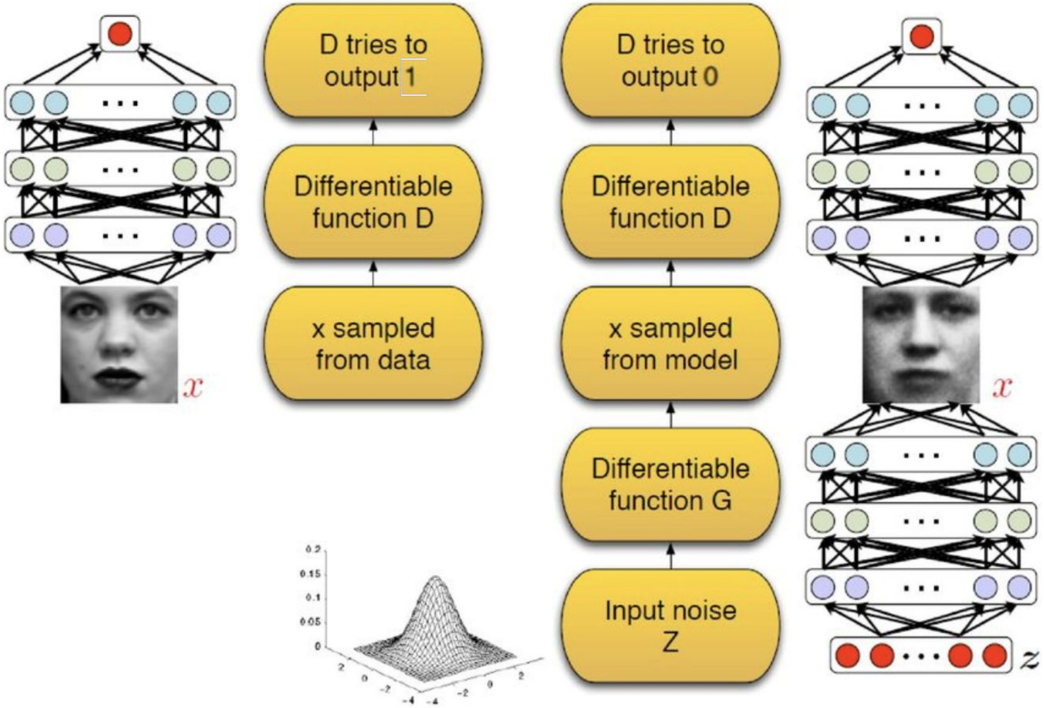


Figure from Goodfellow et al, 2014

Generative Adversarial Imitation Learning (GAIL)

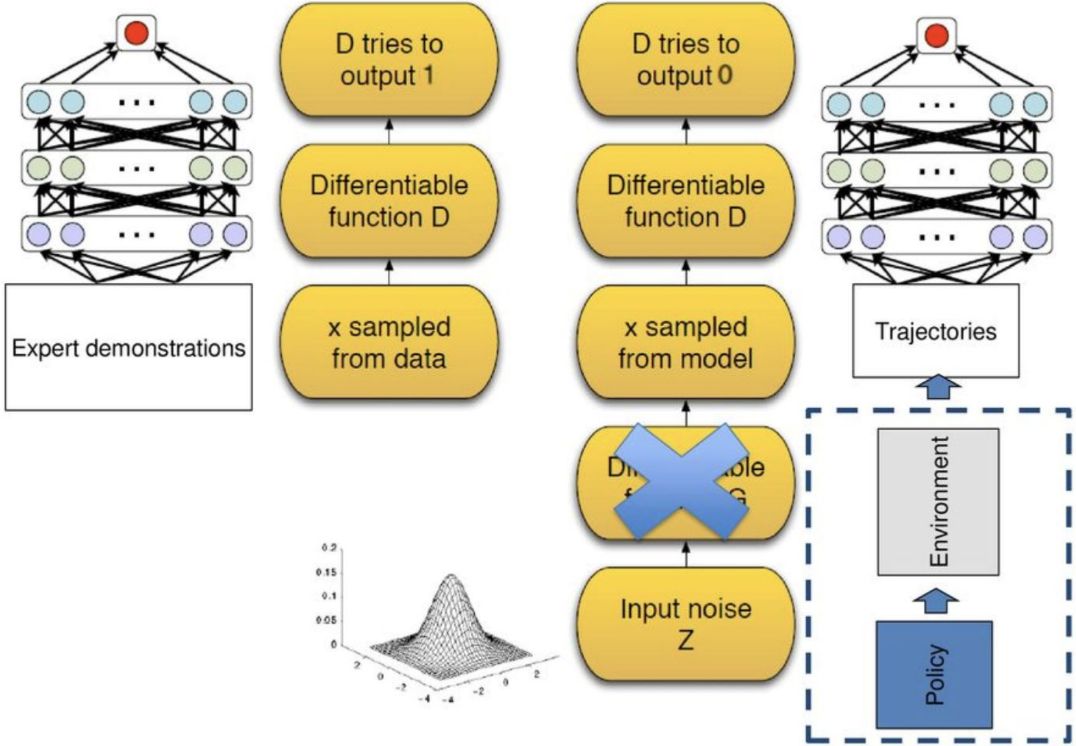


Figure from Ho et al, 2016

GCL ~ GAIL

Connection to Generative Adversarial Networks

(Goodfellow et al. '14)

Inverse RL			GANs
trajectory τ	⟷	**sample x**	
policy π~q(τ)	⟷	**generator** G	
reward r	⟶	**discriminator** D	

data distribution p

Reward/discriminator optimization:

GCL:

$$D^*(\tau) = \frac{p(\tau)}{p(\tau) + q(\tau)}$$

$$D_\psi(\tau) = \frac{\frac{1}{Z}\exp(R_\psi)}{\frac{1}{Z}\exp(R_\psi) + q(\tau)}$$

GAIL:

$D_\psi(\tau) =$ some classifier

Both:

$$\mathcal{L}_{\text{discriminator}}(\psi) = \mathbb{E}_{\tau\sim p}[-\log D_\psi(\tau)] + \mathbb{E}_{\tau\sim q}[-\log(1 - D_\psi(\tau))]$$

(Finn*, Christiano*, et al. '16)

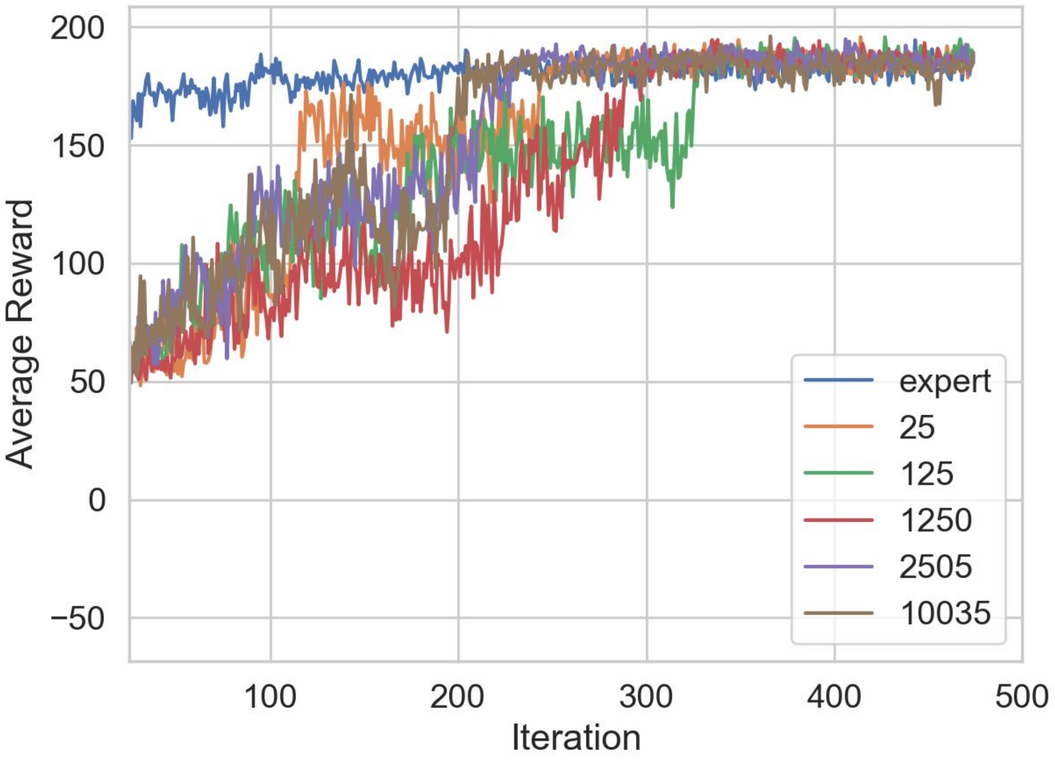
Slides taken from Deep RL Bootcamp at UCB

Overview

- Introduction
  - Robotics task based on human demonstration
- IRL, and why IRL?
- Previous Approaches and why they fail?
- Guided Cost Learning (GCL)
  - Study case: Inverted Pendulum
- Generative Adversarial Imitation Learning (GAIL)
  - Study case: Lunar Lander Continuous
  - Study case: Unmanned Aerial Vehicle
- Future work
- Conclusions

Lunar Lander Continuous (GAIL)

States (discrete + continuous, dim = 8)**:**
position, velocities, leg contact
Actions (continuous, dim = 2)**:**
main and side engines
Reward: landing position and fuel usage



Take away 1: GAIL is sample efficient in terms of number of expert trajectories.

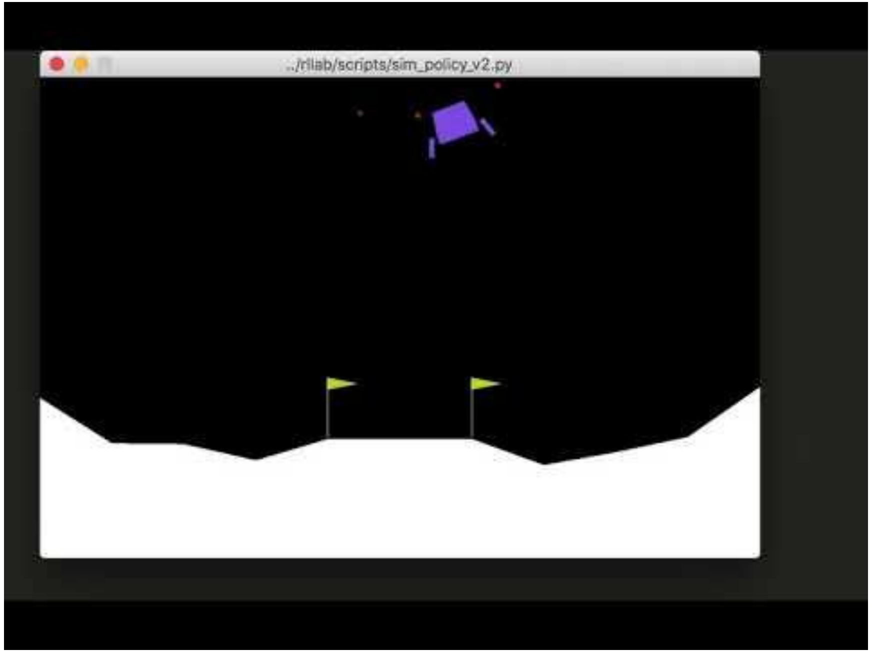
Lunar Lander Continuous (GAIL)

ENGINEERING
TEXAS A&M UNIVERSITY

TRPO Hyperparameters	
Hyperparameter	**Expert**
Policy NN Architecture	20, 20
Batch Size	7000
Max Path Length	200
Discount	0.99

GAIL Hyperparameters					
No. of expert trajectories / **Hyperparameter**	**25**	**125**	**1250**	**2505**	**10035**
Policy NN Architecture	20, 20	30, 30	50, 50, 10	100, 100, 20	200, 200, 20
Batch Size	4000	7000	7000	7000	7000
Max Path Length	200	200	200	200	200
Discriminator Train Iters	100	100	100	100	100
Discount	0.99	0.99	0.99	0.99	0.99

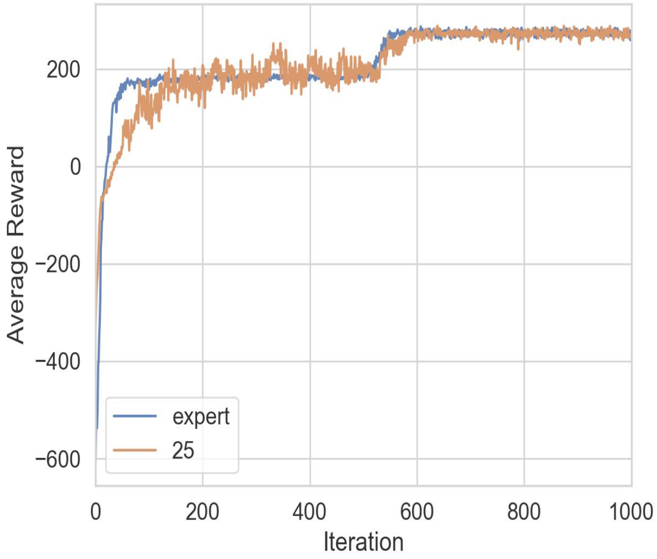
Lunar Lander Continuous (GAIL)

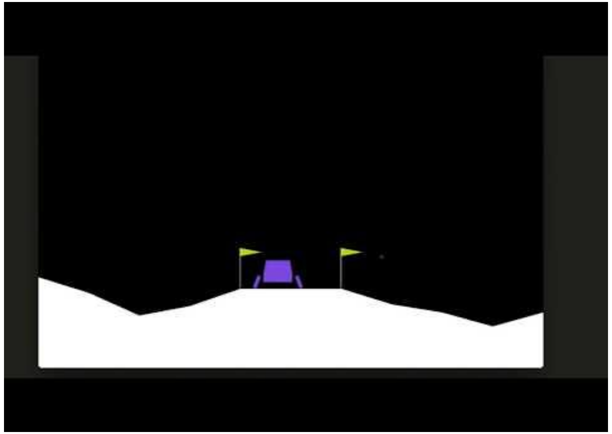


Expert Trajectories

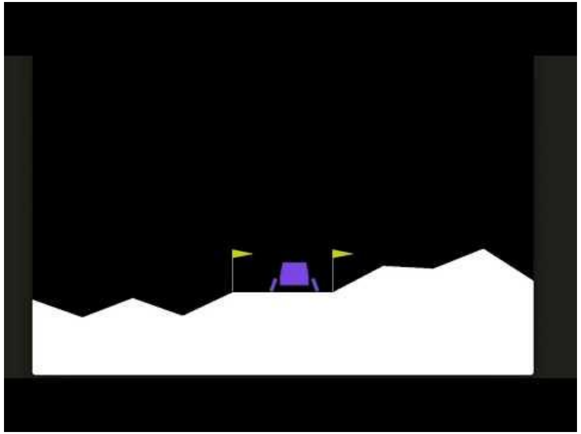
GAIL Performance
(Initial and after 50 and 300 iterations)

Lunar Lander Continuous (GAIL)





Expert Trajectories



GAIL Performance
(after 600 iterations)

Take away 2: But not sample efficient in terms of environment interactions.

Overview

ENGINEERING
TEXAS A&M UNIVERSITY

- Introduction
  - Robotics task based on human demonstration
- IRL, and why IRL?
- Previous Approaches and why they fail?
- Guided Cost Learning (GCL)
  - Study case: Inverted Pendulum
- Generative Adversarial Imitation Learning (GAIL)
  - Study case: Lunar Lander Continuous
  - Study case: Unmanned Aerial Vehicle
- Future work
- Conclusions

Microsoft AirSim: Unmanned Aerial Vehicles Simulation

States (continuous, dim = 15):

positions, velocities, visual features

Actions (continuous, dim = 2):

roll and pitch rates

Reward: distance to landing pad

Data Collection - AirSim GAIL

Human (expert) Collecting Trajectories

Sample of Expert Trajectories

Unmanned Aerial Vehicles (GAIL)

Challenges:

- Training agent in real clock-time, as if it was in hardware.
- Extended training time delays hyperparameter tuning.
- Complex dynamics, observation and action-space.

Unmanned Aerial Vehicles (GAIL)

Additional Challenge:

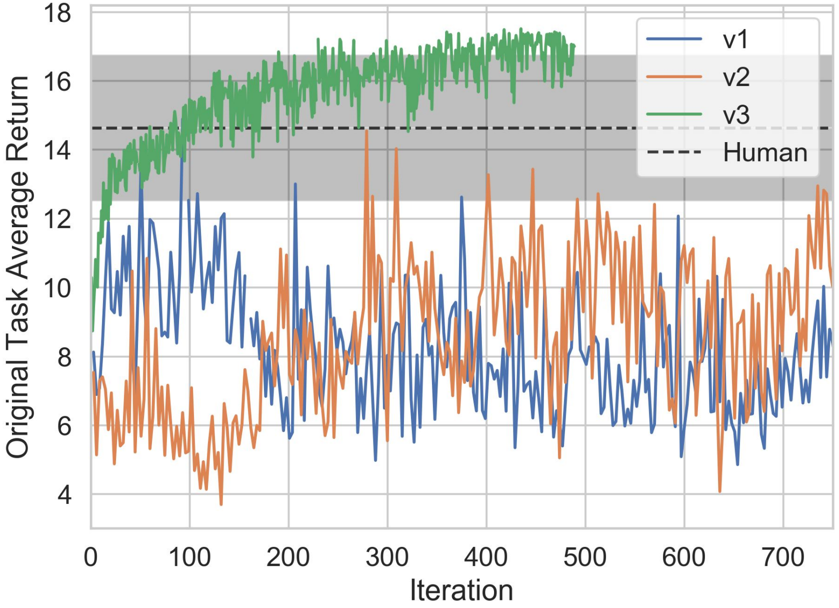
- Narrow region of convergence, which sometimes can be overlooked if only checks convergence based on reward values.

Unmanned Aerial Vehicles (GAIL)

AirSim GAIL - Untrained

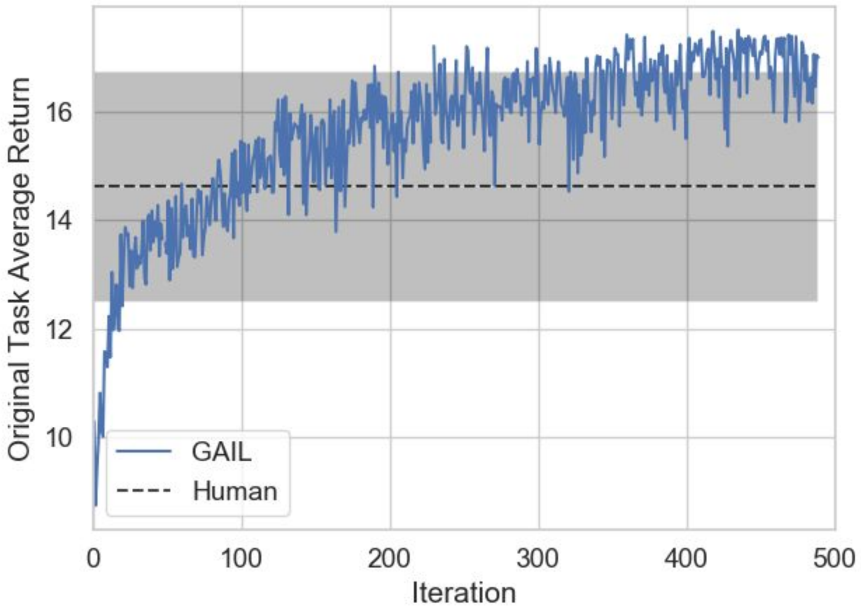
AirSim GAIL - Trained

Unmanned Aerial Vehicles (GAIL)



GAIL Hyperparameters			
Hyperparameter	**v1**	**v2**	**v3**
Policy NN Architecture	50, 50, 10	32, 32	32, 32
Batch Size	60	60	1200
Max Path Length	60	60	60
Discriminator Train Iters	100	100	100
Discount	0.99	0.99	0.99

Unmanned Aerial Vehicles (GAIL)



Final Training Numbers	
Training Time	48.12 hours
Trajectories Generated	10429
GAIL Iterations	489 iterations
Time/GAIL Iterations	~6 minutes
Time/Trajectory (average)	16.61 sec
Human Trajectories	100
Human Time	27.68 min (0.46 hours)

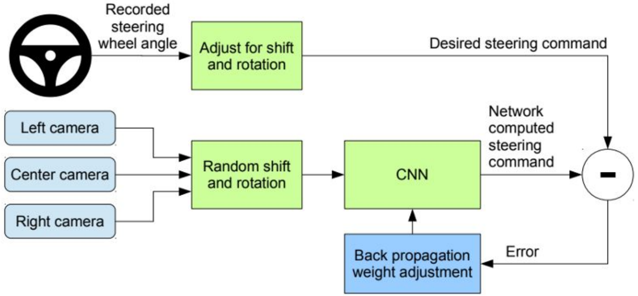
Overview

ENGINEERING
TEXAS A&M UNIVERSITY

- Introduction
  - Robotics task based on human demonstration
- IRL, and why IRL?
- Previous Approaches and why they fail?
- Guided Cost Learning (GCL)
  - Study case: Inverted Pendulum
- Generative Adversarial Imitation Learning (GAIL)
  - Study case: Lunar Lander Continuous
  - Study case: Unmanned Aerial Vehicle
- Future work
- Conclusions

Step 1 : Behavioural Cloning result

- The 1st 3 layers are convolutional Neural Networks. (Kernel-3)
- The convolutional layers are followed by 3 fully connected layers(100 and 50 neurons).



Future work

ENGINEERING
TEXAS A&M UNIVERSITY

Turn GAIL into a mixed-initiative approach (that is, interact with expert in the iterations)

Overview

- Introduction
  - Robotics task based on human demonstration
- IRL, and why IRL?
- Previous Approaches and why they fail?
- Guided Cost Learning (GCL)
  - Study case: Inverted Pendulum
- Generative Adversarial Imitation Learning (GAIL)
  - Study case: Lunar Lander Continuous
  - Study case: Unmanned Aerial Vehicle
- Future work
- Conclusions

Conclusions

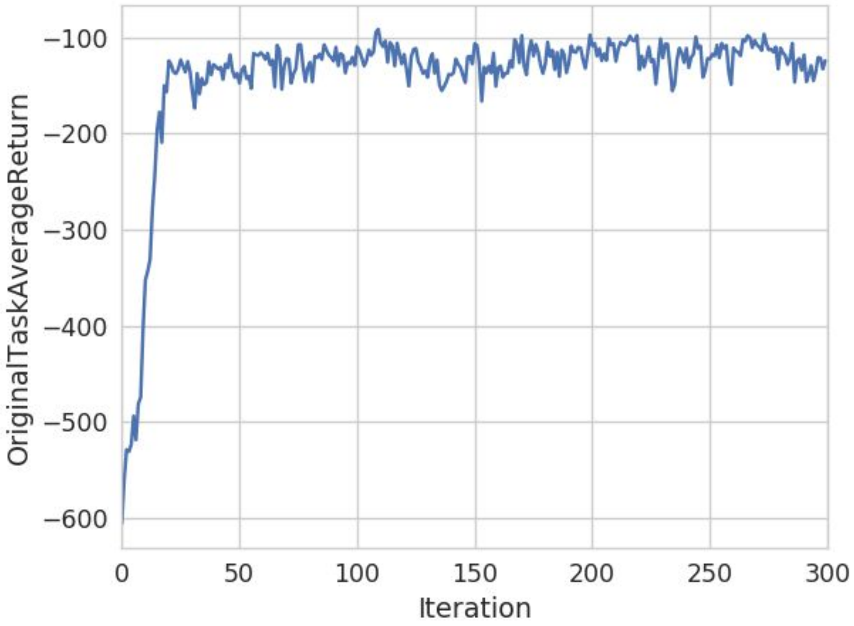
ENGINEERING
TEXAS A&M UNIVERSITY

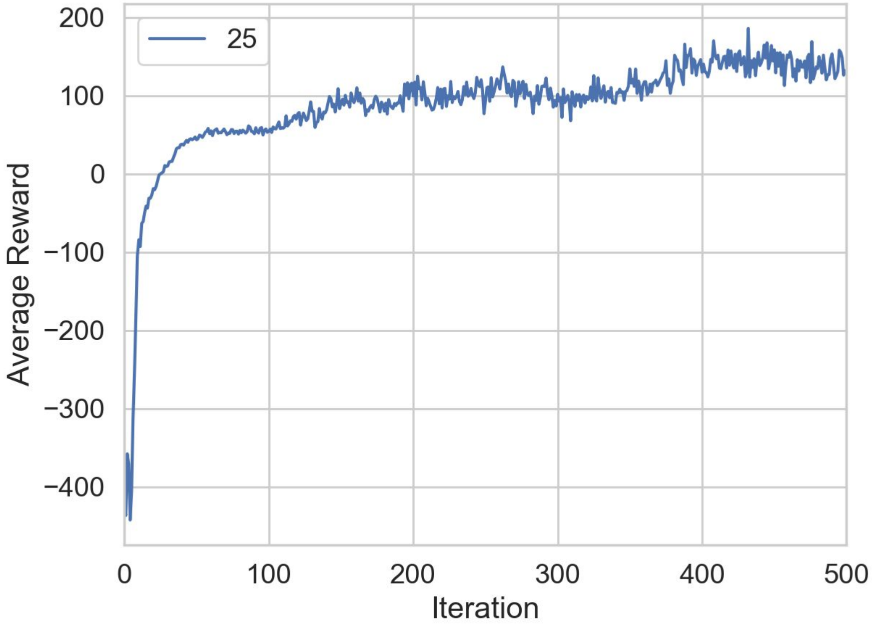
- **Inverse Reinforcement Learning (IRL)** is a viable alternative to learn behavior from expert demonstration.
- For **high-dimensional continuous observation and action-space** environments, limit the application of more traditional IRL approaches.
- For these complex cases, we have demonstrated the application of **Guided Cost Learning (GCL)** and **Generative Adversarial Imitation Learning (GAIL)** solving:
  - Inverted Pendulum after about 125 iterations,
  - Lunar Lander Continuous after about 160 iterations, and
  - Landing an UAV in a high-fidelity simulated environment and consistently **surpassing mean human performance** after about 100 iterations.

ENGINEERING
TEXAS A&M UNIVERSITY

Backup - GCL with Lunar Lander

Need to adjust hyperparameter





Backup - Guided Policy Search (Levine & Abbeel '14)

$$\min_{p(\tau)} \sum_{t=1}^{T} E_{p(\mathbf{x}_t,\mathbf{u}_t)}[c(\mathbf{x}_t,\mathbf{u}_t) + \mathbf{u}_t^T\lambda_t + \rho_t D_{KL}(p(\mathbf{u}_t|\mathbf{x}_t)\|\pi_\theta(\mathbf{u}_t|\mathbf{x}_t))]$$

$$\text{s.t. } D_{KL}(p(\tau)\|\bar{p}(\tau)) \le \epsilon$$

$$\mathcal{L}(p,\eta) = \sum_{t=1}^{T} E_{p(\mathbf{x}_t,\mathbf{u}_t)}[c(\mathbf{x}_t,\mathbf{u}_t)+\mathbf{u}_t^T\lambda_t+\rho_t D_{KL}(p(\mathbf{u}_t|\mathbf{x}_t)\|\pi_\theta(\mathbf{u}_t|\mathbf{x}_t))+\eta D_{KL}(p(\mathbf{u}_t|\mathbf{x}_t\|\bar{p}(\mathbf{u}_t,\mathbf{x}_t))]$$

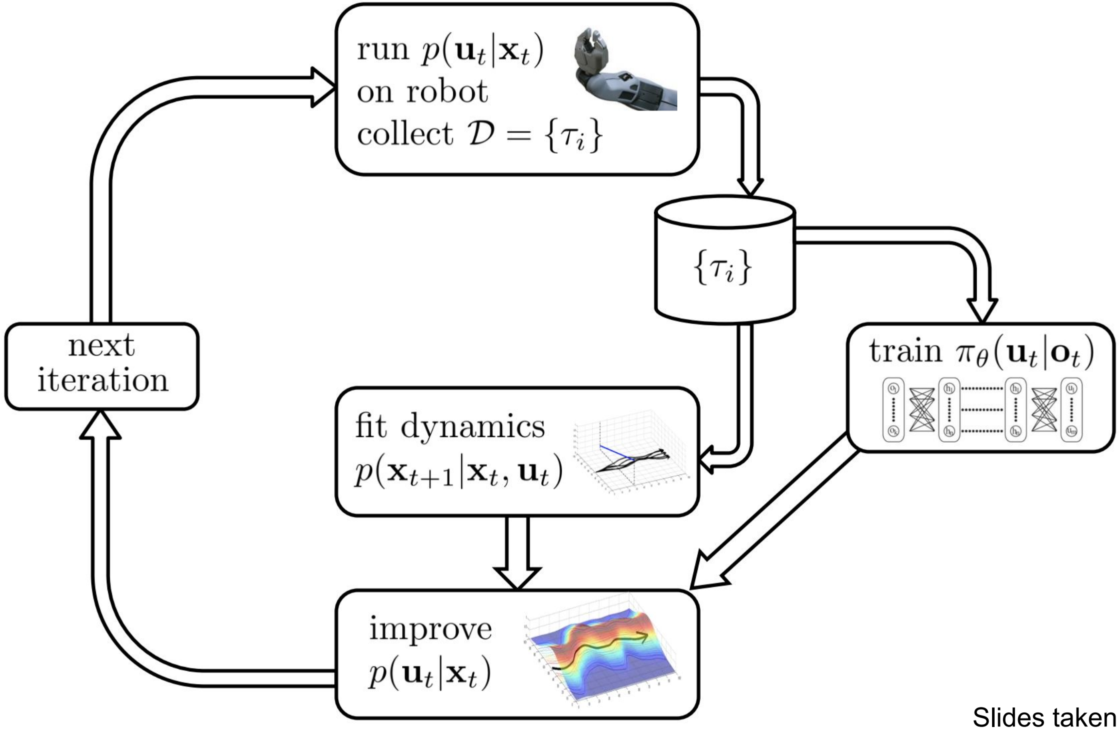
$$\mathcal{L}(p,\eta) = \sum_{t=1}^{T} E_{p(\mathbf{x}_t,\mathbf{u}_t)}[\underbrace{c(\mathbf{x}_t,\mathbf{u}_t)+\mathbf{u}_t^T\lambda_t-\rho_t \log \pi_\theta(\mathbf{u}_t|\mathbf{x}_t)-\eta\bar{p}(\mathbf{u}_t,\mathbf{x}_t)}_{\tilde{c}(\mathbf{x}_t,\mathbf{u}_t)}]-(\rho_t+\eta)\mathcal{H}(p(\mathbf{u}_t|\mathbf{x}_t))$$

$$\mathcal{L}(p,\eta) = \sum_{t=1}^{T} E_{p(\mathbf{x}_t,\mathbf{u}_t)}[\tilde{c}(\mathbf{x}_t,\mathbf{u}_t)] - \nu_t\mathcal{H}(p(\mathbf{u}_t|\mathbf{x}_t))$$

maximum entropy objective

Slides taken from Prof. Abbeel's website

Backup - Guided Policy Search (Levine & Abbeel '14)



Slides taken from Prof. Abbeel's website

Backup - GCL == GAIL

Connection to Generative Adversarial Networks

(Goodfellow et al. '14)

	trajectory τ	⟷	**sample x**	
Inverse RL	**policy** π~q(τ)	⟷	**generator** G	GANs
	reward r	⟶	**discriminator** D	

data distribution p

Policy/generator optimization:

$$\mathcal{L}_{\text{generator}}(\theta) = \mathbb{E}_{\tau \sim q}[\log(1 - D_\psi(\tau)) - \log D_\psi(\tau)]$$
$$= \mathbb{E}_{\tau \sim q}[\log q(\tau) + \log Z - R_\psi(\tau)]$$

entropy-regularized RL

Unknown dynamics: train generator/policy with RL

Baram et al. ICML '17: use learned dynamics model to backdrop through discriminator

(Finn*, Christiano*, et al. '16)

Slides taken from Deep RL Bootcamp at UCB