Generative Adversarial Imitation Learning

A summary of Ho & Ermon (NIPS 2016)

Giuseppe Gabriele Russo

Università di Pisa

June 4, 2025

1. Motivation & Problem Statement

Goal (offline imitation learning):

Given only expert trajectories, learn a policy π that reproduces expert behaviour *without* knowing the reward.

Behavioral Cloning

- Supervised mapping $s \to a$.
- Covariate shift: errors snowball when the agent enters unseen states.

Inverse Reinforcement Learning

- Infer reward, then run RL as an inner loop.
- Heavy computation & poor sample efficiency; struggles with high-dimensional control.

Desired solution:

Direct policy learning that scales to high-dimensional continuous

2. Core Idea: Occupancy-Measure Matching

Match occupancy measures $\rho_\pi(s,a)$ vs. $\rho_E(s,a)$

ρ_π: discounted visitation frequency under policy π

ρ_E: visitation frequency of the expert

Objective: bring ρ_π close to $\rho_E \Rightarrow$ indistinguishable behavior trajectories

Why occupancy and not raw actions?

Matching the entire state–action distribution prevents covariate shift and captures long–term effects of actions.

Minimax formulation

$$\min_\pi \max_D \ \mathbb{E}_{\rho_E}[\log D(s,a)] + \mathbb{E}_{\rho_\pi}[\log(1-D(s,a))] - \lambda H(\pi)$$

3. Key Objective Equation

Definition

Maximum-Entropy GAN Objective

$$\min_{\pi} \; D_{\mathrm{JS}}\big(\rho_\pi \,\|\, \rho_E\big) \;-\; \lambda\, H(\pi)$$

- D_{JS} — Jensen–Shannon divergence induced by the discriminator (zero when $\rho_\pi = \rho_E$).
- $H(\pi) = \mathbb{E}_\pi[\,-\log \pi(a \mid s)\,]$ — causal entropy bonus encourages exploration and smooth policies.
- Optimal value is reached when the agent's occupancy measure matches the expert's: *perfect imitation*.

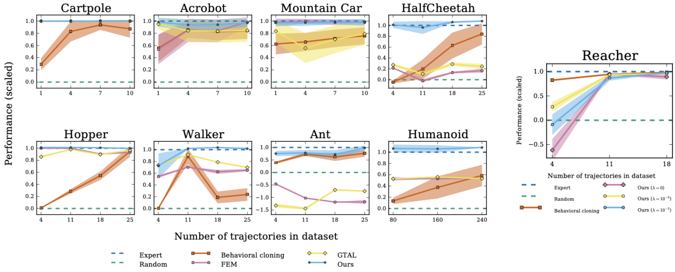
4. GAIL Algorithm

Algorithm 1 Generative Adversarial Imitation Learning

```
Initialise policy π_θ and discriminator D_w
while not converged do
    Collect trajectories τ ~ π_θ
    Update D_w by maximising:
  log D_w(s, a) for (s, a) ~ τ_E & log(1 − D_w(s, a)) for (s, a) ~ τ
    Update π_θ with TRPO on cost:
  c(s, a) = − log D_w(s, a) + λ H(π_θ)
end while
return π_θ
```

5. Empirical Benchmarks

Benchmarks: 9 MuJoCo control tasks (CartPole → Humanoid)



(a) (b)

Figure 1: (a) Performance of learned policies. The y-axis is negative cost, scaled so that the expert achieves 1 and a random policy achieves 0. (b) Causal entropy regularization λ on Reacher.

6. Key Findings

- GAIL attains 70–100% of expert return with $\sim$25 demonstration trajectories.
- Outperforms Behavioral Cloning, FEM, and GTAL on every task.
- Handles 376-dimensional HUMANOID while remaining sample-efficient.

7. Strengths & Limitations

Novelty: bridges IRL and GAN via occupancy-measure matching

Key Strengths

- Direct policy learning - no reward inference.
- Scales to high-dimensional continuous control.
- Requires fewer demos than Behavioral Cloning.

Main Limitations

- Still sample-inefficient vs. model-based methods.
- TRPO/PPO updates are computationally heavy.
- No online expert feedback.

8. Essential References

- J. Ho & S. Ermon, "Generative Adversarial Imitation Learning," in *NeurIPS 29*, 2016.
- P. Abbeel & A. Ng, "Apprenticeship Learning via Inverse Reinforcement Learning," *ICML*, 2004.
- I. Goodfellow *et al.*, "Generative Adversarial Nets," *NeurIPS*, 2014.
- J. Schulman *et al.*, "Trust Region Policy Optimization," *ICML*, 2015.