HG-DAgger: Interactive Imitation Learning with Human Experts (2019)

Authors: Michael Kelly, Chelsea Sidrane,
Katherine Driggs-Campbell, Mykel J. Kochenderfer

Presenter: Adeet Parikh

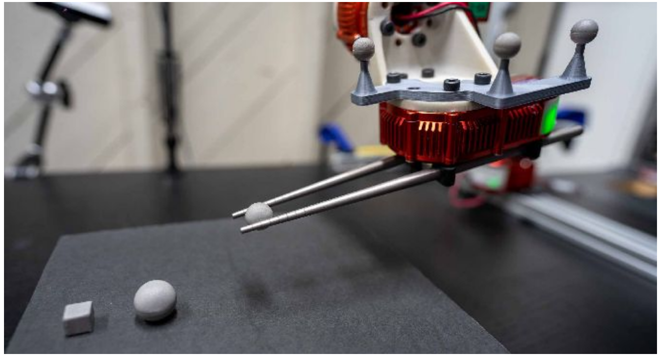
October 18th, 2022

Imitation Learning

Goal is still to solve Markov Decision Process

Why imitation learning?

- Many tasks have sparse reward functions, and designing a fine-grained one can be difficult
- Safer in real world environments



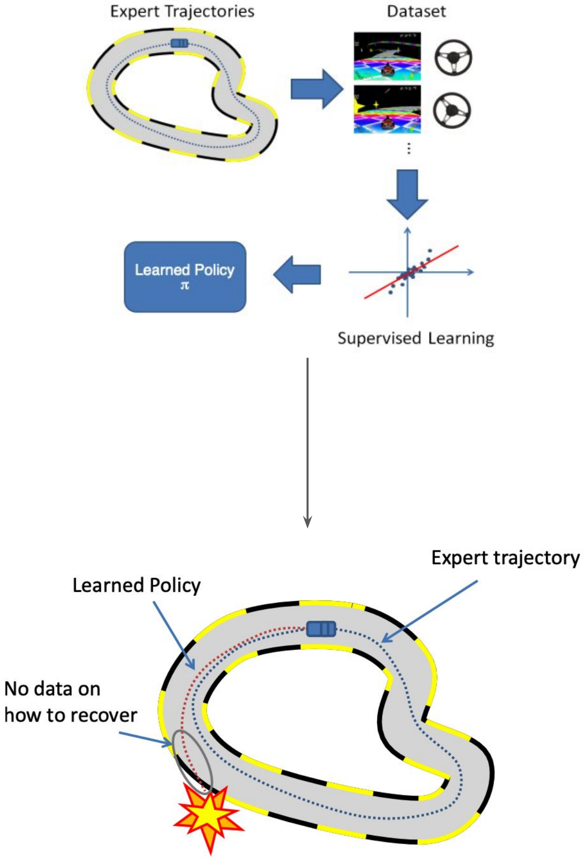
Behavioral Cloning

Treat imitation learning as a supervised learning problem

Issues

- Assumes trajectories and actions are IID
- Accumulating error

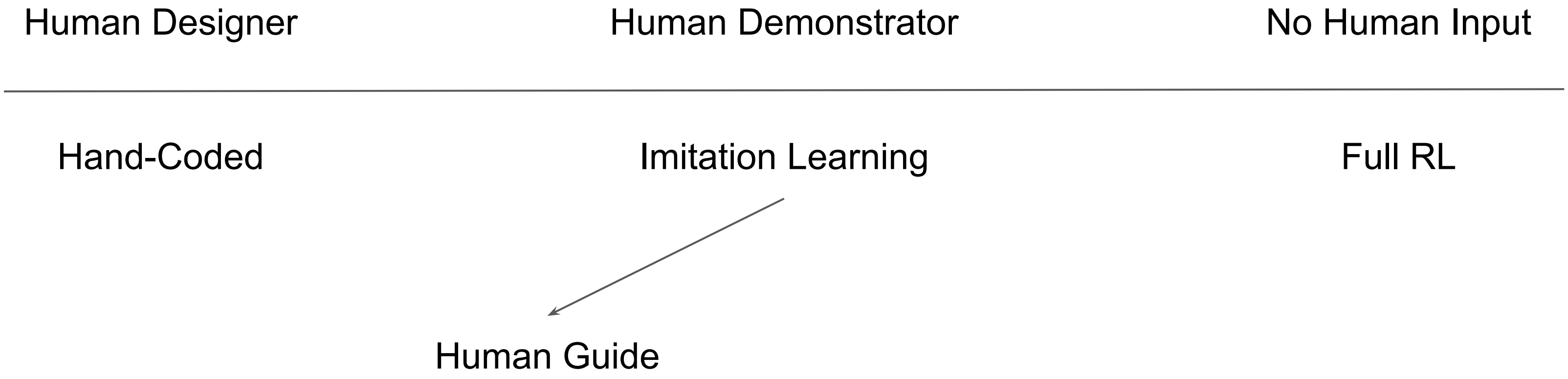
Once the agent leaves the expert path, it has no relevant experience



Human in the Loop Imitation Learning

Human Designer	Human Demonstrator	No Human Input
Hand-Coded	Imitation Learning	Full RL

Human in the Loop Imitation Learning



Human Designer

Human Demonstrator

No Human Input

Hand-Coded

Imitation Learning

Full RL

Human Guide

Human in the Loop Imitation Learning

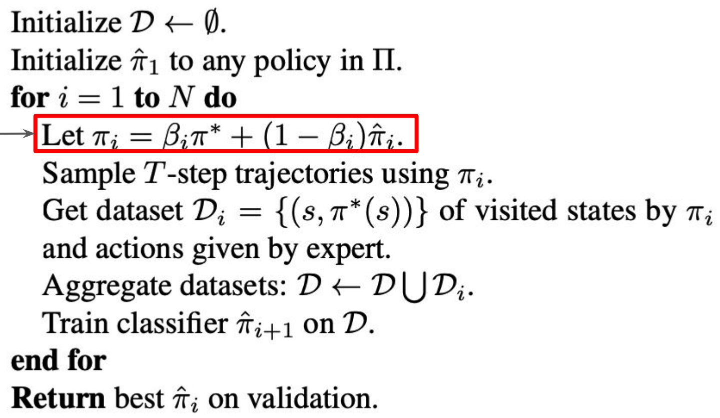
DAgger

Augments dataset with each rollout

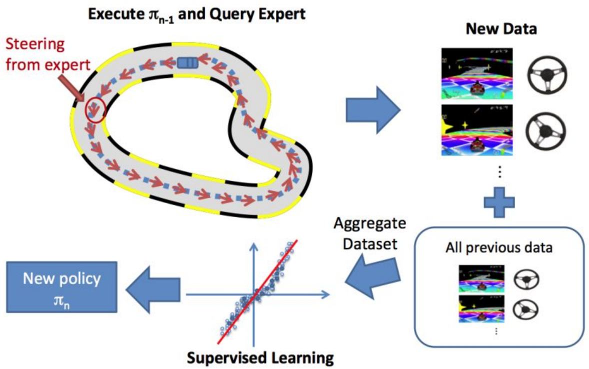
- Uses expert policy with probability β
- β decreases exponentially

π^* is expert policy, $\hat{\pi}$ is agent policy trained on D

DAgger Algorithm



```
Initialize $\mathcal{D} \leftarrow \emptyset$.
Initialize $\hat{\pi}_1$ to any policy in $\Pi$.
**for** $i = 1$ **to** $N$ **do**
    $\boxed{\text{Let } \pi_i = \beta_i \pi^* + (1 - \beta_i)\hat{\pi}_i.}$
    Sample $T$-step trajectories using $\pi_i$.
    Get dataset $\mathcal{D}_i = \{(s, \pi^*(s))\}$ of visited states by $\pi_i$
    and actions given by expert.
    Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \bigcup \mathcal{D}_i$.
    Train classifier $\hat{\pi}_{i+1}$ on $\mathcal{D}$.
**end for**
**Return** best $\hat{\pi}_i$ on validation.
```



Problems with DAgger

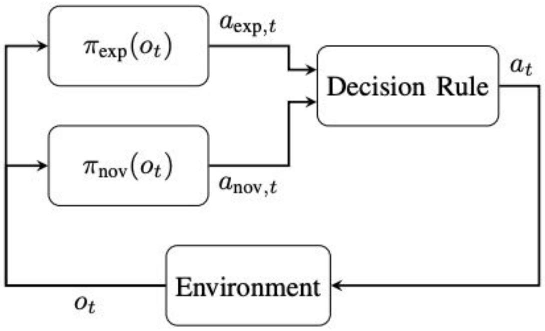
- Shared control scheme can alter expert behavior
  - Human expert does not receive feedback
- β is difficult to tune
  - High beta reduces to behavioral cloning
  - Low beta makes training unsafe and difficult for expert
- Expert demonstration is queried at random

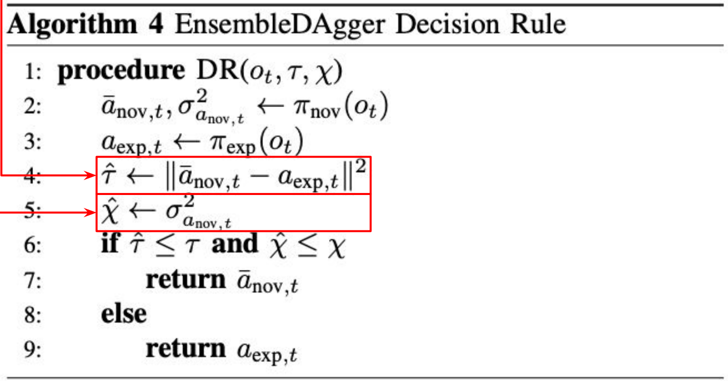
Ensemble DAgger

Novice is an ensemble of Neural Networks

Decision Rule Considerations:

- Discrepancy between novice and expert
- Variance between networks in the ensemble (doubt)





Algorithm 4 EnsembleDAgger Decision Rule

```
procedure DR($o_t, \tau, \chi$)
    $\bar{a}_{\text{nov},t}, \sigma^2_{a_{\text{nov},t}} \leftarrow \pi_{\text{nov}}(o_t)$
    $a_{\text{exp},t} \leftarrow \pi_{\text{exp}}(o_t)$
    $\hat{\tau} \leftarrow \|\bar{a}_{\text{nov},t} - a_{\text{exp},t}\|^2$
    $\hat{\chi} \leftarrow \sigma^2_{a_{\text{nov},t}}$
    if $\hat{\tau} \leq \tau$ and $\hat{\chi} \leq \chi$
        return $\bar{a}_{\text{nov},t}$
    else
        return $a_{\text{exp},t}$
```

Problems with DAgger

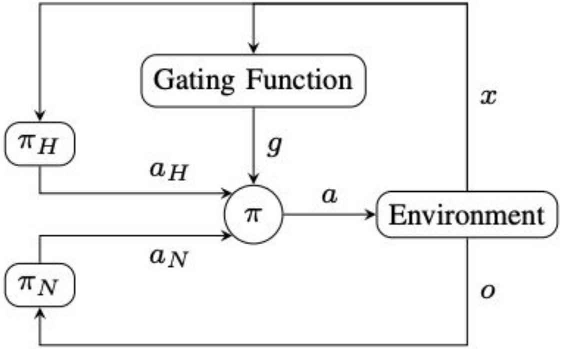
- Shared control scheme can alter expert behavior
  - Human expert does not receive feedback
- ~~β is difficult to tune~~
  - ~~High beta reduces to behavioral cloning~~
  - ~~Low beta makes training unsafe and difficult for expert~~
- ~~Expert demonstration is queried at random~~

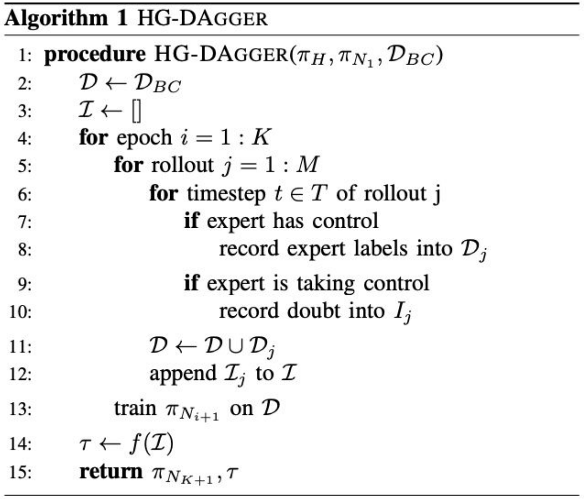
Human-Gated (HG) DAgger

Motivation: Human expert can provide better actions if given control for long stretches

Features

- Expert *is* the gating function
- Novice is represented by NN Ensemble
- Computes doubt when expert takes control





Algorithm 1 HG-DAGGER

```
procedure HG-DAGGER($\pi_H, \pi_{N_1}, \mathcal{D}_{BC}$)
    $\mathcal{D} \leftarrow \mathcal{D}_{BC}$
    $\mathcal{I} \leftarrow []$
    for epoch $i = 1 : K$
        for rollout $j = 1 : M$
            for timestep $t \in T$ of rollout j
                if expert has control
                    record expert labels into $\mathcal{D}_j$
                if expert is taking control
                    record doubt into $I_j$
            $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_j$
            append $\mathcal{I}_j$ to $\mathcal{I}$
        train $\pi_{N_{i+1}}$ on $\mathcal{D}$
    $\tau \leftarrow f(\mathcal{I})$
    return $\pi_{N_{K+1}}, \tau$
```

Doubt in HG-DAgger

$$d_N(o_t) = \|\mathrm{diag}(C_t)\|_2$$

Doubt is the L2 Norm of the Cov matrix

$$\tau = \frac{1}{\mathrm{len}(\mathcal{I})/4} \sum_{i=\lfloor .75N \rfloor}^{N} (\mathcal{I}[i])$$

Compute τ as the mean of the last 25% of entries in the doubt log

Experimental Setup

Experiments are done in simulation and in real world

Environment: Two-lane, one way roadway populated with stationary cars

Task: Ego vehicle must weave through cars without leaving the road

- Additional bike test in simulation

Experimental Setup

Agent Inputs (state)

- Distance to median
- Orientation
- Speed
- Distances to each edge of current lane
- Distances to nearest obstacle in each lane

Policy Outputs (action)

- Steering Angle
- Speed

Metrics

- Road departure rate
- Collision rate
- Bhattacharyya distance (to human driving data)

Baselines

- Behavioral Cloning
- DAgger (β_t=0.85^{t+1})

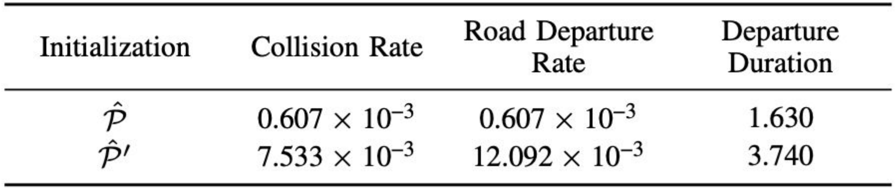
Experimental Setup

Training Method

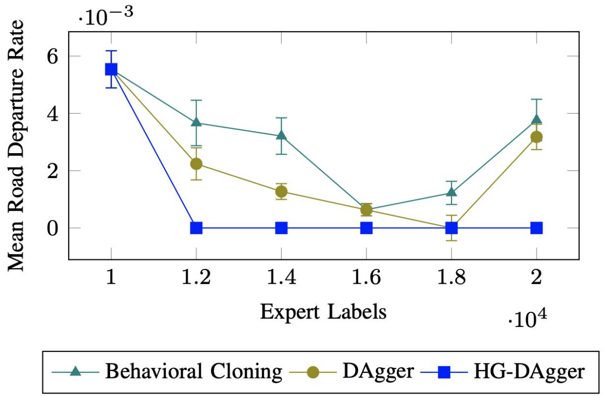
- Initialized with 10,000 samples for behavioral cloning
- Obtains 2,000 more samples per training loop for 5 iterations

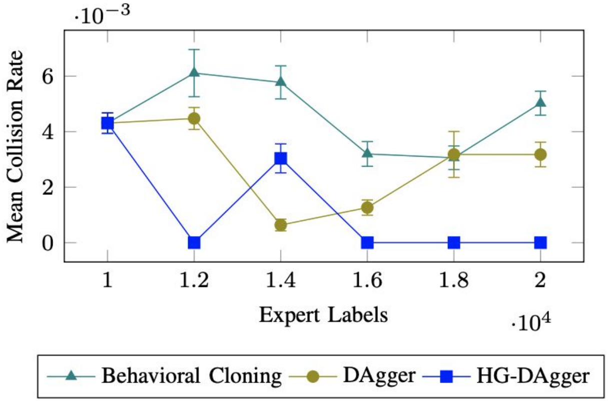
Experimental Results - Sim

- HG-DAgger outperformed DAgger and BC
- Surprisingly, DAgger/BC performed worse with more data
- Rollouts initialized inside the permissible set (states where doubt is below threshold) had much lower failure rates



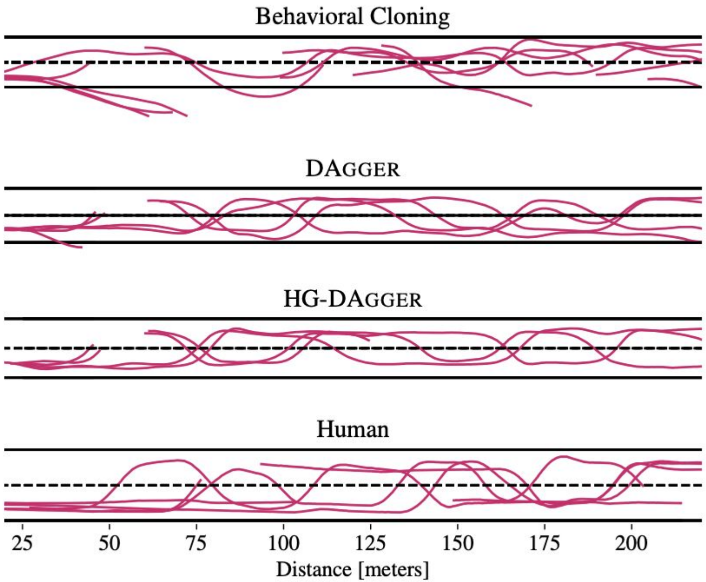
Initialization	Collision Rate	Road Departure Rate	Departure Duration
$\hat{\mathcal{P}}$	0.607×10^{-3}	0.607×10^{-3}	1.630
$\hat{\mathcal{P}}'$	7.533×10^{-3}	12.092×10^{-3}	3.740

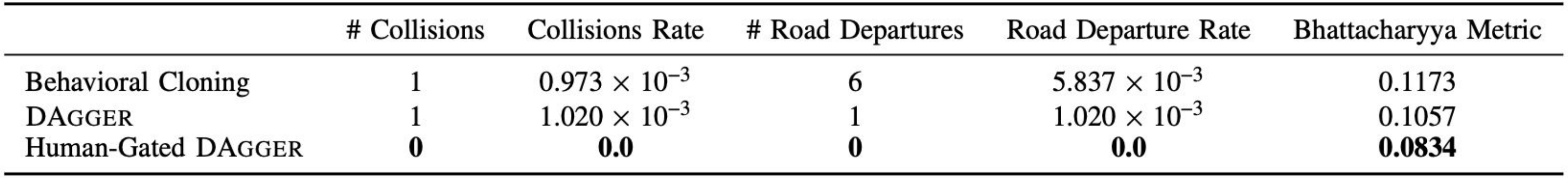




Experimental Results - World

- HG-DAgger behaves the most like a human
- It performs well in real-world testing

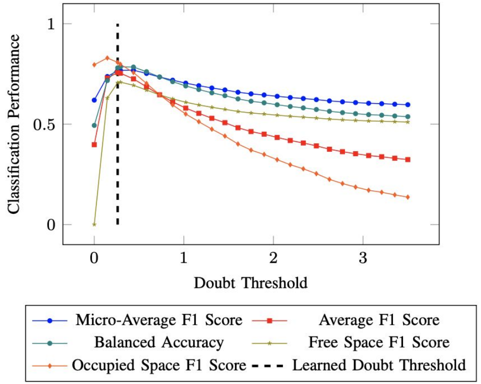


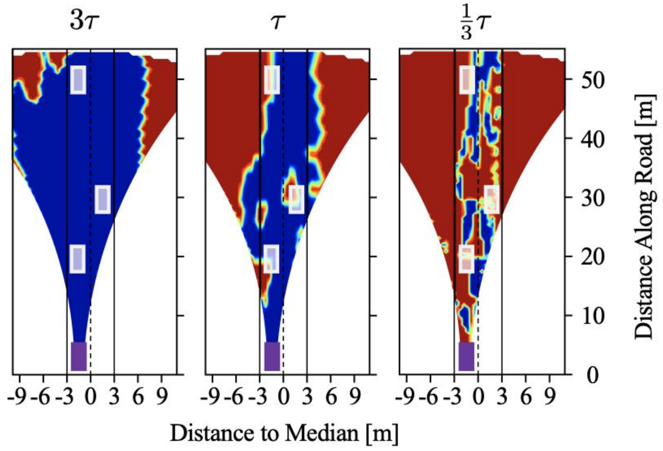


	# Collisions	Collisions Rate	# Road Departures	Road Departure Rate	Bhattacharyya Metric
Behavioral Cloning	1	0.973×10^{-3}	6	5.837×10^{-3}	0.1173
DAgger	1	1.020×10^{-3}	1	1.020×10^{-3}	0.1057
Human-Gated DAgger	**0**	**0.0**	**0**	**0.0**	**0.0834**

Experimental Results - World

- HG-DAgger produces an optimal doubt threshold
- Results demonstrate potential for practical use of doubt threshold generated by HG-DAgger





Critique / Limitations

Critiques

- Mentioned Bicycle test and showed no results
- Only baselines were Behavioral Cloning and DAgger
- Very limited amount of real world data
- Did not compare doubt method to Ensemble DAgger

Limitations

- Gathering real world (or sim) data with a human operator is expensive and potentially risky, and HG-DAgger needs supervision 100% of the time
- Cannot surpass expert performance
    - Is imitation learning from demonstration even useful for driving?

Future Work

- Test further in the real world
- Use a doubt-based risk metric as a gating function
- Find ways to reduce the amount of human input required
    - ThriftyDAgger identifies novel and risky states
- MIND MELD - Perform better than the supervisor

Tesla!

Elon Musk quotes from 2019 interview with ARK Invest

- *“we have just a vast amount of data on interventions. So, effectively, the customers are training the system on how to drive.”*
- *“Every time somebody intervenes - takes over from Autopilot - it saves that information and uploads it to our system ... And we’re really starting to get quite good at not even requiring human labeling. Basically the person, say, drives the intersection and is thereby training Autopilot what to do.”*

Source: https://seekingalpha.com/article/4248919-tesla-waymo-and-autonomous-driving-via-imitation-learning

Extended Readings

- Menda, Kunal, et al. “EnsembleDAgger: A Bayesian Approach to Safe Imitation Learning.” 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2019, pp. 5041–48. DOI.org (Crossref), https://doi.org/10.1109/IROS40897.2019.8968287.

- Hoque, R., Balakrishna, A., Novoseller, E., Wilcox, A., Brown, D. S., & Goldberg, K. (2021). ThriftyDAgger: Budget-aware novelty and risk gating for interactive imitation learning. arXiv preprint arXiv:2109.08273.

- Schrum, Mariah L., et al. “MIND MELD: Personalized Meta-Learning for Robot-Centric Imitation Learning.” 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 2022, pp. 157–65. DOI.org (Crossref), https://doi.org/10.1109/HRI53351.2022.9889616.

Summary

- Imitating human behavior is much easier and safer than reinforcement learning on many tasks
- DAgger fixes the accumulating error problem, but introduces issues with human input
- HG-DAgger uses the expert *as* the gating function, so the human has full control over the collection of training data.
- A new estimate of doubt can be used to determine how confident the novice is about its actions
- Results show that this method of data collection could work in the real world

Image Sources

https://gigazine.net/gsc_news/en/20161128-alvinn/

https://www.semanticscholar.org/paper/Grasping-with-Chopsticks%3A-Combating-Covariate-Shift-Ke-Wang/f518ce9eac0b473883ff6ae3abc3836a75f82ff5

https://www.cs.cmu.edu/~sross1/publications/ross_phdthesis.pdf