DAgger: Taming Covariate Shift with No Regret (Part 2!)

Sanjiban Choudhury
Behavior Cloning

Expert runs away after demonstrations
The Big Problem with BC

Train

$$\sum_{t=0}^{T-1} \mathbb{E}_{s_t \sim d_t^\pi_*} [\ell(s_t, \pi(s_t))]$$

Test

$$\sum_{t=0}^{T-1} \mathbb{E}_{s_t \sim d_t^\pi} [\ell(s_t, \pi(s_t))]$$
The Goal

\[ \sum_{t=0}^{T-1} \mathbb{E}_{s_t \sim d_t^\pi} [\ell (s_t, \pi(s_t))] \]

Can we bound this to \( O(\epsilon T) \) ?
DAGGER: A meta-algorithm for imitation learning

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning

Stéphane Ross  
Robotics Institute  
Carnegie Mellon University  
Pittsburgh, PA 15213, USA  
stephaneross@cmu.edu

Geoffrey J. Gordon  
Machine Learning Department  
Carnegie Mellon University  
Pittsburgh, PA 15213, USA  
ggordon@cs.cmu.edu

J. Andrew Bagnell  
Robotics Institute  
Carnegie Mellon University  
Pittsburgh, PA 15213, USA  
dbagnell@ri.cmu.edu
DAgger: Initializations

Data

Human drives

Policy $\pi_1$
DAgger: Iteration 1

[Ross et al’11]

Robot $\pi_1$ drives

Human corrects!

Data + Old Data

AGGREGATE DATA

Policy $\pi_2$
DAgger: Iteration 2

Old Data + Data → AGGREGATE DATA → Robot $\pi_2$ drives → Policy $\pi_2$
DAgger: Iteration N

[Ross et al’11]

Robot $\pi_N$ drives

After many iterations .... we are able to drive like a human!
**DAgger (Dataset Aggregation)**

Initialize with a random policy $\pi_1$  # Can be BC

Initialize empty data buffer $\mathcal{D} \leftarrow \{\}$

For $i = 1, \ldots, N$

Execute policy $\pi_i$ in the real world and collect data

$$\mathcal{D}_i = \{s_0, a_0, s_1, a_1, \ldots\}$$  # Also called a rollout

Query the expert for the optimal action on learner states

$$\mathcal{D}_i = \{s_0, \pi^*(s_0), s_1, \pi^*(s_1), \ldots\}$$

Aggregate data $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_i$

Train a new learner on this dataset $\pi_{i+1} \leftarrow \text{Train}(\mathcal{D})$

Select the best policy in $\pi_{1:N+1}$
The DAGGER Guarantee

DAGGER returns a policy \( \pi \) such that

\[
J(\pi) - J(\pi^*) \leq O(\epsilon HT)
\]

\( H \) is the recoverability coefficient that says if I make a mistake, how much does an expert have to pay to recover
Many cool applications of DAGGER in robotics

Lee et al, Learning quadrupedal locomotion over challenging terrain (2020)

Choudhury et al, Data Driven Planning via Imitation Learning (2018)

Chen et al Learning by Cheating (2020)

Pan et al Imitation learning for agile autonomous driving (2019)
How do we actually apply DAGGER in practice?

Asking a human expert to label every state the robot visits is hard.
Option 1: Extend DAGGER to different degrees of human feedback

Can we extend DAGGER to handle easier forms of human feedback preferences, interventions, etc?

Yes (*Future lectures!)
Option 2: Use an algorithmic oracle

What if we had a powerful algorithm that we can run in train time but not at test time?
Learning quadrupedal locomotion over challenging terrain

Joonho Lee¹, Jemin Hwangbo¹,² †, Lorenz Wellhausen¹, Vladlen Koltun³, Marco Hutter¹

¹ Robotic Systems Lab, ETH Zurich
² Robotics & Artificial Intelligence Lab, KAIST
³ Intelligent Systems Lab, Intel

†Substantial part of the work was carried out during his stay at 1
But why does aggregating data work?
From Imitation Learning to Interactive No-Regret Learning
Interactive Learning

Learner

\[
\min_{\pi} l(\pi)
\]

Adversary

Choose \[ l(\pi) \]
Interactive Learning

Learner

Initialize policy

Update policy

Adversary

Chooses loss

\[ \pi_1 \text{ [policy]} \]

\[ l_1(\cdot) \text{ [loss]} \]

\[ \pi_2 \]

\[ l_2(\cdot) \]
What is the best that I can do in such an adversarial setting?
From Imitation Learning to Interactive No-Regret Learning
How do we design algorithms that are no-regret?

Regret = \sum_{t=1}^{T} l_t(\pi_t) - \min_{\pi^*} \sum_{t=1}^{T} l_t(\pi^*)

(Learner) - (Best in hindsight)
At every round $t$, choose the best policy in hindsight

$$\pi_t = \arg\min_{\pi} \sum_{i=1}^{t-1} l_i(\pi)$$

(lowest total loss)
\[ \sum l_t \]

<table>
<thead>
<tr>
<th>Policy</th>
<th>( l_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policy 1</td>
<td>1.0</td>
</tr>
<tr>
<td>Policy 2</td>
<td>0.2</td>
</tr>
<tr>
<td>Policy 3</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Avg. Regret: \( - - \)
<table>
<thead>
<tr>
<th>Policy</th>
<th>$l_1$</th>
<th>$l_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Avg. Regret: 0.80
<table>
<thead>
<tr>
<th>$\sum l_t$</th>
<th>$l_1$</th>
<th>$l_2$</th>
<th>$l_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Policy 1</strong></td>
<td>1.5</td>
<td>1.0</td>
<td>0.5</td>
</tr>
<tr>
<td><strong>Policy 2</strong></td>
<td>0.7</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td><strong>Policy 3</strong></td>
<td>0.7</td>
<td>0.5</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Avg. Regret: 0.40
\[ \sum l_t \]

<table>
<thead>
<tr>
<th>Policy 1</th>
<th>Policy 2</th>
<th>Policy 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>0.2</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Avg. Regret: 0.53
<table>
<thead>
<tr>
<th>$\sum l_t$</th>
<th>$l_1$</th>
<th>$l_2$</th>
<th>$l_3$</th>
<th>$l_4$</th>
<th>$l_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policy 1</td>
<td>3.0</td>
<td>1.0</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
</tr>
<tr>
<td>Policy 2</td>
<td>1.9</td>
<td>0.2</td>
<td>0.5</td>
<td>1.0</td>
<td>0.2</td>
</tr>
<tr>
<td>Policy 3</td>
<td>1.4</td>
<td>0.5</td>
<td>0.2</td>
<td>0.2</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Avg. Regret: **0.40**
<table>
<thead>
<tr>
<th>$\sum l_t$</th>
<th>$l_1$</th>
<th>$l_2$</th>
<th>$l_3$</th>
<th>$l_4$</th>
<th>$l_5$</th>
<th>$l_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policy 1</td>
<td>3.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
</tr>
<tr>
<td>Policy 2</td>
<td>2.9</td>
<td>0.2</td>
<td>0.5</td>
<td>1.0</td>
<td>0.2</td>
<td>1.0</td>
</tr>
<tr>
<td>Policy 3</td>
<td>1.6</td>
<td>0.5</td>
<td>0.2</td>
<td>0.2</td>
<td>0.5</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Avg. Regret: **0.32**
\[ \sum l_t \]

<table>
<thead>
<tr>
<th></th>
<th>( l_1 )</th>
<th>( l_2 )</th>
<th>( l_3 )</th>
<th>( l_4 )</th>
<th>( l_5 )</th>
<th>( l_6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policy 1</td>
<td>1.0</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>1.0</td>
</tr>
<tr>
<td>4.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Policy 2</td>
<td>0.2</td>
<td>0.5</td>
<td>1.0</td>
<td>0.2</td>
<td>1.0</td>
<td>0.5</td>
</tr>
<tr>
<td>3.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Policy 3</td>
<td>0.5</td>
<td>0.2</td>
<td>0.2</td>
<td>0.5</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>1.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Avg. Regret: 0.26
Is FTL no-regret?
FTL is no-regret if 

1. We are in the continuous setting

2. Loss is strongly convex
Back to the proof!
Let’s recap!

We can frame interactive imitation learning as online learning

FTL is no-regret if the loss is strongly convex

DAGGER is FTL

No-regret implies $O(\epsilon HT)$
The rabbit hole of online learning

When does FTL break?
Loss = 0.75  Avg. Regret = 0.5

Choose $\pi^1$  Choose $\pi^2$
Loss = 1.0
Avg. Regret = 0.5

Choose $\pi^1$

Choose $\pi^2$
Choose $\pi_1$ for lower loss.

Loss = 1.0

Avg. Regret = 0.5

Choose $\pi_2$ for lower regret.
Loss = 1.0  Avg. Regret = 0.5

Choose $\pi^1$

Choose $\pi^2$
Loss = 1.0  Avg. Regret = 0.5

Choose $\pi^1$  Choose $\pi^2$
Be stable

Slowly change predictions

Achieve no-regret
Follow the **Regularized Leader**

\[ \pi_t = \arg \min_\pi \sum_{i=1}^{t-1} l_i(\pi) + \eta_t R(\pi) \]

Strong regularization!
Loss = 0.5  Avg. Regret = 0.25

Choose $\pi^1$  Choose $\pi^2$
Loss = 0.6  Avg. Regret = 0.17
Loss = 0.78
Avg. Regret = 0.21
Loss = 0.6  Avg. Regret = 0.18

Choose $\pi^1$ 

Choose $\pi^2$
Loss = 0.78  Avg. Regret = 0.2

Choose $\pi^1$  Choose $\pi^2$