Imitation Learning as Inferring Latent Expert Values

Sanjiban Choudhury
Two Core Ideas

Data

“What is the distribution of states?”

Loss

“What is the metric to match to human?”
Two Core Ideas

Data

“What is the distribution of states?”

Loss

“What is the metric to match to human?”
\textbf{DAGGER Algorithm}

\texttt{T}_0 \leftarrow \text{INITIALIZE POLICY WITH BEHAVIOR CLONING.}
\texttt{D} \leftarrow \{\}
\text{INITIALIZE EMPTY DATA BUFFER}

\textbf{FOR} \ i = 1 \ \cdots \ N

\textbf{Rollout} \ \texttt{T}_i
\texttt{(S}_1, a_1, S_2, a_2, \ldots \ldots \text{)}

\textbf{QUERY HUMAN} \ \texttt{A}^* \textbf{FOR}
\textbf{CORRECT ACTIONS}
\texttt{(S}_1, \texttt{A}^*(S)_1, S_2, \texttt{A}^*(S)_2, \ldots \ldots \text{)}

\texttt{D} \leftarrow \texttt{D} \cup \{\texttt{(S}_1, \texttt{A}^*(S)_1, S_2, \texttt{A}^*(S)_2, \ldots \ldots \text{)}\}

\texttt{T}_i \leftarrow \text{TRAIN} \texttt{(D)}
By training on **aggregated** data

$\pi_i$ is playing

**Follow the (Regularized) Leader!**

$$l_i(\pi) = \mathbb{E}_{s \sim d_\pi} 1(\pi(s) \neq \pi^*(s))$$

$$\pi_{i+1} = \arg \min_{\pi} \sum_{j=0}^{i} l_i(\pi) + R(\pi)$$
DAGGER results in an imitation gap of $O(\epsilon T)$

Assume the best policy in our policy class can drive down average loss to $\epsilon$

Then DAGGER finds a policy $\pi_i$

$$J(\pi_i) - J(\pi^*) \leq T l_i(\pi_i) \leq T \epsilon$$
Original results from DAGGER!
DAGGER is a foundation

Imitation under uncertainty
- SAIL
- ExPLORE
- STROLL
- Counterfactual Teaching

Model learning
- Agnostic
- SysID
- DaaD

Imitation learning
- DAEQUIL
- AGGREVATE(D)

Reinforcement Learning
- DPI
- EIL
- NRPI
- HG-DAGGER
- SHIV

Imitation learning
- Query efficient imitation learning

DAGGER
Many cool applications of DAGGER in robotics

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

Chen et al, Learning by Cheating (2020)

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

Pan et al, Imitation learning for agile autonomous driving (2019)
DAGGER is not just for imitation learning!

Agnostic System Identification for Model-Based Reinforcement Learning

Stéphane Ross  
Robotics Institute, Carnegie Mellon University, PA USA

J. Andrew Bagnell  
Robotics Institute, Carnegie Mellon University, PA USA

Model-based Reinforcement Learning
Hidden charges from DAGGER
Hidden Charge #1: Not all errors are equal
Recap: DAGGER

Roll out a learner policy

Collect expert actions

Aggregate data

Update policy

$$\min_{\pi} \mathbb{E}_{s,a^* \sim \mathcal{D}} 1(\pi(s) \neq a^*)$$
What does DAGGER guarantee?

Let's say your policy class $\Pi$ has 2 policies

Policy $\pi_1$:

*Shaky hands,*
*never goes out of racetrack,*
*but can’t recover if it did*
What does DAGGER guarantee?

Let's say your policy class $\Pi$ has 2 policies

Policy $\pi_2$:

Perfect on straight turns,
Perfect when falling off the cliff,
But makes mistake on the curve
What does D AGGER guarantee?

Which policy would you like to learn?

Which policy might D AGGER return?
Activity!
Think-Pair-Share!

Think (30 sec): Which policy would DAGGER return? How would you get it to choose $\pi_1$? Is DAGGER really $O(\epsilon T)$?

Pair: Find a partner

Share (45 sec): Partners exchange ideas
What is theoretically the best we can do in imitation learning?
Performance Difference Lemma
Is there a theoretically best imitation learning algorithm?

AGGREGATE

Reinforcement and Imitation Learning via Interactive No-Regret Learning

Stéphane Ross  J. Andrew Bagnell
stephaneross@cmu.edu  dbagnell@ri.cmu.edu
The Robotics Institute
Carnegie Mellon University,
Pittsburgh, PA, USA
AGGREGATE: Expert provides values

Just like DAGGER

For $i = 0 \ldots N-1$

Roll-in learner $\pi_i$ to get $\{s \sim d_{\pi_i}\}$

Query expert for advantage vector $A^*(s, \cdot)$

Aggregate data $\mathcal{D} \leftarrow \mathcal{D} \cup \{s, A^*(s, \cdot)\}$

Train policy $\pi_{i+1} = \mathbb{E}_{s,A^* \sim \mathcal{D}}(A^*(s, \pi(s)))$
AGGREGATE: Expert provides values

Just like DAGGER

For $i = 0 \ldots N-1$

Roll-in learner $\pi_i$ to get $\{s \sim d_{\pi_i}\}$

Query expert for advantage vector $A^*(s, \cdot)$

Aggregate data $\mathcal{D} \leftarrow \mathcal{D} \cup \{s, A^*(s, \cdot)\}$

Train policy $\pi_{i+1} = \mathbb{E}_{s, A^* \sim \mathcal{D}}(A^*(s, \pi(s)))$
Is Aggrevate even practical?
Yes*! When you are imitating algorithmic oracles

Train in Simulation

Learn Mapping Policy

Choudhury, S. et al Data-driven planning via imitation learning. *IJRR’18*

Test in the real world
Okay ...

But how do we learn from natural human feedback?
Hidden Charge #2: DAGGER queries the human at every state
Behavior Cloning crashes into a wall
DAGGER queries the human at every state!
Impractical: Too much human effort!

Can we learn from minimal human interaction?
Problem: **Impractical** to query expert everywhere

Can we learn from **natural** human interaction, e.g., interventions?
Learn from natural human interventions?

Hands free, no corrections!
Learn from natural human interventions?

Take over and drive back!
HG-DAGGER: Learning from interventions

Roll out a learner policy

Collect expert actions on states where expert intervened

Aggregate data

Update policy

$$\min_{\pi} \mathbb{E}_{s,a^* \sim \mathcal{D}} 1(\pi(s) \neq a^*)$$
Does this work?
Interventions are tell us something about the expert’s latent value function.
Expert Intervention Learning (EIL)

The expert action-value function is latent ...

\[ Q^* (s, a) \]

... and must be inferred from human interventions
Interventions are just constraints on latent action-value function.
Expert Intervention Learning (EIL)

Interventions are just constraints on latent action-value function

\[
\min_{Q \in \mathcal{Q}} \mathbb{E}_{(s^*, a^*) \sim P_{\text{expert}}} \ell(Q(s^*, \cdot), a^*)
\]

classify demonstrations
Expert Intervention Learning (EIL)  

Interventions are just **constraints** on latent action-value function

\[
\min_{Q \in \mathcal{Q}} \mathbb{E}_{(s^*, a^*) \sim P_{\text{expert}}} \ell(Q(s^*, \cdot), a^*) \\
\text{s.t.} \\
Q(s, a) \leq \delta_{\text{good}}
\]

classify demonstrations  
\( \forall (s, a) \in (I) \)  
before expert intervenes
Interventions are just **constraints** on latent action-value function

\[
\min_{Q \in Q} \mathbb{E}_{(s^*, a^*) \sim P_{\text{expert}}} \ell(Q(s^*, a^*)) \\
\text{s.t.} \\
Q(s, a) \leq \delta_{\text{good}} \\
Q(s, a) \geq \delta_{\text{good}}
\]

classify demonstrations

\forall (s, a) \in (I) 
before expert intervenes

\forall (s, a) \in (II) 
after expert intervenes
Interventions are just **constraints** on latent action-value function

\[
\min_{Q \in Q} \mathbb{E}_{(s^*, a^*) \sim P_{\text{expert}}} \ell(Q(s^*, \cdot), a^*) \\
\text{s.t.} \\
Q(s, a) \leq \delta_{\text{good}} \\
Q(s, a) \geq \delta_{\text{good}} \\
Q(s, a) \leq \min_{a'} Q(s, a)
\]

classify demonstrations

∀(s, a) ∈ (I) before expert intervenes

∀(s, a) ∈ (II) after expert intervenes

∀(s, a) ∈ (III) during expert intervention

Reduce to online, convex optimization! $O(\epsilon T)$
EIL is "good-enough" after 60 sec of trials.
EIL drives down error with less expert query
Turning interventions to simulations for learner

https://medium.com/aurora-blog/online-to-offline-turning-real-world-experience-into-virtual-tests-231c1cf8c2cd
The Big Picture

What we really want to solve is:

$$\min_{\pi} \mathbb{E}_{s \sim d_\pi}[Q^*(s, \pi(s)) - Q^*(s, \pi^*(s))]$$

Data

“What is the distribution of states?”

Use interactive online learning!

Loss

“What is the metric to match to human?”

Difference in Q values!
The Big Picture

What we really want to solve is:

$$\min_{\pi} \mathbb{E}_{s \sim d_{\pi}}[Q^*(s, \pi(s)) - Q^*(s, \pi^*(s))]$$

“What is the metric to match to human?”

Difference in Q values!

But Q* is latent!
The Big Picture

Estimate $Q^*$ from demonstrations, interventions, preferences, .. and even E-stops!

$Q^*(s, a)$

$\mathcal{L}(Q^*_\theta)$

Loss

Demonstrations

Interventions

Preferences

E-stops
tl;dr

The Big Picture

What we really want to solve is:

$$\min_{\pi} \mathbb{E}_{s \sim d_s}[Q^*(s, \pi(s)) - Q^*(s, \pi^*(s))]$$

Data

✓ "What is the distribution of states?"

Use interactive online learning!

Loss

✓ "What is the metric to match to human?"

Difference in Q values!

Problem: **Impractical** to query expert everywhere

Can we learn from natural human interaction, e.g., interventions?

Expert Intervention Learning (EIL) [SCB+ RSS'20]

The expert action-value function is latent ...

... and must be inferred from human interventions
Hidden charge #3: Dagger expects at least one policy to be good everywhere.

Learner simply can’t cross the bridge! …

… but can take the long way round.