Imitation Learning as Inferring Latent Expert Values

Sanjiban Choudhury
**Easy**

Expert is **realizable**

\[ \pi^E \in \Pi \]

As \( N \to \infty \), drive down

\[ \epsilon = 0 \] (or Bayes error)

**Solution**

Nothing special. Collect lots of data and do Behavior Cloning

---

**Hard**

Non-realizable expert + limited expert support

Even as \( N \to \infty \), behavior cloning \( O(\epsilon T^2) \)

**Solution**

Requires **interactive** expert (DAGGER) to provide labels \( \Rightarrow O(\epsilon T) \)
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?”
DAGGER queries the human at every state!
Impractical: Too much human effort!

Can we learn from **minimal** human interaction?
Today’s topic: Can we learn from minimal human feedback?

Think of the most minimal feedback:
An E-STOP!

How can we learn from this 1 bit feedback?
Recap: DAGGER

Roll out a learner policy

Collect expert actions

Aggregate data

Update policy

$$\min_{\pi} \mathbb{E}_{s,a^* \sim \mathcal{D}} 1(\pi(s) \neq a^*)$$
Not all errors are equal
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 DAGGER guarantee?

Which policy would you like to learn?

Which policy might DAGGER return?
Activity!
Think-Pair-Share!

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

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?

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

Deeply AggreVaTeD:
Differentiable Imitation Learning for Sequential Prediction

Wen Sun
Arun Venkatraman\textsuperscript{1}
Geoffrey J. Gordon
Byron Boots\textsuperscript{1}
J. Andrew Bagnell\textsuperscript{1}
\textsuperscript{1}School of Computer Science, Carnegie Mellon University, USA
\textsuperscript{2}College of Computing, Georgia Institute of Technology, USA

WENSUN@CS.CMU.EDU
AVENKATRAMAN@CS.CMU.EDU
GGORDON@CS.CMU.EDU
BBOOTS@CC.GATECH.EDU
DBAGNEILL@RI.CMU.EDU
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 $\pi_{i+1} = \mathbb{E}_{s, A^* \sim \mathcal{D}}(A^*(s, \pi(s)))$
AGGREVATE: 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 $A^*(s, . )$

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

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

Train search heuristics by imitating oracular planners
AGGREGVATE for helicopter planning

An autonomous helicopter navigating in a canyon

Learning a heuristic for 4D search (x,y,z,heading)

Dataset of canyons

A* using dubins distance heuristic times out (2531 states, 7000ms)

SAIL expands 18 states in 100 ms
AGGREGVATE for mapping unknown environments

Train Data: Office desks created in Gazebo

Test Data: RGBD data (Sturm et al.)
Okay …
But how do we learn from natural human feedback?
Recap: Learning to drive

Demonstration

[S\text{CB}^+ \text{ RSS’20}]

Learnt policy
Behavior Cloning crashes into a wall
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 ...

... and must be inferred from human interventions
Expert Intervention Learning (EIL)

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

$(s^*, a^*) \sim P_{\text{expert}}$
Expert Intervention Learning (EIL) [SCB+ RSS'20]

Interventions are just \textit{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
Expert Intervention Learning (EIL)

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

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

classify demonstrations

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

\(\forall (s, a) \in (\text{II})\) after expert intervenes

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

Interventions are just constraints on latent action-value function

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

s.t.

\[
Q(s, a) \leq \delta_{\text{good}}
\]

before expert intervenes

\[
Q(s, a) \geq \delta_{\text{good}}
\]

after expert intervenes

\[
Q(s, a) \leq \min_{a'} Q(s, a)
\]

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-231c1cf8c8cd
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?”

Loss

“What is the metric to match to human?”

Use interactive online learning!

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!

$Q^*(s, a)$

Loss

Demonstrations
Interventions
Preferences
E-stops

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

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?”

Loss

✓ “What is the metric to match to human?”

Use interactive online learning!

Problem: Impractical to query expert everywhere

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

Expert Intervention Learning (EIL)

The expert action-value function is latent ...

... and must be inferred from human interventions