Policy Search

- Problem: often feature-based policies that work well aren’t those that approximate $V/Q$ best
- Solution: Find policies that maximize rewards rather than the value that predicts rewards
- Successful
Example: Imitation Learning

- Take examples of experts \( \{(s_1, a_1)\ldots\} \)
- Learn a classifier mapping \( s \rightarrow a \)
- Create loss as the negative reward
Example: Imitation Learning

- Take examples of experts \( \{(s_1, a_1)\ldots\} \)
- Learn a classifier mapping \( s \rightarrow a \)
- Create loss as the negative reward
- What if we diverge?
How do we find a good policy?

- Find optimal policies through dynamic programming $\pi_0 \equiv \pi^*$
- Represent states $s$ through a feature vector $\vec{f}(s)$
How do we find a good policy?

- Find optimal policies through dynamic programming $\pi_0 \equiv \pi^*$
- Represent states $s$ through a feature vector $\vec{f}(s)$
- Until convergence:
  - Generate examples of state action pairs: $(\pi_t(s), s)$
  - Create a classifier that maps states to actions (an apprentice policy) $h_t: f(s) \rightarrow A$
  - Interpolate learned classifier $\pi_{t+1} = \lambda \pi_t + (1 - \lambda) h_t$
How do we find a good policy?

- Find optimal policies through dynamic programming $\pi_0 \equiv \pi^*$
- Represent states $s$ through a feature vector $\vec{f}(s)$
- Until convergence:
  - Generate examples of state action pairs: $(\pi_t(s), s)$
  - Create a classifier that maps states to actions (an apprentice policy) $h_t : f(s) \rightarrow A$ (Loss of classifier is the negative reward)
  - Interpolate learned classifier $\pi_{t+1} = \lambda \pi_t + (1 - \lambda) h_t$
How do we find a good policy?

- Find optimal policies through dynamic programming $\pi_0 \equiv \pi^*$
- Represent states $s$ through a feature vector $\vec{f}(s)$
- Until convergence:
  - Generate examples of state action pairs: $(\pi_t(s), s)$
  - Create a classifier that maps states to actions (an apprentice policy)
    $$ h_t : f(s) \mapsto A $$ (Loss of classifier is the negative reward)
  - Interpolate learned classifier
    $$ \pi_{t+1} = \lambda \pi_t + (1 - \lambda) h_t $$
- DAGGER: Dataset aggregation [Ross, Gordon & Bagnell, 2010]
- searn: Searching to Learn [Daumé & Marcu, 2006]
Applications of Imitation Learning

- Car driving
- Flying helicopters
- Question answering
- Machine translation
Applications of Imitation Learning

- Car driving
- Flying helicopters
- Question answering
- Machine translation
Question Answering

Question 36: 15 points

Two backgrounds of a painting by this artist show either...
Question Answering

- **State**: The words seen, opponent
- **Action**: Buzz or wait
- **Reward**: Points
Why machine translation really hard is

- **State**: The words you’ve seen, output of machine translation system
- **Action**: Take translation, predict the verb
- **Reward**: Translation quality
Comparing Policies
Comparing Policies

Source Sentence

Psychic

Er ist zum Laden gegangen
He went to the store

He to the store

Psychic

He went to the store

Good Translation

Bad Translation

Good Translation

Bad Translation

He went to the store
Comparing Policies
Comparing Policies

Source Sentence

Psychic

Er ist zum Laden gegangen

He to the store

Psychic

He went to the store

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation

He went to the store
Comparing Policies

Source Sentence

Psychic

Er ist zum Laden

He went to the store

Good Translation

Bad Translation
Comparing Policies

Source Sentence

Er ist zum Laden gegangen

Psychic

He went to the store

Good Translation

Bad Translation
Comparing Policies

Source Sentence

Er ist zum Laden gegangen

Psychic

He went to the store

Good Translation

Bad Translation

He to the store

He went to the store

He went to the store

He to the store went

He to the store

He went to the store
Comparing Policies

Source Sentence

- Er
- Psychic
- Monotone
- Batch

Policy Prediction

Good Translation

Bad Translation
Comparing Policies

Source Sentence

Psychic

Monotone

He went to the store

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation
Comparing Policies

Source Sentence

Psychic

Monotone

He went to the store

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation
Comparing Policies

Source Sentence

Psychic: Er ist zum Laden gegangen
Monotone: He went to the store
Bad: He to the store went

Good Translation
Bad Translation
Good Translation
Bad Translation
Good Translation
Comparing Policies

Source Sentence

Er ist zum Laden gegangen
He to the Psychic Monotone Batch Policy Prediction He went

Good Translation
Bad Translation
Good Translation
Bad Translation
Good Translation

Psychic

Monotone

He went to the store
He to the store
He went to the store
He to the store
He to the store went
He to the store
He went to the

Good Translation
Bad Translation
Good Translation
Bad Translation
Good Translation

Comparing Policies

Source Sentence

Er ist zum Laden gegangen

Psychic

He went to the store

Monotone

He to the store went

Bad Translation

Good Translation

Bad Translation

Good Translation
Comparing Policies

Source Sentence

Psychic

Monotone

Bad

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Er ist zum Laden gegangen

He went to the store

He to the store went
Comparing Policies

Source Sentence

Psychic

Monotone

Batch

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation
Comparing Policies

Source Sentence

Psychic

Monotone

Batch

Er ist zum Laden gegangen

He went to the store

He to the store

Psychic

Monotone

Batch

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation
Comparing Policies

Source Sentence

Psychic

Monotone

Batch

Er ist zum Laden gegangen
He went to the store

He went to the store

Psychic
He went to the store

Monotone
He went to the store

Batch

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation
Comparing Policies

Source Sentence

<table>
<thead>
<tr>
<th>Er</th>
</tr>
</thead>
<tbody>
<tr>
<td>ist</td>
</tr>
<tr>
<td>zum</td>
</tr>
</tbody>
</table>

Psychic

He went to the store

Monotone

He went to the

Batch

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation
Comparing Policies

Source Sentence

Psychic

Monotone

Batch

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation

Er ist zum Laden gegangen

He to the Psychic

Monotone

Batch

He went to the store

He to the store
Comparing Policies

Source Sentence

<table>
<thead>
<tr>
<th>Er</th>
<th>ist</th>
<th>zum</th>
<th>Laden</th>
<th>gegangen</th>
</tr>
</thead>
</table>

Psychic

He went to the store

Monotone

He to the store

Batch

He went to the store

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation
Comparing Policies

Source Sentence:
Er ist zum Laden gegangen

Psychic:
- He went to the store

Monotone:
- He to the
- He to the store
- He to the store went

Batch:
- He went to the store

Good Translation:
- He went to the store

Bad Translation:
- He to the
- He to the store went
- He went to the store
Comparing Policies

Source Sentence

Er ist zum Laden gegangen
He went to the store
He to the
He to the store
Psychic
Monotone
Batch
Policy Prediction

Good Translation
Bad Translation
Good Translation
Bad Translation
Good Translation
Bad Translation
Good Translation
Bad Translation
Comparing Policies

Source Sentence

Psychic

Monotone

Batch

Policy Prediction

He went to the store

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation
Comparing Policies

Source Sentence

Psychic

Monotone

Batch

Policy Prediction

He went to the store
Comparing Policies

Source Sentence

Psychic

Monotone

Batch

Policy Prediction

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation

He went to the store
Comparing Policies

Source Sentence:
Er ist zum Laden gegangen
He to the Psychic Monotone Batch Policy Prediction He went β Source Sentence

Good Translation
Bad Translation
Good Translation
Bad Translation
Good Translation
Bad Translation
Good Translation
Bad Translation

He went to the store
He to the store
He went to the store
He to the store went
He to the store
He went to the store

Machine Learning: Jordan Boyd-Graber | UMD
Comparing Policies

Source Sentence

Psychic

Monotone

Batch

Policy Prediction

He went to the store

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation
Comparing Policies

Source Sentence

Psychic

Monotone

Batch

Policy Prediction

Er ist zum Laden gegangen

He to the Psychic

Monotone Batch Policy Prediction

He went

T

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation

Machine Learning: Jordan Boyd-Graber | UMD
Applying SEARN
Applying SEARN

Oracle Example

Training Examples

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>f(s_1)</td>
<td>Wait</td>
</tr>
<tr>
<td>f(s_2)</td>
<td>Predict</td>
</tr>
<tr>
<td>f(s_3)</td>
<td>Wait</td>
</tr>
</tbody>
</table>
Applying SEARN

Oracle Example

Training Examples

\[ \pi : S \rightarrow \alpha \]

Classifier

Cost-Sensitive Classification

\[ x \quad y \]

\[ f(s_1) \quad \text{Wait} \]

\[ f(s_2) \quad \text{Predict} \]

\[ f(s_3) \quad \text{Wait} \]
Applying SEARN

Oracle Example

\[ \pi : s \mapsto a \]

Classifier

Machine Learning: Jordan Boyd-Graber | UMD
Applying SEARN

Oracle Example

\[ \pi : s \mapsto a \]
Classifier

\[ \vdash : s \rightarrow a \]

NO
GOOD!
Applying SEARN

Oracle Example

\[ \pi : s \mapsto a \]

Classifier

\( \Rightarrow \)

Wait
Predict
Wait
Wait
Commit
Wait
Reward
Time

 Wait
Commit
Reward
Time
Wait Wait Wait
Applying SEARN

Oracle Example

Classifier

\[ \pi : s \mapsto a \]

Machine Learning: Jordan Boyd-Graber | UMD
Applying SEARN

Training Examples

\[
x \quad y
\]

\[f(s_1)\]

\[f(s_2)\]

\[f(s_3)\]
Applying SEARN

Machine Learning: Jordan Boyd-Graber

Classifier N

Training Examples

\[ \pi : s \mapsto a \]

Prediction

Wait

Reward

Time

\[ f(s_1) \]

\[ f(s_2) \]

\[ f(s_3) \]

Machine Learning

| UMD

9 / 10
Recap

- Learning from examples: imitation learning
- Role of supervised machine learning
- Room for deep learning . . .