Efficient Robot Skill Learning: Grounded Simulation Learning and Imitation Learning from Observation

Peter Stone

Learning Agents Research Group (LARG)
Department of Computer Science
The University of Texas at Austin

(Also, Cogitai Inc.)
Research Question

To what degree can autonomous intelligent agents learn in the presence of teammates and/or adversaries in real-time, dynamic domains?
Research Question

To what degree can autonomous intelligent agents learn in the presence of teammates and/or adversaries in real-time, dynamic domains?

Research Areas

- Autonomous agents
- Multiagent systems
- Robotics
To what degree can autonomous intelligent agents learn in the presence of teammates and/or adversaries in real-time, dynamic domains?

**Research Areas**
- Autonomous agents
- Multiagent systems
- **Robotics**
- Machine learning
  - Reinforcement learning
Research Question

To what degree can autonomous intelligent agents learn in the presence of teammates and/or adversaries in real-time, dynamic domains?

Research Areas
- Autonomous agents
- Multiagent systems
- Robotics
- Machine learning
  - Reinforcement learning
Learning to interpret natural-language commands through human-robot dialog

Jesse Thomason, Shiqi Zhang, Raymond Mooney, and Peter Stone

Department of Computer Science
The University of Texas at Austin, Austin, TX 78712 USA
Research Question

To what degree can autonomous intelligent agents learn in the presence of teammates and/or adversaries in real-time, dynamic domains?

Research Areas

- Autonomous agents
- Multiagent systems
- Robotics
- Machine learning
  - Reinforcement learning
Research Question

To what degree can autonomous intelligent agents learn in the presence of teammates and/or adversaries in real-time, dynamic domains?

Research Areas
- Autonomous agents
- Multiagent systems
- Robotics
- Machine learning
  - Reinforcement learning
  - Cogitai
Self-learning Actionable AI
More than 60 Years Combined AI R&D

Leadership Team

MARK RING
CEO & Cofounder
“Continual Learning”

PETER STONE
President & COO
Cofounder

PETER WURMAN
VP Engineering

DENNIS CRESPO
VP Marketing & Business Dev

“Brain Trust” Technical Advisory Board — The people who created Reinforcement Learning

SUTTON
U of Alberta

LITTMAN
Brown University

ISBELL
Georgia Tech

ZHANG
U of Haernberg

VAN ROY
Stanford

SATINDER SINGH
Co-founder

BARTO
U. of Mass.

PRECUPE
McGill

BOWLING
U of Alberta

PARKES
Harvard

DAYAN
Gatsby, UCL

Full Time Team

15 PHDs

20 Total
**Continua™ SaaS Platform** improves any process, software bot, system

**First Markets:**
- Automotive Engine Control
- Robotics Control
- Semiconductor Control
Use Cases are Endless
Easy to Replicate Across Industries

- Decision Making
  - Customer service bots
- Web marketing
- Fitness coaches
- Video game agents
- Manufacturing Processes
- Robotic process automation
- Building management
- Self-learning vehicle
CogitAI’s Aggressive Roadmap to Continual Learning

1st Generation
- Human Labeling
  - Supervised Learning
  - Trained from human labels
  - Good for matching patterns
  - (Face Detection)

2nd Generation
- Self-Learns from Reward
  - Reinforcement Learning
  - Self-Learns through trial and reward
  - Good for decision making processes
  - (Robotics, energy management, video games)

3rd Generation
- Continual Learning
  - Acquires Skills, Knowledge on Own
  - Continually accumulates capabilities
  - Good for managing highly complex processes
  - (Multi-use robotics, plant control, autonomous cars, digital assistants)

Continua™
- Continua™ SaaS Platform improves any process, robot, software bot, decision system
Self-learning Actionable AI
Research Question

To what degree can autonomous intelligent agents learn in the presence of teammates and/or adversaries in real-time, dynamic domains?

Research Areas
- Autonomous agents
- Multiagent systems
- Robotics
- Machine learning
  - Reinforcement learning
  - Cogitai
Efficient Robot Skill Learning

Motivation:
Efficient Robot Skill Learning

- **Motivation:** RoboCup
Efficient Robot Skill Learning

- **Motivation:** RoboCup
- **Sim2Real:**
Efficient Robot Skill Learning

- **Motivation:** RoboCup
- **Sim2Real:** Grounded Simulation Learning
Efficient Robot Skill Learning

- **Motivation:** RoboCup
- **Sim2Real:** Grounded Simulation Learning
- **Imitation Learning from Observation:**
Efficient Robot Skill Learning

- **Motivation:** RoboCup
- **Sim2Real:** Grounded Simulation Learning
- **Imitation Learning from Observation:** BCO and GAIfO
RoboCup Soccer
RoboCup Soccer

- Grand challenge: beat World Cup champions by 2050
RoboCup Soccer

- Grand challenge: beat World Cup champions by 2050
- Still in relatively early stages
RoboCup Soccer

- Grand challenge: beat World Cup champions by 2050
- Still in relatively early stages
- Many virtues as a challenge problem:
  - Incremental challenges, closed loop at each stage
  - Robot design to multi-robot systems
  - Relatively easy entry
  - Inspiring to many
RoboCup Soccer

- Grand challenge: beat World Cup champions by 2050
- Still in relatively early stages
- Many virtues as a challenge problem:
  - Incremental challenges, closed loop at each stage
  - Robot design to multi-robot systems
  - Relatively easy entry
  - Inspiring to many
- Visible progress

![Small-sized League, Middle-sized League, Legged Robot League, Simulation League, Humanoid League](image)
RoboCup 1997–1998
RoboCup Soccer

- Grand challenge: beat World Cup champions by 2050
- Still in relatively early stages
- Many virtues as a challenge problem:
  - Incremental challenges, closed loop at each stage
  - Robot design to multi-robot systems
  - Relatively easy entry
  - Inspiring to many
- Visible progress
RoboCup Soccer

- Grand challenge: beat World Cup champions by 2050
- Still in relatively early stages
- Many virtues as a challenge problem:
  - Incremental challenges, closed loop at each stage
  - Robot design to multi-robot systems
  - Relatively easy entry
  - Inspiring to many
- Visible progress
UT Austin Villa
3D Simulation Team
RoboCup 2017 Highlights

World Champions
Record: 23-0
Goals For: 171, Goals Against: 0
RoboCup Soccer

- Grand challenge: beat World Cup champions by 2050
- Still in relatively early stages
- Many virtues as a challenge problem:
  - Incremental challenges, closed loop at each stage
  - Robot design to multi-robot systems
  - Relatively easy entry
  - Inspiring to many
- Visible progress

Small-sized League
Middle-sized League
Legged Robot League
Simulation League
Humanoid League
RoboCup@Home
RoboCup@Home
Open-world Reasoning for Service Robots

Yuqian Jiang*, Nick Walker*, Justin Hart, Peter Stone
RoboCup@Home
Motivation: RoboCup

**Sim2Real**: Grounded Simulation Learning

Imitation Learning from Observation: BCO and GAIfO
Reinforcement Learning for Physical Robots

Patrick MacAlpine
Josiah Hanna
Reinforcement Learning for Physical Robots

Patrick MacAlpine  Josiah Hanna

Learning on physical robots:
- Not data-efficient
- Requires supervision
- Manual resets
- Robots break
- Wear and tear make learning non-stationary
Reinforcement Learning in Simulation

Learning in simulation:
- Thousands of trials in parallel
- No supervision needed
- Automatic resets
- Robots don’t break
Reinforcement Learning in Simulation

Learning in simulation:
- Thousands of trials in parallel
- No supervision needed
- Automatic resets
- Robots don’t break

But, policies learned in simulation often fail in the real world.
Reinforcement Learning in Simulation

Learning in simulation:
- Thousands of trials in parallel
- No supervision needed
- Automatic resets
- Robots don’t break

But, policies learned in simulation often fail in the real world.
Sim2Real
(Cutler and How, “Efficient Reinforcement Learning for Robots using Informative Simulated Priors”);
(Cully et al., “Robots that can adapt like animals”);
(Rusu et al., “Sim-to-real robot learning from pixels with progressive nets”)
(Jakobi, Husbands, and Harvey, “Noise and the reality gap: The use of simulation in evolutionary robotics”); (Peng et al., “Sim-to-Real Transfer of Robotic Control with Dynamics Randomization”); (Tobin et al., “Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World”)

Peter Stone  
Robot Skill Learning  
UT Austin  
9
Grounded Simulation Learning

Real world state-action trajectories → Real World Policy Execution → Policy Improvement in Simulation → Improved Policy → Grounded Simulator → Simulator Grounding → Real world state-action trajectories

Farchy, Barrett, MacAlpine, and Stone, AAMAS 2013
Grounded Simulation Learning

- Real world state-action trajectories
- Simulator Grounding
- Grounded Simulator
- Policy Improvement in Simulation
- Improved Policy
- Real World Policy Execution

Farchy, Barrett, MacAlpine, and Stone, AAMAS 2013
Grounded Simulation Learning

Real world state-action trajectories

Simulator Grounding

Grounded Simulator

Real World Policy Execution

Policy Improvement in Simulation

Improved Policy

Farchy, Barrett, MacAlpine, and Stone, AAMAS 2013
Grounded Simulation Learning

Farchy, Barrett, MacAlpine, and Stone, AAMAS 2013
Grounded Simulation Learning

Farchy, Barrett, MacAlpine, and Stone, AAMAS 2013
Grounded Simulation Learning

Farchy, Barrett, MacAlpine, and Stone, AAMAS 2013
Simulator Grounding

Policy

Simulated Environment

$S, R$  $A$
Sim2Real

Policy

Simulated Environment

Policy

Environment

S, R → A → S, R

(Jakobi, Husbands, and Harvey, “Noise and the reality gap: The use of simulation in evolutionary robotics”);
(Peng et al., “Sim-to-Real Transfer of Robotic Control with Dynamics Randomization”);
(Tobin et al., “Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World”)

Peter Stone  Robot Skill Learning  UT Austin
Sim2Real

Policy

Simulated Environment

Data

Policy

Environment

(Abbeel, Quigley, and Ng, “Using Inaccurate Models in Reinforcement Learning”);
(Ross and Bagnell, “Agnostic System Identification for Model-Based Reinforcement Learning”)
Simulator Grounding

\[ S, R \rightarrow Policy \rightarrow Simulated Environment \rightarrow A \]
Simulator Grounding

\[ S, R \rightarrow \text{Policy} \rightarrow \text{Simulated Environment} \rightarrow \text{A} \]
Grounded Action Transformation

Replace robot’s action $a_t$ with an action that produces a more “realistic” transition.
Grounded Action Transformation

Replace robot’s action $a_t$ with an action that produces a more “realistic” transition.

Learn this action as a function $g(s_t, a_t)$.

Hanna and Stone, AAAI 2017
Grounded Action Transformation

\[ s_t, a_t \rightarrow g \rightarrow \hat{a}_t \]
Grounded Action Transformation

\[ s_t, a_t \]

\[ \hat{a}_t \]

\[ \hat{s}_t \]

\[ f(s_t, a_t) \]
Grounded Action Transformation

\[ s_t, a_t \]

\[ f(s_t, a_t) \]

\[ \hat{s}_t \]

\[ f^{-1}(s_t, \hat{s}_t) \]

\[ \hat{a}_t \]
Grounded Action Transformation

\[ S_{t+1} \]

\[ \text{Policy} \]

\[ g \]

\[ f(s_t, a_t) \]

\[ \hat{s}_t \]

\[ f^{-1}(s_t, \hat{s}_t) \]

\[ \hat{a}_t \]
Supervised Implementation

- Forward model:
  - trained with 15 real world trajectories of 2000 time-steps
Supervised Implementation

- **Forward model:**
  - trained with 15 real world trajectories of 2000 time-steps

- **Inverse model:**
  - trained with 50 simulated trajectories of 1000 time-steps
Supervised Implementation

- Forward model:
  - trained with 15 real world trajectories of 2000 time-steps
- Inverse model:
  - trained with 50 simulated trajectories of 1000 time-steps
- Initial policy in Initial vs. grounded simulator
Empirical Results

(a) Softbank NAO  (b) Gazebo NAO  (c) SimSpark NAO

Applied GAT to learning fast bipedal walks for the Nao robot.
- Initial policy: University of New South Wales Walk Engine.
- Policy Search Algorithm: CMA-ES stochastic search method.
Empirical Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Velocity (cm/s)</th>
<th>% Improve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial policy</td>
<td>19.3</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Empirical Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Velocity (cm/s)</th>
<th>% Improve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial policy</td>
<td>19.3</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Empirical Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Velocity (cm/s)</th>
<th>% Improve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial policy</td>
<td>19.3</td>
<td>0.0</td>
</tr>
<tr>
<td>1st iteration</td>
<td>26.3</td>
<td>34.6</td>
</tr>
</tbody>
</table>
Empirical Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Velocity (cm/s)</th>
<th>% Improve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial policy</td>
<td>19.3</td>
<td>0.0</td>
</tr>
<tr>
<td>1st iteration</td>
<td>26.3</td>
<td>34.6</td>
</tr>
<tr>
<td>2nd iteration</td>
<td><strong>28.0</strong></td>
<td><strong>43.3</strong></td>
</tr>
</tbody>
</table>
GSL Summary

- Introduced **Grounded Simulation Learning** for Sim2Real.

Hanna and Stone, AAAI 2017
GSL Summary

- Introduced **Grounded Simulation Learning** for Sim2Real.
- Improved walk speed of Nao robot by over 40% compared to state-of-the-art walk engine.
- Fastest known stable walk on the Nao

Patrick MacAlpine
Josiah Hanna

Hanna and Stone, AAAI 2017
GSL Summary

- Introduced **Grounded Simulation Learning** for Sim2Real.
- Improved walk speed of Nao robot by over 40% compared to state-of-the-art walk engine.
- Fastest known stable walk on the Nao

**Ongoing Work:**
- Extending to other robotics tasks and platforms
- When does grounding actions work and when does it not?

Hanna and Stone, AAAI 2017
**GSL Summary**

- Introduced **Grounded Simulation Learning** for Sim2Real.
- Improved walk speed of Nao robot by over 40% compared to state-of-the-art walk engine.
- Fastest known stable walk on the Nao

**Ongoing Work:**
- Extending to other robotics tasks and platforms
- When does grounding actions work and when does it not?
- Connecting to **off-policy evaluation** and **safe learning**

---

Hanna and Stone, AAAI 2017
Motivation: RoboCup
Sim2Real: Grounded Simulation Learning
Imitation Learning from Observation:
  - Model-based approach: BCO
GSL Summary

- Introduced **Grounded Simulation Learning** for Sim2Real.
- Improved walk speed of Nao robot by over 40% compared to state-of-the-art walk engine.
- Fastest known stable walk on the Nao

**Ongoing Work:**
- Extending to other robotics tasks and platforms
- When does grounding actions work and when does it not?
- Connecting to **off-policy evaluation** and **safe learning**

---

Hanna and Stone, AAAI 2017
Robot Skill Learning: Real World to Sim and Back

- Motivation: RoboCup
- Sim2Real: Grounded Simulation Learning
- **Imitation Learning from Observation:**
  - Model-based approach: BCO
Robot Skill Learning: Real World to Sim and Back

- Motivation: RoboCup
- Sim2Real: Grounded Simulation Learning
- Imitation Learning from Observation:
  - Model-based approach: BCO
  - Model-free approach: GAIfO
Robot Skill Learning: Real World to Sim and Back

- Motivation: RoboCup
- Sim2Real: Grounded Simulation Learning
- Imitation Learning from Observation:
  - Model-based approach: BCO
  - Model-free approach: GALfO

Faraz Torabi  Garrett Warnell
Imitation Learning

Goal:
- Learn how to make decisions by trying to imitate another agent.
Imitation Learning

Goal:
- Learn how to make decisions by trying to imitate another agent.

Conventional Imitation Learning:
- Observations of other agent (demonstrations) consist of state-action pairs.\(^1\)

\(^1\)Niekum et al., “Learning and generalization of complex tasks from unstructured demonstrations”.

Peter Stone
Robot Skill Learning
UT Austin 20
Imitation Learning

Goal:
- Learn how to make decisions by trying to imitate another agent.

Conventional Imitation Learning:
- Observations of other agent (demonstrations) consist of state-action pairs.\(^1\)
Imitation Learning

Goal:
- Learn how to make decisions by trying to imitate another agent.

Conventional Imitation Learning:
- Observations of other agent (demonstrations) consist of state-action pairs.¹

Challenge:
-Precludes using a large amount of demonstration data where action sequences are not given (e.g. YouTube videos).

¹Niekum et al., “Learning and generalization of complex tasks from unstructured demonstrations”.
Imitation Learning

Algorithms:
Imitation Learning

Algorithms:

- Behavioral Cloning:
Imitation Learning

Algorithms:

- Behavioral Cloning:
  - End to End Learning for Self-Driving Cars.²

²Zhang and Cho, “Query-Efficient Imitation Learning for End-to-End Simulated Driving.”
Imitation Learning

Algorithms:

- Behavioral Cloning:
  - End to End Learning for Self-Driving Cars.\(^2\)

- Inverse Reinforcement Learning:

\(^2\)Zhang and Cho, “Query-Efficient Imitation Learning for End-to-End Simulated Driving.”
Imitation Learning

Algorithms:

- **Behavioral Cloning:**
  - End to End Learning for Self-Driving Cars.\(^2\)

- **Inverse Reinforcement Learning:**
  - Generative Adversarial Imitation Learning.\(^3\)
  - Guided Cost Learning.\(^4\)

---

\(^2\) Zhang and Cho, “Query-Efficient Imitation Learning for End-to-End Simulated Driving.”

\(^3\) Ho and Ermon, “Generative adversarial imitation learning”.

\(^4\) Finn, Levine, and Abbeel, “Guided cost learning: Deep inverse optimal control via policy optimization”.
Imitation from Observation

Goal:

- Learn how to perform a task given state-only demonstrations.
Imitation from Observation

Goal:
- Learn how to perform a task given state-only demonstrations.
Imitation from Observation

Goal:
- Learn how to perform a task given state-only demonstrations.

Formulation:
- Given:
  - $D_{demo} = (s_0, s_1, ...)$
- Learn:
  - $\pi : S \rightarrow A$
Imitation from Observation

Previous work:
Imitation from Observation

Previous work:

- Time Contrastive Networks (TCN).\(^5\)
- Imitation from observation: Learning to imitate behaviors from raw video via context translation.\(^6\)
- Learning invariant feature spaces to transfer skills with reinforcement learning.\(^7\)

\(^6\) Liu et al., “Imitation from observation: Learning to imitate behaviors from raw video via context translation”.
\(^7\) Gupta et al., “Learning invariant feature spaces to transfer skills with reinforcement learning”.
Imitation from Observation

Previous work:

- Time Contrastive Networks (TCN).\textsuperscript{5}
- Imitation from observation: Learning to imitate behaviors from raw video via context translation.\textsuperscript{6}
- Learning invariant feature spaces to transfer skills with reinforcement learning.\textsuperscript{7}

Concentrate on perception; require time-aligned demonstrations.

\textsuperscript{5} Sermanet et al., “Time-contrastive networks: Self-supervised learning from multi-view observation”.
\textsuperscript{6} Liu et al., “Imitation from observation: Learning to imitate behaviors from raw video via context translation”.
\textsuperscript{7} Gupta et al., “Learning invariant feature spaces to transfer skills with reinforcement learning”.
Efficient Robot Skill Learning

- Motivation: RoboCup
- Sim2Real: Grounded Simulation Learning
- Imitation Learning from Observation:
  - Model-based approach: BCO
  - Model-free approach: GAIfO
Model-based Approach

- Imitation Learning: \[ D_{\text{demo}} = \{(s_0, a_0), (s_1, a_1), ...\} \]
Model-based Approach

- Imitation Learning: $D_{\text{demo}} = \{(s_0, a_0), (s_1, a_1), \ldots\}$
- Imitation from Observation: $D_{\text{demo}} = \{(s_0, ?), (s_1, ?), \ldots\}$
Model-based Approach

- Imitation Learning: \(D_{demo} = \{(s_0, a_0), (s_1, a_1), \ldots\}\)
- Imitation from Observation: \(D_{demo} = \{(s_0, ?), (s_1, ?), \ldots\}\)

Model-based Approach:

1. Learn an inverse dynamics model
2. Infer actions
3. Perform a conventional IL method
Behavioral Cloning from Observation (BCO)

Algorithm:

START

Initialize policy $\pi_{\phi}$

Run policy $\pi_{\phi}$

Collect data $T^a_{\pi}, A_{\pi}$

State-only demonstrations

Update policy $\pi_{\phi}$

Infer actions $M_\theta$

Update model $M_\theta$

$D_{demo}$

$T^a_{\pi}, A_{\pi}$

$s^a_t, s^a_{t+1}$

$\alpha_t$

$s_{t+1}$
Behavioral Cloning from Observation (BCO)

Experimental Results:
- **Domain:**
  - Mujoco domain "Ant" with 111 dimensional state space and 8 dimensional action space.
Behavioral Cloning from Observation (BCO)

Experimental Results:

![Graph showing performance over the number of demonstrated trajectories for different methods: Random, NAIL, Expert, BCO(0), FEM, BC. The x-axis represents the number of demonstrated trajectories, and the y-axis represents performance. The graph compares the performance of these methods across different numbers of demonstrations.]
Behavioral Cloning from Observation (BCO)

Experimental Results:

![Graph showing performance of different methods with varying number of Demonstrated trajectories. The x-axis represents the number of demonstrated trajectories ranging from 5 to 25, and the y-axis represents performance. The graph includes lines for Random, CMAIL, Expert, BCO(0), FEM, and BC. The performance values range from -1.5 to 1.0.]
Behavioral Cloning from Observation (BCO)

Experimental Results:

![Ant performance graph]

- Number of demonstrated trajectories
- Performance

Legend:
- Random
- Expert
- BCO(0)
- FEM
- BC
Behavioral Cloning from Observation (BCO(α))

Issue:

- Inverse dynamics model is learned using a random policy.
Behavioral Cloning from Observation ($BCO(\alpha)$)

Issue:

- Inverse dynamics model is learned using a random policy.

Solution: $BCO(\alpha)$
Behavioral Cloning from Observation (BCO(\(\alpha\)))

Issue:
- Inverse dynamics model is learned using a random policy.

Solution: BCO(\(\alpha\))
- Update the model with the learned policy.
Behavioral Cloning from Observation (BCO(\(\alpha\)))

Issue:
- Inverse dynamics model is learned using a random policy.

Solution: BCO(\(\alpha\))
- Update the model with the learned policy.
- Parameter \(\alpha\) controls tradeoff between performance and environment interactions.
Behavioral Cloning from Observation (BCO(\(\alpha\)))

Issue:
- Inverse dynamics model is learned using a random policy.

Solution: BCO(\(\alpha\))
- Update the model with the learned policy.
- Parameter \(\alpha\) controls tradeoff between performance and environment interactions
  - \(\alpha = 0\): no post-demonstration interaction.
Behavioral Cloning from Observation ($\text{BCO}(\alpha)$)

Issue:
- Inverse dynamics model is learned using a random policy.

Solution: $\text{BCO}(\alpha)$
- Update the model with the learned policy.
- Parameter $\alpha$ controls tradeoff between performance and environment interactions
  - $\alpha = 0$: no post-demonstration interaction.
  - Increasing $\alpha$: increasing the number of interactions allowed at each iteration.
Behavioral Cloning from Observation (BCO(α))

Algorithm:

START

Initialize policy $\pi_\phi$

Run policy $\pi_\phi$

Collect data $\tau_\pi^a, A_\pi$

Update model $M_\theta$

Infer actions

Update policy $\pi_\phi$

State-only demonstrations

$D_{demo}$

$S_{demo}$

$\tilde{A}_{demo}$
Behavioral Cloning from Observation (BCO(α))

Algorithm:

START

Initialize policy \( \pi^0 \)

Run policy \( \pi^i \) \( \{ (s^a_t, s^a_{t+1}) \} \)

Append to \( T^a_\pi, A_\pi \)

State-only demonstrations

Update policy \( \pi^i \) \( \{ a_t \} \)

Infer actions

Update model \( M^i_\theta \)

Behavioral Cloning from Observation (BCO)

\[ D_{demo} \]

\[ \tilde{A}_{demo}, S_{demo} \]
Behavioral Cloning from Observation (BCO(α))

Interaction time:

- Pre-demonstration
  - BCO(0)
  - BCO(α)
  - GAIL & FEM

- Post-demonstration
  - α₁ ż₀₀
  - α₁ ż₀₁
  - ... α₁ žₙₙ
  - α₁ ż₀₀ₙ

Time:
- Environment Interactions
- Inverse Model Update
- Policy Learning Update
Behavioral Cloning from Observation (BCO($\alpha$))

Effect of varying $\alpha$ on BCO($\alpha$):
Behavioral Cloning from Observation (BCO(\(\alpha\)))

Effect of varying \(\alpha\) on BCO(\(\alpha\)):

![Graph showing the effect of \(\alpha\) on BCO(\(\alpha\)). The graph plots performance against the number of demonstrated trajectories for different values of \(\alpha\).]
Behavioral Cloning from Observation (BCO($\alpha$))

Effect of varying $\alpha$ on BCO($\alpha$):

![Graph showing the effect of varying $\alpha$ on performance. The x-axis represents the number of demonstrated trajectories, ranging from 5 to 25. The y-axis represents performance, ranging from 0.0 to 1.0. Different lines represent different values of $\alpha$: random, $\alpha = 4\epsilon - 3$, $\alpha = 3\epsilon - 2$, $\alpha = 2\epsilon - 3$, and $\alpha = 0$. The graph shows that performance increases as the number of demonstrated trajectories increases, with different curves representing different $\alpha$ values.]
Efficient Robot Skill Learning

- Motivation: RoboCup
- Sim2Real: Grounded Simulation Learning
- Imitation Learning from Observation:
  - Model-based approach: BCO
  - **Model-free approach**: GAfO
Gen. Adversarial Imitation from Observation (GAIfO)

Motivation:

(a) Random Policy  
(b) Demonstration

Figure: State transition distribution in Hopper domain.
Gen. Adversarial Imitation from Observation (GAILfO)

Algorithm:

![Diagram showing the relationship between Demonstrator and Imitator networks and the environment.](image)
Gen. Adversarial Imitation from Observation (GAIfO)

Comparison against other IfO approaches and GAIL:
Gen. Adversarial Imitation from Observation (GAIIfO)

Comparison against other IfO approaches and GAIL:

![Graph showing comparison](image)
Gen. Adversarial Imitation from Observation (GAIfo)

Comparison against other Ifo approaches and GAIL:

![Bar chart showing comparison between Random, TCN, GAIfo, Expert, BCO, and GAIL methods across different numbers of demonstrated trajectories for Hopper task. The chart displays the final average normalized score.]
Gen. Adversarial Imitation from Observation (GAIfO)

Challenges:
Challenges:
- States are not fully-observable.
Gen. Adversarial Imitation from Observation (GAIfoO)

Challenges:
- States are not fully-observable.
- States are not Markovian.
Gen. Adversarial Imitation from Observation (GAIfoO)

Challenges:
- States are not fully-observable.
- States are not Markovian.

Solution:
Gen. Adversarial Imitation from Observation (GAIfO)

Algorithm:

Policy

Discriminator

Demonstration
Demonstration:
Learned Policy:
Gen. Adversarial Imitation from Observation (GAIfO)

Comparison against other IfO approaches:

![Graph showing comparison](image-url)
Gen. Adversarial Imitation from Observation (GAIfO)

Comparison against other IfO approaches:

![Graph showing comparison against other IfO approaches. The x-axis is labeled 'Number of demonstrated trajectories' ranging from 1 to 15. The y-axis is labeled 'Final Avg. Normalized Score' ranging from 0.0 to 1.0. The legend includes 'Random', 'TRPO', 'BCO', 'TCN', and 'GAIFO'. The graph shows the performance of these methods across different numbers of demonstrated trajectories.](image)
Ongoing Work
Ongoing Work

- Testing algorithms on more domains.
Ongoing Work

- Testing algorithms on more domains.
- Adapt algorithms for physical robots.
Ongoing Work

- Testing algorithms on more domains.
- Adapt algorithms for physical robots.
- Sim-to-real transfer using the algorithms.
Imitation Learning Summary

(a) BCO

(b) GALfO

Faraz Torabi    Garrett Warnell
Research Question

To what degree can autonomous intelligent agents learn in the presence of teammates and/or adversaries in real-time, dynamic domains?

Research Areas
- Autonomous agents
- Multiagent systems
- Machine learning
  - Reinforcement learning
- Robotics
Selected other RL Contributions

- Human interaction
  - Advice, Demonstration
  - Positive/Negative Feedback

- Transfer learning for RL

- Curriculum Learning

- RL for musical playlist recommendation

- TEXPLORE for Robot RL
  - Sample efficient; real-time
  - Continuous state; delayed effects

- Deep RL in continuous action spaces

[References]

-Knox & Stone, ’09
-Taylor & Stone, ’07
-Narvekar et al., ’16
-Liebman et al., ’15
-Hester & Stone, ’13
-Hausknecht & Stone, ’16

Selected MAS Contributions

- Autonomous traffic management
- Trading Agent Competition (PowerTAC)
- Ad Hoc Teamwork
Ad Hoc Teams

- Ad hoc team player is an individual
  - Unknown teammates (programmed by others)
- Teammates likely sub-optimal: no control

**Challenge:** Create a good team player

- Introduced as AAAI Challenge Problem
  - Theory: repeated games, bandits
  - Experiments: pursuit, flocking
  - RoboCup experiments

[AAAI’10] [AIJ’13] [Genter & Stone, ’12] [Genter et al., ’15]
Benchmarking Robot Cooperation without Pre-Coordination in the RoboCup Standard Platform League Drop-In Player Competition

Katie Genter*, Tim Laue°, Peter Stone*

* University of Texas at Austin, Austin, TX, USA
° University of Bremen, Germany
Ad Hoc Teams

- Ad hoc team player is an **individual**
  - Unknown teammates (*programmed by others*)
- Teammates likely **sub-optimal**: no control

**Challenge:** Create a good team player

- Introduced as AAAI Challenge Problem [AAAI’10]
  - Theory: repeated games, bandits [AIJ’13]
  - Experiments: **pursuit, flocking** [Genter & Stone, ’12]
  - **RoboCup experiments** [Genter et al., ’15]
Ad Hoc Teams

- Ad hoc team player is an individual
  - Unknown teammates (programmed by others)
- Teammates likely sub-optimal: no control

**Challenge:** Create a good team player

- Introduced as AAAI Challenge Problem [AAAI’10]
  - Theory: repeated games, bandits [AIJ’13]
  - Experiments: pursuit, flocking [Genter & Stone, ’12]
  - RoboCup experiments [Genter et al., ’15]
- Community: MIPC Workshops, JAAMAS issue
Efficient Robot Skill Learning: GSL and IfO

To what degree can autonomous intelligent agents learn in the presence of teammates and/or adversaries in real-time, dynamic domains?

- **Motivation:** RoboCup
- **Sim2Real:** Grounded Simulation Learning
- **Imitation Learning from Observation:** BCO and GAIfO
Selected other RL Contributions

- Human interaction
  - Advice, Demonstration
  - Positive/Negative Feedback [Knox & Stone, ’09] [Taylor & Stone, ’07]
- Transfer learning for RL
- Curriculum Learning [Narvekar et al., ’16]
- RL for musical playlist recommendation [Liebman et al., ’15]
- TEXPLORE for Robot RL
  - Sample efficient; real-time
  - Continuous state; delayed effects [Hester & Stone, ’13]
- Deep RL in continuous action spaces [Hausknecht & Stone, ’16]
Efficient Robot Skill Learning: GSL and IfO

To what degree can autonomous intelligent agents learn in the presence of teammates and/or adversaries in real-time, dynamic domains?

Motivation: RoboCup
Sim2Real: Grounded Simulation Learning
Imitation Learning from Observation: BCO and GAIfO
Selected other RL Contributions

- Human interaction
  - Advice, Demonstration
  - Positive/Negative Feedback

- Transfer learning for RL

- Curriculum Learning

- RL for musical playlist recommendation

- TEXPLORE for Robot RL
  - Sample efficient; real-time
  - Continuous state; delayed effects

[Knox & Stone, ’09]
[Taylor & Stone, ’07]
[Narvekar et al., ’16]
[Liebman et al., ’15]
[Hester & Stone, ’13]
Efficient Robot Skill Learning: GSL and IfO

To what degree can autonomous intelligent agents learn in the presence of teammates and/or adversaries in real-time, dynamic domains?

- **Motivation:** RoboCup
- **Sim2Real:** Grounded Simulation Learning
- **Imitation Learning from Observation:** BCO and GAIfO
Ad Hoc Teams

- Ad hoc team player is an individual
  - Unknown teammates (programmed by others)
- Teammates likely sub-optimal: no control

**Challenge:** Create a good team player

- Introduced as AAAI Challenge Problem [AAAI’10]
  - Theory: repeated games, bandits [AIJ’13]
  - Experiments: **pursuit, flocking** [Genter & Stone, ’12]
  - **RoboCup experiments** [Genter et al., ’15]
- Community: MIPC Workshops, JAAMAS issue