Reinforcement Learning @ MSR AI


September 2018
Reinforcement Learning: Stunts & Opportunities

Realistic Non-Player Characters

Automated Code Debugging

Advanced, pro-active Cortana
RL is a framework for sequential decision-making under uncertainty.

Agent

Environment

State

Reward

Action

Behaviour: State → Policy → Action

Goal: Find the policy that results in the highest expected sum of rewards.
RL differs from Supervised Learning

*The agent is not told how it should behave, but what it should achieve.*
Our Mission

We conduct ground-breaking research in Reinforcement Learning (RL) to drive real-world AI scenarios.
RL in Simulation

1. RL for next-gen videogame AI
   https://github.com/Microsoft/malmo

2. AirSim
   https://www.microsoft.com/en-us/research/project/aerial-informatics-robotics-platform/

3. Solving Interactive Fiction Games
   https://www.microsoft.com/en-us/research/project/textworld/

4. Grounded vision-language interaction
Next-Generation Game AI

• Create agents that engage and entertain human players, rather than replacing them.

• Build agents capable of learning in open ended worlds like Minecraft.

• Learn policies that we can easily calibrate to specific behaviors/playstyles.
Towards Calibratable Learned Behaviors

Our Goal: Learn policies that we can easily calibrate to specific behaviors/playstyles.
A Solution: Trajectory embedding + Imitation Learning

Frequency of firing

http://atarigrandchallenge.com/data
Our Mission

We conduct ground-breaking research in Reinforcement Learning (RL) to drive real-world AI scenarios.

Foundations

Simulation  |  Digital  |  Physical

Explore  Log  Learn  Deploy
RL in the Digital World

1. Decision Service
   https://ds.microsoft.com

2. Meta-reasoning for pipeline optimization

3. RL for Dialogue Systems

4. Next-gen Web Crawler for Bing
   http://www.pnas.org/content/115/32/8099

5. RL for language generation
Decision Service (https://ds.microsoft.com)

End-to-end RL service on Azure for problems with immediate rewards (contextual bandits)

Significant gains in first and third-party applications

- 26% Lift vs. Editorial
- 40% Lift vs. Editorial
- 30% Revenue-per-click Improvement
Counterfactual dashboard

- RL optimizes decisions online
- Off-policy evaluation and monitoring enable offline analysis

System’s actual online performance

Offline estimate of a baseline’s performance
Our Mission

We conduct ground-breaking research in Reinforcement Learning (RL) to drive real-world AI scenarios.
RL in the Physical World

1. AI for Autonomous Soaring
   https://www.microsoft.com/en-us/research/project/project-frigatebird-ai-for-autonomous-soaring/

2. Optimal Control for Indoor Agriculture

3. Blind Spots in RL
   Ece Kamar, Eric Horvitz

4. Mobile Social Robotics on $\Psi$ ($\\psi$)

5. Programming-by-Demonstration & RL
Project Frigatebird: AI for Autonomous Soaring

Frigatebirds and some other species can stay aloft for hours with hardly a wing flap, on energy they extract from thin air.

**Goal:** Build AI to let sailplane (a.k.a. glider) UAVs fly long distances fully autonomously without active propulsion, using only soaring.

Andrey Kolobov, Iain Guilliard, Rick Rogahn, Chris Lovett, Debadeepta Dey
How do Soaring Birds and Sailplanes Stay Aloft?

- By exploiting 3D wind patterns
- Wind patterns not directly visible, their locations not known with certainty
- Air movement can be sensed with onboard equipment...
- ...but no 2 windfields are alike – limited generalizability

Research challenges: Learn to identify, exploit, predict, and plan for highly probabilistic atmospheric phenomena from little data
If It Doesn’t Fly (Autonomously), It Doesn’t Count!

- Air fleet for various flight test regimes:
- Each carries GPS, airspeed sensor, etc., & onboard compute for autonomous flight
- Use soaring flight simulators (SilentWings, purpose-built) for sanity checks on the ground
Soaring in Thermals and Beyond

- First step: thermal soaring (RSS-2018, IROS-2018)
  - **Thermal**: irregular column of rising air
  - **Approach**: Bayesian RL done in real time aboard the sailplane
  - **Results**: successful thermal exploitation in real-world flights in adverse conditions

- Current research: vision for long-endurance flight planning in uncertain conditions

- More info, code, and data on Project Frigatebird’s webpage. Come talk to the crew!

**Andrey Kolobov, Iain Guilliard, Rick Rogahn, Chris Lovett, Debadeepta Dey**
Our Mission

We conduct ground-breaking research in Reinforcement Learning (RL) to drive real-world AI scenarios.
RL Foundations

1. Interplay of Optimization-Representation-RL

2. Exploration

3. Social and Ethical Aspects

4. Imitation Learning
   e.g. https://www.microsoft.com/en-us/research/publication/learning-gather-information-via-imitation/
SBEED: Convergent RL w/ Function Approximation

SBEED (out of the Deadly Triad):
Convergent Reinforcement Learning with Nonlinear Function Approximation

Bo Dai\textsuperscript{1→2}, Albert Shaw\textsuperscript{1}, Lihong Li\textsuperscript{2}, Lin Xiao\textsuperscript{3}, Niao He\textsuperscript{4}, Zhen Liu\textsuperscript{1}, Jianshu Chen\textsuperscript{5}, Le Song\textsuperscript{1}

\textsuperscript{1}Gatech, \textsuperscript{2}Google Brain, \textsuperscript{3}Microsoft Research, \textsuperscript{4}UIUC, \textsuperscript{5}Tencent AI


Stability/Convergence of RL algorithms

- Impressive empirical success of DeepRL, but,
- No convergence guarantees, often diverges!
- Limited theory and algorithms (e.g. linear)
- Major Open Problem for decades
Smoothed Bellman Error EmbeDing (SBEED)

• First provably convergent ADP/RL algorithm with general nonlinear function approximation

• Tackle RL problems by directly solving the Bellman equation

\[
\min_V \mathbb{E}_s \left[ (V(s) - \max_a (R(s, a) + \gamma \mathbb{E}_{s' | s,a}[V(s')])^2 \right]
\]

• An obvious attempt, but has two difficulties
  • #1: max operator is nonsmooth (hard for analysis and unstable in practice)
    ✓ Solution: smoothing using entropy regularization over policy simplex
  • #2: conditional expectation inside square, causes biased stochastic gradient
    ✓ Solution: primal-dual lifting into minimax problem using Fenchel conjugate
Our Mission

*We conduct ground-breaking research in Reinforcement Learning (RL) to drive real-world AI scenarios.*
Questions?
Appendix - Simulation


September 2018
Grounded Visual Navigation via Imitation Learning

- Long-horizon sequential decision-making.
- Sensing only via vision.
- Photo-realistic real-world indoor datasets (Matterport3D).
- Can setup dialog with human for assistance.
- Requires common-sense reasoning.
- Test-bed for imitation/reinforcement learning.
- Sim-to-real transfer to real world robots.

Agent executes and decides when to ask for help

Khanh Nguyen, Debadeepta Dey, Chris Brockett, Vignesh Shiv, Bill Dolan
RL in Text-based Adventure Games

• Intersection of RL & NLP
• Agents with language understanding
• Commonsense reasoning
• Map building & Memory
Appendix - Digital


September 2018
Bing’s index is a storage of Web content. Web pages change & need to be recrawled to keep Bing’s index fresh.

How do we compute in near-linear time a crawl scheduling policy that maximize Bing index’s content freshness...

• ... while observing web hosts’ and Bing’s own constraints on crawl bandwidth ...
• ... and learning to predict Web page changes ...
• ... over billions of hosts and 100s of billions of pages?
Meta-Reasoning for Pipeline Optimization

Timing & quality tradeoffs, uncertainties with modular pipelines

Loss = f(latency, accuracy)

- Sequential decision-making
- Cost of meta-reasoning should be low
- Real-world RL problem

Debadeepta Dey, Dan Bohus, John Langford, Aditya Modi, Besmira Nushi, Alekh Agarwal, Adith Swaminathan, Sean Andrist, Eric Horvitz
Appendix - Physical

Project Sonoma: Optimal Control for Indoor Farms

- Learn a policy that can optimally control plant growth in indoor farms
- Real-world application that requires advances in model-based RL, transfer RL, POMDP solvers

Kenneth Tran, Ranveer Chandra, Chetan Bansal + external collaborators
**Goal:** Create model of blind spots

**Blind spot:** Systematic input regions with divergence from optimal policy

Complicated by incomplete state representations
$\psi$: Assistive, Mobile, Social Robotics with $\psi$
Programming-by-demonstration and RL

- PbD provides the initial solution
- RI refines the solution