Helping Unlock the Potential of RL

Emma Brunskill
Carnegie Mellon / Stanford University
If you invent a breakthrough in AI, so that machines can learn, that is worth 10 Microsofts
– Bill Gates

If you can invent an AI that helps us all learn to be as motivated, knowledgeable & intelligent as Bill Gates, that is worth 7.1 Billion Microsofts
– me
(Machine) Learning to Improve Learning
From here ....

to education, healthcare...
Towards Unlocking the Potential of RL

• Practitioners in high stakes domains often wary of deploying new policy without some guarantees
• Current system may be insufficient to achieve desired results
Power of Counterfactuals: Use Data from Behavior Policy to Estimate Another Policy’s Performance $E[\Sigma r_i]$
We used off-policy evaluation to find a policy with 30% higher engagement (Mandel et al. 2014)
Key Challenge: Distribution Mismatch

- Rewards $r(s,a,z)$
- Policy maps history $(a,z,r,a',z',...)$ $\rightarrow$ $a$
Prior Work: Estimate Model from Data

**Strengths**
- Low variance estimator of policy performance

**Weaknesses**
- Not unbiased estimator (model may be poor!)
- Not consistent estimator

Historical Data
- History$_1$, $R_1$ = $\sum_i r_{i1}$
- History$_2$, $R_2$ = $\sum_i r_{i2}$
- ...

Build & Estimate a Model
- Define states and actions
- Dynamics $p(s'|s,a)$
- Observation model $p(z|s,a)$
- Rewards $r(s,a,z)$
Prior Work: Importance Sampling (IS)  
(e.g. Precup et al. 2002, Mandel et al. 2014)

**Historical Data**

History₁, \( R₁ = \Sigma_i r_{i₁} \)

History₂, \( R₂ = \Sigma_i r_{i₂} \)

...  

**Estimate of Behavior Policy Performance**  
\[\pi_b \text{ Performance} = \frac{1}{N} \sum_{j=1}^{N} R_j\]

**Estimate of Evaluation Policy Performance**  
\[\pi_e \text{ Performance} = \frac{1}{N} \sum_{j=1}^{N} \frac{p(\text{history}_j | e)}{p(\text{history}_j | b)} R_j\]

\[= \frac{1}{N} \sum_{j=1}^{N} R_j \sum_{t=1}^{H} \frac{p(z_{t+1} | \text{history}_{j,1:t-1}, a_t)p(a_t | e, \text{history}_{j,1:t-1})}{p(z_{t+1} | \text{history}_{j,1:t-1}, a_t)p(a_t | b, \text{history}_{j,1:t-1})}\]

\[= \frac{1}{N} \sum_{j=1}^{N} R_j \sum_{t=1}^{H} \frac{p(a_t | e, \text{history}_{j,1:t-1})}{p(a_t | b, \text{history}_{j,1:t-1})}\]
Prior Work: Importance Sampling (IS)  
(e.g. Precup et al. 2002, Mandel et al. 2014)

**Historical Data**

\[ \text{History}_1, R_1 = \sum_i r_{i1} \]
\[ \text{History}_2, R_2 = \sum_i r_{i2} \]
\[ \ldots \]

\[ \left\{ \begin{array}{l}
\text{Estimated Evaluation Policy} \\
\pi_e \text{ performance} \\
\frac{1}{N} \sum_{j=1}^{N} \frac{p(\text{history}_j | e)}{p(\text{history}_j | b)} R_j
\end{array} \right. \]

**Strengths**

- Unbiased estimator of policy performance
- Strongly consistent estimator of policy perform.

**Weaknesses**

- High variance estimator
Prior Work: Doubly Robust
Dudik et al. 2011, Jiang & Li 2015

• Model + IS-based estimator
• Limitation: for RL made unbiased estimator
Tight Estimate Often Better than Unbiased: Measure with Mean Squared Error

Thomas and B, ICML 2016
Insight: Blend IS-Based & Model Based Estimators to Directly Min Mean Squared Error

Thomas and B, ICML 2016
MSE of Estimated Policy Value

Estimated policy value using particular weighting of model estimate and importance sampling estimate

\[
\text{MSE}(x) = \text{Bias}\left( \sum_{j=0}^{\infty} x_j g^{(j)}(\pi_e | D) \right)^2 + \text{Var}\left( \sum_{j=0}^{\infty} x_j g^{(j)}(\pi_e | D) \right)
\]

Thomas and B, ICML 2016
Estimated Bias & Covariance

- Estimated covariance: use sample covariance
  \[ \Omega_{i,j} = \text{Cov}\left(g^{(i)}(\pi_e|D), g^{(j)}(\pi_e|D)\right) \]

- Estimated bias:
  - May be as hard as estimating true policy value

\[ g^{(j)}(\pi_e|D) \]
Estimated policy value w/ jth weighting of model estimate & importance sampling estimate

Thomas and B, ICML 2016
Quadratic Program Using Estimated Bias and Variance

\[ \text{MSE}(x) = x^T(\Omega + bb^T)x \]

\[ \Omega_{i,j} = \text{Cov} \left( g^{(i)}(\pi_e | D), g^{(j)}(\pi_e | D) \right) \]

\[ b_i = \mathbb{E} \left[ g^{(i)}(\pi_e | D) - \rho(\pi_e) \right] \]
MAGIC Estimator

- **Model And Guided Importance sampling Combining estimator**
- (Purposefully) not necessarily unbiased
- MAGIC is strongly consistent*

Thomas and B, ICML 2016
MAGIC Can Yield Orders of Magnitude Better Estimates of Policy Performance

Log scale!

Thomas and B, ICML 2016
MAGICAL Policy Search

Thomas and B, EWRL (yesterday!)
Let's name fractions using number lines!

Brittany bought a watermelon to share with three of her friends. Each of the watermelon pieces were equal-sized. Brittany ate 1/4 of the watermelon. Use the number line to show how much of the watermelon Brittany ate.

- Importance sampling estimate meaningless
- Work in progress
Summary: Better Off-policy Policy Evaluation

• Estimate alternate policies’ performance from past data
  – Before deploying in important / costly settings
  – Useful online to enable smarter, faster exploration

• MAGIC powerful new off-policy policy evaluation estimator
  – Minimize MSE
  – Strongly consistent
  – Sometimes order of magnitude lower mean squared error estimator
Decisions Made Have Real Impact
Rules of the Game Are Not Fixed
DESCRIPTIONS AND HISTOGRAMS (1/3 points)

The price of airline tickets varies over time. The following is a histogram that could describe the distribution of airplane ticket prices. Select the best option for each of the questions below.

The x-axis should be labeled as

- Time
- Ticket Price
- Frequency [selected]
- Distribution
During Optimization Tutoring System
Stopped Teaching Some Histogram Skills
During Optimization Tutoring System Stopped Teaching Some Histogram Skills

- No improvement in post test → system had learned that some of our content was inadequate so best thing was to skip it!
- Content (action space) insufficient to achieve goals
Humans are Invention Machines

New actions

New sensors
Invention Machines: Creating Systems that Can Evolve Beyond Their Original Capacity

New actions                         New sensors
Problem Formulation

• Maximize expected reward
• Online reinforcement learning
• Directed action invention
  – Where (which states) should we add actions at?
Related Work

• Policy advice / learning from demonstration
• Changing action spaces
  – Almost all work is reactive, not active solicitation

Mandel, Liu, Brunskil & Popovic, AAAI 2017
Online reinforcement learning

Active Domain (Action Space) Adaptation

Environment

Actions

Agent

State selector

Reinforcement Learning Algorithm

Outcomes

State Queries

New Actions

Human

Mandel, Liu, Brunskil & Popovic, AAAI 2017
Requesting New Actions

\[ \arg \max_s \sum_{s_0 \in S_0} V_{A \cup a_n}(s_0)p(s_0) \]

Current action set

New action
Expected Local Improvement

$$\arg\max_s \int_a p_s(a_h)(V_{A \cup a_h}(s) - V_A(s)) da_h$$

Prob. human gives you action $a_h$ for state $s$

Improvement in value at state $s$ if add in action $a_h$
\[
ELI(s) = \int_a p_s(a_h)(V_{A\cup a_h}(s) - V_A(s))da_h \\
\leq \int_{a:V_{A\cup a_h}(s) > V_A(s)} p_s(a_h)(V_{A\cup a_h}(s) - V_A(s))da_h \\
\leq (V_{max} - V_A(s)) \int_{a:V_{A\cup a_h}(s) > V_A(s)} p_s(a_h)da_h
\]

- \(V(s)\) given current action set
- Probability get a new action that will increase \(V(s)\)

Unknown!
What to Use for $V_A(s)$

$$(V_{max} - V_A(s)) \int_{a:V_{A∪a_h}(s) > V_A(s)} p_s(a_h) da_h$$

- Be optimistic (MBIE, Rmax, ...)
- Why?
  - Don’t need to add in new actions if current action set might yield optimal behavior
  - Avoids focusing on highly unlikely states

Mandel, Liu, Brunskil & Popovic, AAAI 2017
Probability of Getting a Better Action

\[
(V_{\text{max}} - V_A(s)) \int_{a : V_A \cup a_h(s) > V_A(s)} p_s(a_h) da_h
\]

• Don’t want to ask for actions at same state forever (maybe no improvement possible)
• Model prob of a better action as \( Beta(1, |A_{s,\ell}| + 1) \)
• Chance of better action decays w/ # of actions

\[
ELI(s) = \frac{1}{|A_{s,\ell}| + 2} (V_{\text{max}} - V_A(s))
\]

Mandel, Liu, Brunskil & Popovic, AAAI 2017
Simulations

• Large action task* (Sallans & Hinton 2004)
  – 13 states
  – 273 outcomes (next possible states per state)
  – $2^{20}$ actions per state
• At start each $s$ has single $a$ (like default $\pi$)
• Every 20 steps can request an action
  – Sample action at random from action set for $s$
  – Compare ELI vs Random $s$ vs High freq $s$
*With best choice of algorithm for choosing current value

Mandel, Liu, Brunskil & Popovic, AAAI 2017
Mostly Bad Human Input

![Graph showing cumulative reward over episodes for different conditions: No addition, ELI-NoLearn, Random, ELI.](image-url)

Mandel, Liu, Brunskil & Popovic, AAAI 2017
Chrissy loves exploring outdoors. Yesterday, she saw a herd of 12 elk being chased by a pack of 8 wolves. How many animals in total did Chrissy see while she was exploring?

- New actions = new hints
- Learning where to ask for new hints
Invention Machines: Creating Systems that Can Evolve Beyond Their Original Capacity

- Arm-acquiring multi-armed bandits (in submission w/Mandel, Liu & Popovic)
- Expected local improvement for action selection in RL (AAAI 2017)
Towards Unlocking the Potential of RL

- Practicioners in high stakes domains often wary of deploying new policy without some guarantees
  → MAGIC policy evaluation
- Current system may be insufficient to achieve desired results
  → Inventing new actions