Contextual Reinforcement Learning

John Langford
Microsoft Research
(In collaboration with many)
$E^3$ Theorem: $\text{Poly}(s, a, \frac{1}{\epsilon}, H)$ samples for good policy.
Let’s solve some problems!

Andrew Moore’s complaint: Exponentially worse than Supervised!
What is a *USEFUL* foundational theory for Reinforcement Learning?
Approach#1: Contextual Bandits

Goal: maximize sum of rewards.
How about **news**?

**Repeatedly:**

1. Observe features of user+articles
2. Choose a news article.
3. Observe click-or-not

**Goal:** Maximize fraction of clicks
A standard pipeline

1. Collect \((user, article, click)\) information.
2. Build \(features(user, article)\)
3. Learn \(\hat{P}(click|features(user, article))\)
4. Act: \(\arg \max_{\{articles\}} \hat{P}(click|features(user, article))\)
5. Deploy in A/B test for 2 weeks
6. A/B test fails 😞 Why?
Q: What goes wrong?

A: Need Right Signal for Right Answer

Is Ukraine interesting to John?
Ex: Which advice?

Repeatedly:
1. Observe features of user+advice
2. Choose an advice.
3. Observe steps walked

Goal: Healthy behaviors
Real-world Applications

News Rec: [LCLS ‘10]
Ad Choice: [BPQCCPRSS ‘12]
Ad Format: [TRSA ‘13]
Education: [MLLBP ‘14]
Music Rec: [WWHW ‘14]
Robotics: [PG ‘16]
Wellness/Health: [ZKZ ‘09, SLLSPM ‘11, NSTWCSM ‘14, PGCRRH ‘14, NHS ‘15, KHSBATM ‘15, HFKMTY ‘16]
Exploration

Policy

[Space Station Image]
Exploration

Randomization
Exploration

Randomization

Policy
Inverse Propensity Score (IPS) [HT ‘52]

Given experience \( \{(x, a, p, r)\} \) and a policy \( \pi: x \rightarrow a \), how good is \( \pi \)?

\[
V_{\text{IPS}}(\pi) = \frac{1}{n} \sum_{(x, a, p, r)} \frac{r I(\pi(x) = a)}{p}
\]
What do we know about IPS?

**Theorem:** For all $\pi$, for all $D(x, \vec{r})$

$$E \left[ r_{\pi(x)} \right] = E[V_{\text{IPS}}(\pi)] = E \left[ \frac{1}{n} \sum_{(x,a,p,r)} \frac{r I(\pi(x)=a)}{p} \right]$$

**Proof:** For all $(x, \vec{r})$, $E_{a \sim \vec{p}} \left[ \frac{r_a I(\pi(x)=a)}{p_a} \right]$

$$= \sum_a p_a \frac{r_a I(\pi(x)=a)}{p_a}$$

$$= r_{\pi(x)}$$
Reward over time

- Offline estimate of system’s performance
- Offline estimate of baseline’s performance
- System’s actual online performance
Better Evaluation Techniques

Double Robust: [DLL ‘11]

Weighted IPS: [K ’92, SJ ‘15]

Clipping: [BL ’08]
Learning from Exploration [‘Z 03]

Given Data \{ (x, a, p, r) \} how to maximize \( E[r_{\pi(x)}] \)?
Maximize \( E[V_{IPS}(\pi)] \) instead!

\[
r_a = \begin{cases} 
  r/p & \text{if } \pi(x) = a \\
  0 & \text{otherwise}
\end{cases}
\]

Equivalent to:

\[
r'_a = \begin{cases} 
  1 & \text{if } \pi(x) = a \\
  0 & \text{otherwise}
\end{cases}
\]

with importance weight \( \frac{r}{p} \)

Importance weighted multiclass classification!
Better Learning from Exploration Data

**Policy Gradient:** [W ’92]

**Offset Tree:** [BL ’09]

**Double Robust for learning:** [DLL ’11]

**Multitask Regression:** Unpublished, but in Vowpal Wabbit

**Weighted IPS for learning:** [SJ ’15]
Evaluating Online Learning

Problem: How do you evaluate an online learning algorithm Offline?

Answer: Use Progressive Validation [BKL ’99, CCG ‘04]

Theorem:
1) Expected PV value = Uniform expected policy value.
2) Trust like a **test set error**.
How do you do Exploration?

Simplest Algorithm: $\epsilon$-greedy.

With probability $\epsilon$ act uniform random

With probability $1 - \epsilon$ act greedily
Better Exploration Algorithms

Better algorithms maintain ensemble and explore amongst actions of this ensemble.

**Thompson Sampling**: [T ‘33]

**EXP4**: [ACFS ‘02]

**Epoch Greedy**: [LZ ‘07]

**Polytime**: [DHKKLRZ ‘11]

**Cover&Bag**: [AHKLLS ‘14]

**Bootstrap**: [EK ‘14]
Vowpal Wabbit: Online/Fast learning

- BSD License, 10 year project
- Mailing List>500, Github>1K forks, >4K stars, >1K issues, >100 contributors
- Command Line/C++/C#/Python/Java/AzureML/Daemon
Decision Service [ABCHLMLMORSS ‘16]

- Open-source on Github
- Host and manage yourself

- Hosted as a Microsoft Cognitive Service
- Logging and model deployment managed
- Data logged to your Azure account

- Contextual bandits optimize decisions online
- Off-policy evaluation and monitoring
Approach #2: Policy Improvement
Approach #2: Policy Improvement
Approach #2: Policy Improvement

Goal: Compete with or improve rewards over policy $\pi$. Applications: Structured/Joint prediction, Sparse RL
Approach #2: Policy Improvement

Theorem (Policy Gradient): Policy improves greedily.
Can we compete with (and improve on) an existing policy?

CPI [KL ‘02] /PSDP [BKNS ‘03]: Compete with best policy given known state distribution.

SEARN [DLM ‘09]: Compete with train-time known policy.

Dagger[RGB ‘11]/Aggravate [RB ‘14]: Via dataset aggregation.

LOLS[CKADL ‘15]: Improve on train-time known policy sometimes.
Approach#3: Global Reinforcement Learning

- Action $a$
- Context $x$
- Reward $r$
- Policy $\pi$
Approach#3: Global Reinforcement Learning

Diagram:
- Action \( a \)
- Context \( x \)
- Reward \( r \)
- Policy \( \pi \)
Approach#3: Global Reinforcement Learning

Goal: maximize sum of rewards.

Applications:
OLIVE: Optimism Led Iterative Value Elimination

Given: Set of value functions \( F = \{ f : X \times A \rightarrow (-\infty, \infty) \} \)

Repeatedly:

1. Pick most optimistic \( f \) at \( h = 1 \)
2. If (predicted value = real value) then return \( f \)
3. Else find horizon \( h \) of large disagreement
   - Act randomly at \( h \)
   - Eliminate all \( f \) with a large bellman error at \( h \)

Theorem: OLIVE needs \( O(B, H, |A|, \log|F|, \frac{1}{\epsilon}) \) samples
Where do we stand?

- MDPs are a poor foundation.
- Contextual Bandits: Use this first!
- Policy Improvement: Trickier but doable
- Contextual Decision Process: Maybe...
Much still to do---Join us 😊

Alekh Agarwal
Alina Beygelzimer
Drew Bagnell
Avrim Blum
Kai-Wei Chang
Christoph Dann
Hal Daume
Geoff Gordon
Daniel Hsu

Nan Jiang
Sham Kakade
Satyen Kale
Nikos Karampatziakis
Akshay Krishnamurthy
Lihong Li
Stephane Ross
Robert Schapire
Tong Zhang
Contextual Bandits(ish) Applications


Education: Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, Zoran Popovic, Offline policy evaluation across representations with applications to educational games. AAMAS 2014: 1077-1084.


Wellness Contextual Bandits Work


Evaluation References


Clipping: Oliver Bembom and Mark J. van der Laan: Data-adaptive selection of the truncation level for inverse probability of treatment-weighted estimators. 2008
Learning from Exploration References

Multitask Regression: Unpublished, but in Vowpal Wabbit.


Exploration Algorithm References


**Cover & Bag**: A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, R. Schapire, Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits, ICML 2014.

**Bootstrap**: D. Eckles and M. Kaptein, Thompson Sampling with Online Bootstrap, arxiv.org/1410.4009