The Contextual Reinforcement Learning Research Program

John Langford
Microsoft Research
(In collaboration with many!)
In 2006, I stopped working on traditional RL.
Traditional RL had become stale

1. Q functions can represent credit assignment.
3. MDP Sample complexity (Kearns&Singh 1998)
4. ??
In 2007, Contextual Bandits started

The Epoch-Greedy Algorithm for Contextual Multi-armed Bandits

John Langford
Yahoo! Research
jl@yahoo-inc.com

Tong Zhang
Department of Statistics
Rutgers University
tongz@rci.rutgers.edu

Abstract

We present Epoch-Greedy, an algorithm for contextual multi-armed bandits (also known as bandits with side information). Epoch-Greedy has the following properties:
What are Contextual Bandits?

Repeatedly:
1. See features $x$
2. Choose actions $a$ in $A$
3. See reward $r$ for action $a$ in context $x$
Goal: maximize sum of rewards.
Why Not Contextual Bandits?

Eh... No credit assignment, easy exploration.

Why Contextual Bandits?

1. Supervised Learning: $\forall$ classifiers $\forall$ data sources: good performance
2. Contextual Bandits: Can we get the same?
3. Contextual RL: Can we get there?
CBs: Actually started in 1995!

The non-stochastic multi-armed bandit problem*

Peter Auer
Institute for Theoretical Computer Science
Graz University of Technology
A-8010 Graz (Austria)
pauer@igi.tu-graz.ac.at

Nicolò Cesa-Bianchi
Department of Computer Science
Università di Milano
I-20135 Milano (Italy)
cesabian@dsi.unimi.it

Yoav Freund
AT&T Labs
180 Park Avenue
Florham Park, NJ 07932-0971
{yoav, schapire}@research.att.com

∀ classifiers ∀ data sources $O\left(\left(\frac{|A| \log |\Pi|}{T}\right)^{0.5}\right)$ regret
Q: How do you make the computation work?
A: Use reduction to Supervised Learning
Can it actually work in practice?

But What about Reinforcement Learning?

Imitation Learning is another plausible island of consistent tractability.
But what about REAL Reinforcement Learning?

PAC Reinforcement Learning with Rich Observations
Akshay Krishnamurthy ¹, Alekh Agarwal ², and John Langford ²
¹University of Massachusetts, Amherst, Amherst, MA 01003
²Microsoft Research, New York, NY 10011

Contextual Decision Processes with Low Bellman Rank are PAC-Learnable
Nan Jiang ² Akshay Krishnamurthy ¹ Alekh Agarwal ¹
nanjiang@umich.edu akshay@cs.umass.edu alekha@microsoft.com
John Langford ² Robert E. Schapire ¹
jcl@microsoft.com schapire@microsoft.com
Contextual Decision Processes

Repeatedly:

For $h = 1$ to $H$

1. See features $x$
2. Choose actions $a$ in $A$
3. See reward $r$ for action $a$ in context $x$ and history $h$

Goal: maximize sum of rewards.
OLIVE: Optimism Led Iterative Value Elimination

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

Repeatedly:

Pick most optimistic \( f \) at \( h = 1 \)

Rollout with \( f \) repeatedly

If (predicted value = real value) then return \( f \)

Else find horizon \( h \) of large disagreement

Rollout with \( f \) except acting randomly at \( h \)

Eliminate all \( f \) with a large bellman error at \( h \)
Bellman Rank = new general notion of tractability

<table>
<thead>
<tr>
<th>Model</th>
<th>tabular MDP</th>
<th>low-rank MDP</th>
<th>reactive POMDP</th>
<th>reactive PSR</th>
<th>LQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bellman rank</td>
<td># states</td>
<td>rank</td>
<td># hidden states</td>
<td>PSR rank</td>
<td># state variables</td>
</tr>
<tr>
<td>PAC Learning</td>
<td>known</td>
<td>new</td>
<td>extended</td>
<td>new</td>
<td>known$^3$</td>
</tr>
</tbody>
</table>

Theorem: $\forall \text{ CDPs}, \forall \text{ self-consistent } F \text{ with Bellman rank } B \text{ with probability } 1 - \delta$, OLIVE requires:

$$\tilde{O}\left(\frac{B^2 H^3 |A| \log \frac{|F|}{\delta}}{\epsilon^2}\right)$$

trajectories to find an $\epsilon$ optimal $f$. 
My History of RL Foundations

1. Q functions can represent credit assignment.
2. Asymptotically valid update rules (Watkins ‘89, Williams ‘92)
3. Contextual Bandits first results (ACFS 1995)
4. MDP Sample complexity (Kearns&Singh 1998)
5. Efficient Contextual Bandit Learning (DHKKLRZ 2011)
6. Imitation w/ Reinforcement (Ross&Bagnell ‘14, CKADL ‘15)
7. Deployable Contextual Bandit System (ABCHLLLMORSS 2016)
8. Contextual Decision Process first results (KAL, JKALS 2016)
9. ... Join us