

# The Contextual Reinforcement Learning Research Program

John Langford

Microsoft Research

(In collaboration with many!)



# In 2006, I stopped working on traditional RL.

---

## PAC Model-Free Reinforcement Learning

---

**Alexander L. Strehl**

STREHL@CS.RUTGERS.EDU

**Lihong Li**

LIHONG@CS.RUTGERS.EDU

Department of Computer Science, Rutgers University, Piscataway, NJ 08854 USA

**Eric Wiewiora**

EWIEWIOR@CS.UCSD.EDU

Computer Science and Engineering Department University of California, San Diego

**John Langford**

JL@HUNCH.NET

TTI-Chicago, 1427 E 60th Street, Chicago, IL 60637 USA

**Michael L. Littman**

MLITTMAN@CS.RUTGERS.EDU

Department of Computer Science, Rutgers University, Piscataway, NJ 08854 USA

# Traditional RL had become stale

1. Q functions can represent credit assignment.
2. Asymptotically valid update rules (Watkins 1989, Williams 1992)
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: **∀ classifiers ∀ 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    Robert E. Schapire**

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

$\forall$  classifiers  $\forall$  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

---

Efficient Optimal Learning for Contextual Bandits

---

2011

Miroslav Dudik  
mdudik@yahoo-inc.com

Daniel Hsu  
djhsu@rci.rutgers.edu

Satyen Kale  
skale@yahoo-inc.com

Nikos Karampatziakis  
nk@cs.cornell.edu

John Langford  
jl@yahoo-inc.com

Lev Reyzin  
lreyzin@cc.gatech.edu

Tong Zhang  
tzhang@stat.rutgers.edu

Taming the Monster:

A Fast and Simple Algorithm for Contextual Bandits

2014

Alekh Agarwal<sup>1</sup>, Daniel Hsu<sup>2</sup>, Satyen Kale<sup>3</sup>, John Langford<sup>1</sup>, Lihong Li<sup>1</sup>, and  
Robert E. Schapire<sup>1,4</sup>



# Can it actually work in practice?

## **A Multiworld Testing Decision Service**

2016

Alekh Agarwal    Sarah Bird    Markus Cozowicz    Luong Hoang    John Langford  
Stephen Lee\*    Jiaji Li\*    Dan Melamed    Gal Oshri\*    Oswaldo Ribas\*  
Siddhartha Sen    Alex Slivkins

Microsoft Research, \*Microsoft

Deployable system optimizing *business* metrics.

Open Source, cloud based.

<http://aka.ms/mwt> for more

# But What about Reinforcement Learning?

---

## Learning to Search Better than Your Teacher

---

**Kai-Wei Chang**

KCHANG10@ILLINOIS.EDU

University of Illinois at Urbana Champaign, IL

**Akshay Krishnamurthy**

AKSHAYKR@CS.CMU.EDU

Carnegie Mellon University, Pittsburgh, PA

**Alekh Agarwal**

ALEKHA@MICROSOFT.COM

Microsoft Research, New York, NY

**Hal Daumé III**

HAL@UMIACS.UMD.EDU

University of Maryland, College Park, MD, USA

**John Langford**

JCL@MICROSOFT.COM

Microsoft Research, New York, NY



**Imitation Learning** is another plausible island of consistent tractability.

# But what about REAL Reinforcement Learning?

## PAC Reinforcement Learning with Rich Observations

NIPS 2016

Akshay Krishnamurthy <sup>\*1</sup>, Alekh Agarwal <sup>†2</sup>, and John Langford <sup>‡2</sup>

<sup>1</sup>University of Massachusetts, Amherst, Amherst, MA 01003

<sup>2</sup>Microsoft Research, New York, NY 10011

## Contextual Decision Processes with Low Bellman Rank are PAC-Learnable

Arxiv 2016

Nan Jiang<sup>‡</sup>      Akshay Krishnamurthy<sup>\*</sup>      Alekh Agarwal<sup>†</sup>  
nanjiang@umich.edu    akshay@cs.umass.edu    alekha@microsoft.com

John Langford<sup>†</sup>      Robert E. Schapire<sup>†</sup>  
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

<i>Model</i>	tabular MDP	low-rank MDP	reactive POMDP	reactive PSR	LQR
<i>Bellman rank</i>	# states	rank	# hidden states	PSR rank	# state variables
<i>PAC Learning</i>	known	new	extended	new	known <sup>3</sup>

Theorem:  $\forall$  CDPs,  $\forall$  self-consistent  $F$  with Bellman rank  $B$  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