

Seminar in Deep RL

Multi Armed Bandits- Contextual Bandits

Orhan Saeedi

20.04.2020

What we saw so far!

k-armed Bandits
(in this case $k=2$)



Exploration
vs.
Exploitation

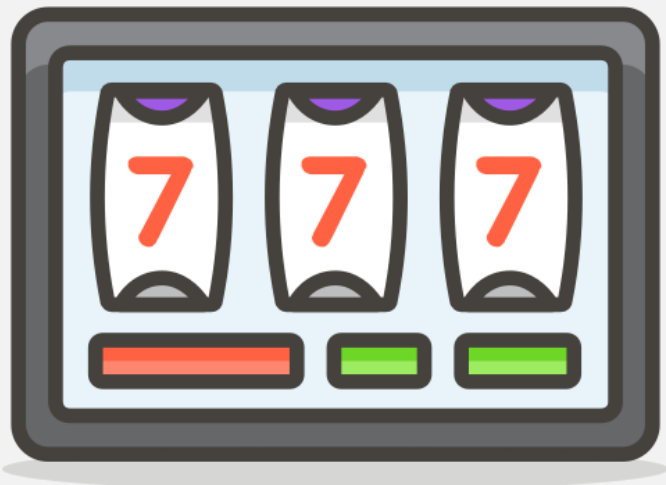


William R. Thompson's
Thompson Sampling

Goal



One-Armed Bandit



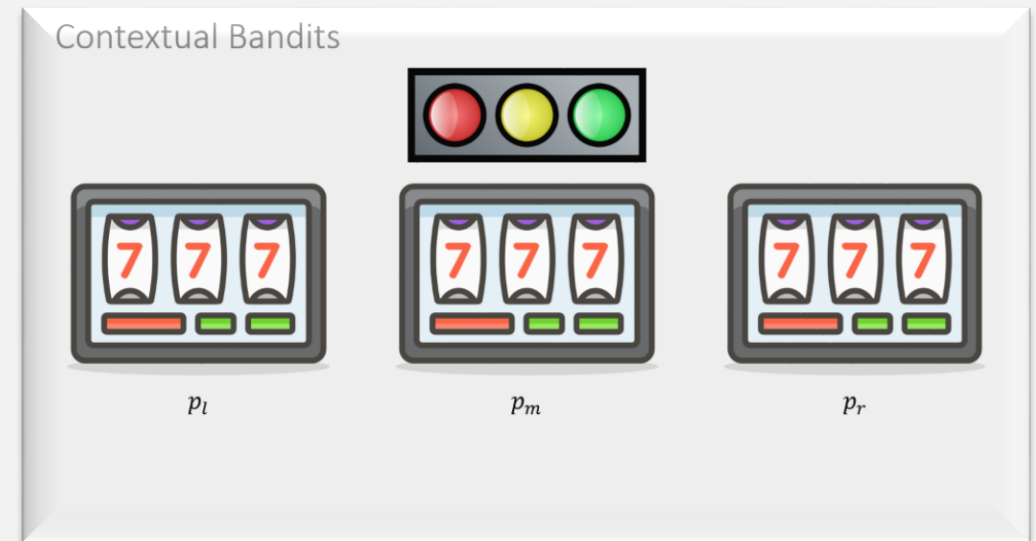
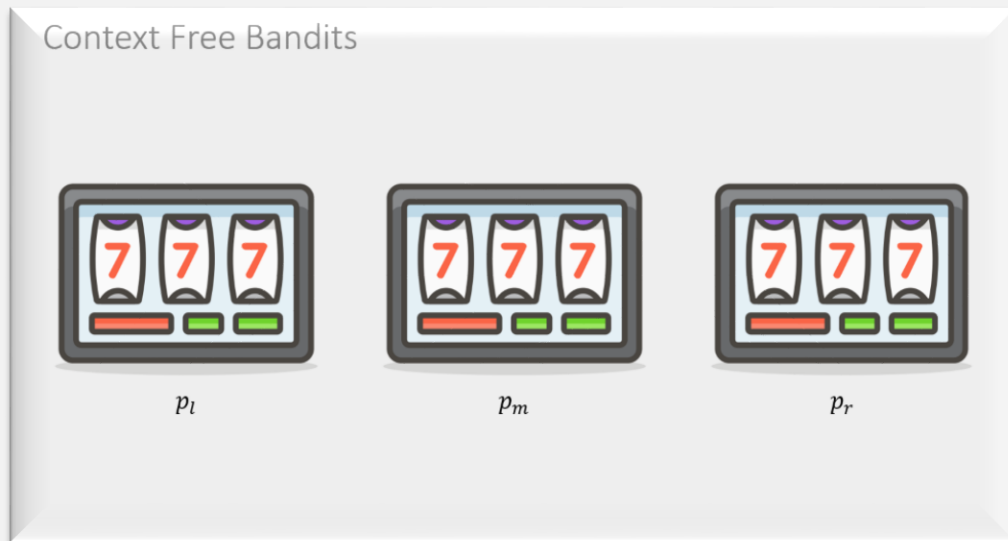
Cost:



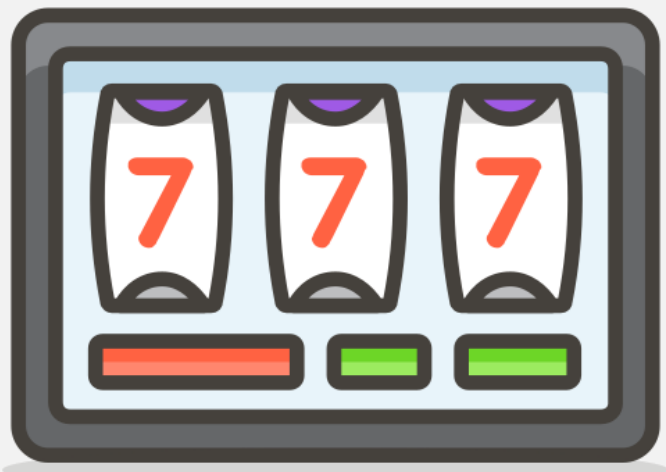
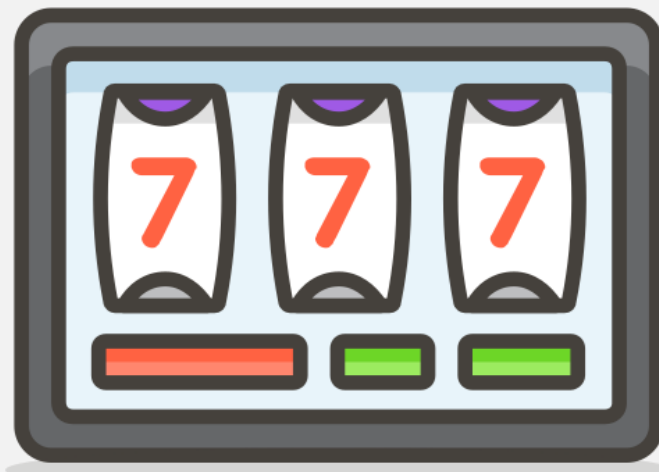
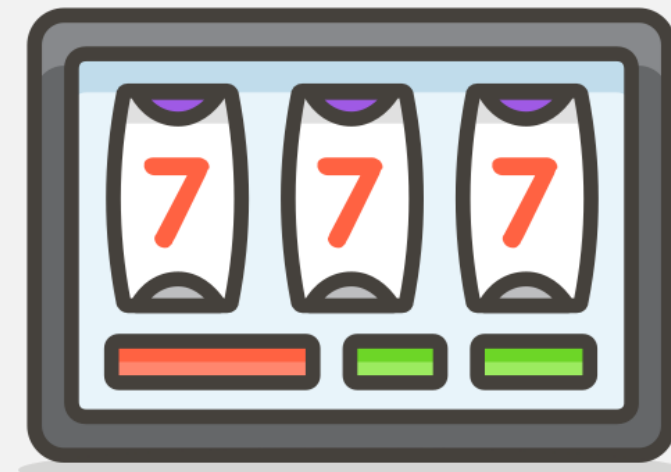
Win:



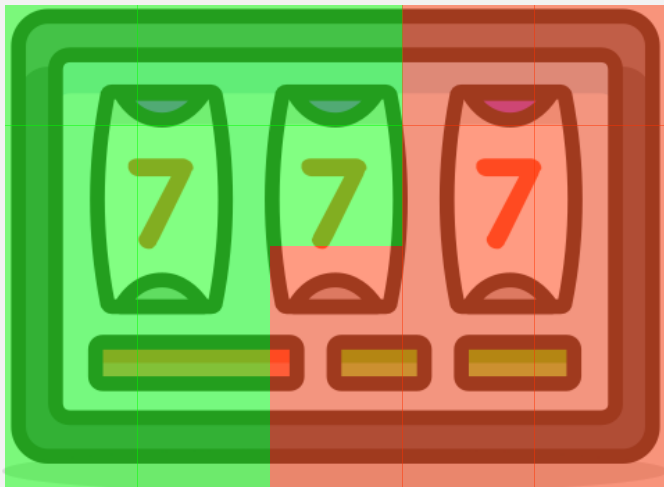
Context Free vs. Context Based/Contextual Bandits



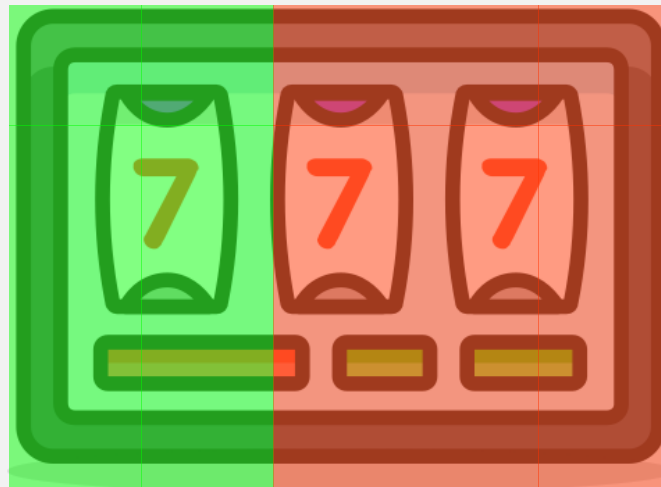
Context Free Bandits

 p_l  p_m  p_r

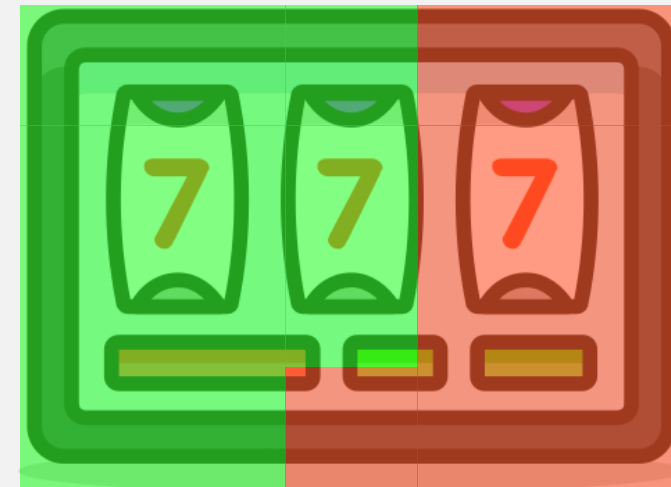
Context Free Bandits



$$p_l = 0.5$$

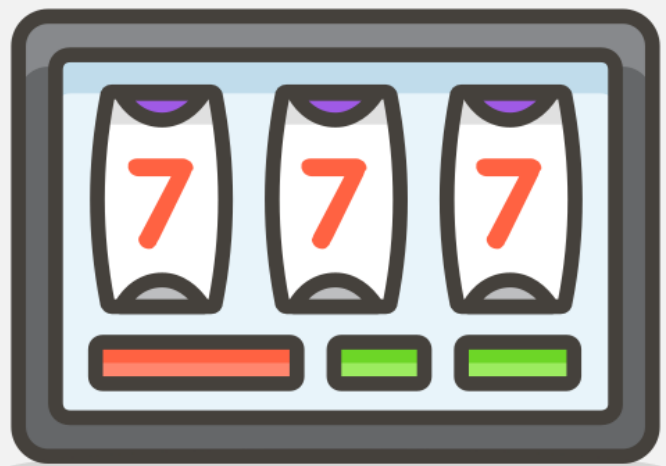
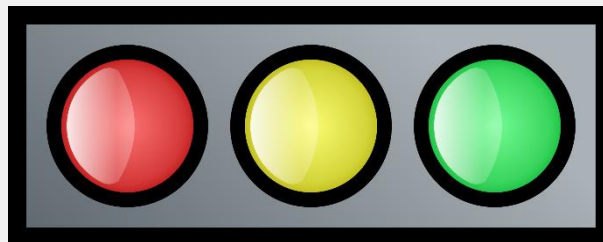
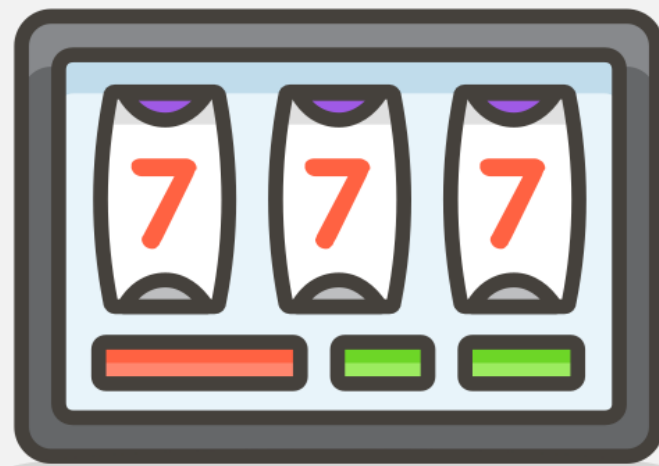
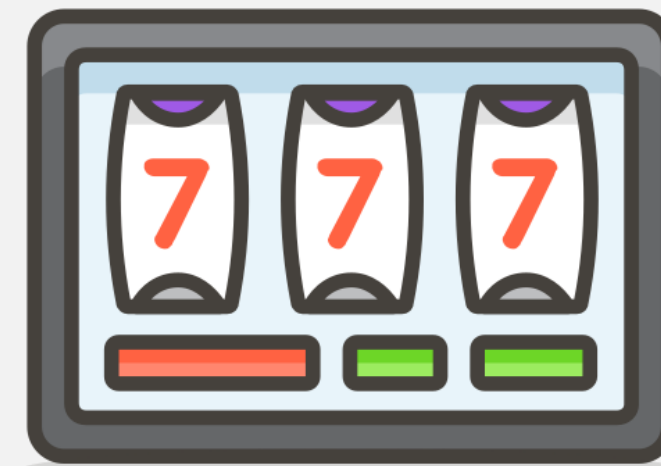


$$p_m = 0.4$$

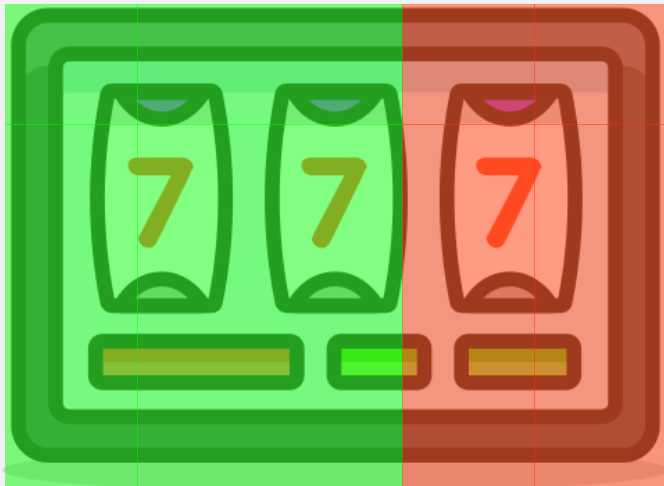
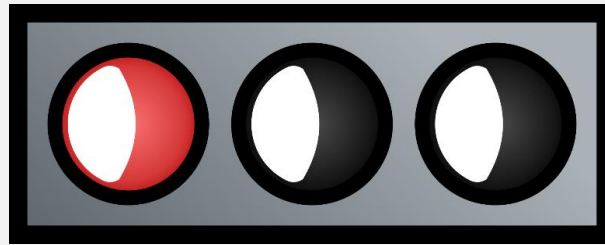


$$p_r = 0.55$$

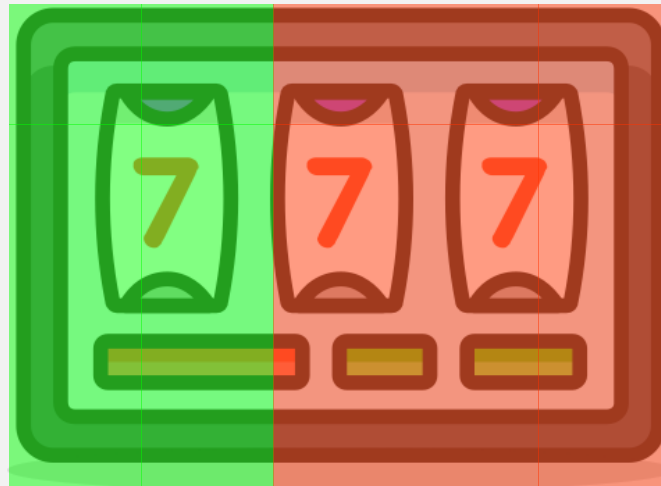
Contextual Bandits

 p_l  p_m  p_r

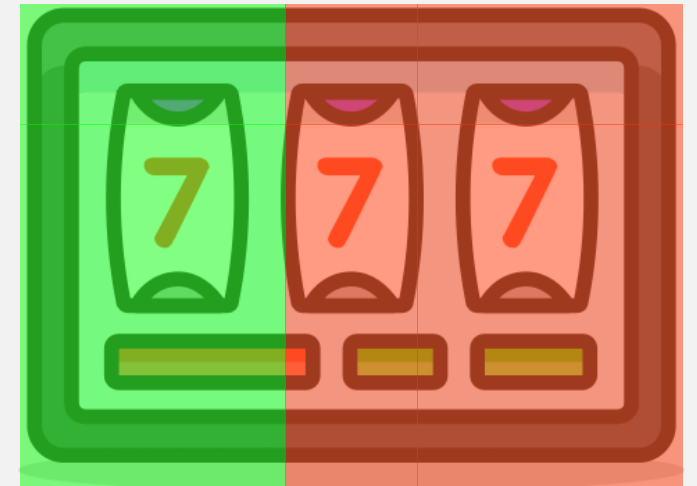
Contextual Bandits



$$p_l = 0.6$$

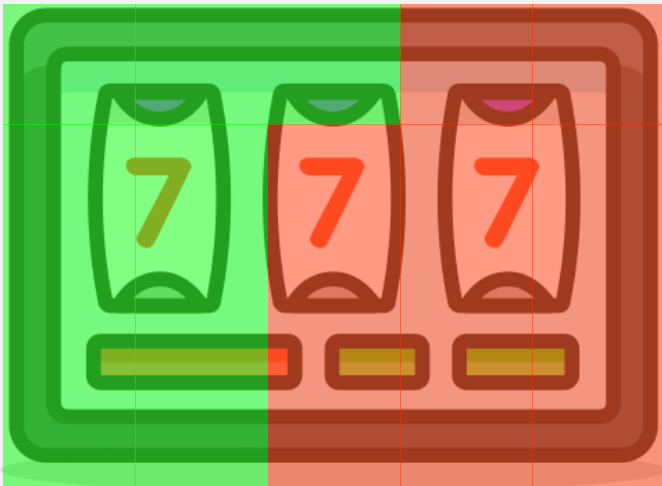
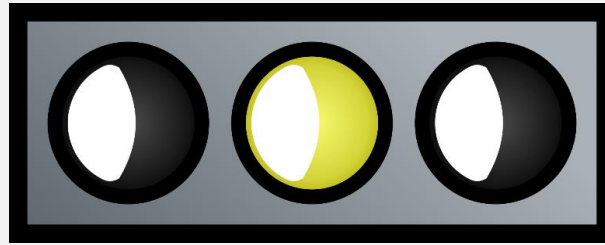


$$p_m = 0.4$$

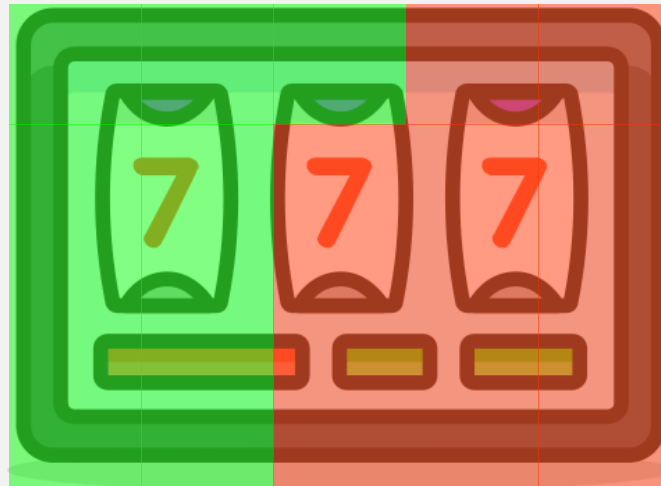


$$p_r = 0.4$$

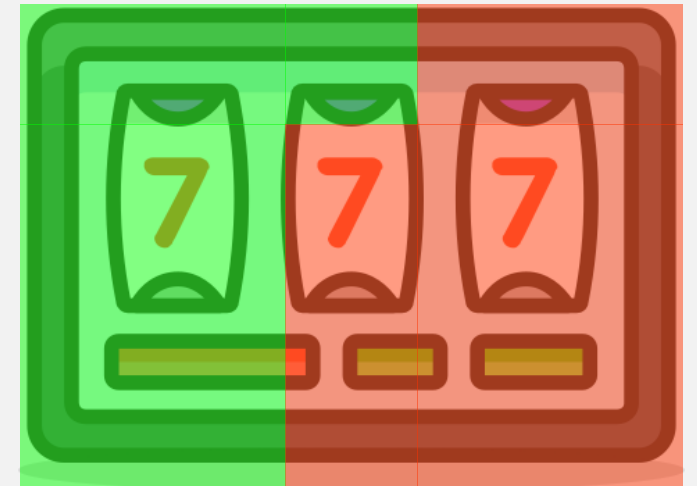
Contextual Bandits



$$p_l = 0.45$$

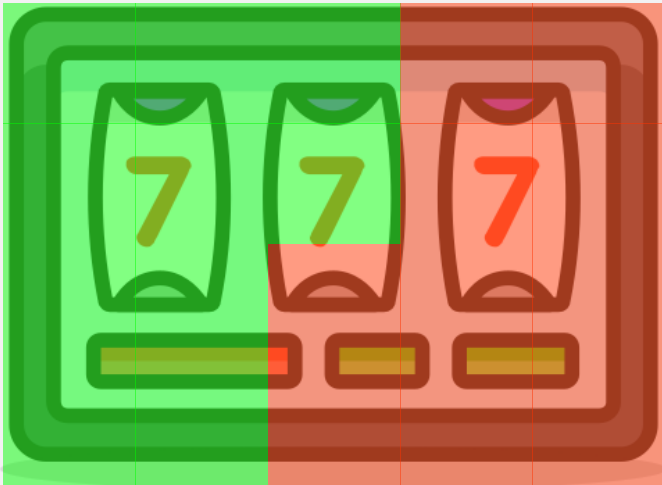
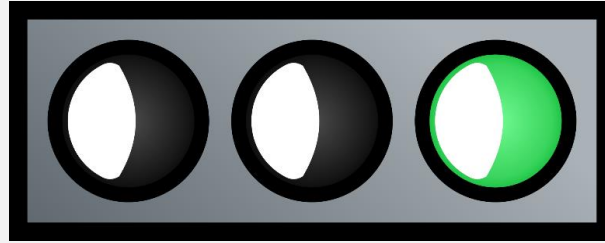


$$p_m = 0.45$$

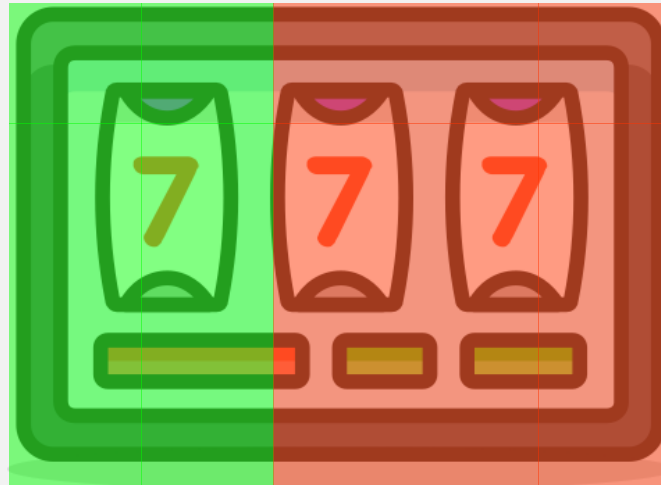


$$p_r = 0.45$$

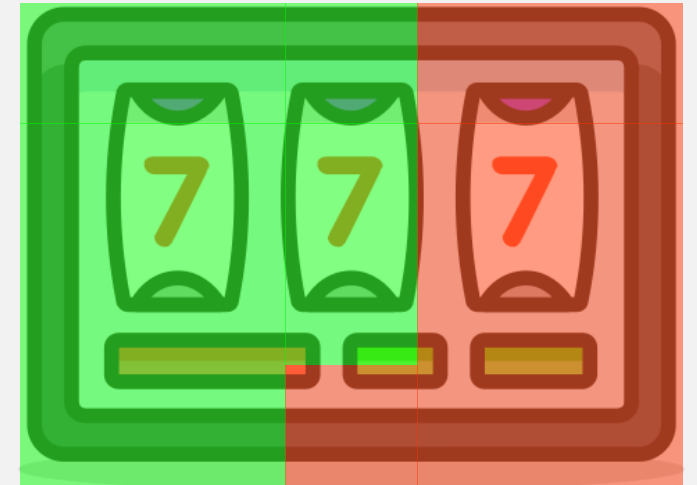
Contextual Bandits



$$p_l = 0.5$$

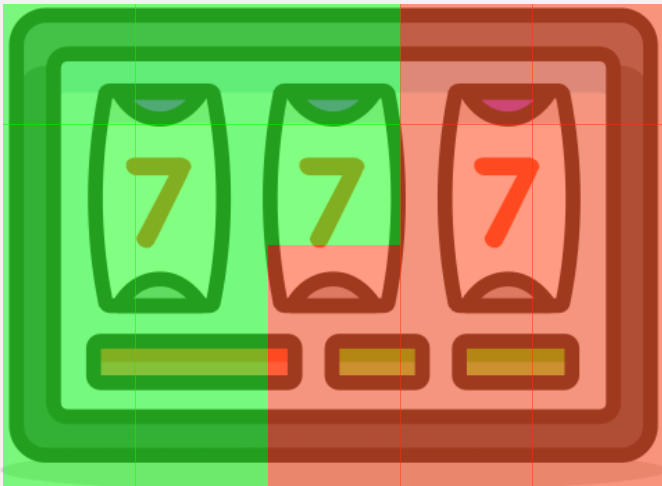


$$p_m = 0.4$$

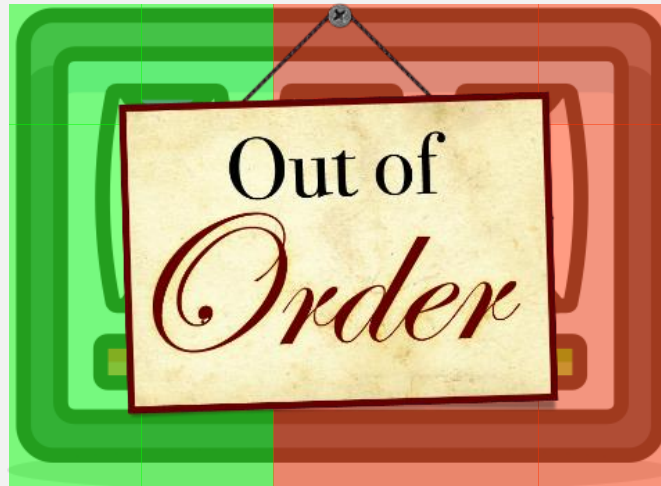


$$p_r = 0.55$$

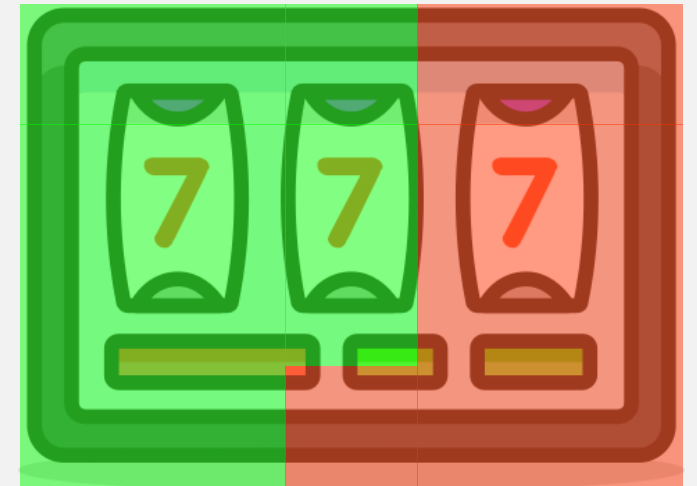
Contextual Bandits



$$p_l = 0.5$$



$$p_m = 0.4$$




$$p_r = 0.55$$

Personalized News Article Recommendation

Menü Zürich 9° -2° **Neue Zürcher Zeitung**

Startseite Neueste Artikel Briefing Meinung International Wirtschaft Technologie Finanzen Feuilleton



Umfassende Antikörpertests könnten einen zweiten generellen Lockdown verhindern

Denn so kann man herausfinden, welche Person sich schon mit Sars-CoV-2 angesteckt hatte und jetzt immun ist. Nach Massentests wüsste man, wie viele Menschen noch gefährdet sind.

Stephanie Lahrtz 26.03.2020

Unsere Spitäler könnten schon Anfang April überlastet sein

Lena Stallmach 25.03.2020

Die Schweiz stoppt Ausschaffungen von Asylbewerbern – erster Corona-Fall im Bundeszentrum Zürich

ENGLISH ESPAÑOL 中文 **The New York Times** Today's Paper

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Books Style Food Travel Magazine T Magazine Real Estate Video

Listen to 'Still Processing' Listen: 'Modern Love' Podcast The Book Review Podcast

U.S. Now Has More Known Cases Than Any Other Country

Markets Shrug Off Jobless Claims as S&P Has Best 3-Day Run Since 1933

- Scientists warned that the United States someday would become the country hardest hit by the coronavirus pandemic. That moment has arrived.
- The latest Wall Street surge came hours after the U.S. reported bleak figures on unemployment claims: More than three million people filed for benefits.
- The Trump administration will publish guidelines to help states decide whether to relax or enhance measures to contain the virus. Here's the latest.

Live 1113 comments



U.S. Cases Jobs Photos Markets

Business Live: Wall Street Shakes Off Unemployment Report as Aid Nears

The S&P 500 had its best three-day run since 1933, buoyed by a huge stimulus plan.

Live 288 comments

Will I Get Some Money? How Much? And When?

Treasury Secretary Steven Mnuchin expects the majority of payments to be made within three weeks. Here's what else taxpayers should know.

552 comments



Jobless Numbers Convey Scale of Economic Wreckage

The weekly figure is among the first data on the economic toll of the vast disruption of normal life and commerce caused by the coronavirus pandemic.



With a Doctor in a Brooklyn E.R.: 'We're in Disaster Mode'

Test kits and protective gear have been in short supply, doctors are falling sick, and every day gets more difficult. But the staff keeps showing up.

413 comments

The Times is providing free access on the coronavirus crisis. Start here for a guide to all our coverage.

Million Menschen infiziert, G20-Staaten wollen Weltwirtschaft mit fünf Billionen US-Dollar stützen – die

Personalized News Article Recommendation



☰ 🔍 ENGLISH ESPAÑOL 中文 LOG IN

The New York Times

Today's Paper

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Books Style Food Travel Magazine T Magazine Real Estate Video

Listen to 'Still Processing'
A new season for a new time.

Listen: 'Modern Love' Podcast
What's a book lover's most embarrassing secret?

The Book Review Podcast
From the archives: Robert Caro on how he gets it done.

Monitor ✔


5156.-
Apple Pro Display XDR - Standard glass
★★★★★ 1

THE SHIFT
Biden Is Losing the Internet. Does That Matter?

The coronavirus has forced the Democrats' presumptive presidential nominee into an all-digital campaign, and he's struggling to break through.

21h ago · By KEVIN ROOSE



TRAVIS DOVE FOR THE NEW YORK TIMES

Monitor ✔


5156.-
Apple Pro Display XDR - Standard glass
★★★★★ 1

THE SHIFT
Welcome to the 'Rabbit Hole'

Introducing an audio series about how the internet is changing, and how it's changing us.

14h ago · By KEVIN ROOSE



Facebook-Backed Libra Cryptocurrency Project Is Scaled Back

After months of criticism, the cryptocurrency is moving ahead with significant revisions.

18h ago · By NATHANIEL POPPER and MIKE ISAAC



ON TECH
FOMO Has Survived the Coronavirus

No one is going anywhere cool, and yet I still feel left out.

20h ago · By SHIRA OVIDE



Notebook ✔


1959.-
Apple MacBook Air – 2019 (13.30", Retina, Intel Core i5-
★★★★★ 2

Notebook 🕒


3423.-
Apple MacBook Pro 16 – 2019 (16", Retina, Intel Core
★★★★★ 1

Personalized News Article Recommendation



☰ 🔍
ENGLISH ESPAÑOL 中文
LOG IN

The New York Times

Today's Paper

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Books Style Food Travel Magazine T Magazine Real Estate Video

Listen to 'Still Processing'
A new season for a new time.

Listen: 'Modern Love' Podcast
What's a book lover's most embarrassing secret?

The Book Review Podcast
From the archives: Robert Caro on how he gets it done.

HIROKO MASUIKE/THE NEW YORK TIMES

Stocks Surge as Washington's Aid Package Advances: Live Updates

Live updates on stock market and business news during the coronavirus outbreak.
5m ago

ANNA MONEYMAKER/THE NEW YORK TIMES

How the Fed's Magic Money Machine Will Turn \$454 Billion Into \$4 Trillion

The central bank takes Treasury Department loan guarantees and uses them to stand up huge programs. Here's how that works.
7h ago • By JEANNA SMIALEK

'It's a Wreck': 3.3 Million File Unemployment Claims as Economy Comes Apart

The weekly figure is among the first data on the economic toll of the vast disruption of normal life and commerce caused by the coronavirus pandemic.
3h ago • By BEN CASSELMAN, PATRICIA COHEN and TIFFANY HSU

Surging Traffic Is Slowing Down Our Internet

With people going online more in the pandemic, internet traffic has exploded. That's taking a toll on our download speeds and video quality.
9h ago • By CECILIA KANG, DAVEY ALBA and ADAM SATARIANO

Personalized News Article Recommendation



☰ 🔍
ENGLISH ESPAÑOL 中文
LOG IN

The New York Times

Today's Paper

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Books Style Food Travel Magazine T Magazine Real Estate Video

Listen to 'Still Processing'
A new season for a new time.

Listen: 'Modern Love' Podcast
What's a book lover's most embarrassing secret?

The Book Review Podcast
From the archives: Robert Caro on how he gets it done.

SCOTT MCINTYRE FOR THE NEW YORK TIMES

ON BASEBALL

It's Opening Day. Baseball Is Closed.

Instead of red, white and blue bunting and crowds flooding ballparks, we have players holed up in homes and endless questions about when baseball will return. When it does, it will be more important than ever.

7h ago · By TYLER KEPNER

Coronavirus Protective Masks to Be Made From M.L.B. Uniform Material

The deal will send masks to health professionals in the Northeast with the distinctive pinstripes of the Yankees and the Philadelphia Phillies.

2h ago · By DAVID WALDSTEIN

W.N.B.A. Will Hold 'Virtual' Draft in April

The league is keeping its scheduled date of April 17 but adapting the format to allow draftees to have their moment in the spotlight no matter where they are.

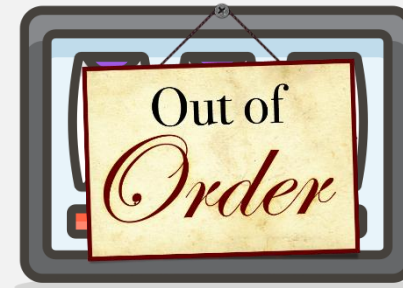
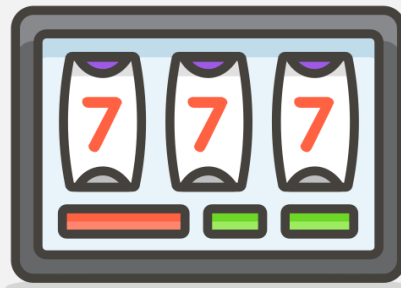
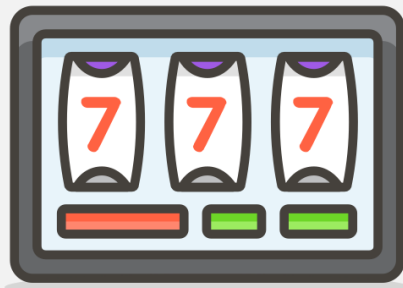
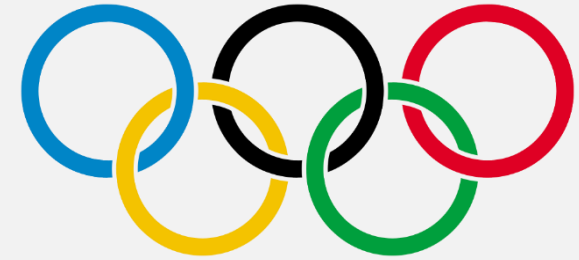
7h ago · By HOWARD MEGDAL

Baseball Season Opens With a Dash of Imagination

Complaints of baseball's slow pace yielded to the universal wish that this game would last forever.

4h ago · By DAN BARRY

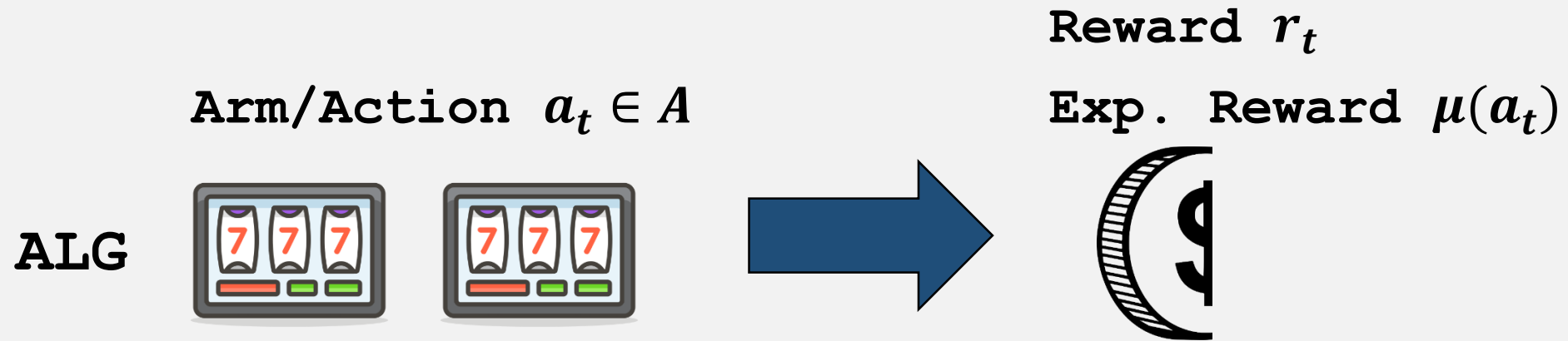
Context



Problem Protocol: Context Free Bandits

For each round $t \in [T]$:
 ALG picks an **Arm/Action** $a_t \in A$.
 Reward r_t is realized.

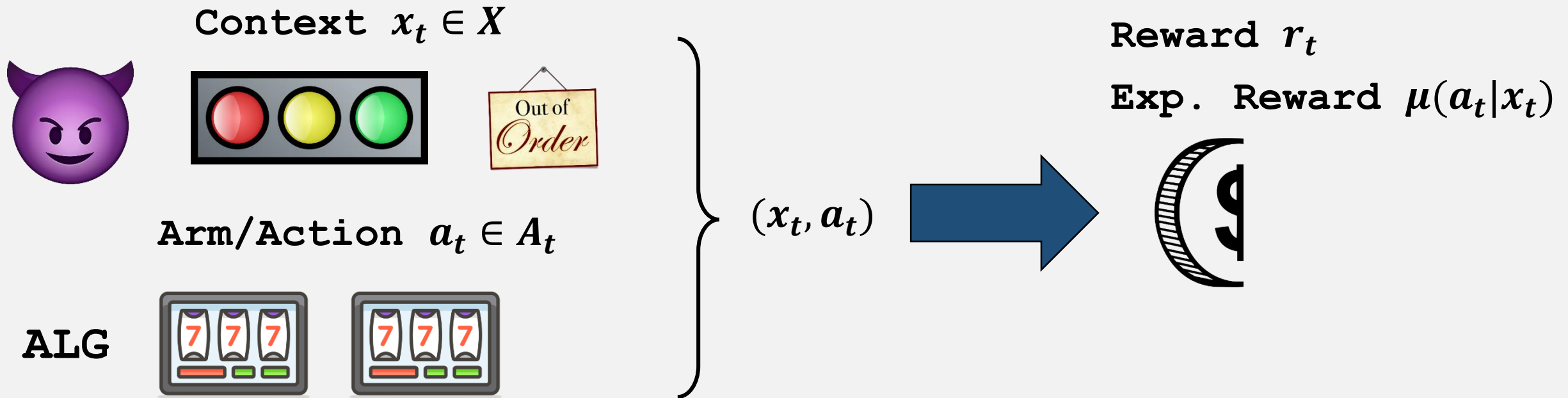
One Round: Context Free Bandits



Problem Protocol: Contextual Bandits

For each round $t \in [T]$:
ALG observes a **Context** $x_t \in X$.
ALG picks an **Arm/Action** $a_t \in A_t$.
Reward r_t is realized.

One Round: Contextual Bandits



Expected Regret: Context Free Bandits



$$\mathbb{E}[R(T)] = \mu^* \cdot T - \sum_{t=1}^T \mu(a_t)$$

with $\mu^* := \max_{a \in A} \mu(a)$

Expected Regret: Contextual Bandits



$$\mathbb{E}[R(T)] = \text{REW}(\pi^*(\mathbf{x})) - \sum_{t=1}^T \mu(\mathbf{a}_t | \mathbf{x}_t)$$

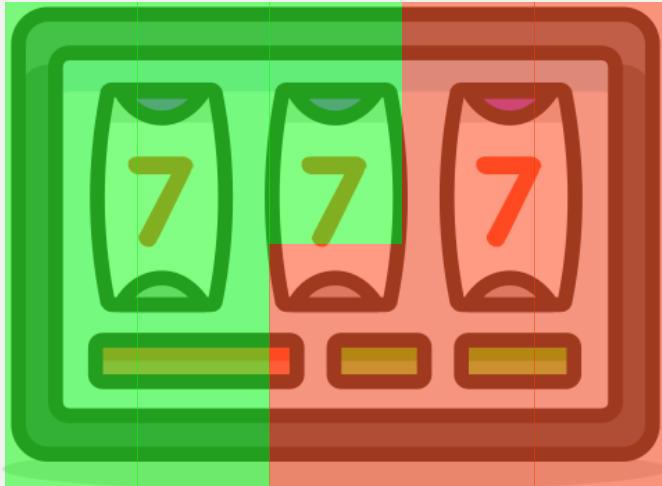
with **Best Response Policy**:

$$\pi^*(\mathbf{x}) = \max_{a \in A} \mu(a_t | \mathbf{x}_t)$$

Recap: Upper Confidence Bound

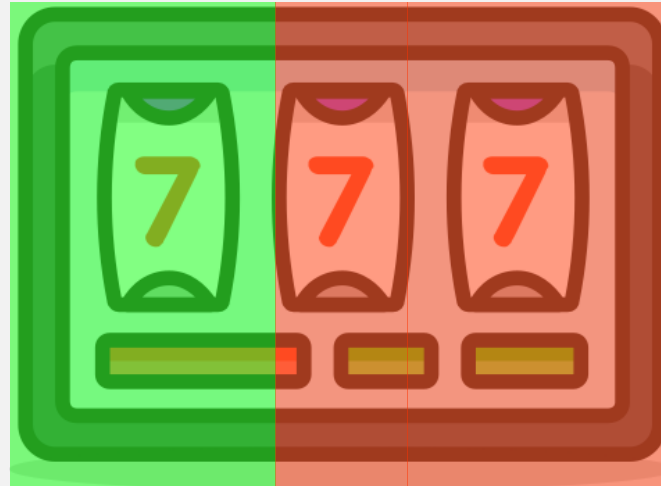
$$r_t(a) = \sqrt{\frac{2 \ln T}{n_t(a)}} = \sqrt{\frac{2 \ln 170}{n_t(a)}}$$

100 × Played
60 × Won



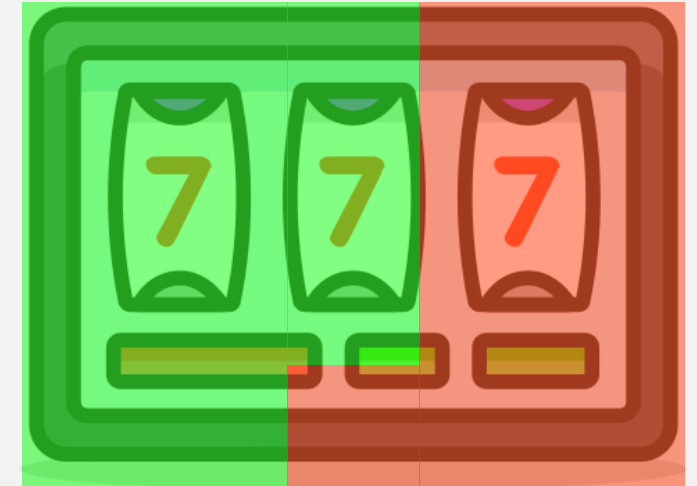
$$\left. \begin{array}{l} p_l = 0.5; \mu(a_l) = 0 \\ \bar{\mu}(a_l) = 0.2 \\ r_t(a_l) \approx 0.32 \end{array} \right\} 0.52$$

50 × Played
15 × Won



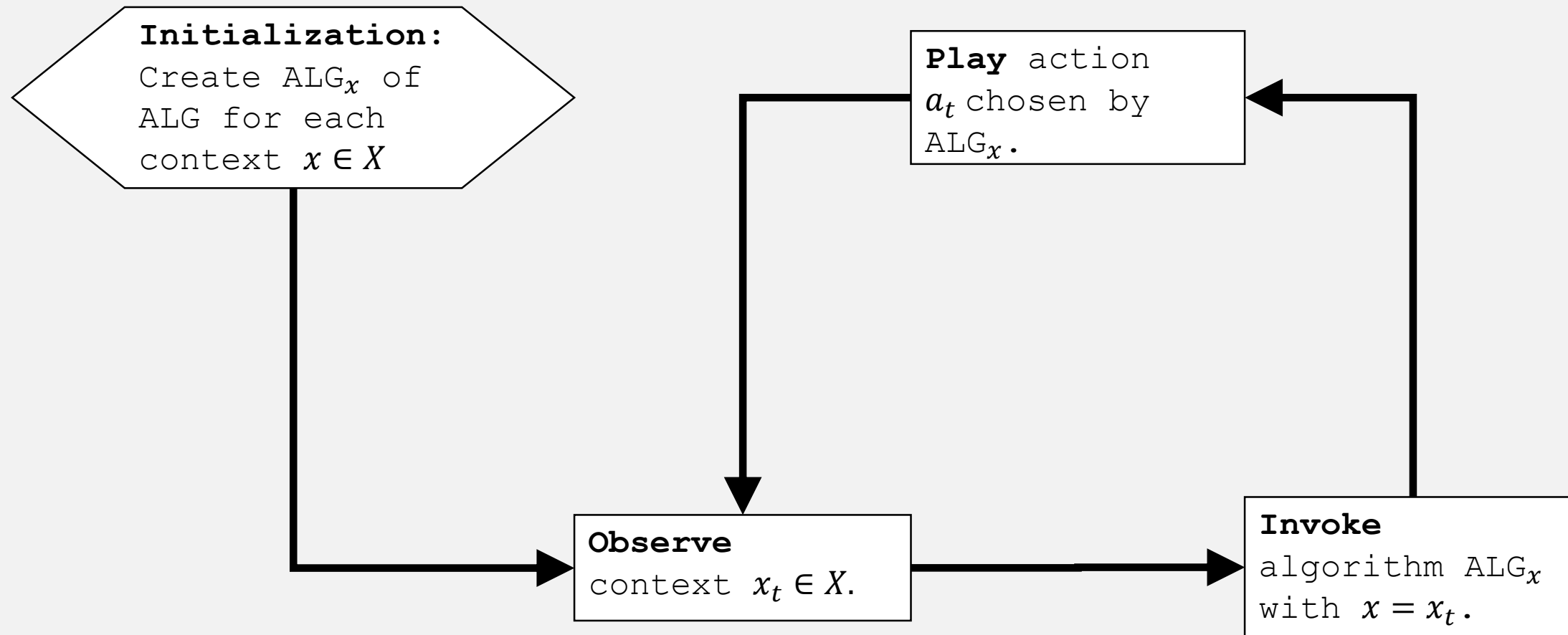
$$\left. \begin{array}{l} p_m = 0.4; \mu(a_m) = -0.2 \\ \bar{\mu}(a_m) = -0.4 \\ r_t(a_m) \approx 0.45 \end{array} \right\} 0.05$$

20 × Played
9 × Won



$$\left. \begin{array}{l} p_r = 0.55; \mu(a_r) = 0.1 \\ \bar{\mu}(a_r) = -0.1 \\ r_t(a_r) \approx 0.71 \end{array} \right\} 0.61$$

Run separate copy of ALG for each context



Example for ALG: Upper Confidence Bound/UCB1

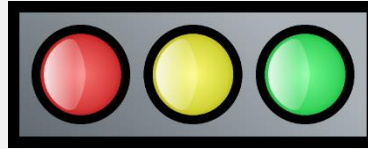
Regret Analysis with $\text{ALG}=\text{UCB1}$

$$\mathbb{E}[\mathbf{R}(T)] = \sum_{x \in X} \mathbb{E}[R_x(T)] = \sum_{x \in X} \underbrace{o\left(\sqrt{Kn_x \log T}\right)}_{\text{Regret of UCB1}} \leq \mathbf{o}\left(\sqrt{KT|X| \log T}\right)$$

Regret Analysis with $ALG=UCB1$

$$\mathbb{E}[R(T)] = o\left(\sqrt{KT|X|\log T}\right)$$

Good for **small** number of contexts:



Bad for **large** number of contexts:

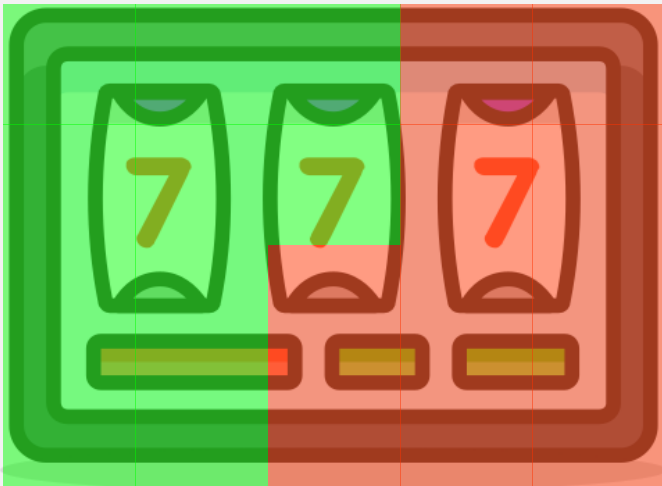


Quick Recap: Lipschitz Bandits

$$A \subseteq [0, 1]$$

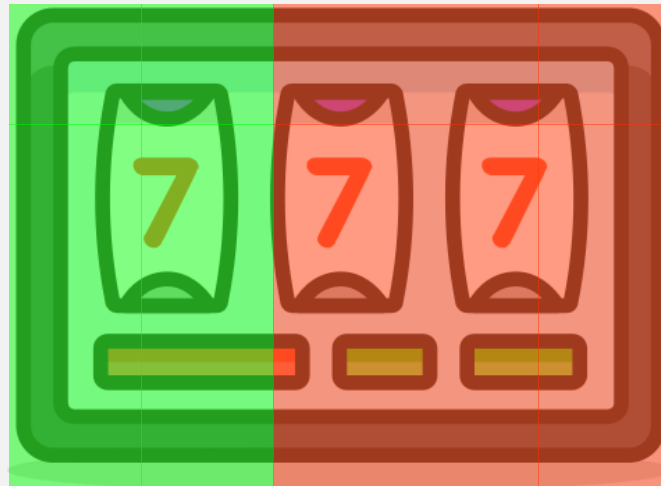


OAB 1



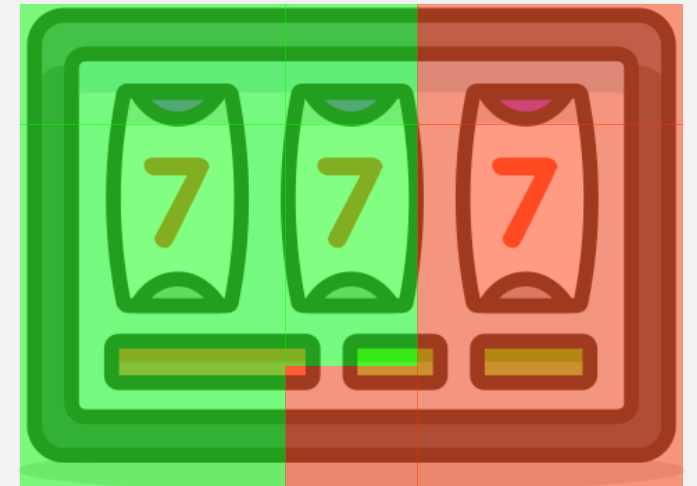
$$p_l = 0.5$$

OAB 2



$$p_m = 0.4$$

OAB 1



$$p_r = 0.55$$


Quick Recap: Lipschitz Bandits

Rewards satisfy a Lipschitz condition:

$|\mu(x) - \mu(y)| \leq L \cdot |x - y|$ for any two arms $x, y \in A$,

$$\left| \text{\$} - \text{\$} \right| \leq L \cdot \left| 1.0 - 0.0 \right|$$

Constant known
to the Algorithm



Quick Recap: Lipschitz Bandits

Simple Solution: **Discretization**
Solve for $S \subseteq A$
with off-the-shelf MAB algorithm
(i.e. UCB1)

Lipschitz Contextual Bandits

$$X \subseteq [0, 1]$$

$$\begin{array}{|c|c|c|} \hline \bullet & \bullet & \bullet \\ \hline \end{array} \in [0, 1]$$



Lipschitz Contextual Bandits

Rewards satisfy a Lipschitz condition:

$|\mu(\mathbf{a}|\mathbf{x}) - \mu(\mathbf{a}|\mathbf{x}')| \leq L \cdot |\mathbf{x} - \mathbf{x}'|$ for any two contexts $\mathbf{x}, \mathbf{x}' \in X$,
where L is the Lipschitz constant known to the `ALG`

$$\left| \text{\$} - \text{\$} \right| \leq L \cdot \left| \text{chart} - \text{target} \right|$$

Constant known
to the Algorithm

Lipschitz Contextual Bandits

Simple Solution: **Discretization (Context)**

Solve for $S \subseteq X$

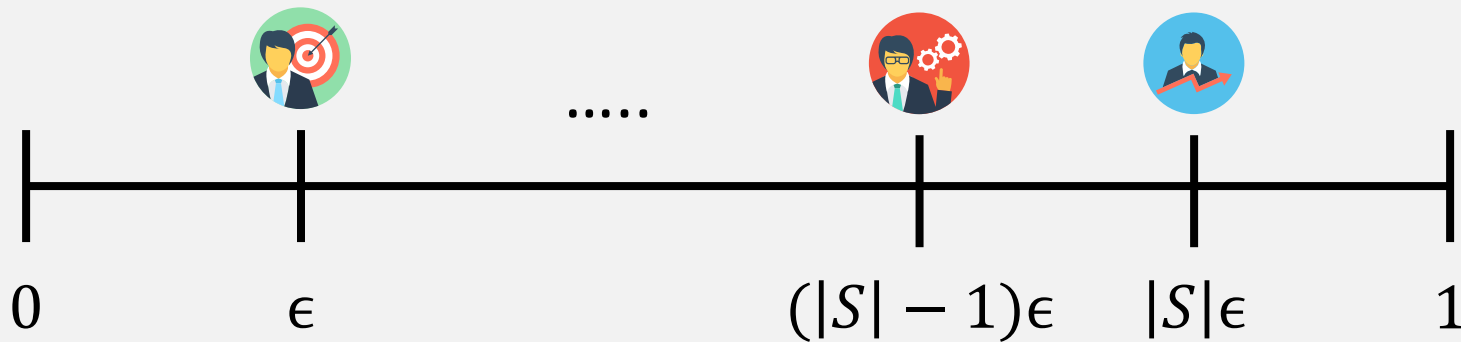
with off-the-shelf MAB algorithm

(i.e. UCB1)

Context Free vs. Contextual Lipschitz Bandits



Lipschitz Contextual Bandits



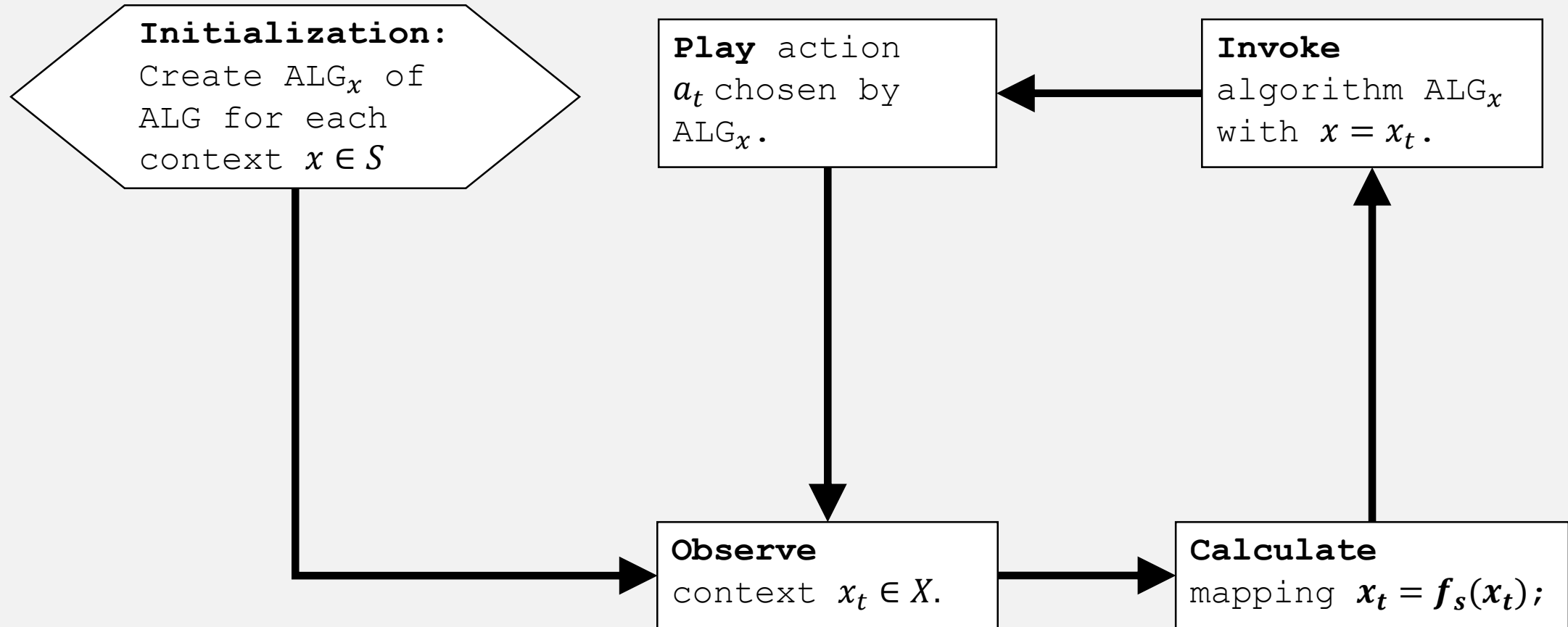
Let S be the ϵ -uniform mesh
 with $\epsilon = \frac{1}{|S|+1}$
 and apply ALG for S

Let $f_S(x)$ be a mapping from context x to the closest point in S :

$$f_S(x) = \min (\arg \min_{x' \in S} |x - x'|)$$

Lipschitz Contextual Bandits - Discretization

ALG_S:



Example for ALG: Upper Confidence Bound/UCB1

Regret Analysis – Lipschitz (ALG = UCB1)

Let us define the **Discretized Best Response** $\pi_S^*: X \rightarrow A$

$$\pi_S^*(x) = \pi^*(f_S(x)) \text{ for each context } x \in X$$

Regret of ALG_S : $R_S(T) = \text{REW}(\pi_S^*) - \text{REW}(\text{ALG}_S)$

Discretization Error: $\text{DE}(S) = \text{REW}(\pi^*) - \text{REW}(\pi_S^*)$

Overall Regret: $R(T) = R_S(T) + \text{DE}(S)$

Overall Expected Regret: $\mathbb{E}[R(T)] = \mathbb{E}[R_S(T)] + \mathbb{E}[\text{DE}(S)]$

Regret Analysis – Lipschitz (ALG = UCB1)

Overall Expected Regret: $\mathbb{E}[R(T)] = \mathbb{E}[R_S(T)] + \mathbb{E}[DE(S)]$

Expected Regret: $\mathbb{E}[R_S(T)] = O\left(\sqrt{KT|S|\log T}\right) = O\left(\sqrt{\frac{1}{\epsilon}KT\log T}\right)$

Exp. Discretization Error: $\mathbb{E}[DE(S)] = \text{REW}(\pi^*) - \mathbb{E}[\text{REW}(\pi_S^*)] \leq \epsilon LT$

$$\begin{aligned} \mu(\pi_S^*(x)|f_S(x)) &\geq \mu(\pi^*(x)|f_S(x)) \\ &\geq \mu(\pi^*(x)|x) - \epsilon L \end{aligned}$$

Optimality of $\pi_S^*(x)$

Lipschitz property

$$\mathbb{E}[\text{REW}(\pi_S^*)] \geq \text{REW}(\pi^*) - \epsilon LT$$

Regret Analysis – Lipschitz (ALG = UCB1)

$$\mathbb{E}[R(T)] \leq O\left(\sqrt{\frac{1}{\epsilon}KT\log T}\right) + \epsilon LT = O\left(T^{\frac{2}{3}}(LK\log T)^{\frac{1}{3}}\right)$$

Linear Contextual Bandits: LinUCB algorithm

WWW 2010 • Full Paper

April 26-30 • Raleigh • NC • USA

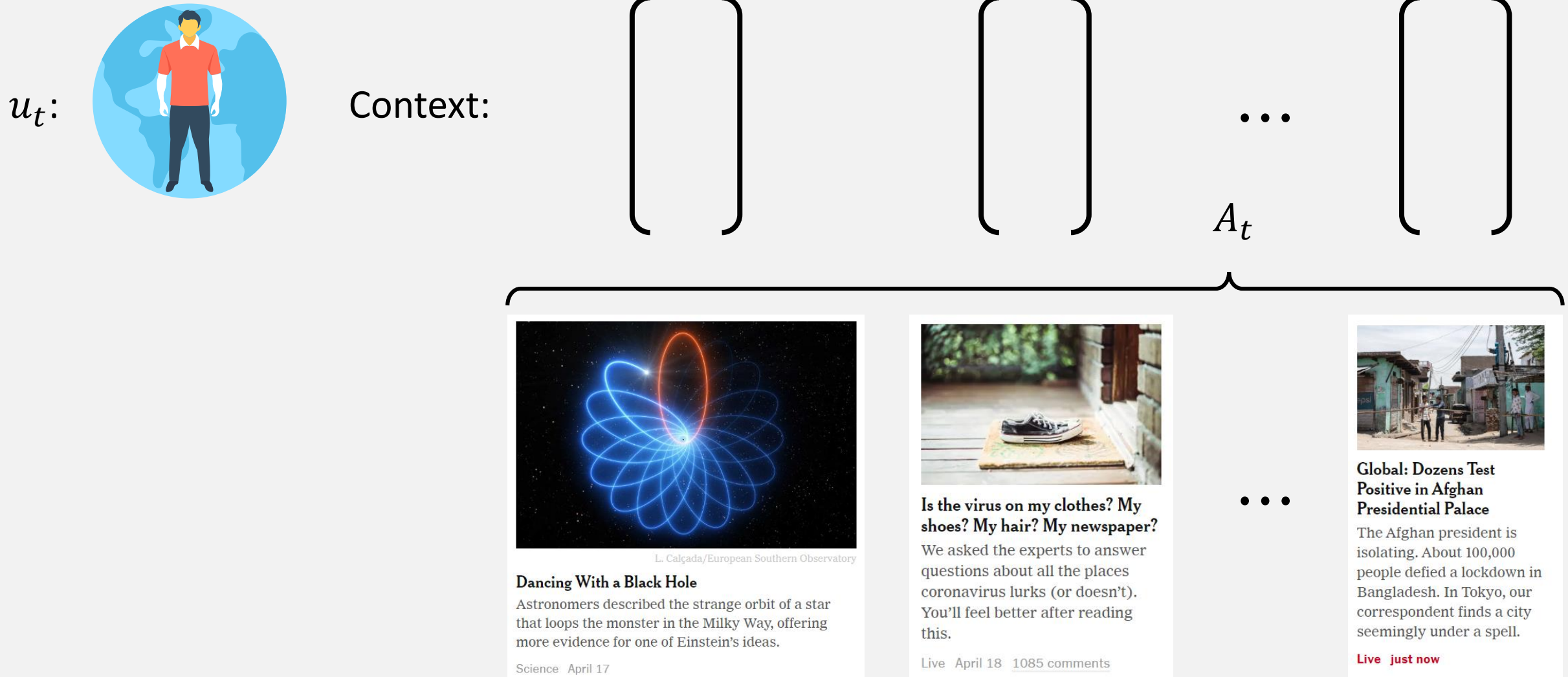
A Contextual-Bandit Approach to Personalized News Article Recommendation

Lihong Li[†], Wei Chu[†],
[†]Yahoo! Labs
lihong,chuwei@yahoo-
inc.com

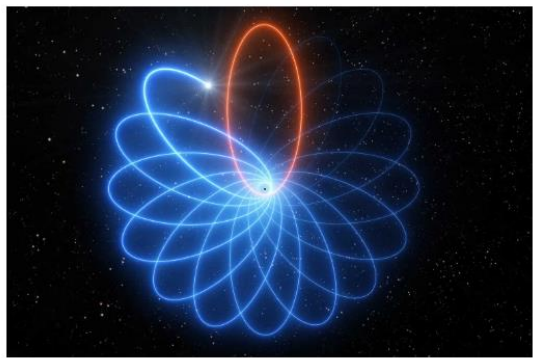
John Langford[‡]
[‡]Yahoo! Labs
jl@yahoo-inc.com

Robert E. Schapire^{+*}
⁺Dept of Computer Science
Princeton University
schapire@cs.princeton.edu

Contextual Bandits in Personalized News Article Recommendation



Contextual Bandits in Personalized News Article Recommendation



L. Calçada/European Southern Observatory

Dancing With a Black Hole

Astronomers described the strange orbit of a star that loops the monster in the Milky Way, offering more evidence for one of Einstein's ideas.

Science April 17



Is the virus on my clothes? My shoes? My hair? My newspaper?

We asked the experts to answer questions about all the places coronavirus lurks (or doesn't). You'll feel better after reading this.

Live April 18 [1085 comments](#)

• • •



Global: Dozens Test Positive in Afghan Presidential Palace

The Afghan president is isolating. About 100,000 people defied a lockdown in Bangladesh. In Tokyo, our correspondent finds a city seemingly under a spell.

Live just now

$$a_t \in A_t$$

Reward r_{t,a_t} : Clicked (1) or Not Clicked (0)

Linear Contextual Bandits

Expected reward is linear in:

$$\mu(a|x_{t,a}) = x_{t,a}^T \theta_a^* \text{ for all arms } a \text{ and contexts } x$$

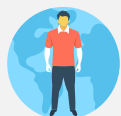
$$\theta_a = (D_a^T D_a + I_d)^{-1} b_a$$

Training input:

$$D_a \in \mathbb{R}^{m \times d}$$

Response vector:

$$b_a \in \mathbb{R}^m$$



Is the virus on my clothes? My shoes? My hair? My newspaper?

We asked the experts to answer questions about all the places coronavirus lurks (or doesn't). You'll feel better after reading this.

Live April 18 1085 comments

Click

Click

No-Click

Confidence Region

When components in b_a are independent to rows of D_a , it can be shown [1] that with probability at least $1 - \delta$:

$$|x_{t,a}^T \theta_a - \mu(a|x_{t,a})| = \alpha \sqrt{x_{t,a}^T (D_a^T D_a + I_d)^{-1} x_{t,a}}$$

for any $\delta > 0$ and $x_{t,a} \in \mathbb{R}^d$ where $\alpha = 1 + \sqrt{\frac{\ln(\frac{2}{\delta})}{2}}$

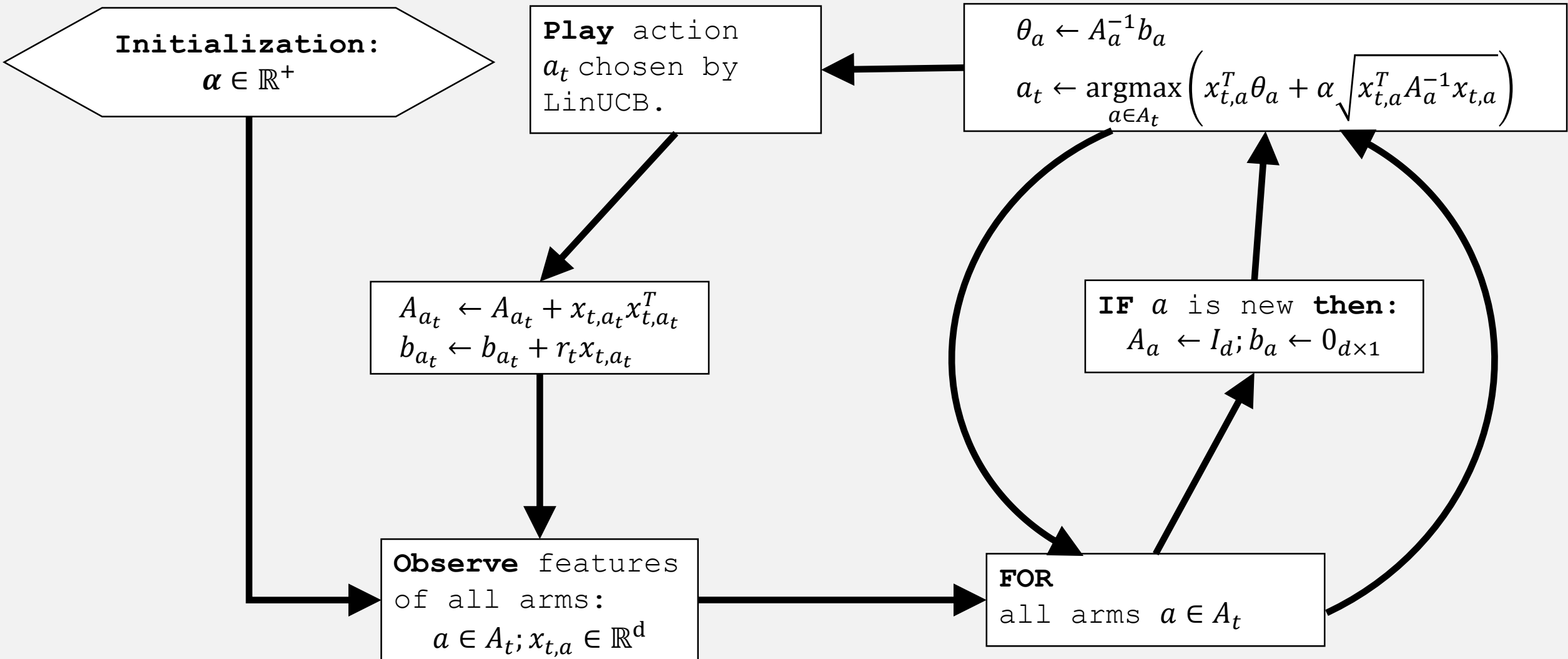
Which gives us following **UCB arm selection policy**:

$$\mathbf{a}_t = \underset{a \in A_t}{\operatorname{argmax}} \left(x_{t,a}^T \boldsymbol{\theta}_a + \alpha \sqrt{x_{t,a}^T A_a^{-1} x_{t,a}} \right)$$

where $A_a = D_a^T D_a + I_d$

Linear Contextual Bandits

LinUCB:



Regret Analysis

$$\mathbb{E}[R(T)] = O\left(d \sqrt{T \ln \frac{1+T}{\delta}}\right)$$

Disjoint vs. Hybrid Linear Models

Disjoint Linear Model:

Expected reward is linear in:

$$\mu(a|x_{t,a}) = x_{t,a}^T \theta_a^*$$

for all arms a and contexts x

Hybrid Linear Model:

Expected reward is linear in:

$$\mu(a|x_{t,a}) = \mathbf{z}_{t,a}^T \boldsymbol{\beta}^* + x_{t,a}^T \theta_a^*$$

for all arms a and contexts x

References

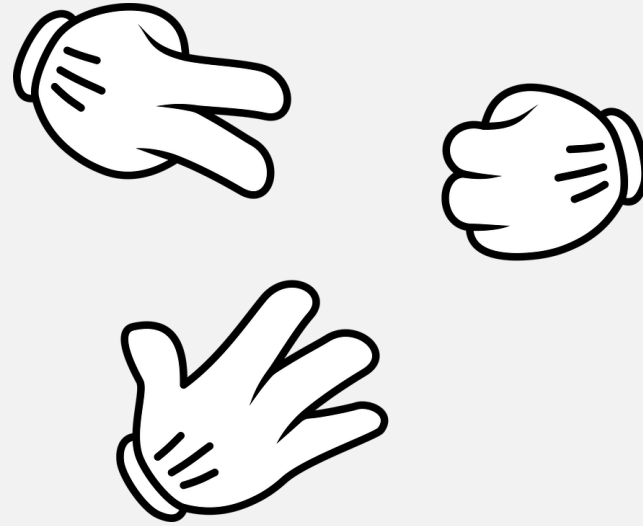
Slivkins, Aleksandrs. "**Introduction to multi-armed bandits.**" *Foundations and Trends® in Machine Learning* 12.1-2 (2019): 1-286.

Li, Lihong, et al. "**A contextual-bandit approach to personalized news article recommendation.**" *Proceedings of the 19th international conference on World wide web*. 2010.

Zhou, Li. "**A survey on contextual multi-armed bandits.**" *arXiv preprint arXiv:1508.03326* (2015)







Other Application for Bandit Algorithms

Bandit & Games



Bandit & Games

Matrix M :

					
	[0	-1	1]
		1	0	-1	
		-1	1	0	

The Game:

ALG chooses row i_t of M ADV chooses col j_t of M



For each round $t \in [T]$:

Simultaneously:

ALG chooses row i_t of M ;

ADV chooses col j_t of M ;

ALG incurs cost $M(i_t, j_t)$

ALG observes feedback $F_t = F(t, i_t, j_t, M)$