Advanced Prediction Models

Deep Learning, Graphical Models and Reinforcement Learning

Beyond Prediction

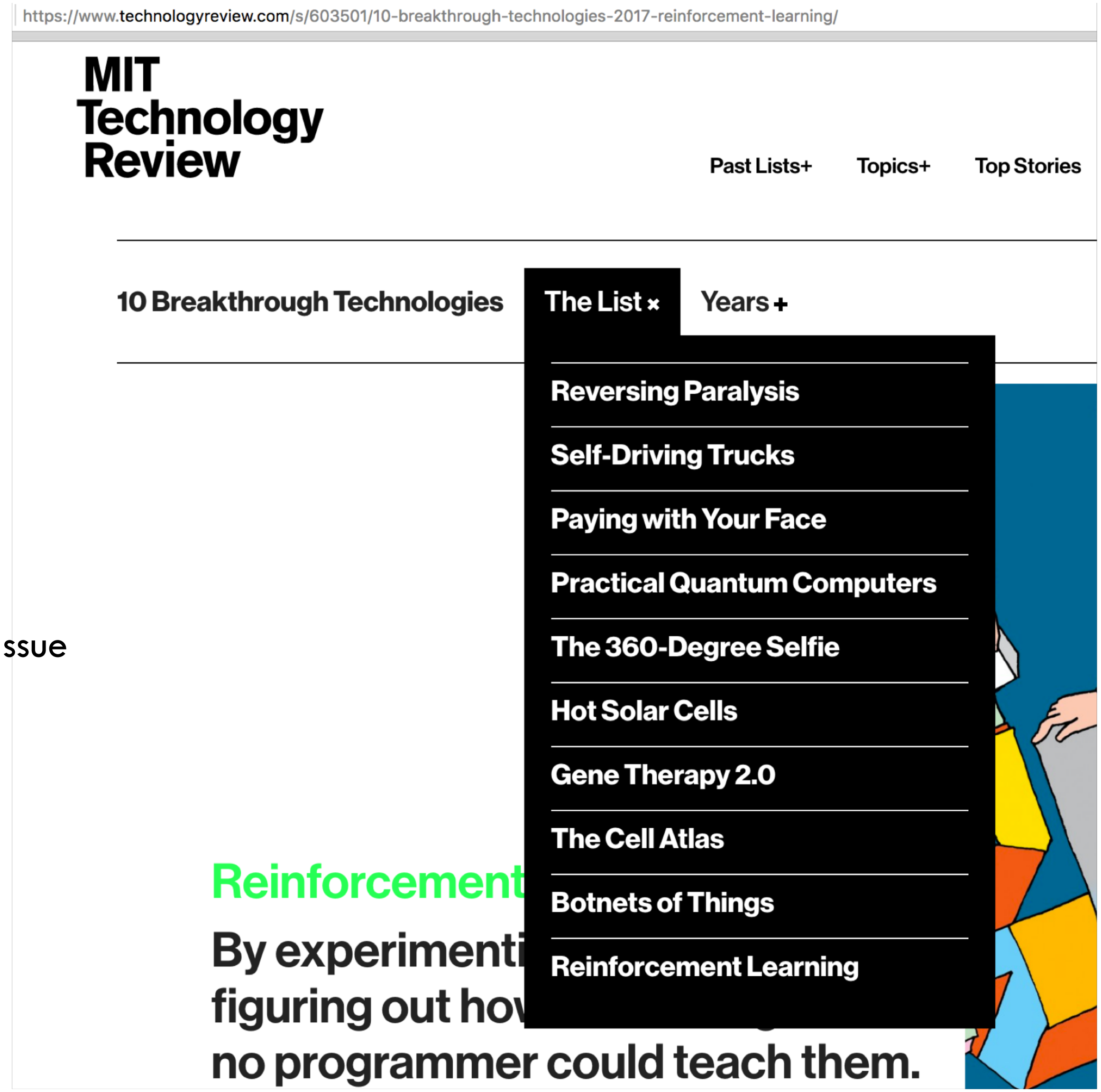
- Recall from the introductory class
  - We discussed complex prediction problems and addressed them using
    - Deep learning architectures
    - Graphical models
  - We also discussed complex decisions, especially in the presence of feedback
- A way to make data-driven decisions: we will look at
  - Online machine learning (this lecture)
  - Reinforcement learning (next)
  - Deep reinforcement learning (next to next)

Examples of Complex Decisions

- Inventory Management
  - Observations: current inventory levels
  - Actions: number of units of each item to purchase
  - Rewards: profit
- Resource allocation: who to provide customer service to first
- Routing problems: in management of shipping fleet, which trucks / truckers to assign to which cargo

[1] Reference: John Schulman, MLSS 2016

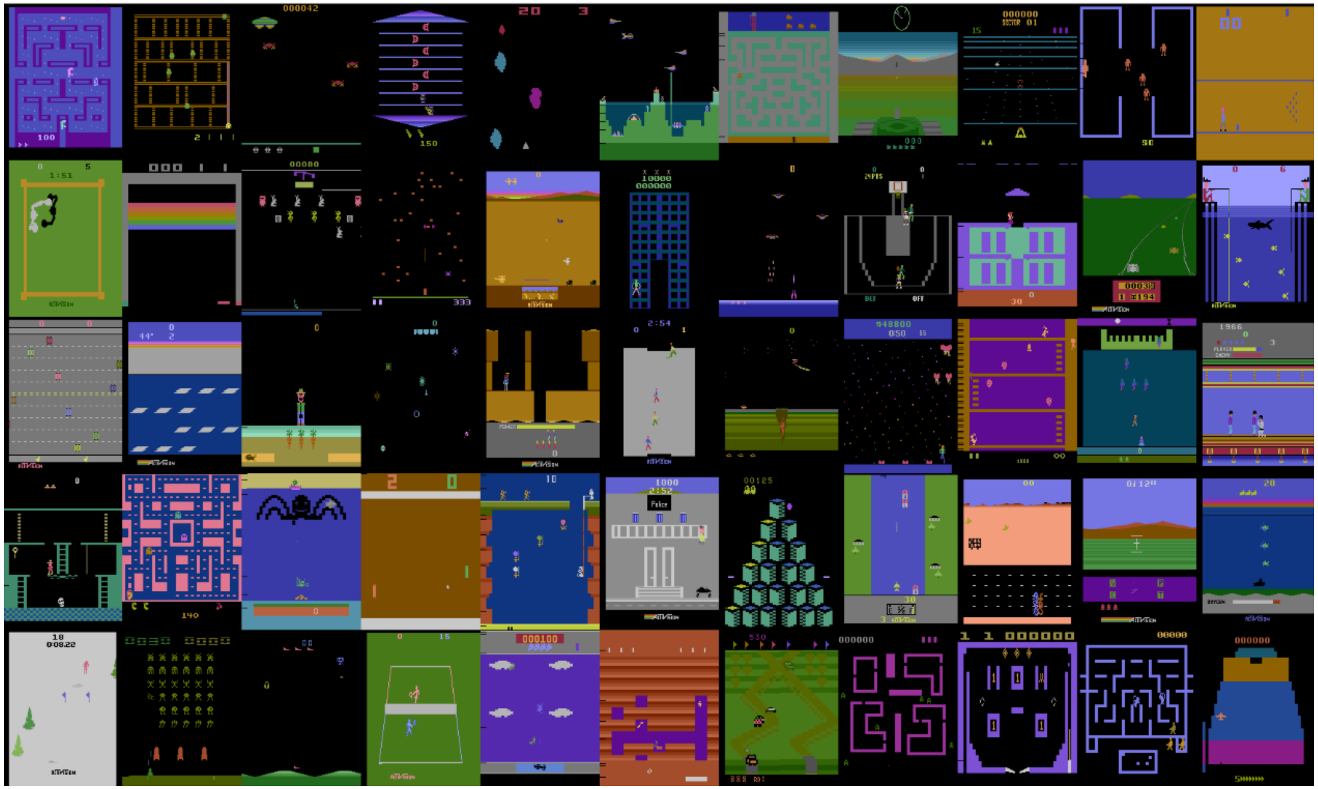
Reinforcement Learning: The Next Frontier in Data Science



March/April 2017 Issue

[1]Reference: technologyreview.com/s/603501/10-breakthrough-technologies-2017-reinforcement-learning/

Reinforcement Learning: The Next Frontier in Data Science



[1]Figure: Defazio Graepel, Atari Learning Environment

Reinforcement Learning: The Next Frontier in Data Science

[1]Reference: DeepMind, March 2016

Today's Outline

- Online Machine Learning
- A/B Testing
- Multi-armed bandits
- Contextual bandits

Online Machine Learning

The Gist of Online (Machine) Learning

1. (Optionally) observe the state of the world (aka **context**)
2. Choose an action
3. Obtain feedback on the chosen action

Repeat

[1]Reference: Alekh Agarwal et al., http://arxiv.org/abs/1606.03966

The Gist of Online (Machine) Learning

1. (Optionally) observe the state of the world (aka **context**)
2. Choose an action
3. Obtain feedback on the chosen action

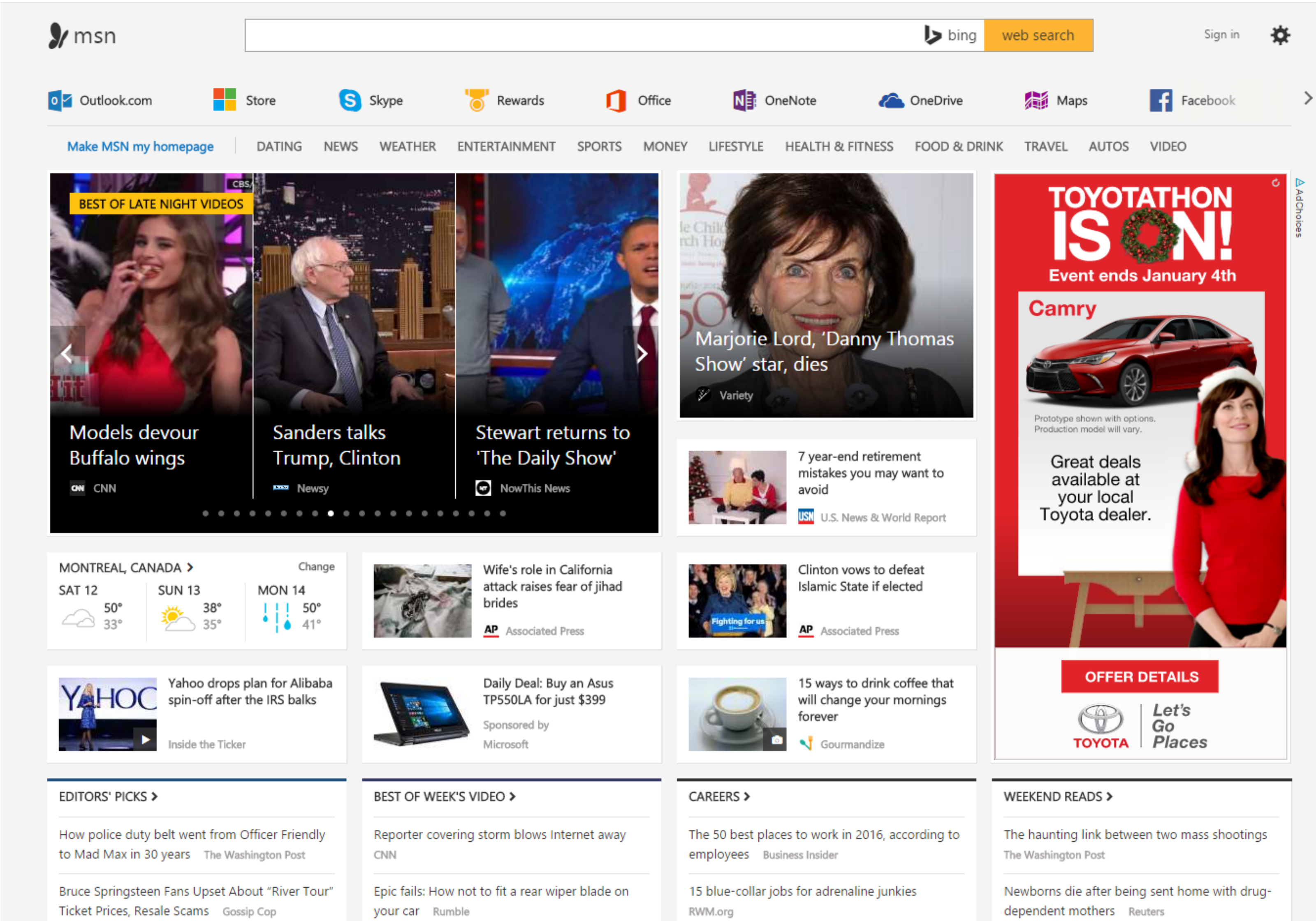
Repeat

Goal: Optimize feedback (e.g. maximize reward) for chosen actions

Assumption: Agent's actions do not influence future contexts

[1]Reference: Alekh Agarwal et al., http://arxiv.org/abs/1606.03966

MSN Deployment for Personalized News

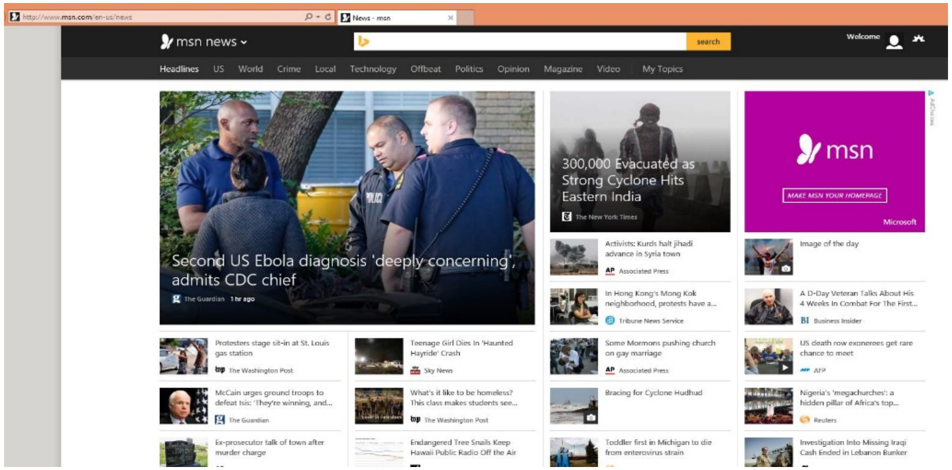


[1]Reference: Alekh Agarwal et al., http://arxiv.org/abs/1606.03966

MSN Deployment for Personalized News

Loop:

1. User **arrives** at MSN with browsing history, user account, previous visits,…
2. Microsoft **chooses** news stories, …
3. User **responds** to content (clicks, navigation, etc)

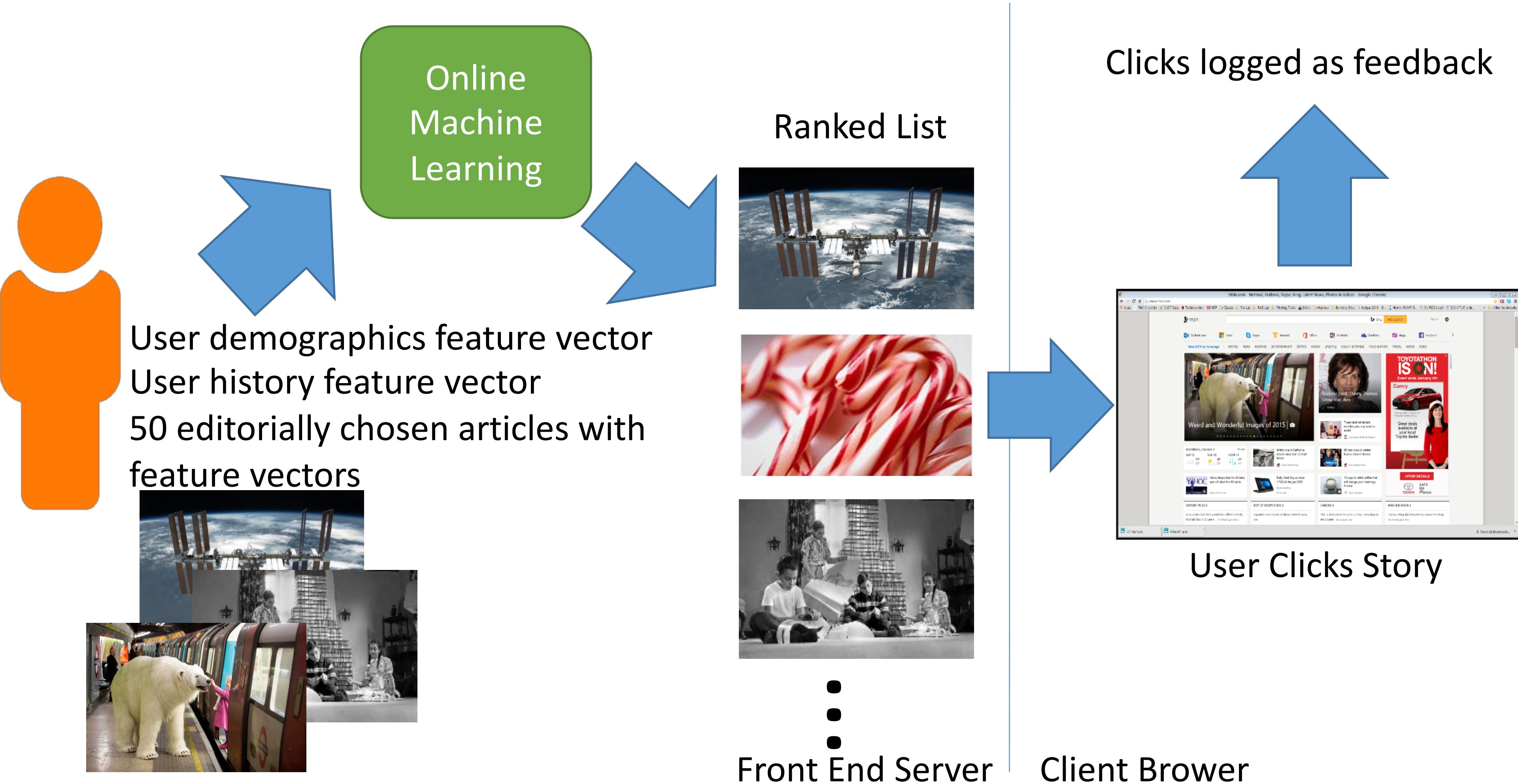


Goal: Choose content to yield desired user behavior

Assumption: Recommendations to one user do not affect other users

[1]Reference: Alekh Agarwal et al., http://arxiv.org/abs/1606.03966

MSN Deployment for Personalized News



[1]Reference: Alekh Agarwal et al., http://arxiv.org/abs/1606.03966

MSN Deployment for Personalized News

- 10 million+ users
- 1000s of requests per second
- 5% overhead on front end machines
- 10s of servers for training
- 5 minute model update frequency

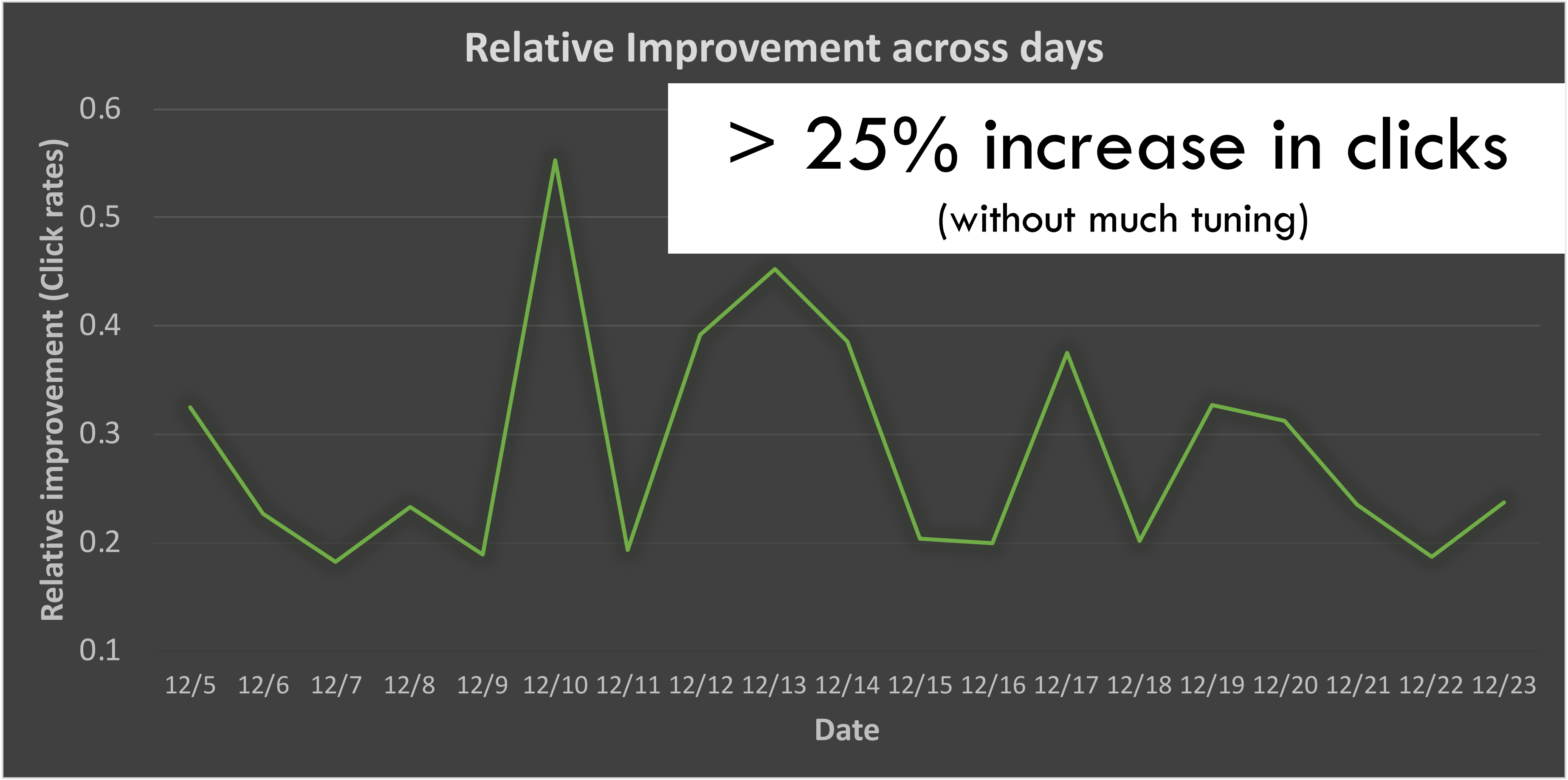


⋮

[1]Reference: Alekh Agarwal et al., http://arxiv.org/abs/1606.03966

MSN Deployment for Personalized News

- Relative gains observed

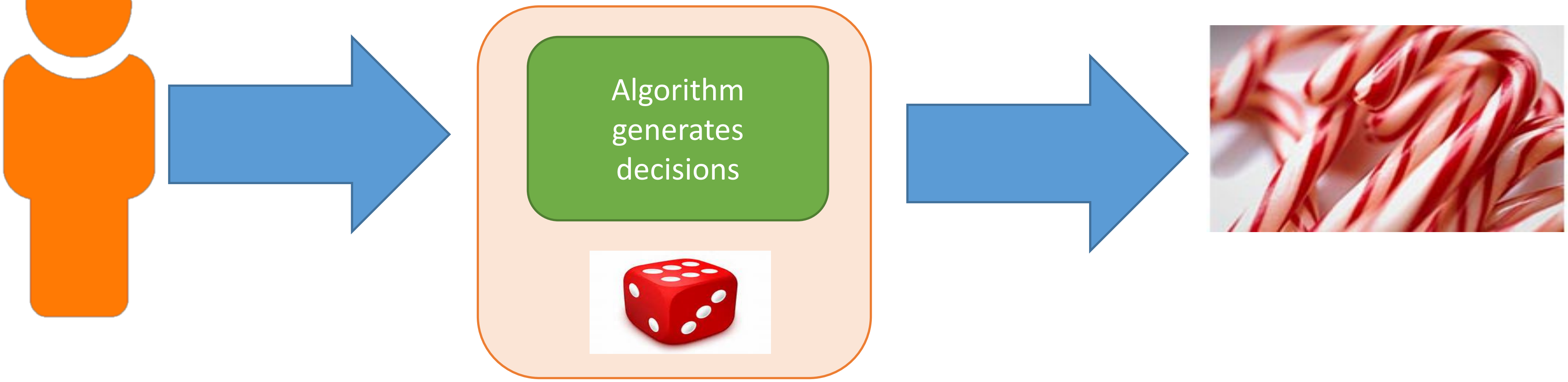


[1]Reference: Alekh Agarwal et al., http://arxiv.org/abs/1606.03966

Multitude of Applications

- Content Recommendation: Apps, Movies, Books, …
- Personalization of search results
- Customer churn prevention
- Adaptive UI personalization
- …

Online Machine Learning



[1]Reference: Alekh Agarwal et al., http://arxiv.org/abs/1606.03966

Questions?

Today's Outline

- Online Machine Learning
- A/B Testing
- Multi-armed bandits
- Contextual bandits

A/B Testing

Motivation for A/B Tests

- Typical business scenario
  - Say there is a meeting to decide on how to improve a product or service
  - Multiple competing ideas emerge
  - Want to make this decision after making some field observations.
  - How to pick one?

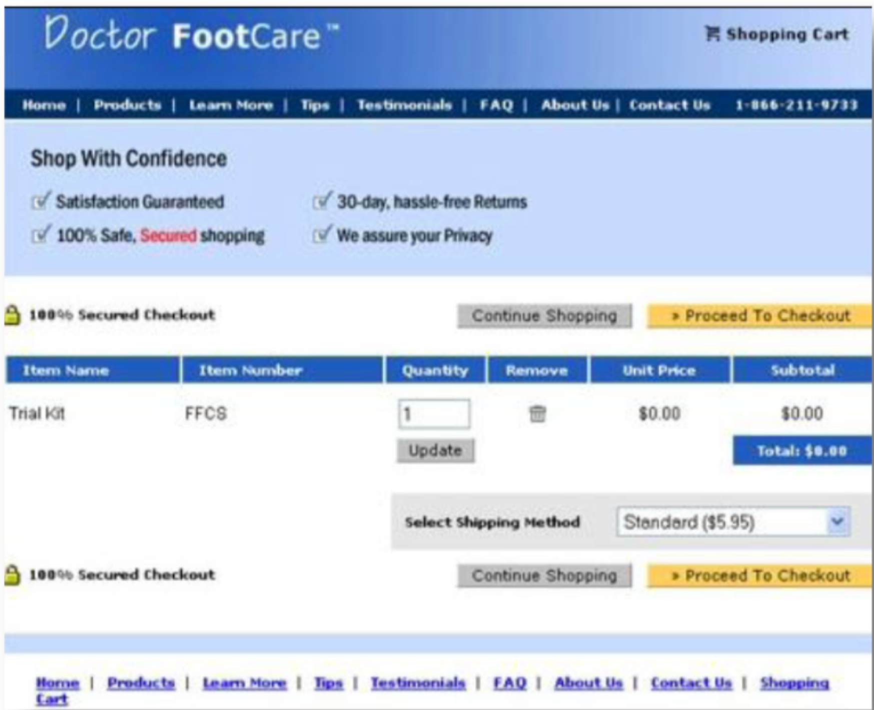
- Use A/B testing (this is related to two-sample hypothesis testing)

Motivation for A/B Tests

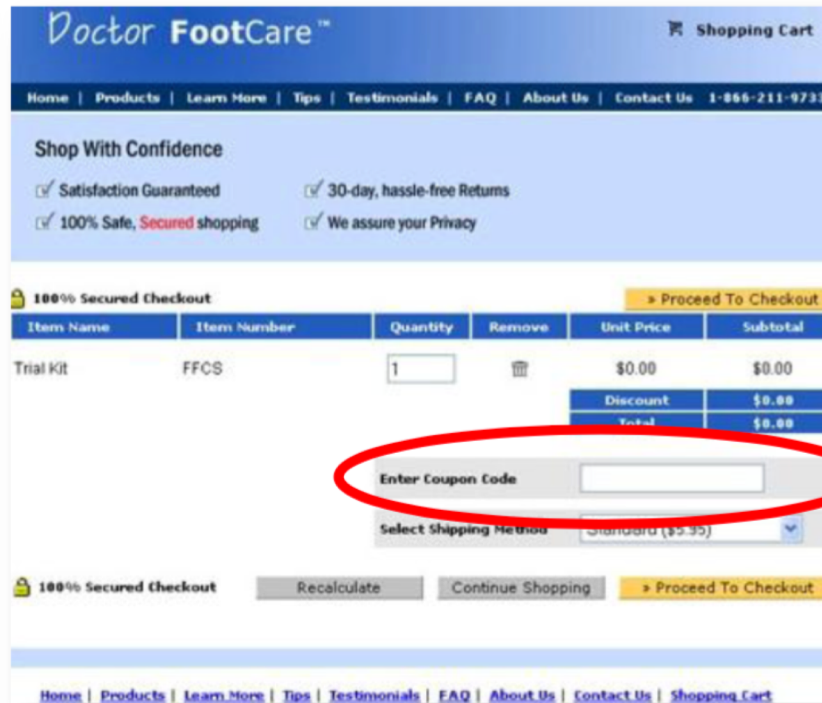
- Full time companies such as Optimizely, Apptimize, APT, Monetate, etc. provide A/B testing services
- Extensively used at
  - Microsoft for Bing.com (see http://exp-platform.com)
  - Google, Facebook, Amazon, Airbnb, Linkedin …
- Marketing tools
- Clinical trials ($11b+ market)

Example with Two Solutions

- Which page has a higher conversion rate?



A



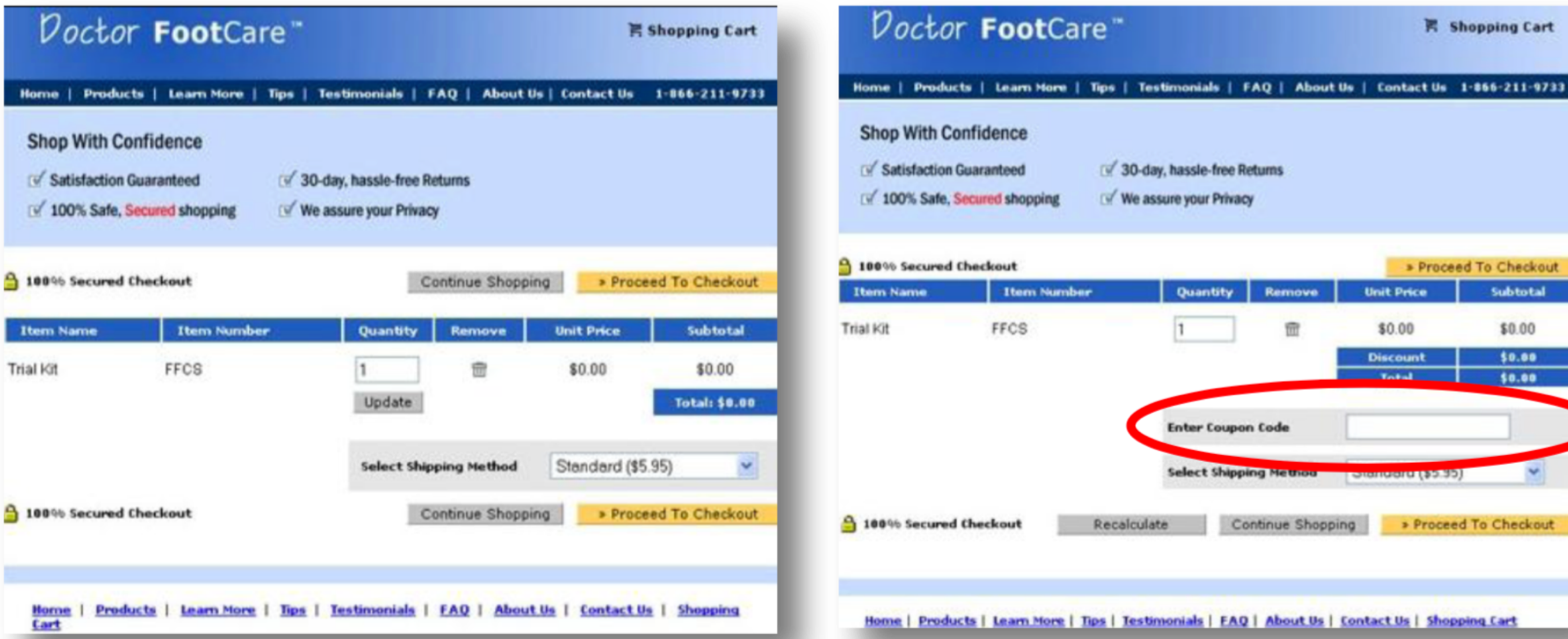
B

Kumar et al. 2009

[1]Reference: Bruno Ribeiro, CS57300 (2016)

Example with Two Solutions

- Which page has a higher conversion rate?



A B

Kumar et al. 2009

- With B, site lost 90% of revenue: users want to find coupons to reduce price

[1]Reference: Bruno Ribeiro, CS57300 (2016)

A/B Testing Setup

- First we will ignore the online aspect of the problem
- That is, we will ignore instantaneous feedback

- We will only use these feedbacks at the end of a period
- In particular,
  - They will be used to decide on good recommendation policies

A/B Testing Setup

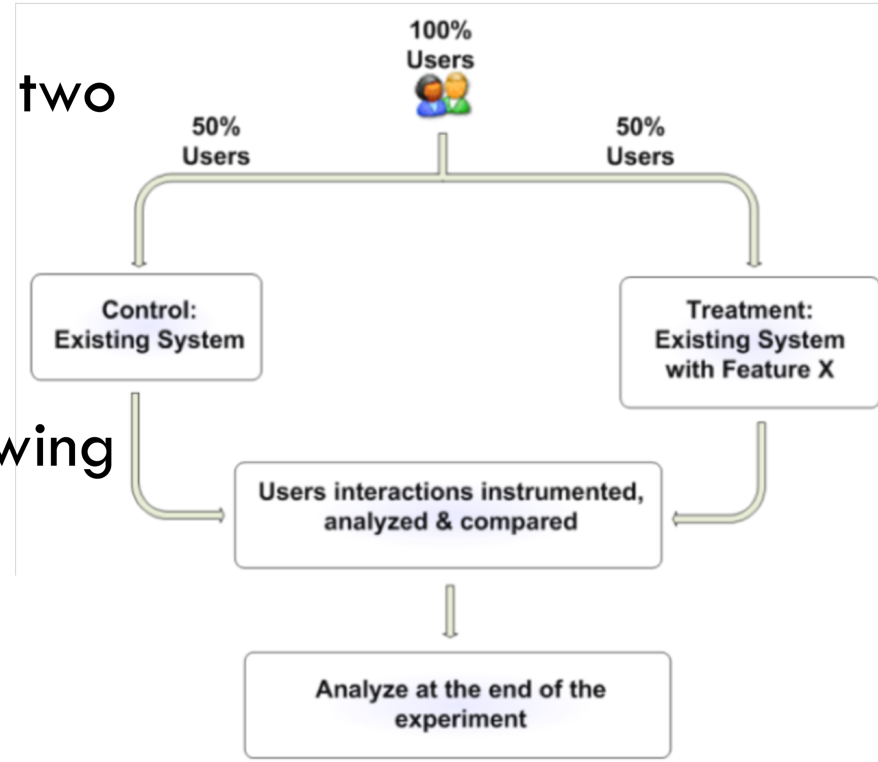
- A/B testing is about showing users two solutions

- And figuring out if solution A is different than solution B

[1]Reference: http://hbx.hbs.edu/blog/post/word-of-the-week-ab-testing

A/B Testing Setup

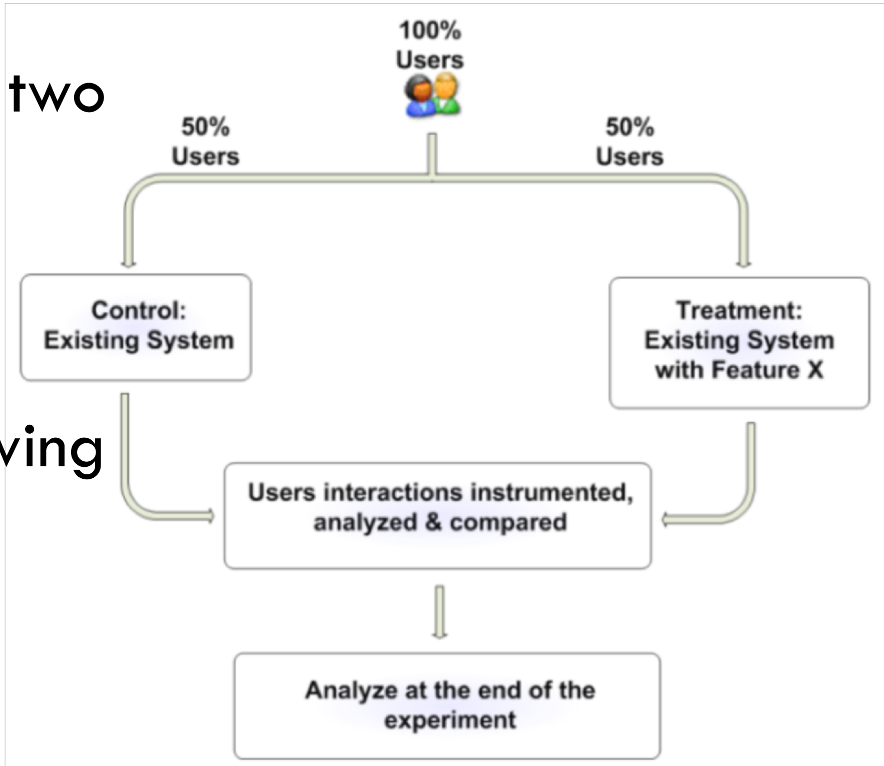
- A/B testing is about showing users two solutions
  - A (control)
  - B (treatment)
- Randomly split the users while showing



[1]Reference: http://alexdeng.github.io/public/files/Amazon%20Tech%20Talk.pdf

A/B Testing Setup

- A/B testing is about showing users two solutions
  - A (control)
  - B (treatment)
- Randomly split the users while showing
- Collect the outcomes and decide which option was better
  - Best scientific way to establish cause-effect relationship
  - Compared to offline data analysis (error prone)



[1] Reference: http://alexdeng.github.io/public/files/Amazon%20Tech%20Talk.pdf

A/B Testing is Two Sample Testing

- A/B testing is about collecting statistics across two groups

- Randomized assignment of the two solutions to each user is a key requirement
  - Eliminates biases and confounding

- Say each group of users has true mean effect μ_1 and μ_2
- From data, we want to infer weather
  - These are different (statistical significance)?
  - Same?
  - Which is larger?

[1]Reference: Bruno Ribeiro, CS57300 (2016)

Types of Hypothesis Tests

- Fisher
  - Reject H_0 (no acceptance as such)
  - More data typically leads to rejection

- Neyman-Pearson
  - Compare H_0 to H_1
  - Find likelihood ratio $P(Data|H_0)/P(Data|H_1)$

- Bayesian
  - Compute $P(H_0|Data)/P(H_1|Data)$
  - Similar to Neyman-Pearson when $P(H_0) = P(H_1)$

[1]Reference: Bruno Ribeiro, CS57300 (2016)

A/B Testing Pros

- Very intuitive setup and conclusions
- Field experiment decides the worth of a feature/offering, not gut instinct
- Most used in industry! (compared to bandit techniques)
  - Also called split or bucket testing
- Need not be a one time process
  - Can repeat if you think users have changed in terms of their preferences

A/B Testing Cons

- Has many bells and whistles to make it work
  - Especially because most treatment effects show small incremental improvement
  - See http://exp-platform.com for an extensive list of issues that affect A/B testing

- What if we can change who sees what treatment (action) dynamically?
  - Leads to Multi-Armed Bandit problems.

- What if we want to optimize over several options dynamically depending on context?
  - Leads to Contextual Bandit problems.

Questions?

Today's Outline

- Online Machine Learning
- A/B Testing
- Multi-armed bandits
- Contextual bandits

Bandit Problems

The Multi-armed Bandit Problem

- Multi-armed bandit (MAB) problem involves the following in each interaction

- pulling an arm = making a choice (which ad/color to display)
- reward/regret = measure of success (user-click, item-buy)

[1]Reference: Sumeet Kataria (2013)

The Formal Setting

Problem Formulation

Consider K arms (actions) each correspond to an unknown distribution $\{\nu_k\}_{k=1}^K$ with values bounded in $[0,1]$.

- At each time t, the agent pulls an arm $I_t \in \{1,\dots,K\}$ and observes a reward $x_t \sim \nu_{I_t}$ (i.i.d. sample from ν_{I_t}).
- The objective is to maximize the expected sum of rewards.

Notations

- mean of each arm: $\mu_k = \mathbb{E}_{X\sim\nu_k}[X]$
- mean of the best arm: $\mu^* = \max_k \mu_k$

MAB Performance

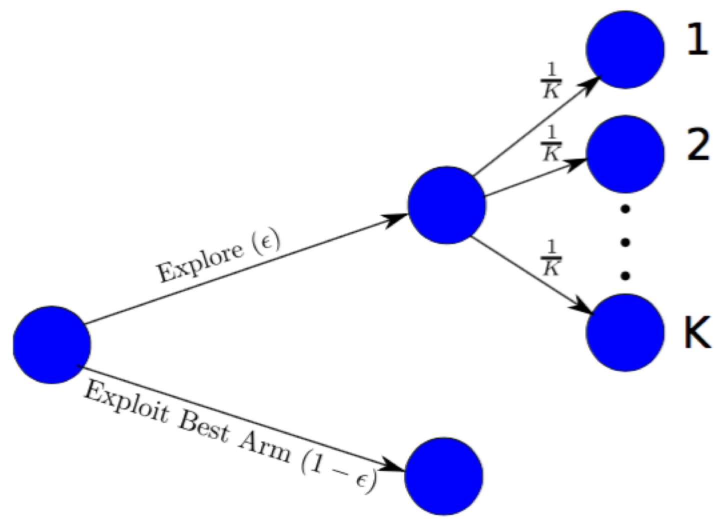
- It is an online problem.
- We need to come up with algorithms/strategies.
  - Example:
    - a round-robin strategy
    - A constant strategy (bad idea!)

To evaluate the performance of a strategy

Cumulative Regret $$R_n = n\mu^* - \sum_{t=1}^{n} x_t$$

Objective: find a strategy with small *expected cumulative regret* $\mathbb{E}[R_n]$

The Epsilon-Greedy Algorithm



Strategy = ϵ·Scientist $+(1-\epsilon)$·Businessman

At each time t

- With probability $1-\epsilon$, pick the subjectively best arm
- With probability $\frac{\epsilon}{K}$, pick a random arm

[1]Reference: Sumeet Kataria (2013)

The Epsilon-Greedy Algorithm Intuition

- How can we do well? We need to explore the arms. We also need to exploit what we have learned so far.

Scientist View

- Explore new ideas

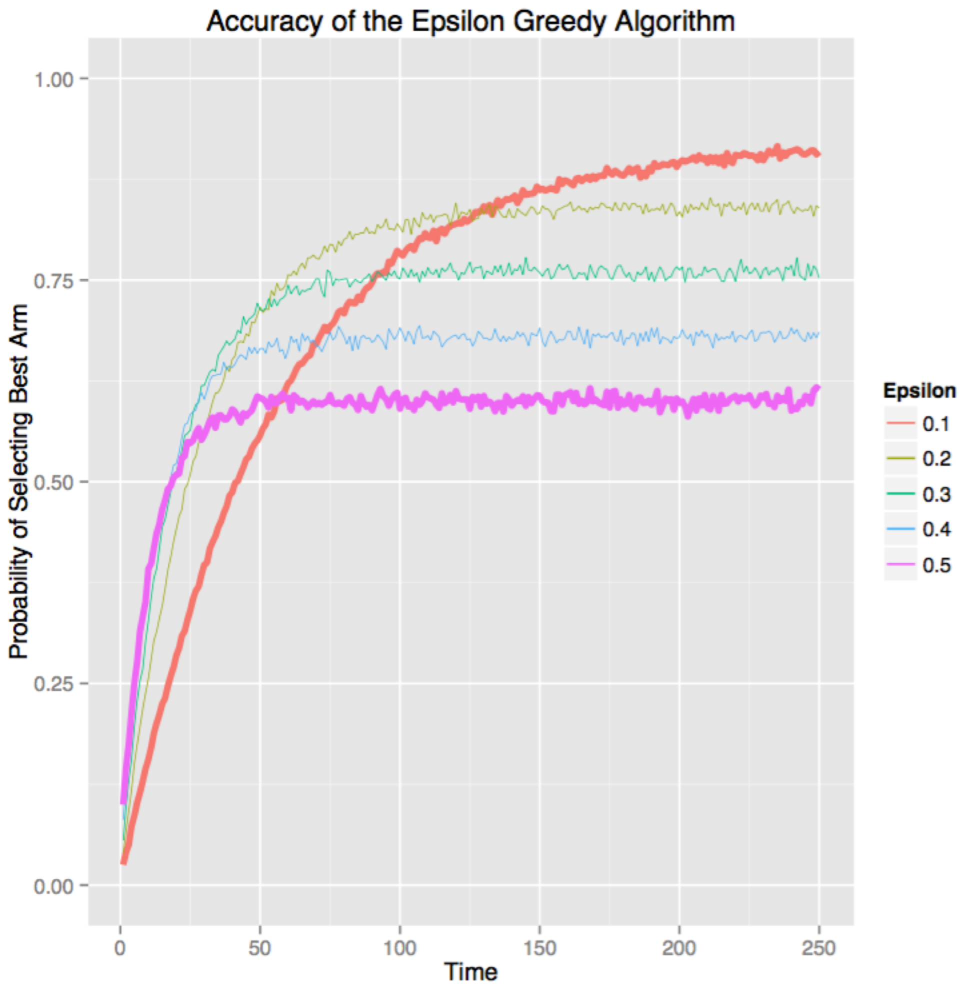
Businessman View

- Exploit best idea found so far

[1]Reference: Sumeet Kataria (2013)

Epsilon-Greedy Synthetic Experiment

5 Bernoulli arms with reward probabilities $0.1, 0.1, 0.1, 0.1, 0.9$



$\epsilon = 0.1$(Businessman)

- Learns slowly
- Does well at the end

$\epsilon = 0.5$(Scientist)

- Learns quickly
- Doesn't exploit at the end

[1]Reference: Sumeet Kataria (2013)

The Upper Confidence Bound (UCB) Algorithm

- Lets look at a slightly more involved algorithm: UCB

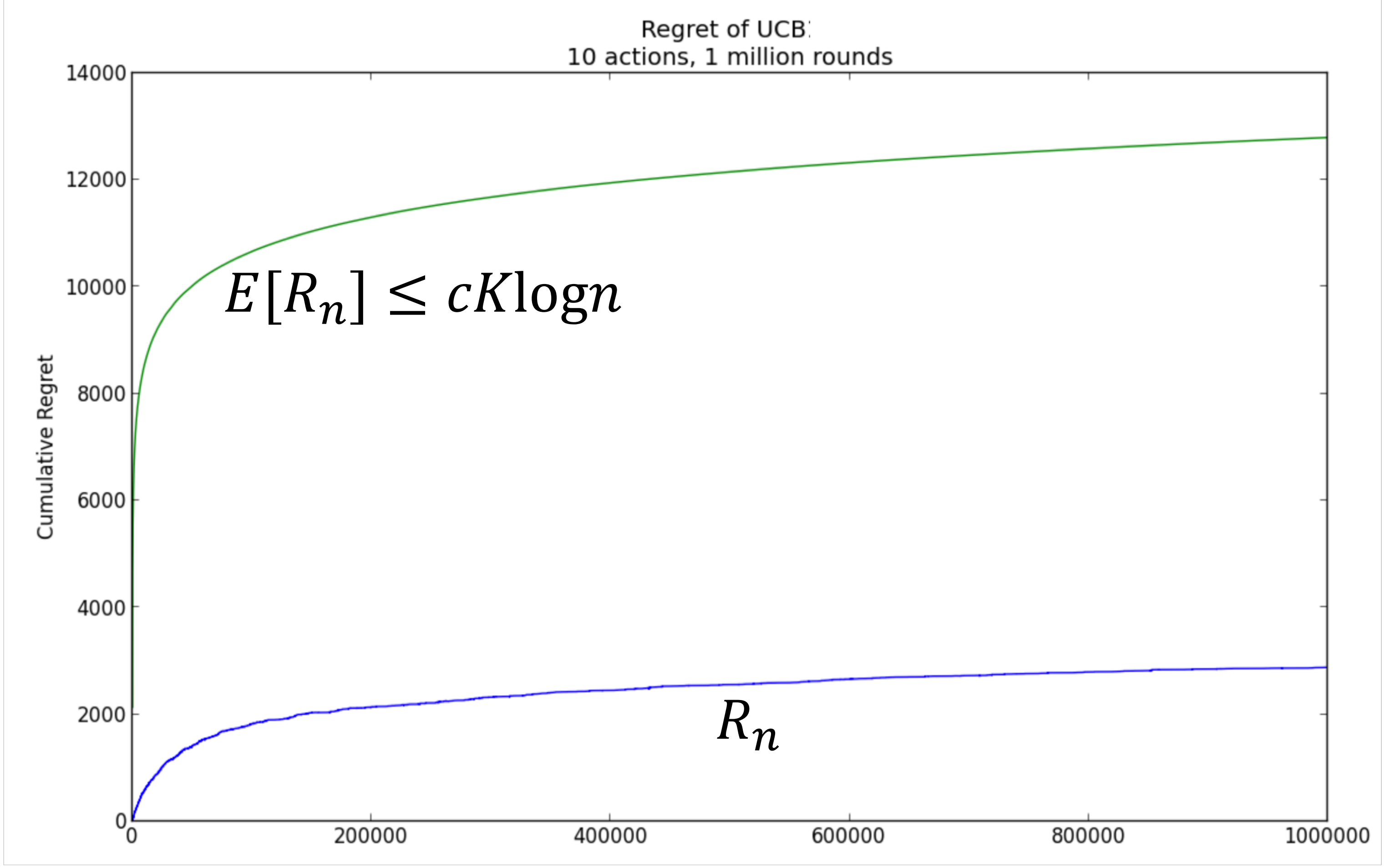
Upper confidence bound (UCB) strategy selects an arm at time t that

$$I_t = \arg\max_k B_{t,T_k(t-1)}(k) \quad , \quad B_{t,s}(k) = \hat{\mu}_{k,s} + \sqrt{\frac{2\log t}{s}}$$

$\hat{\mu}_{k,s} = \frac{1}{s}\sum_{i=1}^{s} x_{k,i}$ is the empirical mean of arm k at time s

UCB Synthetic Experiment

- 10 actions, 10^6 interactions (is this realistic?)
- Reward for each action has mean $0.5/k$ $(5 \leq k \leq 15)$



[1]Reference: https://jeremykun.files.wordpress.com/2013/10/ucb1-simple-example.png

The Thompson Sampling Algorithm

- A Bayesian algorithm for MAB problems is as follows

In Thompson [1933] the following strategy was proposed for the case of Bernoulli distributions:

- Assume a uniform prior on the parameters $\mu_i \in [0, 1]$.
- Let $\pi_{i,t}$ be the posterior distribution for μ_i at the t^{th} round.
- Let $\theta_{i,t} \sim \pi_{i,t}$ (independently from the past given $\pi_{i,t}$).
- $I_t \in \mathrm{argmax}_{i=1,\dots,K}\theta_{i,t}$.

Thompson Sampling: Conjugate Priors

A family of prior distribution

$$\mathcal{P}_A = \{p_\alpha(\theta) \mid \alpha \in A\}$$

is said to be **conjugate** to a model $\mathcal{P}_\Theta$, if, for a sample

$$X^{(1)}, \ldots, X^{(n)} \overset{\text{i.i.d.}}{\sim} p_\theta \qquad \text{with} \qquad p_\theta \in \mathcal{P}_\Theta,$$

the distribution q defined by

$$q(\theta) = p(\theta | x^{(1)}, \ldots, x^{(n)}) = \frac{p_\alpha(\theta) \prod_i p_\theta\left(x^{(i)}\right)}{\int p_\alpha(\theta) \prod_i p_\theta\left(x^{(i)}\right) d\theta}$$

is such that

$$q \in \mathcal{P}_A.$$

[1] Reference: http://imagine.enpc.fr/%7Eobozinsg/stats_review.html

Thompson Sampling: Conjugate Priors

We say that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ follows the Dirichlet distribution and note

$$\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$$

for $\boldsymbol{\theta}$ in the simplex $\triangle_K = \{\mathbf{u} \in \mathbb{R}_+^K \mid \sum_{k=1}^K u_k = 1\}$ and

[1]Reference: http://imagine.enpc.fr/%7Eobozinsg/stats_review.html

Thompson Sampling: Conjugate Priors

We say that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ follows the Dirichlet distribution and note

$$\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$$

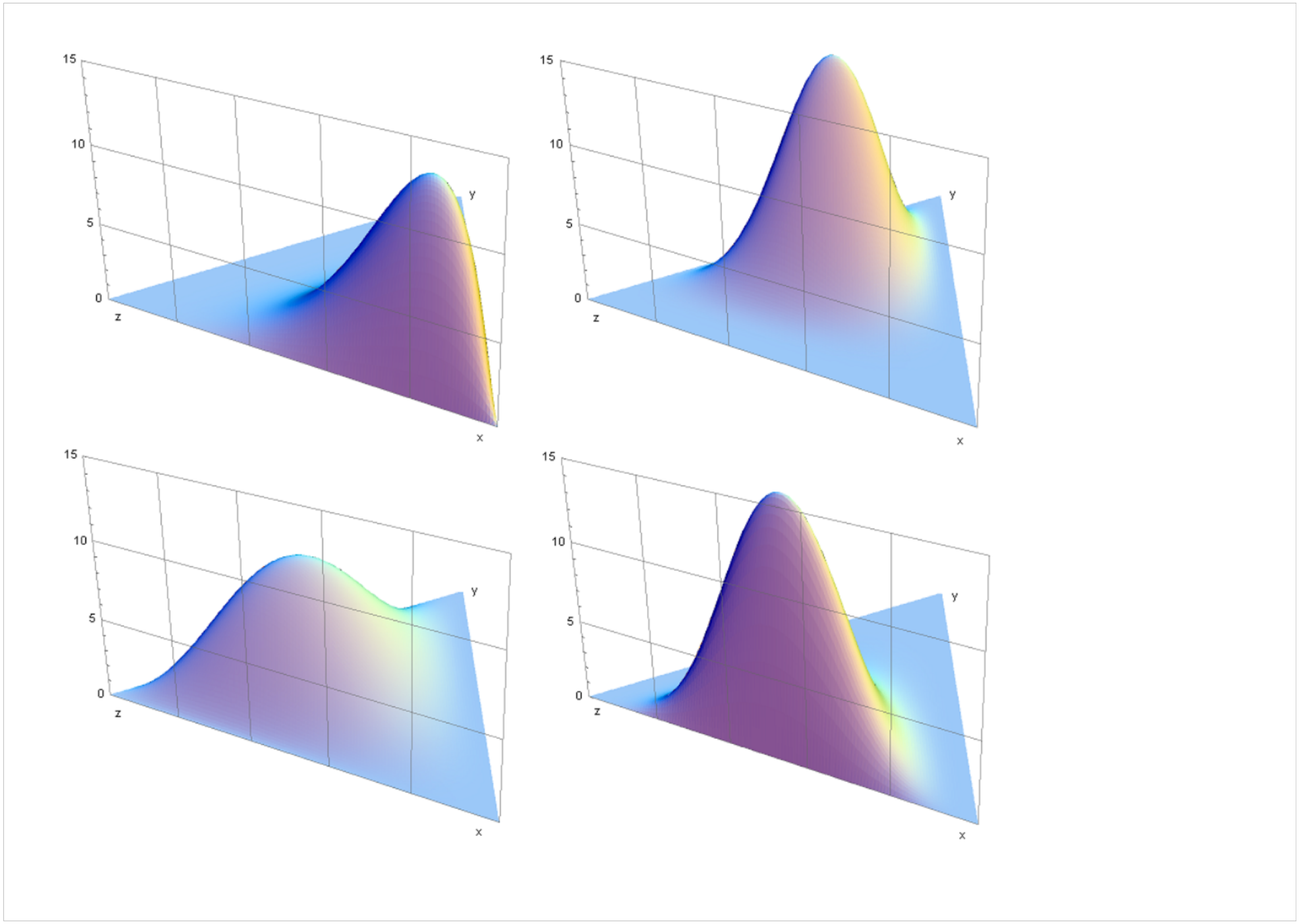
for $\boldsymbol{\theta}$ in the simplex $\triangle_K = \{\mathbf{u} \in \mathbb{R}_+^K \mid \sum_{k=1}^K u_k = 1\}$ and admitting the density

$$p(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_k \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \ldots \theta_K^{\alpha_K - 1}$$

$$\alpha_0 = \sum_k \alpha_k \quad \text{and} \quad \Gamma(x) := \int_0^\infty t^{x-1} e^{-t} dt$$

[1] Reference: http://imagine.enpc.fr/%7Eobozinsg/stats_review.html

Thompson Sampling: Conjugate Priors



[1]Reference: http://imagine.enpc.fr/%7Eobozinsg/stats_review.html

Thompson Sampling: Categorical-Dirichlet Conjugacy

Consider the simple Bayesian Dirichlet-Multinomial model with

- A Dirichlet prior on the parameter of the multinomial: $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
- A multinomial random variable $\mathbf{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad \text{and} \qquad p(\mathbf{z}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{z_k}$$

[1] Reference: http://imagine.enpc.fr/%7Eobozinsg/stats_review.html

Thompson Sampling: Categorical-Dirichlet Conjugacy

Consider the simple Bayesian Dirichlet-Multinomial model with

- A Dirichlet prior on the parameter of the multinomial: $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
- A multinomial random variable $\mathbf{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad \text{and} \qquad p(\mathbf{z}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{z_k}$$

Let $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$ be an i.i.d. sample distributed like $\mathbf{z}$.
We have

$$p(\boldsymbol{\theta}|\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}) = \frac{p(\boldsymbol{\theta}) \prod_n p(\mathbf{z}^{(n)}|\theta)}{p(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)})}$$

[1]Reference: http://imagine.enpc.fr/%7Eobozinsg/stats_review.html

Thompson Sampling: Categorical-Dirichlet Conjugacy

Consider the simple Bayesian Dirichlet-Multinomial model with

- A Dirichlet prior on the parameter of the multinomial: $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
- A multinomial random variable $\mathbf{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad \text{and} \qquad p(\mathbf{z}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{z_k}$$

Let $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$ be an i.i.d. sample distributed like $\mathbf{z}$.
We have

$$p(\boldsymbol{\theta}|\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}) = \frac{p(\boldsymbol{\theta}) \prod_n p(\mathbf{z}^{(n)}|\theta)}{p(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)})} \propto \prod_k \theta_k^{\alpha_k + \sum_n z_{nk} - 1}$$

So that $(\theta|(Z)) \sim \mathrm{Dir}((\alpha_1 + N_1, \ldots, \alpha_K + N_K))$ with $N_k = \sum_n z_{nk}$

Non-Probabilistic Setting

- Why do we need to assume that the rewards are i.i.d.?
- Can we drop the stochastic assumptions on the rewards?

- Reason #1: These rewards may be the output of a complex process
- Reason #2: These rewards may be generated by an 'adversary' (someone who is not random)

Non-Probabilistic Setting

- We can in fact drop the probabilistic reward assumption!

- Template
  - Adversary selects rewards $x_t(1), \dots, x_t(K)$, which are not known to the player (us)
  - Player selects arm I_t
  - In full information, player sees $x_t(1), \dots, x_t(K)$
  - In bandit information setup, player only sees $x_t(I_t)$

Exp3 Algorithm

Initialization: $w_1(k) = 1$ for all $k = 1, \ldots, K$

Exp3 Algorithm

Initialization: $w_1(k) = 1$ for all $k = 1, \ldots, K$

At each time $t = 1, \ldots, n$: the player selects an arm $I_t \sim p_t$, where

$$p_t(k) = (1-\gamma)\frac{w_t(k)}{\sum_{i=1}^{K} w_t(i)} + \frac{\gamma}{K}$$

Exp3 Algorithm

Initialization: $w_1(k) = 1$ for all $k = 1, \ldots, K$

At each time $t = 1, \ldots, n$: the player selects an arm $I_t \sim p_t$, where

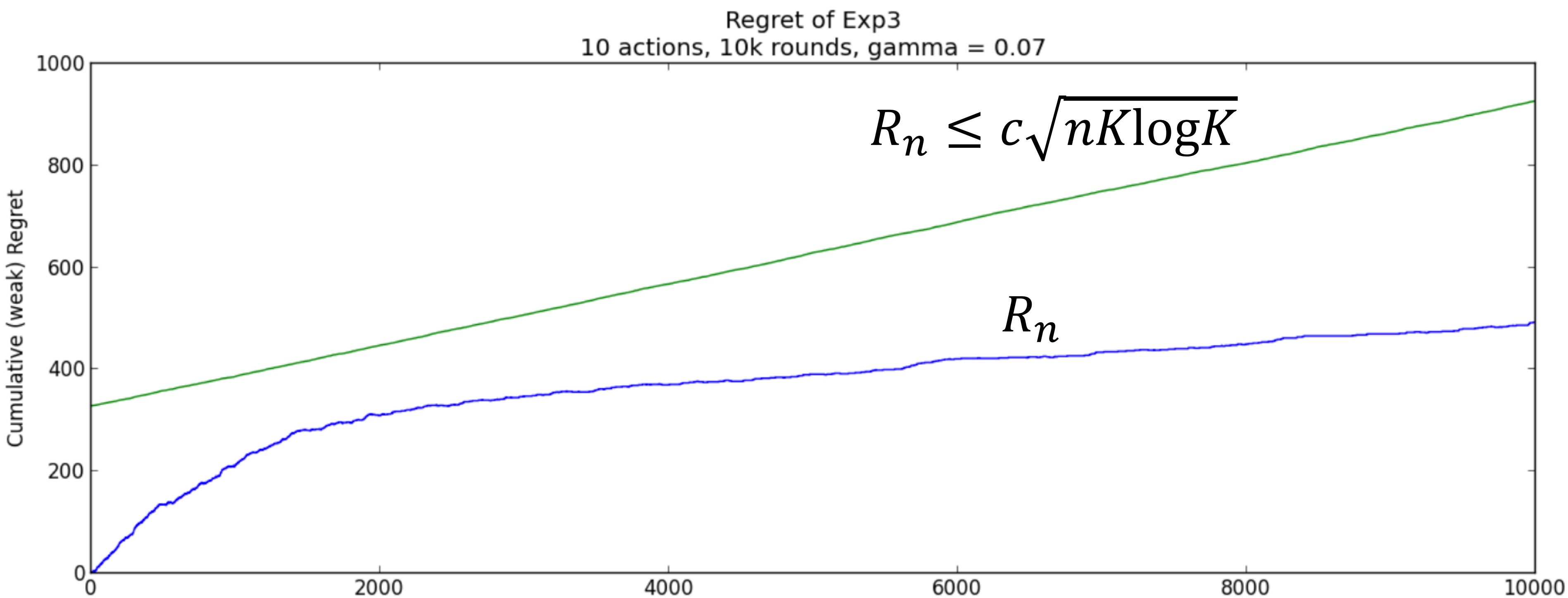
$$p_t(k) = (1-\gamma)\frac{w_t(k)}{\sum_{i=1}^{K} w_t(i)} + \frac{\gamma}{K}$$

with $w_t(k) = e^{\eta \sum_{s=1}^{t-1} \tilde{x}_s(k)}$, where $\tilde{x}_s(k) = \frac{x_s(k)}{p_s(k)} \mathbf{1}\{I_s = k\}$.

$\eta > 0$ and $\gamma > 0$ are the parameters of the algorithm.

Exp3 Synthetic Experiment

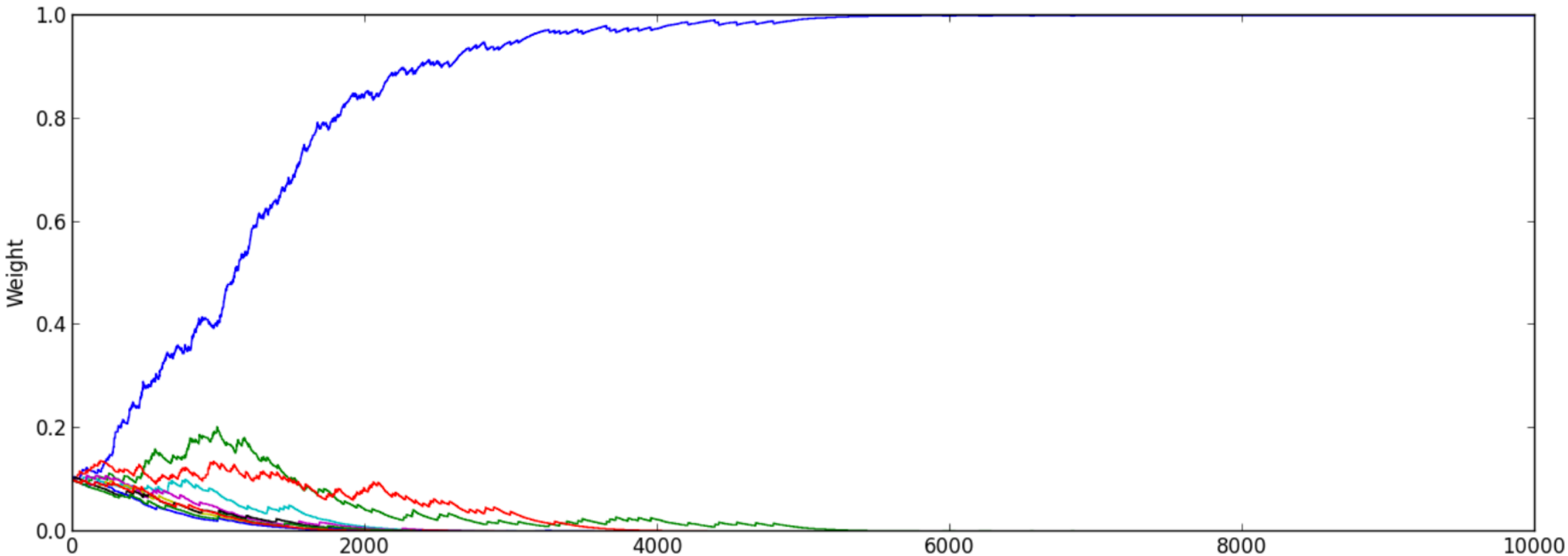
- 10 actions, 10^3 interactions
- Reward for each action is Bernoulli with means $1/k$ $(2 \leq k < 12)$



[1]Reference: https://jeremykun.files.wordpress.com/2013/11/exp3-regret-graph.png

Exp3 Synthetic Experiment

- 10 actions, 10^3 interactions
- Reward for each action is Bernoulli with means $1/k$ $(2 \leq k < 12)$



[1]Reference: https://jeremykun.files.wordpress.com/2013/11/exp3-regret-graph.png

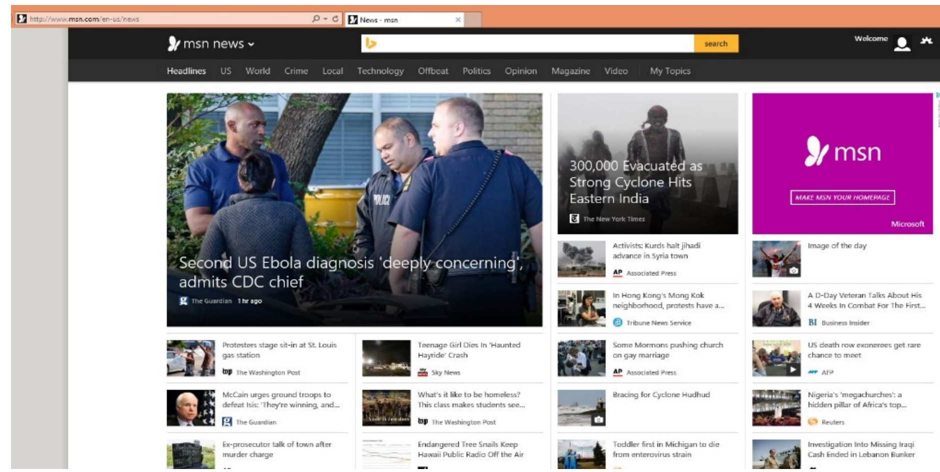
Questions?

Bandits with Contexts

Recall: MSN Deployment for Personalized News

Loop:

1. User **arrives** at MSN with browsing history, user account, previous visits,…
2. Microsoft **chooses** news stories, …
3. User **responds** to content (clicks, navigation, etc)



Goal: Choose content to yield desired user behavior

Assumption: Recommendations to one user do not affect other users

[1]Reference: Alekh Agarwal et al., http://arxiv.org/abs/1606.03966

Previous Bandit Models are not Enough

- No context!

- No-carry over effect from one interaction to the next
  - Say users can change behavior by seeing recommendations
  - Can be captured by Reinforcement Learning

The Contextual Bandit Problem

- In the Contextual Bandit problem,
  - Every round, we get context
  - We want to find the best policy (what to do in each context)
  - May not see the same context twice!

- Different from MAB setting because in MAB problems
  - No context
  - We were finding a single best action

[1]Reference: John Langford (2011)

Benefit of Context

- Say we have 5 ads

$$a_1 = \text{“buy pet lizards”}$$
$$a_2 = \text{“1-800-petunias”}$$
$$a_3 = \text{“cheap mp3 players”}$$
$$a_4 = \text{“find local florists”}$$
$$a_5 = \text{“affordable dragon souls”},$$

- Say we have 4 policies
  - These map context to ads
- Now, lets look at one round of Exp3
  - For Exp3, it is as if it has 4 “arms” (one per policy)

[1]Reference: http://courses.cs.washington.edu/courses/cse599s/12sp/scribes/lecture13.pdf

Benefit of Context

- In round t say the policies recommend the following:

e_1 **chose** a_2

e_2 **chose** a_2

e_3 **chose** a_4

e_4 **chose** a_4

$a_1 =$ **"buy pet lizards"**

$a_2 =$ **"1-800-petunias"**

$a_3 =$ **"cheap mp3 players"**

$a_4 =$ **"find local florists"**

$a_5 =$ **"affordable dragon souls"**

- Say Exp3 chose "arm" e_1 by sampling from weights
- And, say e_1's ad choice a_2 was clicked

[1] Reference: http://courses.cs.washington.edu/courses/cse599s/12sp/scribes/lecture13.pdf

Benefit of Context

- Exp3 assigns reward $\tilde{x}_s(e_1) = \frac{x_s(e_1)}{p_s(e_1)}$
- Rest of the arms all get reward 0

e_1	**chose**	a_2
e_2	**chose**	a_2
e_3	**chose**	a_4
e_4	**chose**	a_4

- Can we do better?
  - Yes! e_2 also was recommending a_2
  - We should better estimate reward of e_2

[1]Reference: http://courses.cs.washington.edu/courses/cse599s/12sp/scribes/lecture13.pdf

Exp4 Algorithm

Initialization: $\forall \pi \in \Pi : w_t(\pi) = 1$

For each $t = 1, 2, \ldots$:

1. Observe x_t and let for $a = 1, \ldots, K$

$$p_t(a) = \frac{\sum_\pi \mathbf{1}[\pi(x_t) = a]\, w_t(\pi)}{\sum_\pi w_t(\pi)}$$

[1]Reference: John Langford (2011)

Exp4 Algorithm

Initialization: $\forall \pi \in \Pi : w_t(\pi) = 1$

For each $t = 1, 2, \ldots$:

1. Observe x_t and let for $a = 1, \ldots, K$

$$p_t(a) = \frac{\sum_\pi \mathbf{1}[\pi(x_t) = a]\, w_t(\pi)}{\sum_\pi w_t(\pi)} + p_{\min},$$

[1]Reference: John Langford (2011)

Exp4 Algorithm

Initialization: $\forall \pi \in \Pi : w_t(\pi) = 1$

For each $t = 1, 2, \ldots$:

1. Observe x_t and let for $a = 1, \ldots, K$

$$p_t(a) = (1 - Kp_{\min}) \frac{\sum_\pi \mathbf{1}[\pi(x_t) = a]\, w_t(\pi)}{\sum_\pi w_t(\pi)} + p_{\min},$$

[1] Reference: John Langford (2011)

Exp4 Algorithm

Initialization: $\forall \pi \in \Pi : w_t(\pi) = 1$

For each $t = 1, 2, \ldots$:

1. Observe x_t and let for $a = 1, \ldots, K$

$$p_t(a) = (1 - Kp_{\min}) \frac{\sum_\pi \mathbf{1}[\pi(x_t) = a]\, w_t(\pi)}{\sum_\pi w_t(\pi)} + p_{\min},$$

where $p_{\min} = \sqrt{\frac{\ln |\Pi|}{KT}}$.

[1]Reference: John Langford (2011)

Exp4 Algorithm

Initialization: $\forall \pi \in \Pi : w_t(\pi) = 1$

For each $t = 1, 2, \ldots$:

1. Observe x_t and let for $a = 1, \ldots, K$

$$p_t(a) = (1 - Kp_{\min}) \frac{\sum_\pi \mathbf{1}[\pi(x_t) = a] \, w_t(\pi)}{\sum_\pi w_t(\pi)} + p_{\min},$$

   where $p_{\min} = \sqrt{\frac{\ln |\Pi|}{KT}}$.

2. Draw a_t from p_t, and observe reward $r_t(a_t)$.

[1]Reference: John Langford (2011)

Exp4 Algorithm

Initialization: $\forall \pi \in \Pi : w_t(\pi) = 1$

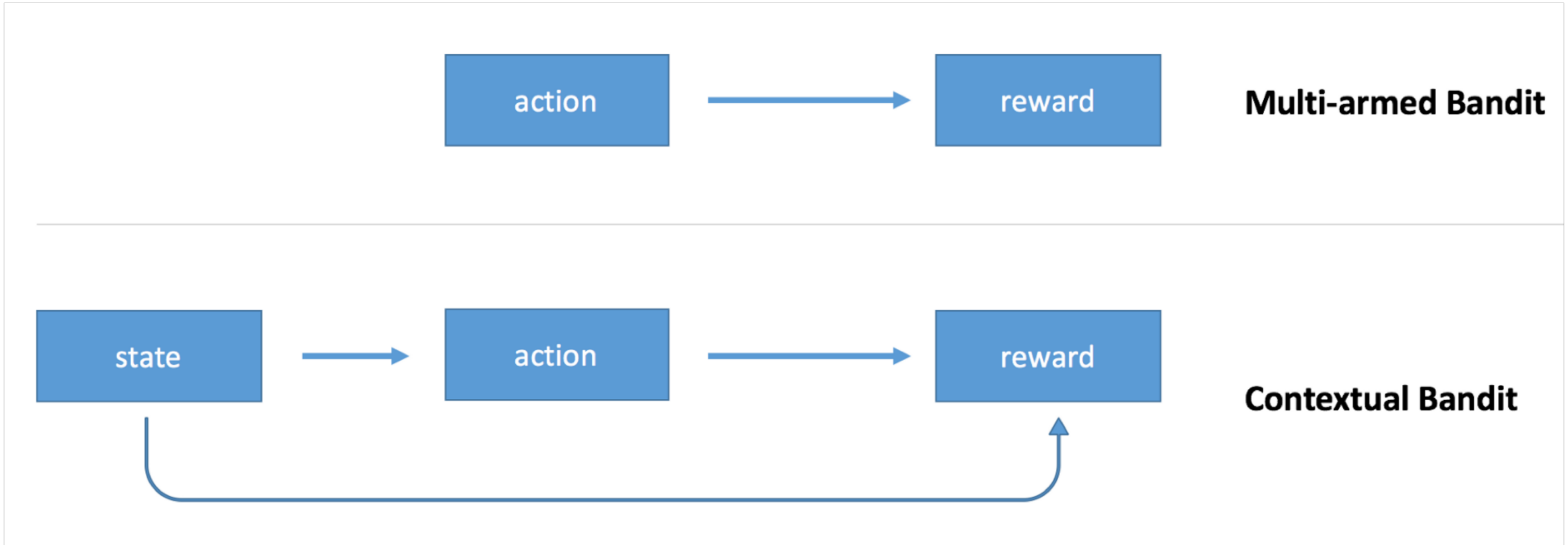
For each $t = 1, 2, \ldots$:

1. Observe x_t and let for $a = 1, \ldots, K$
$$p_t(a) = (1 - Kp_{\min}) \frac{\sum_\pi \mathbf{1}[\pi(x_t) = a] \, w_t(\pi)}{\sum_\pi w_t(\pi)} + p_{\min},$$
where $p_{\min} = \sqrt{\frac{\ln |\Pi|}{KT}}$.
2. Draw a_t from p_t, and observe reward $r_t(a_t)$.
3. Update for each $\pi \in \Pi$
$$w_{t+1}(\pi) = \begin{cases} w_t(\pi) \exp\left(p_{\min} \frac{r_t(a_t)}{p_t(a_t)}\right) & \text{if } \pi(x_t) = a_t \\ w_t(\pi) & \text{otherwise} \end{cases}$$

[1]Reference: John Langford (2011)

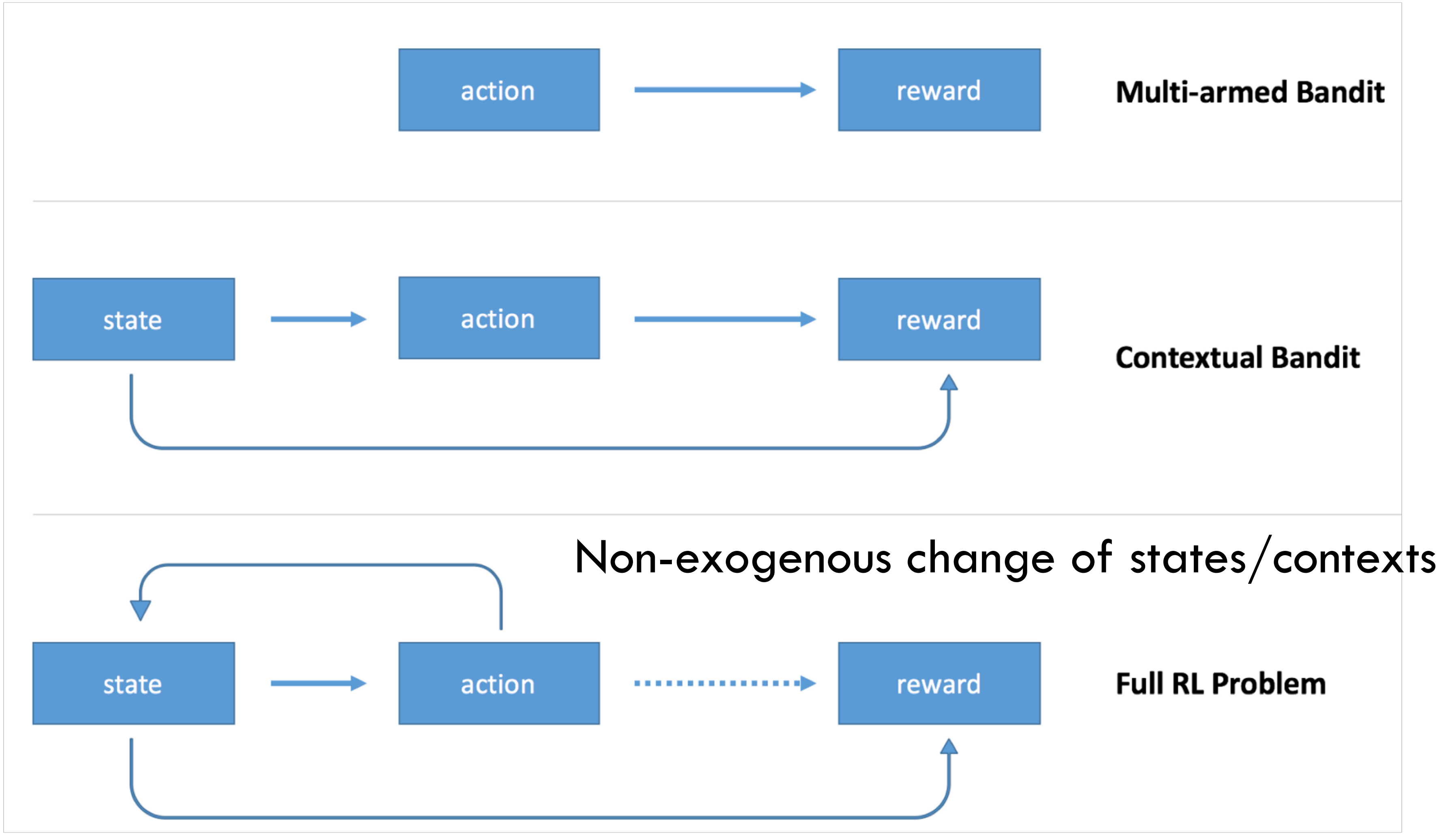
Questions?

Reinforcement Learning: Because Contextual Bandit Formulation is not Enough



[1]Reference: https://medium.com/@awjuliani/simple-reinforcement-learning-with-tensorflow-part-1-5-contextual-bandits-bff01d1aad9c

Reinforcement Learning: Because Contextual Bandit Formulation is not Enough



[1]Reference: https://medium.com/@awjuliani/simple-reinforcement-learning-with-tensorflow-part-1-5-contextual-bandits-bff01d1aad9c

Summary

- We looked at A/B testing as a way to introduce enhancements in a business product/service
  - May need a lot of examples
  - Is based on the idea of randomized control trials

- We also looked at two new online ML problems
  - Multi-Armed Bandits
  - Contextual Bandits

- Contextual bandits are a special case of reinforcement learning, which we will study next time.

Appendix

Sample Exam Questions

- What is the difference between A/B testing and Multi-armed bandits?
- Can we do A/B testing when we have more than two options?
- What is the role of exploration in the Bandit problems?
- Can Exp3 be used in a stochastic setting?
- How does the contextual problem differ from the non-contextual problem?

Online ML is Difficult to Deploy

- Separate teams for each part of the process
- Faulty logging
  - Logging just choice, not probabilities
  - Features not logged and change in time
- Runtime behavior incompatible with the ML
  - Business logic overriding randomization
  - Using the probability as feature for downstream ML
- Subtle errors that are difficult to find in complex systems!

[1]Reference: Alekh Agarwal et al., http://arxiv.org/abs/1606.03966

Contextual Bandit: Website Example

Repeatedly:

1. A user comes to Yahoo! (with history of previous visits, IP address, data related to his Yahoo! account)

[1] Reference: John Langford (2011)

Contextual Bandit: Website Example

Repeatedly:

1. A user comes to Yahoo! (with history of previous visits, IP address, data related to his Yahoo! account)

2. Yahoo! chooses information to present (from urls, ads, news stories)

[1]Reference: John Langford (2011)

Contextual Bandit: Website Example

Repeatedly:

1. A user comes to Yahoo! (with history of previous visits, IP address, data related to his Yahoo! account)
2. Yahoo! chooses information to present (from urls, ads, news stories)
3. The user reacts to the presented information (clicks on something, clicks, comes back and clicks again, et cetera)

[1]Reference: John Langford (2011)

Contextual Bandit: Website Example

Repeatedly:

1. A user comes to Yahoo! (with history of previous visits, IP address, data related to his Yahoo! account)

2. Yahoo! chooses information to present (from urls, ads, news stories)

3. The user reacts to the presented information (clicks on something, clicks, comes back and clicks again, et cetera)

Yahoo! wants to interactively choose content and use the observed feedback to improve future content choices.

[1]Reference: John Langford (2011)

Contextual Bandit: Clinical Example

"I stopped taking the medicine because I prefer the original disease to the side effects."

Repeatedly:

1. A patient comes to a doctor with symptoms, medical history, test results
2. The doctor chooses a treatment
3. The patient responds to it

The doctor wants a policy for choosing targeted treatments for individual patients.

[1] Reference: John Langford (2011)

Additional Resources

- Course at UWash:
  - http://courses.cs.washington.edu/courses/cse599s/12sp/scribes.html (lectures 13,14)
- Course at UCSD:
  - http://cseweb.ucsd.edu/~kamalika/teaching/CSE291W11/ (lecture5)
- Tutorial by Bygelzimer and Langford:
  - http://hunch.net/~exploration_learning/
- Course at UAlberta:
  - https://sites.ualberta.ca/~szepesva/CMPUT654/

Note: These are optional. May be slightly theoretical in nature.