

Online and Offline Experimentation in Complex Systems

Akshay Krishnamurthy
Microsoft Research, NYC
akshay@cs.umass.edu

Online Personalization

The screenshot shows a user interface for the msn lifestyle website. At the top, there's a navigation bar with the msn logo, a dropdown menu for 'lifestyle', a search icon, the user's name 'Akshay' with a profile icon, and a settings gear icon. Below the navigation, there are category links: Home, Cooking, Restaurants & News (which is underlined), Beverages, and Video. The main content area features several news cards. One card on the left shows a group of people at a social gathering with the caption 'Food & Wine 8 hrs ago'. Another card in the center is a sponsored post from 'The Motley Fool' about a new smartphone, featuring the Apple logo and the headline 'Say Goodbye to iPhone: This Could Be 40X Better'. A third card on the right shows a box of Girl Scout Cookies with the headline 'Girl Scout Cookies Are Officially on Sale' and the source 'Food & Wine'. At the bottom left, there's a snippet for 'Trader Joe's new bagged french fry chip mashup totally lives...' from 'TODAY'.

- Learn from interacting with users in production

Online Personalization

The screenshot shows a user interface for the msn lifestyle website. At the top, there's a navigation bar with the msn logo, a dropdown menu for 'lifestyle', a search icon, the user's name 'Akshay' with a profile icon, and a settings gear icon. Below the navigation, there are category links: Home, Cooking, Restaurants & News (which is underlined), Beverages, and Video. The main content area features several news cards. One card on the left shows a group of people at a social gathering with the caption 'Food & Wine 8 hrs ago'. Another card in the center is a sponsored post from 'The Motley Fool' about a new smartphone, with the Apple logo and the headline 'Say Goodbye to iPhone: This Could Be 40X Better'. A third card on the right shows a box of Girl Scout Cookies with the headline 'Girl Scout Cookies Are Officially on Sale' and the source 'Food & Wine'. At the bottom left, there's a snippet for 'Trader Joe's new bagged french fry chip mashup totally lives...' from 'TODAY'.

- Learn from interacting with users in production
- No counterfactuals

Online Personalization

The screenshot shows a user interface for the msn lifestyle website. At the top, there's a navigation bar with the msn logo, a dropdown menu for 'lifestyle', a search icon, the user's name 'Akshay' with a profile icon, and a settings gear icon. Below the navigation, there are category links: Home, Cooking, Restaurants & News (which is underlined), Beverages, and Video. The main content area features several news cards. One card on the left shows a group of people at a social gathering with the caption 'Food & Wine 8 hrs ago'. Another card in the center is a sponsored post from 'The Motley Fool' titled 'Say Goodbye to iPhone: This Could Be 40X Better', featuring the Apple logo. A third card on the right is from 'Food & Wine' about Girl Scout Cookies. At the bottom left, there's a snippet from 'TODAY' about Trader Joe's new bagged french fry chip mashup.

- Learn from interacting with users in production
- No counterfactuals
- Exploration vs Exploitation

Online Personalization

The screenshot shows a personalized news feed on the msn lifestyle website. At the top, there's a navigation bar with the msn logo, a search icon, the user name 'Akshay' with a profile icon, and a settings gear icon. Below the navigation, there are category links: Home, Cooking, Restaurants & News (which is underlined), Beverages, and Video. The main content area displays several news items:

- A large image of a group of people at a social gathering, with the caption "Food & Wine 8 hrs ago".
- A sponsored post from "The Motley Fool" featuring the Apple logo and the headline "Say Goodbye to iPhone: This Could Be 40X Better".
- A news item from "TODAY" about "Trader Joe's new bagged french fry chip mashup totally lives...".
- A news item from "Food & Wine" about "Girl Scout Cookies Are Officially on Sale".

- Learn from interacting with users in production
- No counterfactuals
- Exploration vs Exploitation
- Optimize whole-page layout

Industry Standard: A/B Testing

The screenshot shows the msn lifestyle website interface. At the top, there's a navigation bar with the msn logo, a search icon, and a user profile for "Akshay". Below the navigation, there are category links: Home, Cooking, Restaurants & News (which is underlined), Beverages, and Video. The main content area displays a news article with a photo of people at a social gathering. The photo has a caption: "Food & Wine 8 hrs ago". To the right of this is a sponsored advertisement for Apple, featuring a large glowing Apple logo and the text "Say Goodbye to iPhone: This Could Be 40X Better". Below the ad is a sponsored post from "The Motley Fool". At the bottom of the screen, there's another news item about Trader Joe's bagged french fry chip mashup, followed by a snippet about Girl Scout Cookies.

msn | lifestyle

Home Cooking Restaurants & News Beverages Video

Food & Wine 8 hrs ago

Sponsored

Say Goodbye to iPhone: This Could Be 40X Better

Sponsored
The Motley Fool

Trader Joe's new bagged french fry chip mashup totally lives... TODAY

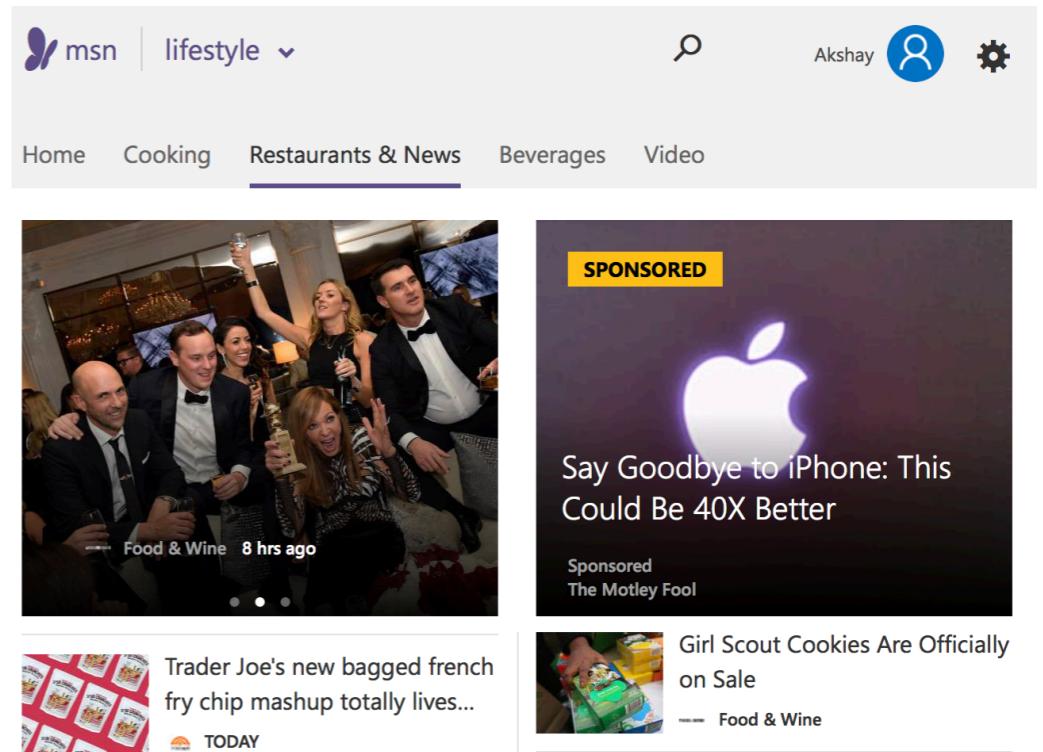
Girl Scout Cookies Are Officially on Sale

Food & Wine

Industry Standard: A/B Testing

Given policy π :

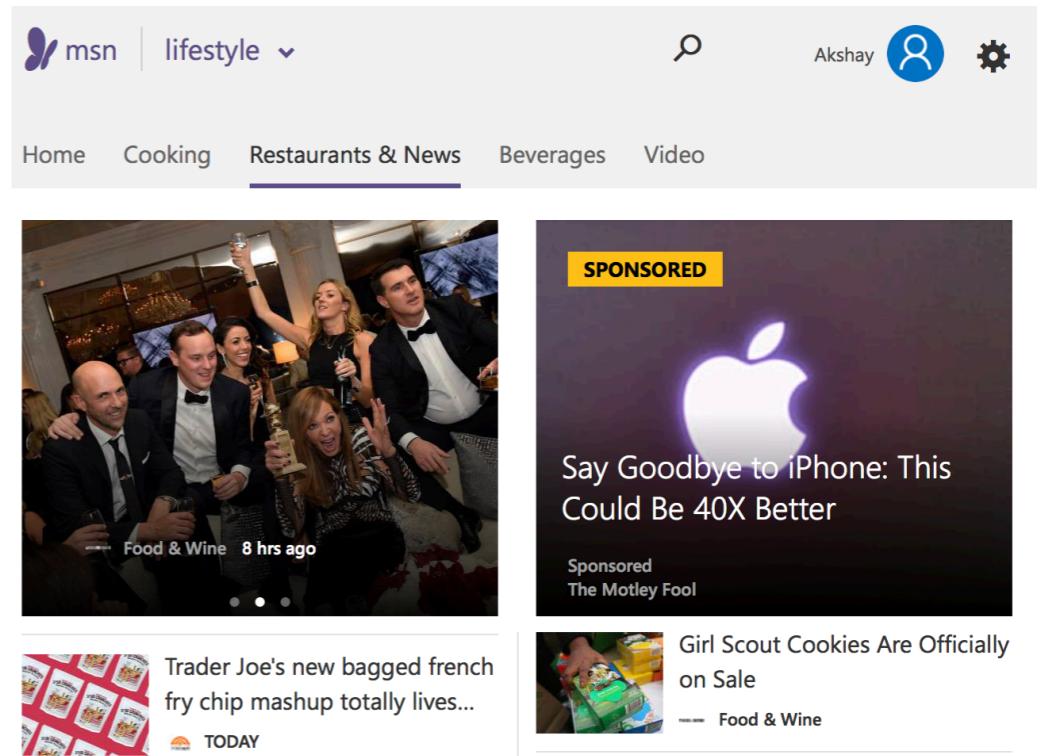
1. Use π for 1/2 of traffic (at random)
2. Evaluate π 's quality (click prob.)



Industry Standard: A/B Testing

Given policy π :

1. Use π for 1/2 of traffic (at random)
2. Evaluate π 's quality (click prob.)

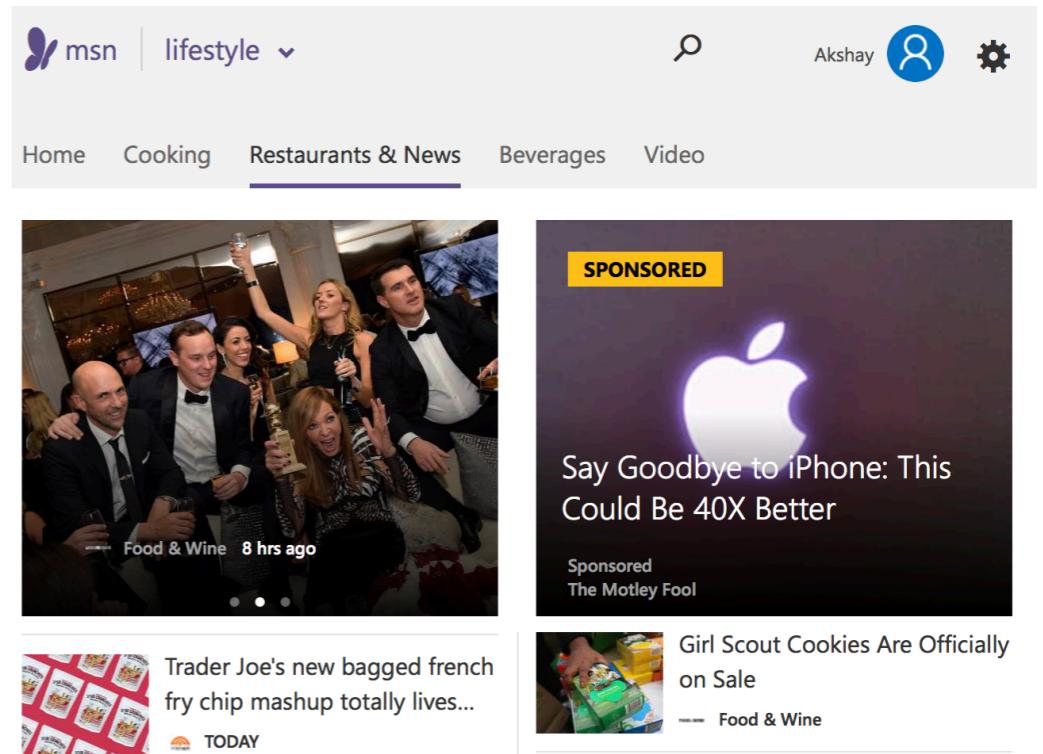


Two main issues:

Industry Standard: A/B Testing

Given policy π :

1. Use π for 1/2 of traffic (at random)
2. Evaluate π 's quality (click prob.)



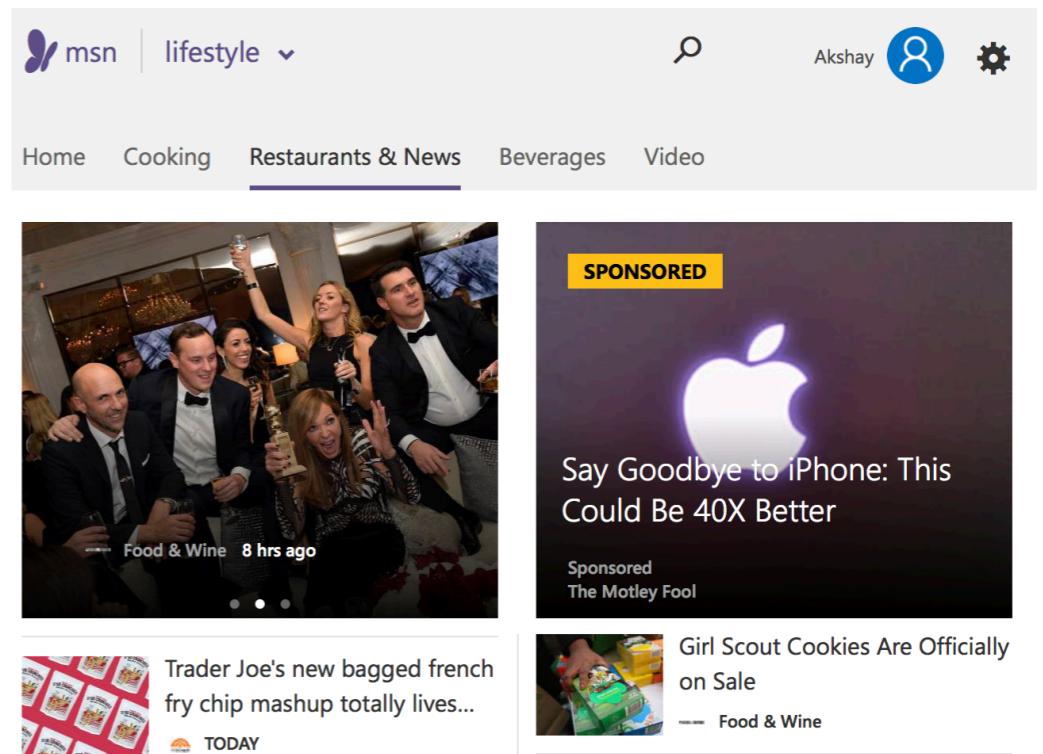
Two main issues:

1. Poor performance while evaluating policies

Industry Standard: A/B Testing

Given policy π :

1. Use π for 1/2 of traffic (at random)
2. Evaluate π 's quality (click prob.)



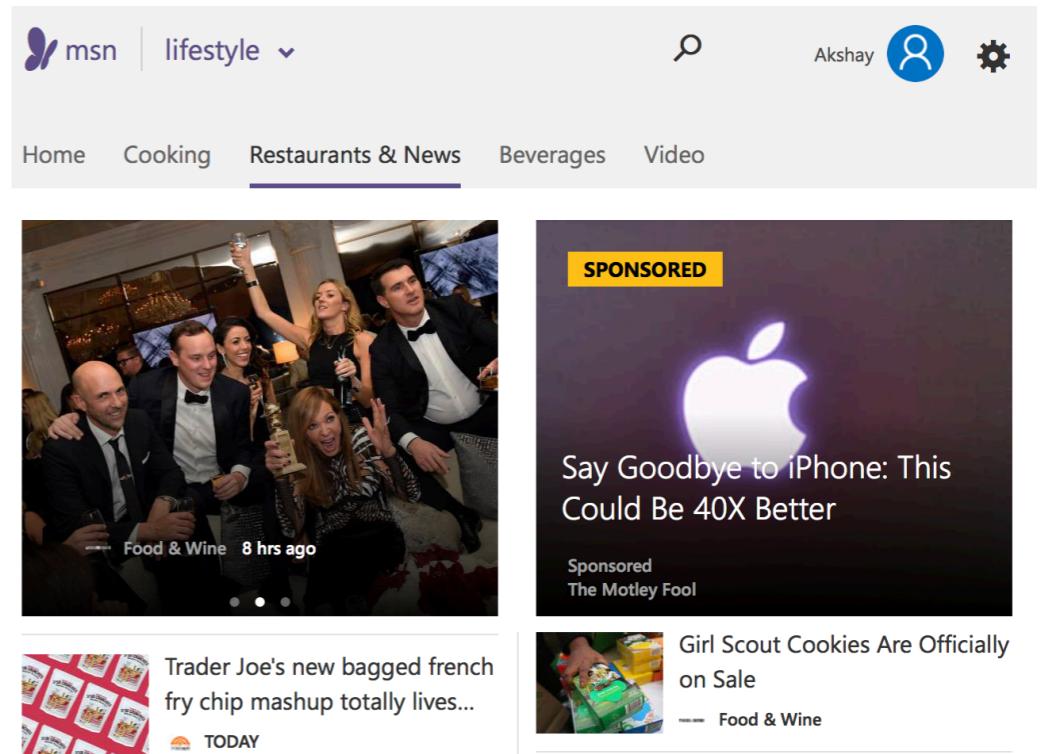
Two main issues:

1. Poor performance while evaluating policies
2. Requires $O(|\Pi|)$ samples to evaluate $|\Pi|$ policies

Industry Standard: A/B Testing

Given policy π :

1. Use π for 1/2 of traffic (at random)
2. Evaluate π 's quality (click prob.)



Two main issues:

1. Poor performance while evaluating policies
2. Requires $O(|\Pi|)$ samples to evaluate $|\Pi|$ policies

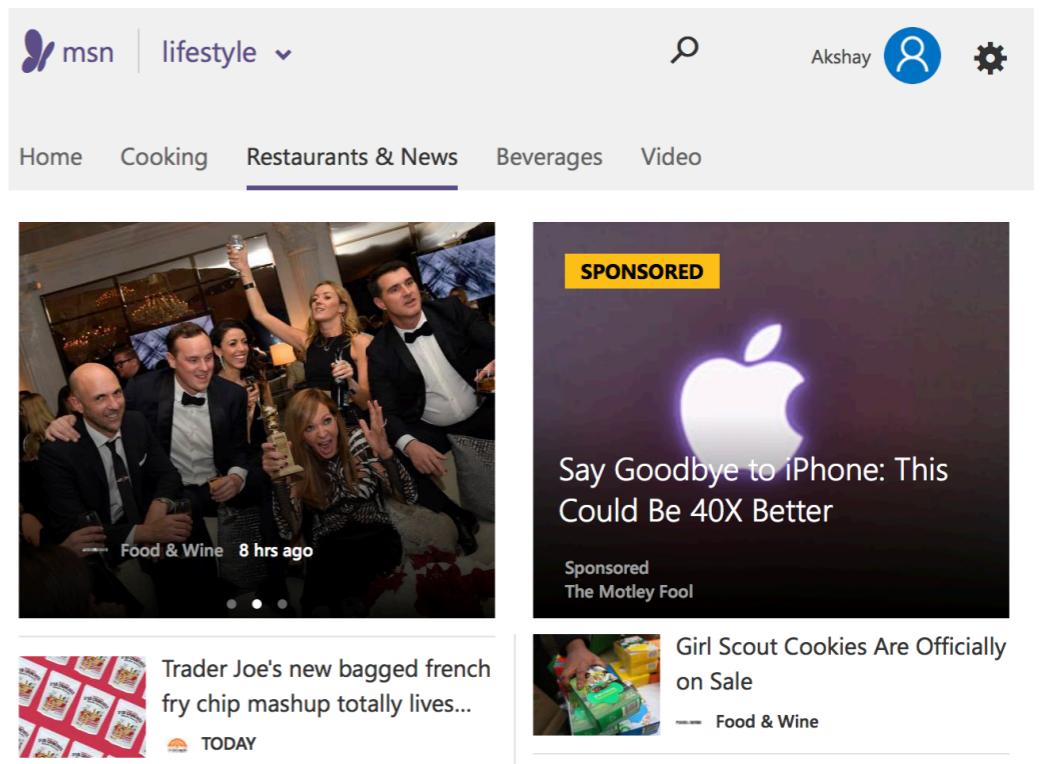
Can do **exponentially** better with contextual bandits!

Exploration + Offline Evaluation

The screenshot shows the msn lifestyle website interface. At the top, there is a navigation bar with the msn logo, a search icon, the user name 'Akshay' with a profile icon, and a settings gear icon. Below the navigation bar, there is a secondary navigation menu with links for Home, Cooking, Restaurants & News (which is underlined), Beverages, and Video. The main content area displays a news feed. The first item is a sponsored post from 'The Motley Fool' titled 'Say Goodbye to iPhone: This Could Be 40X Better', featuring a large Apple logo and a photo of people at a social gathering. The second item is a news article from 'TODAY' titled 'Trader Joe's new bagged french fry chip mashup totally lives...', accompanied by a small image of the product. The third item is a news article titled 'Girl Scout Cookies Are Officially on Sale', with a small image of Girl Scout cookies. The overall layout is clean and modern, typical of a digital news platform.

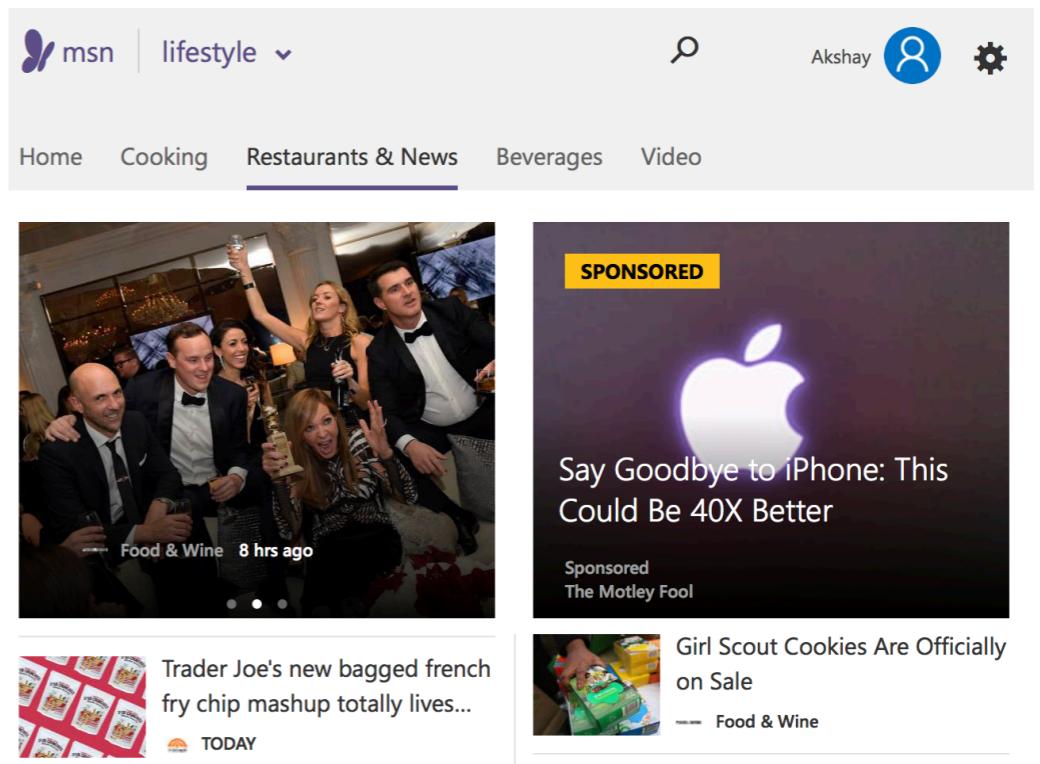
Exploration + Offline Evaluation

1. Collect dataset by serving content at random



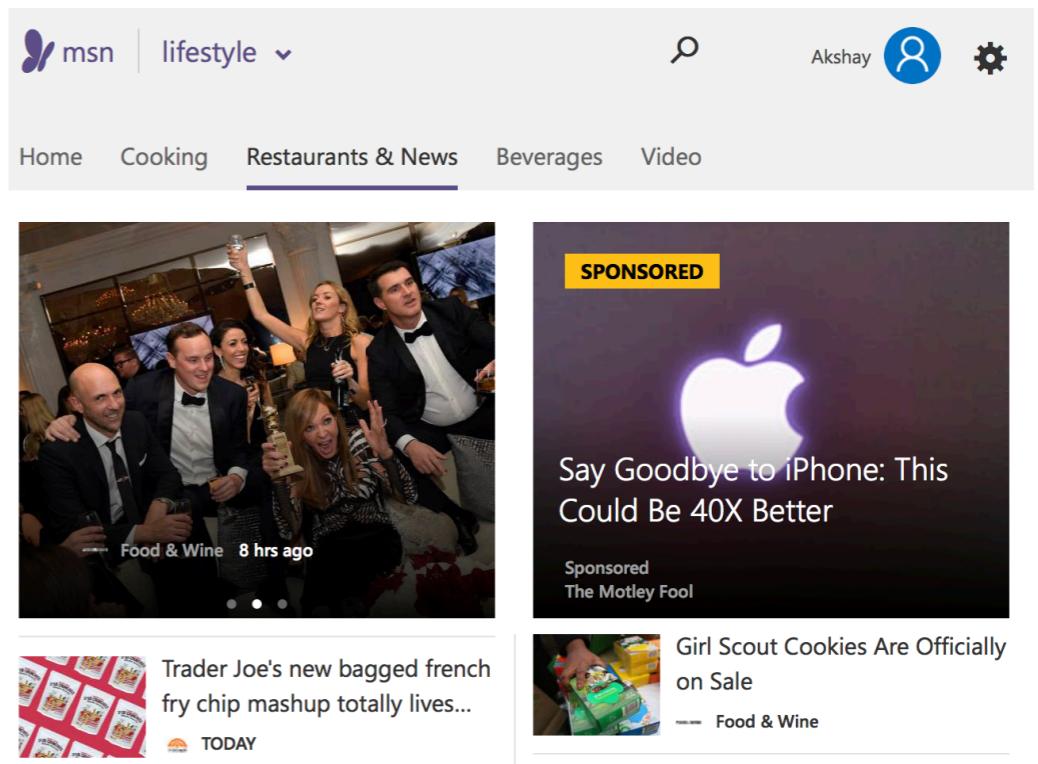
Exploration + Offline Evaluation

1. Collect dataset by serving content at random
2. For each policy, estimate performance by taking samples where we used its recommendation



Exploration + Offline Evaluation

1. Collect dataset by serving content at random
2. For each policy, estimate performance by taking samples where we used its recommendation



With K actions and $|\Pi|$ policies, we need $O(K \log |\Pi|)$ samples

Contextual Bandits

The screenshot shows a web browser window for the msn lifestyle site. The top navigation bar includes the msn logo, a dropdown menu for 'lifestyle', a search icon, a user profile for 'Akshay' with a blue circular icon, and a gear icon for settings.

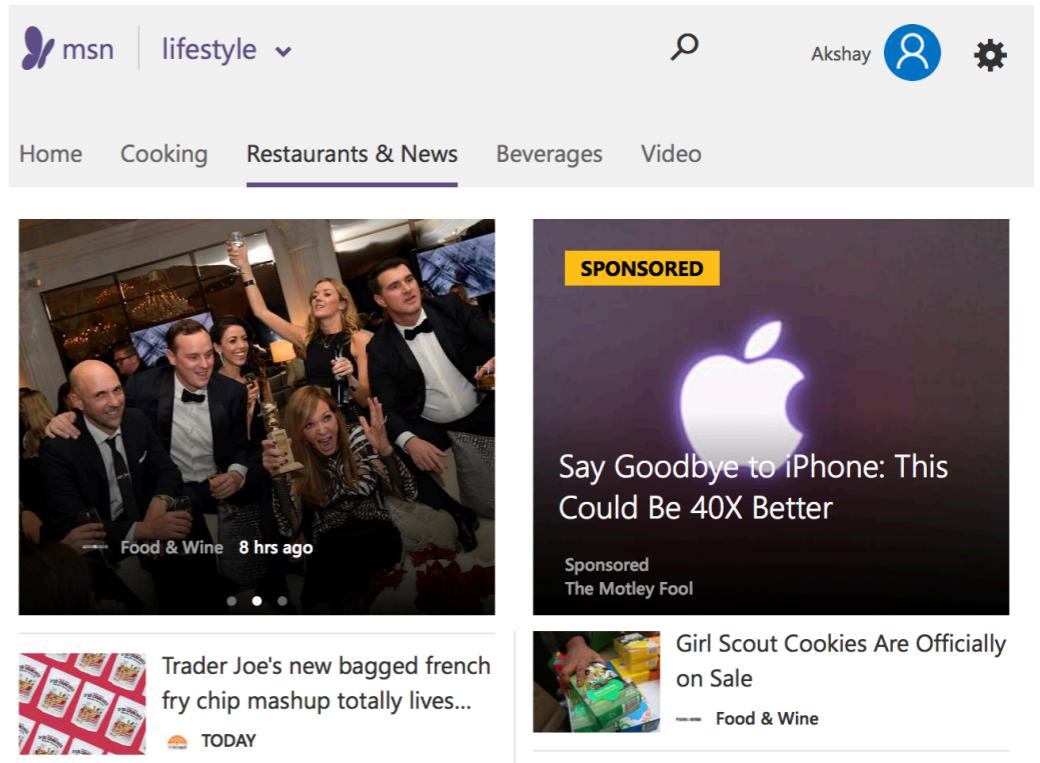
The main content area displays a news feed:

- Sponsored Content:** A large image of a group of people in formal attire at a social gathering, with a yellow 'SPONSORED' banner at the top. Below it, the headline reads "Say Goodbye to iPhone: This Could Be 40X Better" and is attributed to "Sponsored The Motley Fool".
- Food & Wine:** A smaller image of a person holding a bag of Trader Joe's bagged french fry chip mashup. The headline says "Trader Joe's new bagged french fry chip mashup totally lives..." and is dated "TODAY".
- Food & Wine:** A small image of Girl Scout cookies. The headline says "Girl Scout Cookies Are Officially on Sale" and is dated "Food & Wine".

Contextual Bandits

On each of T rounds:

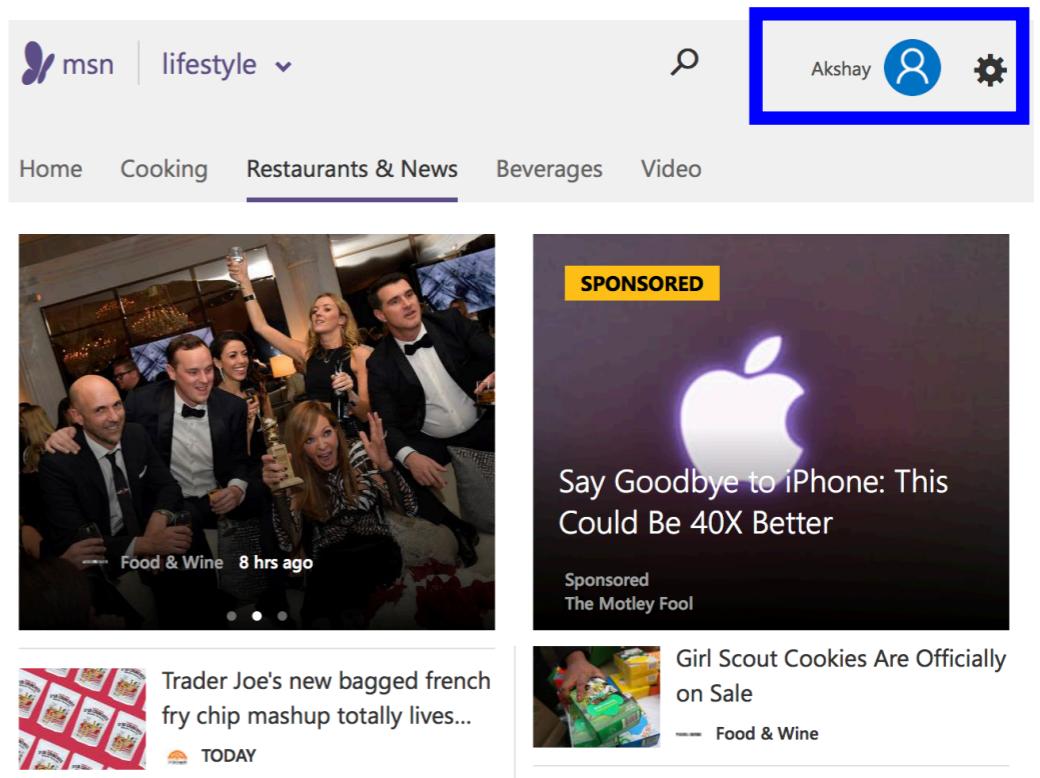
1. Observe context
2. Play action
3. Observe reward



Contextual Bandits

On each of T rounds:

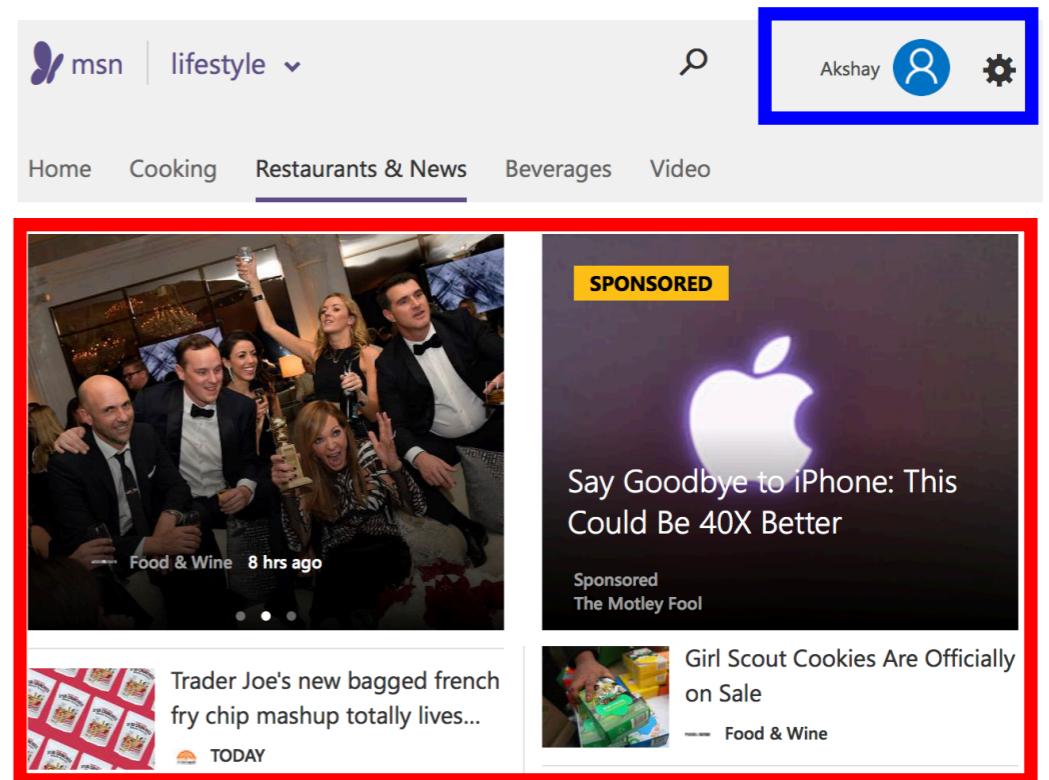
1. Observe context x_t
2. Play action
3. Observe reward



Contextual Bandits

On each of T rounds:

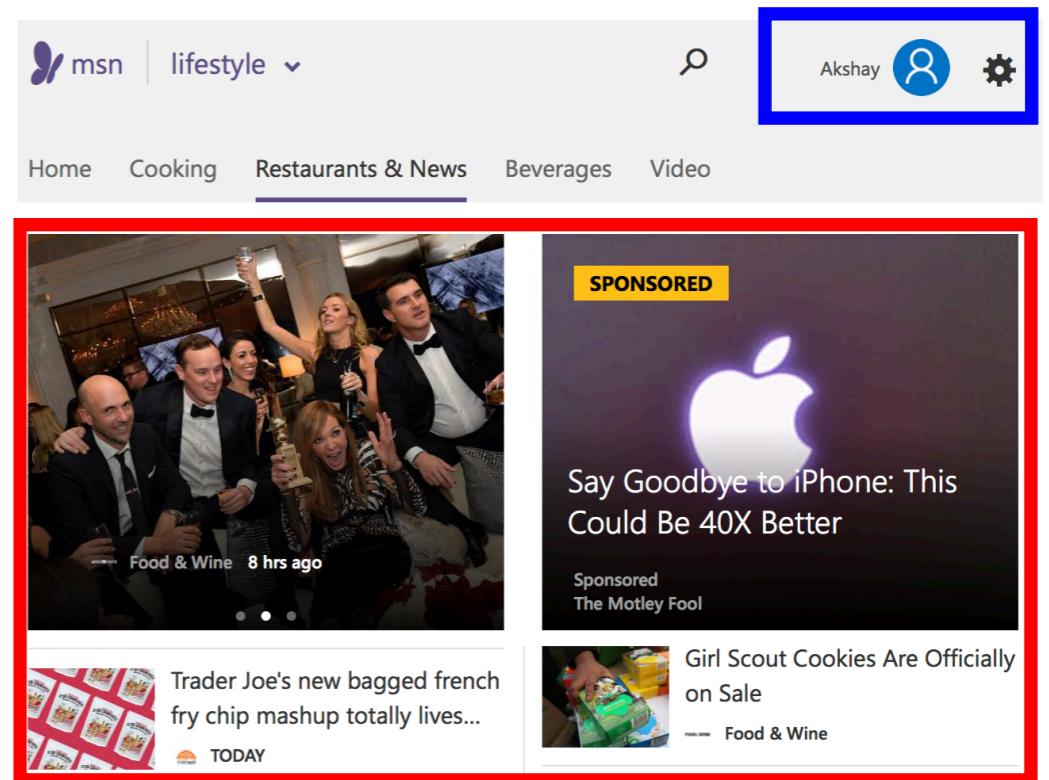
1. Observe context x_t
2. Play action a_t
3. Observe reward



Contextual Bandits

On each of T rounds:

1. Observe context x_t
2. Play action a_t
3. Observe reward $r_t(a_t, x_t)$



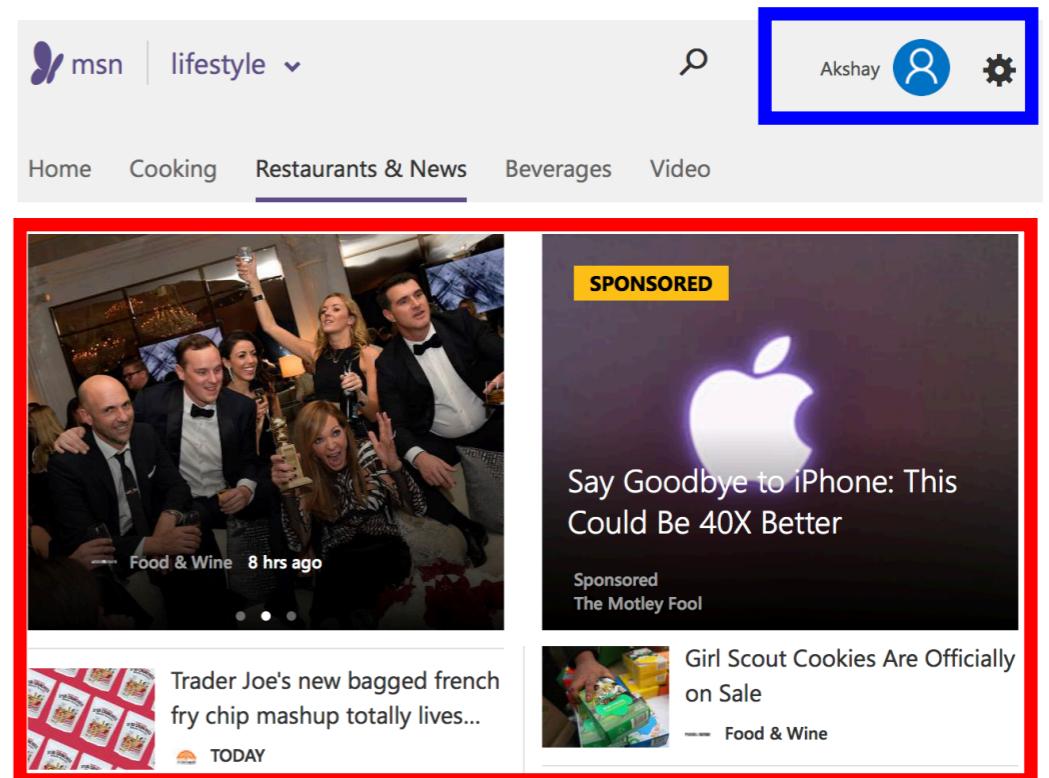
$$r_t = \# \text{ clicks}$$

Contextual Bandits

On each of T rounds:

1. Observe context x_t
2. Play action a_t
3. Observe reward $r_t(a_t, x_t)$

K = number of actions



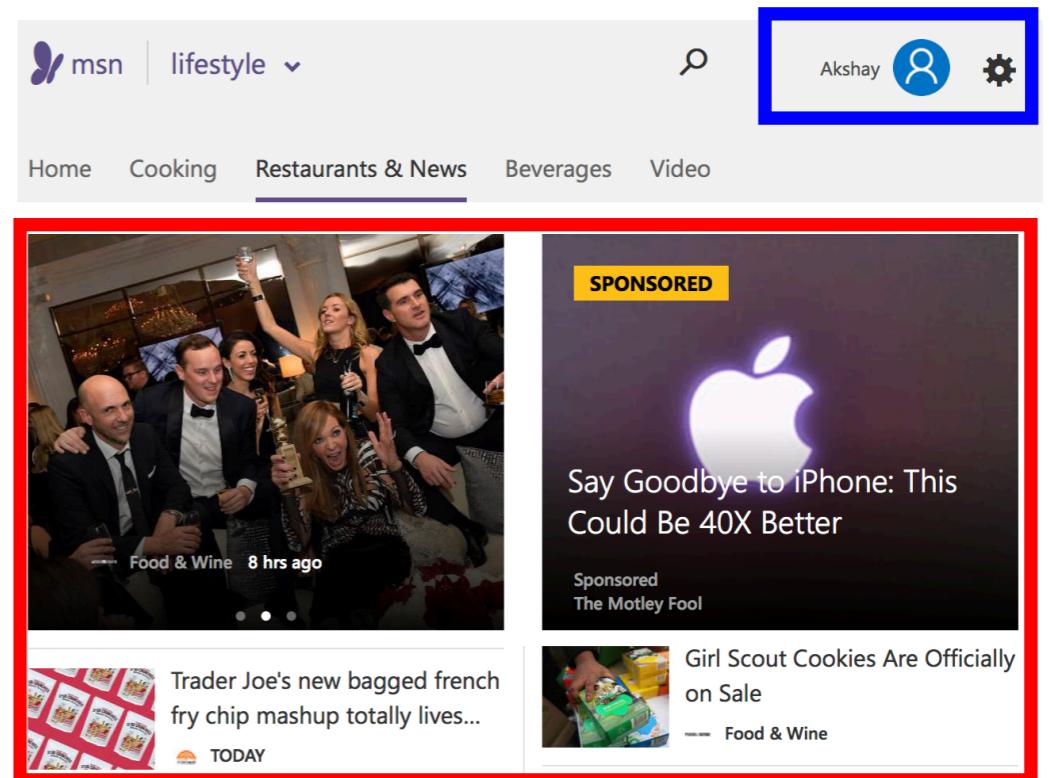
$$r_t = \# \text{ clicks}$$

Contextual Bandits

On each of T rounds:

1. Observe context x_t
2. Play action a_t
3. Observe reward $r_t(a_t, x_t)$

K = number of actions



$$r_t = \# \text{ clicks}$$

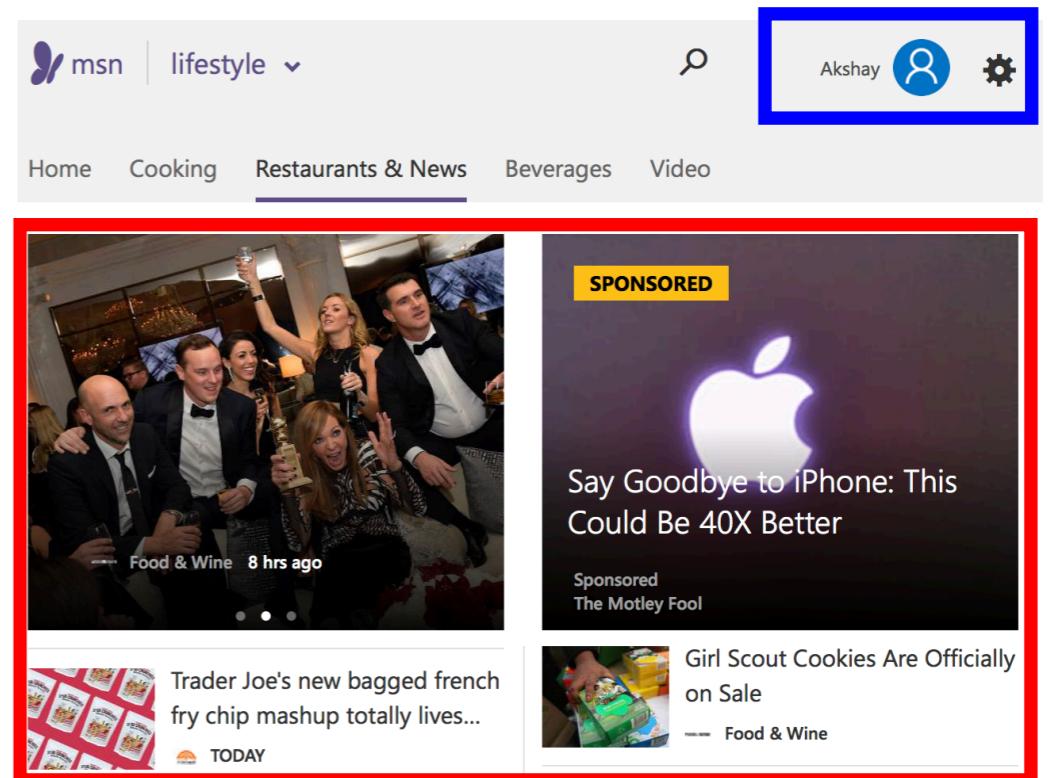
$$\text{Regret}(T, \Pi) = \max_{\pi \in \Pi} \text{Reward}(T, \pi) - \text{LearnerReward}(T)$$

Contextual Bandits

On each of T rounds:

1. Observe context x_t
2. Play action a_t
3. Observe reward $r_t(a_t, x_t)$

K = number of actions



$$r_t = \# \text{ clicks}$$

$$\text{Regret}(T, \Pi) = \max_{\pi \in \Pi} \text{Reward}(T, \pi) - \text{LearnerReward}(T)$$

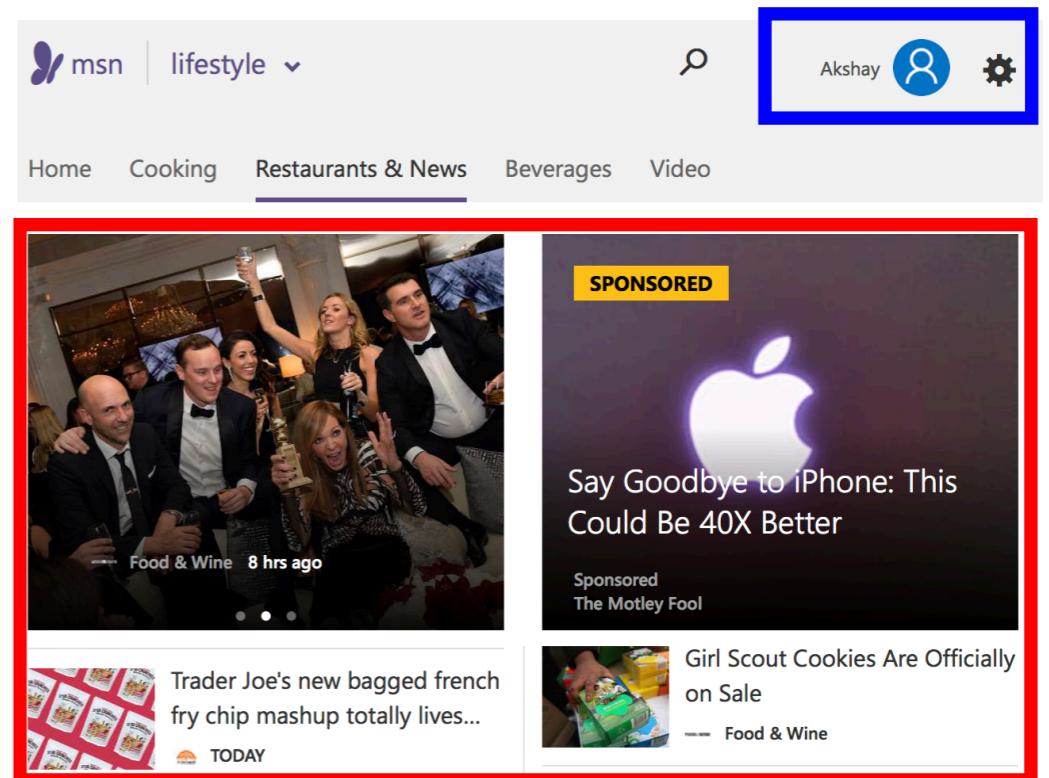
Fact: Can get $\sqrt{KT \log |\Pi|}$ regret.

Contextual Bandits

On each of T rounds:

1. Observe context x_t
2. Play action a_t
3. Observe reward $r_t(a_t, x_t)$

K = number of actions



$$r_t = \# \text{ clicks}$$

$$\text{Regret}(T, \Pi) = \max_{\pi \in \Pi} \text{Reward}(T, \pi) - \text{LearnerReward}(T)$$

Fact: Can get $\sqrt{KT \log |\Pi|}$ regret. Offline Eval gets $(K \log |\Pi|)^{1/3} T^{2/3}$

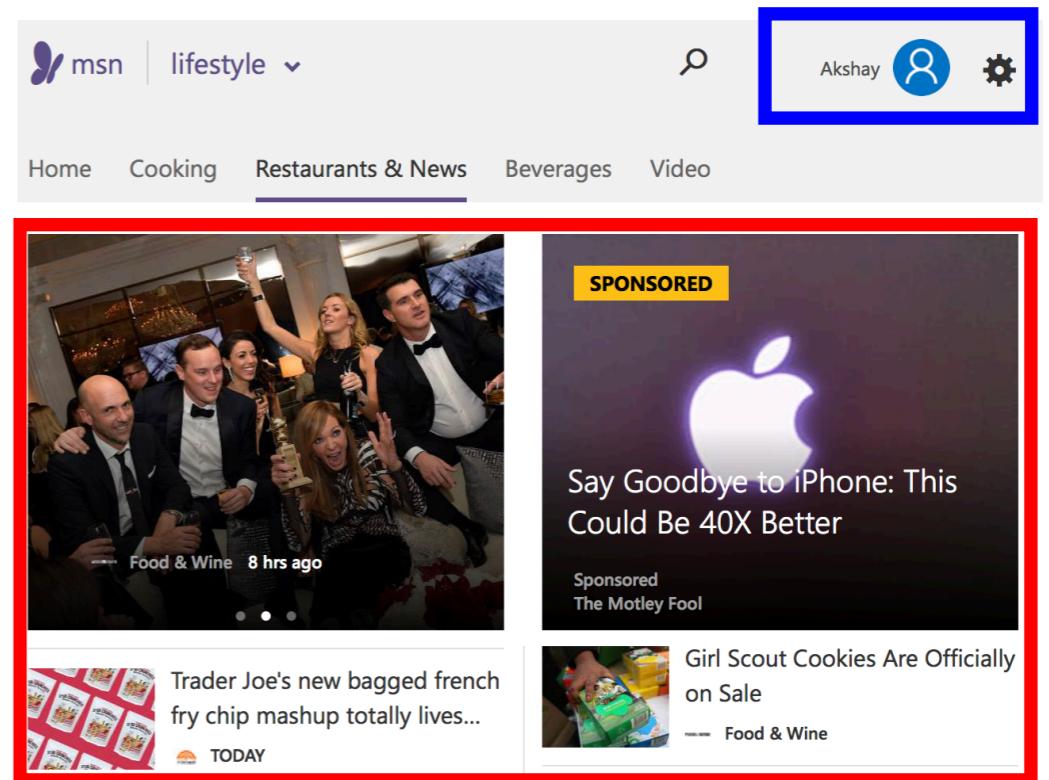
A/B testing gets $(|\Pi|)^{1/3} T^{2/3}$

Contextual Bandits

On each of T rounds:

1. Observe context x_t
2. Play action a_t
3. Observe reward $r_t(a_t, x_t)$

K = number of actions



$$r_t = \# \text{ clicks}$$

$$\text{Regret}(T, \Pi) = \max_{\pi \in \Pi} \text{Reward}(T, \pi) - \text{LearnerReward}(T)$$

A/B testing gets $(|\Pi|)^{1/3} T^{2/3}$
Fact: Can get $\sqrt{KT \log |\Pi|}$ regret. Offline Eval gets $(K \log |\Pi|)^{1/3} T^{2/3}$

Exponential with combinatorial action space!

Contextual Semibandits

The screenshot shows a web browser displaying the msn lifestyle website. The top navigation bar includes the msn logo, a search icon, and a user profile for "Akshay". Below the bar, there are five menu items: Home, Cooking, Restaurants & News (which is underlined), Beverages, and Video. The main content area features several news cards. The first card, titled "Food & Wine 8 hrs ago", shows a group of people in formal attire at a social gathering. The second card, titled "Sponsored Say Goodbye to iPhone: This Could Be 40X Better Sponsored The Motley Fool", features a large Apple logo and text about a new product. The third card, titled "Trader Joe's new bagged french fry chip mashup totally lives... TODAY", shows a bag of Trader Joe's chips. The fourth card, titled "Girl Scout Cookies Are Officially on Sale Food & Wine", shows a box of Girl Scout cookies.

msn | lifestyle

Home Cooking Restaurants & News Beverages Video

SPONSORED

Say Goodbye to iPhone: This Could Be 40X Better

Sponsored
The Motley Fool

Food & Wine 8 hrs ago

Trader Joe's new bagged french fry chip mashup totally lives... TODAY

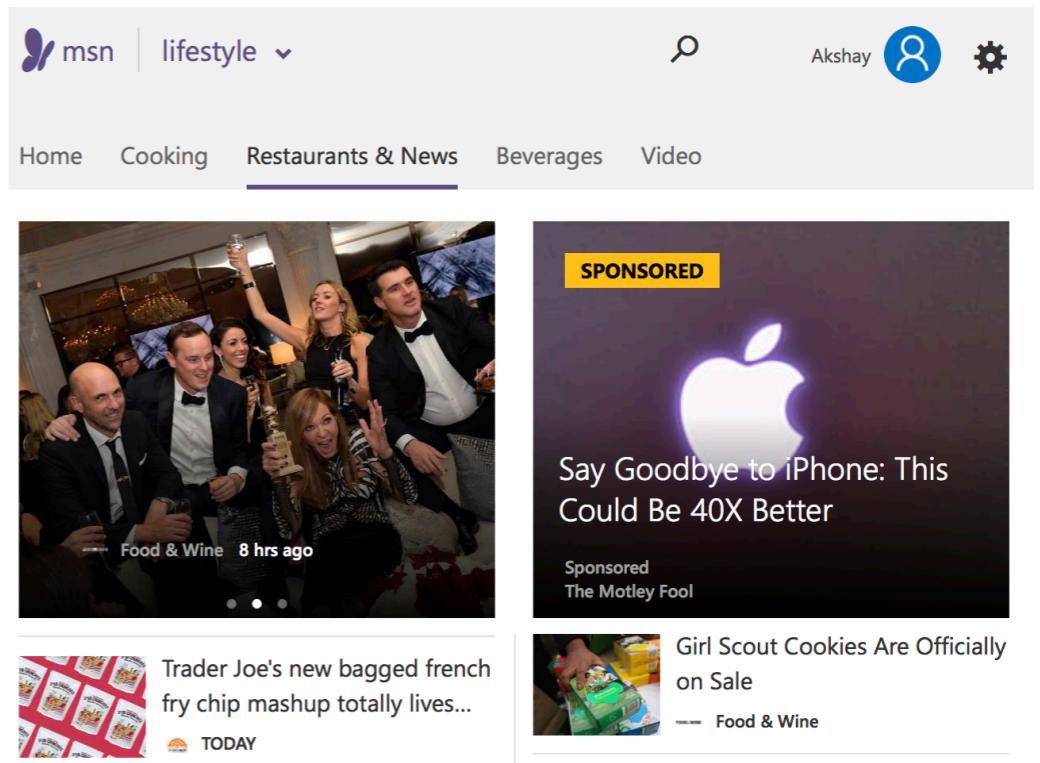
Girl Scout Cookies Are Officially on Sale

Food & Wine

Contextual Semibandits

On each of T rounds:

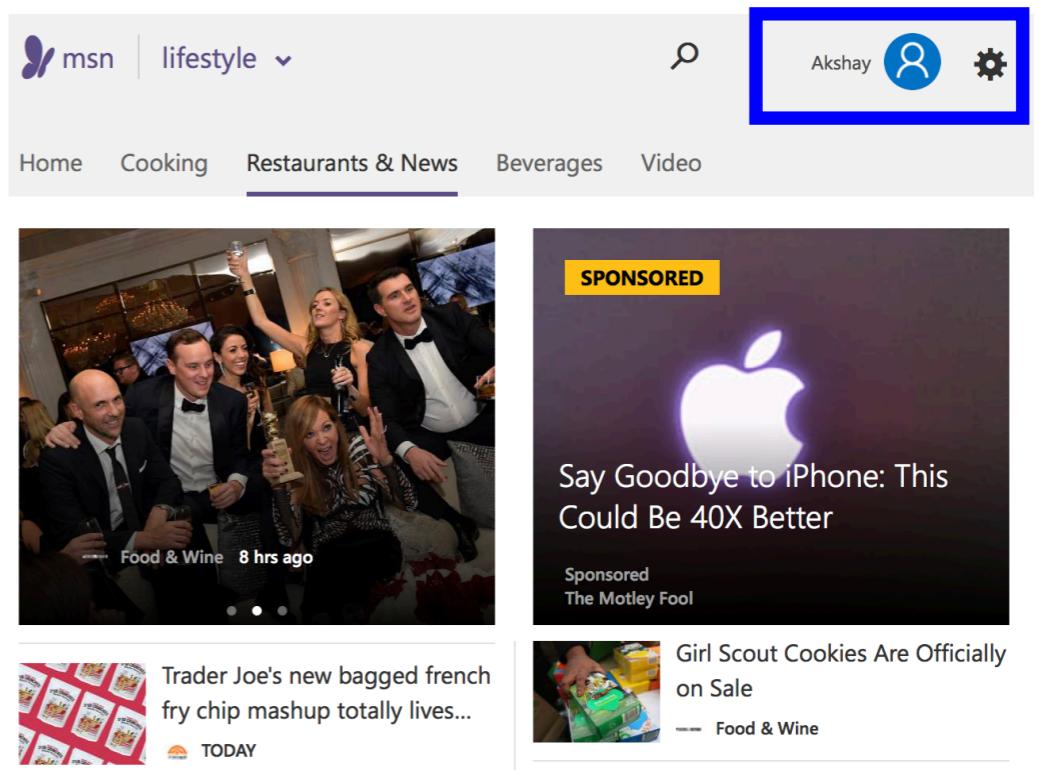
1. Observe context
2. Play action
3. Observe features
4. Observe reward



Contextual Semibandits

On each of T rounds:

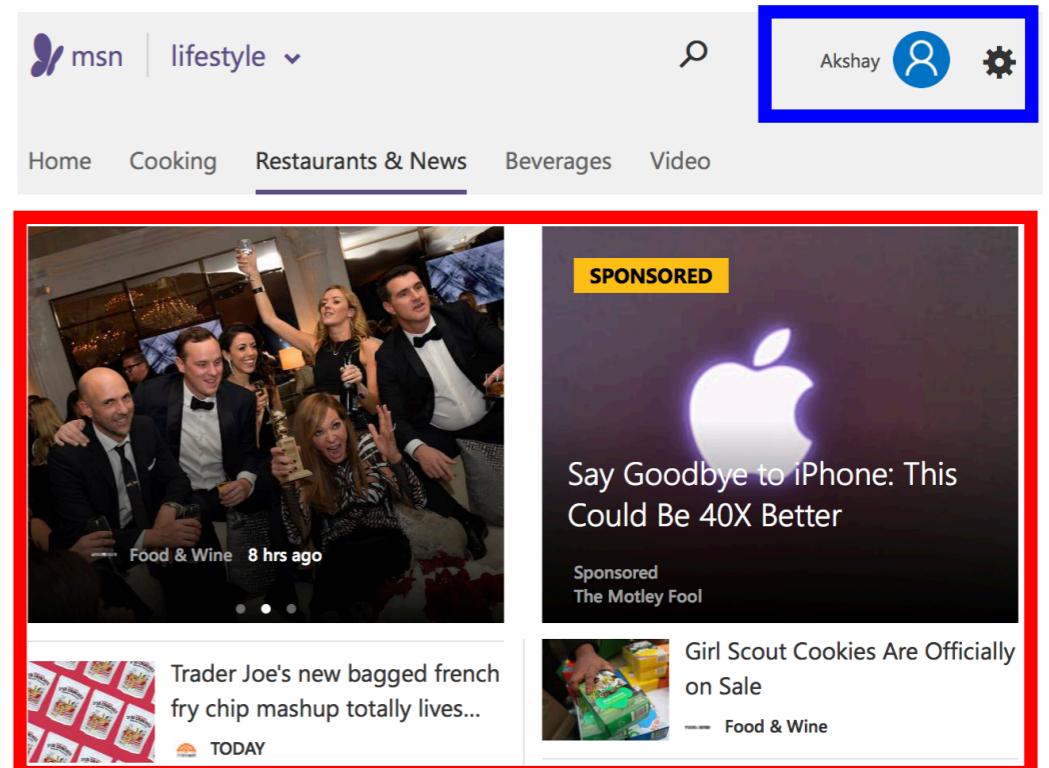
1. Observe context x_t
2. Play action
3. Observe features
4. Observe reward



Contextual Semibandits

On each of T rounds:

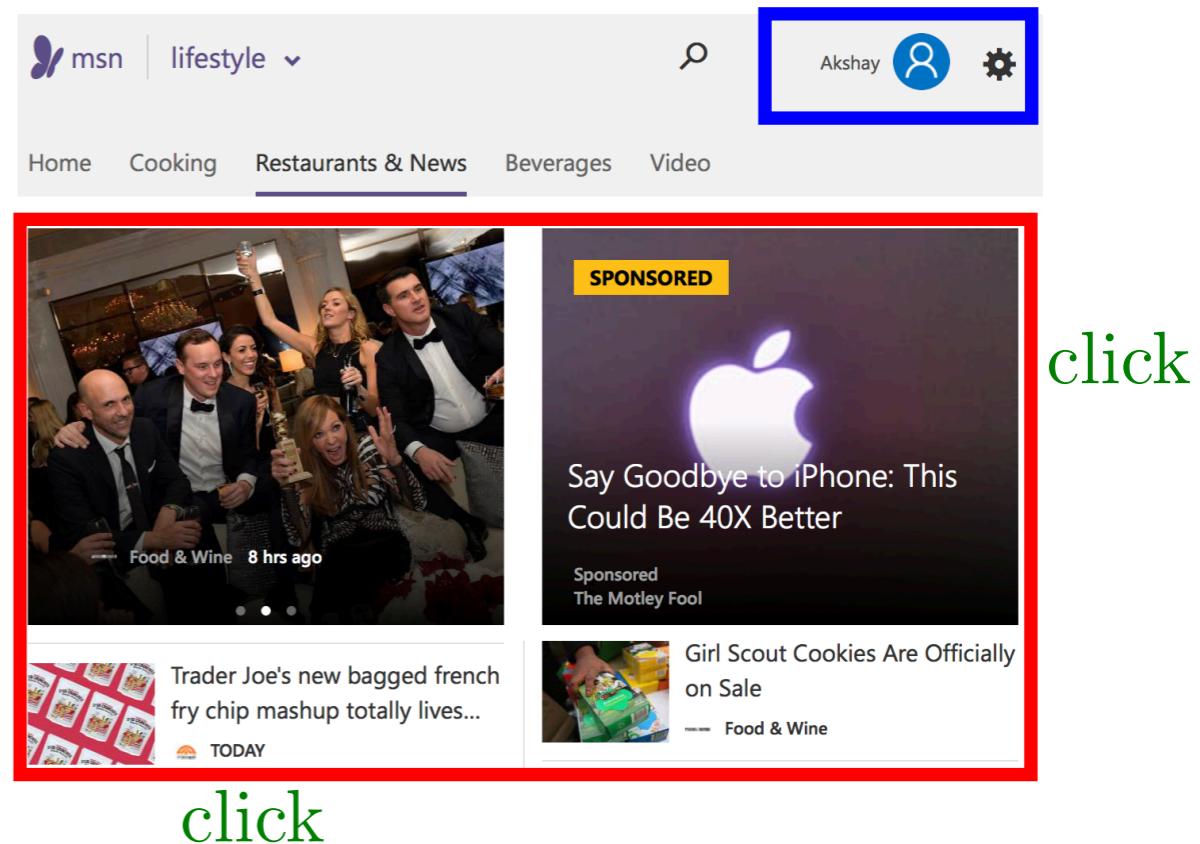
1. Observe context x_t
2. Play action $A_t = (a_1, \dots, a_L)$
3. Observe features
4. Observe reward



Contextual Semibandits

On each of T rounds:

1. Observe context x_t
2. Play action $A_t = (a_1, \dots, a_L)$
3. Observe features $\{y(a_\ell)\}_{\ell=1}^L$
4. Observe reward

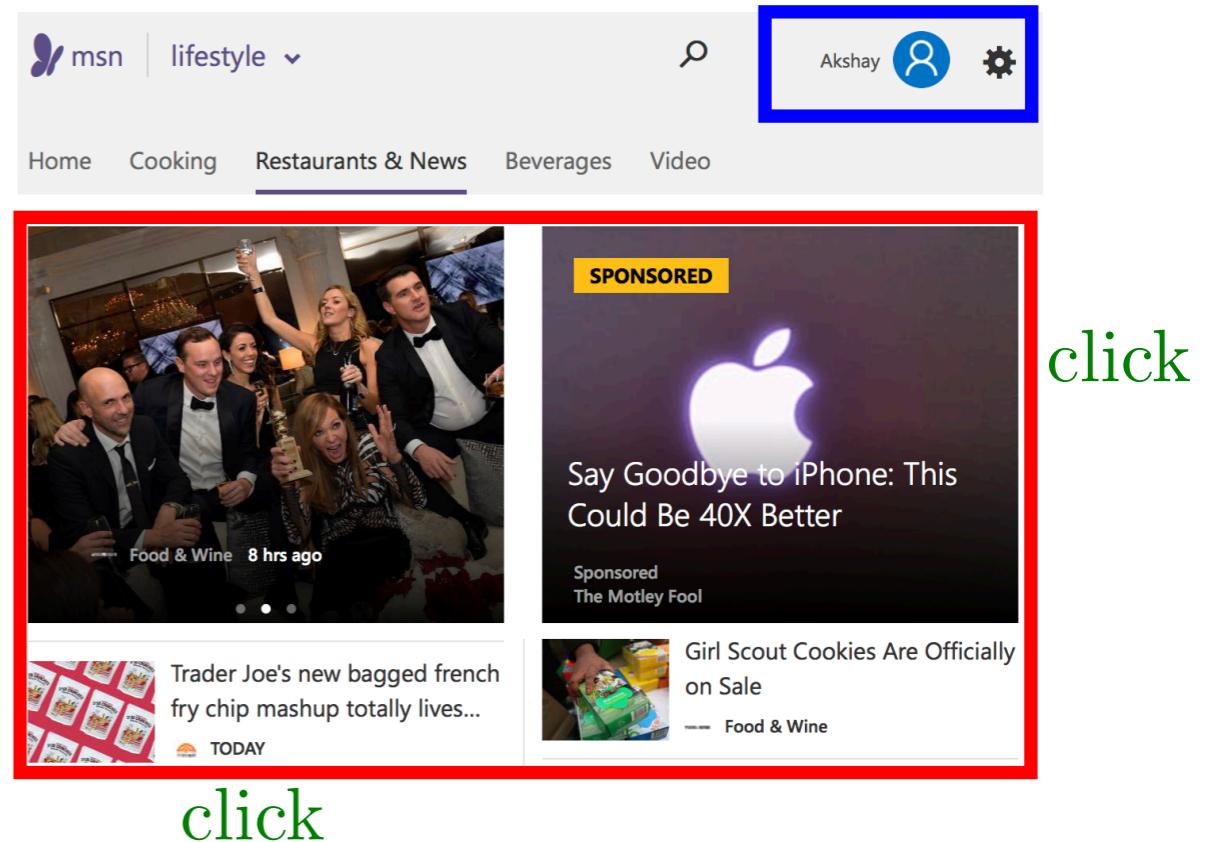


Contextual Semibandits

On each of T rounds:

1. Observe context x_t
2. Play action $A_t = (a_1, \dots, a_L)$
3. Observe features $\{y(a_\ell)\}_{\ell=1}^L$
4. Observe reward

$$r_t(A_t, x_t) = \sum_{\ell} y(a_\ell) + \text{noise}$$



Contextual Semibandits

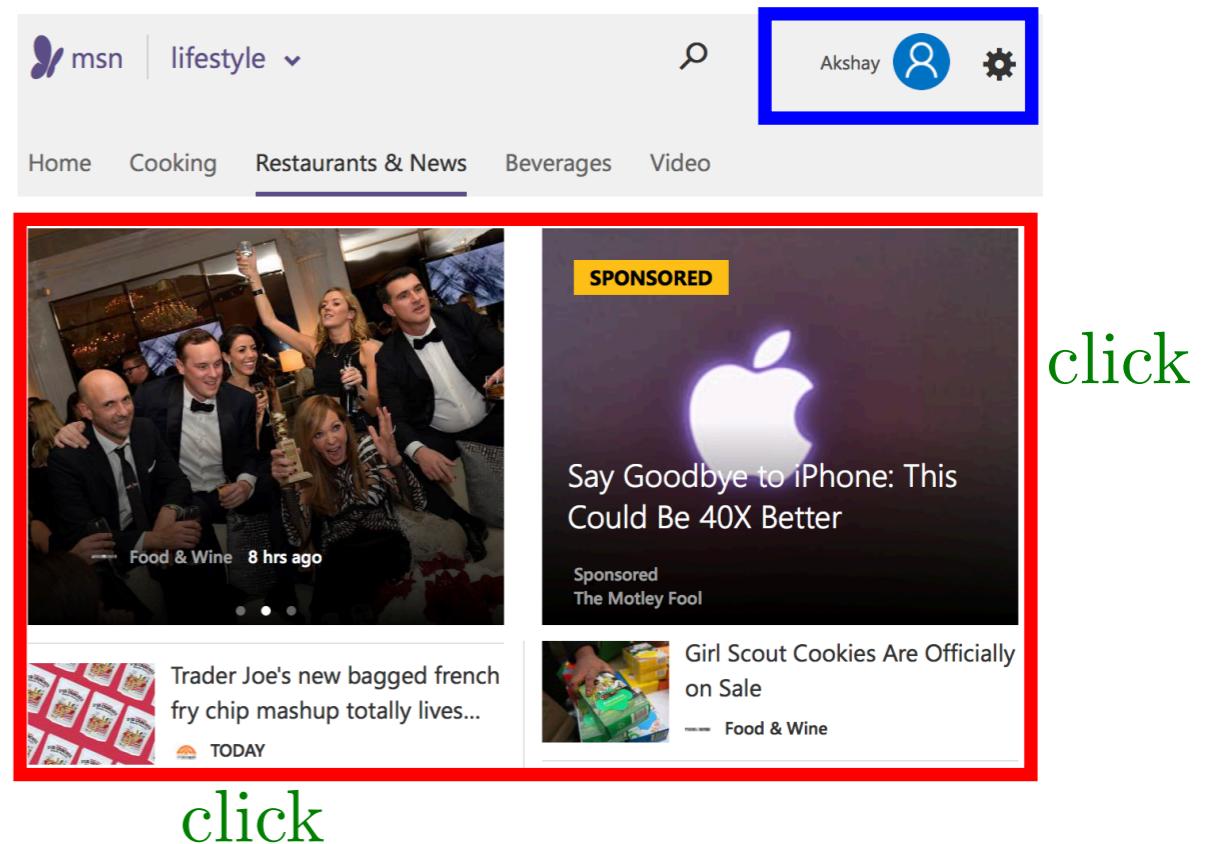
On each of T rounds:

1. Observe context x_t
2. Play action $A_t = (a_1, \dots, a_L)$
3. Observe features $\{y(a_\ell)\}_{\ell=1}^L$
4. Observe reward

$$r_t(A_t, x_t) = \sum_{\ell} y(a_\ell) + \text{noise}$$

B = number of simple actions

L = composite action length



Contextual Semibandits

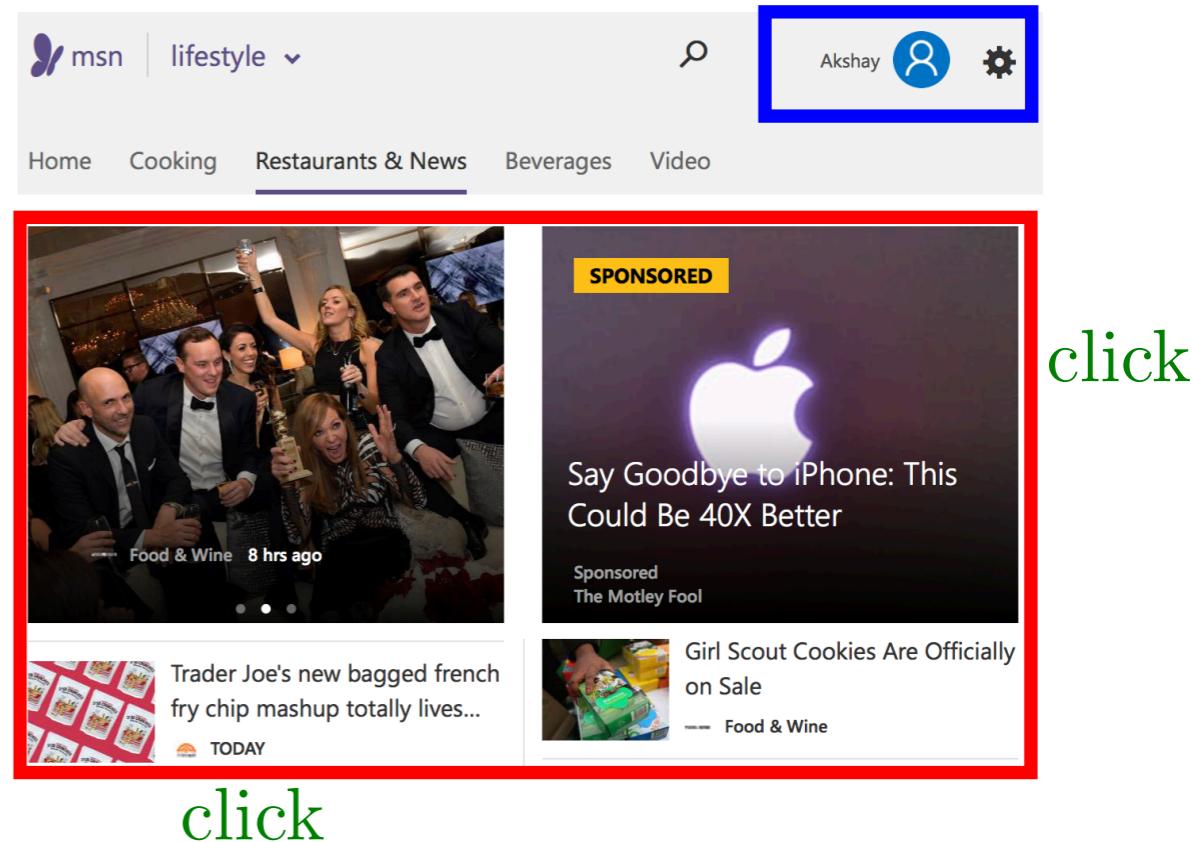
On each of T rounds:

1. Observe context x_t
2. Play action $A_t = (a_1, \dots, a_L)$
3. Observe features $\{y(a_\ell)\}_{\ell=1}^L$
4. Observe reward

$$r_t(A_t, x_t) = \sum_{\ell} y(a_\ell) + \text{noise}$$

B = number of simple actions

L = composite action length



click

Question: Improve performance by leveraging reward structure + additional feedback?

Contextual Semibandits

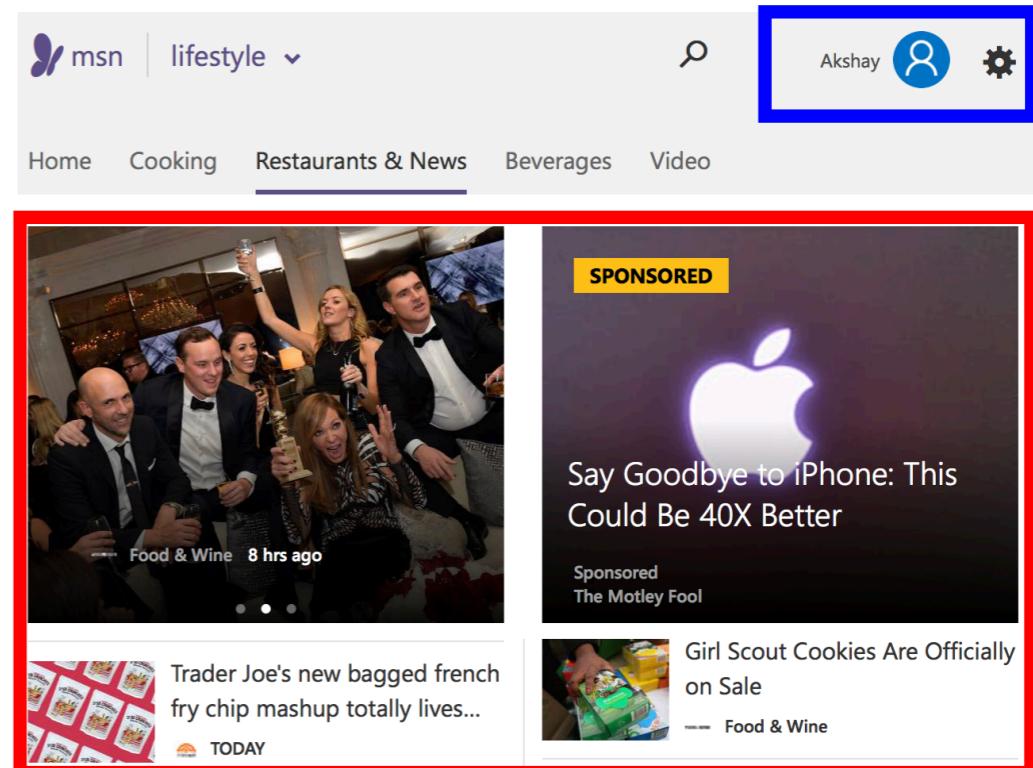
On each of T rounds:

1. Observe context x_t
2. Play action $A_t = (a_1, \dots, a_L)$
3. Observe features $\{y(a_\ell)\}_{\ell=1}^L$
4. Observe reward

$$r_t(A_t, x_t) = \sum_{\ell} y(a_\ell) + \text{noise}$$

B = number of simple actions

L = composite action length



click

Question: Improve performance by leveraging reward structure + additional feedback?

Challenges:

Contextual Semibandits

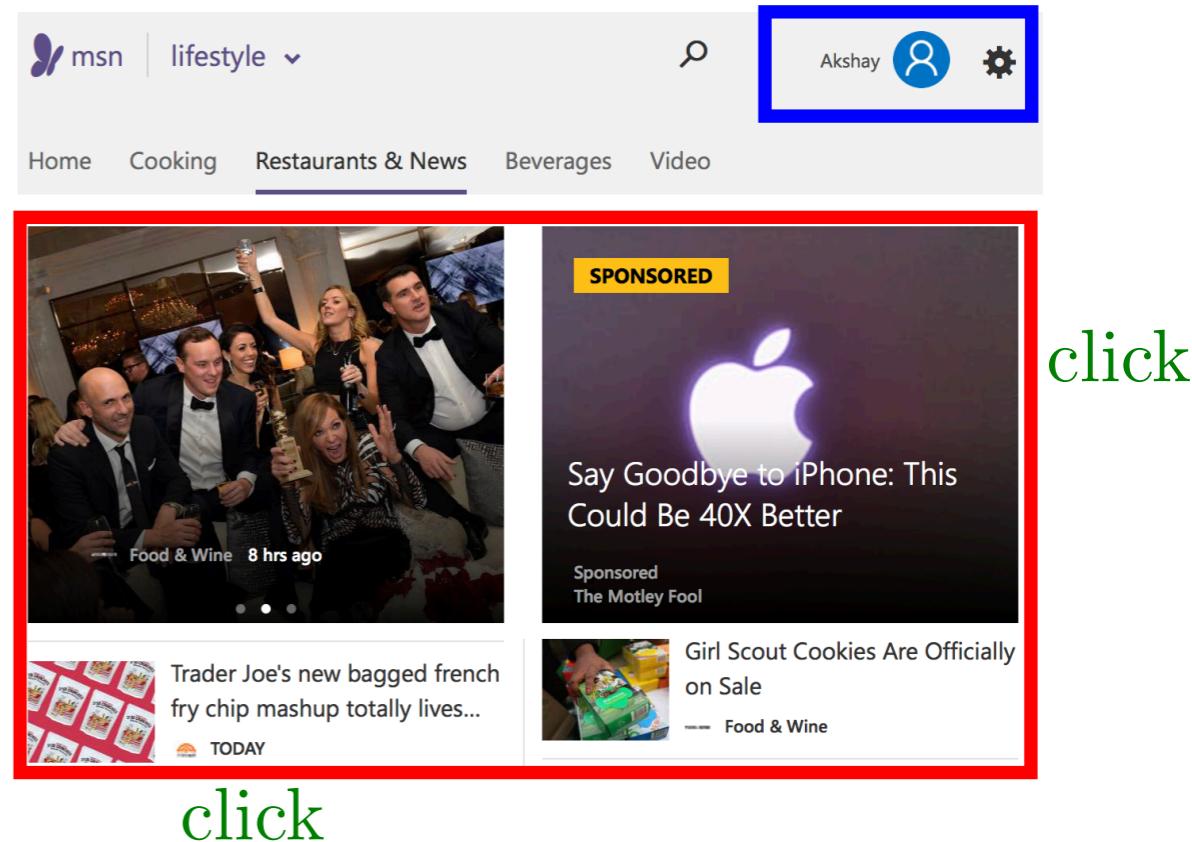
On each of T rounds:

1. Observe context x_t
2. Play action $A_t = (a_1, \dots, a_L)$
3. Observe features $\{y(a_\ell)\}_{\ell=1}^L$
4. Observe reward

$$r_t(A_t, x_t) = \sum_{\ell} y(a_\ell) + \text{noise}$$

B = number of simple actions

L = composite action length



click

Question: Improve performance by leveraging reward structure + additional feedback?

Challenges:

- Off-policy evaluation?

Contextual Semibandits

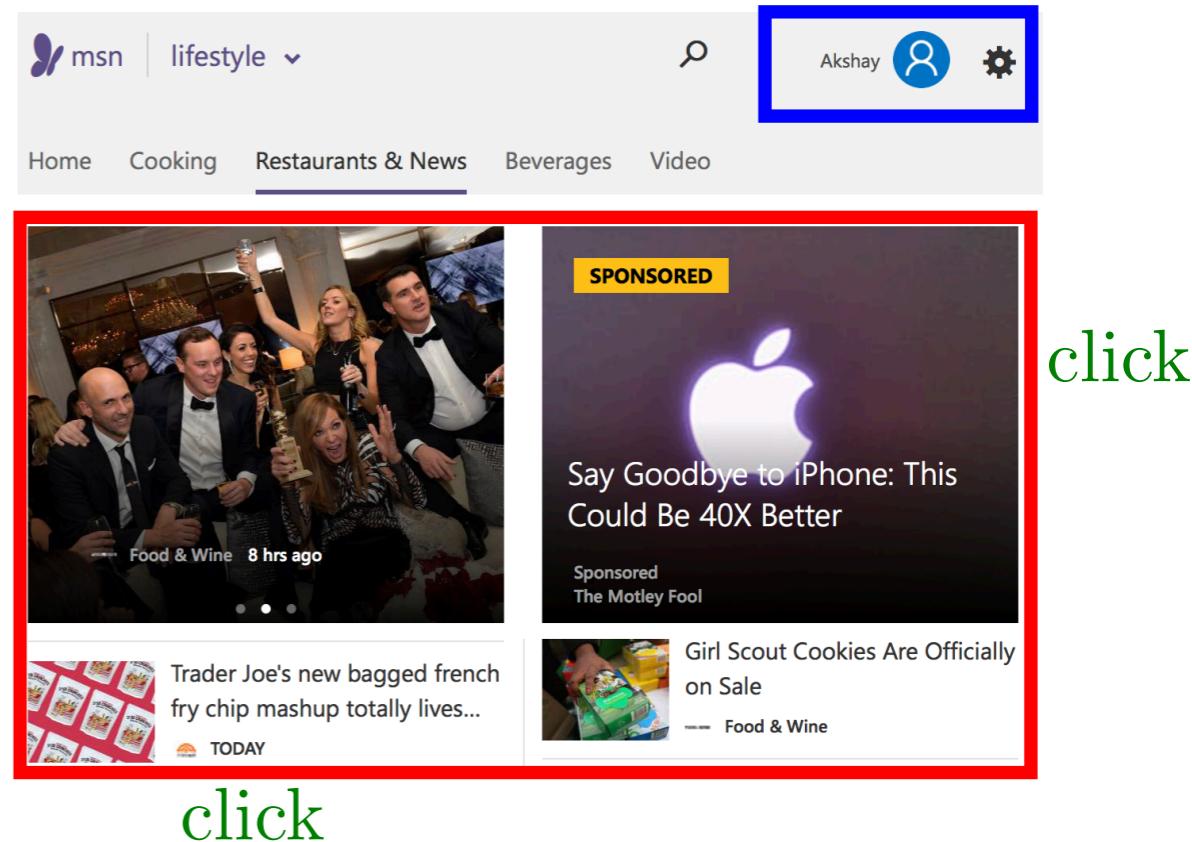
On each of T rounds:

1. Observe context x_t
2. Play action $A_t = (a_1, \dots, a_L)$
3. Observe features $\{y(a_\ell)\}_{\ell=1}^L$
4. Observe reward

$$r_t(A_t, x_t) = \sum_{\ell} y(a_\ell) + \text{noise}$$

B = number of simple actions

L = composite action length



click

Question: Improve performance by leveraging reward structure + additional feedback?

Challenges:

- Off-policy evaluation?
- Explore vs Exploit?

Contextual Semibandits

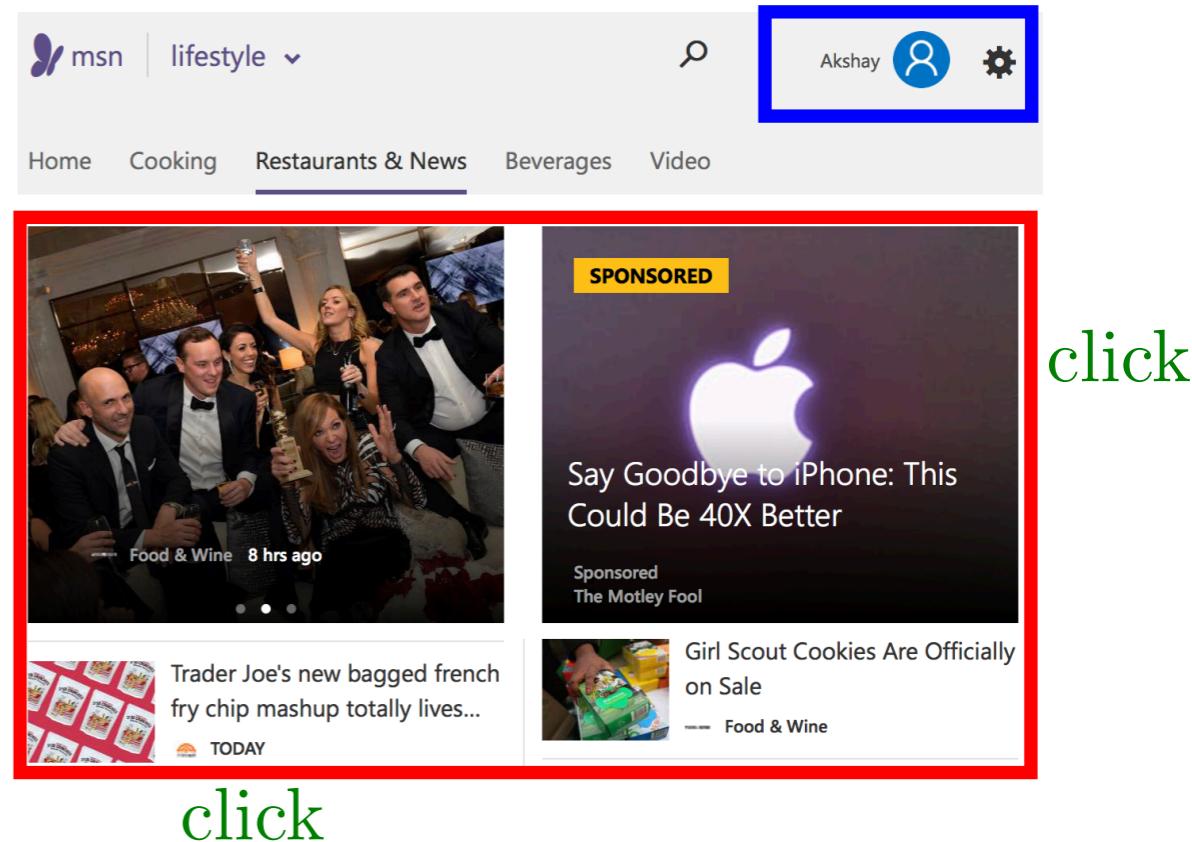
On each of T rounds:

1. Observe context x_t
2. Play action $A_t = (a_1, \dots, a_L)$
3. Observe features $\{y(a_\ell)\}_{\ell=1}^L$
4. Observe reward

$$r_t(A_t, x_t) = \sum_{\ell} y(a_\ell) + \text{noise}$$

B = number of simple actions

L = composite action length



Question: Improve performance by leveraging reward structure + additional feedback?

Challenges:

- Off-policy evaluation?
- Explore vs Exploit?
- Computational Efficiency?

Results

[Krishnamurthy, Agarwal, Dudik. NeurIPS 2016]

Results

Theorem: Efficient algorithm with $\sqrt{BT \log(|\Pi|)}$ regret

Parameters: T rounds, B simple actions, composite action length L

Results

Theorem: Efficient algorithm with $\sqrt{BT \log(|\Pi|)}$ regret

Parameters: T rounds, B simple actions, composite action length L

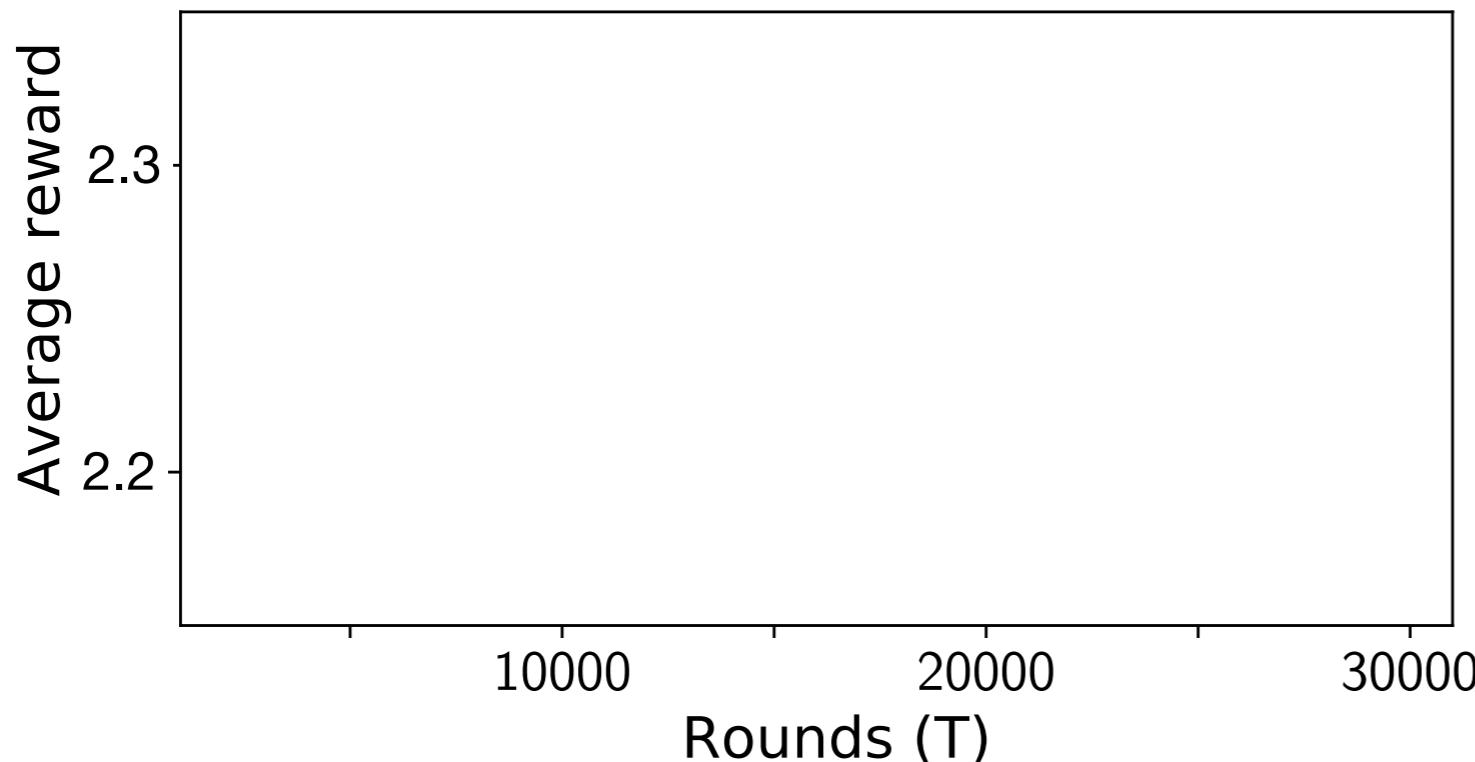
- Exponentially better than $\sqrt{B^L T \log(|\Pi|)}$ for naive contextual bandits
- Computationally efficient with rich policy classes

Results

Theorem: Efficient algorithm with $\sqrt{BT \log(|\Pi|)}$ regret

Parameters: T rounds, B simple actions, composite action length L

- Exponentially better than $\sqrt{B^L T \log(|\Pi|)}$ for naive contextual bandits
- Computationally efficient with rich policy classes

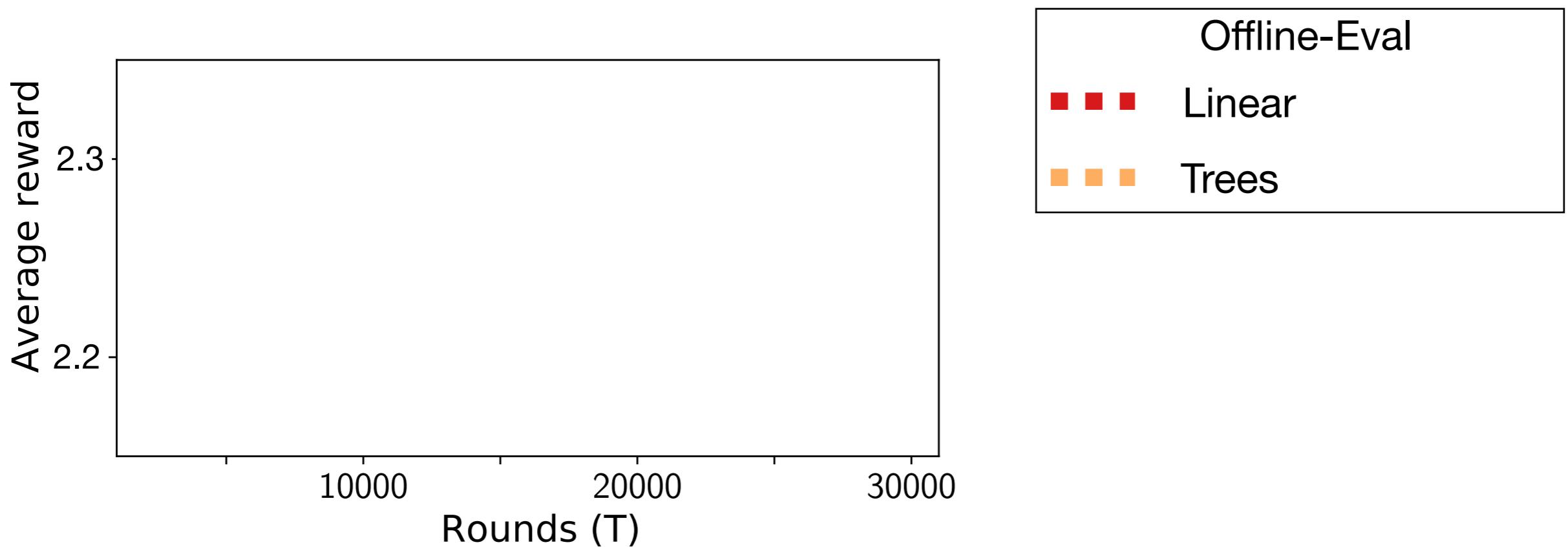


Results

Theorem: Efficient algorithm with $\sqrt{BT \log(|\Pi|)}$ regret

Parameters: T rounds, B simple actions, composite action length L

- Exponentially better than $\sqrt{B^L T \log(|\Pi|)}$ for naive contextual bandits
- Computationally efficient with rich policy classes

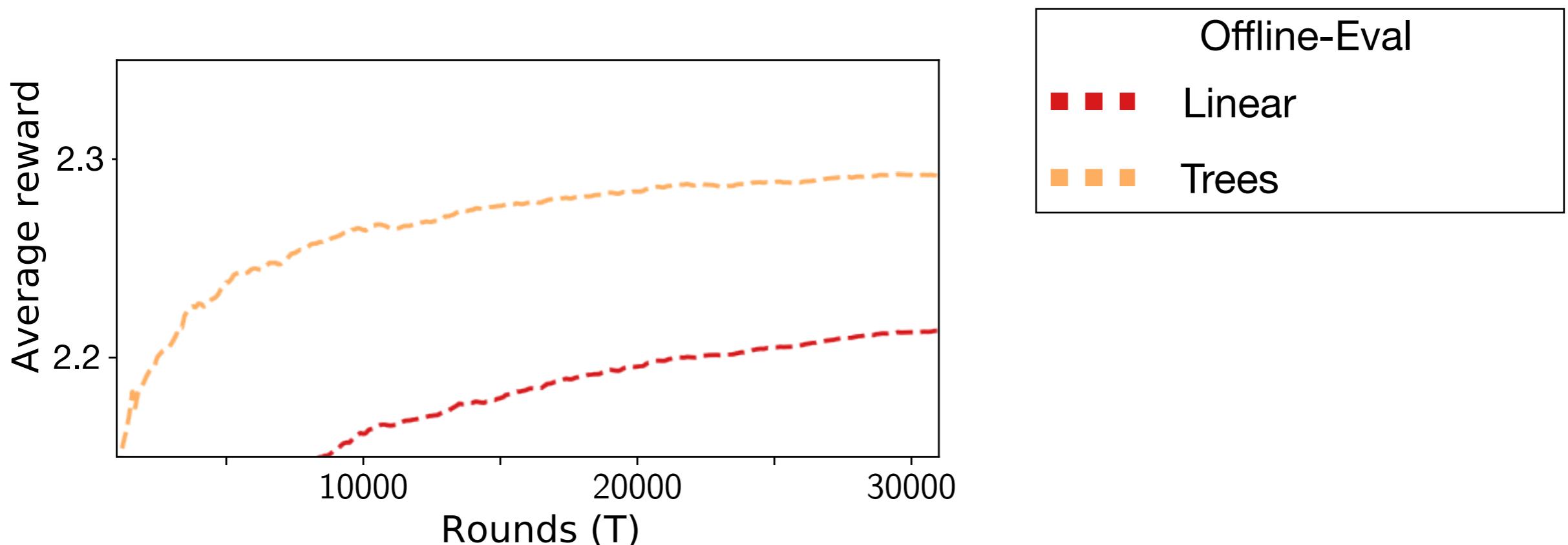


Results

Theorem: Efficient algorithm with $\sqrt{BT \log(|\Pi|)}$ regret

Parameters: T rounds, B simple actions, composite action length L

- Exponentially better than $\sqrt{B^L T \log(|\Pi|)}$ for naive contextual bandits
- Computationally efficient with rich policy classes

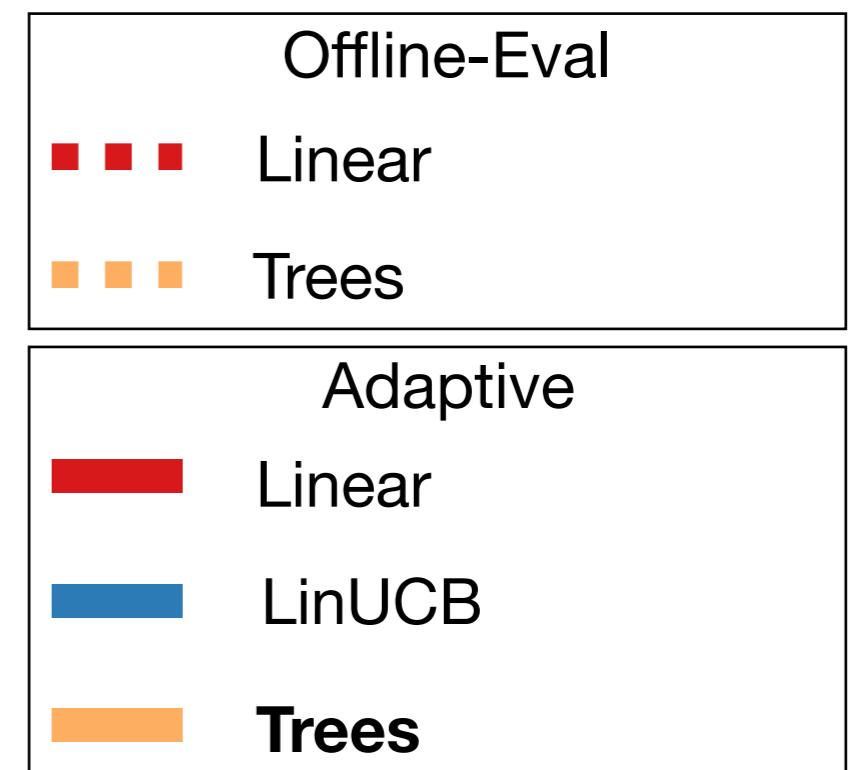
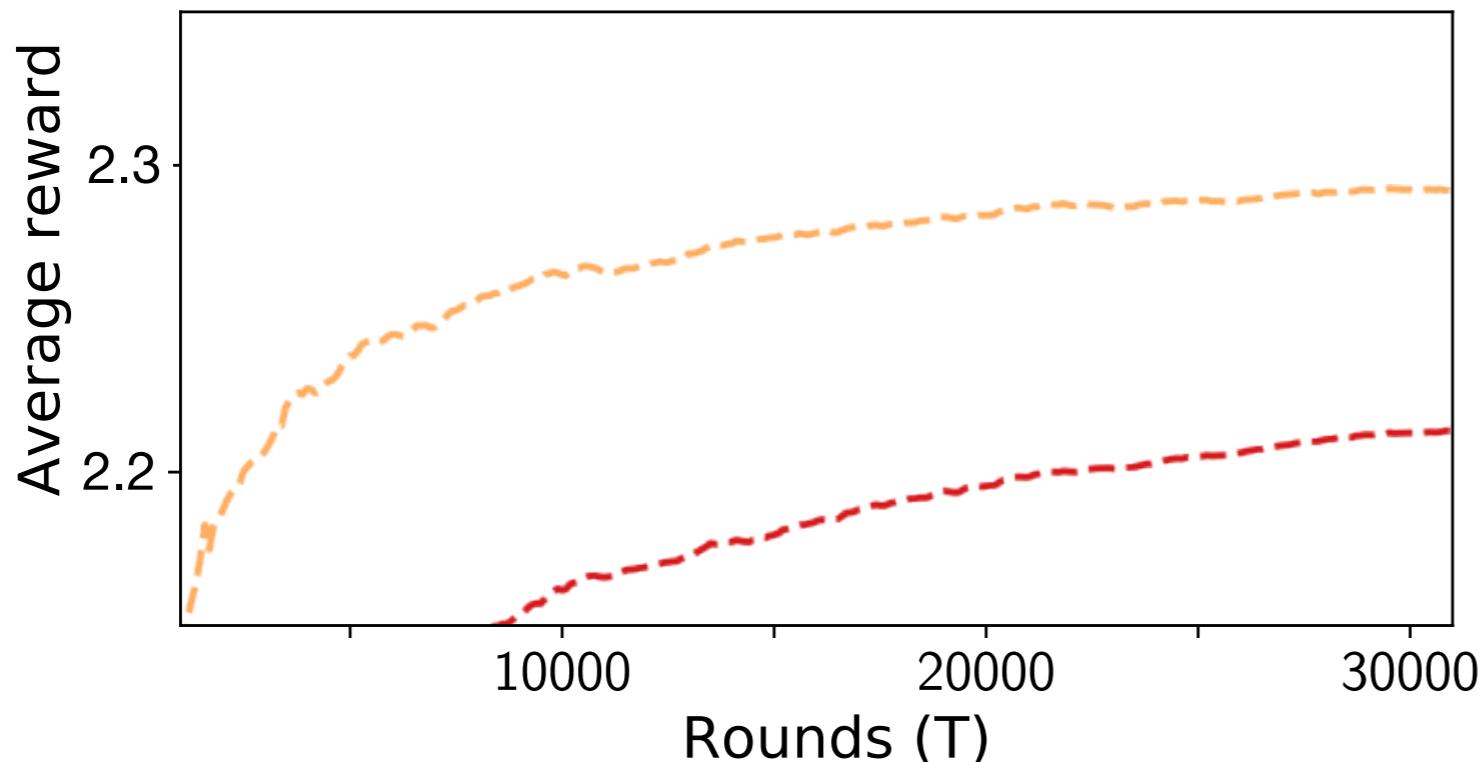


Results

Theorem: Efficient algorithm with $\sqrt{BT \log(|\Pi|)}$ regret

Parameters: T rounds, B simple actions, composite action length L

- Exponentially better than $\sqrt{B^L T \log(|\Pi|)}$ for naive contextual bandits
- Computationally efficient with rich policy classes

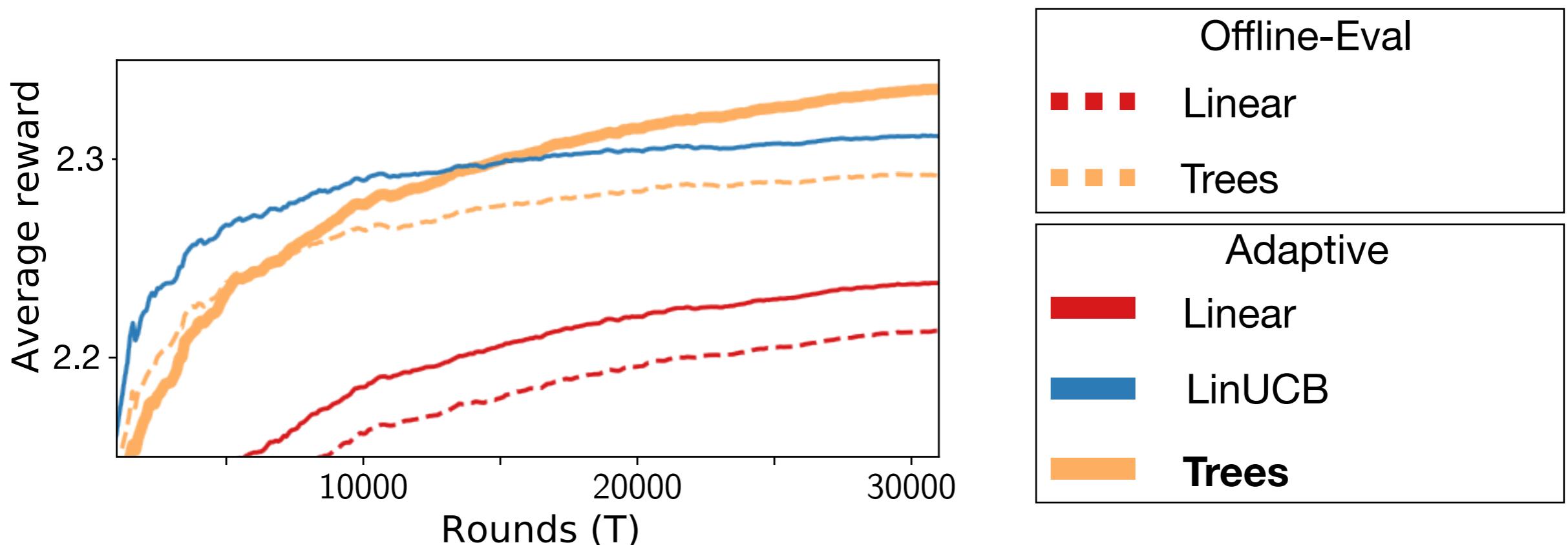


Results

Theorem: Efficient algorithm with $\sqrt{BT \log(|\Pi|)}$ regret

Parameters: T rounds, B simple actions, composite action length L

- Exponentially better than $\sqrt{B^L T \log(|\Pi|)}$ for naive contextual bandits
- Computationally efficient with rich policy classes

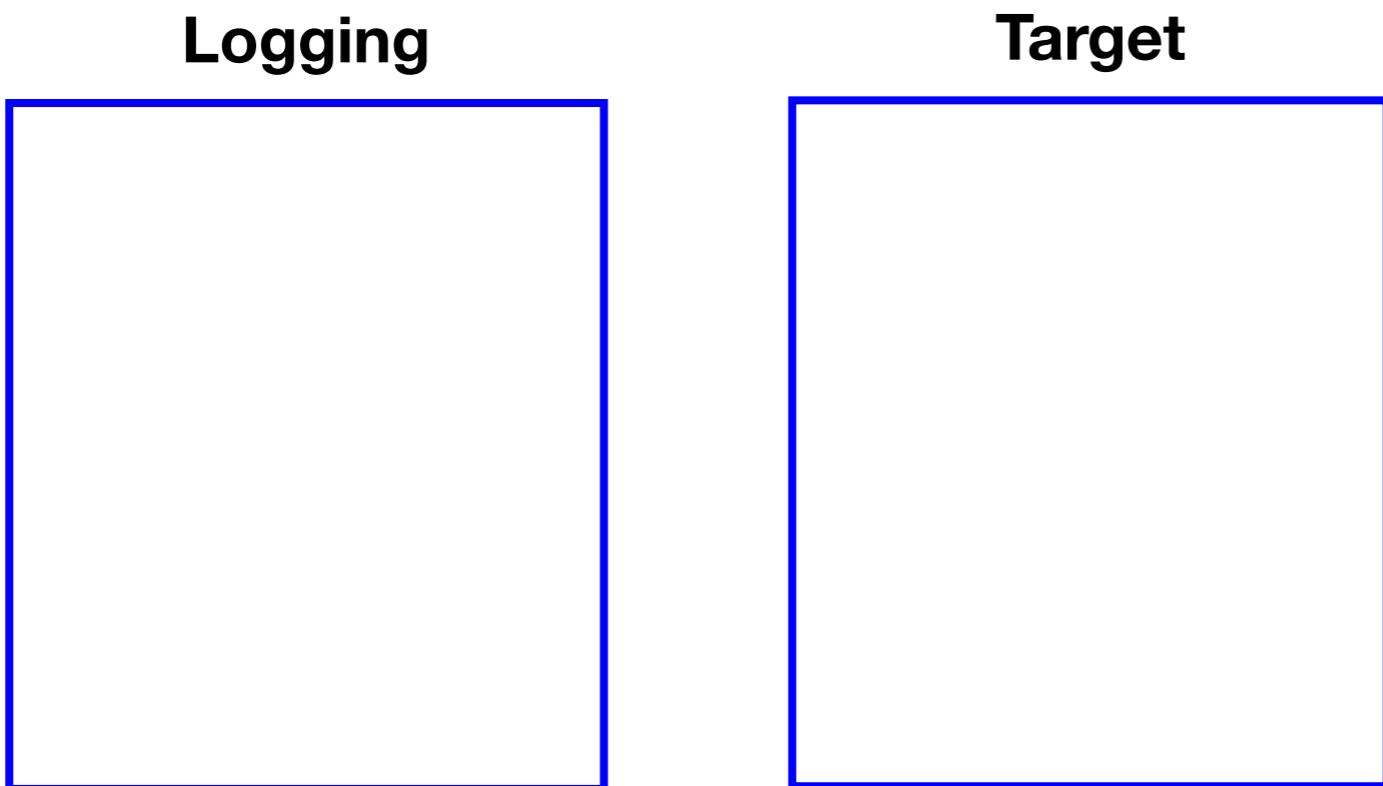


Techniques – Off-policy evaluation

Subproblem: Given data collected by a logging policy, estimate reward of a target policy

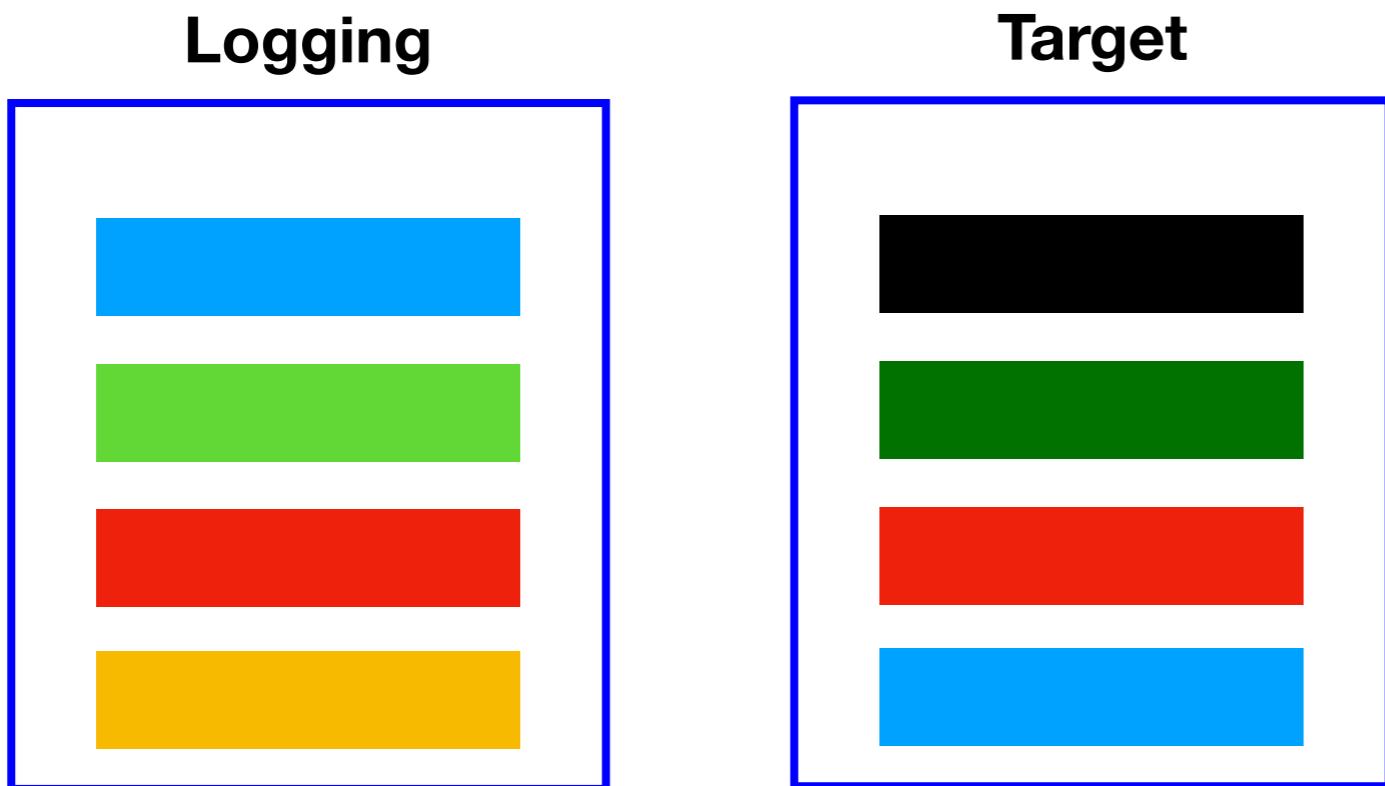
Techniques – Off-policy evaluation

Subproblem: Given data collected by a logging policy, estimate reward of a target policy



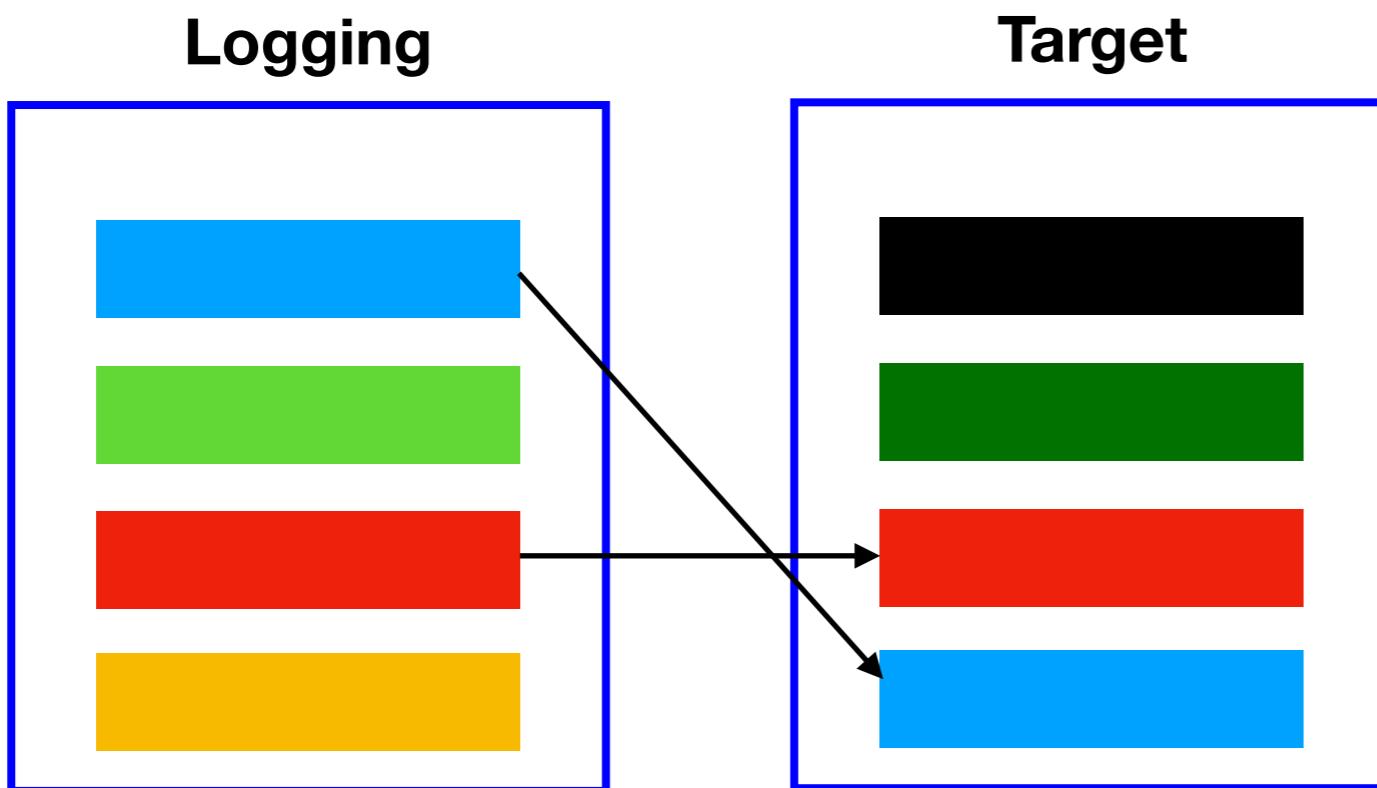
Techniques – Off-policy evaluation

Subproblem: Given data collected by a logging policy, estimate reward of a target policy



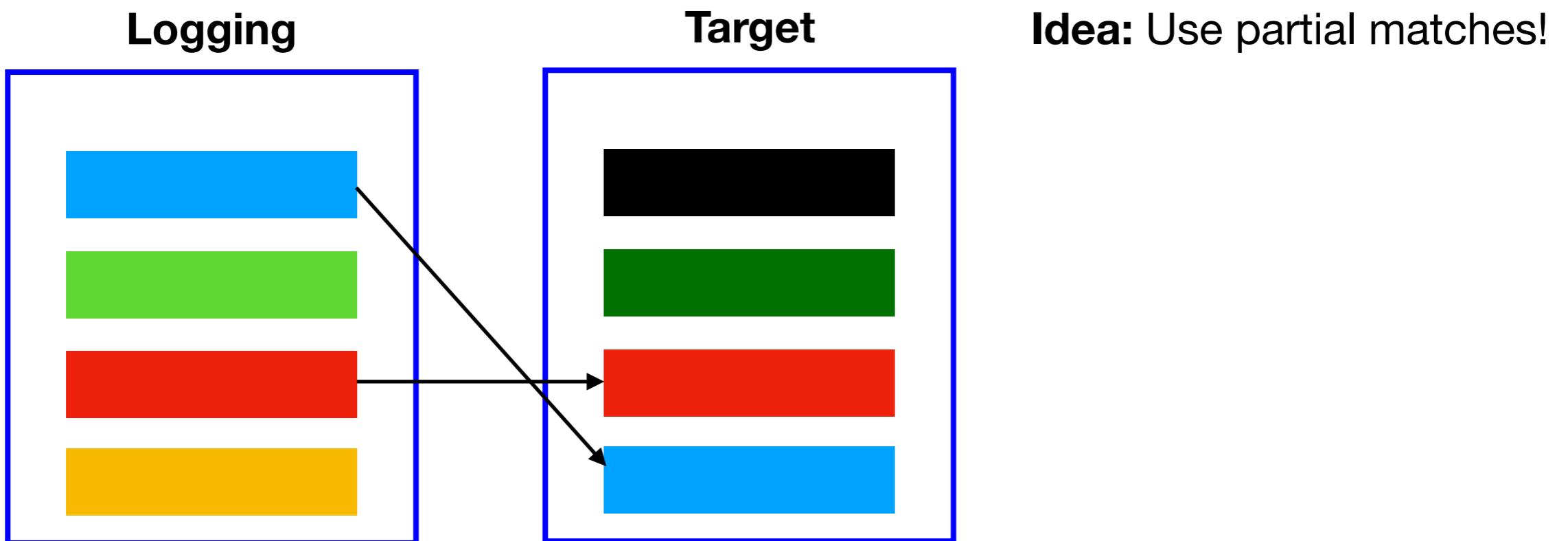
Techniques – Off-policy evaluation

Subproblem: Given data collected by a logging policy, estimate reward of a target policy



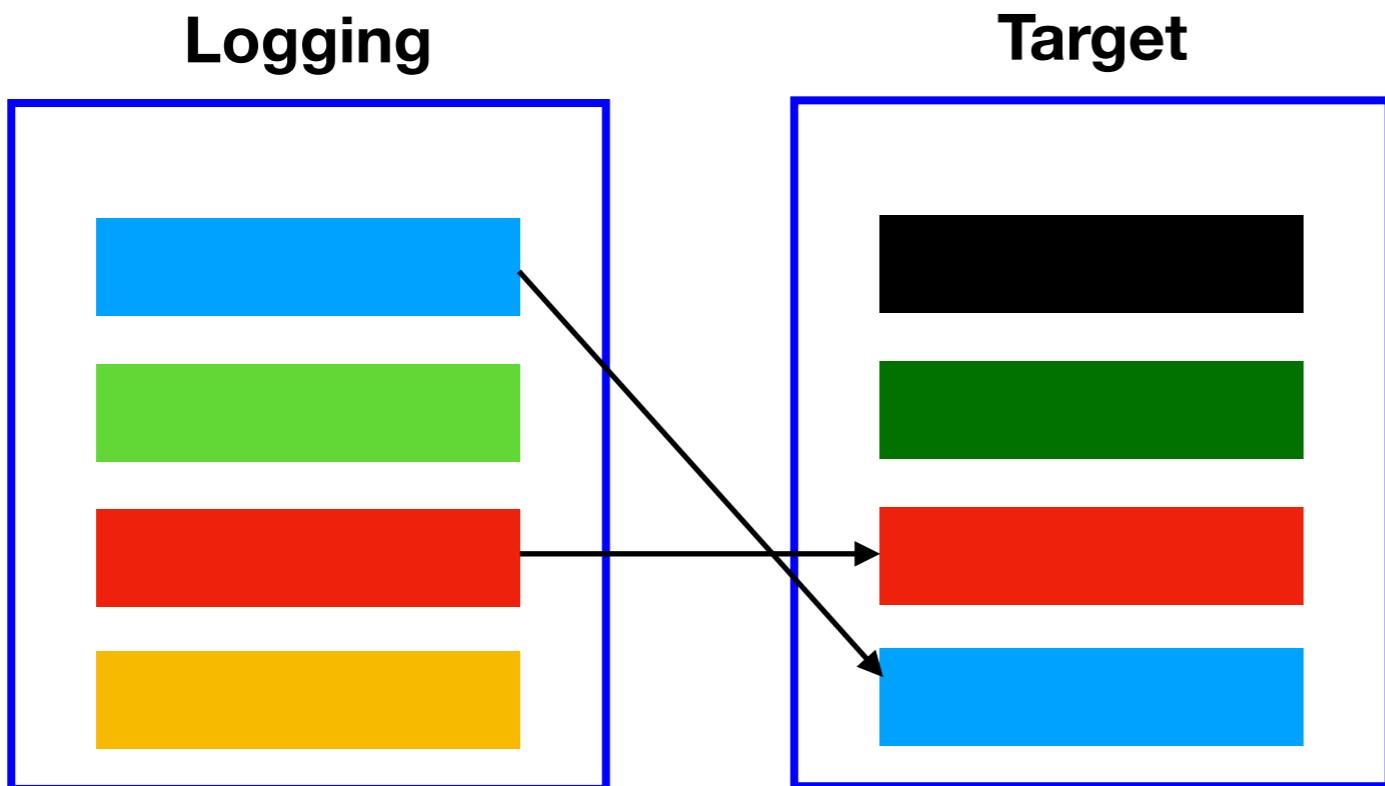
Techniques – Off-policy evaluation

Subproblem: Given data collected by a logging policy, estimate reward of a target policy



Techniques – Off-policy evaluation

Subproblem: Given data collected by a logging policy, estimate reward of a target policy



Idea: Use partial matches!

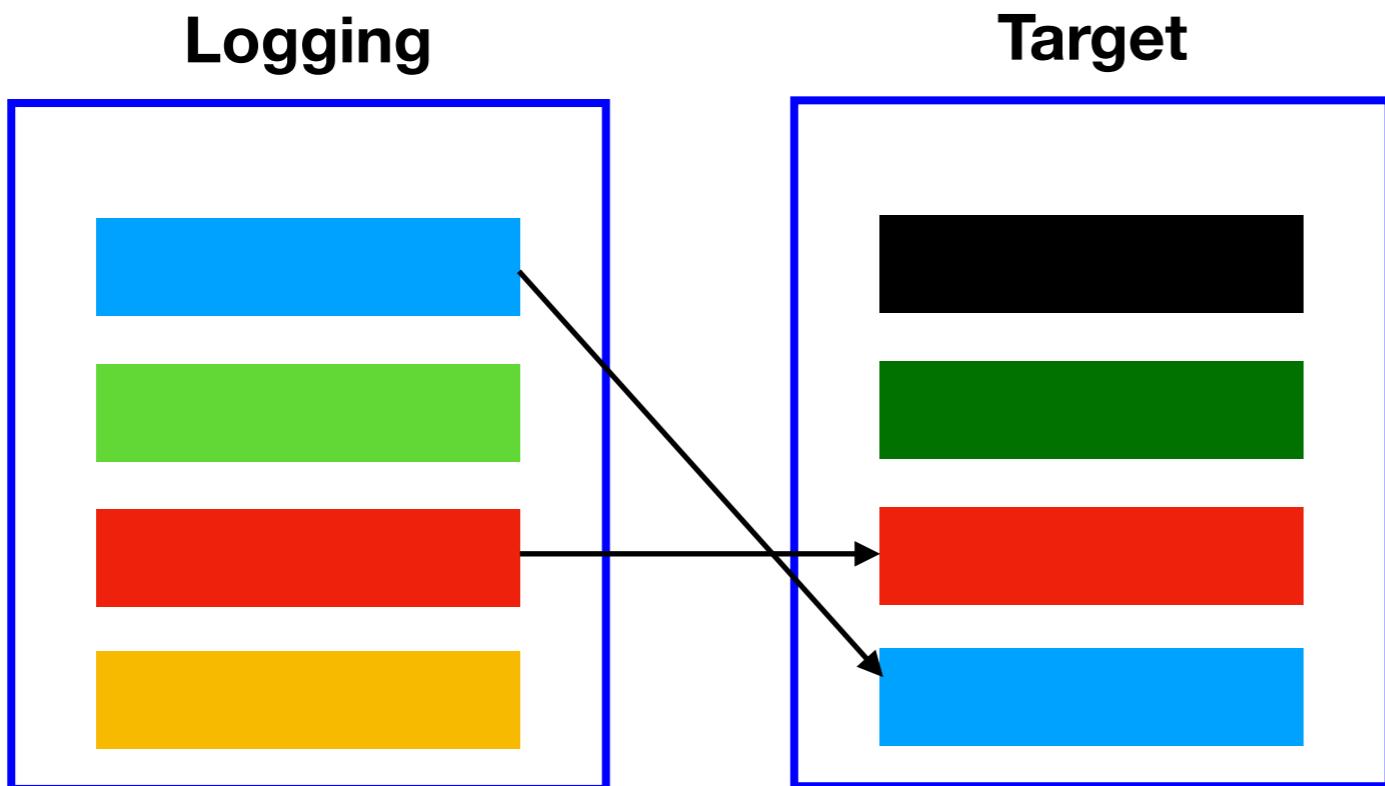
If $A \sim Q(\cdot|x)$

$$\hat{y}(a) = \frac{y(a)\mathbf{1}(a \in A)}{Q(a \in A|x)}$$

$$\hat{r}(\pi, x) = \sum_{a \in \pi(x)} \hat{y}(a)$$

Techniques – Off-policy evaluation

Subproblem: Given data collected by a logging policy, estimate reward of a target policy



Idea: Use partial matches!

If $A \sim Q(\cdot|x)$

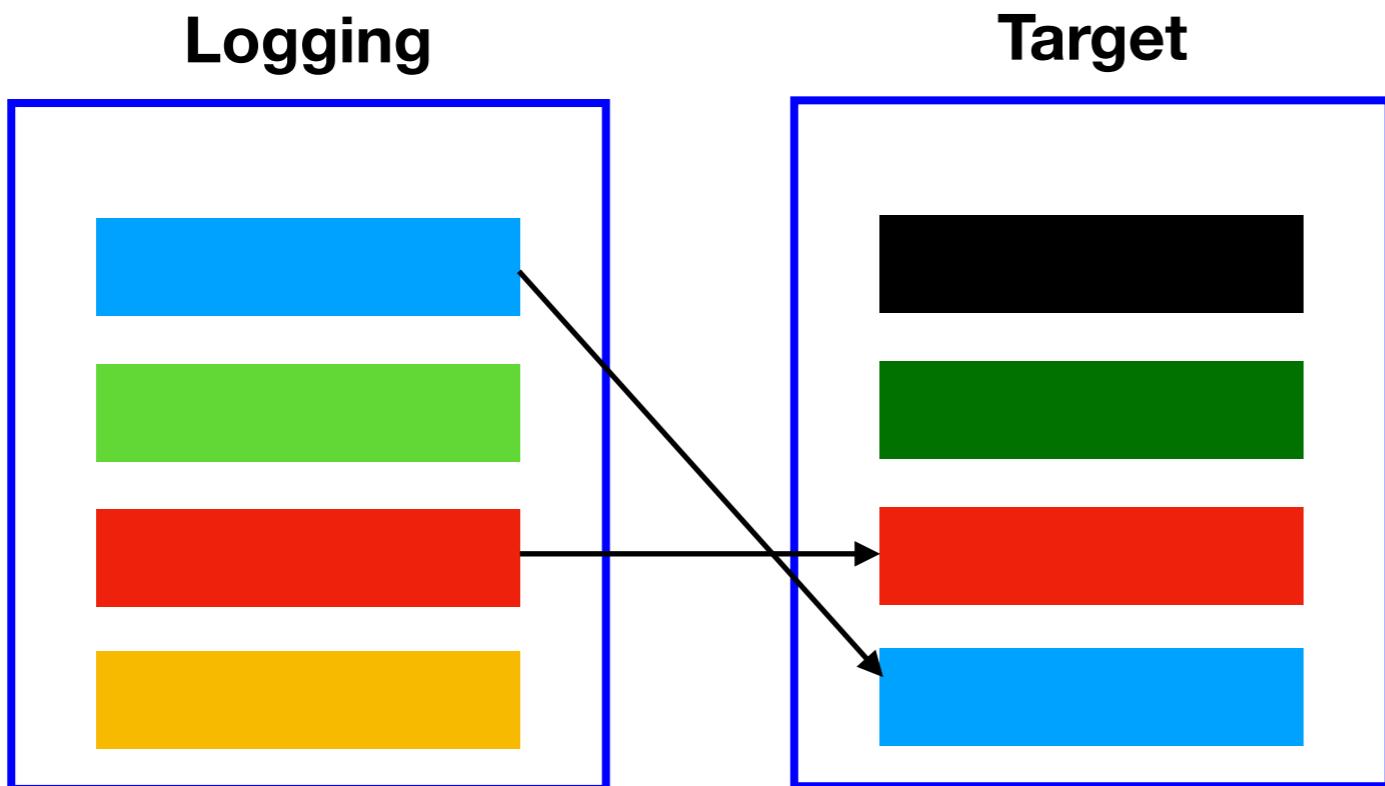
$$\hat{y}(a) = \frac{y(a)\mathbf{1}(a \in A)}{Q(a \in A|x)}$$

$$\hat{r}(\pi, x) = \sum_{a \in \pi(x)} \hat{y}(a)$$

- Uniform Q gives O(B) variance

Techniques – Off-policy evaluation

Subproblem: Given data collected by a logging policy, estimate reward of a target policy



Idea: Use partial matches!

If $A \sim Q(\cdot|x)$

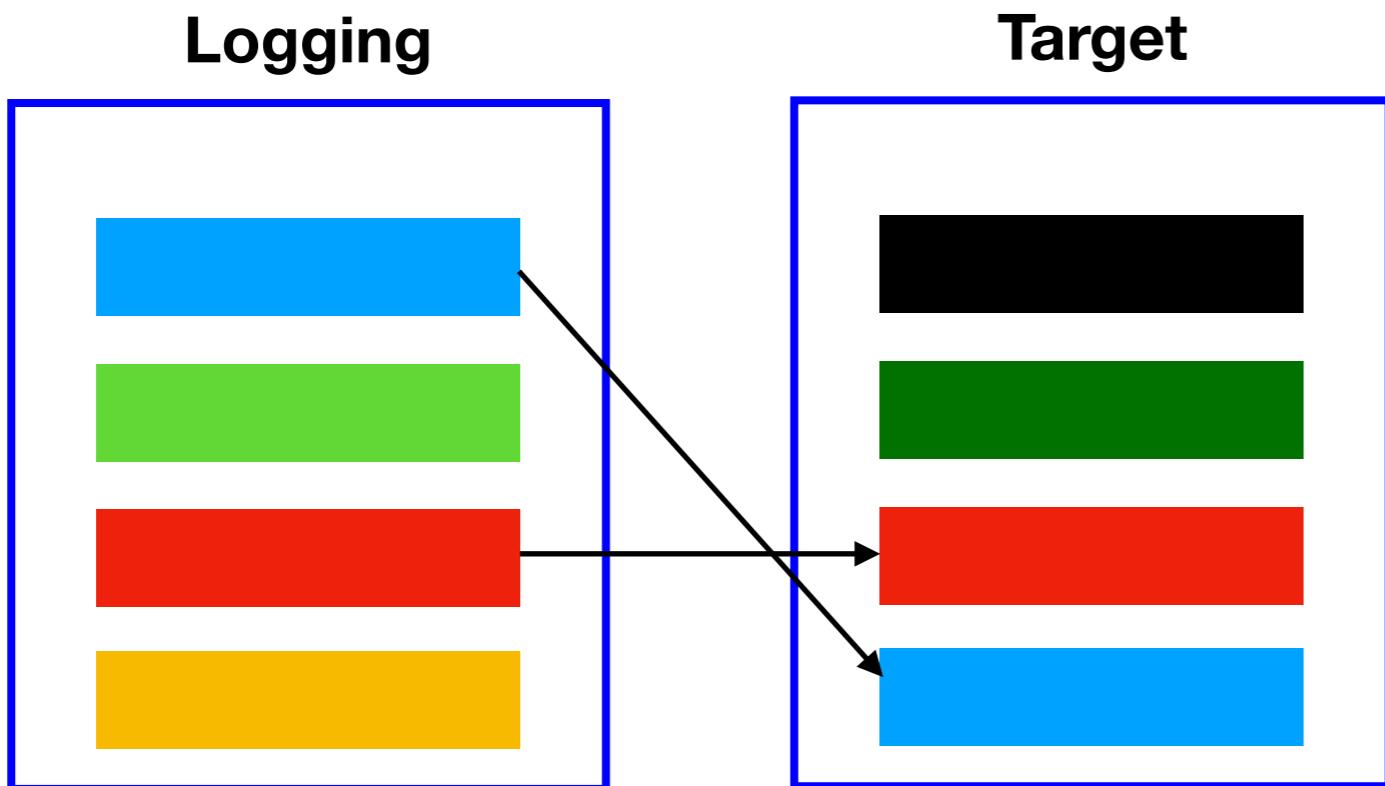
$$\hat{y}(a) = \frac{y(a)\mathbf{1}(a \in A)}{Q(a \in A|x)}$$

$$\hat{r}(\pi, x) = \sum_{a \in \pi(x)} \hat{y}(a)$$

- Uniform Q gives O(B) variance
- Immediately gives decent algorithm (eps-greedy)

Techniques – Off-policy evaluation

Subproblem: Given data collected by a logging policy, estimate reward of a target policy



Idea: Use partial matches!

If $A \sim Q(\cdot|x)$

$$\hat{y}(a) = \frac{y(a)\mathbf{1}(a \in A)}{Q(a \in A|x)}$$

$$\hat{r}(\pi, x) = \sum_{a \in \pi(x)} \hat{y}(a)$$

- Uniform Q gives O(B) variance
- Immediately gives decent algorithm (eps-greedy)
- We need more refined approach

Combinatorial Contextual Bandits

The screenshot shows a web browser window for the msn lifestyle website. The top navigation bar includes the msn logo, a search icon, a user profile for "Akshay" (with a blue circular icon), and a gear icon. Below the bar, there are menu links: Home, Cooking, Restaurants & News (which is underlined in purple), Beverages, and Video.

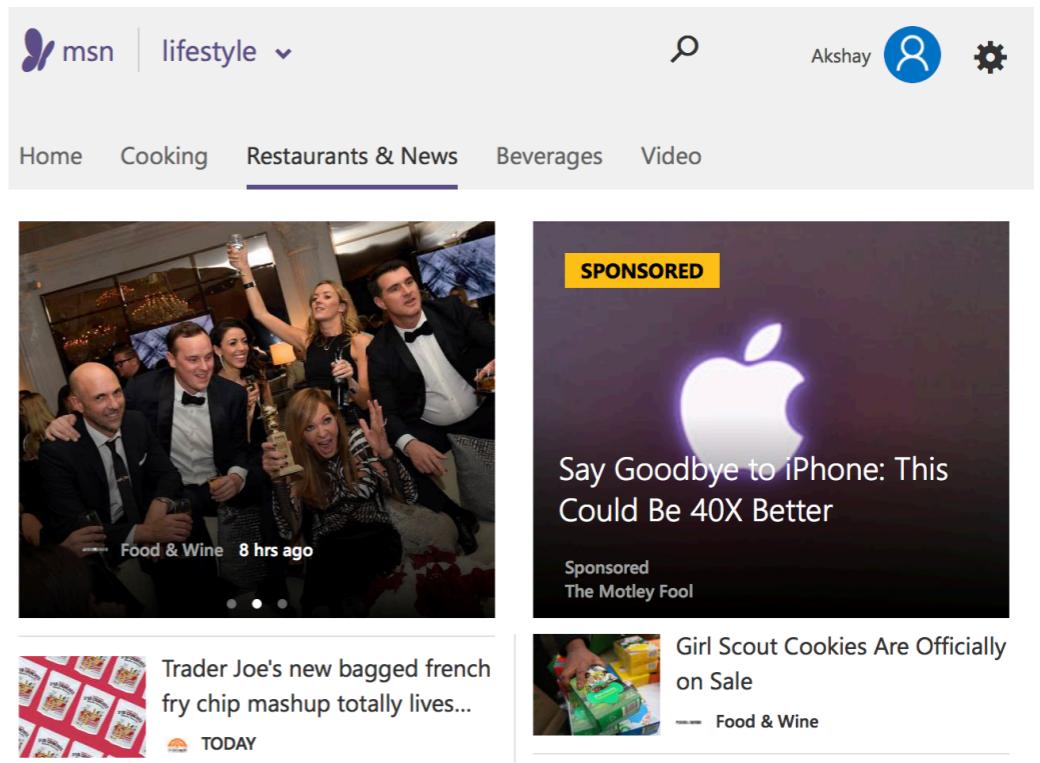
The main content area displays a news feed:

- A large image of a group of people in formal attire at a social gathering, with the caption "Food & Wine 8 hrs ago".
- A sponsored post from "The Motley Fool" titled "Say Goodbye to iPhone: This Could Be 40X Better", featuring a glowing Apple logo.
- A news item from "Food & Wine" about Trader Joe's new bagged french fry chip mashup.
- A news item from "Food & Wine" about Girl Scout Cookies being officially on sale.

Combinatorial Contextual Bandits

On each of T rounds:

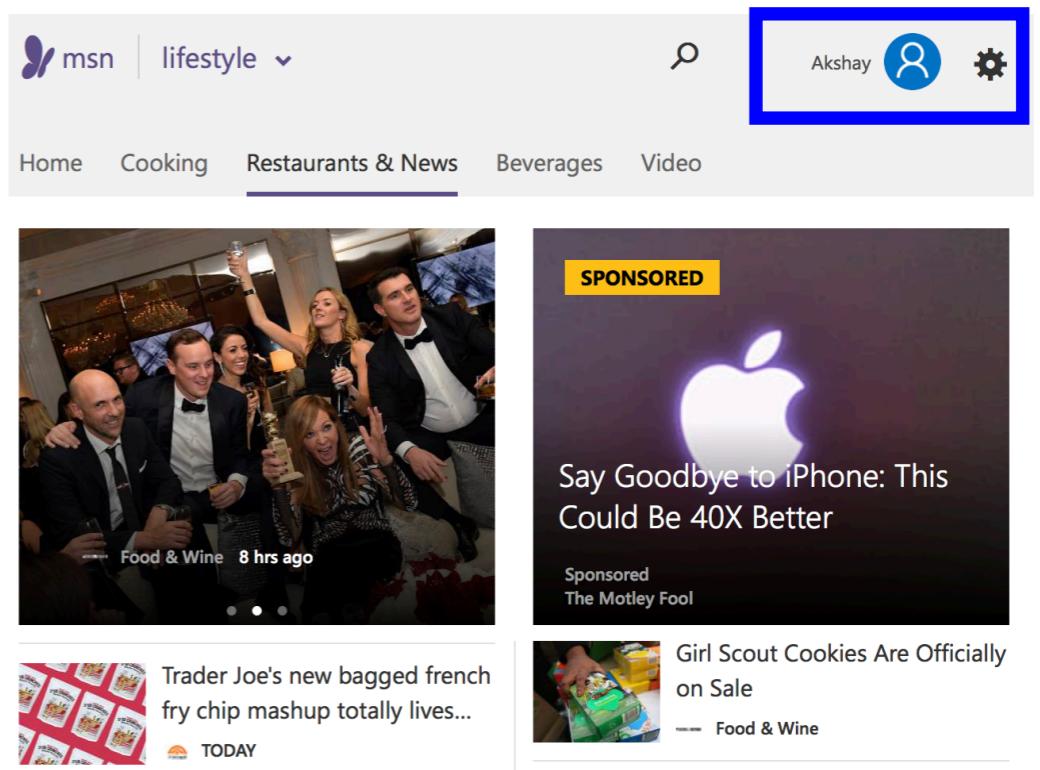
1. Observe context
2. Play action
3. Unobserved features
4. Observe reward



Combinatorial Contextual Bandits

On each of T rounds:

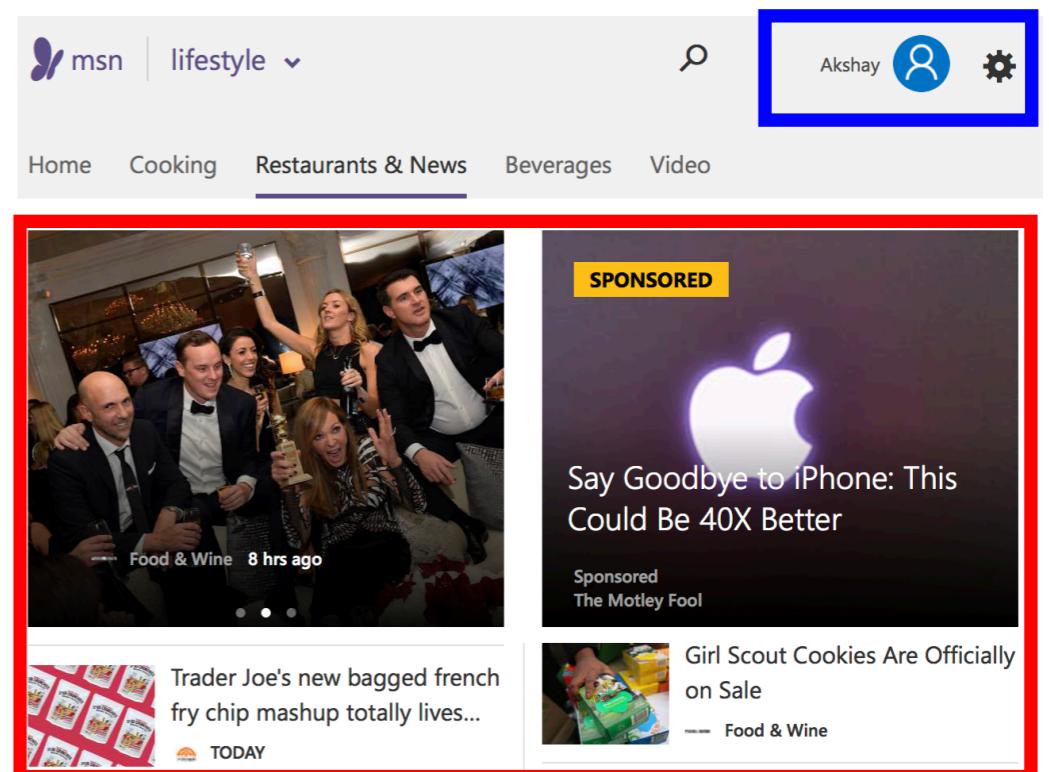
1. Observe context x_t
2. Play action
3. Unobserved features
4. Observe reward



Combinatorial Contextual Bandits

On each of T rounds:

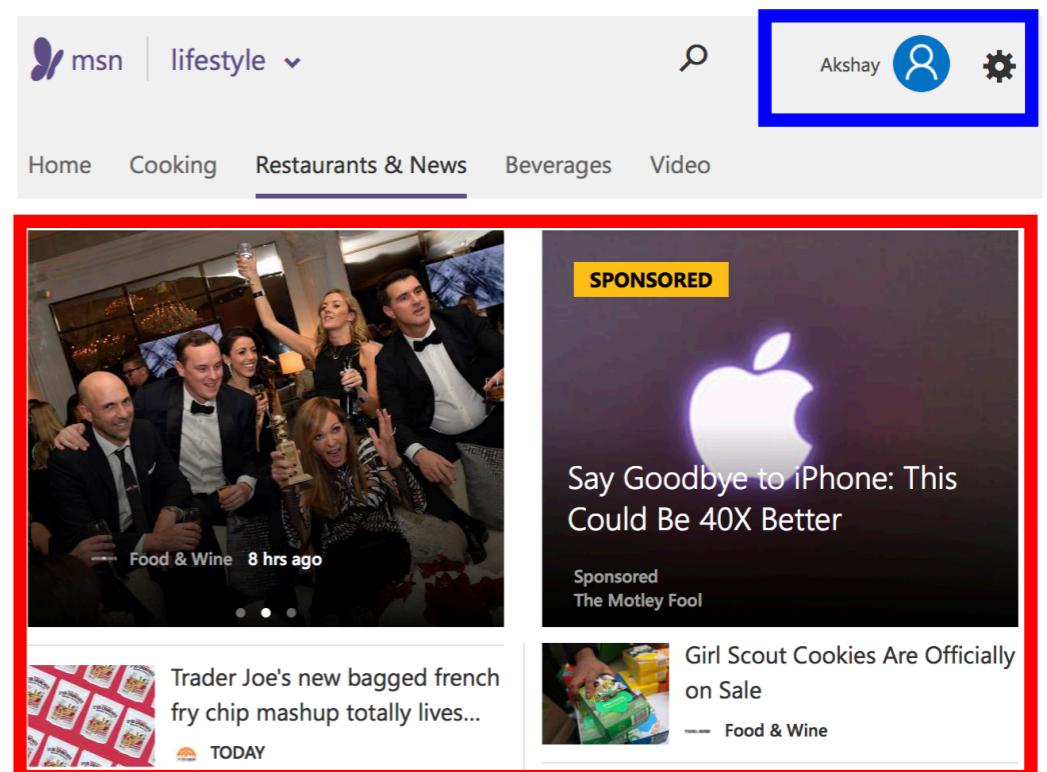
1. Observe context x_t
2. Play action $A_t = (a_1, \dots, a_L)$
3. Unobserved features
4. Observe reward



Combinatorial Contextual Bandits

On each of T rounds:

1. Observe context x_t
2. Play action $A_t = (a_1, \dots, a_L)$
3. Unobserved features $\{y(\ell, a_\ell)\}_{\ell=1}^L$
4. Observe reward

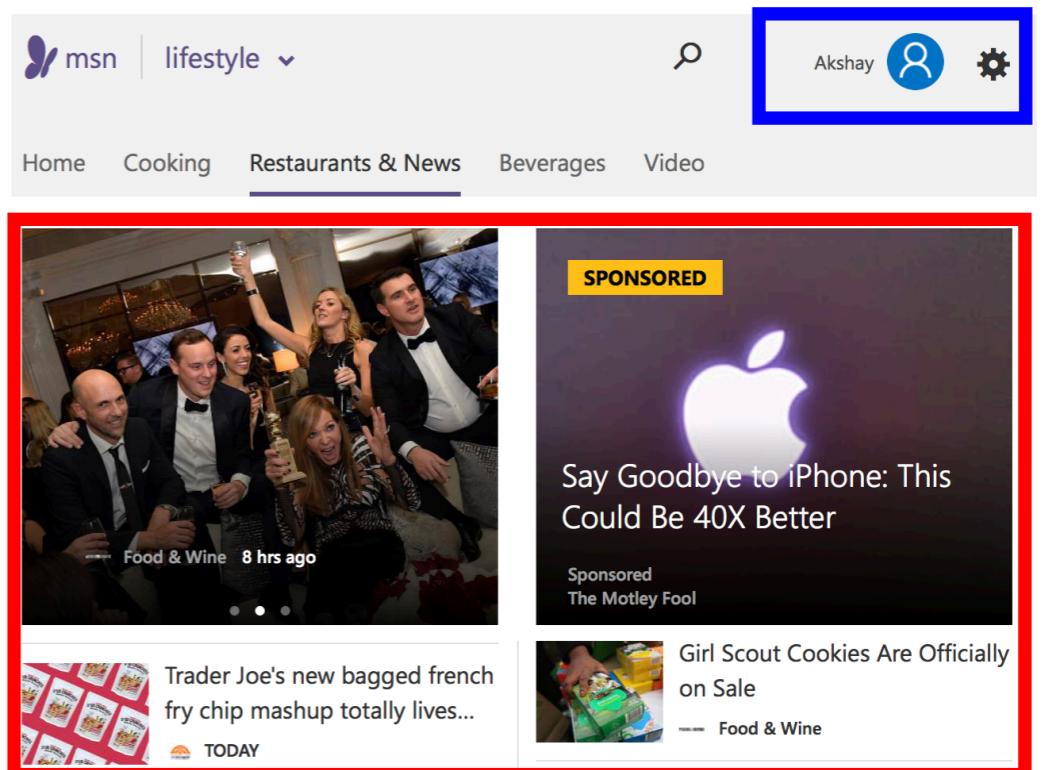


Combinatorial Contextual Bandits

On each of T rounds:

1. Observe context x_t
2. Play action $A_t = (a_1, \dots, a_L)$
3. Unobserved features $\{y(\ell, a_\ell)\}_{\ell=1}^L$
4. Observe reward

$$r_t(A_t, x_t) = \sum_{\ell} y(\ell, a_\ell) + \text{noise}$$



Combinatorial Contextual Bandits

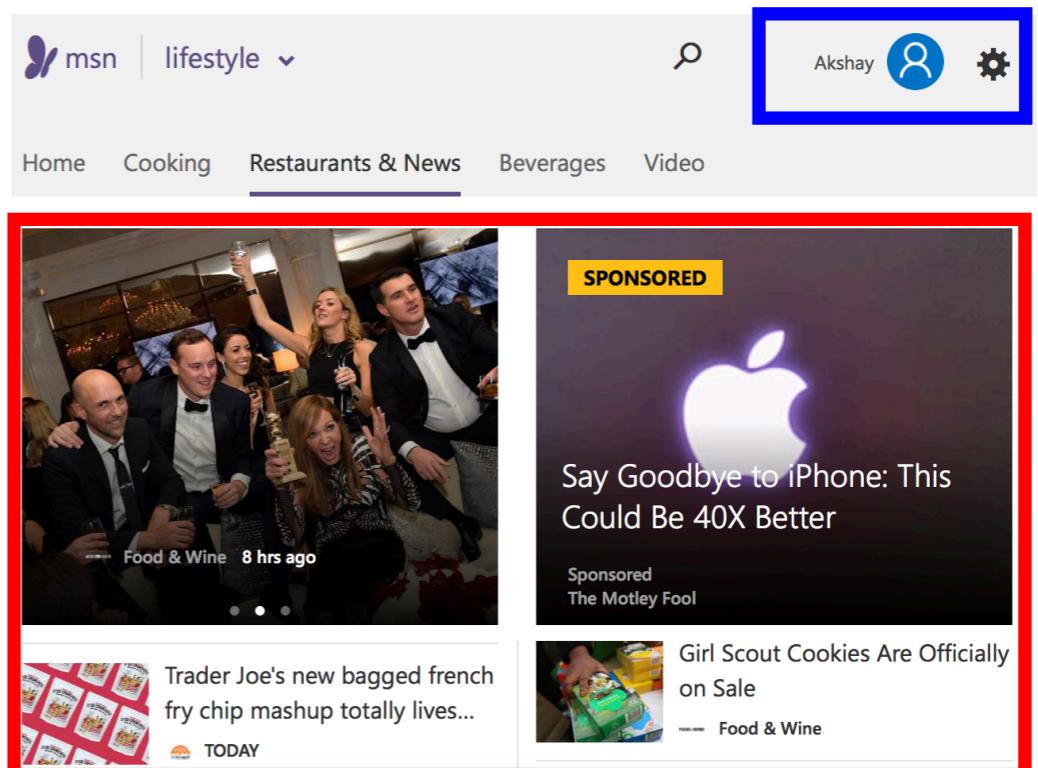
On each of T rounds:

1. Observe context x_t
2. Play action $A_t = (a_1, \dots, a_L)$
3. Unobserved features $\{y(\ell, a_\ell)\}_{\ell=1}^L$
4. Observe reward

$$r_t(A_t, x_t) = \sum_{\ell} y(\ell, a_\ell) + \text{noise}$$

B = number of simple actions

L = composite action length



Combinatorial Contextual Bandits

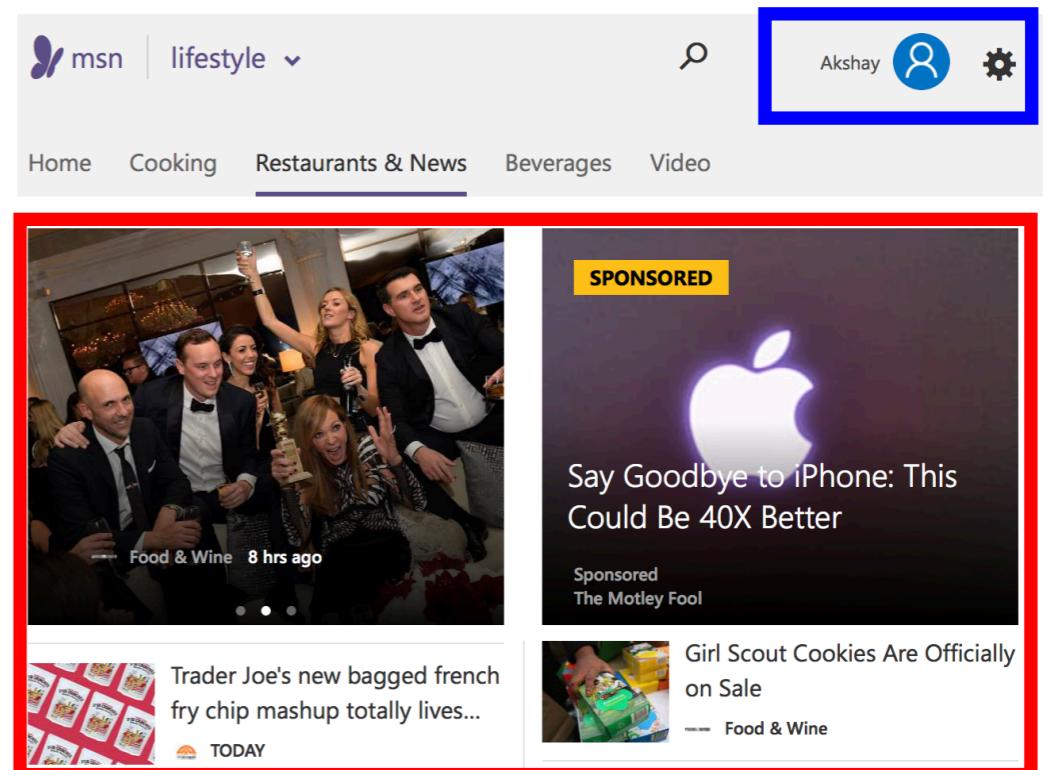
On each of T rounds:

1. Observe context x_t
2. Play action $A_t = (a_1, \dots, a_L)$
3. Unobserved features $\{y(\ell, a_\ell)\}_{\ell=1}^L$
4. Observe reward

$$r_t(A_t, x_t) = \sum_{\ell} y(\ell, a_\ell) + \text{noise}$$

B = number of simple actions

L = composite action length



Question: Improve performance by leveraging reward structure?

Combinatorial Contextual Bandits

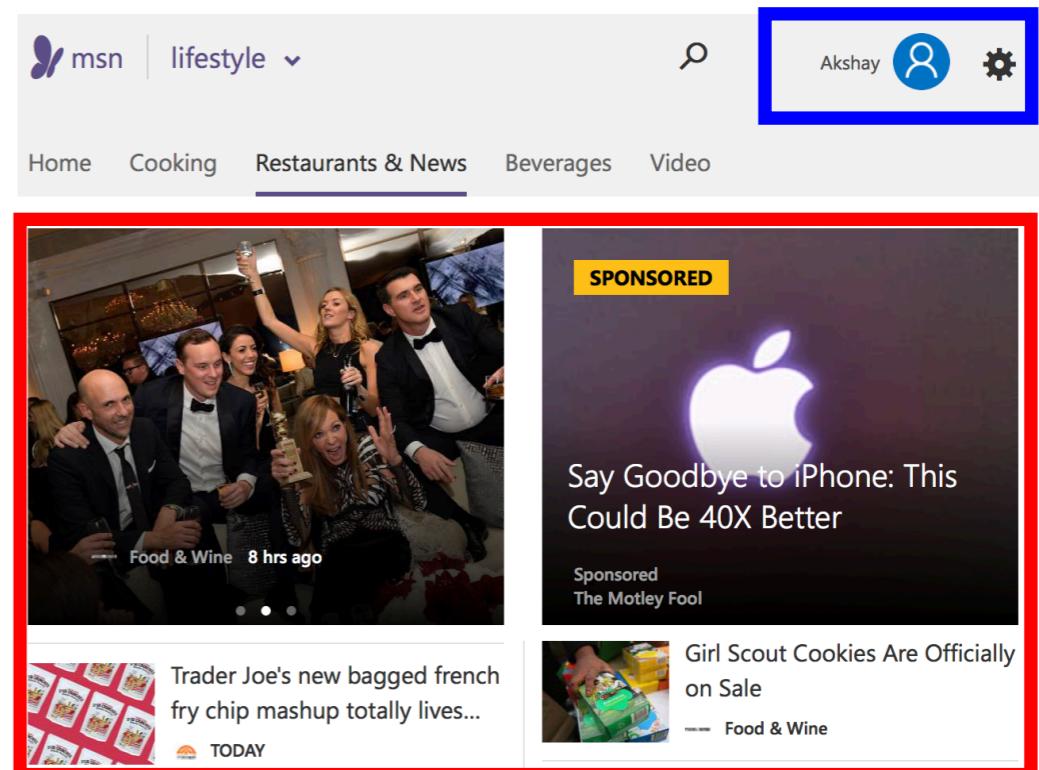
On each of T rounds:

1. Observe context x_t
2. Play action $A_t = (a_1, \dots, a_L)$
3. Unobserved features $\{y(\ell, a_\ell)\}_{\ell=1}^L$
4. Observe reward

$$r_t(A_t, x_t) = \sum_{\ell} y(\ell, a_\ell) + \text{noise}$$

B = number of simple actions

L = composite action length



Question: Improve performance by leveraging reward structure?

Challenges:

- Off-policy evaluation?
- Explore vs Exploit?
- Computational Efficiency?

Results – Off-policy evaluation

Subproblem: Given data collected by logging policy, estimate reward of a target policy

Results – Off-policy evaluation

Subproblem: Given data collected by logging policy, estimate reward of a target policy

Theorem: If logging close to uniform, can estimate target with BL/ϵ^2 samples

Parameters: B simple actions, composite action length L

Results – Off-policy evaluation

Subproblem: Given data collected by logging policy, estimate reward of a target policy

Theorem: If logging close to uniform, can estimate target with BL/ϵ^2 samples

Parameters: B simple actions, composite action length L

- Compare with $O(B^L)$ naively and $O(B)$ with semibandit feedback

Results – Off-policy evaluation

Subproblem: Given data collected by logging policy, estimate reward of a target policy

Theorem: If logging close to uniform, can estimate target with BL/ϵ^2 samples

Parameters: B simple actions, composite action length L

- Compare with $O(B^L)$ naively and $O(B)$ with semibandit feedback
- Gives decent eps-greedy algorithm with $T^{2/3}(BL)^{1/3}$ regret

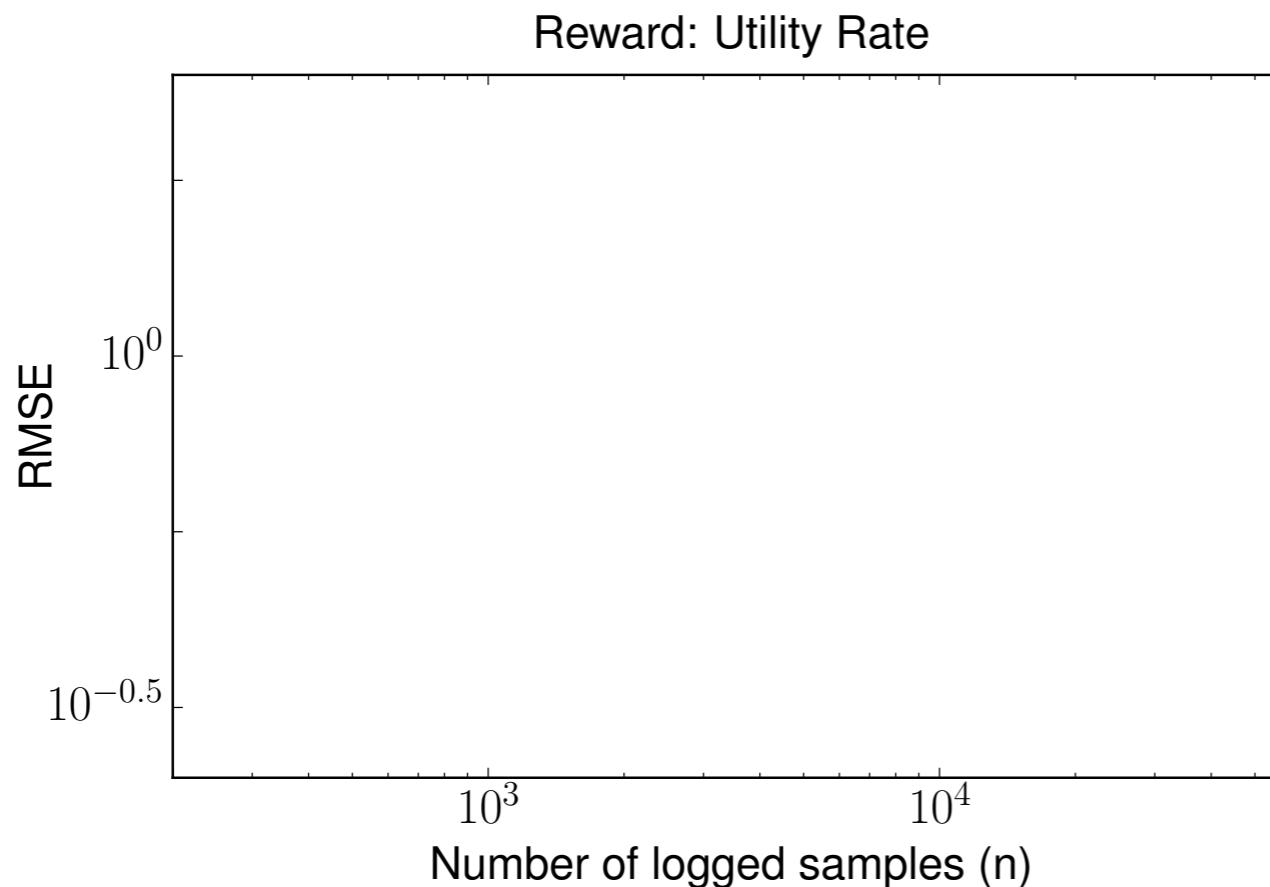
Results – Off-policy evaluation

Subproblem: Given data collected by logging policy, estimate reward of a target policy

Theorem: If logging close to uniform, can estimate target with BL/ϵ^2 samples

Parameters: B simple actions, composite action length L

- Compare with $O(B^L)$ naively and $O(B)$ with semibandit feedback
- Gives decent eps-greedy algorithm with $T^{2/3}(BL)^{1/3}$ regret



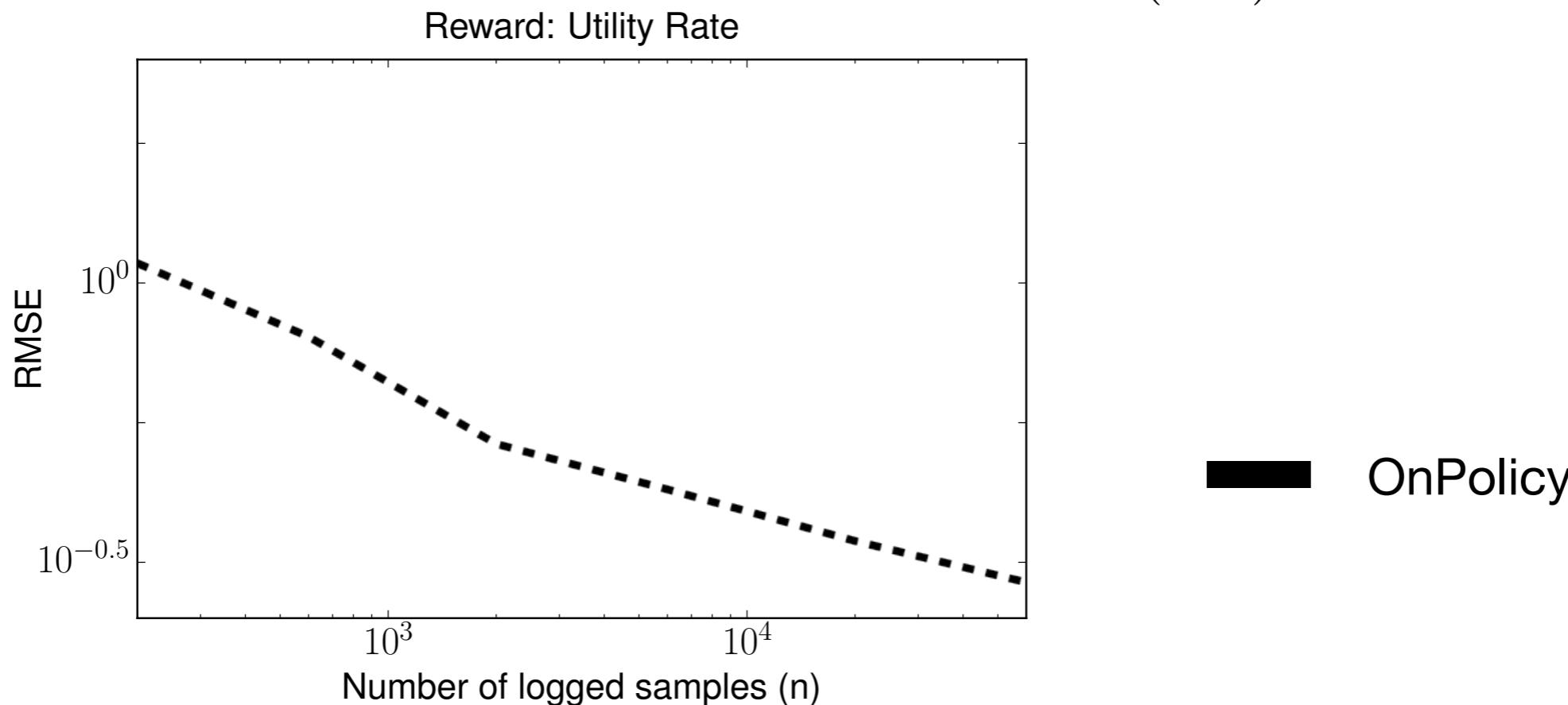
Results – Off-policy evaluation

Subproblem: Given data collected by logging policy, estimate reward of a target policy

Theorem: If logging close to uniform, can estimate target with BL/ϵ^2 samples

Parameters: B simple actions, composite action length L

- Compare with $O(B^L)$ naively and $O(B)$ with semibandit feedback
- Gives decent eps-greedy algorithm with $T^{2/3}(BL)^{1/3}$ regret



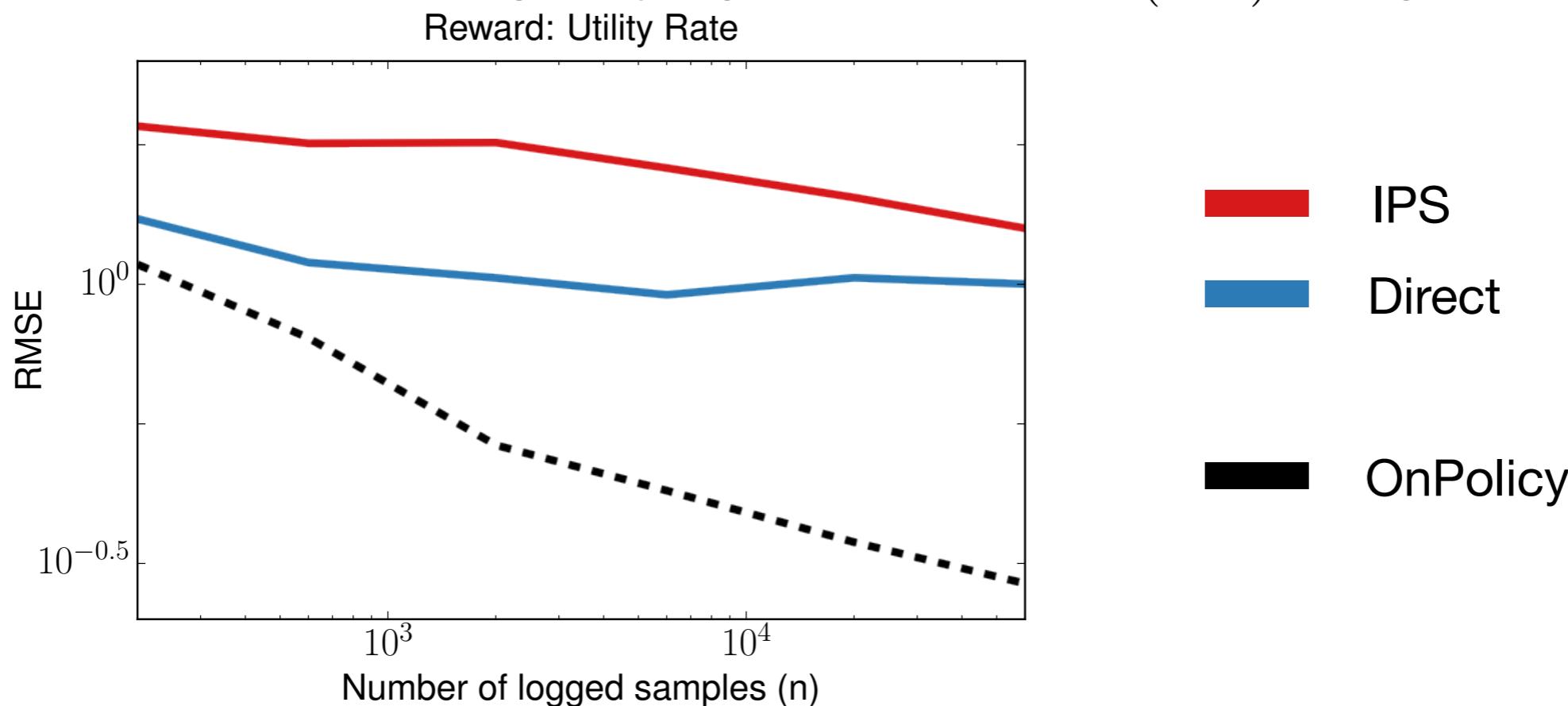
Results – Off-policy evaluation

Subproblem: Given data collected by logging policy, estimate reward of a target policy

Theorem: If logging close to uniform, can estimate target with BL/ϵ^2 samples

Parameters: B simple actions, composite action length L

- Compare with $O(B^L)$ naively and $O(B)$ with semibandit feedback
- Gives decent ϵ -greedy algorithm with $T^{2/3}(BL)^{1/3}$ regret



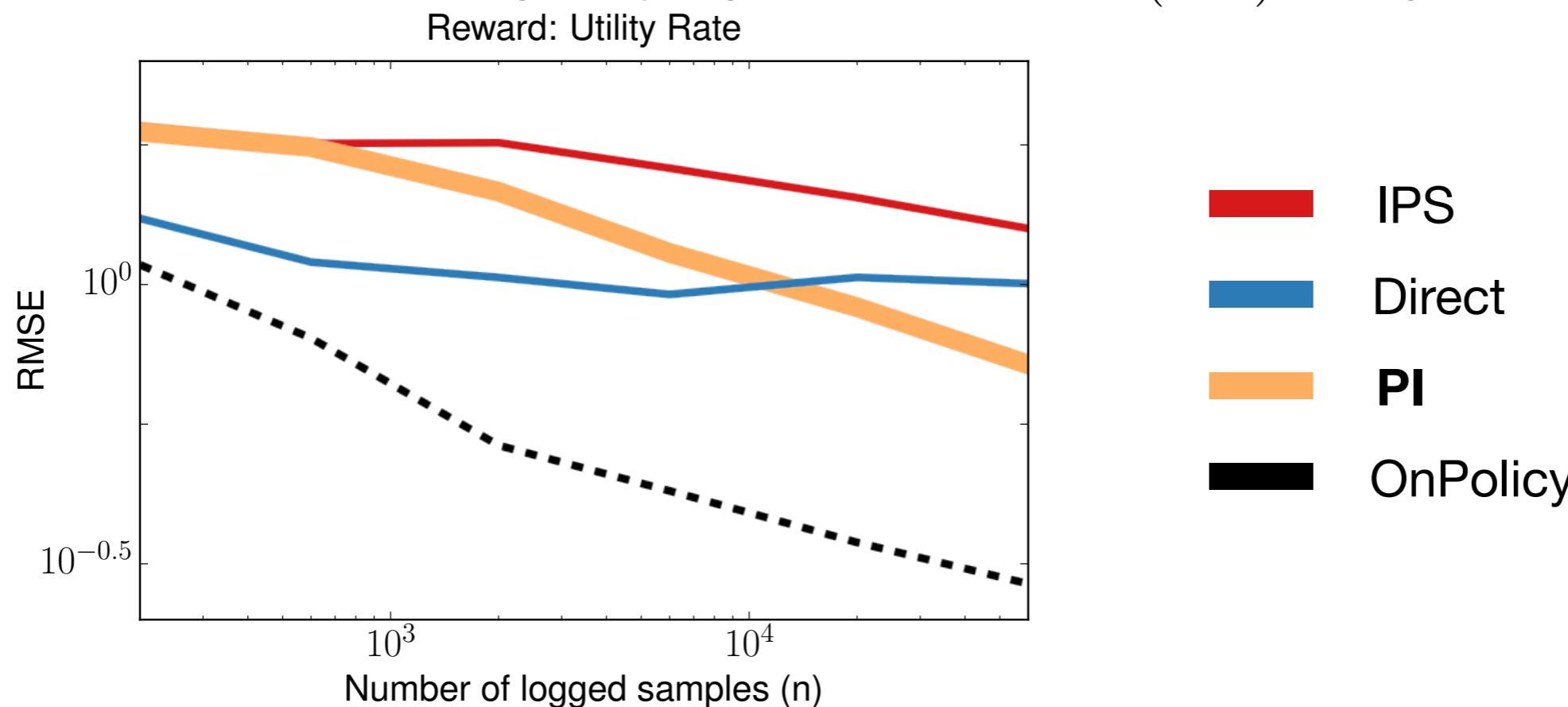
Results – Off-policy evaluation

Subproblem: Given data collected by logging policy, estimate reward of a target policy

Theorem: If logging close to uniform, can estimate target with BL/ϵ^2 samples

Parameters: B simple actions, composite action length L

- Compare with $O(B^L)$ naively and $O(B)$ with semibandit feedback
- Gives decent ϵ -greedy algorithm with $T^{2/3}(BL)^{1/3}$ regret



Techniques

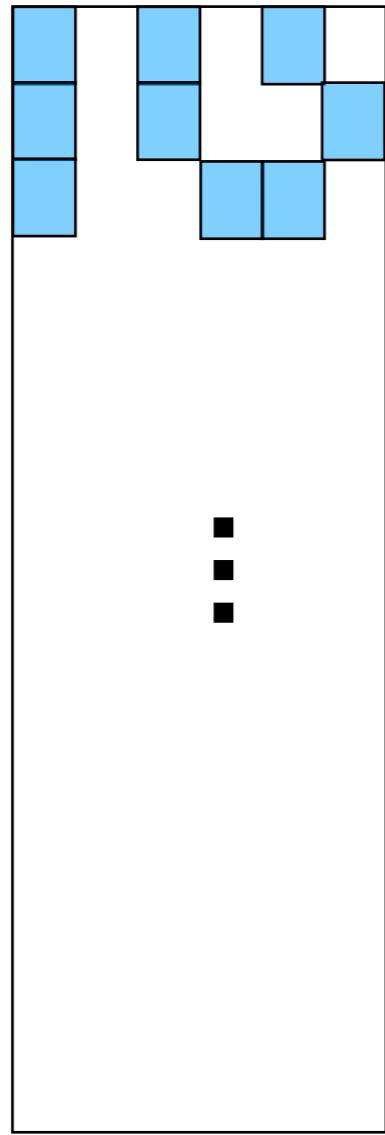
$$r_t(A_t, x_t) = \sum_{\ell} y(a_{\ell}) + \text{noise}$$

Techniques

$r(A)$



=



$y(\ell, a_\ell)$

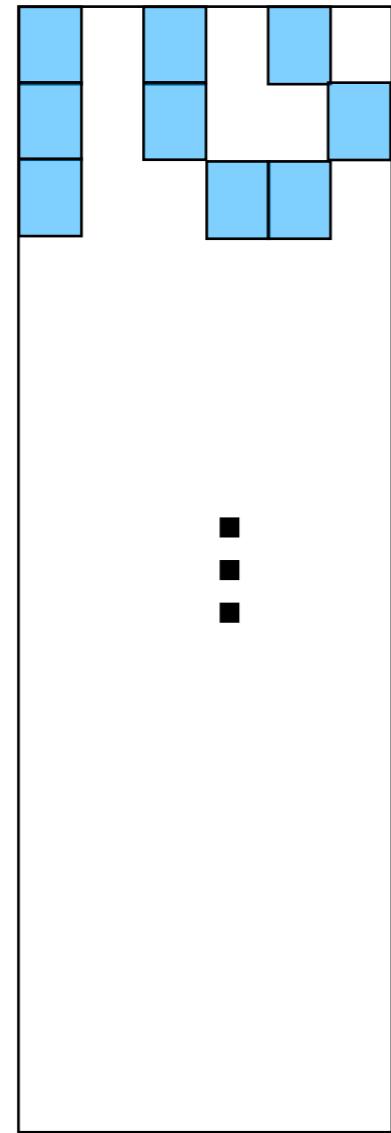
$$r_t(A_t, x_t) = \sum_{\ell} y(a_\ell) + \text{noise}$$

Techniques

$r(A)$



=



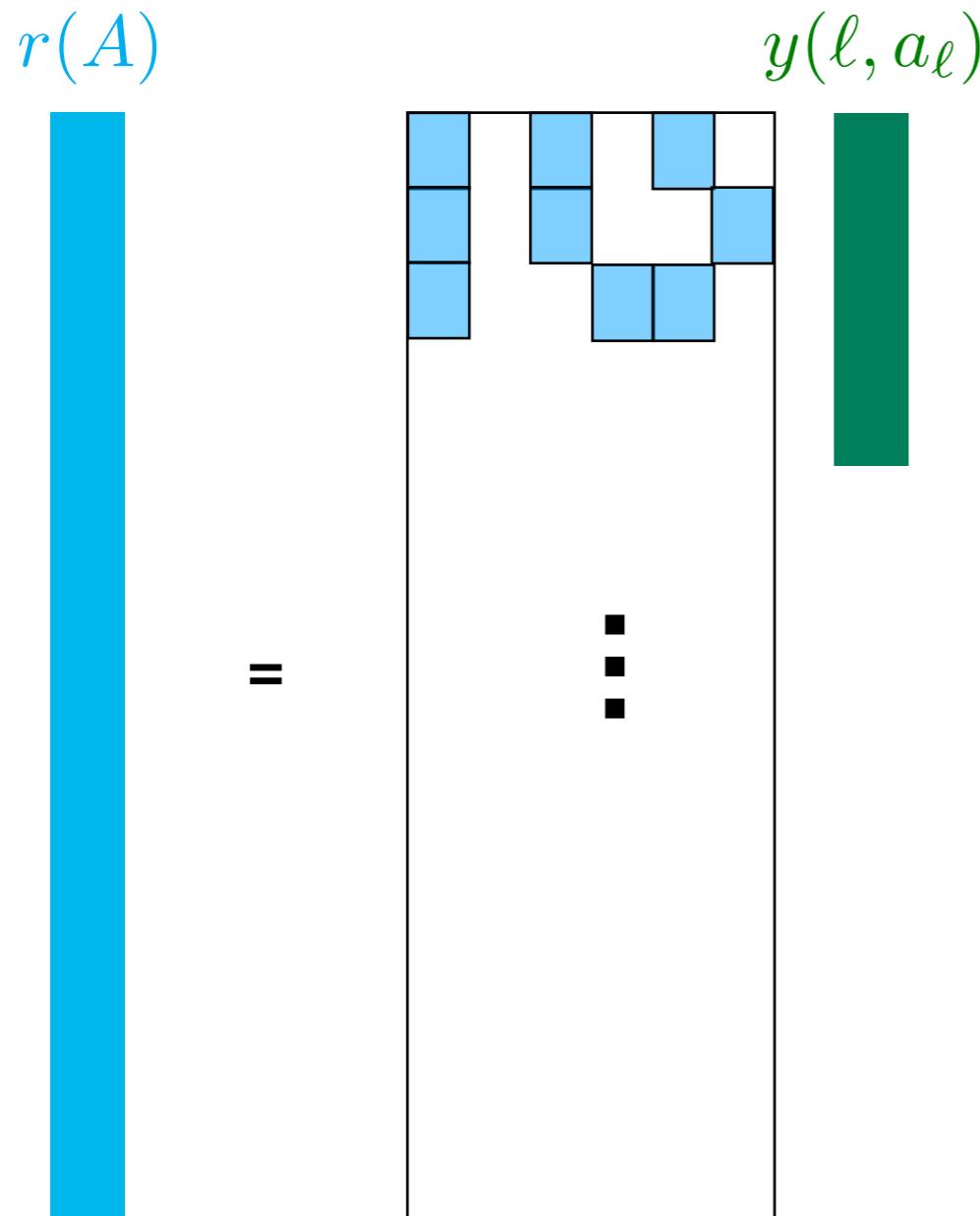
$y(\ell, a_\ell)$

$$r_t(A_t, x_t) = \sum_{\ell} y(a_\ell) + \text{noise}$$

With logging μ can write

$$\bar{y} = \arg \min_w \mathbb{E}_\mu[(\mathbf{1}_A^T w - r)^2 | x]$$

Techniques



$$r_t(A_t, x_t) = \sum_{\ell} y(a_\ell) + \text{noise}$$

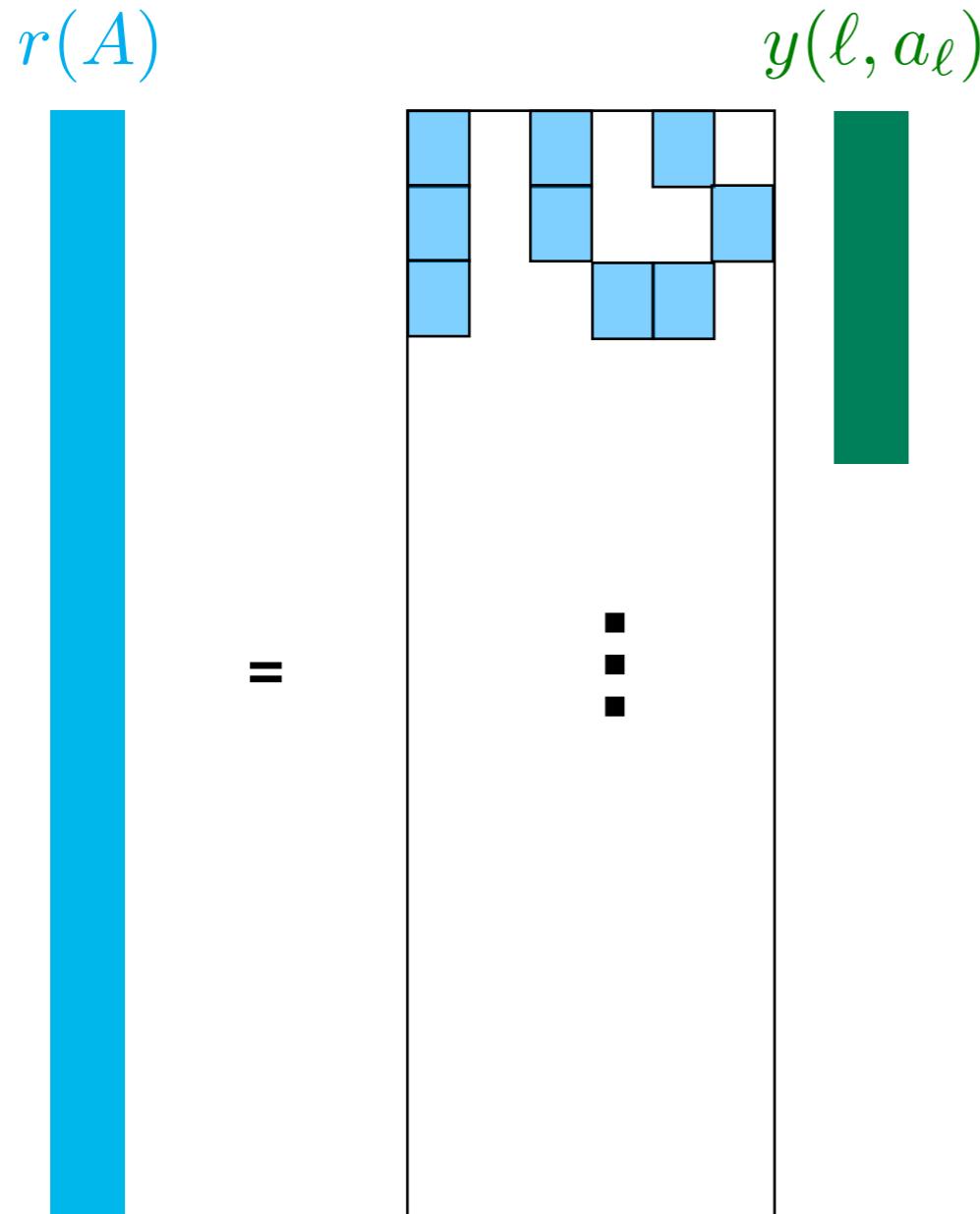
With logging μ can write

$$\bar{y} = \arg \min_w \mathbb{E}_\mu[(\mathbf{1}_A^T w - r)^2 | x]$$

So with (A_t, r_t) estimate

$$\hat{y}_t = (\mathbb{E}_\mu[\mathbf{1}_A \mathbf{1}_A^T])^\dagger \mathbf{1}_{A_t} r_t$$

Techniques



$$r_t(A_t, x_t) = \sum_{\ell} y(a_\ell) + \text{noise}$$

With logging μ can write

$$\bar{y} = \arg \min_w \mathbb{E}_\mu[(\mathbf{1}_A^T w - r)^2 | x]$$

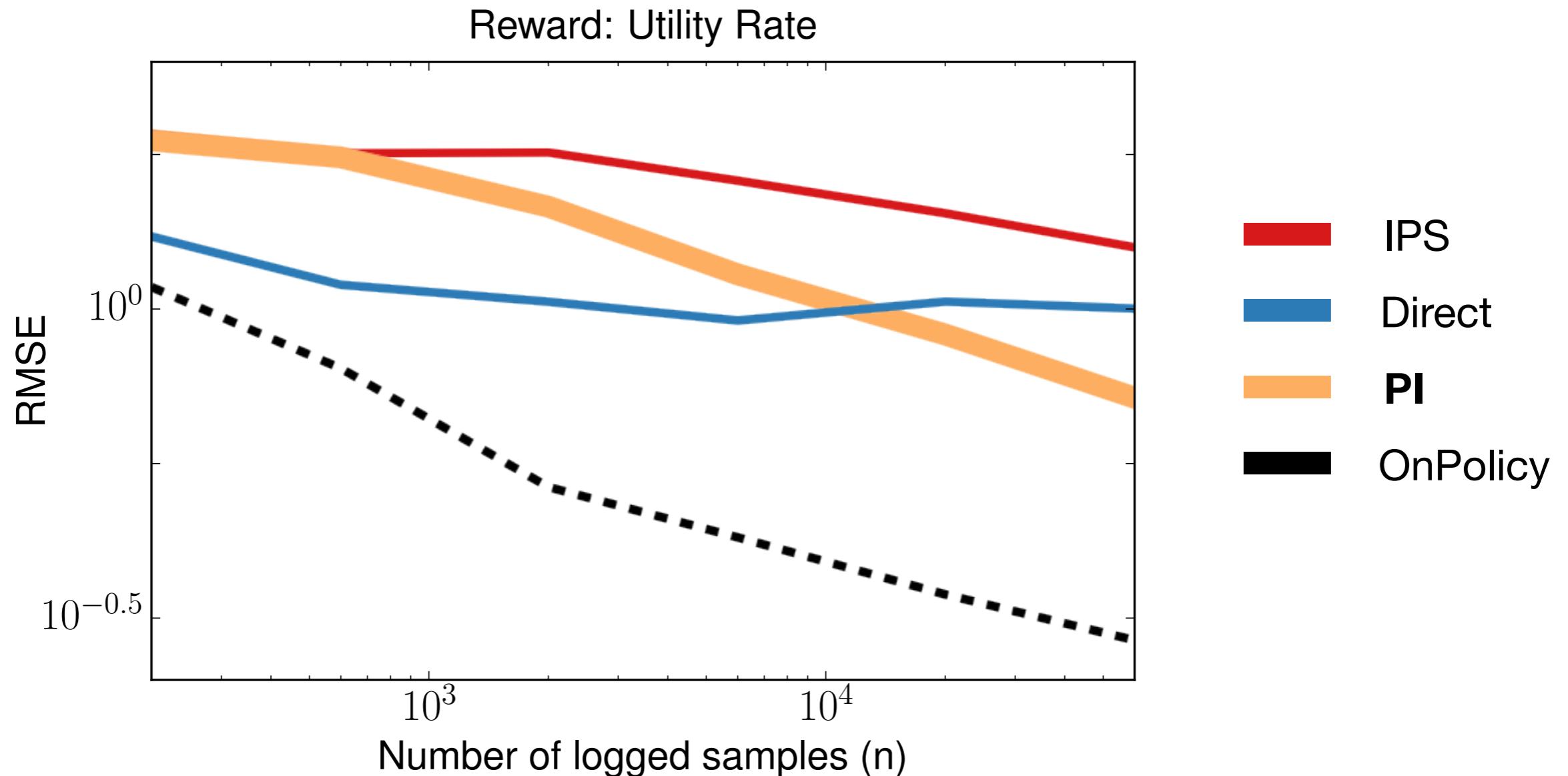
So with (A_t, r_t) estimate

$$\hat{y}_t = (\mathbb{E}_\mu[\mathbf{1}_A \mathbf{1}_A^T])^\dagger \mathbf{1}_{A_t} r_t$$

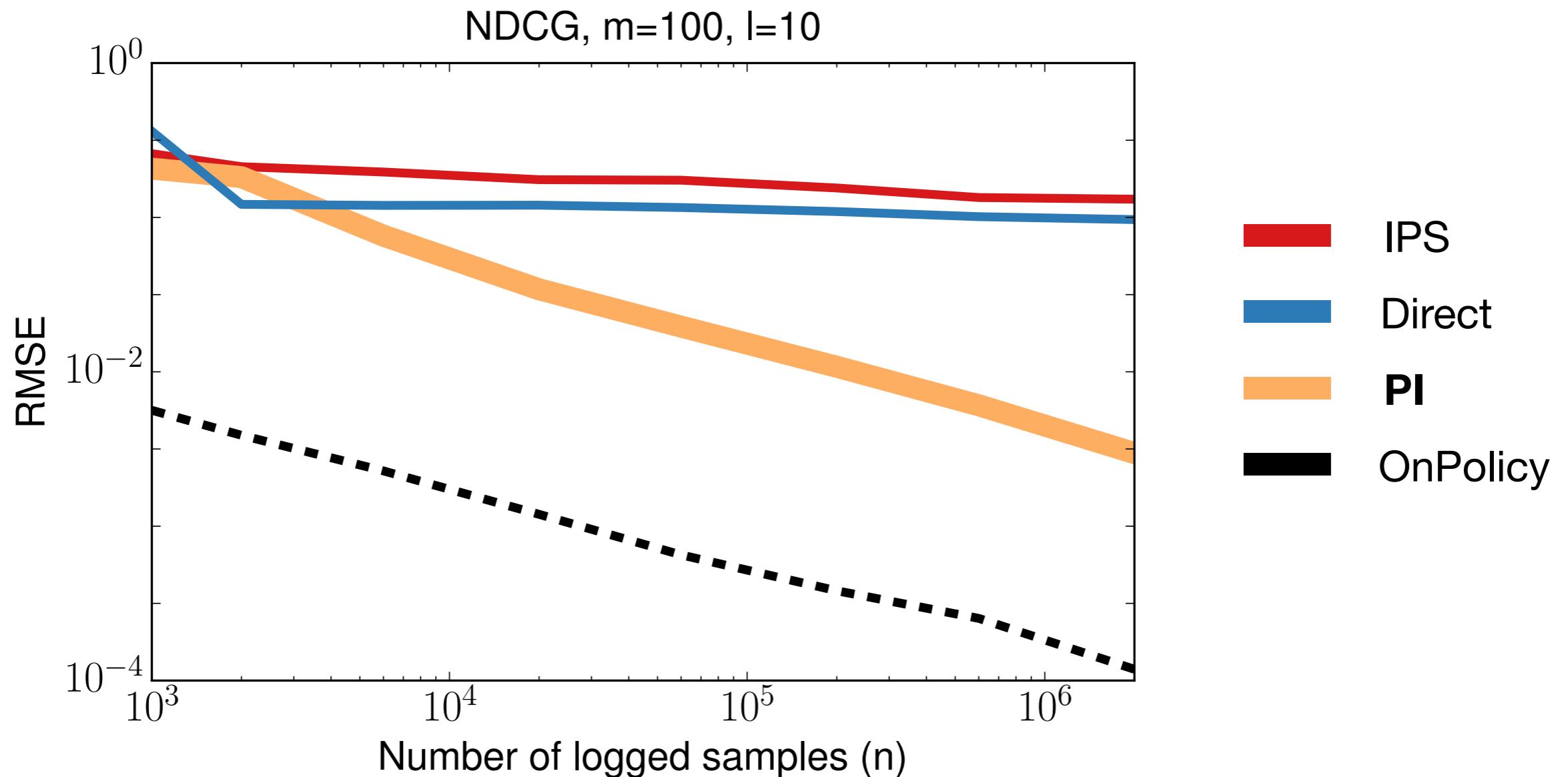
For policy π

$$\hat{r}(\pi, x_t) = \mathbf{1}_{\pi(x_t)}^T \hat{y}_t$$

Experiment



Experiment



Policy Optimization

- Use PI estimator to obtain, with x_t

$$\hat{y}_t = (\mathbb{E}_{\mu}[\mathbf{1}_A \mathbf{1}_A^T])^\dagger \mathbf{1}_{A_t} r_t$$

Policy Optimization

- Use PI estimator to obtain, with x_t

$$\hat{y}_t = (\mathbb{E}_{\mu}[\mathbf{1}_A \mathbf{1}_A^T])^\dagger \mathbf{1}_{A_t} r_t$$

- Akin to supervised learning to rank dataset

Policy Optimization

- Use PI estimator to obtain, with x_t

$$\hat{y}_t = (\mathbb{E}_{\mu}[\mathbf{1}_A \mathbf{1}_A^T])^\dagger \mathbf{1}_{A_t} r_t$$

- Akin to supervised learning to rank dataset
- Train L2R model via regression

Policy Optimization

- Use PI estimator to obtain, with x_t

$$\hat{y}_t = (\mathbb{E}_{\mu}[\mathbf{1}_A \mathbf{1}_A^T])^\dagger \mathbf{1}_{A_t} r_t$$

- Akin to supervised learning to rank dataset
- Train L2R model via regression

Metric	LambdaMART	Random	SUP	PI

Policy Optimization

- Use PI estimator to obtain, with x_t

$$\hat{y}_t = (\mathbb{E}_{\mu}[\mathbf{1}_A \mathbf{1}_A^T])^\dagger \mathbf{1}_{A_t} r_t$$

- Akin to supervised learning to rank dataset
- Train L2R model via regression

Metric	LambdaMART	Random	SUP	PI
NDCG	0.457	0.152	0.438	0.421

Policy Optimization

- Use PI estimator to obtain, with x_t

$$\hat{y}_t = (\mathbb{E}_{\mu}[\mathbf{1}_A \mathbf{1}_A^T])^\dagger \mathbf{1}_{A_t} r_t$$

- Akin to supervised learning to rank dataset
- Train L2R model via regression

Metric	LambdaMART	Random	SUP	PI
NDCG	0.457	0.152	0.438	0.421
ERR	—	0.096	0.311	0.321

Policy Optimization

- Use PI estimator to obtain, with x_t

$$\hat{y}_t = (\mathbb{E}_{\mu}[\mathbf{1}_A \mathbf{1}_A^T])^\dagger \mathbf{1}_{A_t} r_t$$

- Akin to supervised learning to rank dataset
- Train L2R model via regression

Metric	LambdaMART	Random	SUP	PI
NDCG	0.457	0.152	0.438	0.421
ERR	—	0.096	0.311	0.321

PI finds good targets to optimize metric!

Summary

Summary

Naive CB

Semibandits

Combinatorial

Parameters: T rounds, B simple actions, composite action length L

Summary

	Naive CB	Semibandits	Combinatorial
Off-Policy Eval	B^L	B	BL
On-Policy Eval	B^L	B	BL

Parameters: T rounds, B simple actions, composite action length L

Summary

	Naive CB	Semibandits	Combinatorial
Off-Policy Eval	B^L	B	BL
Explore/Exploit	$\sqrt{B^L T \log(\Pi)}$	$\sqrt{BT \log(\Pi)}$	$T^{2/3} (BL \log(\Pi))^{1/3}$

Parameters: T rounds, B simple actions, composite action length L

Summary

	Naive CB	Semibandits	Combinatorial
Off-Policy Eval	B^L	B	BL
Explore/Exploit	$\sqrt{B^L T \log(\Pi)}$	$\sqrt{BT \log(\Pi)}$	$T^{2/3} (BL \log(\Pi))^{1/3}$

Parameters: T rounds, B simple actions, composite action length L

Empirically

Summary

	Naive CB	Semibandits	Combinatorial
Off-Policy Eval	B^L	B	BL
Explore/Exploit	$\sqrt{B^L T \log(\Pi)}$	$\sqrt{BT \log(\Pi)}$	$T^{2/3} (BL \log(\Pi))^{1/3}$

Parameters: T rounds, B simple actions, composite action length L

Empirically

- Semibandits — With rich policy class, best performance

Summary

	Naive CB	Semibandits	Combinatorial
Off-Policy Eval	B^L	B	BL
Explore/Exploit	$\sqrt{B^L T \log(\Pi)}$	$\sqrt{BT \log(\Pi)}$	$T^{2/3} (BL \log(\Pi))^{1/3}$

Parameters: T rounds, B simple actions, composite action length L

Empirically

- Semibandits — With rich policy class, best performance
- Off-Policy Eval — Works in practice, even without linearity

Summary

	Naive CB	Semibandits	Combinatorial
Off-Policy Eval	B^L	B	BL
Explore/Exploit	$\sqrt{B^L T \log(\Pi)}$	$\sqrt{BT \log(\Pi)}$	$T^{2/3} (BL \log(\Pi))^{1/3}$

Parameters: T rounds, B simple actions, composite action length L

Empirically

- Semibandits — With rich policy class, best performance
- Off-Policy Eval — Works in practice, even without linearity
- Off-Policy Opt — Finds better targets than supervision!

Summary

	Naive CB	Semibandits	Combinatorial
Off-Policy Eval	B^L	B	BL
Explore/Exploit	$\sqrt{B^L T \log(\Pi)}$	$\sqrt{BT \log(\Pi)}$	$T^{2/3} (BL \log(\Pi))^{1/3}$

Parameters: T rounds, B simple actions, composite action length L

Empirically

- Semibandits — With rich policy class, best performance
- Off-Policy Eval — Works in practice, even without linearity
- Off-Policy Opt — Finds better targets than supervision!

Open

- Efficient CCB with \sqrt{T} regret