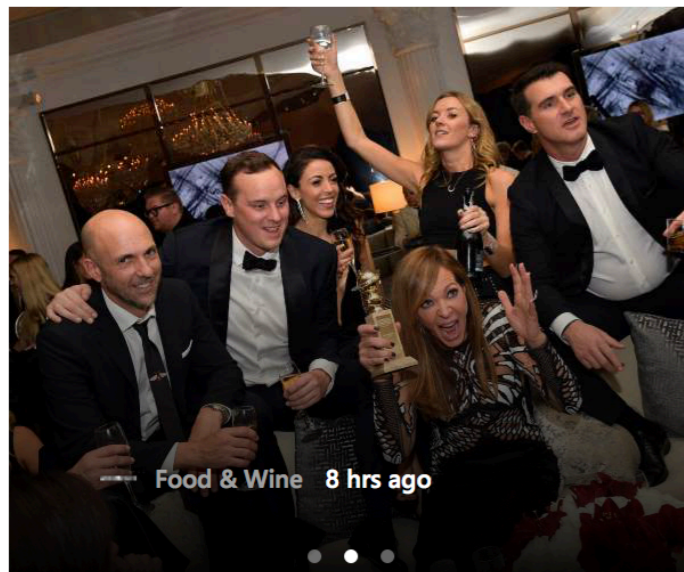
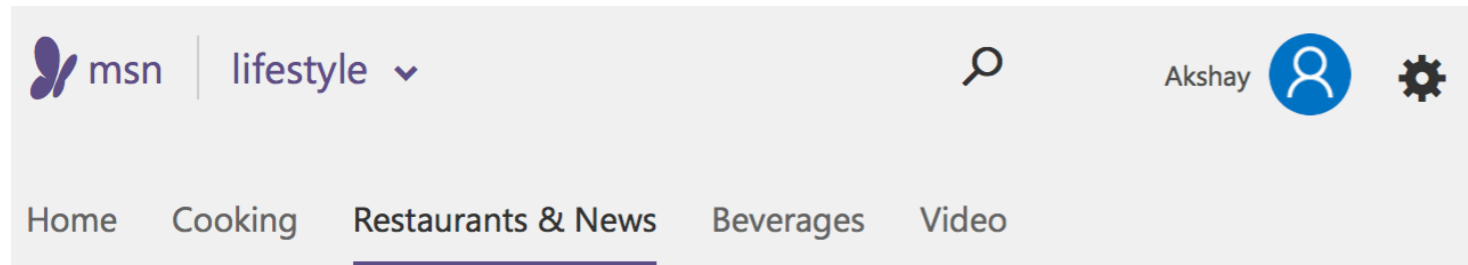


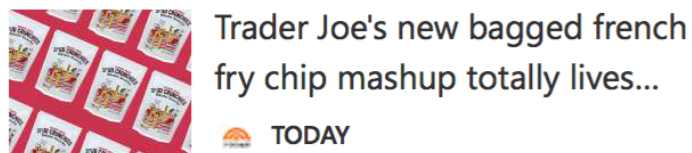
Online and Offline Experimentation in Complex Systems

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

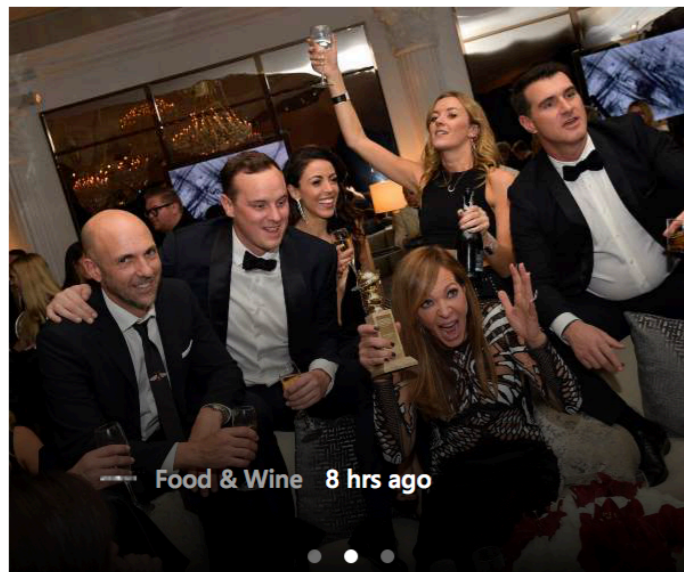
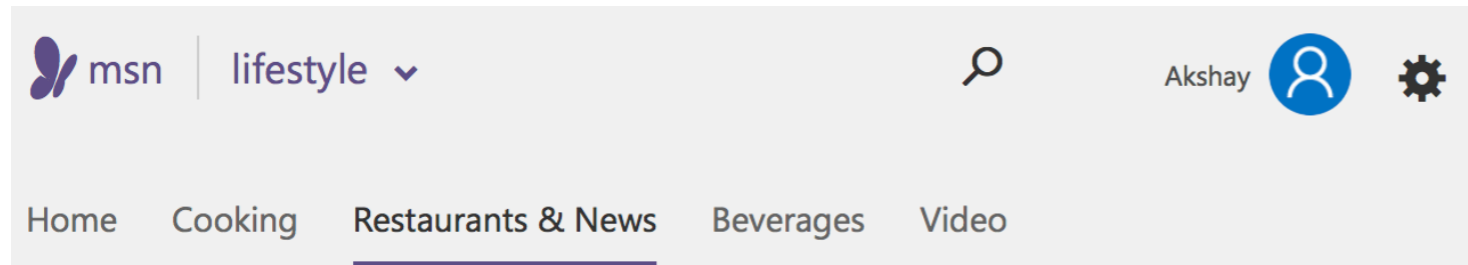
Online Personalization



- Learn from interacting with users in production



Online Personalization



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

TODAY

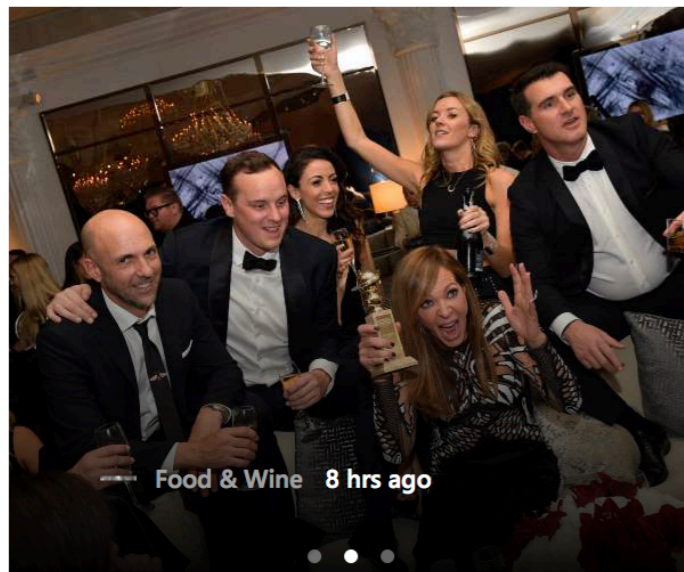
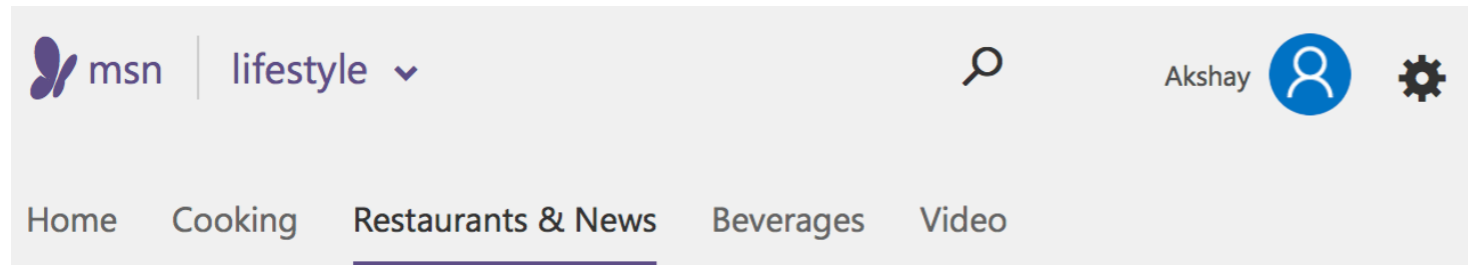


Girl Scout Cookies Are Officially on Sale

Food & Wine

- Learn from interacting with users in production
- No counterfactuals

Online Personalization



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



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

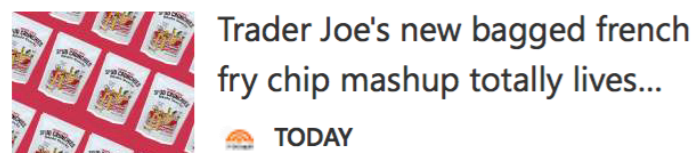
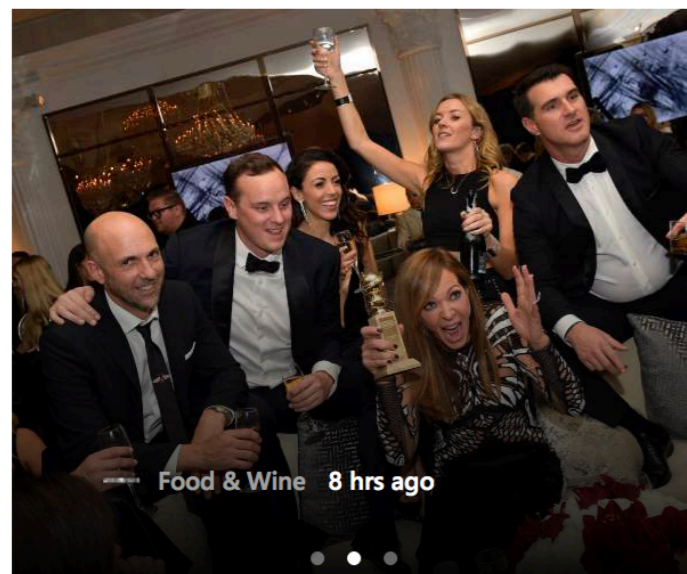
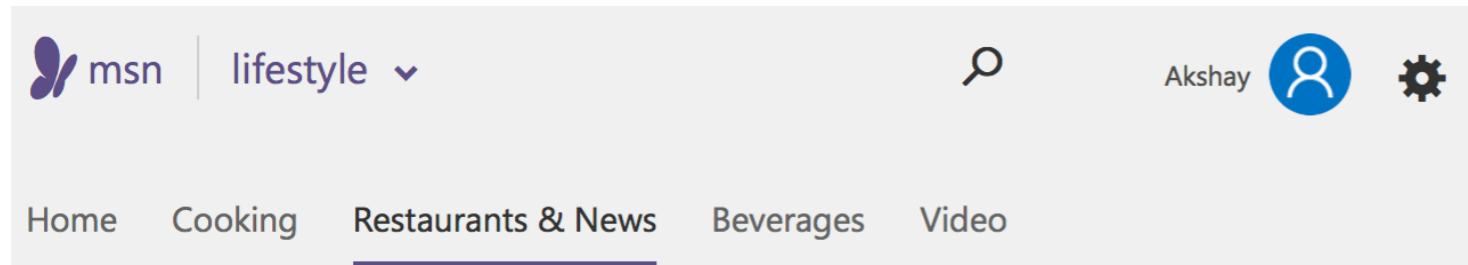
TODAY



Girl Scout Cookies Are Officially on Sale

Food & Wine

Online Personalization



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

Industry Standard: A/B Testing

The screenshot displays the MSN lifestyle website interface. At the top, the MSN logo and 'lifestyle' category are visible, along with a search icon, the user name 'Akshay', and a settings gear. Below the navigation bar, there are links for 'Home', 'Cooking', 'Restaurants & News', 'Beverages', and 'Video'. The main content area features a large photo of a group of people at a party, with a 'Food & Wine' tag and '8 hrs ago' timestamp. To the right is a sponsored advertisement for 'Say Goodbye to iPhone: This Could Be 40X Better' by 'The Motley Fool'. Below the party photo is an article snippet about 'Trader Joe's new bagged french fry chip mashup totally lives...' from 'TODAY'. To the right of this is another article snippet about 'Girl Scout Cookies Are Officially on Sale' from 'Food & Wine'.

msn | lifestyle

Akshay

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.)

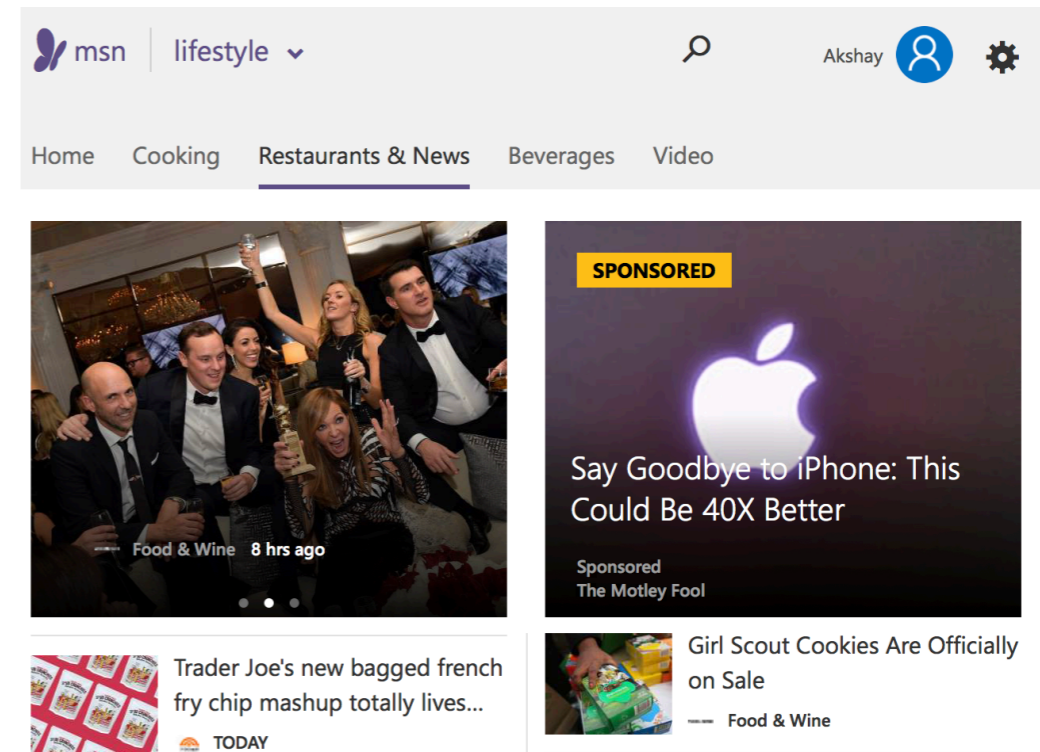
The screenshot shows the MSN lifestyle page with a navigation bar at the top. The main content area is divided into two columns. The left column features a large image of a group of people at a party, with the caption 'Food & Wine 8 hrs ago'. The right column features a sponsored article with a black background and a glowing Apple logo. The article title is 'Say Goodbye to iPhone: This Could Be 40X Better' and it is sponsored by 'The Motley Fool'. Below the sponsored article is another article titled 'Trader Joe's new bagged french fry chip mashup totally lives...' from 'TODAY'. At the bottom right, there is a small article titled 'Girl Scout Cookies Are Officially on Sale' from '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.)

Two main issues:



The screenshot shows the MSN lifestyle page interface. At the top, there is a navigation bar with the MSN logo, a search icon, and a user profile icon labeled 'Akshay'. Below the navigation bar, there are several article cards. The first card on the left is a photo of a group of people at a party, with the text 'Food & Wine 8 hrs ago'. To its right is a sponsored article with a black background and a white Apple logo, titled 'Say Goodbye to iPhone: This Could Be 40X Better' by 'Sponsored The Motley Fool'. Below these, there are two more article cards: one for 'Trader Joe's new bagged french fry chip mashup totally lives...' from 'TODAY', and another for 'Girl Scout Cookies Are Officially on Sale' from 'Food & Wine'.

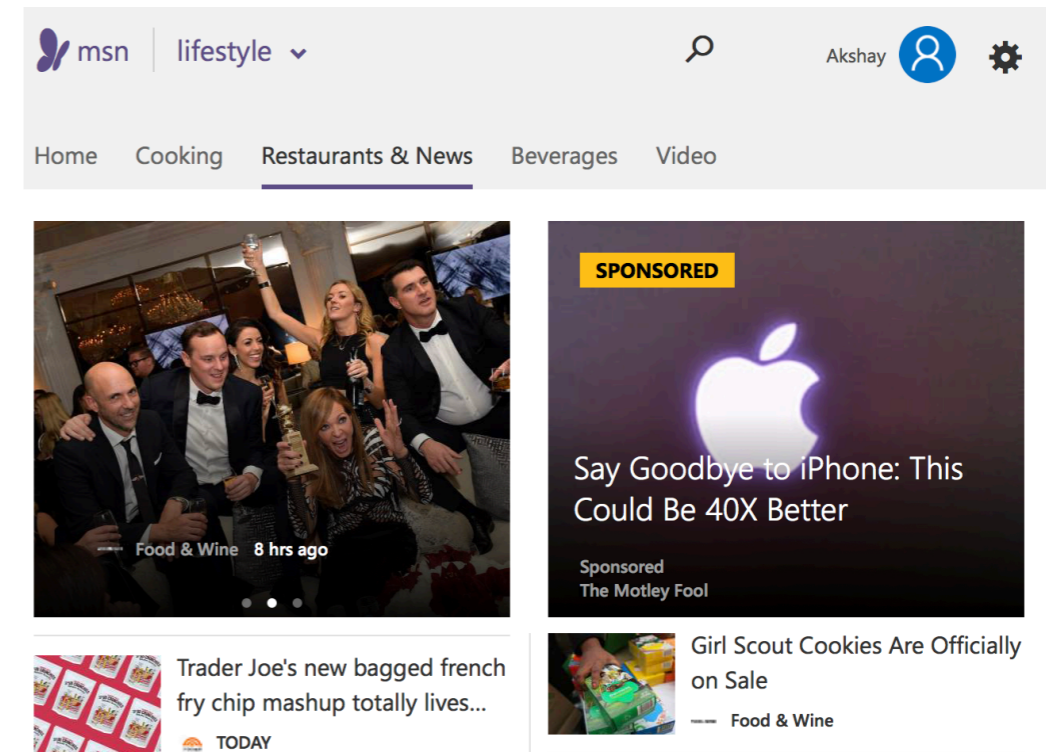
Industry Standard: A/B Testing

Given policy π :

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

Two main issues:

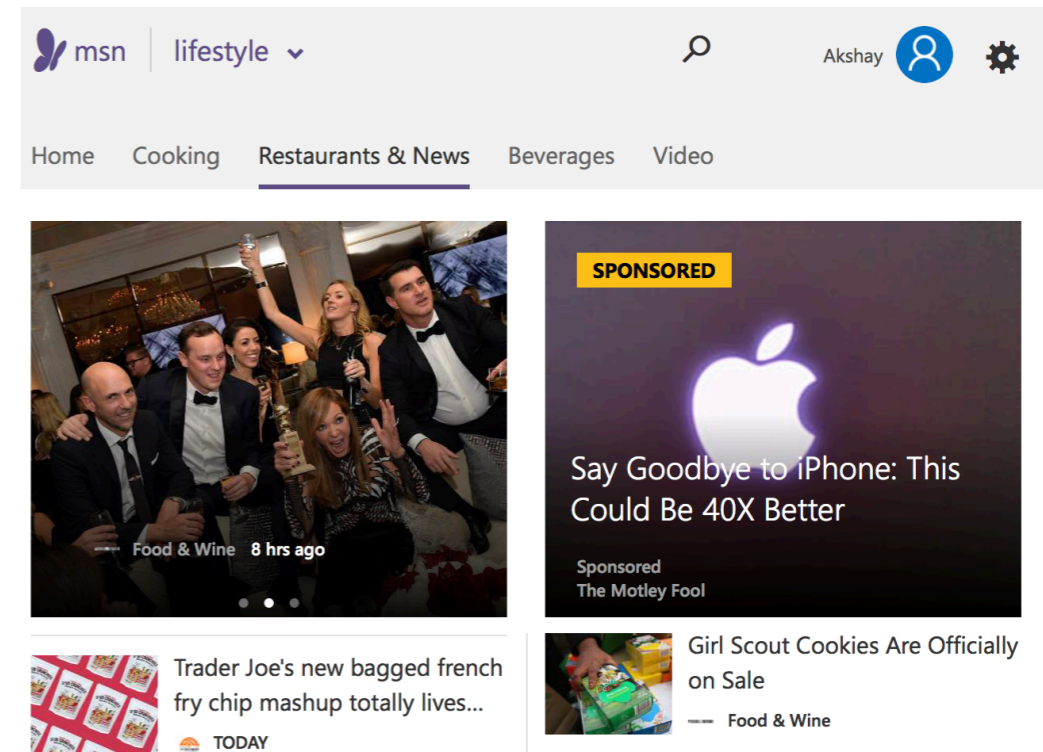
1. Poor performance while evaluating policies



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.)



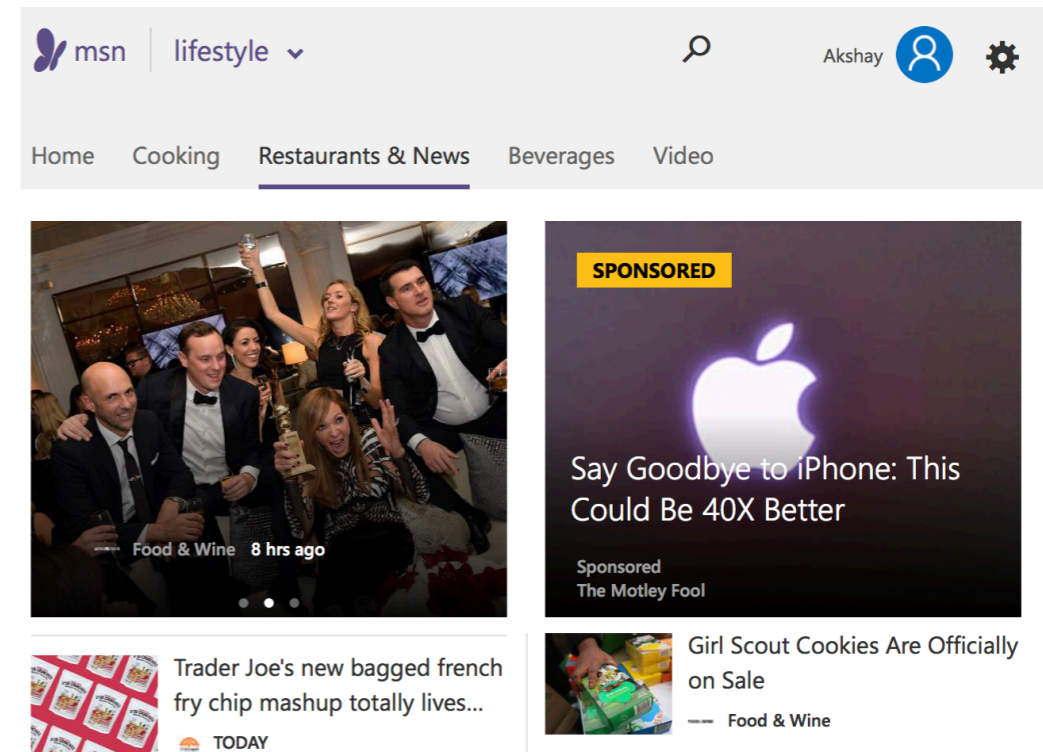
Two main issues:

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 displays the MSN lifestyle website interface. At the top, the MSN logo and 'lifestyle' category are visible, along with a search icon, the user name 'Akshay', and a settings gear. Below the navigation bar, there are links for 'Home', 'Cooking', 'Restaurants & News', 'Beverages', and 'Video'. The main content area features a carousel of articles. The first article in the carousel shows a group of people at a party, with the text 'Food & Wine 8 hrs ago'. To the right of the carousel is a sponsored advertisement for 'Say Goodbye to iPhone: This Could Be 40X Better' by 'The Motley Fool'. Below the carousel, there are two more article thumbnails: one for 'Trader Joe's new bagged french fry chip mashup totally lives...' from 'TODAY' and another for 'Girl Scout Cookies Are Officially on Sale' from 'Food & Wine'.

msn | lifestyle

Home Cooking Restaurants & News Beverages Video

Food & Wine 8 hrs ago

SPONSORED

Say Goodbye to iPhone: This Could Be 40X Better

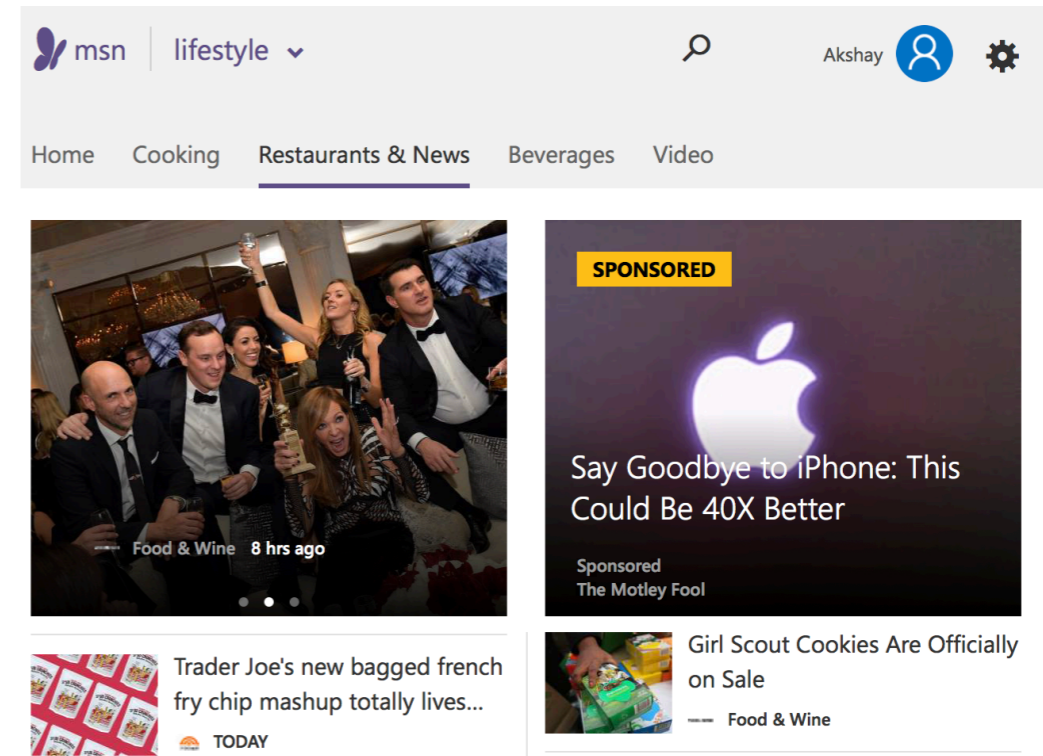
Sponsored
The Motley Fool

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

Girl Scout Cookies Are Officially on Sale
Food & Wine

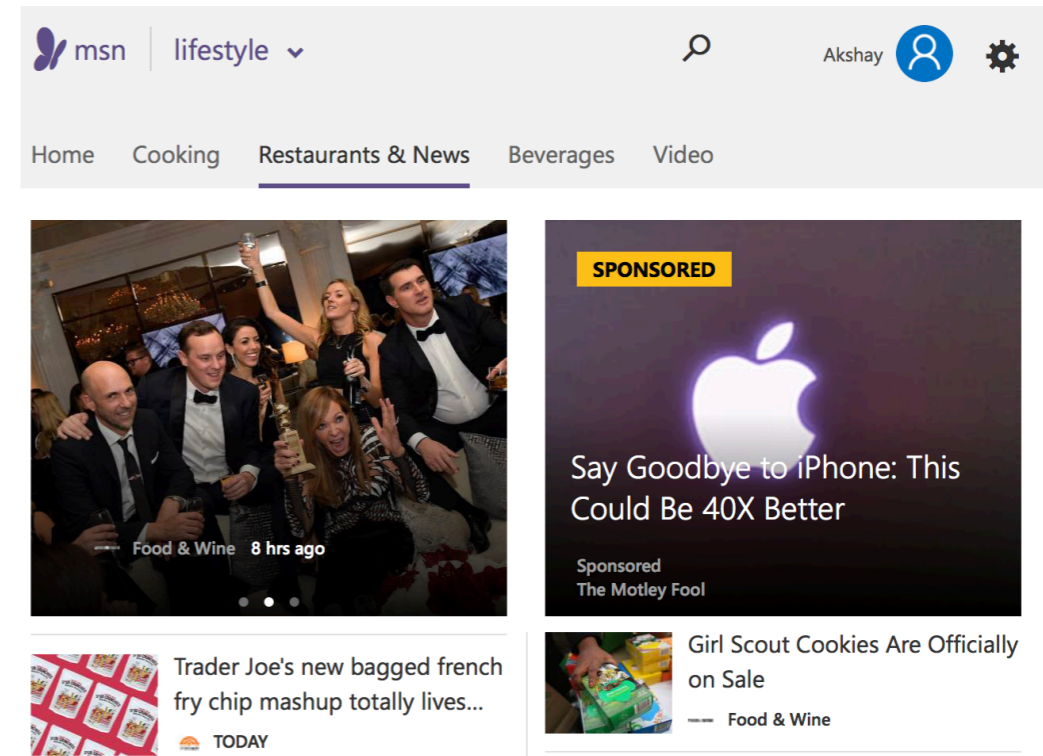
Exploration + Offline Evaluation

1. Collect dataset by serving content at random



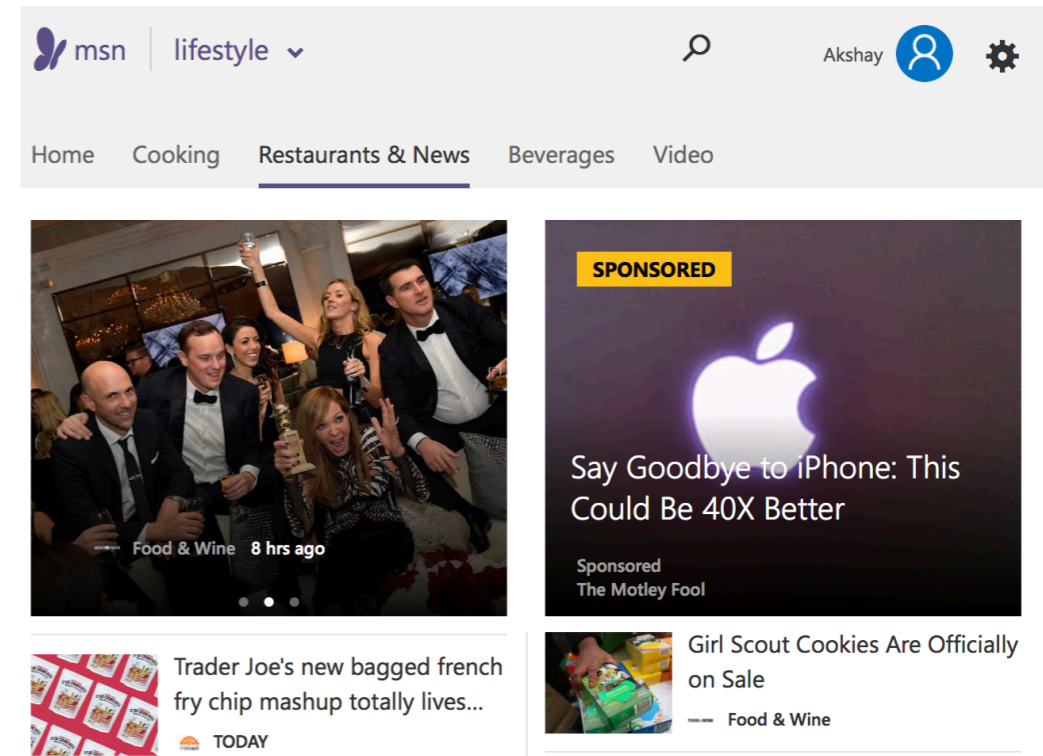
Exploration + Offline Evaluation

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 the MSN lifestyle page with a navigation bar at the top. The navigation bar includes the MSN logo, a dropdown menu for 'lifestyle', a search icon, the user name 'Akshay', and a settings icon. Below the navigation bar are tabs for 'Home', 'Cooking', 'Restaurants & News', 'Beverages', and 'Video'. The main content area features three items:

- A carousel image showing a group of people in formal attire at a party, with the text 'Food & Wine 8 hrs ago' at the bottom.
- A sponsored article with a black background and a glowing Apple logo, titled 'Say Goodbye to iPhone: This Could Be 40X Better', sponsored by 'The Motley Fool'.
- A sponsored article with a red and white checkered background, titled 'Trader Joe's new bagged french fry chip mashup totally lives...', sponsored by 'TODAY'.
- A sponsored article with a green and white background, titled 'Girl Scout Cookies Are Officially on Sale', sponsored by 'Food & Wine'.

Contextual Bandits

On each of T rounds:

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

The screenshot displays the MSN lifestyle website interface. At the top, there is a navigation bar with the MSN logo, a 'lifestyle' dropdown menu, a search icon, and a user profile for 'Akshay' with a settings gear icon. Below the navigation bar, there are tabs for 'Home', 'Cooking', 'Restaurants & News', 'Beverages', and 'Video'. The main content area is a grid of four items:

- Top Left:** A photograph of a group of people in formal attire at a party. Below the image, it says 'Food & Wine 8 hrs ago'.
- Top Right:** A sponsored advertisement with a black background and a glowing Apple logo. The text reads 'Say Goodbye to iPhone: This Could Be 40X Better'. It is labeled 'SPONSORED' in a yellow box and 'Sponsored The Motley Fool' at the bottom.
- Bottom Left:** An article snippet with a red and white patterned image. The text reads 'Trader Joe's new bagged french fry chip mashup totally lives...' and is dated 'TODAY'.
- Bottom Right:** An article snippet with a green and yellow image. The text reads 'Girl Scout Cookies Are Officially on Sale' and is dated 'Food & Wine'.

Contextual Bandits

On each of T rounds:

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

The screenshot shows the MSN lifestyle website interface. At the top, there is a navigation bar with the MSN logo, the word "lifestyle" with a dropdown arrow, a search icon, and a user profile section for "Akshay" with a profile icon and a settings gear icon. Below the navigation bar are menu items: "Home", "Cooking", "Restaurants & News" (which is underlined), "Beverages", and "Video".

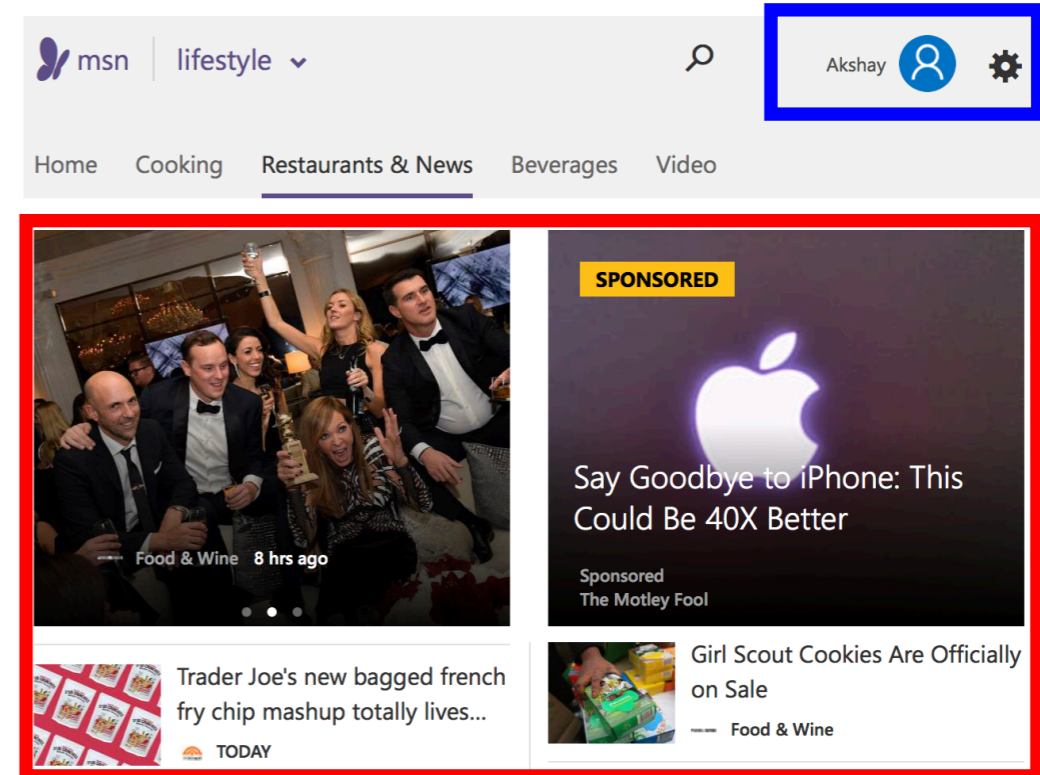
The main content area features a large featured article with a photo of a group of people at a party. The article is titled "Food & Wine" and is dated "8 hrs ago". To the right of this article is a sponsored advertisement with a yellow "SPONSORED" label at the top. The ad features a glowing Apple logo and the text "Say Goodbye to iPhone: This Could Be 40X Better". Below the ad, it says "Sponsored The Motley Fool".

Below the featured article, there is a news snippet with a small image of a bag of french fries and the text "Trader Joe's new bagged french fry chip mashup totally lives...". The source is "TODAY". To the right of this snippet is another news snippet with a small image of Girl Scout cookies and the text "Girl Scout Cookies Are Officially on Sale". The source is "Food & Wine".

Contextual Bandits

On each of T rounds:

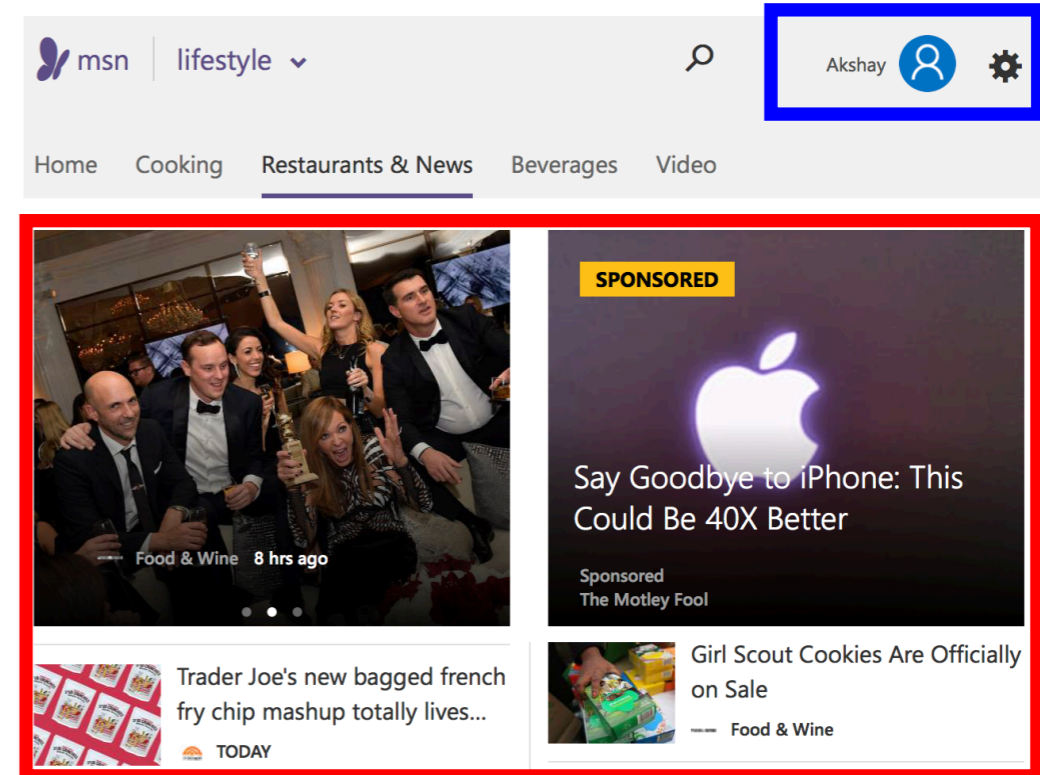
1. Observe context x_t
2. Play action a_t
3. Observe reward



Contextual Bandits

On each of T rounds:

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



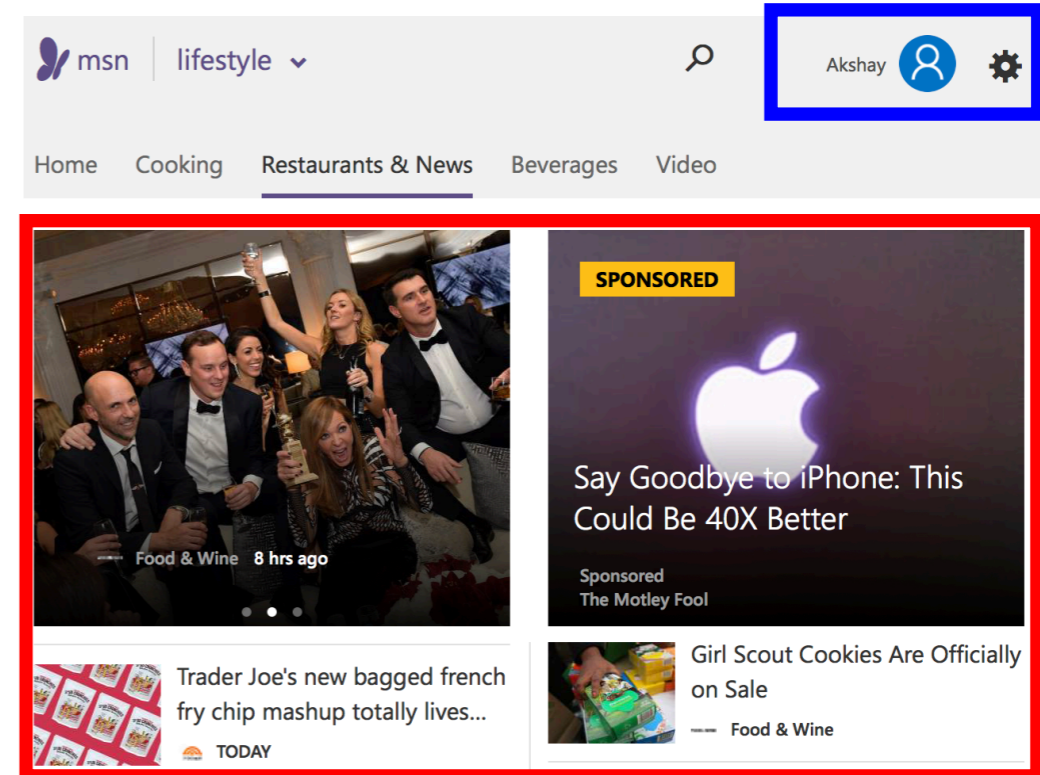
$$r_t = \# \text{ clicks}$$

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



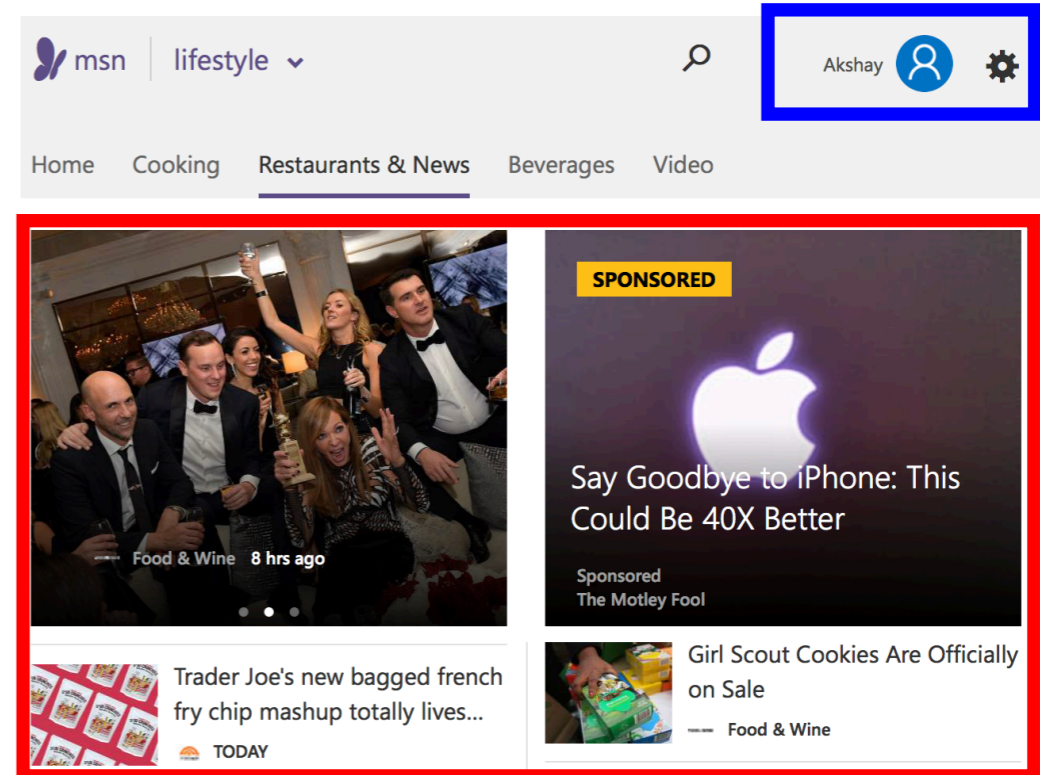
$$r_t = \# \text{ clicks}$$

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



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

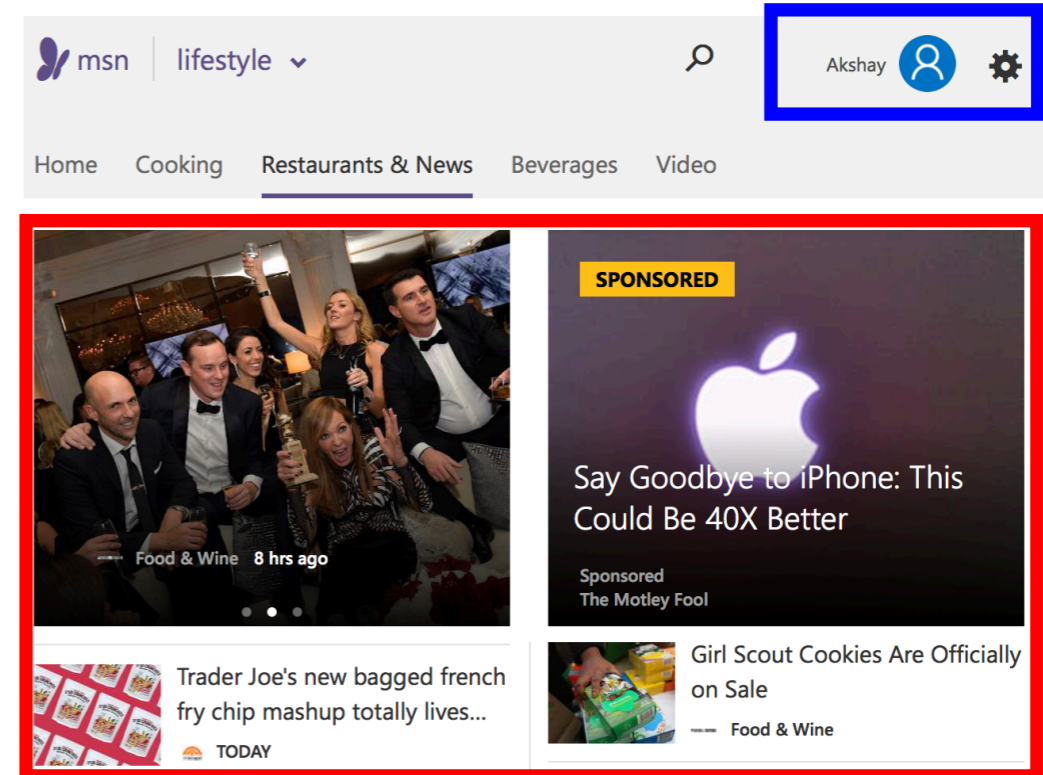
$$\text{Regret}(T, \Pi) = \max_{\pi \in \Pi} \text{Reward}(T, \pi) - \text{LearnerReward}(T)$$

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



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

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

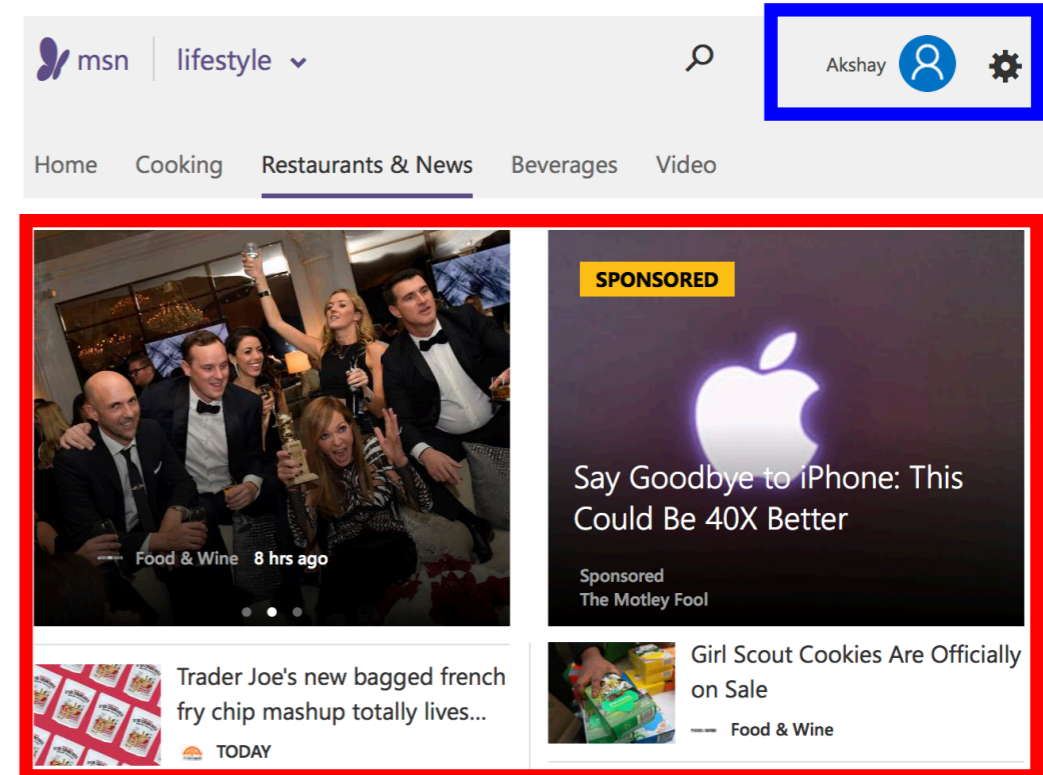
Fact: Can get $\sqrt{KT \log |\Pi|}$ regret.

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



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

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

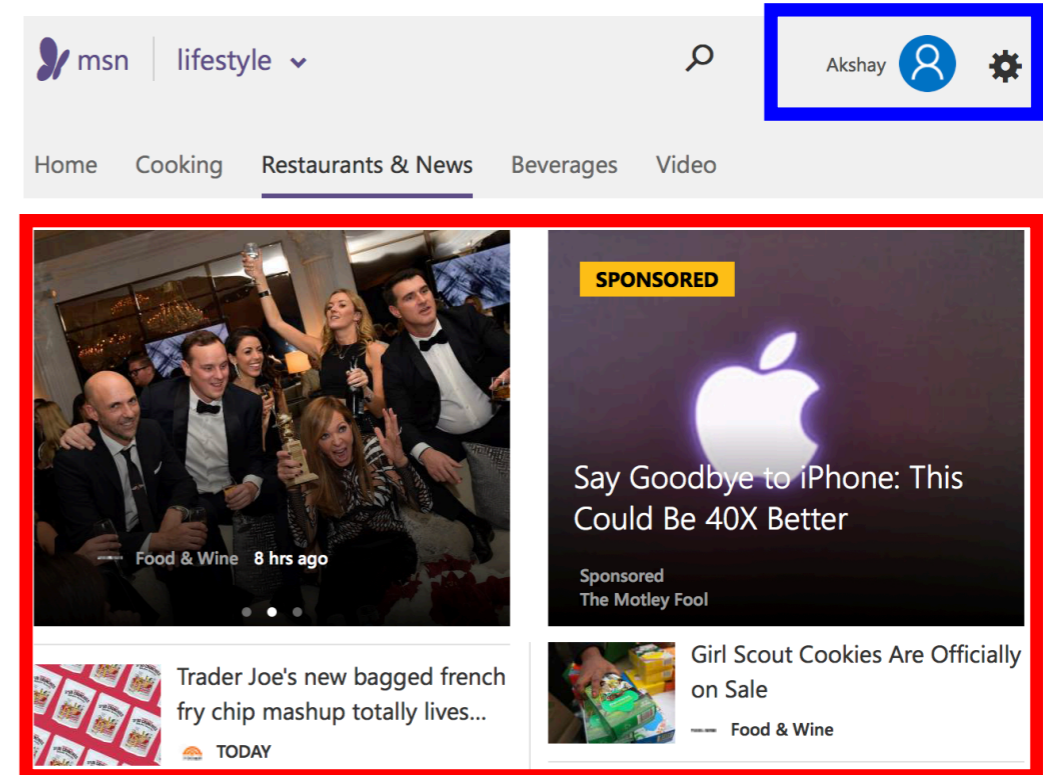
Fact: Can get $\sqrt{KT \log |\Pi|}$ regret. A/B testing gets $(|\Pi|)^{1/3} T^{2/3}$. Offline Eval gets $(K \log |\Pi|)^{1/3} T^{2/3}$

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



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

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

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

Exponential with combinatorial action space!

Contextual Semibandits

The screenshot displays the MSN lifestyle website interface. At the top, the MSN logo and 'lifestyle' category are visible, along with a search icon, the user name 'Akshay', and a settings gear. Below the navigation bar, there are four tabs: 'Home', 'Cooking', 'Restaurants & News' (which is underlined), 'Beverages', and 'Video'. The main content area features several articles and sponsored ads. On the left, there is a large photo of a group of people at a party, with the text 'Food & Wine 8 hrs ago' overlaid. Below this is a smaller article titled 'Trader Joe's new bagged french fry chip mashup totally lives...' from 'TODAY'. On the right, there is a large sponsored advertisement for an Apple product, titled 'Say Goodbye to iPhone: This Could Be 40X Better', sponsored by 'The Motley Fool'. Below the sponsored ad is another article titled 'Girl Scout Cookies Are Officially on Sale' from 'Food & Wine'.

msn | lifestyle

Akshay

Home Cooking Restaurants & News Beverages Video

Food & Wine 8 hrs ago

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

SPONSORED

Say Goodbye to iPhone: This Could Be 40X Better

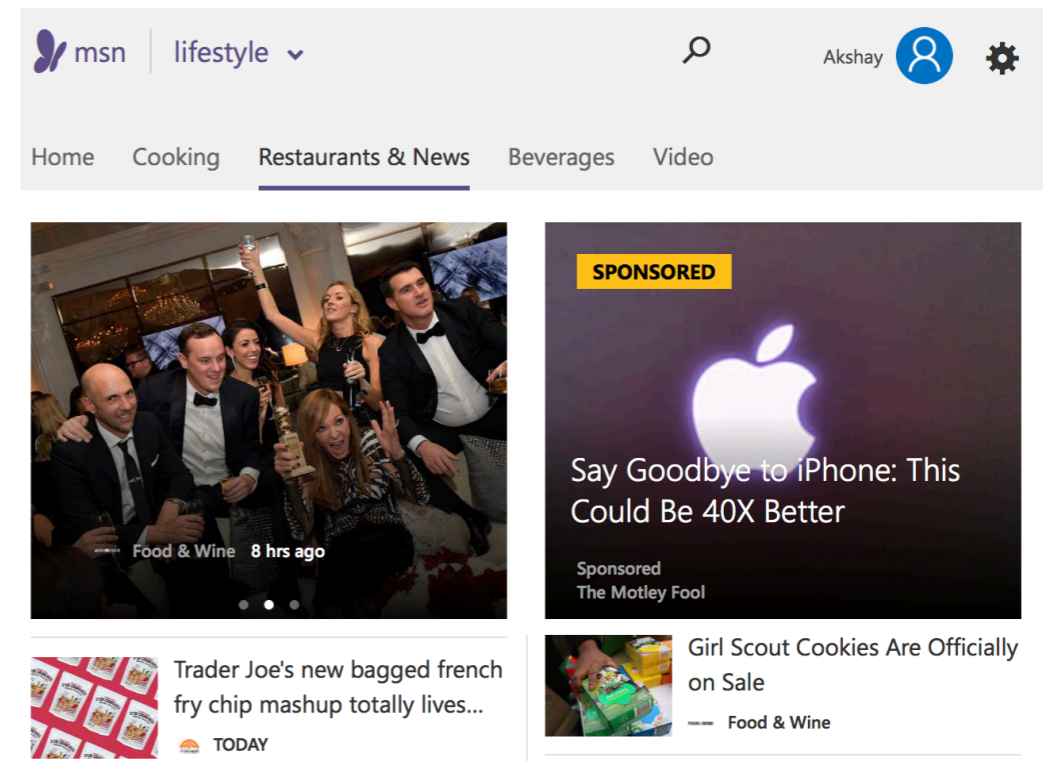
Sponsored The Motley Fool

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



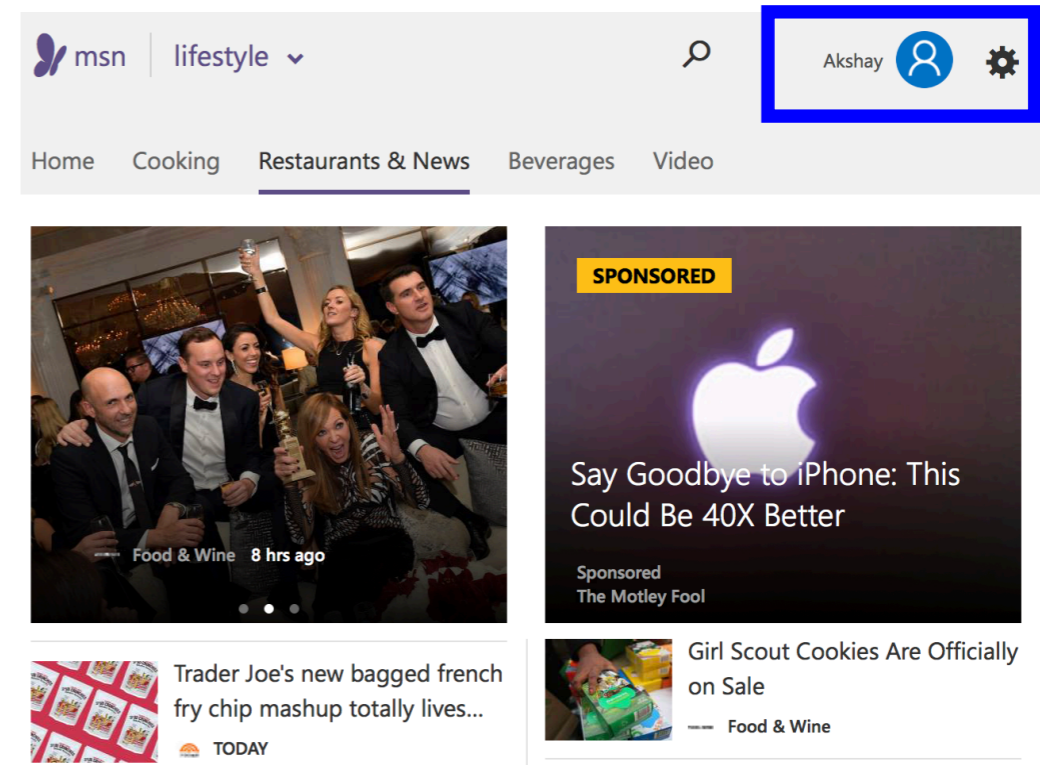
The screenshot shows the MSN lifestyle website interface. At the top, there is a navigation bar with the MSN logo, the word "lifestyle" with a dropdown arrow, a search icon, and a user profile icon labeled "Akshay" with a settings gear. Below the navigation bar are tabs for "Home", "Cooking", "Restaurants & News" (which is underlined), "Beverages", and "Video". The main content area is a grid of four items:

- Top Left:** A photograph of a group of people in formal attire at a party. Below the photo, it says "Food & Wine" and "8 hrs ago".
- Top Right:** A sponsored advertisement with a yellow "SPONSORED" label at the top. It features a glowing Apple logo and the text "Say Goodbye to iPhone: This Could Be 40X Better". Below the text, it says "Sponsored The Motley Fool".
- Bottom Left:** A small image of a bag of french fries. To its right, the text reads "Trader Joe's new bagged french fry chip mashup totally lives..." and "TODAY".
- Bottom Right:** A small image of a box of Girl Scout cookies. To its right, the text reads "Girl Scout Cookies Are Officially on Sale" and "Food & Wine".

Contextual Semibandits

On each of T rounds:

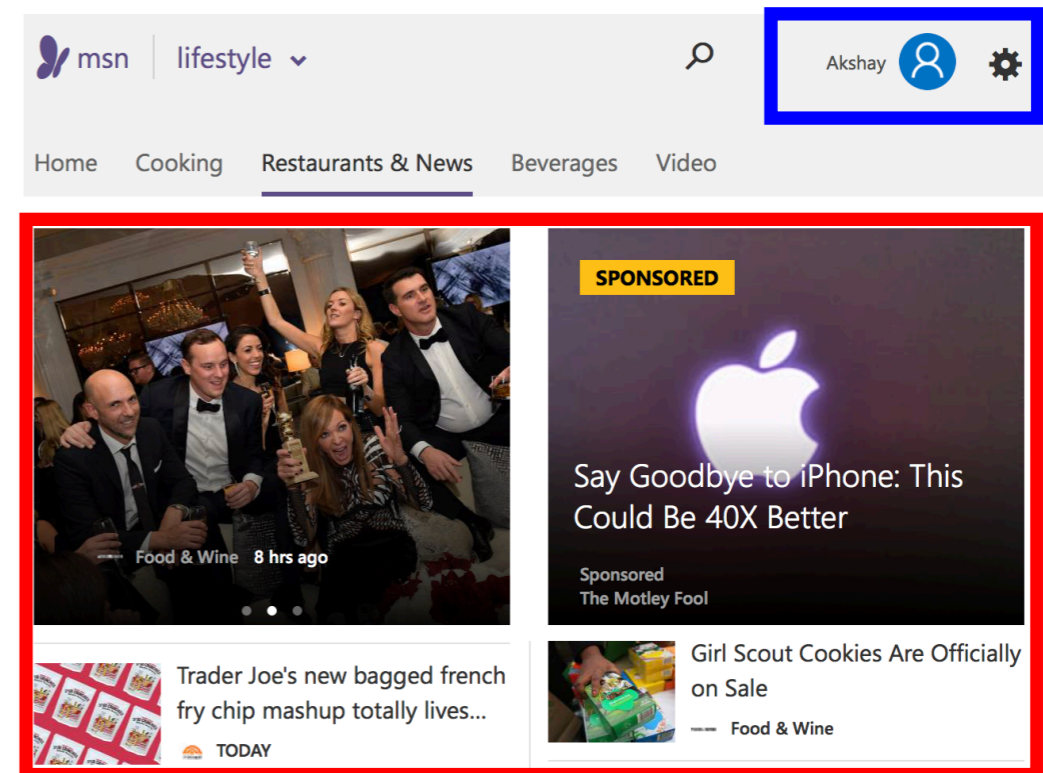
1. Observe context x_t
2. Play action
3. Observe features
4. Observe reward



Contextual Semibandits

On each of T rounds:

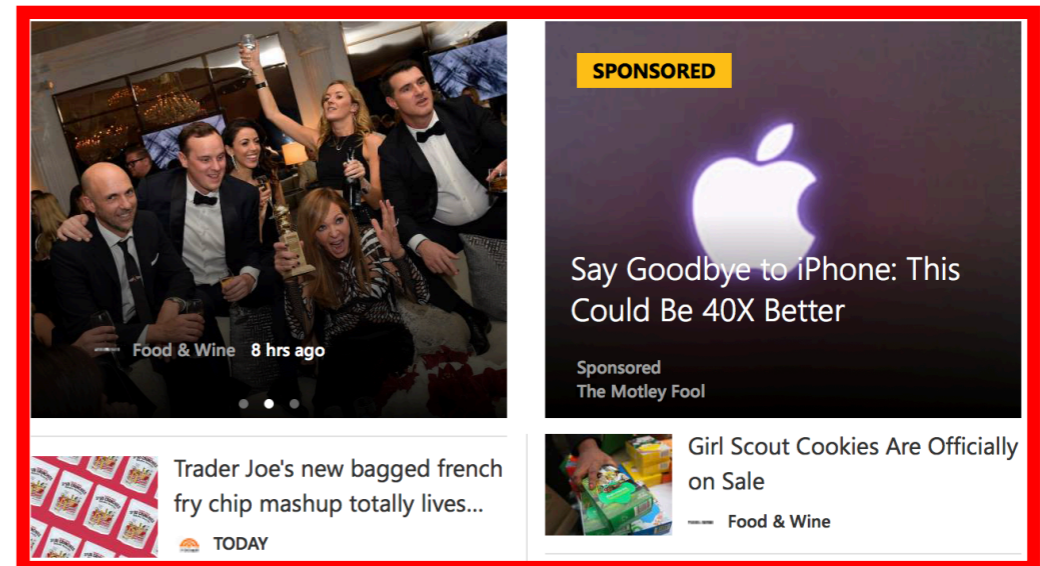
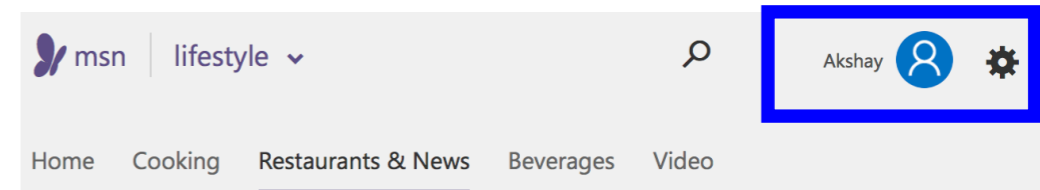
1. Observe context x_t
2. Play action $A_t = (a_1, \dots, a_L)$
3. Observe features
4. Observe reward



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



click

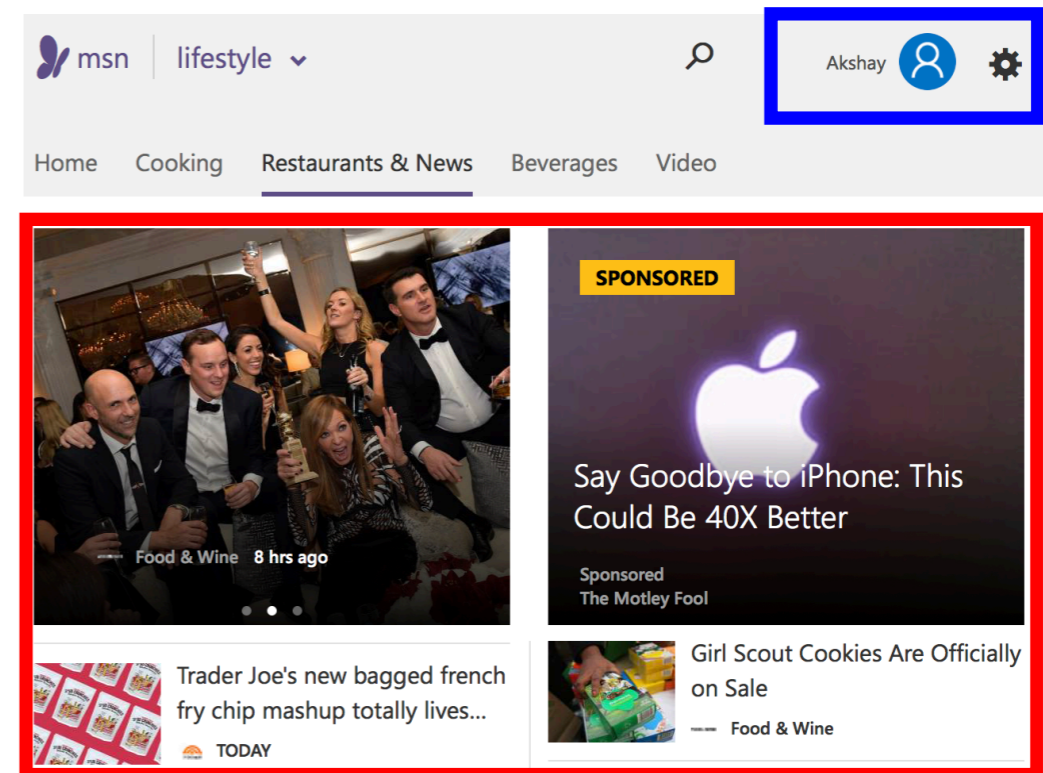
click

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

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



click

click

Contextual Semibandits

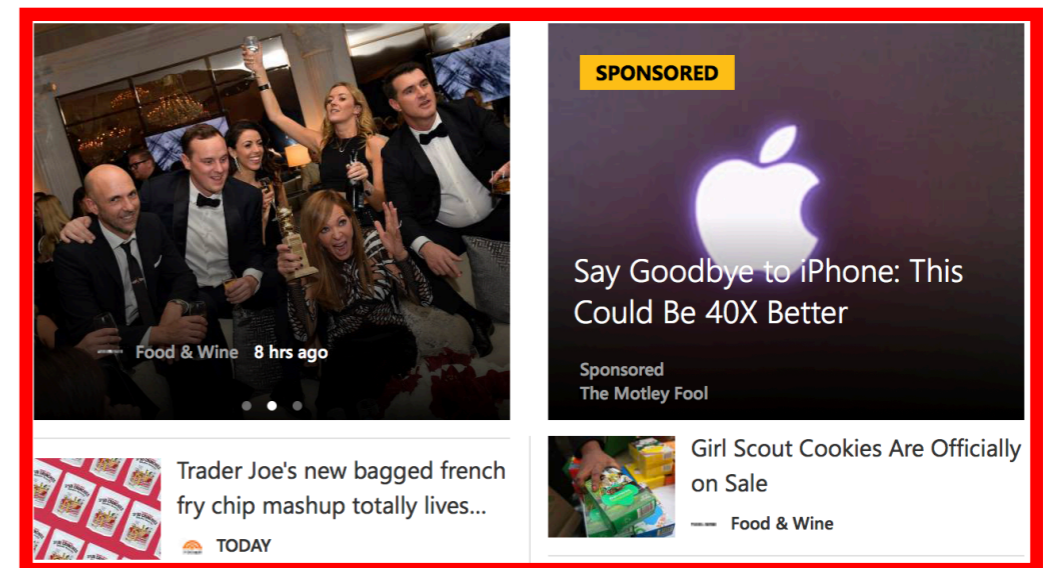
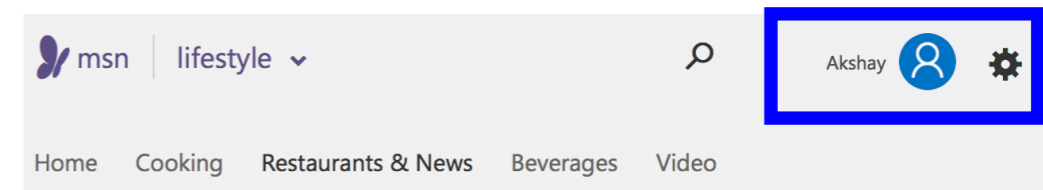
On each of T rounds:

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

click

Contextual Semibandits

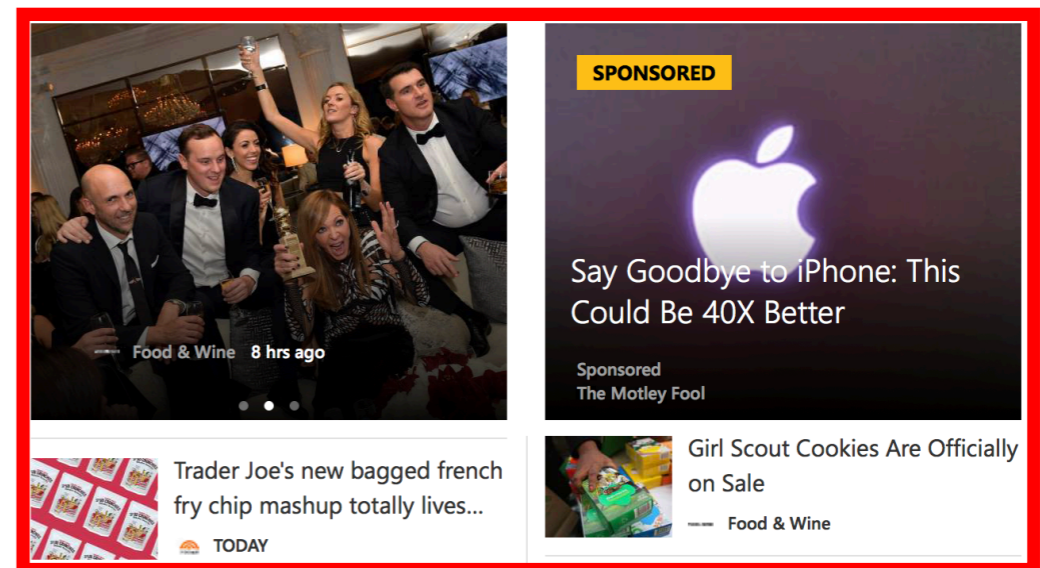
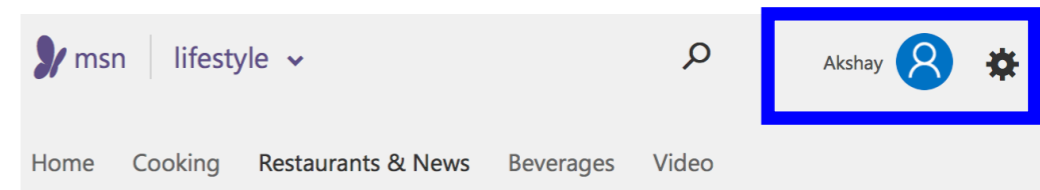
On each of T rounds:

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

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

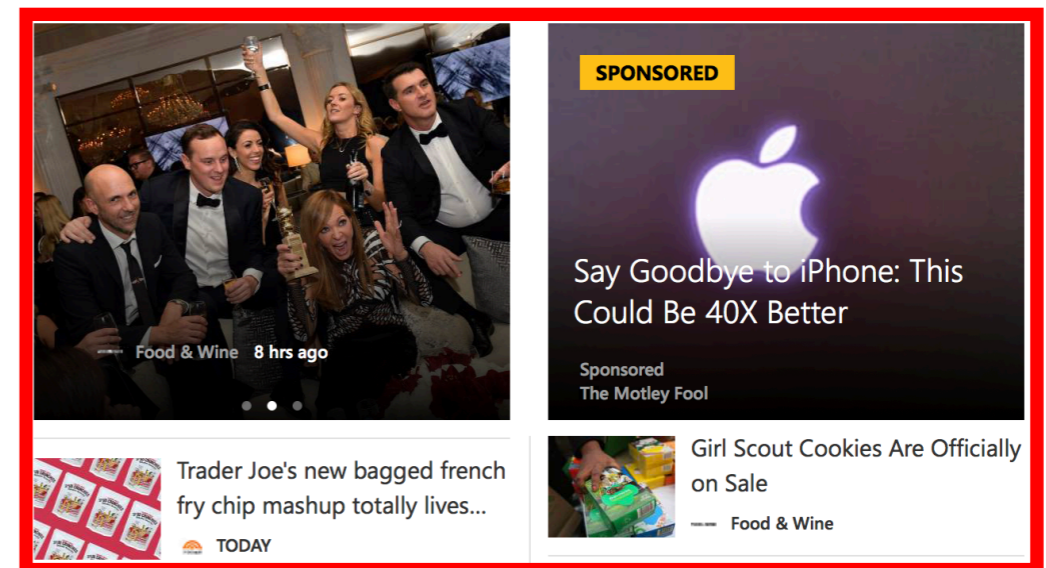
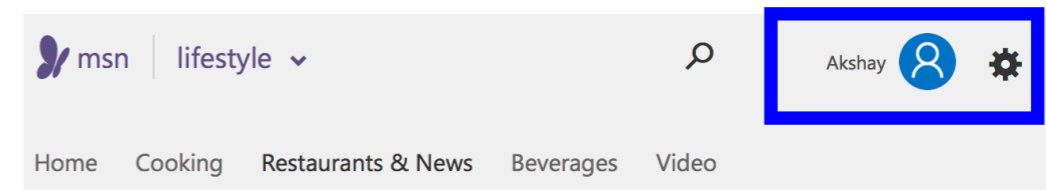
$$r_t(A_t, x_t) = \sum_{\ell} y(a_\ell) + \text{noise}$$

B = number of simple actions

L = composite action length

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

Challenges:



click

click

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

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

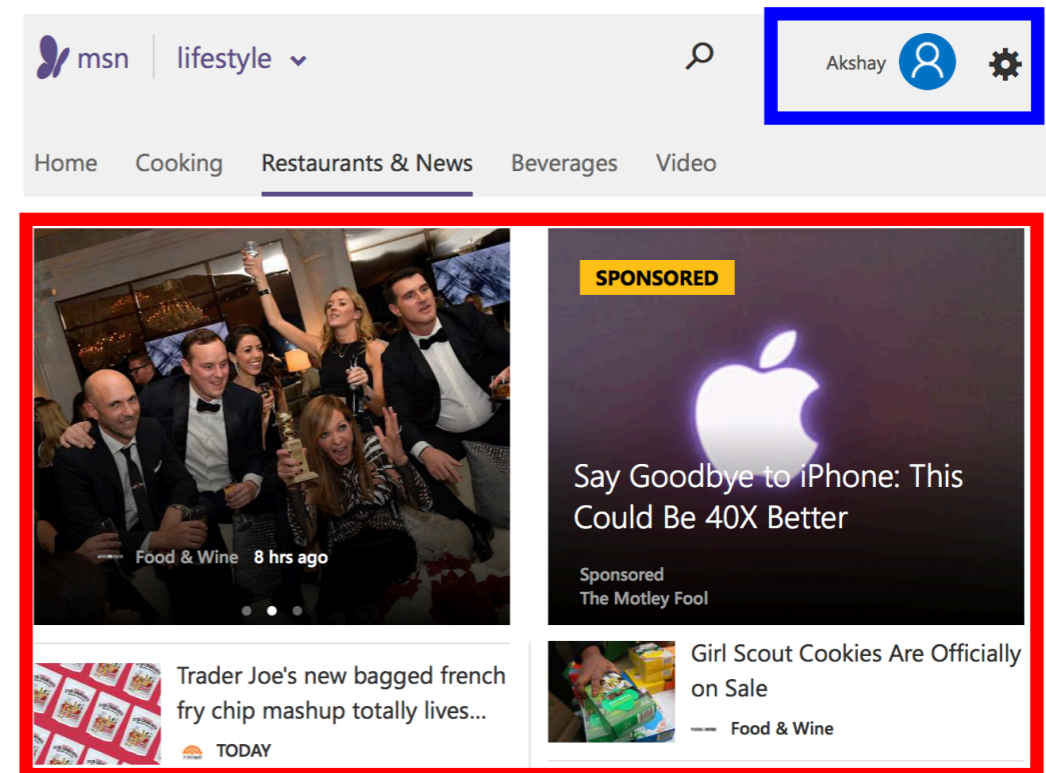
B = number of simple actions

L = composite action length

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

Challenges:

- Off-policy evaluation?



click

click

Contextual Semibandits

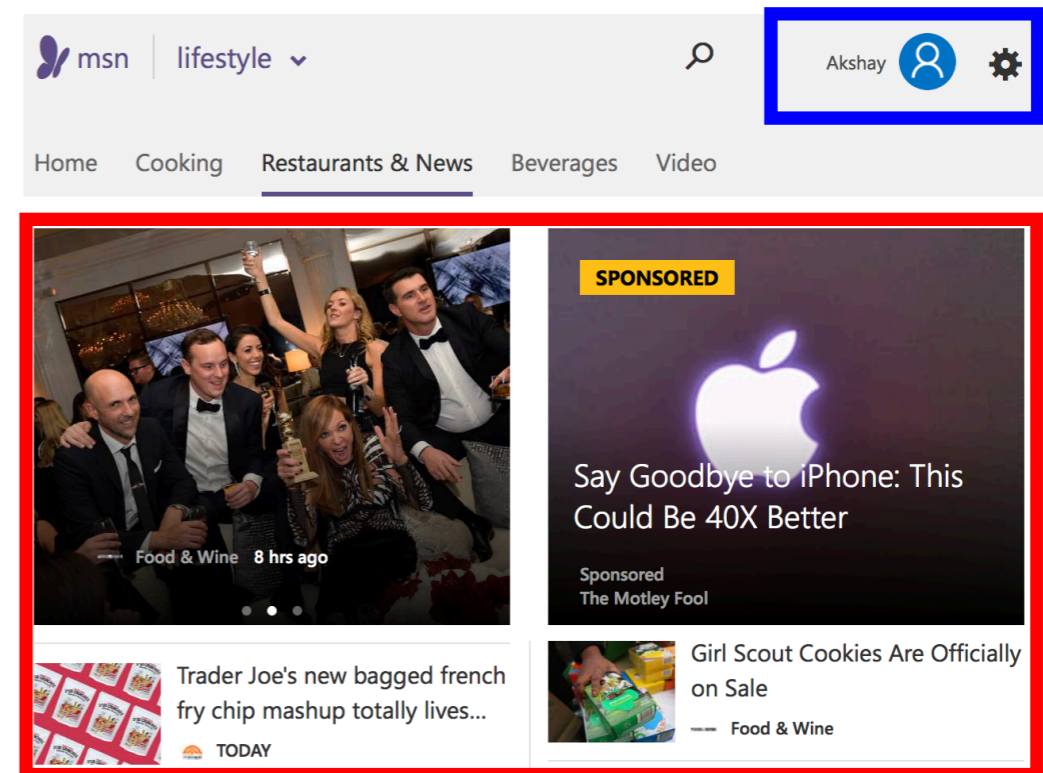
On each of T rounds:

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

click

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

Challenges:

- Off-policy evaluation?
- Explore vs Exploit?

Contextual Semibandits

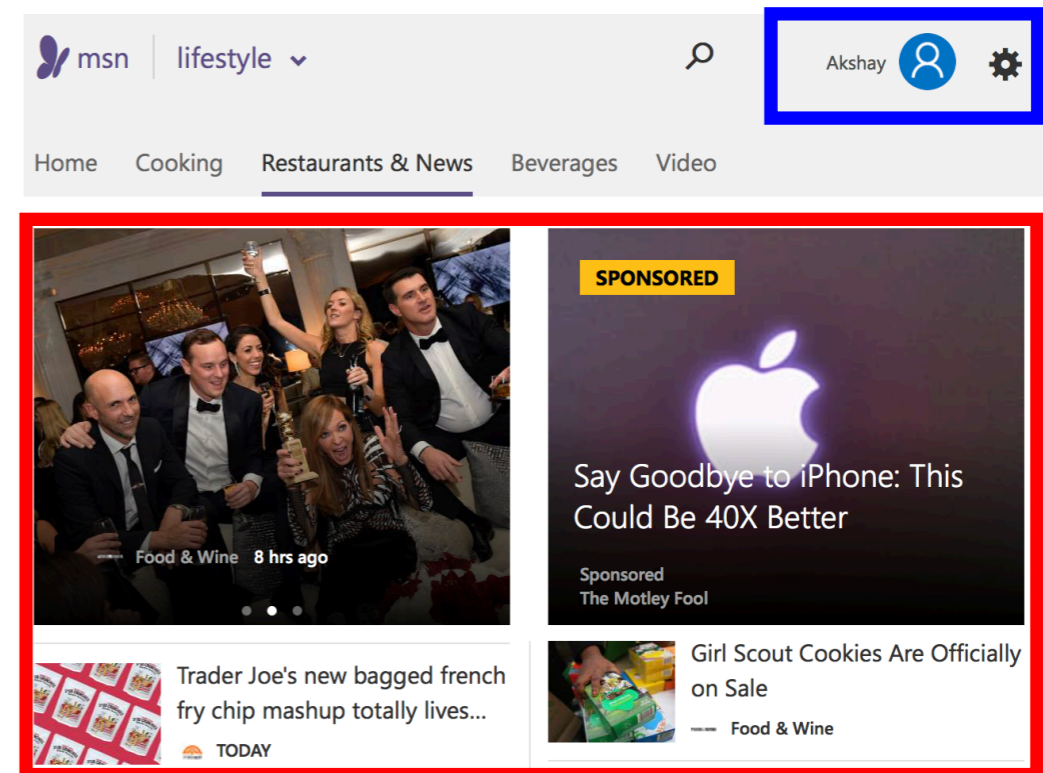
On each of T rounds:

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



click

click

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

Challenges:

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

Results

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

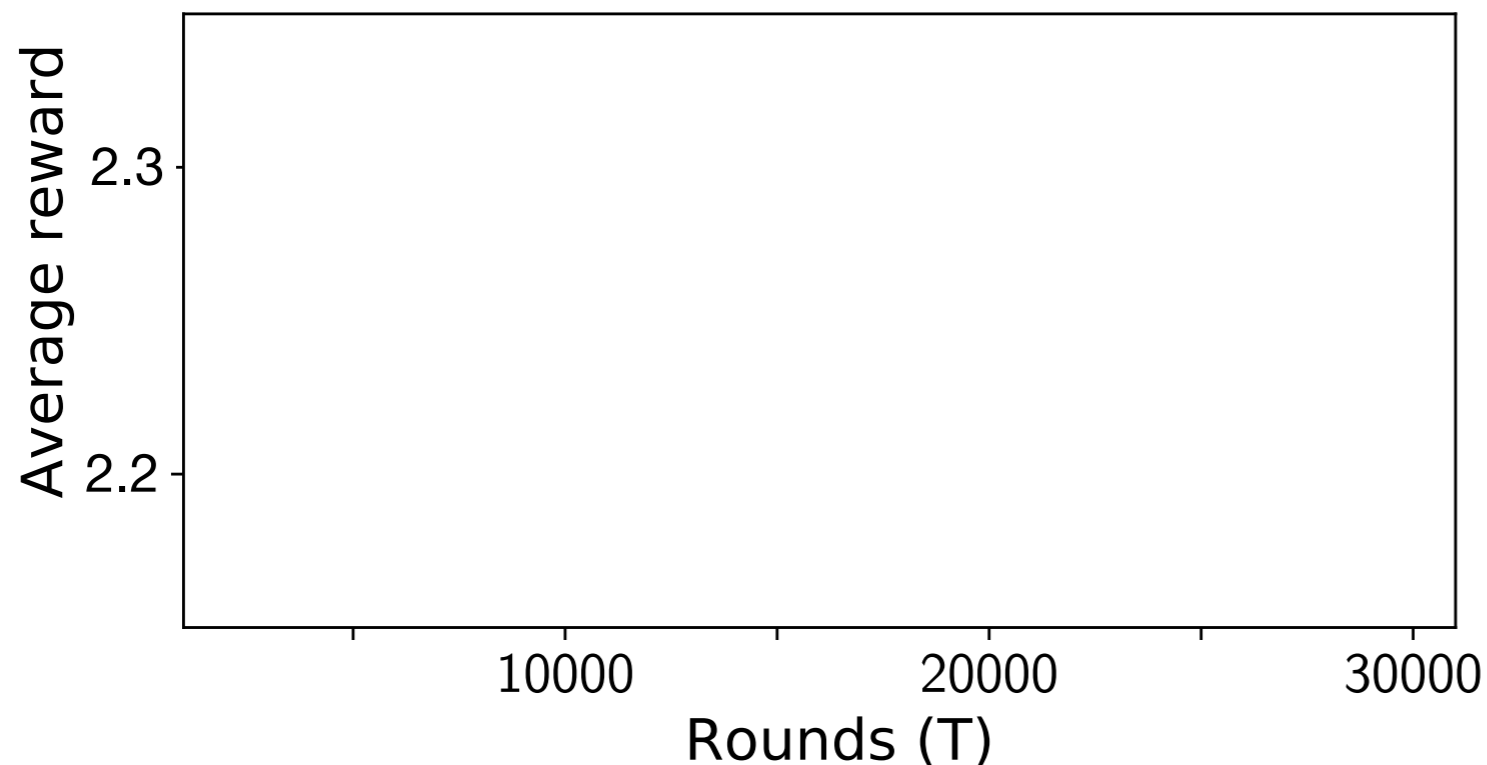
- Exponentially better than $\sqrt{B^L T \log(|\Pi|)}$ for naive contextual bandits
- Computationally efficient with rich policy classes

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



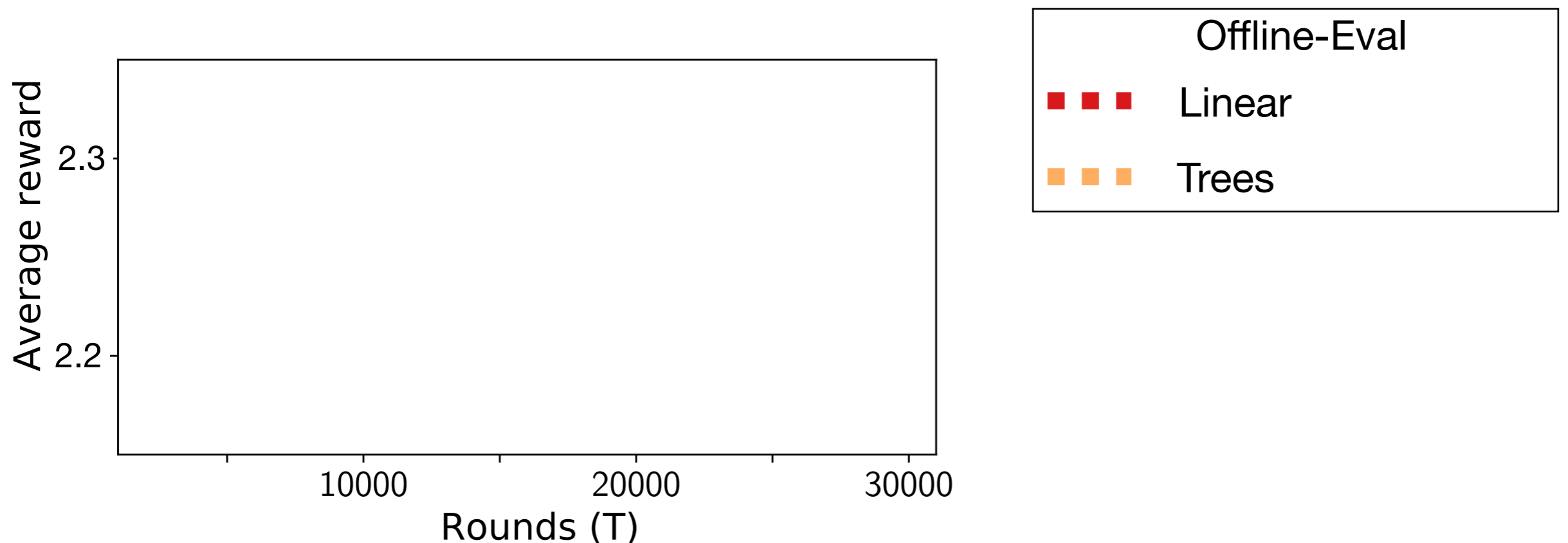
[Krishnamurthy, Agarwal, Dudik. NeurIPS 2016]

Results

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

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

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

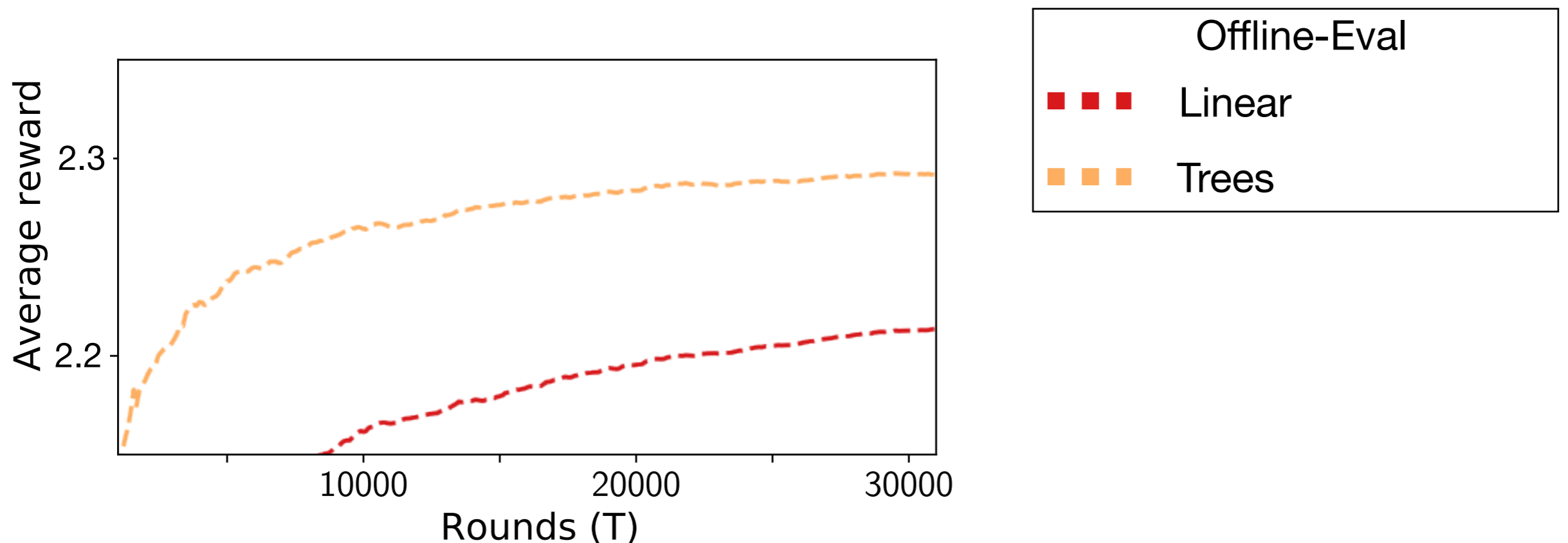


Results

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



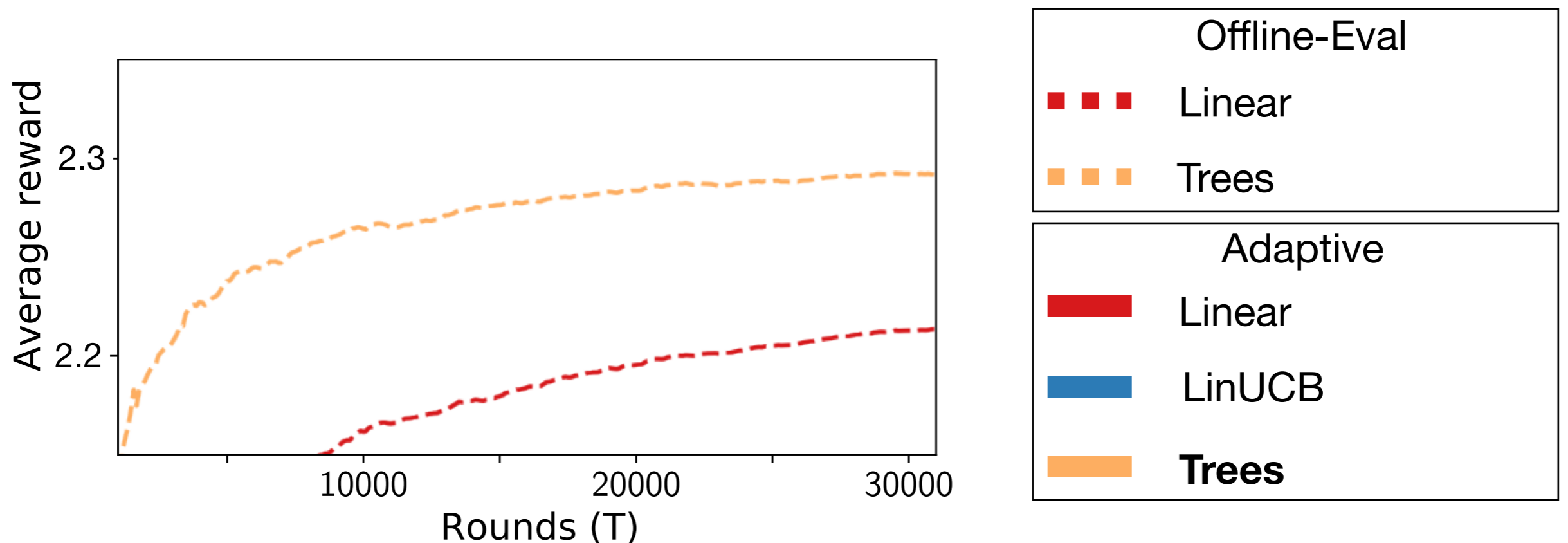
[Krishnamurthy, Agarwal, Dudik. NeurIPS 2016]

Results

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

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

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

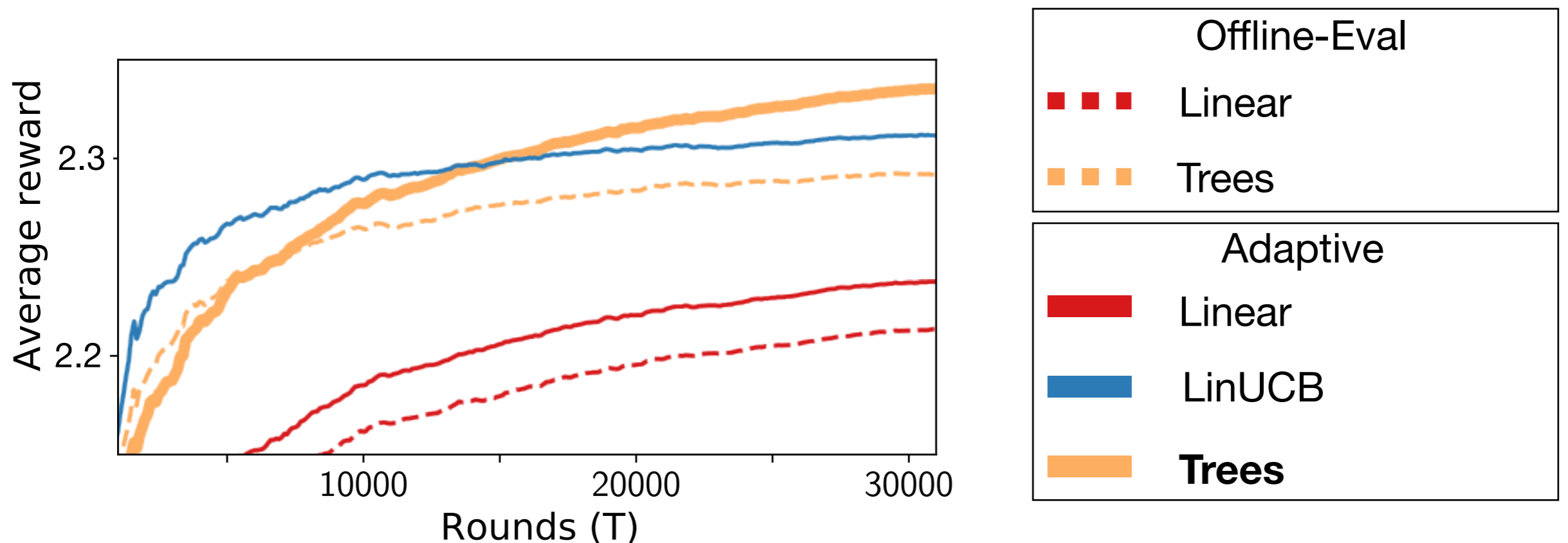


Results

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

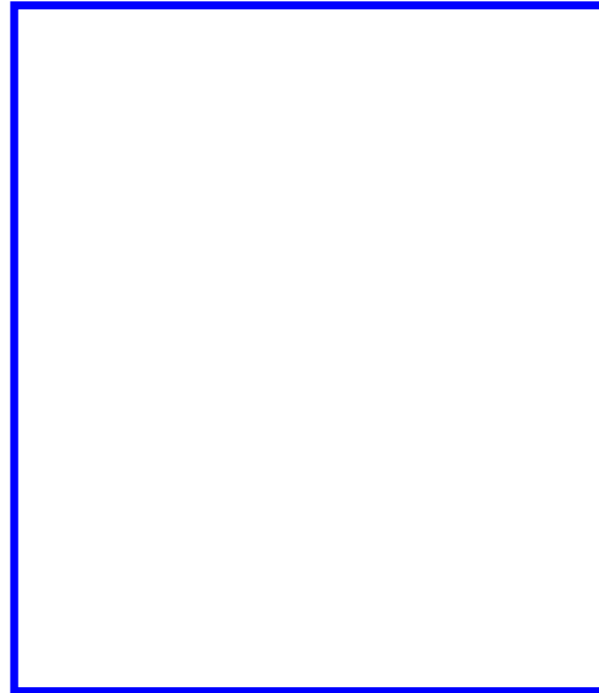
Techniques — Off-policy evaluation

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

Logging



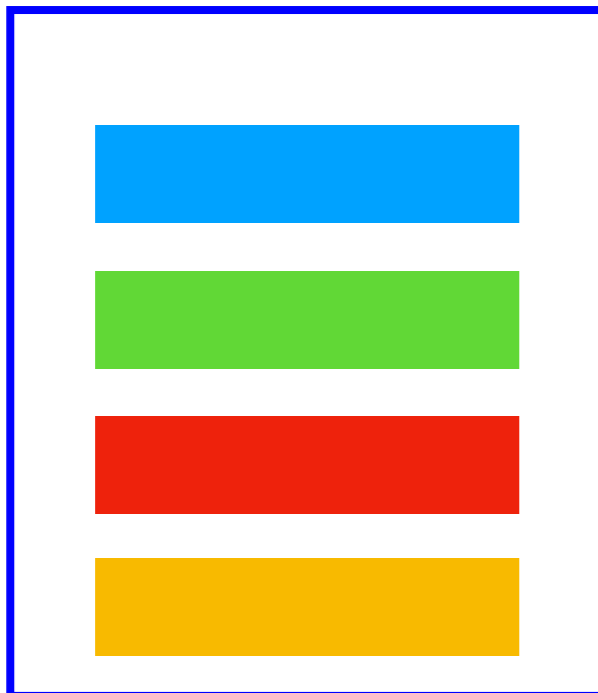
Target



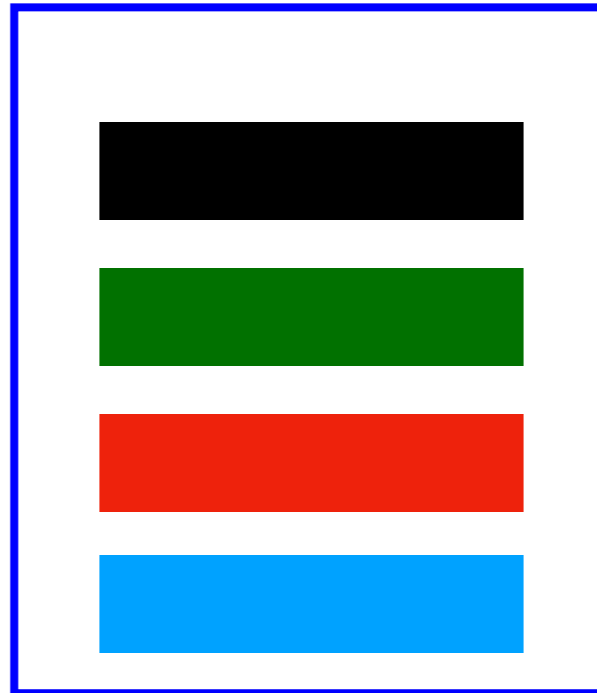
Techniques – Off-policy evaluation

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

Logging

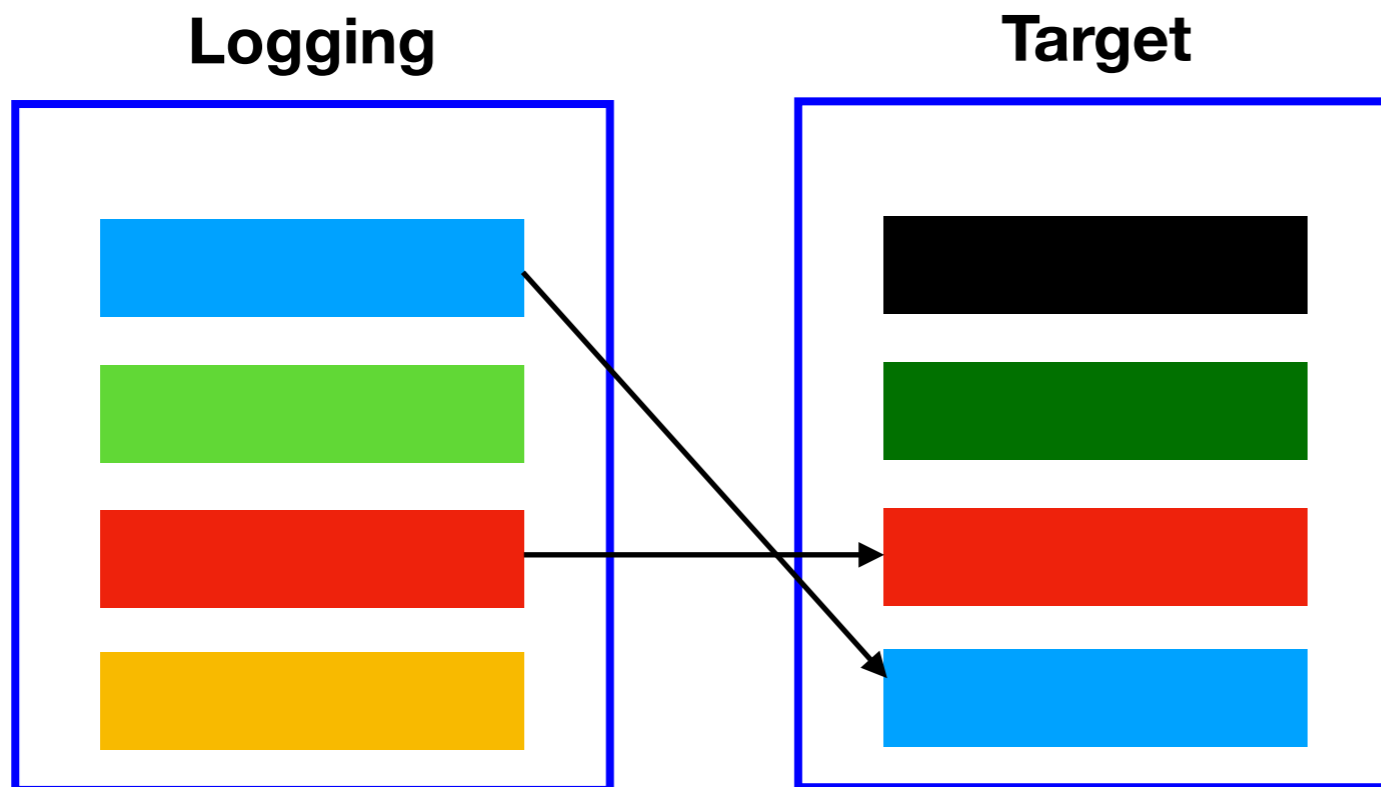


Target



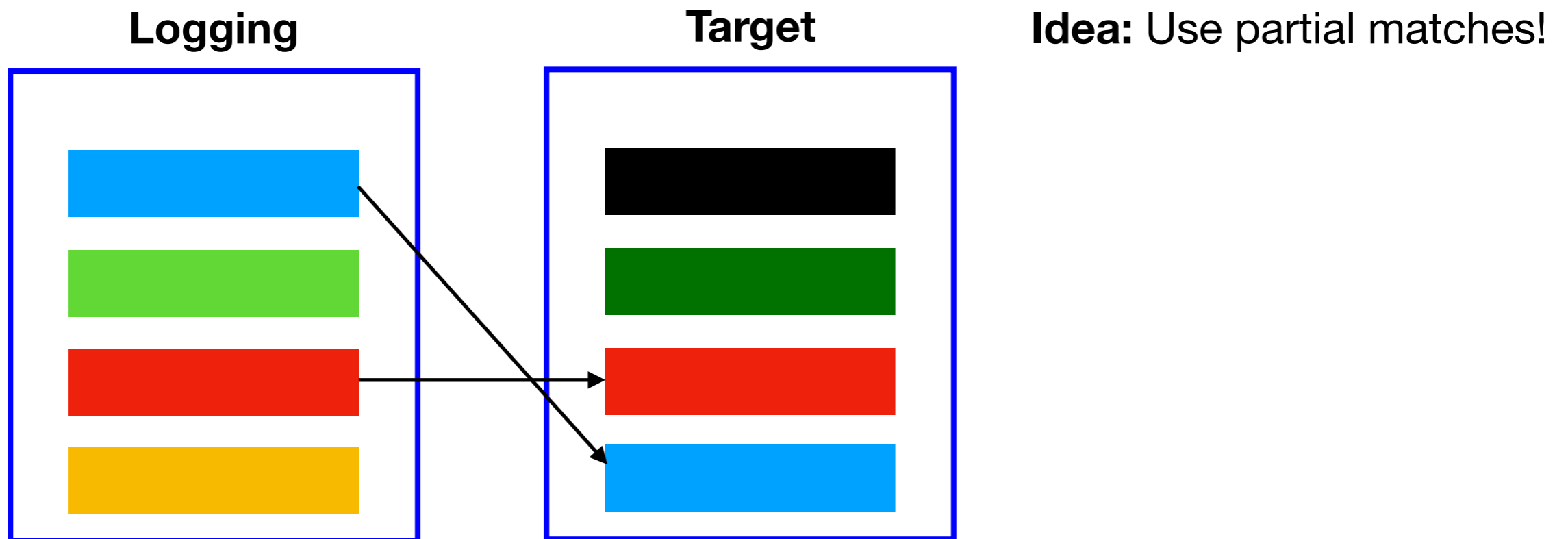
Techniques – Off-policy evaluation

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



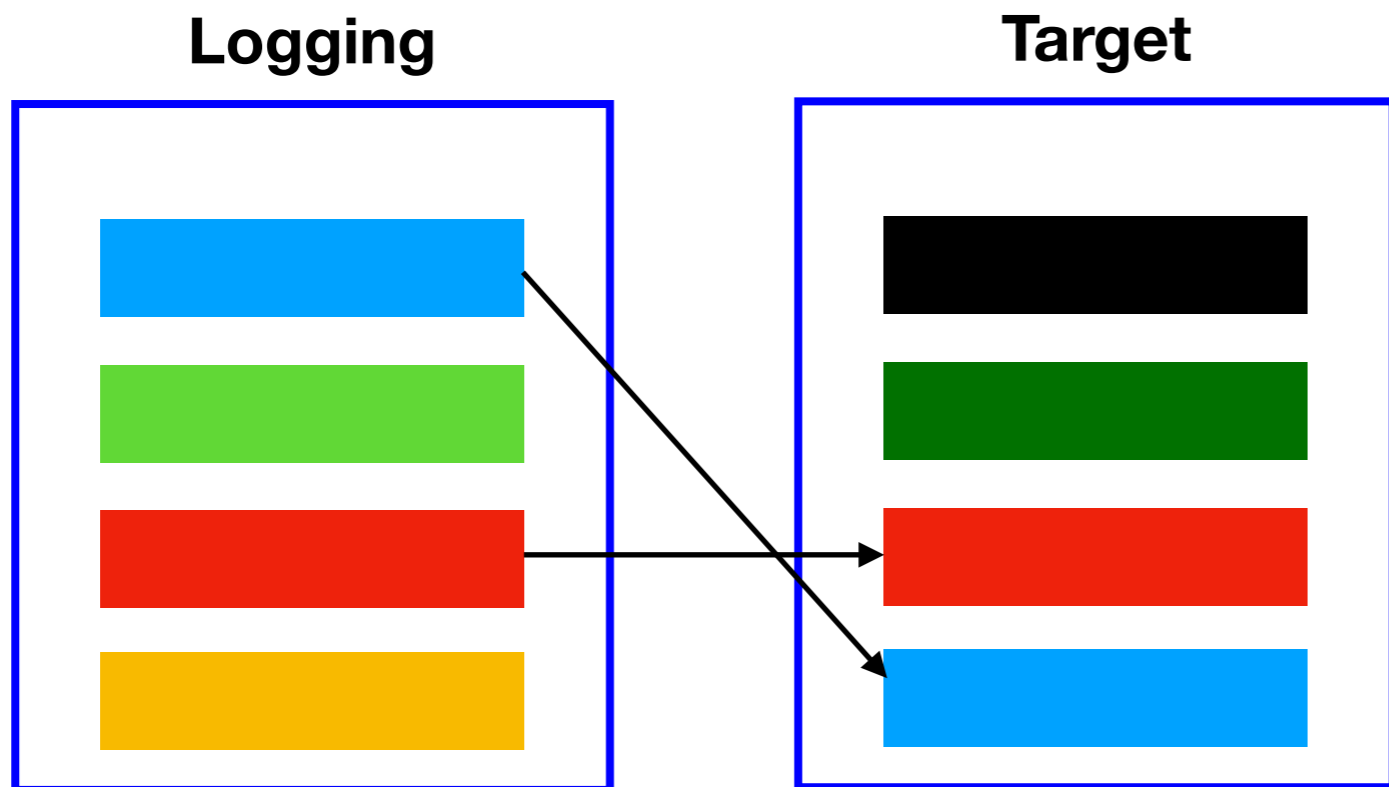
Techniques – Off-policy evaluation

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!

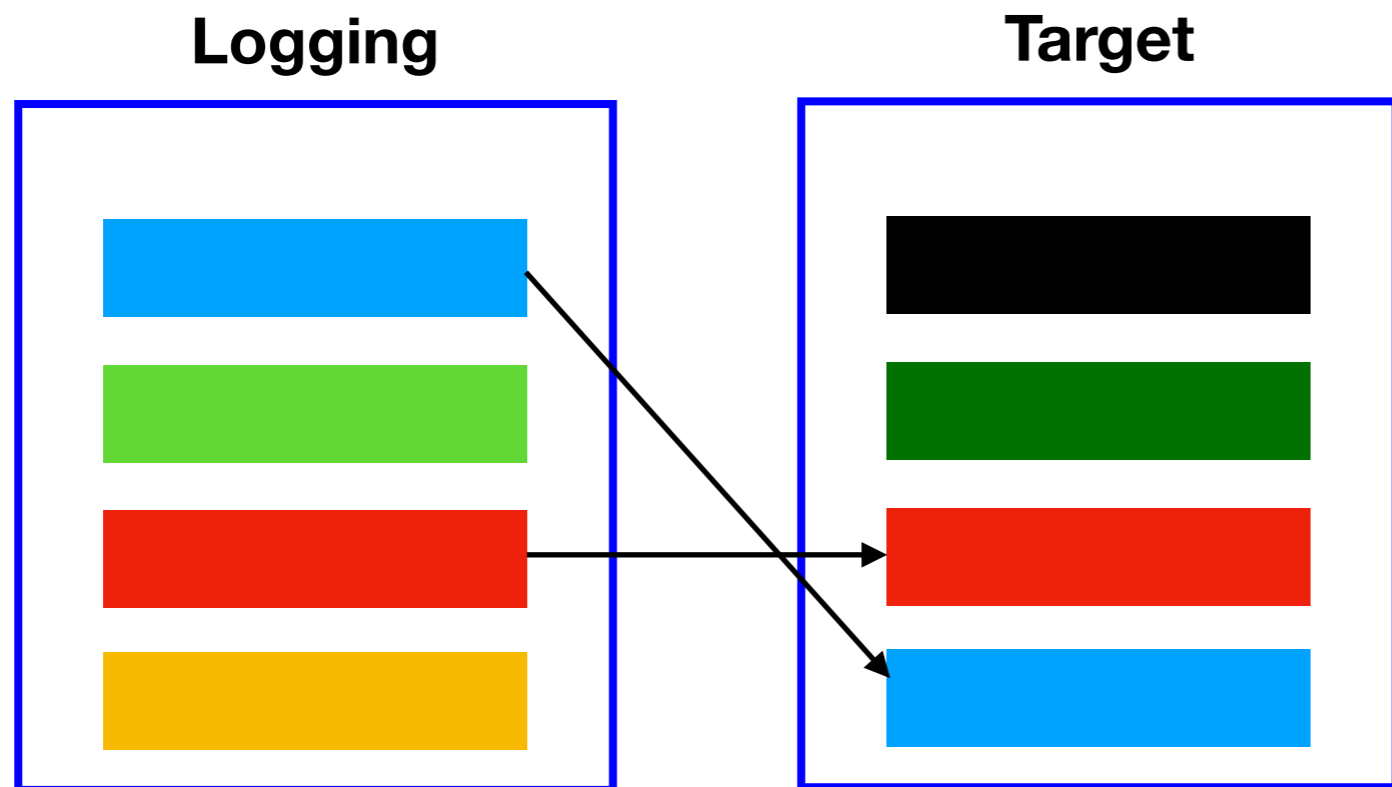
If $A \sim Q(\cdot|x)$

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

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

Techniques – Off-policy evaluation

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



Idea: Use partial matches!

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

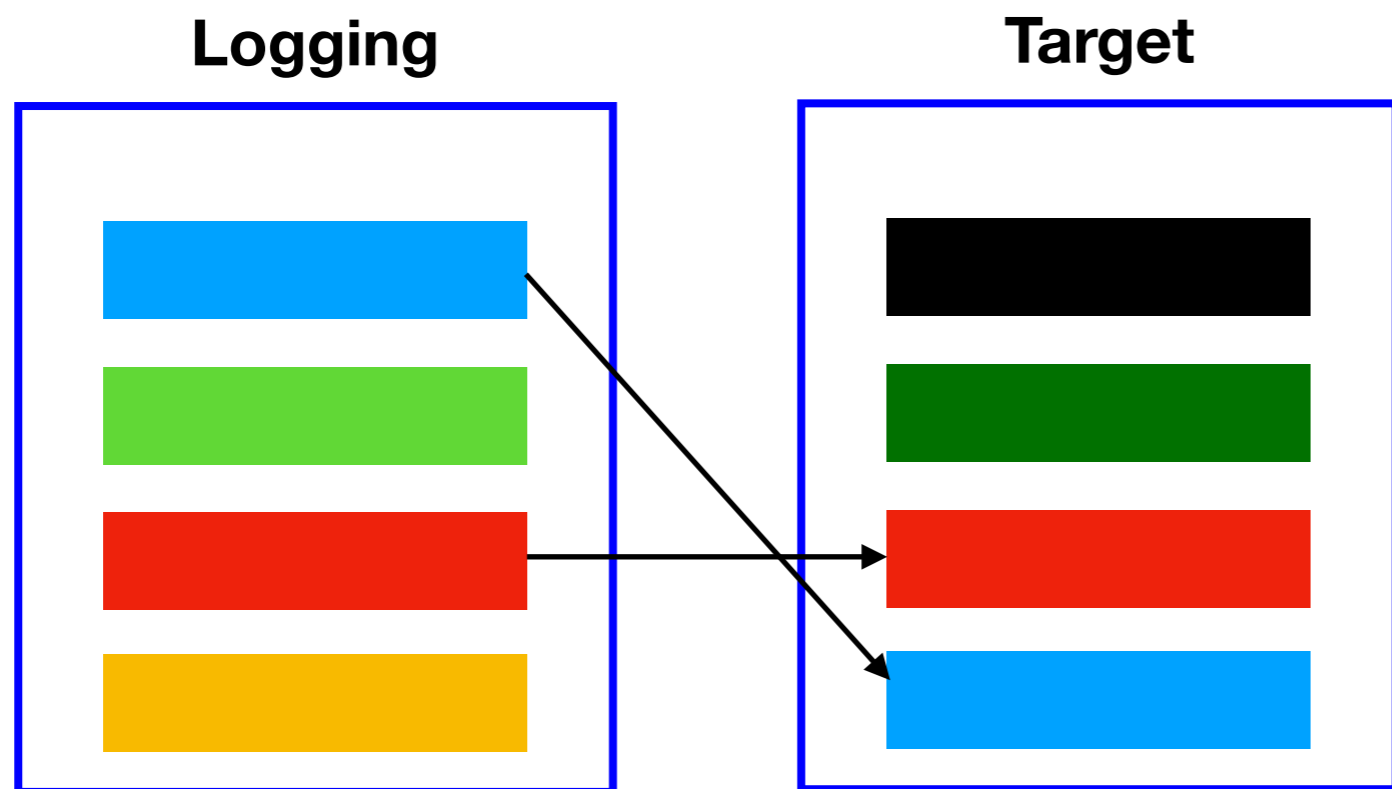
$$\hat{y}(a) = \frac{y(a)\mathbf{1}(a \in A)}{Q(a \in A|x)}$$

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

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

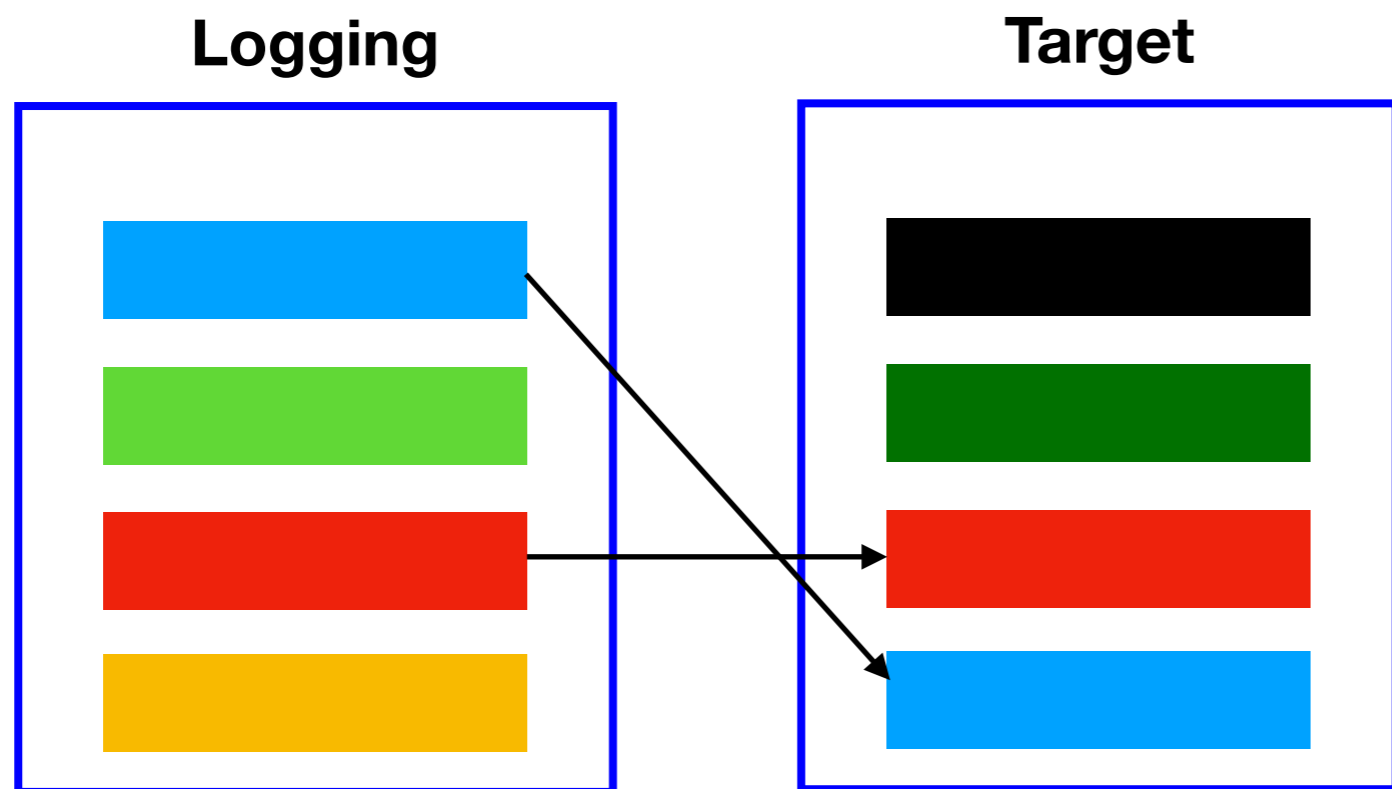
$$\hat{y}(a) = \frac{y(a)\mathbf{1}(a \in A)}{Q(a \in A|x)}$$

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

- 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 the MSN lifestyle website interface. At the top, there is a navigation bar with the MSN logo, the word "lifestyle" with a dropdown arrow, a search icon, and a user profile icon labeled "Akshay" with a settings gear. Below the navigation bar is a horizontal menu with the following items: "Home", "Cooking", "Restaurants & News" (which is underlined), "Beverages", and "Video".

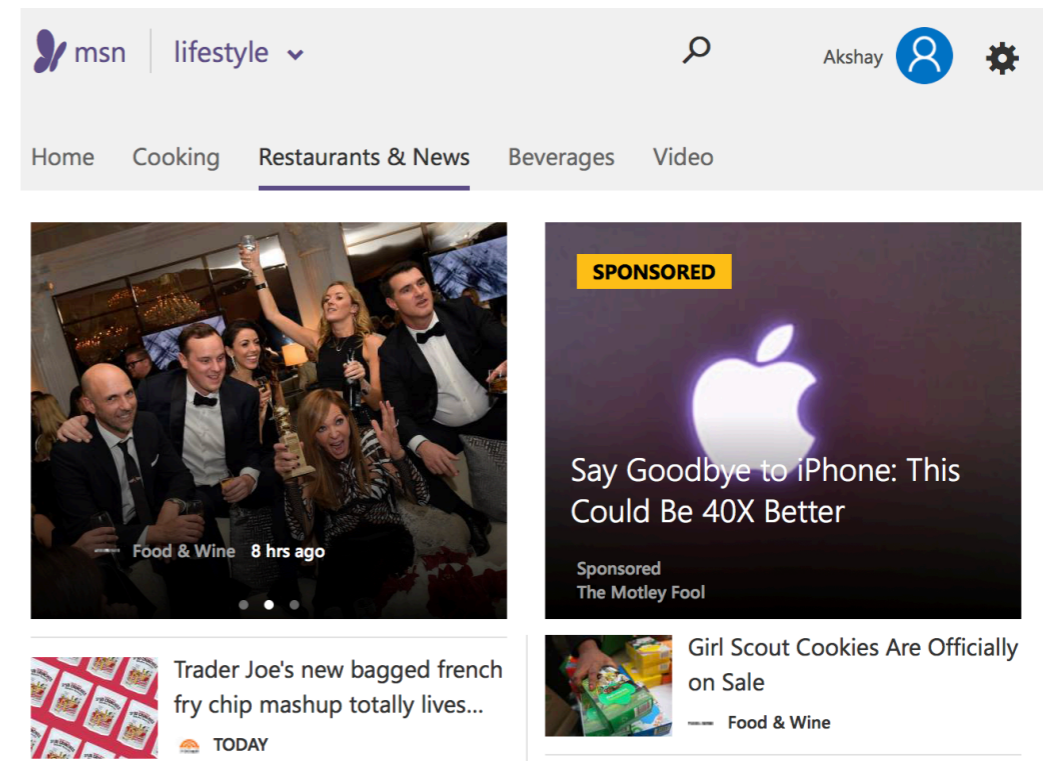
The main content area is a grid of four items:

- Top Left:** A photograph of a group of people in formal attire at a party. Below the image, it says "Food & Wine" and "8 hrs ago".
- Top Right:** A sponsored advertisement with a black background and a glowing Apple logo. The text reads "Say Goodbye to iPhone: This Could Be 40X Better". Below the text, it says "Sponsored" and "The Motley Fool".
- Bottom Left:** A small image of a red and white bagged product. The text reads "Trader Joe's new bagged french fry chip mashup totally lives..." and "TODAY".
- Bottom Right:** A small image of a box of Girl Scout cookies. The text reads "Girl Scout Cookies Are Officially on Sale" and "Food & Wine".

Combinatorial Contextual Bandits

On each of T rounds:

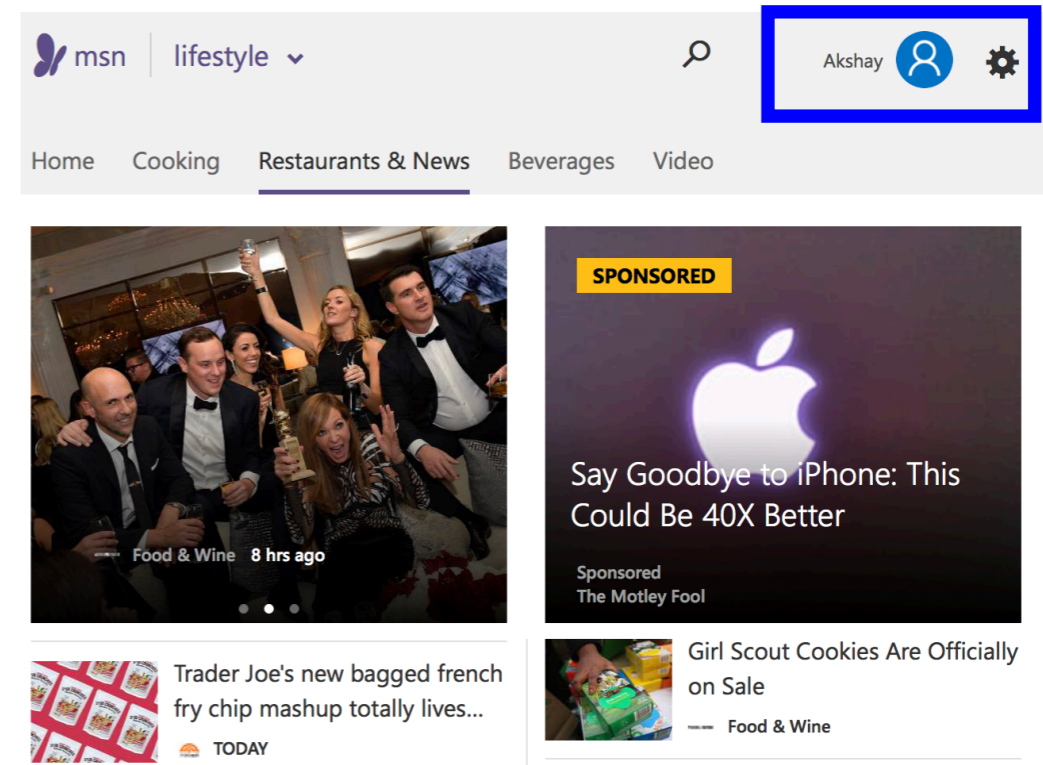
1. Observe context
2. Play action
3. *Unobserved* features
4. Observe reward



Combinatorial Contextual Bandits

On each of T rounds:

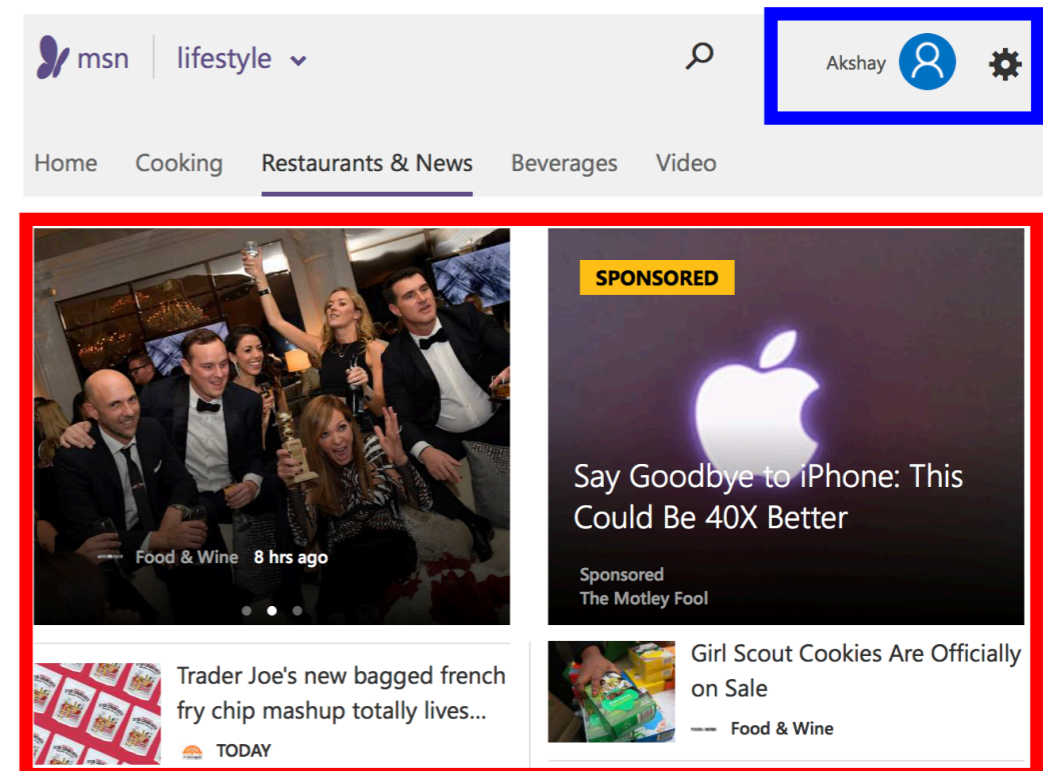
1. Observe context x_t
2. Play action
3. *Unobserved* features
4. Observe reward



Combinatorial Contextual Bandits

On each of T rounds:

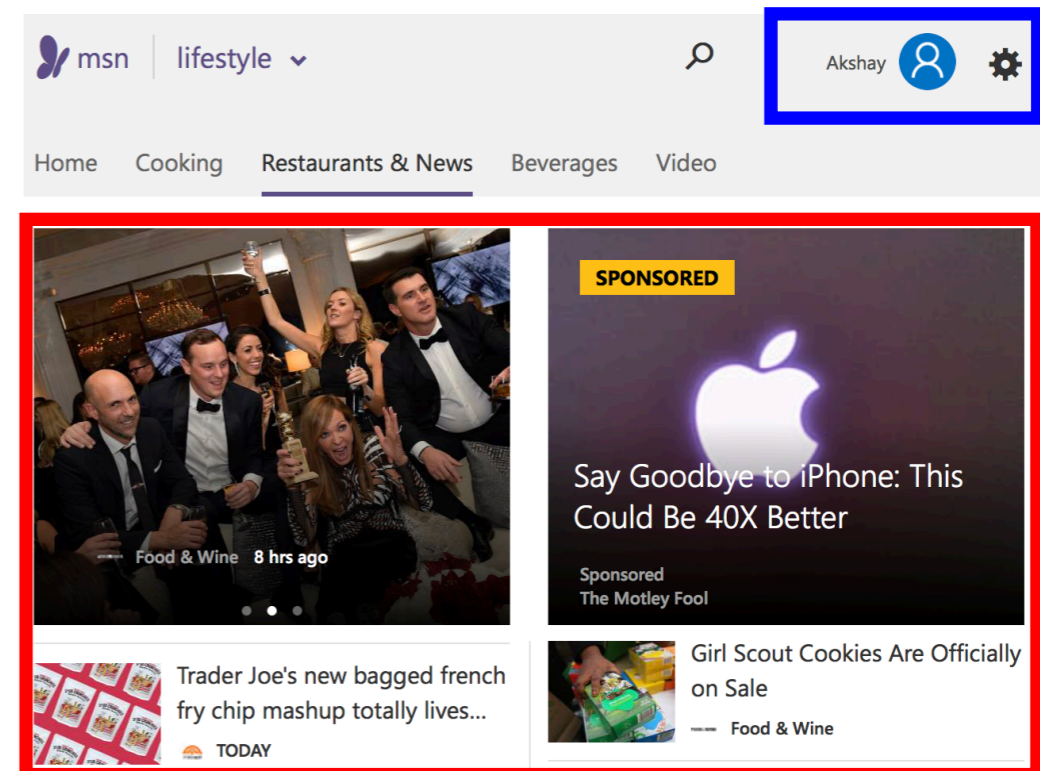
1. Observe context x_t
2. Play action $A_t = (a_1, \dots, a_L)$
3. *Unobserved* features
4. Observe reward



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

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

The image shows a screenshot of the MSN lifestyle website. At the top, there is a navigation bar with the MSN logo, the word "lifestyle", and a search icon. To the right of the search icon, the user's name "Akshay" is displayed next to a profile icon and a settings gear icon, all enclosed in a blue rectangular box. Below the navigation bar, there are several menu items: "Home", "Cooking", "Restaurants & News", "Beverages", and "Video". The main content area is divided into several sections. On the left, there is a large image of a group of people in formal attire at a party, with the text "Food & Wine 8 hrs ago" below it. To the right of this image is a sponsored advertisement for Apple, featuring the Apple logo and the text "Say Goodbye to iPhone: This Could Be 40X Better". Below the Apple ad, there is another sponsored advertisement for Girl Scout Cookies, with the text "Girl Scout Cookies Are Officially on Sale". At the bottom left, there is a small advertisement for Trader Joe's, with the text "Trader Joe's new bagged french fry chip mashup totally lives...".

Combinatorial Contextual Bandits

On each of T rounds:

1. Observe context x_t

2. Play action $A_t = (a_1, \dots, a_L)$

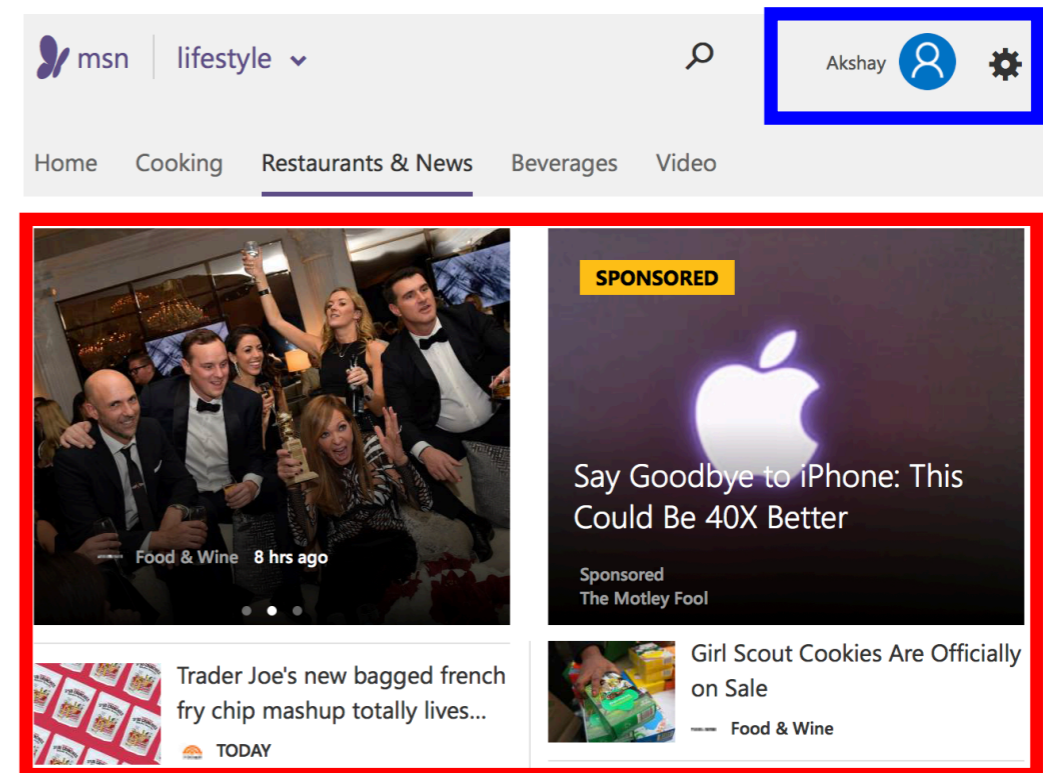
3. Unobserved features $\{y(\ell, a_\ell)\}_{\ell=1}^L$

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

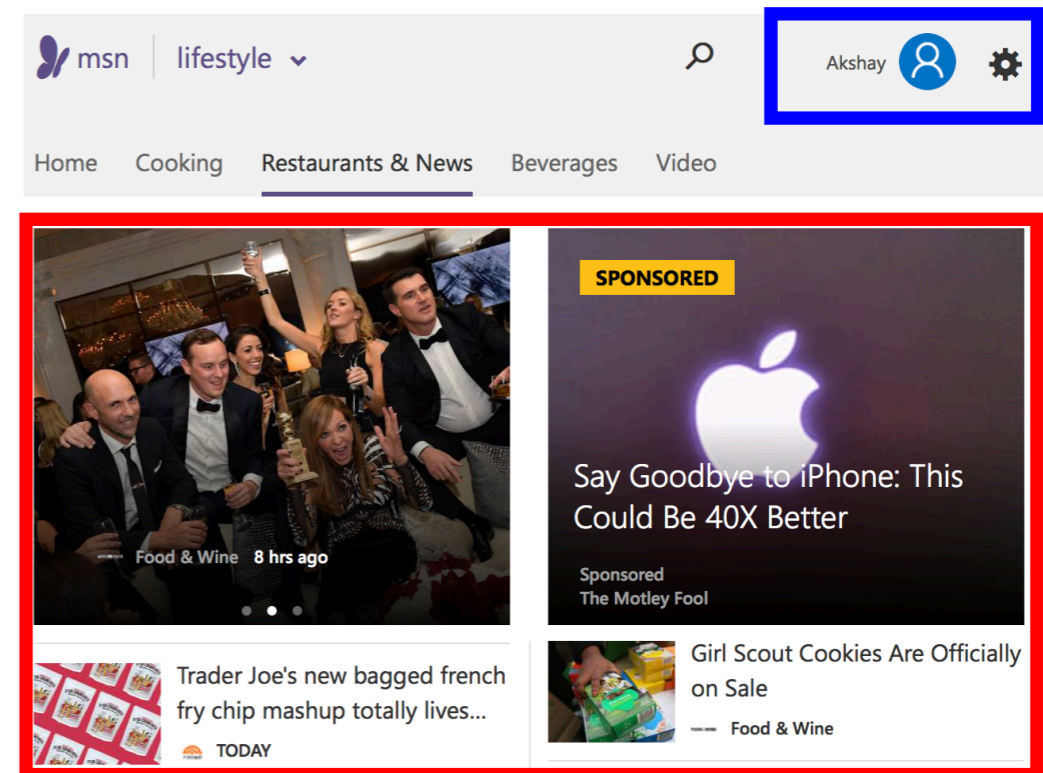
On each of T rounds:

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

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

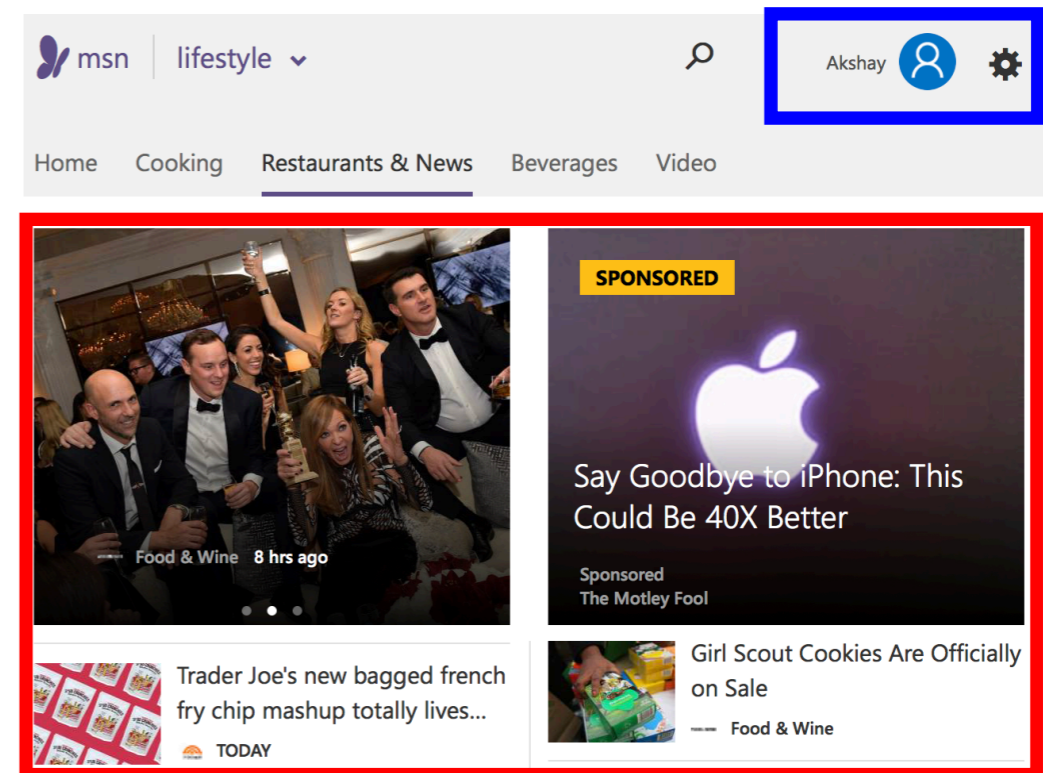
B = number of simple actions

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

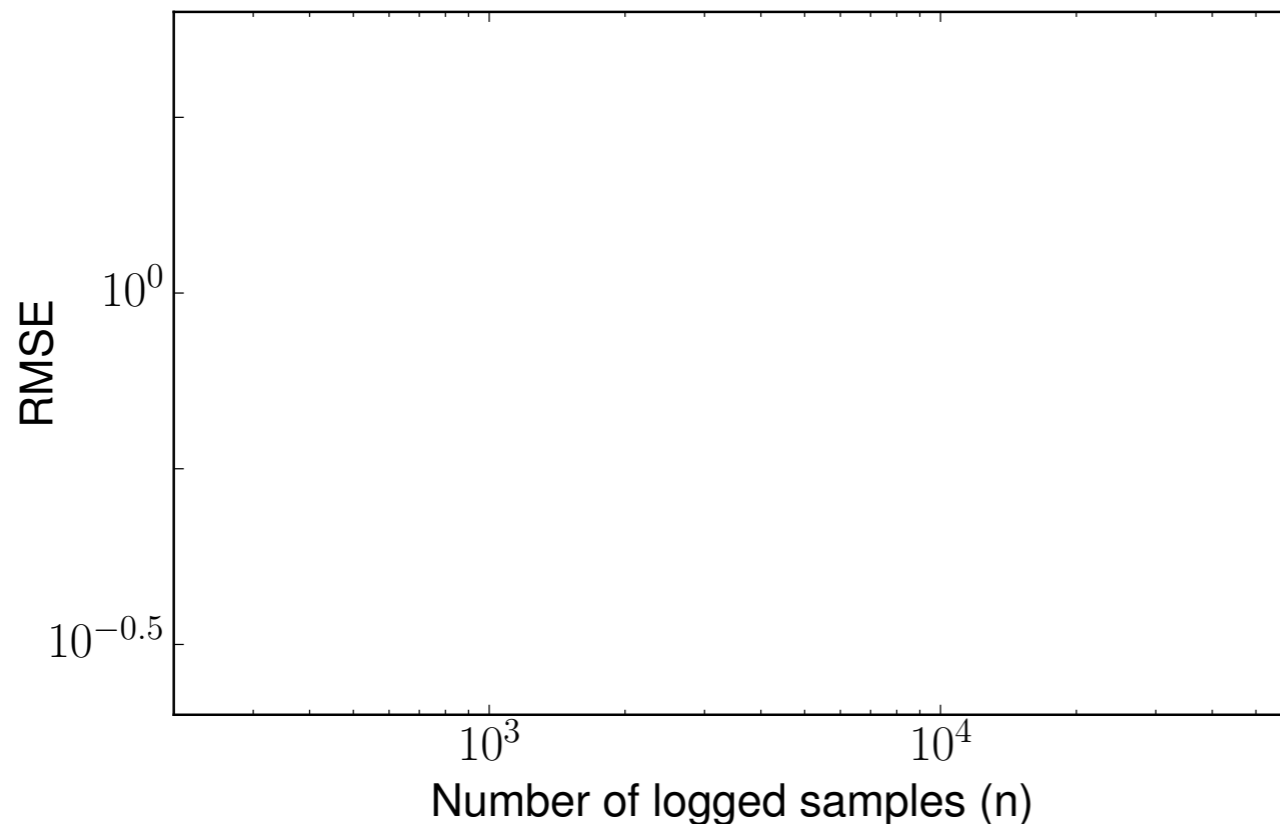
Subproblem: Given data collected by logging policy, estimate reward of a target policy

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

Reward: Utility Rate



Results — Off-policy evaluation

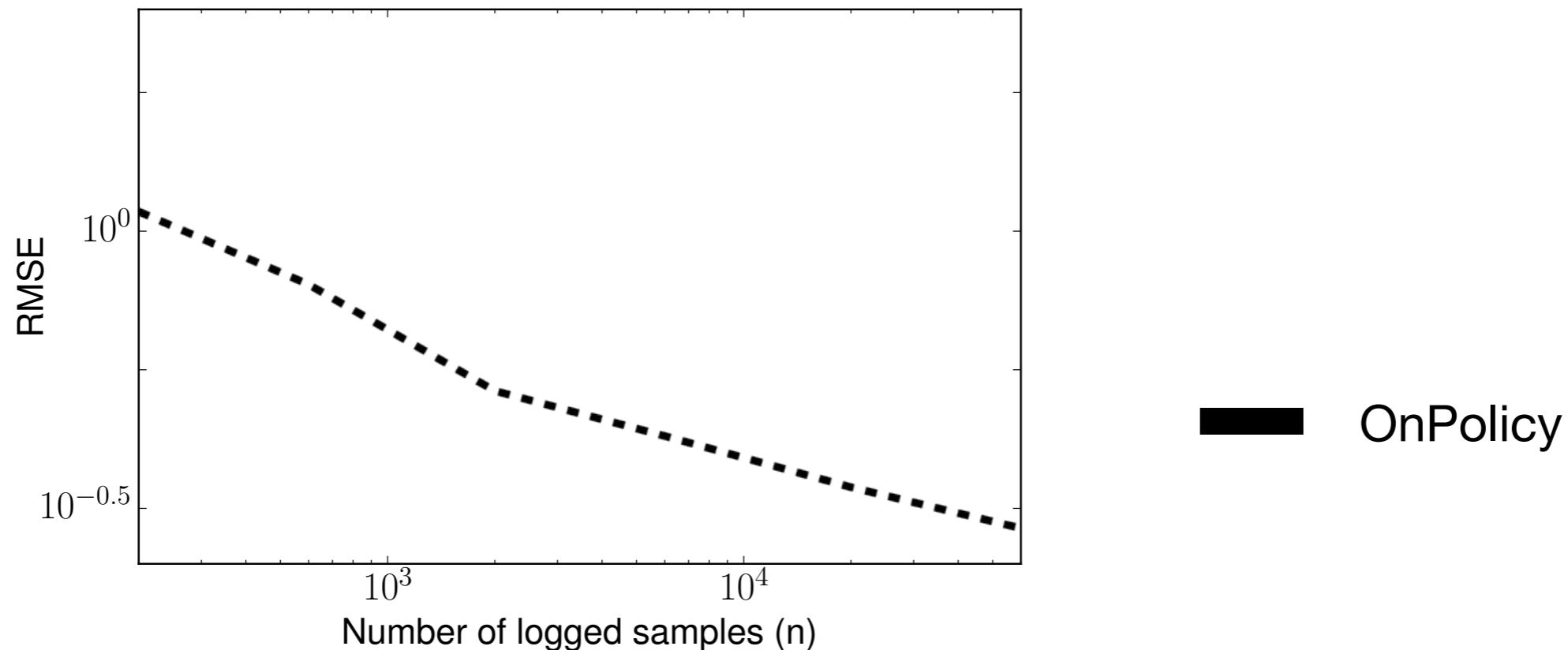
Subproblem: Given data collected by logging policy, estimate reward of a target policy

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

Reward: Utility Rate



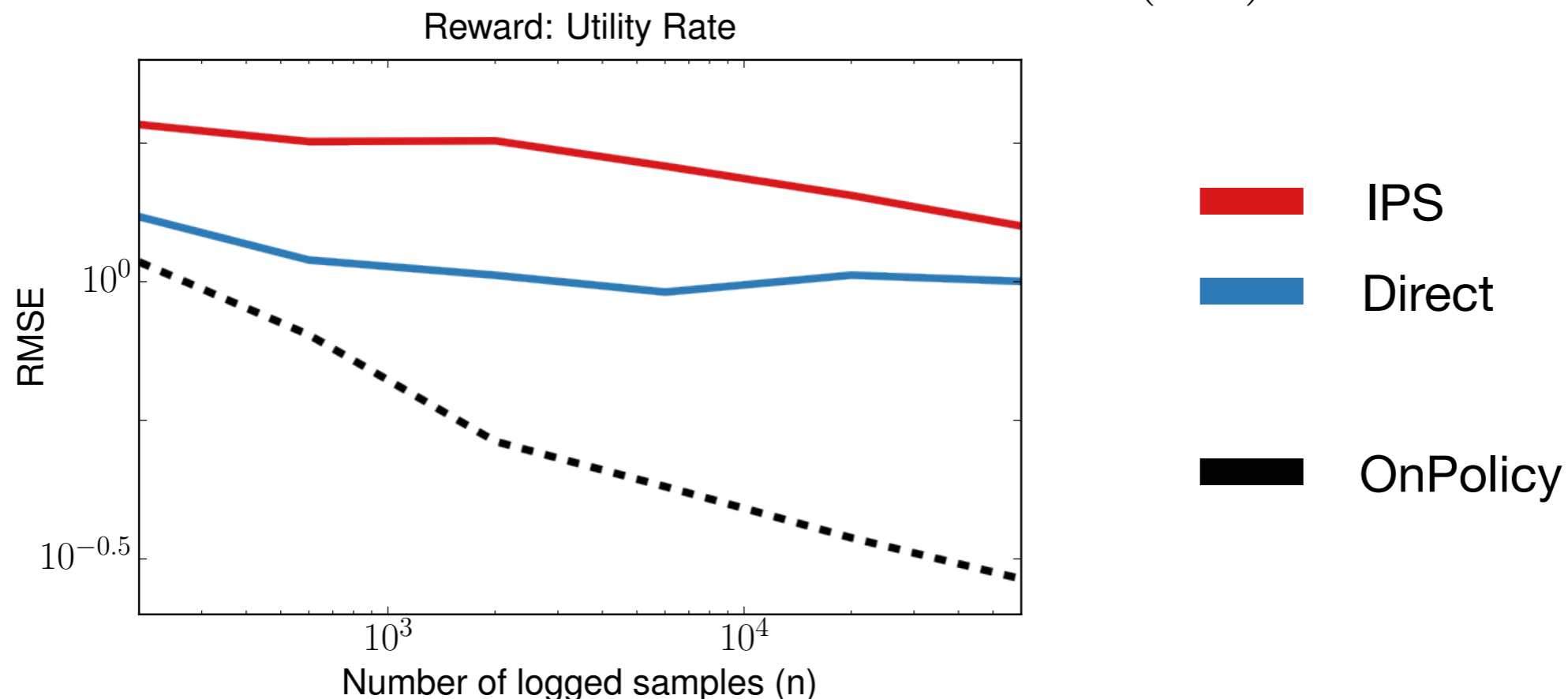
Results – Off-policy evaluation

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



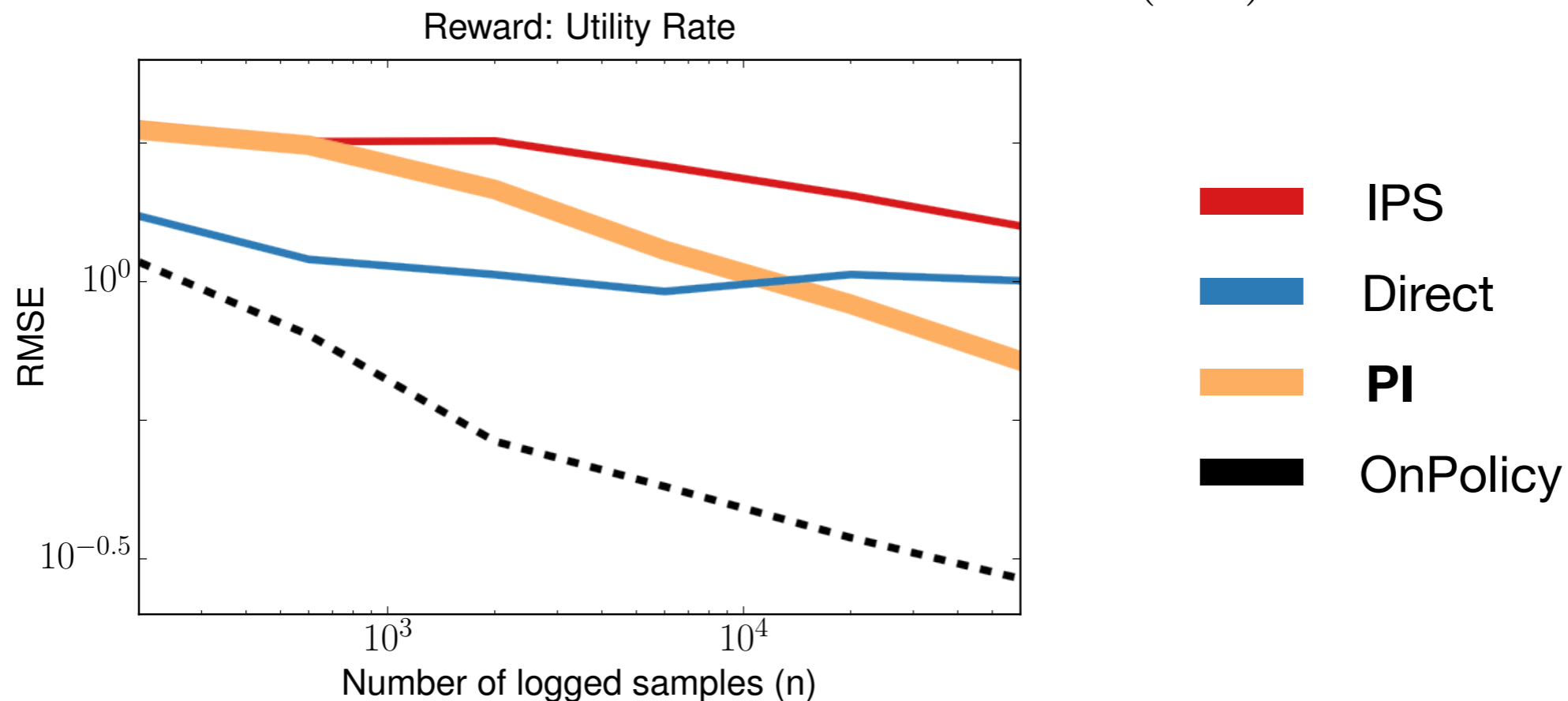
Results – Off-policy evaluation

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



Techniques

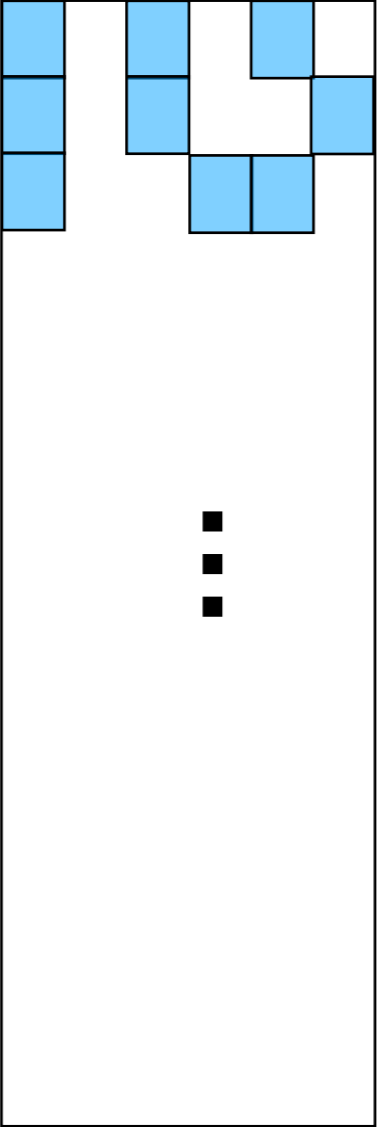
$$r_t(A_t, x_t) = \sum_{\ell} y(a_{\ell}) + \text{noise}$$

Techniques

$r(A)$



=



$y(\ell, a_\ell)$



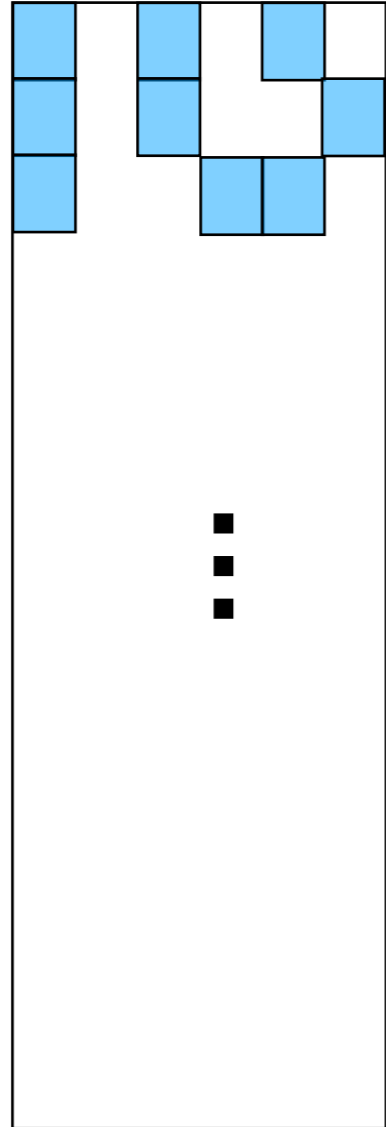
$$r_t(A_t, x_t) = \sum_{\ell} y(a_\ell) + \text{noise}$$

Techniques

$r(A)$



=



$y(l, a_l)$

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

With logging μ can write

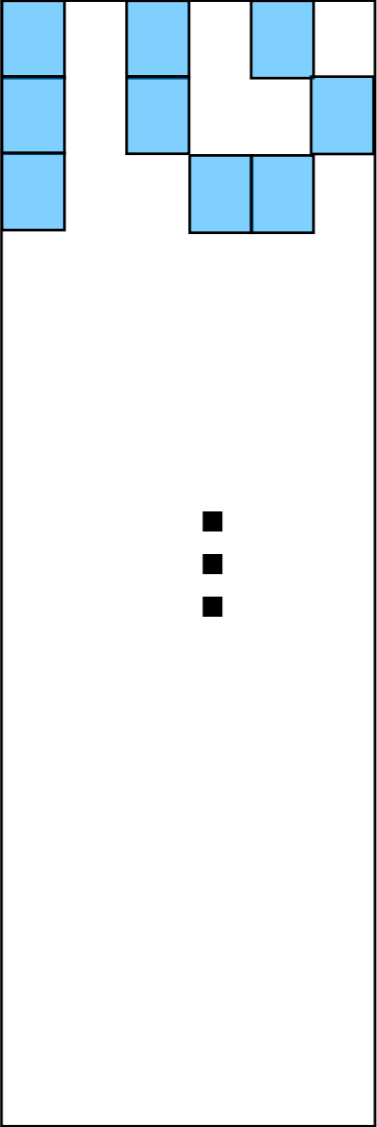
$$\bar{y} = \arg \min_w \mathbb{E}_{\mu}[(\mathbf{1}_A^T w - r)^2 | x]$$

Techniques

$r(A)$



=



$y(\ell, a_\ell)$



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

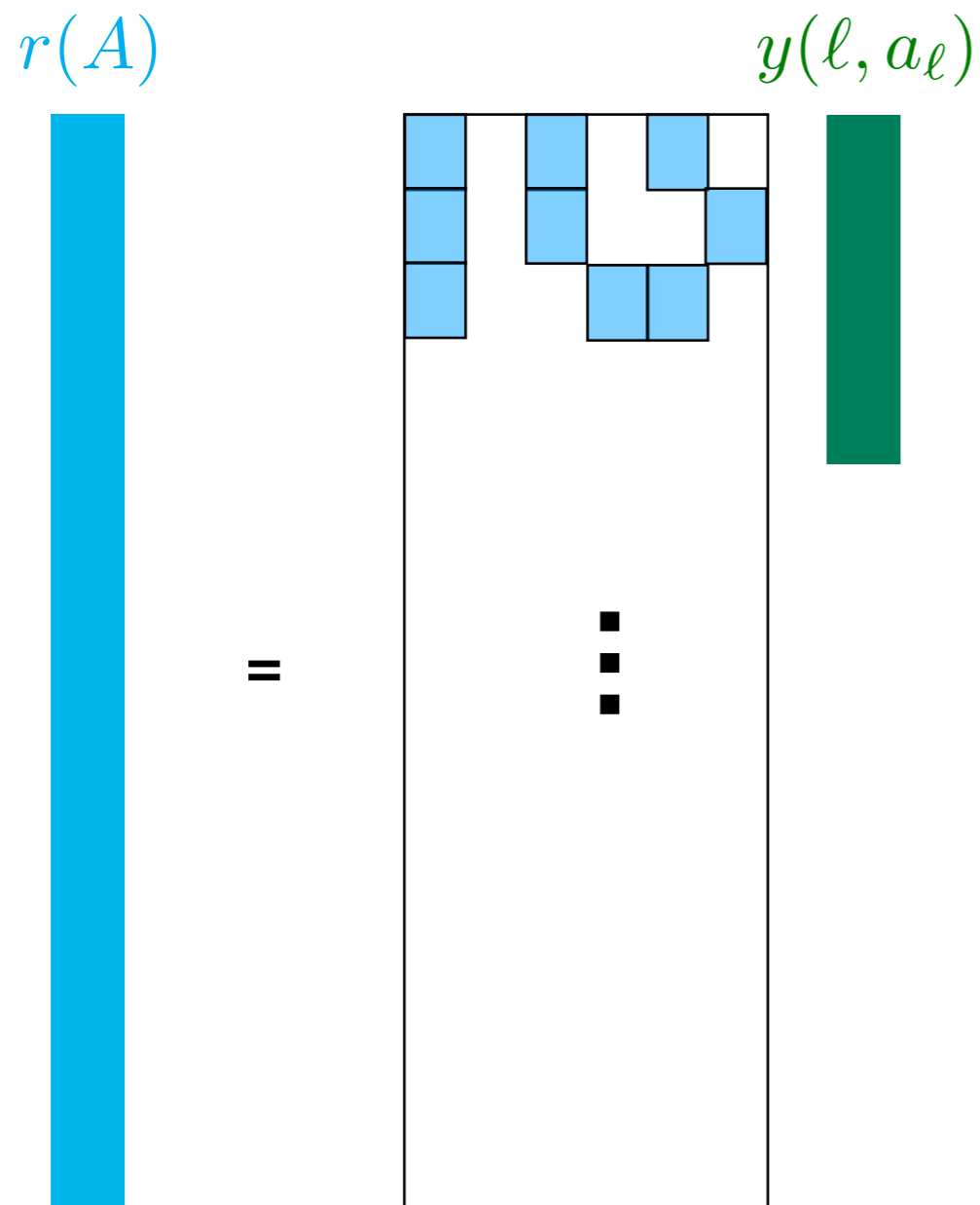
With logging μ can write

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

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

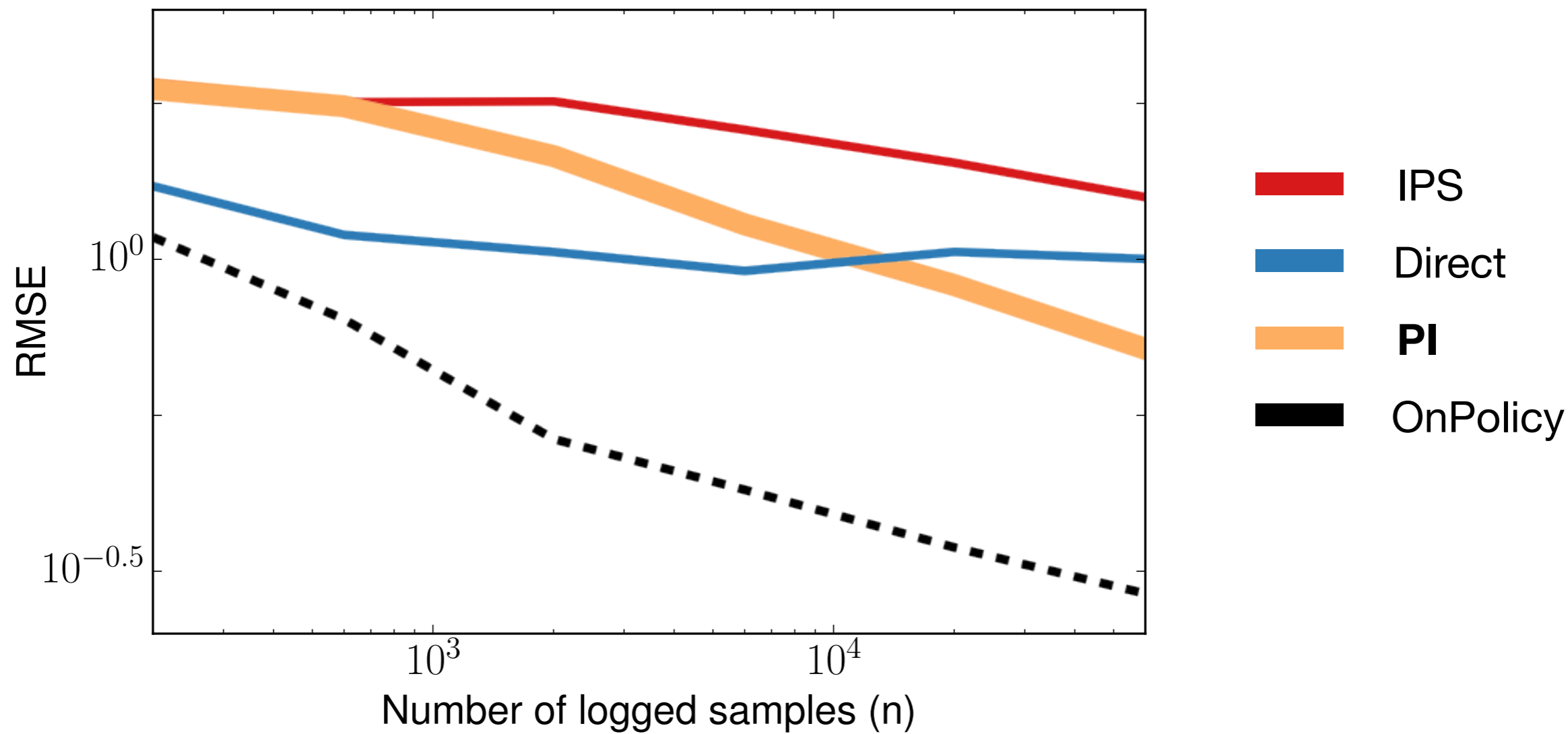
$$\hat{y}_t = (\mathbb{E}_{\mu} [\mathbf{1}_A \mathbf{1}_A^T])^{\dagger} \mathbf{1}_{A_t} r_t$$

For policy π

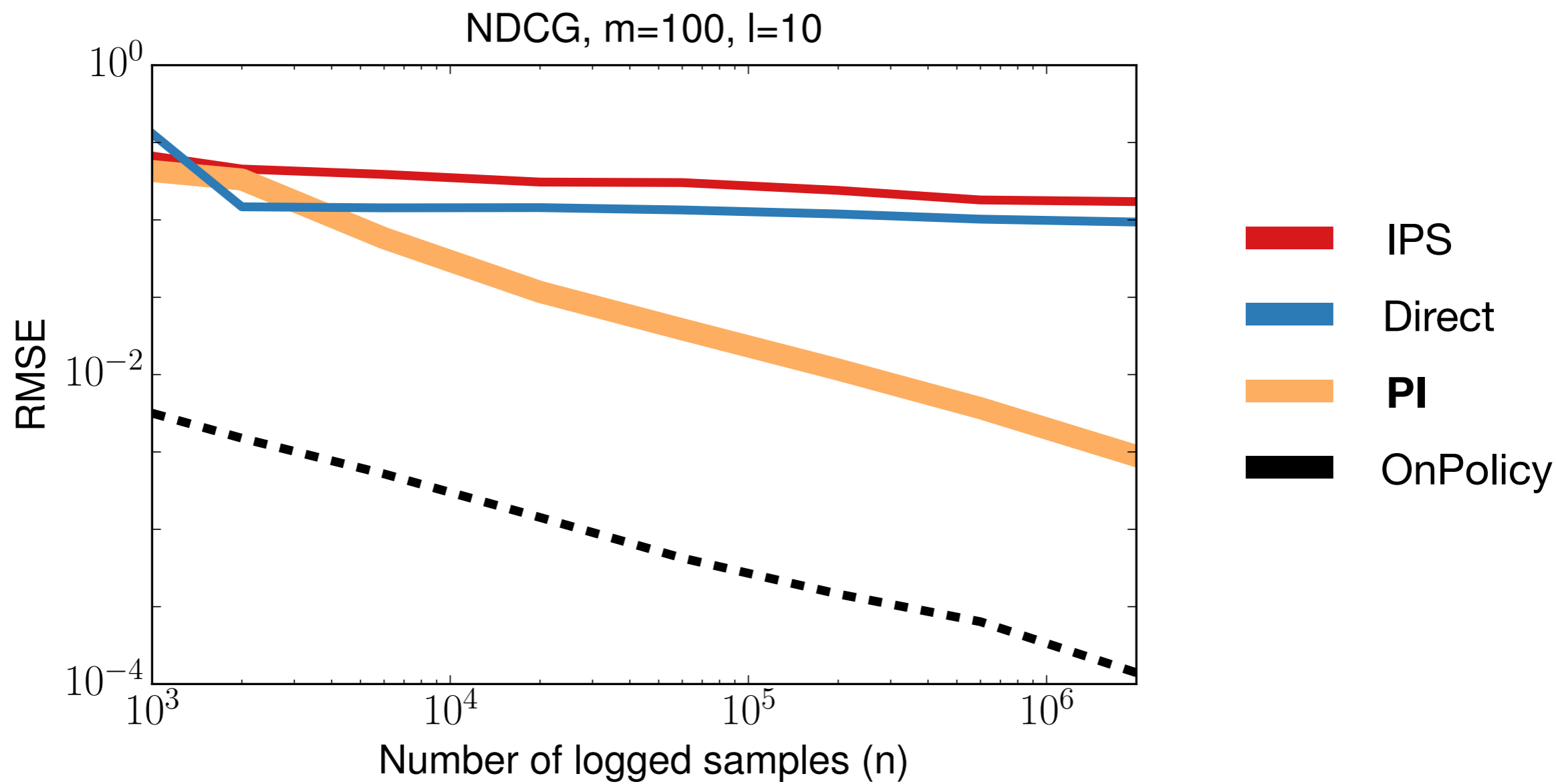
$$\hat{r}(\pi, x_t) = \mathbf{1}_{\pi(x_t)}^T \hat{y}_t$$

Experiment

Reward: Utility Rate



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

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