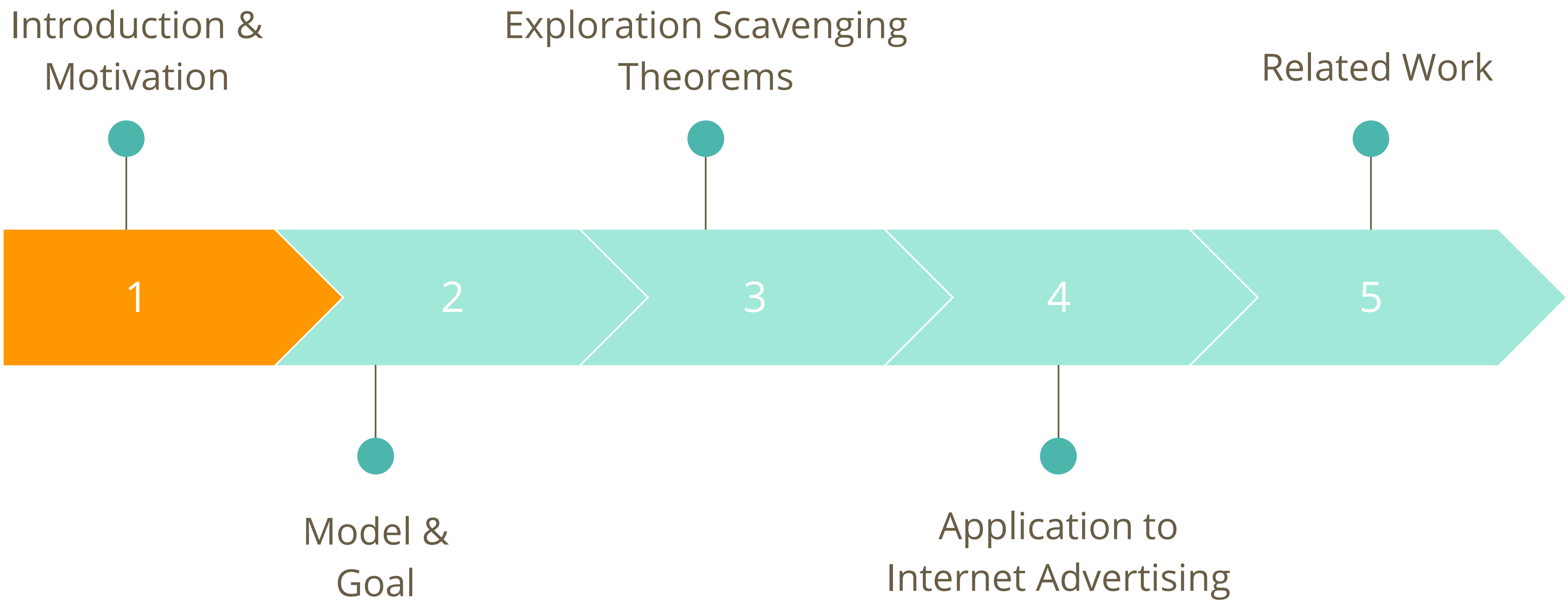
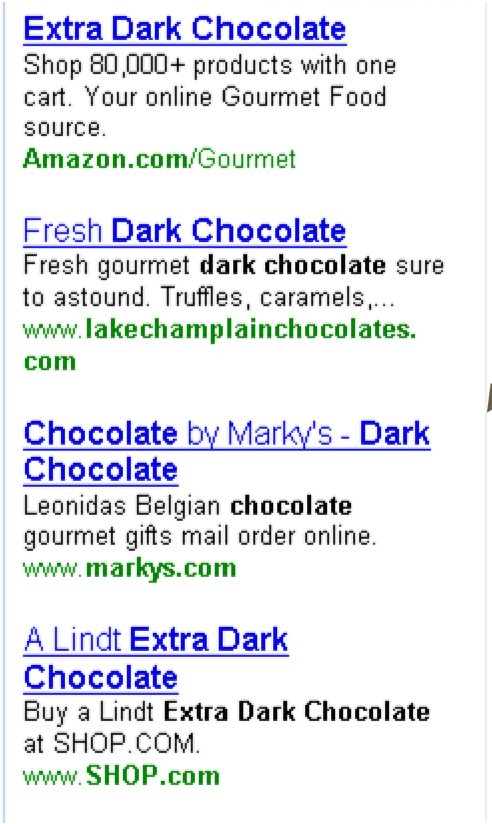
Off-Policy Evaluation

Miguel Aroca-Ouellette, Akshata Athawale, Mannat Singh


Introduction &
Motivation
Exploration Scavenging
Theorems
Related Work
1
2
3
4
5
Model &
Goal
Application to
Internet Advertising

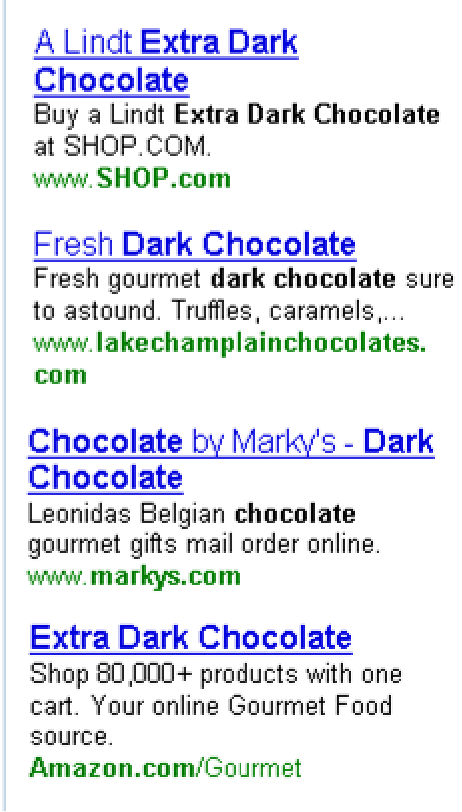
Motivation: Internet Advertising (Again)



Old Ad Serving Policy
(We have data!)

New (Better?) Policy
(No data)

Can we determine the value of our new policy using only our old data?



Policy evaluation

Definition: The problem of evaluating a new strategy for behavior, or *policy*, using only observations collected during the execution of another policy.

How can we evaluate the value of a new policy if we have no control over the available data?

Exploration Scavenging!

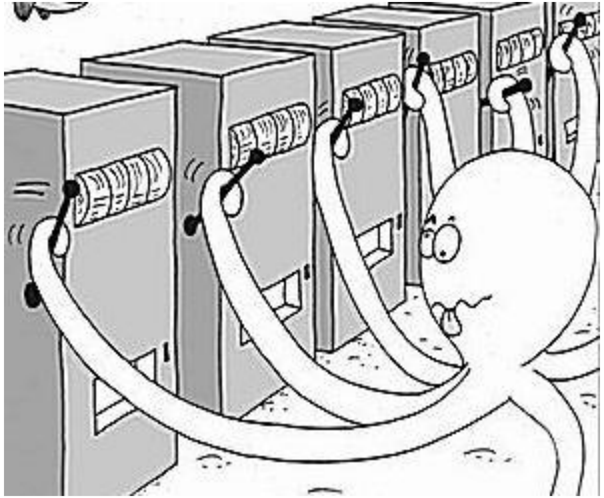
- Principled method for policy evaluation when the following (very) **restrictive** assumption holds true

The original exploration policy does not depend on the current input.

- Other assumption: Each action is chosen sufficiently often.
- Given this assumption, the technique can be used to accurately estimate the value of a new policy.
  - Bonus: Can be used even if the exploration policy is deterministic.
  - Bonus: A trick allows us to evaluate between multiple policies, even if they depend on the input. More on this later.

Contextual Bandit Setting (Again)

- Recall: k-armed bandit.
  - Formalization of exploration vs exploitation dilemma by autonomous agents.
- Recall: Contextual bandit
  - Generalization of standard k-armed bandit.
  - Allows agent to first observe *side information* or *context* before choosing an arm.
- In Advertisement:
  - Choose ad or set of ads to display.
  - Contextual information about user and page/article.
  - Reward in the form of CTR.



Context + Our Old Friend The Gambling Octopus

Why Some Other Models Fail

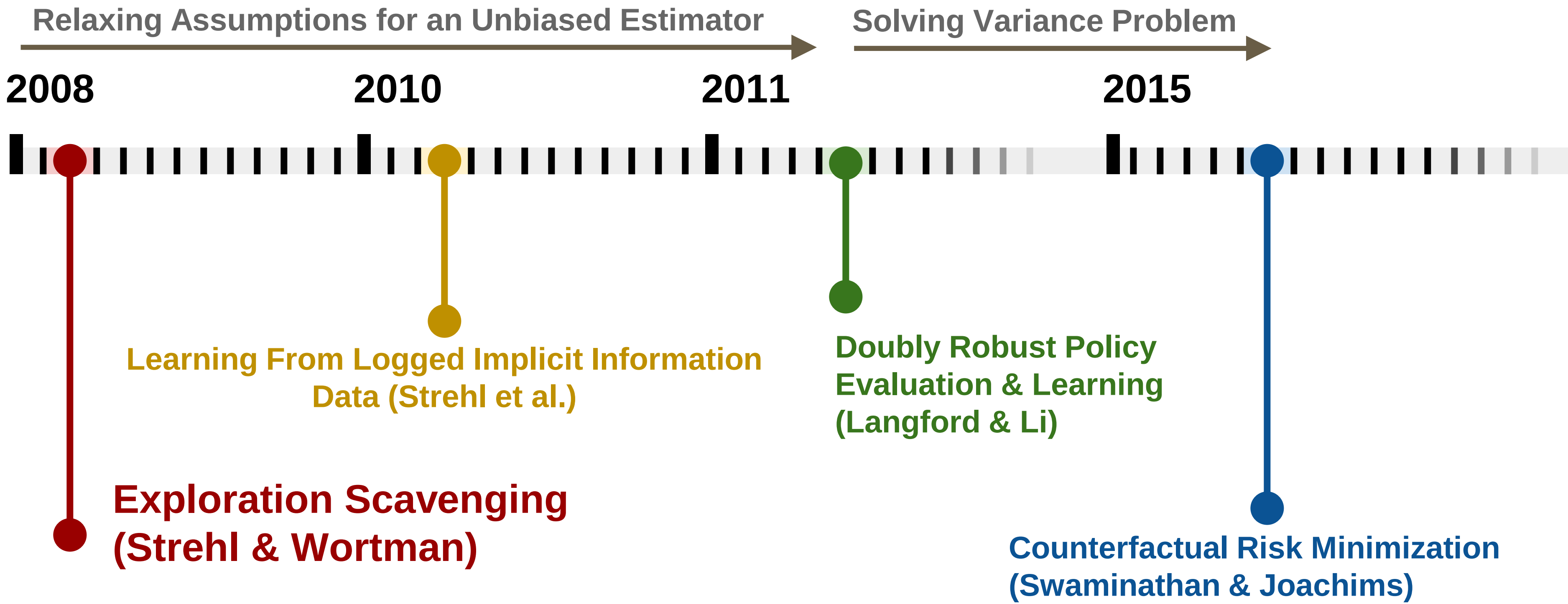
- **Recall:** Want to find a new policy maximizing our expected reward given a previous dataset. This is a “warm start” problem
- **Supervised learning using a regressor:** Generalizes poorly because it may include choices not in the data. Distribution mismatch.
- **Standard bandit:** Curse of dimensionality, because it requires condition on the context.
- **Contextual bandit:** Requires interaction or probability across the actions of the policy.

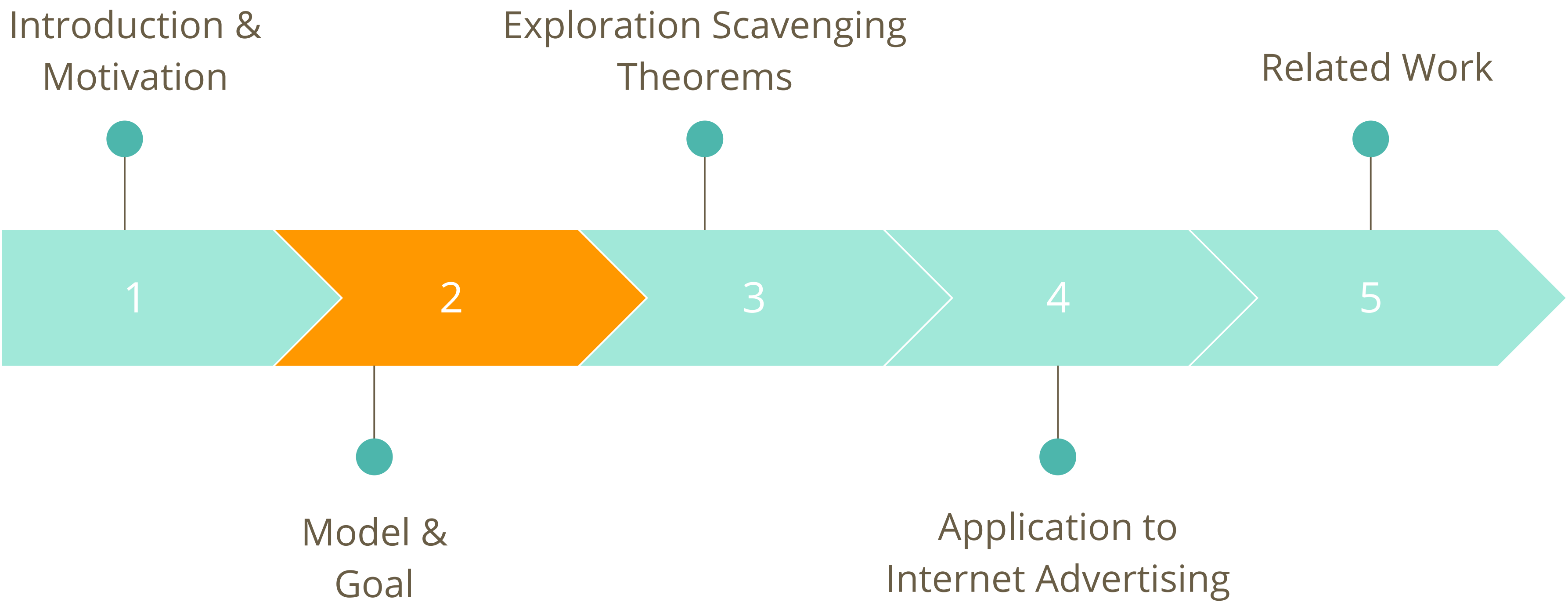
Exploration Scavenging provides a solution given our independence assumption.

Why Should I Care? (aka. Why Businesses Care)

- Want to evaluate new method without incurring the risk and cost of actually implementing this new method/policy.
- Existing logs containing huge amounts of historical data based on existing policies.
- It makes **economical** sense to, if possible, use these logs.
- It makes **economical** sense to, if possible, not risk the loss of testing out a new potentially bad policy.
- Online ad placement is once again a good example.

Literature Timeline




Introduction & Motivation
Exploration Scavenging Theorems
Related Work
1
2
3
4
5
Model & Goal
Application to Internet Advertising

Contextual Bandit Model

- Input Space: $\mathcal{X}$
- Action Set: $\mathcal{A}$
- Distribution of (input, reward) tuples:$(x, \vec{r}\ ) \sim D$
- Where $x \in \mathcal{X}$ and $\vec{r} \in [0,1]^k$
- Note: Non-contextual bandit is simply the case where $|\mathcal{X}| = 1$

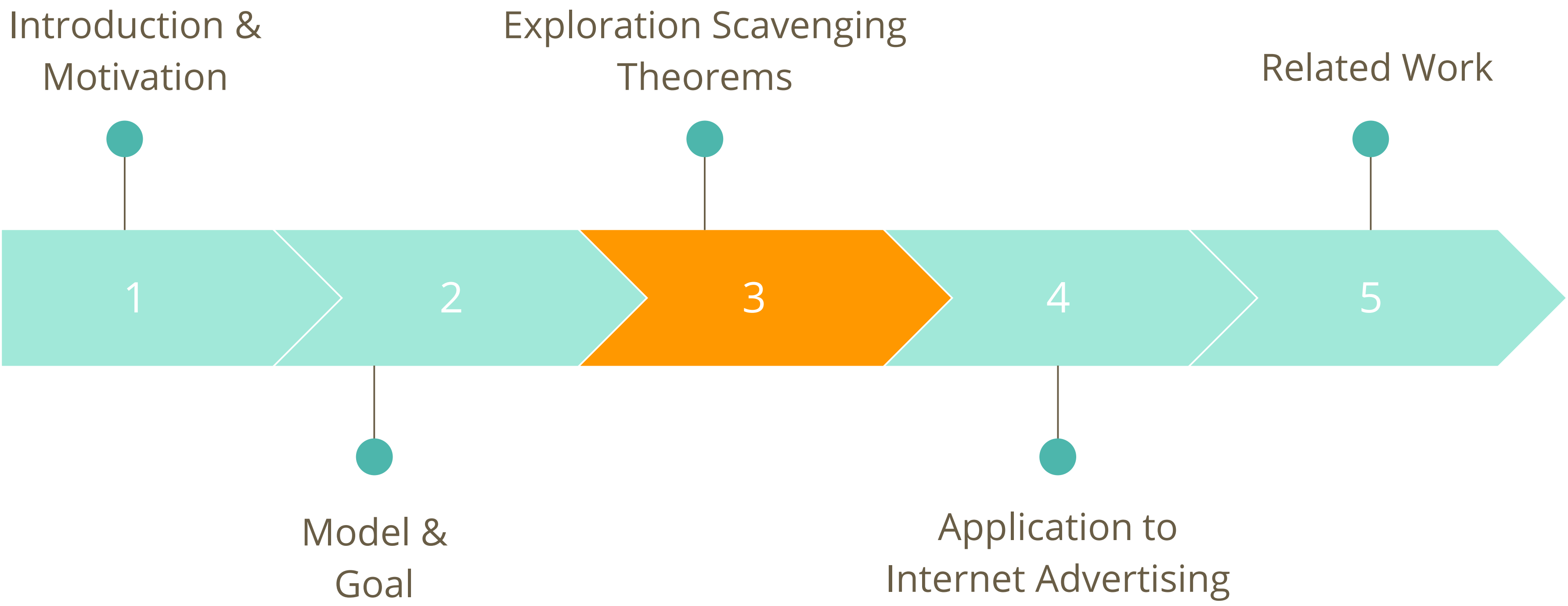
Contextual Bandit Model

- Events occur on a round by round basis where at each round t:
  - The world draws $(x, \vec{r}\,) \sim D$ and announces x_t
  - The algorithm chooses an action $a_t \in \mathcal{A}$
  - The world announces the reward r_{t,a_t} of action a_t
- The algorithm does not learn what reward it would have received had it chosen some other action.

Goal

- In the general bandit setting the goal is to maximize the sum of rewards over the rounds of interaction.
- However, our focus here is the subgoal of **policy evaluation**.
- Explicitly, given a data set $S \in (\mathcal{X} \text{ x } \mathcal{A} \text{ x } [0,1])^T$ which is generated by following some fixed policy π for T steps.
- Given a different policy $h : \mathcal{X} \rightarrow \mathcal{A}$ we want to estimate the value of policy h, where value is defined as

$$V_D(h) = E_{(x,\vec{r})\sim D}[r_{h(x)}]$$


Introduction &
Motivation
Exploration Scavenging
Theorems
Related Work
1
2
3
4
5
Model &
Goal
Application to
Internet Advertising

Impossibility Results

- Policy evaluation not possible when the exploration policy π chooses some action a with zero probability

Natural Question: Is it possible to have an evaluation procedure as long as π chooses each action sufficiently often?

- If π depends on the current input, there are cases when new policies h cannot be evaluated, even if each action is chosen frequently by π
- If input-dependent exploration policies are disallowed, policy evaluation becomes possible

Proving that Evaluation is not possible in general

Theorem 1: There exist contextual bandit problems D and D' with $k = 2$ actions, a hypothesis h, and a policy π dependent on the current observation x_t with each action visited with probability ½, such that the observations of π on D are statistically indistinguishable from observations of π on D', yet $|V_D(h) - V_{D'}(h)| = 1$

Proof of Theorem 1

Proof: The proof is by construction. Suppose x_t takes on the values 0 and 1, each with probability ½ under both D and D'. Let $\pi(x) = x$ be the exploration policy, and let $h(x) = 1 - x$ be the policy we wish to evaluate. Suppose that rewards are deterministic given x_t.

	Under D		**Under D'**	
	$r_{t,0}$	$r_{t,1}$	$r_{t,0}$	$r_{t,1}$
$x_t = 0$	0	0	0	1
$x_t = 1$	0	1	1	1

Here, $V_D(h) = 0$, while $V_{D'}(h) = 1$, but observations collected using exploration policy π are indistinguishable for D and D'

Techniques for Policy Evaluation

- We have established that policy evaluation can be impossible in general
- Cannot perform policy evaluation when –
    - The exploration ploicy π depends on the current input
    - π fails to choose each action sufficiently often

Next Question: Can policy evaluation be done when this is not the case?

- We now discuss techniques for policy evaluation under these special circumstances

Exact Theoretical Estimator for the Value

Theorem 2: For any contextual bandit distribution D over $(x, \vec{r})$, any policy h, any exploration policy π such that

- For each action a, there is a constant $T_a > 0$ for which $|\{t : a_t = a\}| = T_a$ with probability 1
- π chooses a_t independent of x_t,

$$V_D(h) = E_{\{x_t, \vec{r_t}\} \sim D^T}\left[\sum_{t=1}^{T} \frac{r_{t,a_t} I(h(x_t) = a_t)}{T_{a_t}}\right]$$

Proof of Theorem 2

Proof: $E_{\{x_t,\vec{r_t}\}\sim D^T}\left[\sum_{t=1}^{T}\frac{r_{t,a_t}I(h(x_t)=a_t)}{T_{a_t}}\right]$

Reordering terms in the summation, we can write $t = \{1, \ldots, T\} = \cup_{a\in\{1,\ldots,k\}}\{t : a_t = a\}$

$$= E_{\{x_t,\vec{r_t}\}\sim D^T}\left[\sum_{a=1}^{k}\sum_{\{t:a_t=a\}}\frac{r_{t,a}\,I(h(x_t)=a)}{T_a}\right]$$

By linearity of expectation,

$$=\sum_{a=1}^{k} E_{\{x_t,\vec{r_t}\}\sim D^T}\left[\sum_{\{t:a_t=a\}}\frac{r_{t,a}\,I(h(x_t)=a)}{T_a}\right]$$

The exploration policy π chooses a_t independent of x_t, and T_a is fixed. So, $\frac{r_{t,a}\,I(h(x_t)=a)}{T_a}$ is identically distributed for all t such that $a_t = a$.

Proof of Theorem 2

So, we get,

$$E_{\{x_t,\vec{r_t}\}\sim D^T}\left[\sum_{t=1}^{T}\frac{r_{t,a_t} I(h(x_t)=a_t)}{T_{a_t}}\right]$$

$$=\sum_{a=1}^{k} E_{\{x_t,\vec{r_t}\}\sim D^T}\left[\sum_{\{t:a_t=a\}}\frac{r_{t,a}\, I(h(x_t)=a\,)}{T_a}\right]$$

$$=\sum_{a=1}^{k} E_{(x,\vec{r})\sim D}\left[T_a\frac{r_a\, I(h(x)=a\,)}{T_a}\right]$$

By linearity of expectation again,

$$=E_{(x,\vec{r})\sim D}\left[\sum_{a=1}^{k} r_a\, I(h(x)=a\,)\right]$$

$$=V_D(h)$$

The Practical Estimator

Theorem 3: For every contextual bandit distribution D over $(x, \vec{r}\,)$ with rewards $r_a \in [0,1]$, for every sequence of T actions a_t chosen by an exploration policy π that may be a function of history but does not depend on x_t, for every hypothesis h, and for any $\delta \in (0,1)$, with probability $1-\delta$,

$$\left| V_D(h) - \sum_{t=1}^{T} \frac{r_{t,a_t} I(h(x_t) = a_t)}{T_{a_t}} \right| \leq \sum_{a=1}^{k} \sqrt{\frac{2\ln(2kT/\delta)}{T_a}}$$

Proof of Theorem 3

Proof: $V_D(h) = E_{(x,\vec{r})\sim D}\left[\sum_{a=1}^{k} r_a\, I(h(x) = a\,)\right]$

$$\left|V_D(h) - \sum_{t=1}^{T} \frac{r_{t,a_t} I(h(x_t) = a_t)}{T_{a_t}}\right| = \left|E_{(x,\vec{r})\sim D}\left[\sum_{a=1}^{k} r_a\, I(h(x) = a\,)\right] - \sum_{t=1}^{T} \frac{r_{t,a_t} I(h(x_t) = a_t)}{T_{a_t}}\right|$$

By linearity of Expectation,

$$= \left|\sum_{a=1}^{k} E_{(x,\vec{r})\sim D}\left[r_a\, I(h(x) = a\,)\right] - \sum_{t=1}^{T} \frac{r_{t,a_t} I(h(x_t) = a_t)}{T_{a_t}}\right|$$

Fix an action a. Let t_i denote the i^{th} time step that action a was taken, with i ranging from 1 to T_a.

Again, we can use the fact that $t = \{1, \ldots, T\} = \bigcup_{a\in\{1,\ldots,k\}}\{t : a_t \; = a\} = \bigcup_{a\in\{1,\ldots,k\}}\{t_i : i \in \{1, \ldots, T_a\}\}$ to get

Proof of Theorem 3

$$\left|\sum_{a=1}^{k} E_{(x,\vec{r})\sim D}\left[r_a\, I(h(x)=a\,)\right] - \sum_{t=1}^{T} \frac{r_{t,a_t} I(h(x_t)=a_t)}{T_{a_t}}\right|$$

$$= \left|\sum_{a=1}^{k}\left[E_{(x,\vec{r})\sim D}\left[r_a\, I(h(x)=a\,)\right] - \frac{1}{T_a}\sum_{i=1}^{T_a} r_{t_i,a} I\left(h\left(x_{t_i}\right)=a\right)\right]\right|$$

From the triangle inequality, we get,

$$\leq \sum_{a=1}^{k}\left|E_{(x,\vec{r})\sim D}\left[r_a\, I(h(x)=a\,)\right] - \frac{1}{T_a}\sum_{i=1}^{T_a} r_{t_i,a} I\left(h\left(x_{t_i}\right)=a\right)\right|$$

R.H.S. of original bound $= \sum_{a=1}^{k}\sqrt{\frac{2\ln(2kT/\delta)}{T_a}}$

Proof of Theorem 3

So, if we are able to prove for all actions $a \in \{1, \dots, k\}$ that, with probability at most δ/k,

$$\left| E_{(x,\vec{r})\sim D}\left[r_a\, I(h(x) = a\,)\right] - \frac{1}{T_a}\sum_{i=1}^{T_a} r_{t_i,a} I\left(h\left(x_{t_i}\right) = a\right) \right| > \sqrt{\frac{2\ln(2kT/\delta)}{T_a}}$$

we can use the union bound to show that with probability $1 - \sum_{a\in\{1,\dots,k\}}\left(\frac{\delta}{k}\right) = 1 - \delta$,

$$\left| V_D(h) - \sum_{t=1}^{T} \frac{r_{t,a_t} I(h(x_t) = a_t)}{T_{a_t}} \right| \le \sum_{a=1}^{k} \left| E_{(x,\vec{r})\sim D}\left[r_a\, I(h(x) = a\,)\right] - \frac{1}{T_a}\sum_{i=1}^{T_a} r_{t_i,a} I\left(h\left(x_{t_i}\right) = a\right) \right|$$

$$\le \sum_{a=1}^{k} \sqrt{\frac{2\ln(2kT/\delta)}{T_a}}$$

Proof of Theorem 3

Fix an action a. Let us define for $i \in \{1, \dots T\}$,

$$Z_i = \begin{cases} r_{t_i,a} I\left(h\left(x_{t_i}\right) = a\right) - E_{(x,\vec{r})\sim D}\left[r_a\, I(h(x) = a\,)\right], & if\ i \le T_a \\ 0, & otherwise \end{cases}$$

Note that $Z_i \in [-1, 1]$ and $E[Z_i] = 0$.

Fix $t \in \{1, \dots, T\}$. Applying Azuma's inequality, we get that for any $\delta' \in (0,1)$, with probability $1 - \delta'$,

$$\frac{1}{t}\left|\sum_{i=1}^{t} Z_i\right| \le \sqrt{\frac{2\ln(2/\delta')}{t}}$$

So, if $t \le T_a$,

$$\frac{1}{t}\left|\sum_{i=1}^{t} Z_i\right| = \frac{1}{t}\left|\sum_{i=1}^{t} r_{t_i,a} I\left(h\left(x_{t_i}\right) = a\right) - E_{(x,\vec{r})\sim D}\left[r_a\, I(h(x) = a\,)\right]\right|$$

Proof of Theorem 3

$$\frac{1}{t}\left|\sum_{i=1}^{t} r_{t_i,a} I\left(h\left(x_{t_i}\right)=a\right)-E_{(x,\vec{r})\sim D}\left[r_a\, I(h(x)=a\,)\right]\right|$$

$= \left|E_{(x,\vec{r})\sim D}\left[r_a\, I(h(x)=a\,)\right] - \frac{1}{t}\sum_{i=1}^{t} r_{t_i,a} I\left(h\left(x_{t_i}\right)=a\right)\right| \le \sqrt{\frac{2\ln(2/\delta\prime)}{t}}$ with probability $1-\delta'$.

Taking $\delta' = \delta/(Tk)$,

$\left|E_{(x,\vec{r})\sim D}\left[r_a\, I(h(x)=a\,)\right] - \frac{1}{t}\sum_{i=1}^{t} r_{t_i,a} I\left(h\left(x_{t_i}\right)=a\right)\right| > \sqrt{\frac{2\ln(2kT/\delta)}{t}}$ with probability $\delta/(Tk)$

We have that this equation holds for $t \in \{1, \dots, T\}$. So, we have T inequalities.

For the above equation to hold for all $t \in \{1, \dots, T\}$, all the T inequalities have to hold. Using the union bound, the probability of that happening is upper bounded by $\sum_{t\in\{1,\dots,T\}} \delta/(Tk) \ = T\delta/(Tk) = \frac{\delta}{k}$

Proof of Theorem 3

Since the inequality holds for all t with probability δ/k, it holds for $t = T_a$ with probability δ/k.

$$\left|E_{(x,\vec{r})\sim D}[r_a\, I(h(x) = a\,)] - \frac{1}{T_a}\sum_{i=1}^{T_a} r_{t_i,a} I\left(h\left(x_{t_i}\right) = a\right)\right| > \sqrt{\frac{2\ln(2kT/\delta)}{T_a}} \text{ with probability } \delta/k$$

Note that we can’t directly say that this happens with a probability of δ/kT by taking $t = T_a$, as T_a is a random variable and our analysis is for a fixed t, it doesn’t hold when t is a random variable.

Practical Estimator Reaching the Exact Value

Corollary 4: For every contextual bandit distribution D over $(x, \vec{r})$, for every exploration policy π choosing action a_t independent of the current input, for every hypothesis h, is every action $a \in \{1, \dots, k\}$ is guaranteed to be chosen by π at least a constant fraction of the time, then as $T \to \infty$, the estimator

$$\widehat{V_D}(h) = \sum_{t=1}^{T} \frac{r_{t,a_t} I(h(x_t) = a_t)}{T_{a_t}}$$

goes arbitrarily close to $V_D(h)$ with probability 1.

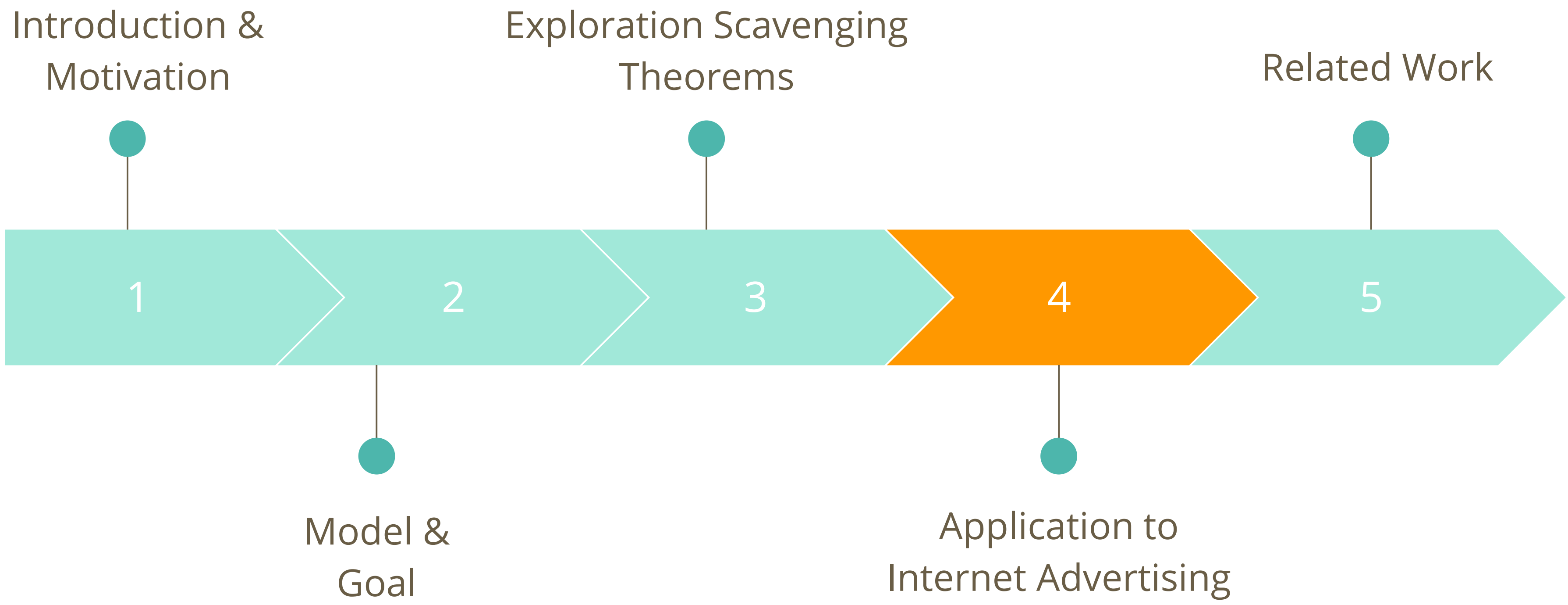
Multiple Exploration Policies

- The results we have discussed require the exploration policy to choose actions independent of the current input, which is very limiting
- There exist some special cases when exploration data can prove to be useful even when the exploration policy depends on the context

One such scenario : Multiple Exploration Policies

Multiple Exploration Policies

- Suppose we have collected data from a system that has rotated through K known exploration policies $\pi_1, \pi_2, \ldots, \pi_K$ over time
- Each policy π_i may depend on the context, but the choice of picking a policy at any given time may not
- Can redefine the action of the bandit problem as a choice of following one of the K policies, i.e. policy h's action is to choose which of $\pi_1, \pi_2, \ldots, \pi_K$ to follow
- Since historically the decision to choose amongst $\pi_1, \pi_2, \ldots, \pi_K$ was content independent, Theorem 3 holds
- Here, h can make a context dependent decision about which policy to follow, potentially achieving better performance than any single policy


Introduction &
Motivation
Exploration Scavenging
Theorems
Related Work
1
2
3
4
5
Model &
Goal
Application to
Internet Advertising

Application to Internet Advertising

- Technology companies are interested in finding better ways to search over the increasingly large selection of potential ads to display
- Evaluating ad-serving policies can be costly
    - This cost grows linearly with the number of candidate policies
- We can tackle the problem of evaluating a new ad-serving policy using data logged from an existing system using exploration scavenging

Internet Advertising as a Contextual Bandit Problem

- Each time a user visits a web page, an advertising engine places a limited number of ads in a slate on the page
    - Slate of ads has a limited number of selected ads
    - Every ad is placed on a specific position, selected by the algorithm
- Online advertising problem can be mapped to contextual bandit problem
    - Choosing an ad or set of ads to display corresponds to choosing an arm to pull.
    - Content of the web page provides context
- For our problem, we have, reward as a bit vector that identifies whether or not each returned ad was clicked

The Direct Approach

- The bit vector can be converted to a single real-valued reward r in a number of ways, for instance by summing the components and normalizing
- The we compute $r\frac{I(h(x)=s)}{\text{count}(s)}$, where
  - s is a slate of ads
  - count(s) is the number of times the state s was displayed during all trials
- Summing this quantity over all trials yields a good estimator of the value of the new policy h

Drawbacks of The Direct Approach

- For a large set of ads and a large slate size, the number of possible slates is very large
- Due to the indicator variable the contribution to the sum for a single example is zero unless same slate is chosen by h and the current system π
- But because of the large number of possible slates it's unlikely that the same state is chosen many times
- Therefore, the resulting estimator has a large variance

The Factoring Assumption

- The above problem can be avoided by making the following assumption
- **Assumption:** Probability of clicking can be decomposed into two terms, an intrinsic click-through rate (CTR) that depends only on the web page x and the ad a, and a position-dependent multiplier C_i for position i, called the attention decay coefficient (ADC)
- Formally: We assume that $\mathcal{P}(x, a, i) = C_i \mathcal{P}(x, a)$ where
  - $\mathcal{P}(x, a, i)$ is the probability that ad a is clicked when placed at position i, on web page x,
  - $\mathcal{P}(x, a)$is the position independent click through rate
  - C_i is the position dependent constant (ADC). We have, $C_1 = 1$, so $\mathcal{P}(x, a, 1) = \mathcal{P}(x, a)$

The Factoring Assumption

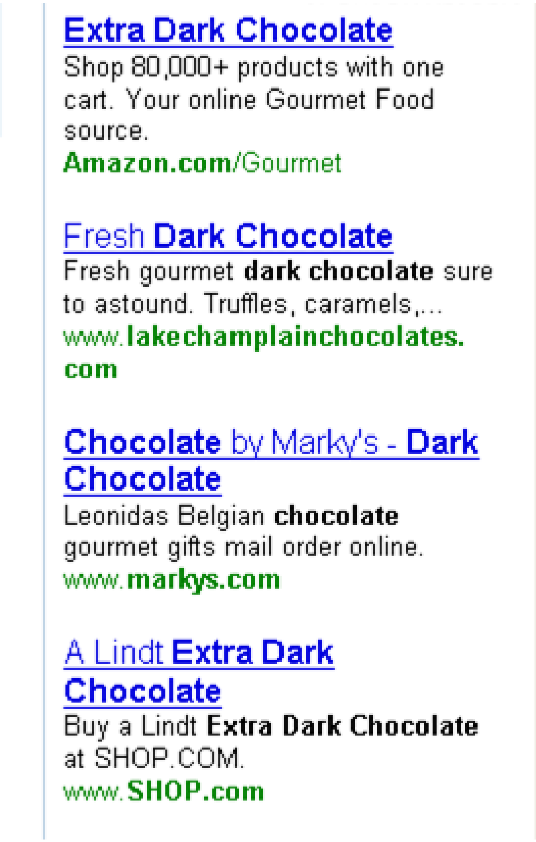
Probability of being clicked → $\boldsymbol{P(x, a, i)}$

=

Position Independent Click Through Rate → $\boldsymbol{P(x, a)}$

*

Position Dependent Multiplier (ADC) → $\boldsymbol{C_i}$



Attention Decay Coefficient (C_i) Decreases

The Factoring Assumption

- The assumption allows us to transition from dealing with slates of ads to individual ads

- For a slate $(x, s, \vec{r})$, with l ads, we can form l examples of the form $\left(x, a_{i,} r_i'\right)$
  - where a^i is the i^{th} ad in the slate
  - $r_i' = \frac{r_i}{C_i}$, where $r_i = 1\ or\ 0$

- We will now define a new estimator.

- Let $\sigma(a, x)$ be the slot in which the evaluation policy h places ad a on input x

The Factoring Assumption

- If h does not display a on input x, then $\sigma(a, x) = 0$. We define $C_0 = 0$.
- Define a new estimator of the value of h as

$$\hat{V}_D(h) = \sum_{t=1}^{T} \sum_{i=1}^{l} \frac{r_i' C_{\sigma(a,x)}}{T_{a_i}}$$

- Where T_{a_i} is the total number of times ad a is displayed and l be the number of ads shown in a slate

The Factoring Assumption

- Here, $C_{\sigma(a_i,x)}$ takes place of the indicator
    - $C_{\sigma(a_i,x)}$ is zero when h does not place a on the page x
    - It gives higher weights to the reward of an ad that h places in a better slot
- This estimator is consistent as long as the current ad-serving policy does not depend on the input webpage x and every ad is displayed enough
- We require the knowledge of the ADCs to use the above. We will now discuss how to estimate them

Estimating Attention Decay Coefficients (ADCs)

- Assume that a data set S includes observations $(x_t, \overrightarrow{a_t}, \overrightarrow{r_{t,a_t}})$, for a policy π that chooses the t^{th} slate of ads to display independent of the input x_t, for $t = \{1,2,3 \ldots T\}$, $\overrightarrow{a_t}$ is the slate of ads displayed at time t and $\overrightarrow{r_{t,a_t}}$ is the reward vector.

- Let $C(a, i)$ be the number of clicks on ad a observed during rounds in which it is displayed in position i, and $M(a, i)$be the number of impressions of a in slot i. Finally, $CTR(a, i) = \frac{C(a,i)}{M(a,i)}$, be the observed CTR of ad a in slot i

The Naive Estimator

- One might think that the ADCs can be calculated by taking the ratio between the global empirical click-through rate for each position i and the global empirical click-through rate for position 1

$$Est_{naive}(i) = \frac{\sum_a C(a,i) \,/\, \sum_a M(a,i)}{\sum_a C(a,1) / \sum_a M(a,1)}$$

- Unfortunately, this method has a bias which is often quite large in practice
- It underestimates the ratios C_i due to the fact that existing policies generally already place better ads (with higher $\mathcal{P}(x,a)$) in the better slot

A New Estimator

- For a fixed ad a and a fixed position i, it is possible to estimate the probability of a being clicked in position i fairly accurately, if it happens sufficiently many times. Similarly, for ad a in position 1.

- We may estimate C_i as $C_i = \frac{E_{x \sim D}[\mathcal{P}(x,a,i)]}{E_{x \sim D}[\mathcal{P}(x,a,1)]}$. If we do this for all ads, we can average the resulting estimates to form a single estimate.

$$EST_{\overrightarrow{\alpha_t}}(i) = \frac{\sum_a \alpha_a CTR(a,i)}{\sum_a \alpha_a CTR(a,1)}$$

- where $\vec{\alpha}$ is a vector of nonnegative constants α_a for each ad $a \in A$.

Consistency of the Estimator

Theorem 6: If the ad-display policy chooses slates independent of input and $\vec{\alpha}$ has all positive entries, then the estimator $Est_{\vec{\alpha}}$ in Equation 4 is consistent.

- Next question, How do we choose the values for α
- If every component of $\vec{\alpha}$ is set to the same value, then the estimate for C_i can be viewed as the mean of all estimates of C_i for each ad
- If the estimates for certain ads are more accurate than others, we'd like to weight those more heavily
- We want to pick $\vec{\alpha}$ to minimize the variance of our final estimator

Minimizing the Variance

Theorem 7: The variance of the expression

$$\sum_a \alpha_a CTR\,(a,i) + \sum_a \alpha_a CTR(a,1)$$

subject to $\sum_a \alpha_a = 1$ is minimized when

$$\alpha_a := \frac{2\,M(a,i).\,M(a,1)}{M(a,i)\sigma_{a,i}^2 + M(a,1)\sigma_{a,1}^2}$$

where $\sigma_{a,i}^2$ is the variance of the indicator random variable that is 1 when ad a is clicked given that ad a is placed in position i

A New Estimator

- Most current ad serving algorithm violate the assumption that the policy has to be independent of the web page.
  - Exploration scavenging is no longer guaranteed to work.
- Luckily, in practice, it is generally not the case that extreme scenarios like the counterexample in the proof of Theorem 1 arise.
- It is more likely that the algorithms choose among the same small set of ads to display for any given context
- In practise, major difference is the order in which these ads are displayed

Empirical comparison

- A common technique for estimating ADCs borrowed from the information retrieval literature is discounted cumulative gain
- Given parameter b, DCG would suggest defining $C_i = 1/\log_b(b+i)$ for all i
- The coefficients discussed below were computed from training on about 20 million examples obtained from the logs of "Content Match", Yahoo!'s online advertisement engine
- For the new estimator we use $\alpha_a = M(a,p)M(a,1)/(M(a,p)+M(a,1))$

Empirical comparison

- The following table summarizes the coefficients computed for the first four slots using the naive estimator and the new estimator, and the DCG

	c_1	c_2	c_3	c_4
Naive	1.0	0.512090	0.369638	0.271847
New	1.0	0.613387	0.527310	0.432521
DCG	1.0	0.630930	0.5	0.430677

- As suspected, the coefficients for the new estimator are larger than the old, suggesting a reduction in bias

Towards A Realistic Application

- Unfortunately the new estimator may still have an unacceptably large variance

- The method only benefits from examples in which the exploration policy and the new policy h choose overlapping sets of ads to display
    - A rare event in large databases

- Instead, consider policies h_π to reorder the ads chosen by π

- A good reordering policy plausibly provides useful information to guide the choice of a new ranking policy.

Towards A Realistic Application

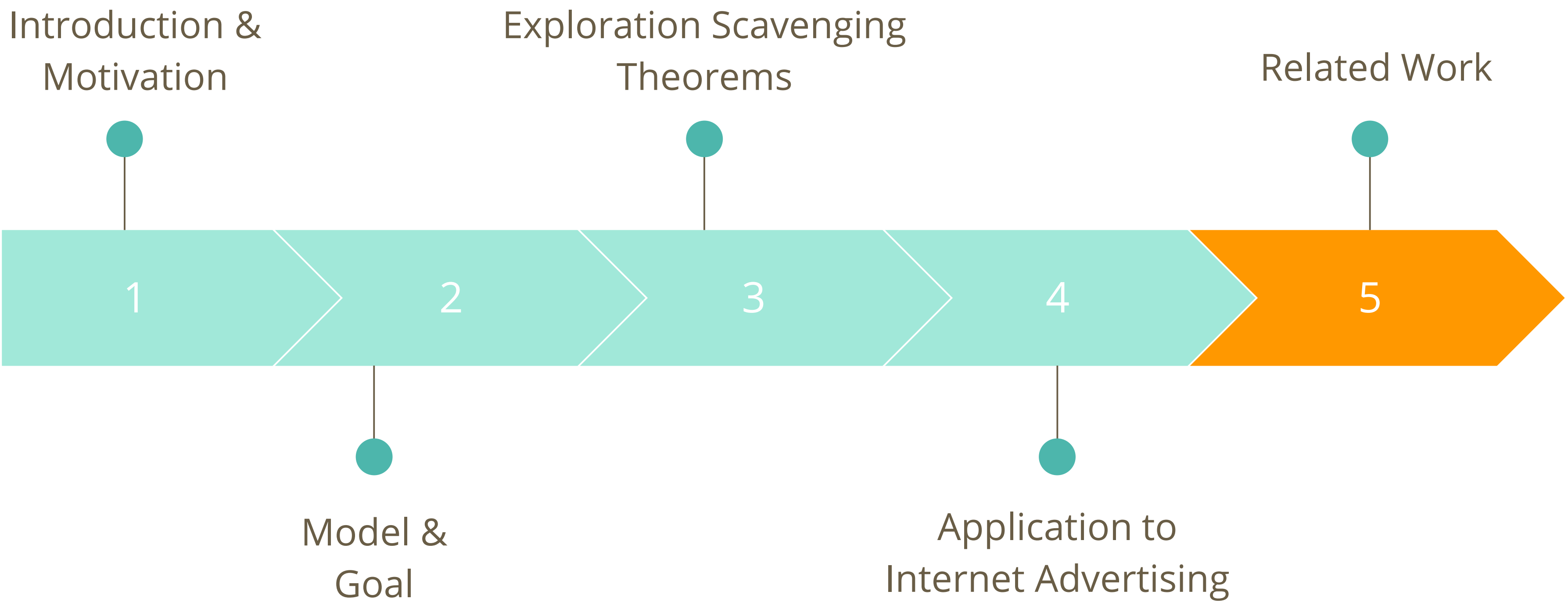
We define a new estimator

$$\widehat{V_D}(h_\pi) = \sum_{t=1}^{T} \sum_{i=1}^{l} r_i' C_{\sigma'(a_i, x)}$$

- Where $\sigma'(a_i, x)$ is the slot that h_π would assign to ad a_i in this new model.
- This approach has small variance and quickly converges

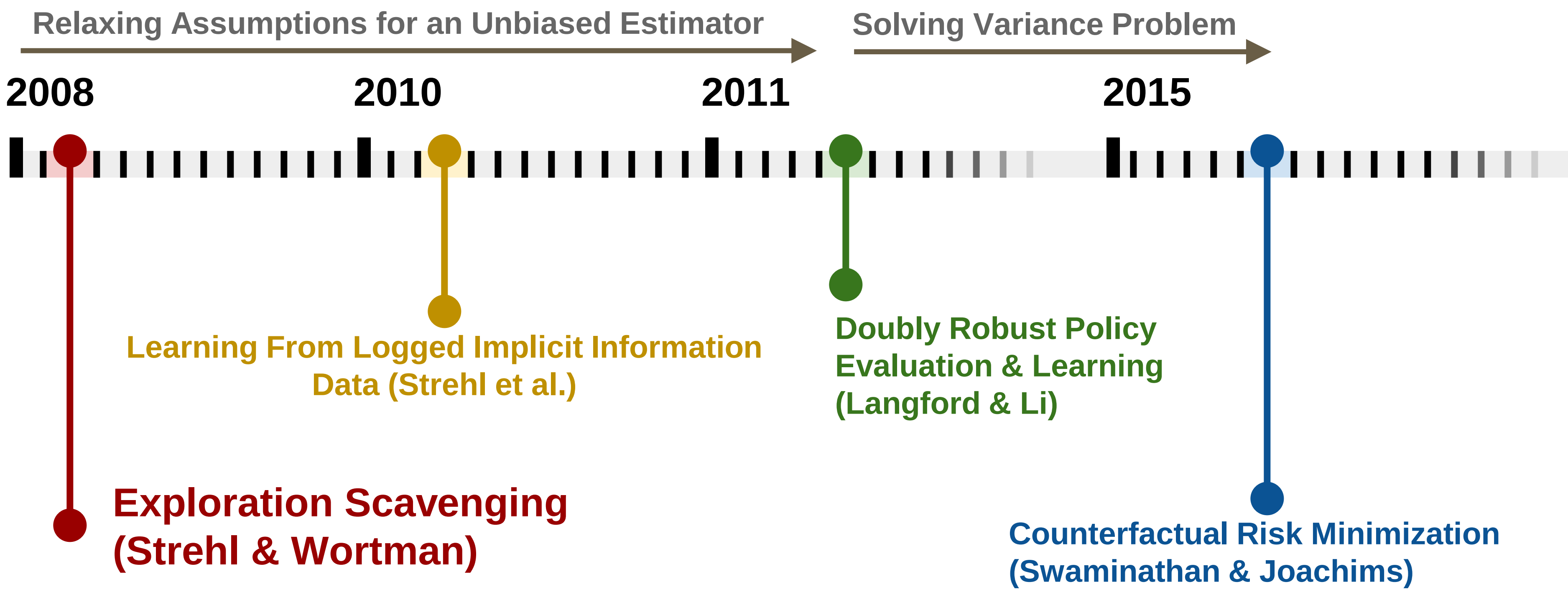
Results

- To illustrate the method a training set of 20 million examples gathered using Yahoo!'s current ad serving algorithm π is used
- We let the policy h_π be the policy that reorders ads to display those with the highest empirical click-through rate first, ignoring the context x.
- This policy was then compared to policy $\mathrm{h'}_\pi$ which reorders ads at random
- Number of clicks we expect the new policies to receive per click of the old policy π were computed in the two cases for comparison, which we call r
- **Result:** For $h_{\pi,}\ r = 1.086$ and for $h'_\pi, r = 1.016$
  - Thus, exploration scavenging strongly suggests using policy h_π over h'_π



Introduction & Motivation
1
Model & Goal
2
Exploration Scavenging Theorems
3
Application to Internet Advertising
4
Related Work
5

Literature Timeline



Learning from Logged Implicit Exploration Data

- Same setup as the previous paper – Contextual bandit problem
- But remove the tight assumption that the exploration policy π can only take context independent actions
- Goal is to maximize the sum of rewards r_a over the rounds of interaction
  - Use previously recorded events to form a good policy on the first round of interaction
- Formally, given a dataset $S = (x, a, r_a)^*$ generated by the interaction of an uncontrolled logging policy, address the problem of constructing a policy h which tries to maximize $V(h) = E_{(x,\vec{r})\sim D}[r_{h(x)}]$

Approach

- For each event (x, a, r_a), estimate the probability $\hat{\pi}(a|x)$ that the logging policy chooses a, using regression
- For each (x, a, r_a), create a synthetic controlled contextual bandit setting according to $(x, a, r_a, 1/\max\{\hat{\pi}(a|x), \tau\})$
  - $1/\max\{\hat{\pi}(a|x), \tau\}$ is an importance weight that specifies how important the current event is for training
- Apply an offline contextual bandit algorithm to these generated events to evaluate the performance of any hypothesis h

$$\hat{V}_{\hat{\pi}}^{h}(S) = \frac{1}{|S|} \sum_{(x,a,r)\in S} \frac{r_a I(h(x) = a)}{\max\{\hat{\pi}(a|x), \tau\}}$$

  - Evaluation works well as long as the actions chosen by h have adequate support over π and $\hat{\pi}$ is a good estimate for π
- Using the evaluation model, find the best hypothesis $\hat{h}$

Double Robust Policy Evaluation and Learning

- Propose a method of policy evaluation that is more robust compared to earlier methods
- Main problem with policy evaluation is that we cannot directly simulate our policy over the data set and we have only partial information about the reward
- Two approaches to overcome this limitation, direct method (DM) and inverse propensity score (IPS)
- We first introduce these methods, which were used in the previous two papers

Direct Method

- **Direct Method:** Form an estimate $\hat{\rho}_a(x)$ of the expected reward conditioned on the context and action, $\rho_a(x) = E_{(x,\vec{r})\sim D}[r_a|x]$. The policy value is estimated by

$$\hat{V}^h{}_{DM} = \frac{1}{|S|}\sum_{x\in S}\hat{\rho}_{h(x)}(x)$$

- If $\hat{\rho}_a(x)$ is a good estimate of the true expected reward, $\rho_a(x)$, then the DM estimate is close to V^h
- Problem is that the estimate $\hat{\rho}$ is formed without the knowledge of h and hence might focus on approximating ρ mainly in areas that are not relevant for V^h and not sufficiently in the areas that are important for V^h
- So, DM suffers from problems with bias

Inverse Propensity Score

- **Inverse Propensity Score:** Instead of approximating the reward, IPS forms an approximation $\hat{p}(a|x,\pi)$ of $p(a|x,\pi)$, and uses this estimate to correct for the shift in action proportions between the data collection policy and the new policy

$$\hat{V}^h{}_{IPS} = \frac{1}{|S|} \sum_{(x,\pi,a,r_a)\in S} \frac{r_a I(h(x) = a)}{\hat{p}(a|x,\pi)}$$

- If $\hat{p}(a|x,\pi) \approx p(a|x,\pi)$, then the IPS estimate will be approximately an unbiased estimate of V^h
- Since we typically have a good (or even accurate) understanding of the data collection policy, it is easier to get a good estimate $\hat{p}$ making IPS less susceptible to problems with bias
- However, due to the range of the random variable increasing, suffers from variance issues, with the problem getting exacerbated when $p(a|x,\pi)$ gets smaller

Approach

- Doubly Robust estimators take advantage of both, the estimate of the expected reward $\hat{\rho}_a(x)$ and the estimate of the action probabilities $\hat{p}(a|x,\pi)$

$$\hat{V}^h{}_{DR} = \frac{1}{|S|} \sum_{(x,\pi,a,r_a)\in S} \left[\frac{(r_a - \hat{\rho}_a(x)) I(h(x) = a)}{\hat{p}(a|x,\pi)} + \hat{\rho}_{h(x)}(x) \right]$$

- Informally, the estimator uses $\hat{\rho}$ as a baseline and if there is data available, a correction is applied
- It is shown that this estimator is accurate if at least one of the estimators, $\hat{\rho}$ and $\hat{p}$ is accurate, hence the name doubly robust

Counterfactual Risk Minimization

- Uses clipped version of the IPS and regularization for reducing variance

$$\hat{h}^{CRM} = \arg\min_{h \in H}\{R^M(h) + \lambda\sqrt{\left(\frac{Var_h(u)}{n}\right)}\}$$

  - $R^M(h)$ is the clipped version of IPS
  - Where the second term serves as a data-dependent regularizer
  - Var is defined in terms of $M, h, p_i, \delta_i and\ n$
- The results in the paper show that CRM is beneficial. They have derived a learning algorithm called POEM (Policy Optimizer for Exponential Models) for structured output prediction which is shown to work better than IPS.

References

- Langford, John, Alexander Strehl, and Jennifer Wortman. "Exploration scavenging." *Proceedings of the 25th international conference on Machine learning*. ACM, 2008.
- Strehl, Alex, et al. "Learning from logged implicit exploration data." *Advances in Neural Information Processing Systems*. 2010.
- Dudík, Miroslav, John Langford, and Lihong Li. "Doubly robust policy evaluation and learning." *arXiv preprint arXiv:1103.4601* (2011).
- Swaminathan, Adith, and Thorsten Joachims. "Counterfactual Risk Minimization." *Proceedings of the 24th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2015.