Exploration & Exploitation

in Computational Advertising

Hulu ADI

Zhe Wang

Outline

- What is E&E?
- Main Approaches
- E&E applications in Computational Advertising
- Summary

What is E&E?

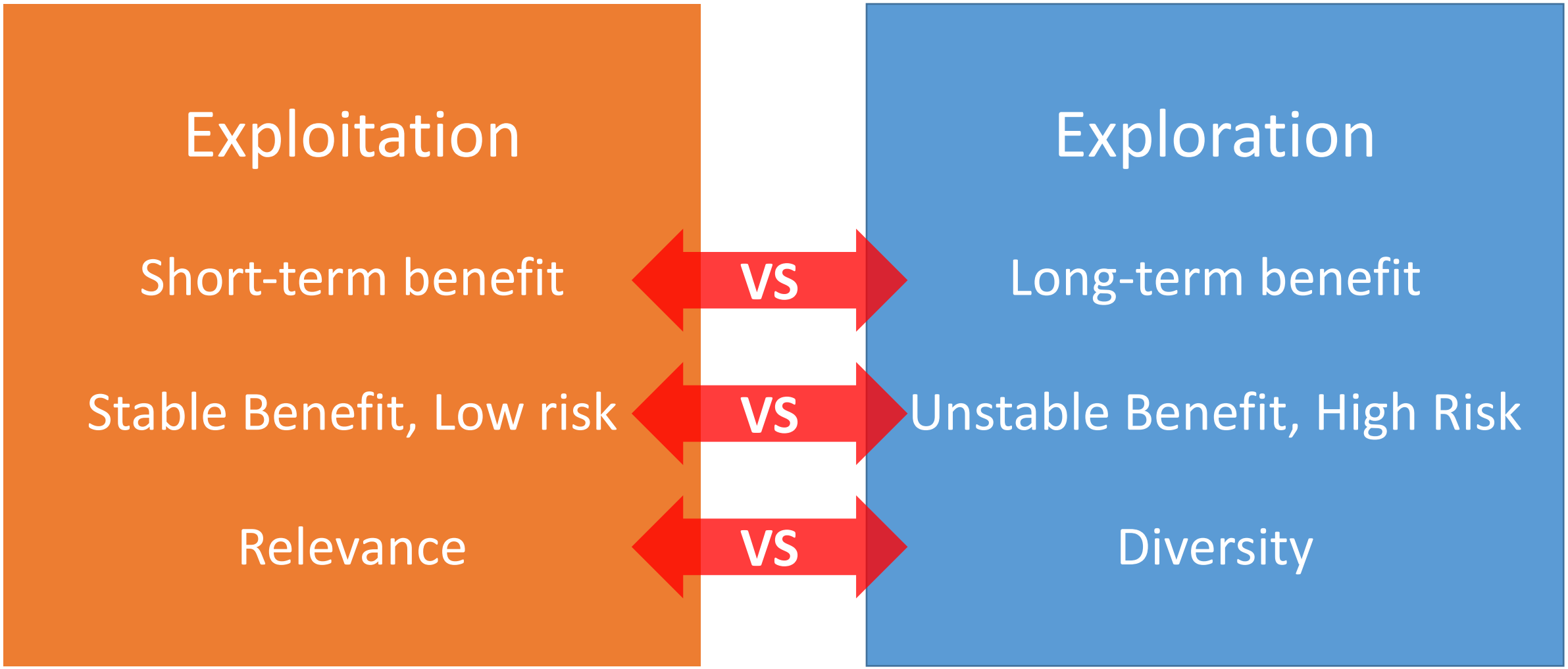
Exploitation

Maximize its expected revenue according to its current knowledge in short term

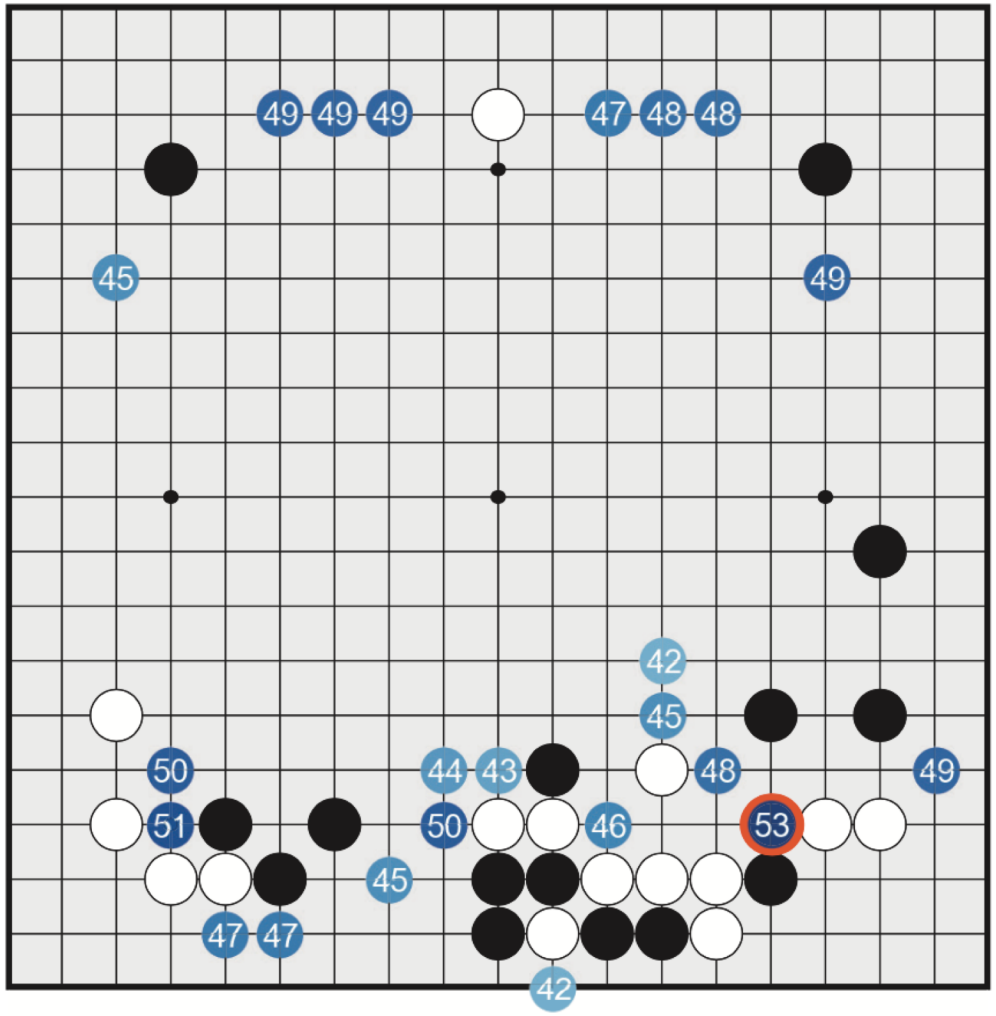
Exploration

learn more about the unknown to improve its knowledge, since the latter might increase its revenue in long term

What is E&E?


Exploitation
Exploration
Short-term benefit
VS
Long-term benefit
Stable Benefit, Low risk
VS
Unstable Benefit, High Risk
Relevance
VS
Diversity

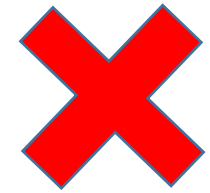
E&E Example – AlphaGo



How to select next move from several candidate locations?

Benefit VS Risk

E&E Example – Performance Ads Optimization

hulu	1,000,000 impressions	10,000 clicks	CTR=1%	
NETFLIX	1,000,000 impressions	1,000 clicks	CTR=0.1%	
YouTube	10,000 impressions	10 clicks	CTR=0.1%	

E&E Example – Multi-armed Bandits

Definition of Multi-armed Bandits Problem

Real Reward Distributions $B = \{R_1, \ldots, R_K\}$

Expected Reward $\mu_1, \ldots, \mu_K$

Maximal Expected Reward $\mu^* = \max_k \{\mu_k\}$

Regret $\rho = T\mu^* - \sum_{t=1}^{T} \hat{r}_t$

Main Approaches – Epsilon greedy

$1-\epsilon$ Exploit, select the best arm

ϵ Explore, randomly select a arm

Main Approaches – UCB(Upper Confidence Bound)

UCB1:

Play each of the K actions once, giving initial values for empirical mean payoffs $\overline{x}_i$ of each action i.

For each round $t = K, K+1, \ldots$:

Let n_j represent the number of times action j was played so far.

Play the action j maximizing $\overline{x}_j + \sqrt{2 \log t / n_j}$. ?

Observe the reward $X_{j,t}$ and update the empirical mean for the chosen action.

Main Approaches – UCB(Upper Confidence Bound)

Chernoff - Hoeffding Inequality

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n). \quad 0 \le X_i \le 1$$

$$\mathbb{P}(\overline{X} - \mathbf{E}[\overline{X}] \ge t) \le e^{-2nt^2}$$

$$\sqrt{2\log t/n_j}$$

$$\mathbb{P}(\overline{X} - \mathbf{E}[\overline{X}] \ge \sqrt{2\log t/n_j}\,) \le \; t^{-4}$$

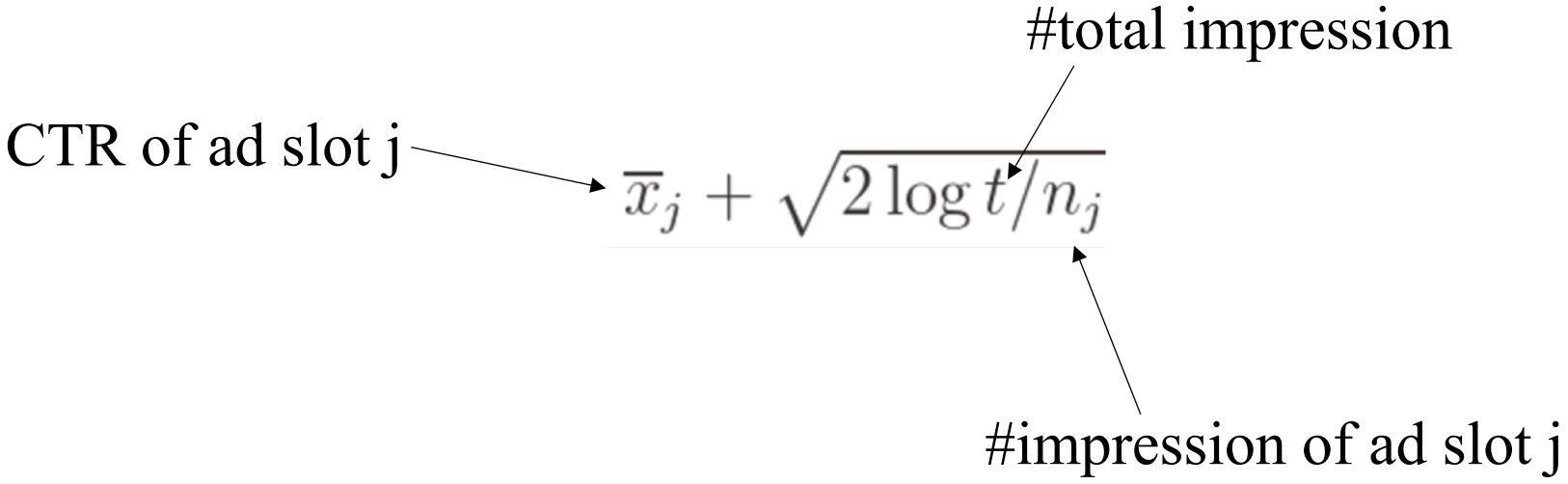
Main Approaches – UCB

Theorem: Suppose UCB1 is run as above. Then its expected cumulative regret $\mathbb{E}(R_{\mathrm{UCB1}}(T))$ is at most

$$8 \sum_{i:\mu_i<\mu^*} \frac{\log T}{\Delta_i} + \left(1 + \frac{\pi^2}{3}\right) \left(\sum_{j=1}^{K} \Delta_j\right)$$

E&E Application – CTR Exploration

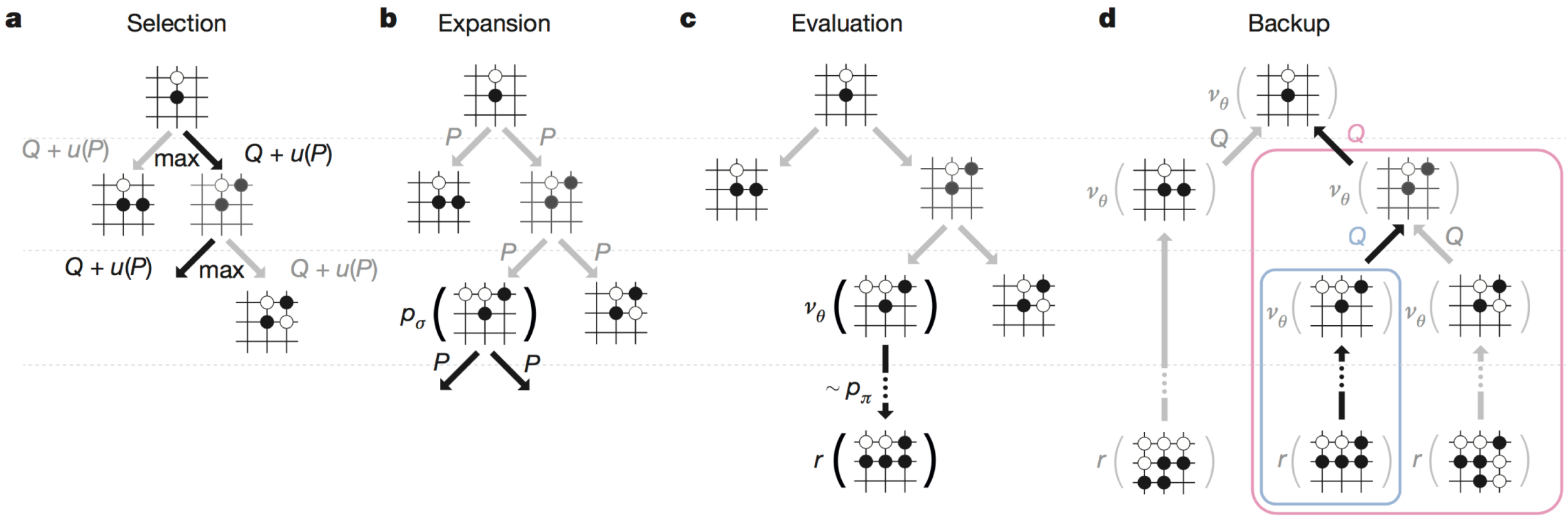
- Explore more high CTR ad slots, and more accurate CTR prediction
  - Reward: Click
  - Arm: Each Ad Slot
  - Action j: Impression on ad slot j



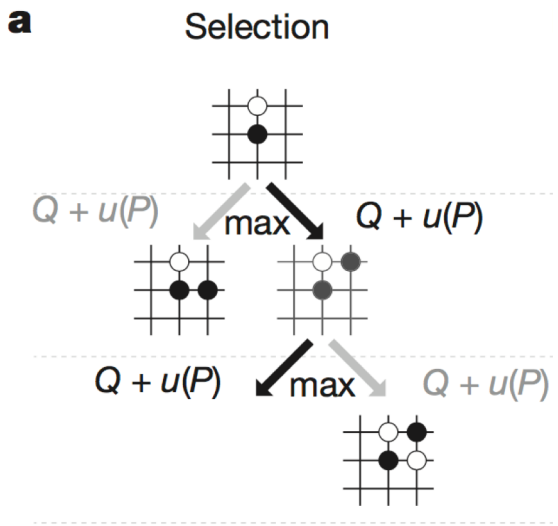
Main Approaches – UCB extension

- UCB1-Normal
- UCB1-Tuned
- UCB2
- UCT(Upper Confidence Bound Apply to Tree)
  - UCB+MCTS(Monte Carlo Tree Search)

Main Approaches – UCT



Main Approaches – UCT



$$a_t = \text{argmax}(Q(s_t, a) + u(s_t, a))$$

$$u(s, a) \propto \frac{P(s, a)}{1 + N(s, a)}$$

$$N(s, a) = \sum_{i=1}^{n} 1(s, a, i)$$

$$Q(s, a) = \frac{1}{N(s, a)} \sum_{i=1}^{n} 1(s, a, i) V(s_L^i)$$

Main Approaches – Thompson Sampling

Algorithm 1 Thompson sampling

$D = \emptyset$
for $t = 1, \ldots, T$ **do**
  Receive context x_t
  Draw θ^t according to $P(\theta|D)$
  Select $a_t = \arg\max_a \mathbb{E}_r(r|x_t, a, \theta^t)$
  Observe reward r_t
  $D = D \cup (x_t, a_t, r_t)$
end for

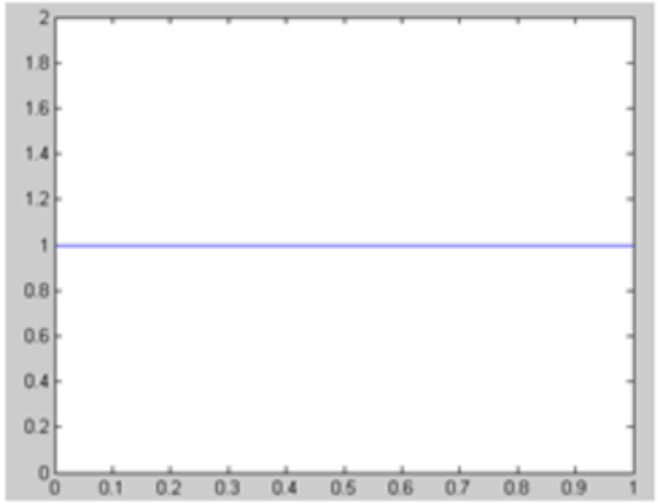
Main Approaches – Thompson Sampling

Algorithm 2 Thompson sampling for the Bernoulli bandit

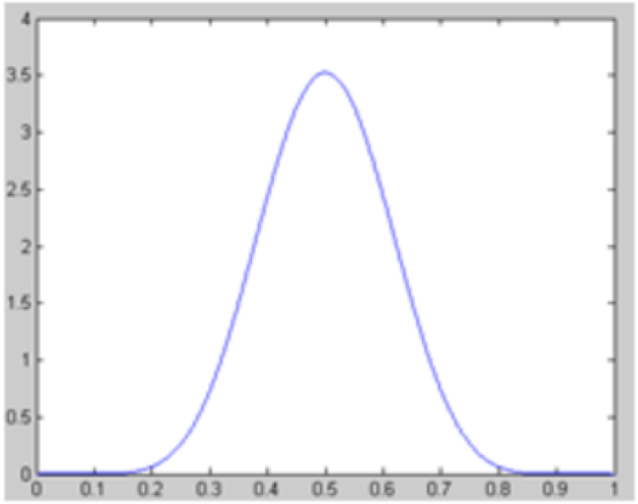
Require: α, β prior parameters of a Beta distribution
  $S_i = 0, F_i = 0, \quad \forall i$. {Success and failure counters}
  for $t = 1, \ldots, T$ **do**
    for $i = 1, \ldots, K$ **do**
      Draw θ_i according to $\text{Beta}(S_i + \alpha, F_i + \beta)$. ?
    end for
    Draw arm $\hat{\imath} = \arg\max_i \theta_i$ and observe reward r
    if $r = 1$ **then**
      $S_{\hat{\imath}} = S_{\hat{\imath}} + 1$
    else
      $F_{\hat{\imath}} = F_{\hat{\imath}} + 1$
    end if
  end for

Main Approaches – Beta Distribution Examples

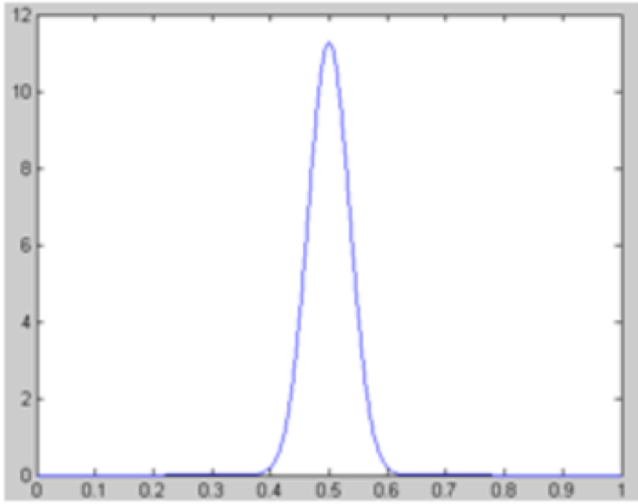
α=1,β=1



α=10,β=10



α=100,β=100



Main Approaches – Thompson Sampling

- Why Beta distribution?
  - the beta distribution is the conjugate prior probability distribution for the Bernoulli

$p(D|\vartheta) = \vartheta^{N_1}(1-\vartheta)^{N_0}$ Bernoulli likelihood

$p(\vartheta) = \frac{1}{B(a,\ b)}\vartheta^{a-1}(1-\vartheta)^{b-1}$ Beta(a,b)

$$p(\vartheta|\mathrm{D}) = \frac{p(D|\vartheta) * p(\vartheta)}{p(D)} \propto p(D|\vartheta) * p(\vartheta) = \vartheta^{N_1}(1-\vartheta)^{N_0} * p(\vartheta)$$

$p(\vartheta|\mathrm{D})$=Beta(a+N_1,b+N_0)

Main Approaches – Thompson Sampling

$$
\begin{aligned}
f(x;\alpha,\beta) &= \text{constant} \cdot x^{\alpha-1}(1-x)^{\beta-1} \\
&= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1} \\
&= \frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}
\end{aligned}
$$

E&E Application - Programmatic Creative Optimization

Which Creative is the Best?

Main Approaches – Contextual Bandit Approach

- How to integrate Explore and Exploit with contextual model?
  - Content
  - Topic
  - Labels / Attributes
  - Demographic
  - User Historical Behavior
  - Etc.

Main Approaches – Contextual Bandit Approach

- How to integrate Explore and Exploit with contextual model?
  - Content
  - Topic
  - Labels / Attributes
  - Demographic
  - User Historical Behavior
  - Etc.

d-dimensional feature $\mathbf{x}_{t,a}$

coefficient vector $\boldsymbol{\theta}_a^*$

$$\mathbf{E}[r_{t,a}|\mathbf{x}_{t,a}] = \mathbf{x}_{t,a}^{\top}\boldsymbol{\theta}_a^* \quad ?$$

Main Approaches – LinUCB

$$\mathbf{E}[r_{t,a}|\mathbf{x}_{t,a}] = \mathbf{x}_{t,a}^{\top}\boldsymbol{\theta}_a^*$$?

UCB $\overline{x}_j + \sqrt{2\log t / n_j}$

LinUCB $p_{t,a} \leftarrow \hat{\boldsymbol{\theta}}_a^{\top}\mathbf{x}_{t,a} + \alpha\sqrt{\mathbf{x}_{t,a}^{\top}\mathbf{A}_a^{-1}\mathbf{x}_{t,a}}$

? ?

Main Approaches – LinUCB

LinUCB

$$p_{t,a} \leftarrow \hat{\boldsymbol{\theta}}_a^\top \mathbf{x}_{t,a} + \alpha\sqrt{\mathbf{x}_{t,a}^\top \mathbf{A}_a^{-1} \mathbf{x}_{t,a}}$$

training data $(\mathbf{D}_a, \mathbf{c}_a)$

Ridge Regression

$$\hat{\boldsymbol{\theta}}_a = (\mathbf{D}_a^\top \mathbf{D}_a + \mathbf{I}_d)^{-1} \mathbf{D}_a^\top \mathbf{c}_a$$

Main Approaches – LinUCB

LinUCB

$$p_{t,a} \leftarrow \hat{\boldsymbol{\theta}}_a^\top \mathbf{x}_{t,a} + \alpha\sqrt{\mathbf{x}_{t,a}^\top \mathbf{A}_a^{-1} \mathbf{x}_{t,a}}$$

$$P\left(\left|\mathbf{x}_{t,a}^\top \hat{\boldsymbol{\theta}}_a - \mathbf{E}[r_{t,a}|\mathbf{x}_{t,a}]\right| \leq \alpha\sqrt{\mathbf{x}_{t,a}^\top (\mathbf{D}_a^\top \mathbf{D}_a + \mathbf{I}_d)^{-1} \mathbf{x}_{t,a}}\right) > 1 - \delta$$

$$\alpha = 1 + \sqrt{\ln(2/\delta)/2}$$

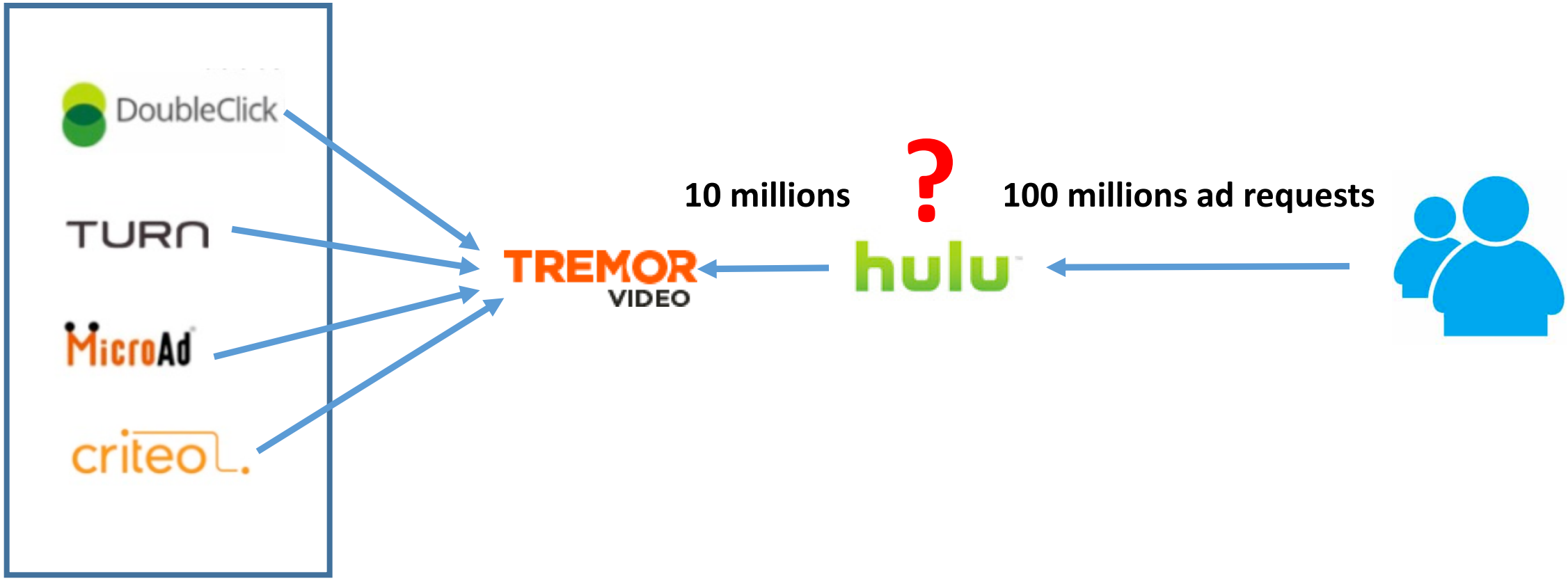
$$\mathbf{A}_a \stackrel{\text{def}}{=} \mathbf{D}_a^\top \mathbf{D}_a + \mathbf{I}_d$$

Main Approaches – LinUCB

Algorithm 1 LinUCB with disjoint linear models.

Inputs: $\alpha \in \mathbb{R}_+$
for $t = 1, 2, 3, \ldots, T$ **do**
  Observe features of all arms $a \in \mathcal{A}_t$: $\mathbf{x}_{t,a} \in \mathbb{R}^d$
  for all $a \in \mathcal{A}_t$ **do**
    if a is new **then**
      $\mathbf{A}_a \leftarrow \mathbf{I}_d$ (d-dimensional identity matrix)
      $\mathbf{b}_a \leftarrow \mathbf{0}_{d \times 1}$ (d-dimensional zero vector)
    end if
    $\hat{\boldsymbol{\theta}}_a \leftarrow \mathbf{A}_a^{-1} \mathbf{b}_a$
    $p_{t,a} \leftarrow \hat{\boldsymbol{\theta}}_a^\top \mathbf{x}_{t,a} + \alpha \sqrt{\mathbf{x}_{t,a}^\top \mathbf{A}_a^{-1} \mathbf{x}_{t,a}}$
  end for
  Choose arm $a_t = \arg\max_{a \in \mathcal{A}_t} p_{t,a}$ with ties broken arbitrarily, and observe a real-valued payoff r_t
  $\mathbf{A}_{a_t} \leftarrow \mathbf{A}_{a_t} + \mathbf{x}_{t,a_t} \mathbf{x}_{t,a_t}^\top$
  $\mathbf{b}_{a_t} \leftarrow \mathbf{b}_{a_t} + r_t \mathbf{x}_{t,a_t}$
end for

E&E Applications – Auction Traffic Allocation



Recent Research

•**NeuralBandit algorithm**: In this algorithm several neural networks are trained to modelize the value of rewards knowing the context, and it uses a multi-experts approach to choose online the parameters of multi-layer perceptrons. (2014)

•**Bandit Forest algorithm**: a random forest is built and analyzed w.r.t the random forest built knowing the joint distribution of contexts and rewards. (2016)

•**Prior Sensitivity of Thompson Sampling**: One important benefit of Thompson Sampling is that it allows domain knowledge to be conveniently encoded as a prior distribution to balance exploration and exploitation more effectively. (2016)

•**Generalized Linear Contextual Bandits:** Propose an upper confidence bound based algorithm for generalized linear contextual bandits (2017)

•**Exploration and Intrinsic Motivation:** consider an agent's uncertainty about its environment and the problem of generalizing this uncertainty across states. (2016)

Summary

- **Exploitation and Exploration Tradeoff**

- **Main approaches**
  - Epsilon Greedy
  - UCB, UCT
  - Thompson Sampling
  - Contextual Bandit

- **Applications in Computational Advertising**

Thanks