

Contextual Bandits with Linear Payoff Functions

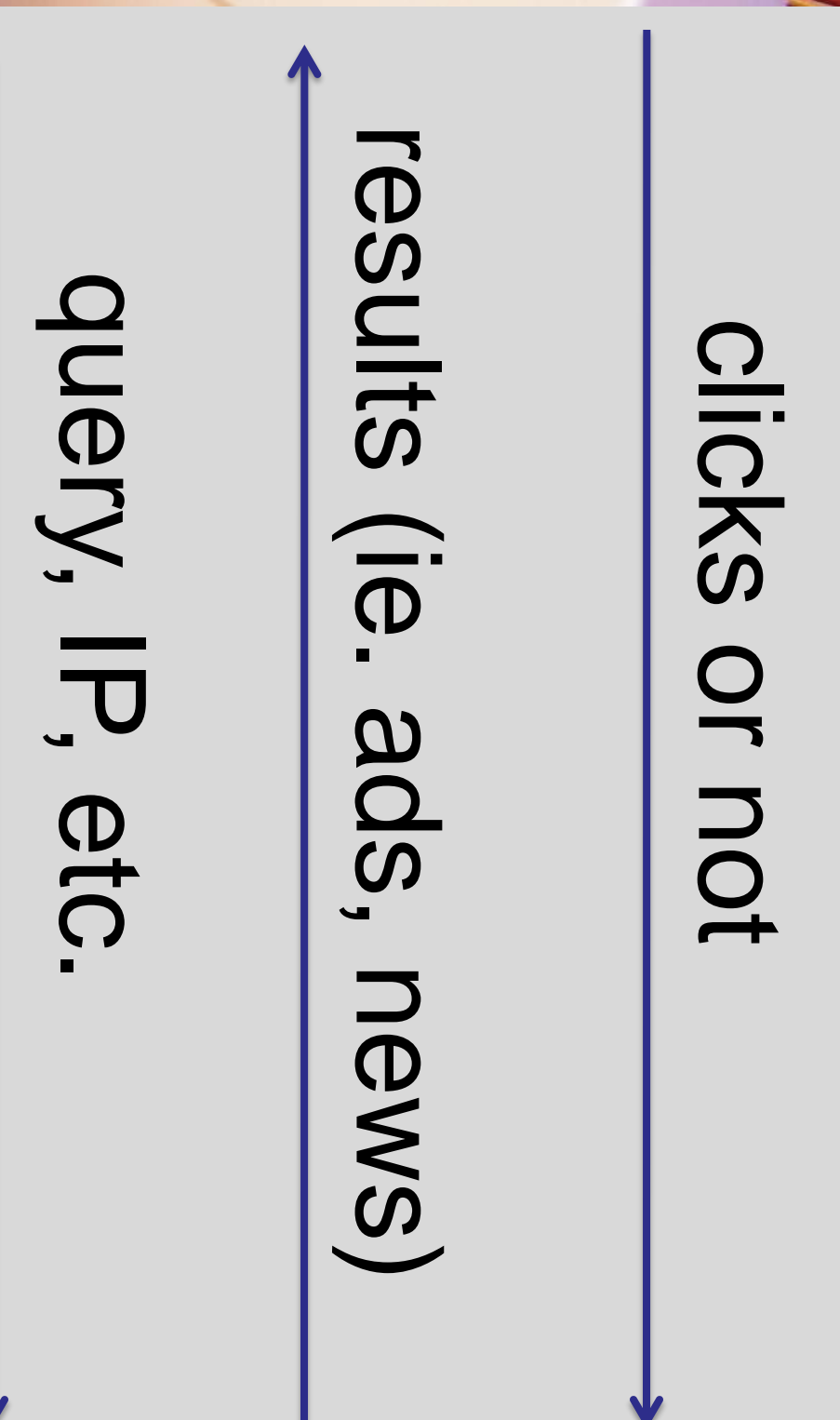
Wei Chu
Yahoo! Labs

Lihong Li
Yahoo! Research

Lev Reyzin
Georgia Institute of Technology

Robert E. Schapire
Princeton University

Motivation



Setting: Contextual Bandits with Linear Payoffs

Notation: T rounds, K actions, features in $x_{t,a}$ in \mathbb{R}^d (and $\|x_{t,a}\| \leq 1$), and action rewards $r_{t,a}$ in $[0,1]$

Game: for $t=1$ to T
world reveals $x_{t,1}, x_{t,2}, \dots, x_{t,K}$ (these are “action features”)
learner selects action a_t
learner receives reward r_{t,a_t} from the world

Assumption: exists an unknown vector θ^* , w/ $\|\theta^*\| \leq 1$ s.t. for all a and t ,
 $E[r_{t,a}|x_{t,a}] = x_{t,a}^\top \theta^*$.

The $r_{t,a}$ s are independent random variables with expectation $x_{t,a}^\top \theta^*$.

Goal: to minimize regret w.r.t. to the choice of $a_t^* = \operatorname{argmax}_a x_{t,a}^\top \theta^*$,

$$\text{regret} = \sum_{t=1}^T (r_{t,a^*} - r_{t,a_t})$$

LinUCB (Li et al. 2010)

Algorithm 1 LinUCB: UCB with Linear Hypotheses

```
0: Inputs:  $\alpha \in \mathbb{R}_+, K, d \in \mathbb{N}$ 
1:  $A \leftarrow I_d$  {The  $d$ -by- $d$  identity matrix}
2:  $b \leftarrow \mathbf{0}_d$ 
3: for  $t = 1, 2, 3, \dots, T$  do
4:    $\theta_t \leftarrow A^{-1}b$ 
5:   Observe  $K$  features,  $x_{t,1}, x_{t,2}, \dots, x_{t,K} \in \mathbb{R}^d$ 
6:   for  $a = 1, 2, \dots, K$  do
7:      $p_{t,a} \leftarrow \theta_t^\top x_{t,a} + \alpha \sqrt{x_{t,a}^\top A^{-1} x_{t,a}}$  {Computes upper confidence bound}
8:   end for
9:   Choose action  $a_t = \operatorname{argmax}_a p_{t,a}$  with ties broken arbitrarily
10:  Observe payoff  $r_t \in \{0, 1\}$ 
11:   $A \leftarrow A + x_{t,a_t} x_{t,a_t}^\top$ 
12:   $b \leftarrow b + x_{t,a_t} r_t$ 
13: end for
```

Previous Work

The contextual bandit with linear payoffs setting was introduced by Auer (2002). He proved a regret of $O(Td)^{1/2}$ for his algorithm LinRel.

Abe et al. (2003) gave a lower bound on regret of $\Omega(T^{3/4}K^{1/4})$ for this setting, and gave algs with regret $O(T^{3/4}K^{1/2})$ and $O(T^{4/5}K^{1/4})$.

Li et al. (2010) introduced LinUCB, a simpler and more efficient algorithm than LinRel, and showed its effectiveness in experiments.

Our Results

We prove a variant of LinUCB has regret

$$O\left(\sqrt{Td \ln^3(KT \ln(T) / \delta)}\right)$$

This gives evidence for the effectiveness of LinUCB.

We decompose LinUCB into algorithms SupLinUCB and BaseLinUCB using a trick from Auer (2002). This makes predicted rewards be independent rand vars.

We also give an almost-matching lower bound for this setting of $\Omega(T^{1/2}d^{1/2})$ for $d^2 \leq T$.

Open Problems

An analysis of the original (non-decomposed) LinUCB is still needed, and would be interesting, as LinUCB remains the simplest algorithm we know for this setting.

Can we get a $O(T^{1/2}d^{1/2})$ algorithm for this setting, (without the log factors)?

Can we say anything about the agnostic case, where we want to compete with the best θ^* , even if no vector θ^* predicts the expected reward?