Contextual Bandits with Linear Payoff Functions

Wei Chu
Yahoo! Labs
Santa Clara, CA, USA
chuwei@yahoo-inc.com

Lihong Li
Yahoo! Labs
Santa Clara, CA, USA
lihong@yahoo-inc.com

Lev Reyzin
Georgia Institute of Tech.
Atlanta, GA, USA
lreyzin@cc.gatech.edu

Robert E. Schapire
Princeton University
Princeton, NJ, USA
schapire@cs.princeton.edu

Abstract

In this paper we study the contextual bandit problem (also known as the multi-armed bandit problem with expert advice) for linear payoff functions. For T rounds, K actions, and d dimensional feature vectors, we prove an $O\left(\sqrt{Td\ln^3(KT\ln(T)/\delta)}\right)$ regret bound that holds with probability $1-\delta$ for the simplest known (both conceptually and computationally) efficient upper confidence bound algorithm for this problem. We also prove a lower bound of $\Omega(\sqrt{Td})$ for this setting, matching the upper bound up to logarithmic factors.

1 INTRODUCTION

In the contextual bandit problem, on each of T rounds a learner is presented with the choice of taking one of K actions. Before making the choice of action, the learner sees a feature vector associated with each of its possible choices. In this setting the learner has access to a hypothesis class, in which the hypotheses take in action features and predict which action will give the best reward. If the learner can guarantee to do nearly as well as the prediction of the best hypothesis in hindsight (to have low regret), the learner is said to successfully compete with that class.

In this paper, we study the contextual bandit setting with linear payoffs. This setting was introduced by Abe et al. [2003] and developed by Auer [2002]. In this contextual bandit setting, the learner competes with the set of all linear predictors on the feature vectors. The set of linear predictors is both expressive enough to yield good real-world performance, yet yields to a succinct representation that makes it a tractable case.

An example application for contextual bandits with linear payoffs is the Internet advertisement selection problem [Abe et al., 2003], where advertisement and webpage features are used to construct a linear function to predict the probability of a user clicking on a given advertisement. This setting has been used for other applications including making article recommendations on web portals [Agarwal et al., 2009, Li et al., 2010].

In this paper, we give a theoretical analysis of a variant of LinUCB, a natural upper confidence bound algorithm introduced and experimentally demonstrated to be effective by Li et al. [2010]. We use a technique first developed by Auer [2002] by decomposing LinUCB into two algorithms: BaseLinUCB and SupLinUCB (which uses BaseLinUCB as a subroutine). We then show a $O\left(\sqrt{Td\ln^3(KT\ln(T)/\delta)}\right)$ high-probability regret bound for SupLinUCB. Finally, a lower bound of $\Omega(\sqrt{Td})$ for the contextual bandit problem with linear payoffs are also given.

2 PREVIOUS WORK

In the traditional, non-contextual, multiarmed bandit problem, the learner has no access to arm features and simply competes with pulling the best of K arms in hindsight. In this setting, when the rewards are i.i.d. from round to round, upper confidence bound (UCB) algorithms proved both efficient and optimal [Lai and Robbins, 1985, Agrawal, 1995, Auer et al., 2002a]. The idea of confidence bound algorithms is to keep upper bounds on the plausible rewards of the arms and to pull the arm with the highest UCB. These approaches are known to give near-optimal algorithms that provide high probability guarantees on the regret suffered by the learner.

Our setting, in comparison, focuses on bandit prob-

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

lems with features on actions. As opposed to traditional K-armed bandit problems, action features in contextual bandits may be useful to infer the conditional average payoff of an action, which allows a sequential experimenter to even improve the average payoff over time. The name contextual bandit is borrowed from Langford and Zhang [2008], but this setting is also known by other names such as bandit problems with covariates [Woodroofe, 1979, Sarkar, 1991], associative reinforcement learning [Kaelbling, 1994], associative andit problems [Auer, 2002, Strehl et al., 2006], and bandit problems with expert advice [Auer et al., 2002b], among others.

It is useful to mention that one could imagine trying to solve the contextual bandit problem with linear payoffs using Exp4-type approaches [Auer et al., 2002b] that are made to work with an arbitrary hypothesis set. For N experts, Exp4 gives a $O(\sqrt{KT \ln N})$ bound on the regret. One could imagine discretizing the linear hypotheses into an epsilon-net of experts and running Exp4 on them. However, this would have two disadvantages compared to our approach. First, the algorithm would run in time exponential in d, the number of features. Second, its regret bound would have a K dependence (this may be due to the fact that Exp4 works in an adversarial setting while UCB algorithms require rewards to be i.i.d.). For these reasons, it is useful to explicitly take advantage of the linear structure of the predictors in the contextual bandit problem with linear payoffs and to tackle it directly.

Most notably, Auer [2002] considered the contextual bandit problem with linear payoffs (under the name "associative reinforcement learning with linear value functions") and presented LinRel, the first $\tilde{O}(\sqrt{Td})$ algorithm for this problem. LinUCB has a couple of practical advantages over LinRel: first, it is simpler to state and implement; second, it uses ridge regression as its core operation, which is often easier to solve with standard software packages and may be less prone to numerical instability issues compared to the eigen decomposition step used in LinRel.[1] While LinUCB has been experimentally tested [Li et al., 2010, Pavlidis et al., 2008], no theoretical analysis has been carried out. Indeed, it was previously conjectured that LinUCB does not enjoy the same regret bound as LinRel [Auer, 2002, footnote 5]. Our analysis thus answers the question affirmatively.

Our regret lower bound improves the earlier result $\Omega(T^{3/4}K^{1/4})$ by Abe et al. [2003]. This new result matches the best known regret upper bounds up to logarithmic factors.

[1]Even in the worst case, the computation complexity of matrix inversion is no worse than that of eigen decomposition.

Recently, the study of stochastic linear optimization under bandit feedback has received a lot of attention (*e.g.*, Dani et al. [2008], and Rusmevichientong and Tsitsiklis [2010]). Their setting is similar to ours but is different. On each round, instead of having to pick one of K arms with given contextual information, their algorithms are allowed to pick a point in an infinitely large context space (called decision space in the literature). In contrast to the results discussed above, the matching upper and lower regret bounds for this setting is on the order of $d\sqrt{T}$.

3 PROBLEM SETTING

Let T be the number of rounds and K the number of possible actions. Let $r_{t,a} \in [0,1]$ be the reward of action a on round t. On each round t for each action a the learner observes K feature vectors $x_{t,a} \in \mathbb{R}^d$, with $\|x_{t,a}\| \le 1$, where $\|\cdot\|$ denotes the ℓ_2-norm. After observing the feature vectors the learner then selects an action a_t and receives reward r_{t,a_t}.

We operate under the linear realizability assumption; that is, there exists an unknown weight vector $\theta^* \in \mathbb{R}^d$ with $\|\theta^*\| \le 1$ so that

$$\mathbf{E}[r_{t,a} \mid x_{t,a}] = x_{t,a}^\top \theta^*$$

for all t and a. Hence, we assume that the $r_{t,a}$ are independent random variables with expectation $x_{t,a}^\top \theta^*$.

Let $x_{t,a}$ be the feature vector of action a at step t, and define the regret of an algorithm $\mathcal{A}$ to be

$$\sum_{t=1}^{T} r_{t,a_t^*} - \sum_{t=1}^{T} r_{t,a_t}\,,$$

where $a_t^* = \arg\max_a x_{t,a}^\top \theta^*$ is the best action at step t according to θ^* and a_t is the action selected by $\mathcal{A}$ at step t.

We note that in our setting, the context vectors $x_{t,a}$ can be chosen arbitrarily by an oblivious adversary as long as the rewards $r_{t,a}$ are independent given the context.

4 LINUCB

The LinUCB algorithm is motivated by the UCB algorithm of Auer et al. [2002a] and the KWIK algorithm of Walsh et al. [2009]. LinUCB is also similar to the LinRel algorithm of Auer [2002]—the main idea of both algorithms is to compute the expected reward of each arm by finding a linear combination of the previous rewards of the arm. To do this, LinUCB decomposes the feature vector of the current round into

Algorithm 1 LinUCB: UCB with Linear Hypotheses

Inputs: $\alpha \in \mathbb{R}_+, K, d \in \mathbb{N}$
$A \leftarrow I_d$ {The d-by-d identity matirx}
$b \leftarrow \mathbf{0}_d$
for $t = 1, 2, 3, \ldots, T$ **do**
  $\theta_t \leftarrow A^{-1}b$
  Observe K features, $x_{t,1}, x_{t,2}, \cdots, x_{t,K} \in \mathbb{R}^d$
  for $a = 1, 2, \ldots, K$ **do**
    $p_{t,a} \leftarrow \theta_t^\top x_{t,a} + \alpha\sqrt{x_{t,a}^\top A^{-1} x_{t,a}}$ {Computes upper confidence bound}
  end for
  Choose action $a_t = \arg\max_a p_{t,a}$ with ties broken arbitrarily
  Observe payoff $r_t \in \{0, 1\}$
  $A \leftarrow A + x_{t,a_t} x_{t,a_t}^\top$
  $b \leftarrow b + x_{t,a_t} r_t$
end for

a linear combination of feature vectors seen on previous rounds and uses the computed coefficients and rewards on previous rounds to compute the expected reward on the current round.

LinRel, however, is not only a more complicated algorithm but also requires solving an SVD (or eigendecomposition) of a symmetric matrix, while LinUCB only requires inverting the same matrix.

5 REGRET ANALYSIS

In this section we introduce a modified version of LinUCB, and prove its regret is bounded by $\tilde{O}(\sqrt{dT})$. While experiments show LinUCB is probably sufficient in practice Li et al. [2010], there is technical difficulty in analyzing it. In our analysis, we need the predicted set of rewards on the current round to be computed from a linear combination of rewards that are *independent* random variables, in order to apply the Azuma/Hoeffding inequality.

However, LinUCB has the problem that predictions in later rounds are made using previous outcomes. To handle this problem, we modify the algorithm into BaseLinUCB (Algorithm 2) which assumes statistical independence among the samples, and then use a master algorithm SupLinUCB (Algorithm 3) to ensure the assumption holds. This technique is similar to the LinRel/SupLinRel decomposition by Auer [2002].

Algorithm 2 BaseLinUCB: Basic LinUCB with Linear Hypotheses at Step t

Inputs: $\alpha \in \mathbb{R}_+, \Psi_t \subseteq \{1, 2, \cdots, t-1\}$
$A_t \leftarrow I_d + \sum_{\tau \in \Psi_t} x_{\tau,a_\tau}^\top x_{\tau,a_\tau}$
$b_t \leftarrow \sum_{\tau \in \Psi_t} r_{\tau,a_\tau} x_{\tau,a_\tau}$
$\theta_t \leftarrow A_t^{-1} b_t$
Observe K arm features, $x_{t,1}, x_{t,2}, \cdots, x_{t,K} \in \mathbb{R}^d$
for $a \in [K]$ **do**
  $w_{t,a} \leftarrow \alpha\sqrt{x_{t,a}^\top A_t^{-1} x_{t,a}}$
  $\hat{r}_{t,a} \leftarrow \theta_t^\top x_{t,a}$
end for

5.1 Analysis for BaseLinUCB

For convenience, define

$$\begin{aligned} s_{t,a} &= \sqrt{x_{t,a}^\top A_t^{-1} x_{t,a}} \in \mathbb{R}_+ \\ D_t &= \left[x_{\tau,a_\tau}^\top\right]_{\tau \in \Psi_t} \in \mathbb{R}^{|\Psi_t| \times d} \\ y_t &= [r_{\tau,a_\tau}]_{\tau \in \Psi_t} \in \mathbb{R}^{|\Psi_t| \times 1} \end{aligned}$$

Then,

$$A_t = I_d + D_t^\top D_t \quad \text{and} \quad b_t = D_t^\top y_t$$

Furthermore, we denote the eigendecomposition of A_t by $A_t = U_t \Delta_t U_t^\top$, where $\Delta_t = \mathrm{diag}(\lambda_{t,1}, \lambda_{t,2}, \ldots, \lambda_{t,d})$ contains eigenvalues of A_t in the diagonal entries, and $U_t \in \mathbb{R}^{d \times d}$ is a unitary matrix. Note that $\lambda_{1,j} = 1$ for all j as A_t is initialized to I_d.

Lemma 1. *Suppose the input index set Ψ_t in BaseLinUCB is constructed so that for fixed x_{τ,a_τ} with $\tau \in \Psi_t$, the rewards r_{τ,a_τ} are independent random variables with means $\mathbf{E}[r_{\tau,a_\tau}] = x_{\tau,a_\tau}^\top \theta^*$. Then, with probability at least $1 - \delta/T$, we have for all $a \in [K]$ that*

$$\left|\hat{r}_{t,a} - x_{t,a}^\top \theta^*\right| \le (\alpha + 1) s_{t,a}.$$

Proof. Using notation in the algorithm descriptions of BaseLinUCB, we have

$$\begin{aligned} &\hat{r}_{t,a} - x_{t,a}^\top \theta^* \\ &= x_{t,a}^\top \theta_t - x_{t,a}^\top \theta^* \\ &= x_{t,a}^\top A_t^{-1} b_t - x_{t,a}^\top A_t^{-1} \left(I_d + D_t^\top D_t\right) \theta^* \\ &= x_{t,a}^\top A_t^{-1} D_t^\top y_t - x_{t,a}^\top A_t^{-1} \left(\theta^* + D_t^\top D_t \theta^*\right) \\ &= x_{t,a}^\top A_t^{-1} D_t^\top \left(y_t - D_t \theta^*\right) - x_{t,a}^\top A_t^{-1} \theta^*, \end{aligned}$$

and since $\|\theta^*\| \le 1$,

$$\begin{aligned} \left|\hat{r}_{t,a} - x_{t,a}^\top \theta^*\right| \le{}& \left|x_{t,a}^\top A_t^{-1} D_t^\top \left(y_t - D_t \theta^*\right)\right| \\ &+ \left\|A_t^{-1} x_{t,a}\right\|. \end{aligned} \quad (1)$$

Algorithm 3 SupLinUCB (adapted from Auer [2002])

Inputs: $T \in \mathbb{N}$
$S \leftarrow \ln T$
$\Psi_t^s \leftarrow \emptyset$ for all $s \in [T]$
for $t = 1, 2, \ldots, T$ **do**
  $s \leftarrow 1$ and $\hat{A}_1 \leftarrow [K]$
  repeat
    Use BaseLinUCB with Ψ_t^s to calculate the width, $w_{t,a}^s$, and upper confidence bound, $\hat{r}_{t,a}^s + w_{t,a}^s$, for all $a \in \hat{A}_s$.
    if $w_{t,a}^s \le 1/\sqrt{T}$ for all $a \in \hat{A}_s$ **then**
      Choose $a_t = \arg\max_{a \in \hat{A}_s} \left(\hat{r}_{t,a}^s + w_{t,a}^s\right)$
      Keep the same index sets at all levels: $\Psi_{t+1}^{s'} \leftarrow \Psi_t^{s'}$ for all $s' \in [S]$.
    else if $w_{t,a}^s \le 2^{-s}$ for all $a \in \hat{A}_s$ **then**
      $\hat{A}_{s+1} \leftarrow \{a \in \hat{A}_s \mid \hat{r}_{t,a}^s + w_{t,a}^s \ge \max_{a' \in \hat{A}_s} \left(\hat{r}_{t,a'}^s + w_{t,a'}^s\right) - 2^{1-s}\}$
      $s \leftarrow s + 1$.
    else
      Choose $a_t \in \hat{A}_s$ such that $w_{t,a_t}^s > 2^{-s}$.
      Update the index sets at all levels:

$$\Psi_{t+1}^{s'} \leftarrow \begin{cases} \Psi_t^{s'} \cup \{t\}, & \text{if } s = s' \\ \Psi_t^{s'}, & \text{otherwise} \end{cases}.$$

    end if
  until an action a_t is found.
end for

The right-hand side above decomposes the prediction error into a variance term (first) and a bias term (second). Due to statistical independence of samples indexed in Ψ_t, we have $\mathbf{E}[y_t - D_t\theta^*] = 0$, and, by Azuma's inequality,

$$\begin{aligned} &\Pr\left(\left|x_{t,a}^\top A_t^{-1} D_t^\top (y_t - D_t\theta^*)\right| > \alpha s_{t,a}\right) \\ &\le 2\exp\left(-\frac{2\alpha^2 s_{t,a}^2}{\left\|D_t A_t^{-1} x_{t,a}\right\|^2}\right) \\ &\le 2\exp(-2\alpha^2) \\ &= \frac{\delta}{TK}, \end{aligned}$$

where the last inequality is due to the following fact:

$$\begin{aligned} s_{t,a}^2 &= x_{t,a}^\top A_t^{-1} x_{t,a} \\ &= x_{t,a}^\top A_t^{-1}(I_d + D_t^\top D_t) A_t^{-1} x_{t,a} \\ &\ge x_{t,a}^\top A_t^{-1} D_t^\top D_t A_t^{-1} x_{t,a} \\ &= \left\|D_t A_t^{-1} x_{t,a}\right\|^2. \end{aligned}$$

Now applying a union bound, we can guarantee, with probability at least $1-\delta/T$, that for all actions $a \in [K]$,

$$\left|x_{t,a}^\top A_t^{-1} D_t^\top (y_t - D_t\theta^*)\right| \le \alpha s_{t,a}.$$

We next bound the second term in Equation 2:

$$\begin{aligned} \left\|A_t^{-1} x_{t,a}\right\| &= \sqrt{x_{t,a}^\top A_t^{-1} I_d A_t^{-1} x_{t,a}} \\ &\le \sqrt{x_{t,a}^\top A_t^{-1}(I_d + D_t^\top D_t) A_t^{-1} x_{t,a}} \\ &= \sqrt{x_{t,a}^\top A_t^{-1} x_{t,a}} = s_{t,a}. \end{aligned}$$

Combining the two upper bounds above finishes the proof. □

Lemma 2 (Auer [2002], Lemma 11)**.** *Suppose $\Psi_{t+1} = \Psi_t \cup \{t\}$ in BaseLinUCB. Then, the eigenvalues of A_{t+1} can be arranged so that $\lambda_{t,j} \le \lambda_{t+1,j}$ for all j and*

$$s_{t,a_t}^2 \le 10 \sum_{j=1}^{d} \frac{\lambda_{t+1,j} - \lambda_{t,j}}{\lambda_{t,j}}.$$

Lemma 3. *Using notation in BaseLinUCB and assuming $|\Psi_{T+1}| \ge 2$, we have*

$$\sum_{t \in \Psi_{T+1}} s_{t,a_t} \le 5\sqrt{d\,|\Psi_{T+1}| \ln |\Psi_{T+1}|}.$$

Proof. The proof is similar to the proof of Auer [2002, Lemma 13], but modified to handle the difference between our algorithms. For convenience, define $\psi = |\Psi_{T+1}|$. Lemma 2 implies

$$\sum_{t \in \Psi_{T+1}} s_{t,a_t} = \sum_{t \in \Psi_{T+1}} \sqrt{10 \sum_{j=1}^{d} \left(\frac{\lambda_{t+1,j}}{\lambda_{t,j}} - 1\right)}.$$

The function

$$f = \sum_{t \in \Psi} \sqrt{\sum_{j=1}^{d} (h_{tj} - 1)}$$

is maximized under the constraints

$$h_{tj} \ge 1 \text{ and } \sum_{j=1}^{d} \prod_{t \in \Psi} h_{tj} \le C,$$

when

$$h_{tj} = \left(\frac{C}{d}\right)^{1/|\Psi|}$$

for all $t \in \Psi$ and $j \in [d]$, according to Lemma 8.[2]

In our context, we have

$$\frac{\lambda_{t+1,j}}{\lambda_{t,j}} \ge 1$$

[2]The claim was made by Auer [2002] without proof.

and

$$\begin{aligned}\sum_{j=1}^{d}\prod_{t\in\Psi_{T+1}}\frac{\lambda_{t+1,j}}{\lambda_{t,j}} &= \sum_{j=1}^{d}\lambda_{T+1,j}\\ &= \sum_{t\in\Psi_{T+1}}\|x_{ta_t}\|^2+d\\ &\le \psi+d,\end{aligned}$$

and so

$$\begin{aligned}\sum_{t\in\Psi_{T+1}}s_t &\le \psi\sqrt{10d}\sqrt{\left(\frac{\psi+d}{d}\right)^{1/\psi}-1}\\ &\le \psi\sqrt{10d}\sqrt{(\psi+1)^{1/\psi}-1}.\end{aligned}$$

When $\psi\ge 2$, according to Lemma 9, we have

$$(\psi+1)^{1/\psi}-1\le\frac{2.5}{\psi}\ln\psi,$$

and hence

$$\sum_{t\in\Psi_{T+1}}s_t\le\sqrt{25d\psi\ln\psi}=5\sqrt{d\psi\ln\psi}.$$

□

5.2 Analysis for SupLinUCB

In this section, we make use of lemmas in the previous subsections and complete the regret analysis for SupLinUCB. The following lemma is critical for our application of BaseLinUCB.

Lemma 4 (Auer [2002], Lemma 14)**.** *For each $s\in[S]$, each $t\in[T]$, and any fixed sequence of feature vectors x_{t,a_t} with $t\in\Psi_t^s$, the corresponding rewards r_{t,a_t} are independent random variables such that $\mathbf{E}[r_{t,a_t}]=x_{t,a_t}^\top\theta^*$.*

Lemma 5 (Auer [2002], Lemma 15)**.** *With probability $1-\delta S$, for any $t\in[T]$ and any $s\in[S]$, the following hold:*

1. *$|\hat{r}_{t,a}-\mathbf{E}[r_{t,a}]|\le w_{t,a}$ for any $a\in[K]$,*
2. *$a_t^*\in\hat{A}_s$, and*
3. *$\mathbf{E}[r_{t,a_t^*}]-\mathbf{E}[r_{t,a}]\le 2^{3-s}$ for any $a\in\hat{A}_s$.*

Lemma 6. *For all $s\in[S]$,*

$$\left|\Psi_{T+1}^s\right|\le 5\cdot 2^s\left(1+\alpha^2\right)\sqrt{d\left|\Psi_{T+1}^s\right|}.$$

Proof. The proof is a small modification of that in Auer [2002, Lemma 16]. □

Observing that $2^{-s}\le\frac{1}{\sqrt{T}}$, given the previous lemmas, the main theorem follows using a similar argument as Auer [2002].

Theorem 1. *If SupLinUCB is run with*

$$\alpha=\sqrt{\frac{1}{2}\ln\frac{2TK}{\delta}},$$

then with probability at least $1-\delta$, the regret of the algorithm is

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

6 A MATCHING LOWER BOUND

In this section, we prove the following lower bound that matches the upper bound in Theorem 1 up to logarithmic factors.

Theorem 2. *For the contextual bandit problem with linear payoff functions, for any number of trials T and K actions (where $T\ge K\ge 2$), for any any algorithm $\mathcal{A}$ choosing action a_t at time t, there is a constant $\gamma>0$, for $d^2\le T$ a sequence of d-dimensional vectors $x_{t,a}$, such that*

$$\mathbf{E}\left[\sum_{t=1}^{T}\max_a x_{t,a}^\top\theta^*-\sum_{t=1}^{T}r_{t,a_t}\right]\ge\gamma\sqrt{Td}.$$

The following lemma will be useful for our lower bound proof:

Lemma 7 (Auer et al. [2002b], Theorem 5.1)**.** *There is a constant $\gamma>0$ s.t. for any bandit algorithm $\mathcal{A}$ choosing action a_t at time t and any $T\ge K\ge 2$, if arm a is chosen uniformly at random and the probability slot machine pays 1 is set to $p_i=\frac{1}{2}+\frac{1}{4}\sqrt{\frac{K}{T}}$ and the rest pay 1 w.p. $\frac{1}{2}$ then*

$$\boldsymbol{E}\left(p_iT-\sum_{t=1}^{T}r_{t,a_t}\right)\ge\gamma\sqrt{KT}.$$

Proof. We will use a technique similar to Abe et al. [2003] to prove the lower bound for our setting.

We divide our T rounds into $m=(d-1)/2$ groups of $T'=\lfloor 2T/(d-1)\rfloor$ rounds such that each group $1,2,\ldots,m$ has a different best action. We will then use Lemma 7 for each group independently.

We say time step t belongs in group r if $\lfloor t/m\rfloor=r$. For all t and all actions, we let $x_{t,a}$ have a $1/2$ in the first component and also $1/2$ in components $2r$ and $2r+1$, and 0 in the remaining components.

The true vector θ^* will have a $1/2$ in its first coordinate, and values $\sqrt{1/T'}$ for either coordinate $2r$ or

$2r+1$ for each group r (and 0 in the other)—setting θ^* in this manner constrains $d^2 \leq T$, as otherwise we would violate the condition that $\|\theta^*\| \leq 1$.

Our choice of θ^* induces expected rewards that corresponds to a bandit problem where, in each group, one of two arms (either the rth or $r+1$st) pays off with probability $1/4+1/(2T')$ and the other with probability $1/4$.

Scaling by 2 and applying Lemma 7 (with K=2) independently for each group, we get a per-group regret of $\gamma'\sqrt{T'}$ or $\gamma''\sqrt{\frac{T}{d}}$ for some constants $\gamma', \gamma'' > 0$. Summing over the $(d-1)/2$ groups finishes the proof. □

We note that this bound is within a polylogarithmic factors of LinUCB and LinRel.

We can compare this with work of Abe et al. [2003] that gives regret of $\Omega(K^{1/4}T^{3/4})$ for this problem; they also give an algorithm with an almost-matching upper bound on regret of $O(K^{1/2}T^{3/4})$—this dependence on $T^{3/4}$ shows that our $\Omega(\sqrt{dT})$ lower bound requires the constraint $d^2 \leq T$.

7 CONCLUSIONS

In this paper we analyze a polynomial-time algorithm for the contextual bandit problem with linear payoffs. The algorithm is simpler and more robust in practice than its precursor. We also give a lower bound showing that a regret of $\tilde{O}\left(\sqrt{Td}\right)$ cannot be improved, modulo logarithmic factors.

However, like Auer [2002] and Abe et al. [2003], we make the realizability assumption that there exists a vector θ^* for which $\mathbf{E}[r_t \mid x_{t,a}] = x_{t,a}^\top \theta^*$. Ideally, we would like to be able to handle the agnostic case. Competing with linear predictors without assuming one perfectly predicts the expected rewards remains an interesting open problem.

Acknowledgements

We thank an anonymous referee from an earlier version of this paper for correcting our proof of Theorem 2.

This work was done while Lev Reyzin and Robert E. Schapire were at Yahoo! Research, New York. Lev Reyzin acknowledges that this material is based upon work supported by the National Science Foundation under Grant #0937060 to the Computing Research Association for the Computing Innovation Fellowship program.

References

Naoki Abe, Alan W. Biermann, and Philip M. Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003.

Deepak Agarwal, Bee-Chung Chen, Pradheep Elango, Nitin Motgi, Seung-Taek Park, Raghu Ramakrishnan, Scott Roy, and Joe Zachariah. Online models for content optimization. In *Advances in Neural Information Processing Systems 21 (NIPS-08)*, pages 17–24, 2009.

Rajeev Agrawal. Sample mean based index policies with $o(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4): 1054–1078, 1995.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32 (1):48–77, 2002b.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the Twenty-First Annual Conference on Learning Theory (COLT-08)*, pages 355–366, 2008.

Leslie Pack Kaelbling. Associative reinforcement learning: Functions in k-DNF. *Machine Learning*, 15(3):279–298, 1994.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Advances in Neural Information Processing Systems 20*, pages 1096–1103, 2008.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *WWW*, pages 661–670, 2010.

Nicos G. Pavlidis, Dimitris K. Tasoulis, and David J. Hand. Simulation studies of multi-armed bandits with covariates. In *Proceedings on the Tenth International Conference on Computer Modeling and Simulation (UKSIM-08)*, pages 493–498, 2008. Invited paper.

Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Jyotirmoy Sarkar. One-armed bandit problems with covariates. *The Annals of Statistics*, 19(4):1978–2002, 1991.

Alexander L. Strehl, Chris Mesterharm, Michael L. Littman, and Haym Hirsh. Experience-efficient learning in associative bandit problems. In *Proceedings of the Twenty-Third International Conference on Machine Learning (ICML-06)*, pages 889–896, 2006.

Thomas J. Walsh, István Szita, Carlos Diuk, and Michael L. Littman. Exploring compact reinforcement-learning representations with linear regression. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI-09)*, pages 591–598, 2009. A corrected version is available as Technical Report DCS-tr-660, Department of Computer Science, Rutgers University, December, 2009.

Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistics Association*, 74(368):799–806, 1979.

A TECHNICAL LEMMAS

We state and prove two technical lemmas that were useful for the regret analysis in Section 5.

Lemma 8. *The function*

$$f = \sum_{t\in\Psi} \sqrt{\sum_{j=1}^{d} c_{tj}}$$

is maximized, under the constraints

$$c_{tj} \geq 0 \text{ and } \sum_{j=1}^{d} \prod_{t\in\Psi} (c_{tj}+1) \leq C,$$

for some constant $C > d$ when

$$c_{tj} = \left(\frac{C}{d}\right)^{1/|\Psi|} - 1$$

for all $t \in \Psi$ and $j \in [d]$.

Proof. Compute the Lagrangian function using the second constraint, and we have

$$L(\lambda) = \sum_{t\in\Psi} \sqrt{\sum_{j=1}^{d} c_{tj}} + \lambda \left(\sum_{j=1}^{d} \prod_{t\in\Psi} (c_{tj}+1) - C \right).$$

Letting the partial derivative w.r.t. $c_{\tau i}$ be 0, we have

$$\frac{\partial L}{\partial c_{\tau i}} = \frac{1}{2} \left(\sum_{j=1}^{d} c_{\tau j} \right)^{-1/2} + \frac{\lambda}{c_{\tau i}+1} \prod_{t\in\Psi} (c_{ti}+1) = 0.$$

It can then be seen that at a stationary point, all c_{tj} are identical, yielding

$$c_{tj} = \left(\frac{C}{d}\right)^{1/|\Psi|} - 1.$$

It can be seen easily that $c_{tj} > 0$.

We now compute the Hessian matrix, $\nabla^2 f$, which is clearly negative definite at the stationary point, thus proving the stationary point is indeed a maximizer of the constrained optimization problem. □

Lemma 9. *If $\psi \geq 2$, then*

$$(\psi+1)^{1/\psi} - 1 \leq \frac{2.5}{\psi} \ln \psi.$$

Proof. It is equivalent to show

$$\frac{\ln(\psi+1)}{\psi} \leq \ln\left(1 + \frac{2.5}{\psi} \ln \psi\right).$$

Since $\psi \geq 2$ and $\ln(\psi+1)/\ln\psi$ is a decreasing function of ψ, it suffices to show

$$\frac{\ln 3}{\ln 2} \frac{\ln \psi}{\psi} \leq \ln\left(1 + \frac{2.5}{\psi} \ln \psi\right),$$

which can be verified. □