Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

CBRAP: Contextual Bandits with RAndom Projection

Xiaotian Yu, Michael R. Lyu, Irwin King
Department of Computer Science and Engineering
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
Email: {xtyu,lyu,king}@cse.cuhk.edu.hk

Abstract

Contextual bandits with linear payoffs, which are also known as linear bandits, provide a powerful alternative for solving practical problems of sequential decisions, e.g., online advertisements. In the era of big data, contextual data usually tend to be high-dimensional, which leads to new challenges for traditional linear bandits mostly designed for the setting of low-dimensional contextual data. Due to the curse of dimensionality, there are two challenges in most of the current bandit algorithms: the first is high time-complexity; and the second is extreme large upper regret bounds with high-dimensional data. In this paper, in order to attack the above two challenges effectively, we develop an algorithm of Contextual Bandits via RAndom Projection (CBRAP) in the setting of linear payoffs, which works especially for high-dimensional contextual data. The proposed CBRAP algorithm is time-efficient and flexible, because it enables players to choose an arm in a low-dimensional space, and relaxes the sparsity assumption of constant number of non-zero components in previous work. Besides, we prove an upper regret bound for the proposed algorithm, which is associated with reduced dimensions. By comparing with three benchmark algorithms, we demonstrate improved performance on cumulative payoffs of CBRAP during its sequential decisions on both synthetic and real-world datasets, as well as its superior time-efficiency.

Introduction

The Multi-Armed Bandit (MAB) problem was proposed and investigated by Robbins in 1952, which has attracted great interests from numerous researchers in operation research and computer science (Robbins 1952; Auer, Cesa-Bianchi, and Fischer 2002; Bubeck and Cesa-Bianchi 2012). The fundamental issue in the MAB problem and its variants focuses on the exploration-exploitation trade-off, which refers to an algorithm trying to maximize cumulative rewards in sequential decisions but the algorithm has only limited knowledge about the mechanism of generating the rewards (Auer 2002).

As a natural and important variant of the basic MAB problem, contextual bandits with linear payoffs, which are also known as linear bandits, are sequential decision-making problems with side information (Wang, Kulkarni, and Poor 2005; Dani, Hayes, and Kakade 2008; Abbasi-Yadkori, Pál, and Szepesvári 2011; Chu et al. 2011). Specifically, given feature information of arm space for each of T rounds, a learner is required to choose one of K arms. Linear bandits contain a basic assumption of linearly mapping from the arm space to the reward space (Filippi et al. 2010), which should be the most common case in reality.

Recently, contextual bandits with linear payoffs have been successfully applied into many practical applications, such as recommender systems (Tang et al. 2013), social network analysis (Zhao, McAuley, and King 2014; Zhao and King 2016b) and information retrieval (Zhao and King 2016a). In (Tang et al. 2013), the demonstration of advertisements was based on users' input information on web pages. The authors formulated the problem of automatic layout selection in online advertisements as a contextual bandit problem. These personalized advertisements are expected to improve click-through rates of web links. For these models of sequential decisions, recommendation algorithms always receive additional contextual information from users, which could be greatly useful for the online sequential decisions.

In the big data era, it is pretty common to encounter high-dimensional and/or sparse contextual information. In this case, traditional bandit algorithms, which are mostly designed in the setting of low-dimensional data, are facing new challenges in applications. Due to the curse of dimensionality, there are two challenges in most the of current bandit algorithms. The first is high time-complexity; and the second is extremely large upper regret bounds with high-dimensional data. Specifically, traditional contextual bandits (e.g., LinUCB in (Chu et al. 2011)) contain inverse operations in the original contextual space, which will be time-consuming for computations. Besides, the regret bounds of linear bandits in (Chu et al. 2011; Abbasi-Yadkori, Pál, and Szepesvári 2011) are related to the original dimension of contextual data, which can lead to increasing regret bounds with the curse of dimensionality. This will be even worse when the original dimension of contextual data is larger than the the total sequential rounds of playing bandits.

There have been some efforts on context bandits with linear payoffs in high-dimensional and/or sparse contextual data (Deshpande and Montanari 2012; Carpentier, Munos, and others 2012). The corresponding bandit algorithms are

named BallExp in (Deshpande and Montanari 2012), and SLUCB in (Carpentier, Munos, and others 2012).

However, in SLUCB, the authors assumed that the contextual data contain S non-zero components, which may not be flexible in applications, especially for cases with high-dimensional dense data. In BallExp, the upper and lower regret bounds are relatively loose, and are still closely related to the original dimension of contextual data. Besides, both SLUCB and BallExp adopted the technique of ball exploration in the high-dimensional space, which will be time-consuming in applications. For rigorous analysis of time complexity for these two algorithms, interested readers can refer to (Dani, Hayes, and Kakade 2008).

Random projection is a powerful and popular technique to deal with high-dimensional data (Fern and Brodley 2003; Zhang et al. 2016), which maps high-dimensional data onto a low-dimensional space. Note that random projection does not contain the assumption of sparsity in high-dimensional data. The most common case in random projection is to construct a Gaussian random matrix, where each element is an i.i.d. sample following a standard normal distribution. It has been proved to preserve the Euclidean distance within an error ball (Dasgupta and Gupta 1999). Besides, the error bounds for inner products in random projection have been investigated (Kabán 2015), which will be an effective tool for analysis of upper regret bounds in contextual bandits.

In this paper, to tackle the aforementioned two challenges effectively, we propose an algorithm of Contextual Bandits via RAndom Projection (CBRAP) in the setting of linear payoffs, which works especially for high-dimensional data. Note that, for simplicity in the work, we assume that contextual bandits have linear payoff functions. But the framework of our bandit algorithm can be easily generalized to the case of relaxing the linear assumption. Specifically, our proposed algorithm adopts random projection to map the high-dimensional contextual information onto a low-dimension space, where we should design a random matrix. The proposed CBRAP algorithm is time-efficient and flexible, because it enables players to choose an arm in a low-dimensional space, and relaxes the sparsity assumption of constant number of non-zero components in previous work. Besides, we prove an upper regret bound for the proposed algorithm, and show the bound to be better than the traditional ones with appropriate reduced dimensions. By comparing with three benchmark algorithms (i.e., LinUCB, BallExp and SLUCB), we demonstrate improved performance on cumulative payoffs of the CBRAP algorithm during its sequential decisions on both synthetic and real-world datasets, as well as its superior time-efficiency.

In summary, we make the following contributions.

- For contextual bandits in the setting of linear payoffs, we develop an efficient and practical algorithm named CBRAP by taking advantage of random projection.
- We derive an upper regret bound for the proposed CBRAP algorithm, which guarantee the worst case is associated with the reduced dimensions. Besides, our algorithm is more flexible and time-efficient than BallExp and SLUCB in high-dimensional settings.
- We evaluate the CBRAP algorithm via a series of experiments with synthetic and real-world datasets. Compared with the three benchmarks, we demonstrate the proposed algorithm's improved performance of cumulative payoffs during sequential decisions, as well as its time-efficiency.

Preliminary and Related Work

In this section, we first introduce notions and the definition of sub-Gaussian of a random variable, which will be used in this paper. Then, we present the process of contextual bandits with linear payoffs, as well as the metric of bandits. Finally, we provide a brief survey on random projection.

Notations and Definition of Sub-Gaussian

The total sequential rounds of playing bandits is T. For each round $t \in [T]$ with $[T] = \{1, 2, \cdots, T\}$, a learner receives contextual information from the set of $\mathcal{X} \in \mathbb{R}^n$, where n can be an extremely large integer representing a high-dimensional space. In this work, high-dimensional contextual data precisely mean that $T \leq n$ or even $T \ll n$, which is the case mentioned in (Carpentier, Munos, and others 2012). Let $K \in \mathbb{N}_+$ be the number of arms and $\pi_{t,y} \in [0, 1]$ the reward of arm y on round t with $y \in [K]$ and $[K] = \{1, 2, \cdots, K\}$. We adopt $\|\cdot\|_2$ to denote the ℓ^2 norm of a vector $\mathbf{x} \in \mathbb{R}^n$, and $\mathbf{I}_{m\times m}$ to denote the identity matrix with dimensions of $m \times m$. For a positive definite matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$, the weighted norm of vector $\mathbf{x}$ is defined as $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}}$. The inner product is represented as $\langle\cdot,\cdot\rangle$, and the weighted inner product is $\mathbf{x}\mathbf{A}^{\mathrm{T}}\mathbf{y} = \langle\mathbf{x}, \mathbf{y}\rangle_{\mathbf{A}}$.

Mathematically, we given the following definition on the sub-Gaussian of a random variable.

Definition 1 ((Buldygin and Kozachenko 1980))**.** *A random variable ξ is sub-Gaussian if there exists an $R \geq 0$ such that*

$$\mathbf{E}[\exp(\lambda\xi)] \leq \exp(\frac{\lambda^2 R^2}{2}), \tag{1}$$

where $\lambda \in \mathbb{R}$, $\mathbf{E}[\cdot]$ is the expectation of a random variable and $\exp(\cdot)$ denotes the exponential operation.

Given a set $\mathcal{F}$, ξ is conditionally R-sub-Gaussian if $\forall\lambda \in \mathbb{R}$ and a fixed $R \geq 0$, we have $\mathbf{E}[\exp(\lambda\xi)|\mathcal{F}] \leq \exp(\frac{\lambda^2 R^2}{2})$.

Contextual Bandits with Linear Payoffs

As shown in (Auer 2002; Chu et al. 2011), an algorithm (denoted by $\mathcal{A}$) for contextual bandits with linear payoffs usually contains the following three steps at round t:

1) contextual information $\mathbf{x}_{t,y} \in \mathcal{X}$ for all $y \in [K]$ is revealed to the bandit algorithm $\mathcal{A}$;
2) the bandit algorithm $\mathcal{A}$ chooses an arm $a_t \in [K]$, which follows an underlying distribution $\Pi(\mathbf{x}_{t,a_t}, \boldsymbol{\theta}^*)$ with $\boldsymbol{\theta}^* \in \mathbb{R}^n$ being the unknown true parameter vector; and
3) a stochastic payoff $\pi_{t,a_t} \in [0, 1]$ is revealed to the bandit algorithm $\mathcal{A}$.

In the above stochastic setting of step 3, for the chosen arm a_t at round t, we usually assume that there is an underlying distribution $\Pi(\mathbf{x}_{t,a_t}, \boldsymbol{\theta}^*)$ with the first moment information being $\langle\mathbf{x}_{t,a_t}, \boldsymbol{\theta}^*\rangle$, so that π_{t,a_t} is a sample from $\Pi(\mathbf{x}_{t,a_t}, \boldsymbol{\theta}^*)$. Thus, we have

$$\pi_{t,a_t} = \mathbf{x}_{t,a_t}^{\mathrm{T}}\boldsymbol{\theta}^* + \eta_t, \tag{2}$$

where η_t is a random noise satisfying the assumption of conditionally R-sub-Gaussian. That is, $\forall\lambda \in \mathbb{R}$, we have

$$\mathbf{E}[\exp(\lambda\eta_t)|\mathcal{F}_t] \leq \exp(\frac{\lambda^2 R^2}{2}), \tag{3}$$

where $\mathcal{F}_t$ is the σ-algebra of $\sigma(\{\mathbf{x}_{i,a_i}\}_{i\in[t]}, \{\eta_i\}_{i\in[t-1]})$ and $R \geq 0$. Eq. (3) implies that $\mathbf{E}[\eta_t|\mathcal{F}_t] = 0$.

A popular measure in demonstrating the performance of an algorithm for solving MAB problems is regret, which is defined as the difference between the expected payoff of the optimal decision in hindsight and that of the algorithm. Mathematically, the regret of the algorithm $\mathcal{A}$ is defined as

$$Regret(T) \triangleq \mathbf{E}[\sum_{t=1}^{T} \max_{y\in[K]} \mathbf{x}_{t,y}^{\mathrm{T}}\boldsymbol{\theta}^* - \sum_{t=1}^{T} \pi_{t,a_t}]. \tag{4}$$

Random Projection

One common technique for dimensionality reduction is to perform linear random projection (Baraniuk, Cevher, and Wakin 2010; Fodor 2002). In this paper, we consider projecting the contextual data of $\mathcal{X} \in \mathbb{R}^n$ onto a low-dimensional space of $\mathcal{Z} \in \mathbb{R}^m$. Without loss of generality, we denote the random projection matrix by $\mathbf{M} \in \mathbb{R}^{m\times n}$. Then, we have

$$\mathbf{z} = \mathbf{M}\mathbf{x}, \tag{5}$$

where $\mathbf{z} \in \mathcal{Z}$ and $\mathbf{x} \in \mathcal{X}$.

In (Blum 2006), $\mathbf{M}$ is constructed as a random matrix where each element follows a normal distribution of $\mathcal{N} \sim (0, \hat{\sigma}^2)$. By setting $\hat{\sigma}^2 = 1/m$ in the next section, we name our algorithm of `CBRAP` with Standard Gaussian (SG) matrix (abbreviated as `CBRAP.SG`).

In (Achlioptas 2003), the authors proposed new methods for constructing sparse random sign matrix for dimensionality reduction. In the ensuing section, we name our algorithm of `CBRAP` with Random Sign (RS) matrix (abbreviated as `CBRAP.RS`).

In addition to the above work, there have been other ways of constructing a matrix for random projection (Li, Hastie, and Church 2006; Ailon and Chazelle 2006; Clarkson and Woodruff 2013; Lu et al. 2013). These investigations consider how to speed up the dimensionality reduction, or how to conduct the random projection with the assumption of low-rank matrix. In this paper, since our focus is to conduct the dimensionality reduction of contextual bandits in a general way, we only consider the construction of random matrix in (Blum 2006; Achlioptas 2003).

Related Work

Contextual bandits are important variants of traditional MAB problems and match many real applications (Langford and Zhang 2008; Li et al. 2010; Tang et al. 2013; Wang, Kulkarni, and Poor 2005). Contextual bandits with linear payoffs have been intensively investigated in previous work (Abbasi-Yadkori, Pal, and Szepesvari 2012; Abe, Biermann, and Long 2003; Abe and Long 1999; Chu et al. 2011; Kaelbling 1994). As shown in (Chu et al. 2011), the traditional upper regret bound for the `LinUCB` algorithm is

$$Regret(T) \leq \mathcal{O}(\sqrt{Tn\ln^3(KT\ln(T)/\delta)}), \tag{6}$$

where $\delta \in (0, 1)$ is a confidence parameter. We know that the dimension of contextual information of n in Eq. (6) will increase when the dimension of context space increases. Roughly, we have the upper regret bound as $R(T) \leq \mathcal{O}(T)$ when $n = T$. This will be even worse when the dimension of context data becomes larger, especially for $n \ll T$.

In (Abbasi-Yadkori, Pal, and Szepesvari 2012), Abbasi-Yadkori et al. studied a sparse variant of stochastic linear bandits. For high-dimensional bandits, Carpentier and Munos (Carpentier, Munos, and others 2012) attacked high-dimensional stochastic linear bandits with the sparsity assumption of S non-zero component, where the algorithm is named `SLUCB`. The upper regret bound in (Carpentier, Munos, and others 2012) is $\mathcal{O}(S\sqrt{T})$. In real applications, the sparsity assumption may be unreasonable, especially for high-dimensional dense data.

In (Deshpande and Montanari 2012), the authors proposed an algorithm named `BallExp` for high-dimensional linear bandits, where the regret bound is relatively loose, and is directly related to the dimension of data.

Recently, by adopting additional assumptions of margin and compatibility conditions in (Bastani and Bayati 2015), the authors investigated high-dimensional covariates in online decision-marking.

From prior work, it is urgent and important to develop a flexible and practical algorithm for contextual bandits with high-dimensional data, where we do not have additional assumptions (e.g., sparsity or the margin condition). This motivates our proposed `CBRAP` algorithm in the next section.

The `CBRAP` Algorithm

In this section, we firstly present the overview of `CBRAP`, and then provide theoretical analyses of a practical upper regret bound and time complexity for the algorithm.

Overview of `CBRAP`

Our proposed bandit algorithm is shown in Algorithm 1, which is named `CBRAP`. As depicted in Algorithm 1, the basic idea of `CBRAP` algorithm is to project the high-dimensional data onto a low-dimensional space, and maintains a confidence set of the unknown optimal parameter $\boldsymbol{\theta}_{\mathbf{z}}^* \in \mathbb{R}^m$ in a low-dimensional space. $\boldsymbol{\theta}_{\mathbf{z}}^*$ is corresponding to the original true parameter $\boldsymbol{\theta}^*$ in n-dimensional space.

Our main contribution in the `CBRAP` algorithm is twofold. First, we construct a random matrix for contextual bandits from Step 2 to Step 7 in Algorithm 1. Note that the designed random matrix is flexible, and can be revised based on users' needs. Here we just consider the random matrix in (Blum 2006; Achlioptas 2003). Second, via the designed random matrix $\mathbf{M}$, we conduct dimensionality reduction for the contextual information in Algorithm 1.

Theoretical Analyses of Regret Bound

In this section, we provide theoretical results on upper regret bound of the proposed `CBRAP` algorithm. First of all, we show that errors resulting from projecting high-dimensional data onto low-dimensional data are condition-

Algorithm 1 CBRAP

input: m, T, $\beta \in \mathbb{R}_+$ and $\alpha \in \mathbb{R}_+$
for $p = 1, 2, \cdots, m$ **do**
  for $q = 1, 2, \cdots, n$ **do**
    generate a random value b_{pq} based on SG or RS
    $\mathbf{M}(p, q) \leftarrow b_{pq}$
  end for
end for
$\mathbf{A}_0 \leftarrow \mathbf{I}_{m \times m}$
$\mathbf{b}_0 \leftarrow \mathbf{0}_m$
for $t = 1, 2, \cdots, T$ **do**
  observe context $\mathbf{x}_{t,y} \in \mathbb{R}^n$ for all $y \in [K]$
  $\mathbf{z}_{t,y} \leftarrow \mathbf{M}\mathbf{x}_{t,y}$ for all $y \in [K]$
  if $t == 1$ **then**
    $\mathbf{A}_t \leftarrow \mathbf{A}_{t-1}$
    $\mathbf{b}_t \leftarrow \mathbf{b}_{t-1}$
  else
    $\mathbf{A}_t \leftarrow \mathbf{A}_{t-1} + \mathbf{z}_{t-1,a_{t-1}}\mathbf{z}_{t-1,a_{t-1}}^{\mathrm{T}}$
    $\mathbf{b}_t \leftarrow \mathbf{b}_{t-1} + \pi_{t-1,a_{t-1}}\mathbf{z}_{t-1,a_{t-1}}$
  end if
  $\boldsymbol{\theta}_{\mathbf{z}}^t \leftarrow \mathbf{A}_t^{-1}\mathbf{b}_t$
  for $y \in [K]$ **do**
    $v_{t,y} \leftarrow \beta\|\mathbf{z}_{t,y}\|_{\mathbf{A}_t^{-1}}$
    $\hat{r}_{t,y} \leftarrow \langle \boldsymbol{\theta}_{\mathbf{z}}^t, \mathbf{z}_{t,y} \rangle$
    $\mathrm{ucb}_{t,y} \leftarrow \hat{r}_{t,y} + v_{t,y}$
  end for
  choose the arm $a_t \leftarrow \arg\max_{y \in [K]} \mathrm{ucb}_{t,y}$ (break ties arbitrarily)
  observe the reward π_{t,a_t}
end for

ally sub-Gaussian associated with the reduced dimension m. Then, we derive a practical upper regret bound for CBRAP.

Incurred sub-Gaussian errors in CBRAP

We focus on the errors due to dimensionality reduction in CBRAP. The following theorem shows that the incurred errors are sub-Gaussian, which can be combined with the original independent noise of η_t to be a new sub-Gaussian random noise.

Theorem 1. *Considering a linear stochastic payoff model as $\pi_t = \langle \mathbf{x}_t, \boldsymbol{\theta}^* \rangle + \eta_t$, where $\mathbf{x}_t \in \mathbb{R}^n$, $\boldsymbol{\theta}^* \in \mathbb{R}^n$ and η_t is conditionally R-sub-Gaussian with the σ-algebra as $\mathcal{F}_t = \sigma(\{\mathbf{x}_i\}_{i \in [t]}, \{\eta_i\}_{i \in [t-1]})$, we can design a random matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ such that $\pi_t = \langle \mathbf{z}_t, \boldsymbol{\theta}_{\mathbf{z}}^* \rangle + \eta_t + \rho_t$ and $\forall \lambda \in \mathbb{R}$*

$$\mathbf{E}[\exp(\lambda(\eta_t + \rho_t))|\mathcal{F}_t'] \leq \exp(\frac{\lambda^2(\sqrt{4/m} + R)^2}{2}), \tag{7}$$

where $\mathcal{F}_t' = \sigma(\{\mathbf{x}_i\}_{i \in [t]}, \{\eta_i\}_{i \in [t-1]}, \{\rho_i\}_{i \in [t-1]})$ is σ-algebra, $\mathbf{z}_t = \mathbf{M}\mathbf{x}_t$ and $\boldsymbol{\theta}_{\mathbf{z}}^* \in \mathbb{R}^m$ is the unknown parameter in m-dimensional space. In other words, with the designed random matrix $\mathbf{M}$, $\eta_t + \rho_t$ is conditionally $(\sqrt{4/m} + R)$-sub-Gaussian.

Proof. Without loss of generality, we assume that $\|\mathbf{x}\|_2 \leq 1$ and $\|\boldsymbol{\theta}^*\|_2 \leq 1$. Otherwise, we can conduct normalization. Given a random matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ with normal distribution as $\mathcal{N} \sim (0, \hat{\sigma}^2)$, the mapping from n-dimension to m-dimension for $\boldsymbol{\theta} \in \mathbb{R}^n$ is denoted by

$$\boldsymbol{\theta}_{\mathbf{z}} = \mathbf{M}\boldsymbol{\theta}, \tag{8}$$

where $\boldsymbol{\theta}_{\mathbf{z}} \in \mathbb{R}^m$.

Since the random noise η_t is conditionally R-sub-Gaussian with $\mathcal{F}_t = \sigma(\{\mathbf{x}_i\}_{i \in [t]}, \{\eta_i\}_{i \in [t-1]})$, we are ready to have $\mathbf{E}[\eta_t|\mathcal{F}_t] = 0$, and $\mathbf{E}[\eta_t|\mathcal{F}_t'] = 0$. Besides, we have

$$\mathbf{E}[\langle \mathbf{x}_t, \boldsymbol{\theta}^* \rangle + \eta_t|\mathcal{F}_t] = \mathbf{E}[\langle \mathbf{z}_t, \boldsymbol{\theta}_{\mathbf{z}}^* \rangle + \eta_t + \rho_t|\mathcal{F}_t']. \tag{9}$$

Due to the mean preservation of inner product for random projection, which has been shown in Lemma 4 of (Shi et al. 2012), we have $\mathbf{E}[\langle \mathbf{x}_t, \boldsymbol{\theta}^* \rangle|\mathcal{F}_t] = \mathbf{E}[\langle \mathbf{z}_t, \boldsymbol{\theta}_{\mathbf{z}}^* \rangle|\mathcal{F}_t']$. Thus,

$$\mathbf{E}[\rho_t|\mathcal{F}_t'] = 0. \tag{10}$$

Based on (Kabán 2015), given $\epsilon \in (0, 1)$, we have

$$\Pr\{\mathbf{z}^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{z}} < \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}m\hat{\sigma}^2 - \epsilon m\hat{\sigma}^2\|\mathbf{x}\|_2\|\boldsymbol{\theta}\|_2\} < \exp(-\frac{m\epsilon^2}{8}),$$

$$\Pr\{\mathbf{z}^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{z}} > \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}m\hat{\sigma}^2 + \epsilon m\hat{\sigma}^2\|\mathbf{x}\|_2\|\boldsymbol{\theta}\|_2\} < \exp(-\frac{m\epsilon^2}{8}).$$

We focus on the error bounds of $\mathbf{z}^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{z}} - \mathbf{x}\boldsymbol{\theta}$. In practice, we can set $m\hat{\sigma}^2 = 1$. Note that $\|\mathbf{x}\|_2 \leq 1$ and $\|\boldsymbol{\theta}\|_2 \leq 1$. Thus, we have

$$\Pr\{\mathbf{z}^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{z}} < \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta} - \epsilon\} < \exp(-\frac{m\epsilon^2}{8}), \tag{11}$$

$$\Pr\{\mathbf{z}^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{z}} > \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta} + \epsilon\} < \exp(-\frac{m\epsilon^2}{8}). \tag{12}$$

Note that $\rho_t = \langle \mathbf{x}_t, \boldsymbol{\theta}^* \rangle - \langle \mathbf{z}_t, \boldsymbol{\theta}_{\mathbf{z}}^* \rangle = \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}^* - \mathbf{z}^{\mathrm{T}}\mathbf{M}\boldsymbol{\theta}^*$. Thus, with high probability of $1 - 2\exp(-\frac{m\epsilon^2}{8})$, we have $|\rho_t| \leq \epsilon$. Since we can adopt the σ-algebra $\mathcal{F}_t'$ in Eqs. (11) and (12),

$$\Pr\{|\rho_t| > \epsilon|\mathcal{F}_t'\} \leq 2\exp(-\frac{m\epsilon^2}{8}). \tag{13}$$

Based on Eqs. (10) and (13), and Theorem 3.1 in (Rivasplata 2012), we have $\mathbf{E}[\exp(\lambda\rho_t)|\mathcal{F}_t'] \leq \exp(2\lambda^2/m)$ for all $\lambda \in \mathbb{R}$. In light of the definition of sub-Gaussian, we have that ρ_t is conditionally $\sqrt{4/m}$-sub-Gaussian.

Now we know η_t and ρ_t are both conditionally sub-Gaussian. Thus, in light of Theorem 2.7 in (Rivasplata 2012), we have $\mathbf{E}[\exp(\lambda(\eta_t + \rho_t))|\mathcal{F}_t'] \leq \exp(\lambda^2(\sqrt{4/m} + R)^2/2)$ for all $\lambda \in \mathbb{R}$, which means that $\eta_t + \rho_t$ is conditionally $(\sqrt{4/m} + R)$-sub-Gaussian. □

Upper regret bound for CBRAP

In this subsection, we derive an upper regret bound for the proposed CBRAP algorithm.

Theorem 2. *If CBRAP is run, then with probability at least $(1 - \delta)(1 - 2\exp(-\frac{m\epsilon^2}{8}))$, the upper bound of regret of the algorithm is*

$$Regret(T) \leq 2\sqrt{mT\log(1 + \frac{TL_2^2}{\gamma m})}\sqrt{\beta_T^2(\delta) + \beta_T(\delta)\epsilon},$$

where $\beta_T(\delta) = (\sqrt{4/m + R})\sqrt{m\log(\frac{1 + TL_2^2/\gamma}{\delta})} + \gamma^{1/2}L_1$, with L_1 and L_2 being parameters associated with the reduced dimension.

Table 1: Statistics of used datasets.

dataset	name	#items	#users	#dim.	{(#feedback; sparsity)}	rewards
synthetic 1	s1	100	10	1,000	{(200;0.10),(201;0.30),(194;0.50),(333;0.70),(238;0.90)}	{0,1}
synthetic 2	s2	200	20	2,000	{(1,098;0.50)}	{0,1}
synthetic 3	s3	200	20	5,000	{(1,282;0.50)}	{0,1}
synthetic 4	s4	100	30	10,000	{(592;0.50)}	{0,1}
Movielens	ML	668	10,329	9,689	{(105,339;0.99)}	{0,1,2,3,4,5}
Jester	JR	24,983	100	1,458	{(1,810,455;0.98)}	[0,10]

Table 2: Time complexity for bandit algorithms.

CBRAP	SLUCB	BallExp	LinUCB
$\mathcal{O}(m^3KT)$	$\mathcal{O}(n^2KT)$	$\mathcal{O}(n^2KT)$	$\mathcal{O}(n^3KT)$

Proof. We provide a sketch of the proof here. The detailed proof can be found in the appendix.

First, we adopt the theoretical result in (Abbasi-Yadkori, Pál, and Szepesvári 2011) for the unknown parameter in the reduced m-dimension with a confidence bound.

Second, we divide the regret between payoffs of m-dimension and optimal payoffs into two parts. The first is the error resulting from the dimension reduction. The second is the error due to adopting the confidence bound.

Third, by adding the errors together, we can obtain the final regret of the algorithm in the m-dimension space. □

Time Complexity Analysis

We analyze the time complexity for the CBRAP algorithm, and the three benchmarks, which is shown in Table 2. Note that, in the CBRAP algorithm and the LinUCB algorithm, the most time-consuming operation is the inverse of matrix $\mathbf{A}_t$. For SLUCB and BallExp, the analysis of time complexity can be found in (Dani, Hayes, and Kakade 2008). From the table, we know that the proposed CBRAP algorithm is time efficient, especially for the case of $m \ll n$.

Experiments

In this section, we conduct experiments based on the proposed CBRAP algorithm, and three benchmarks of LinUCB, BallExp and SLUCB. Note that, in the CBRAP algorithm, by designing different random matrices, we have two variants of CBRAP.SG and CBRAP.RS. Our algorithm and used datasets are all publicly available[1].

Datasets

For verifications, we adopt six datasets in the experiments, of which statistics are shown in Table 1. Specifically, we first construct four synthetic datasets with different dimensions and sparsity, which are named from s1 to s4. Then, we conduct experiments on two real-world datasets, i.e., Movielens[2] and Jester[3]. In Table 1, the sparsity is defined as the percentage of zero components divided by the total number of contextual dimension. For example, given a 0.10 sparsity of s1, the zero components in contextual vectors should be $1000 \times 0.10 = 100$. For non-zero components in s1, we generate values from a standard Gaussian distribution. Note that, for the sparsity of the real-world datasets, we count the percentage of zero components divided by the number of dimension for each feature vector, and then show the average percentage among the whole contextual vectors.

For comparisons, all the datasets are repeated for 10 times in the experiments. Besides, the performance metric of bandit algorithms is the cumulative payoffs with T=1000.

Setting

We conduct all experiments on a server installed with Ubuntu 12.04.5 LTS, which contains 24 processors of each core being Intel CPU@2.60GHz, and has a total memory of 200GB. In experiments, we investigate cumulative payoffs for the CBRAP algorithm with different values of n and m.

With three benchmarks of LinUCB, BallExp and SLUCB, we evaluate CBRAP.SG and CBRAP.RS via the following three questions.

1) Given a fixed high-dimensional space n, do different values of m affect the performance of CBRAP?
2) What is the performance of CBRAP when the sparsity of contextual data increases?
3) What is the performance of CBRAP when it is compared with the three benchmarks?

Results

For the first question, we set the reduced dimension (RD) as $m = 10, 20, 30, 40, 50$ in synthetic datasets. We show the cumulative payoffs for dataset of s1 in Figure 1. From the figure, we find the proposed CBRAP algorithm is flexible with different RD, especially for CBRAP.SG. Similar results of other datasets can be obtained via the source codes[1].

For the second question, we investigate the performance of CBRAP algorithm via the synthetic dataset of s1. The experimental results are shown in Figure 2. We find that CBRAP is stable, which means the sparsity assumption in the previous work can be relaxed.

For the third question, we show the results in Tables 3 and 4. From the table, we know that, when the dimension is high, the proposed CBRAP algorithm outperforms other three benchmarks. Due to space limitation, we just show the results with $m = 20, 50$. For the real-world datasets, we find that the CBRAP algorithm greatly outperforms the SLUCB and BallExp, and slightly better than the LinUCB.

[1]https://github.com/Aaronyxt/CBRAP

[2]http://grouplens.org/datasets/movielens/

[3]http://www.ieor.berkeley.edu/~goldberg/jester-data/

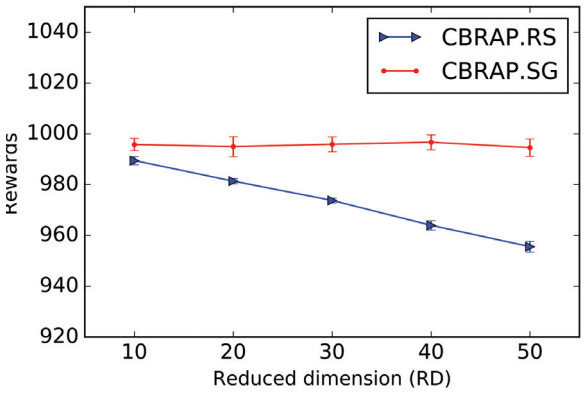

Figure 1: Cumulative payoffs with different reduced dimensions in s1 by adopting `CBRAP.SG` and `CBRAP.RS`.

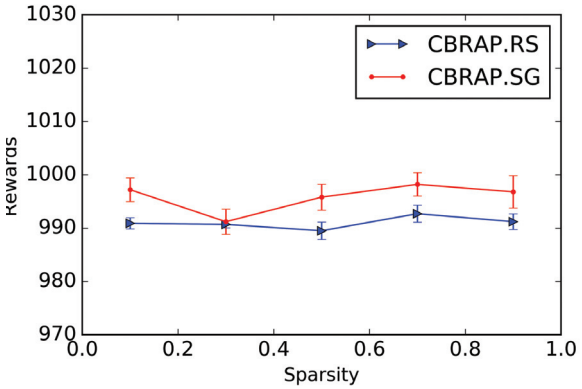

Figure 2: Cumulative payoffs with reduced dimension $m = 10$ and different sparsity in s1.

Besides, we show comparisons of the average of time consumption in experiments in Table 5. From the table, we find that the `CBRAP` algorithm is the most time-efficient.

Overall, the proposed `CBRAP` algorithm is suitable and flexible for bandits with high-dimensional contextual data.

Conclusion

In this paper, we have investigated contextual bandits with linear payoffs by adopting the technique of random projection. There are two main challenges in the most of the current bandit algorithms due to the curse of dimensionality in the big data era. The first is high time-complexity; and the second is increasing upper regret bounds. To solve these two challenges, we proposed an algorithm named `CBRAP` for contextual bandits with linear payoffs. We adopted the technique of random projection in `CBRAP`. For theoretical results, we have derived a practical upper regret bound for `CBRAP`. Finally, experimental results have demonstrated the improved payoffs of `CBRAP`, as well as its time-efficiency.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. The work described in this paper was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CUHK14205214 and No. CUHK14208815 of the General Research Fund), and 2015 Microsoft Research Asia

Table 3: Cumulative payoffs for synthetic datasets with the sparsity being 0.50.

name	algorithm	RD (m)	payoffs ($\mu \pm \tilde{\sigma}$)
s1	`CBRAP.SG`	20	981.4 ± 1.0
		50	955.6 ± 2.2
	`CBRAP.RS`	20	995.0 ± 3.9
		50	994.6 ± 3.4
	`BallExp`	NA	206.8 ± 8.7
	`SLUCB`	NA	312.5 ± 10.1
	`LinUCB`	NA	989.7 ± 2.4
s2	`CBRAP.SG`	20	982.5 ± 1.6
		50	955.6 ± 3.9
	`CBRAP.RS`	20	996.3 ± 3.1
		50	998.8 ± 1.0
	`BallExp`	NA	183.2 ± 5.0
	`SLUCB`	NA	350.6 ± 4.7
	`LinUCB`	NA	975.4 ± 3.1
s3	`CBRAP.SG`	20	982.7 ± 1.3
		50	957.3 ± 2.5
	`CBRAP.RS`	20	991.6 ± 2.9
		50	998.0 ± 2.2
	`BallExp`	NA	280.4 ± 19.0
	`SLUCB`	NA	321.6 ± 5.1
	`LinUCB`	NA	981.2 ± 4.5
s4	`CBRAP.SG`	20	982.4 ± 1.6
		50	957.0 ± 2.6
	`CBRAP.RS`	20	989.4 ± 4.8
		50	998.1 ± 1.8
	`BallExp`	NA	201.5 ± 19.4
	`SLUCB`	NA	298.1 ± 7.5
	`LinUCB`	NA	967.7 ± 6.3

Table 4: Cumulative payoffs for real-world datasets.

name	algorithm	RD (m)	payoffs ($\mu \pm \tilde{\sigma}$)
ML	`CBRAP.SG`	20	2693.0 ± 30.3
		50	2302.1 ± 14.9
	`CBRAP.RS`	20	2303.3 ± 46.2
		50	2034.5 ± 17.8
	`BallExp`	NA	628.6 ± 9.8
	`SLUCB`	NA	500.1 ± 19.0
	`LinUCB`	NA	2595.2 ± 8.5
JR	`CBRAP.SG`	20	1016.5 ± 10.1
		50	1008.9 ± 1.8
	`CBRAP.RS`	20	1028.0 ± 11.2
		50	1065.1 ± 3.7
	`BallExp`	NA	441.2 ± 7.6
	`SLUCB`	NA	323.5 ± 5.3
	`LinUCB`	NA	998.2 ± 10.2

Table 5: Comparisons of the average of time consumption for synthetic dataset of s1 with the sparsity being 0.50.

name	algorithm	RD (m)	time (second)
s1	CBRAP.SG	20	$1,424.9$
		50	$3,285.3$
	CBRAP.RS	20	$1,410.4$
		50	$3,046.9$
	BallExp	NA	$359,555.7$
	SLUCB	NA	$337,197.2$
	LinUCB	NA	$1,633,656.4$

Collaborative Research Program (Project No. FY16-RES-THEME-005).

References

Abbasi-Yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved algorithms for linear stochastic bandits. In *NIPS*, 2312–2320.

Abbasi-Yadkori, Y.; Pal, D.; and Szepesvari, C. 2012. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *AISTATS*, 1–9.

Abe, N., and Long, P. M. 1999. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, 3–11.

Abe, N.; Biermann, A. W.; and Long, P. M. 2003. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica* 37(4):263–293.

Achlioptas, D. 2003. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of computer and System Sciences* 66(4):671–687.

Ailon, N., and Chazelle, B. 2006. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *STOC*, 557–563.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47(2-3):235–256.

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

Baraniuk, R. G.; Cevher, V.; and Wakin, M. B. 2010. Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective. *Proceedings of the IEEE* 98(6):959–971.

Bastani, H., and Bayati, M. 2015. Online decision-making with high-dimensional covariates. *Available at SSRN 2661896*.

Blum, A. 2006. Random projection, margins, kernels, and feature-selection. In *Subspace, Latent Structure and Feature Selection*. Springer. 52–68.

Bubeck, S., and Cesa-Bianchi, N. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*.

Buldygin, V. V., and Kozachenko, Y. V. 1980. Sub-gaussian random variables. *Ukrainian Mathematical Journal* 32(6):483–489.

Carpentier, A.; Munos, R.; et al. 2012. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In *AISTATS*, 190–198.

Chu, W.; Li, L.; Reyzin, L.; and Schapire, R. E. 2011. Contextual bandits with linear payoff functions. In *AISTATS*, 208–214.

Clarkson, K. L., and Woodruff, D. P. 2013. Low rank approximation and regression in input sparsity time. In *STOC*, 81–90.

Dani, V.; Hayes, T. P.; and Kakade, S. M. 2008. Stochastic linear optimization under bandit feedback. In *COLT*, 355–366.

Dasgupta, S., and Gupta, A. 1999. An elementary proof of the johnson-lindenstrauss lemma. *International Computer Science Institute, Technical Report* 99–006.

Deshpande, Y., and Montanari, A. 2012. Linear bandits in high dimension and recommendation systems. In *CCC*, 1750–1754.

Fern, X. Z., and Brodley, C. E. 2003. Random projection for high dimensional data clustering: A cluster ensemble approach. In *ICML*, 186–193.

Filippi, S.; Cappe, O.; Garivier, A.; and Szepesvári, C. 2010. Parametric bandits: The generalized linear case. In *NIPS*, 586–594.

Fodor, I. K. 2002. A survey of dimension reduction techniques. *Technical Report*.

Kabán, A. 2015. Improved bounds on the dot product under random projection and random sign projection. In *SIGKDD*, 487–496.

Kaelbling, L. P. 1994. Associative reinforcement learning: A generate and test algorithm. *Machine Learning* 15(3):299–319.

Langford, J., and Zhang, T. 2008. The epoch-greedy algorithm for multi-armed bandits with side information. In *NIPS*, 817–824.

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *WWW*, 661–670.

Li, P.; Hastie, T. J.; and Church, K. W. 2006. Very sparse random projections. In *SIGKDD*, 287–296.

Lu, Y.; Dhillon, P.; Foster, D. P.; and Ungar, L. 2013. Faster ridge regression via the subsampled randomized hadamard transform. In *NIPS*, 369–377.

Rivasplata, O. 2012. Subgaussian random variables: an expository note. *Internet publication*.

Robbins, H. 1952. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 58(5):527–535.

Shi, Q.; Shen, C.; Hill, R.; and Hengel, A. 2012. Is margin preserved after random projection? In *ICML*, 591–598.

Tang, L.; Rosales, R.; Singh, A.; and Agarwal, D. 2013. Automatic ad format selection via contextual bandits. In *CIKM*, 1587–1594.

Wang, C.-C.; Kulkarni, S. R.; and Poor, H. V. 2005. Bandit problems with side observations. *IEEE Transactions on Automatic Control* 50(3):338–355.

Zhang, W.; Zhang, L.; Jin, R.; Cai, D.; and He, X. 2016. Accelerated sparse linear regression via random projection. In *AAAI*, 2337–2343.

Zhao, T., and King, I. 2016a. Constructing reliable gradient exploration for online learning to rank. In *CIKM*, 1643–1652.

Zhao, T., and King, I. 2016b. Locality-sensitive linear bandit model for online social recommendation. In *ICONIP*, 80–90.

Zhao, T.; McAuley, J. J.; and King, I. 2014. Leveraging social connections to improve personalized ranking for collaborative filtering. In *CIKM*, 261–270.

Appendix

In this appendix, we show the proof of Theorem 2, and discuss the lower regret bound of the proposed algorithm.

Proof of Theorem 2

Before we show the proof of Theorem 2, we show the following lemma, which is directly adopted from (Abbasi-Yadkori, Pál, and Szepesvári 2011).

Lemma 1. *Let* $\pi_t = \mathbf{z}_{t,a_t}^T \boldsymbol{\theta}_{\mathbf{z}}^* + \eta_t + \rho_t$*, where* $\eta_t + \rho_t$ *is* $(\sqrt{4/m} + R)$*-sub-Gaussian. Assume* $\|\boldsymbol{\theta}_{\mathbf{z}}^*\|_2 \leq L_1$ *and* $\|\mathbf{z}_{t,a_t}\|_2 \leq L_2$*. By defining* $\mathbf{A}_t = \gamma \mathbf{I}_{m\times m} + \sum_{i=1}^t \mathbf{z}_{i,a_i}\mathbf{z}_{i,a_i}^T$*,* $\mathbf{b}_t = \sum_{i=1}^t \pi_t \mathbf{z}_{i,a_i}$ *and* $\hat{\boldsymbol{\theta}}_{\mathbf{z}}^t = \mathbf{A}_t^{-1}\mathbf{b}_t$*, with probability at least* $1-\delta$*, for all rounds* $t \geq 0$*, we have*

$$\|\hat{\boldsymbol{\theta}}_{\mathbf{z}}^t - \boldsymbol{\theta}_{\mathbf{z}}^*\|_{\mathbf{A}_t} \leq (\sqrt{4/m+R})\sqrt{m\log(\frac{1+tL_2^2/\gamma}{\delta})} + \gamma^{1/2}L_1. \quad (14)$$

Now we provide the proof of Theorem 2.

Proof. Consider random projection of the contextual vector and the unknown weight vector as

$$\mathbf{z}_{t,a} = \mathbf{M}\mathbf{x}_{t,a}, \quad (15)$$
$$\boldsymbol{\theta}_{\mathbf{z}}^* = \mathbf{M}\boldsymbol{\theta}^*, \quad (16)$$

where $\mathbf{z}_{t,a} \in \mathbb{R}^m$, $\mathbf{x}_{t,a} \in \mathbb{R}^n$, $\boldsymbol{\theta}_{\mathbf{z}}^* \in \mathbb{R}^m$ and $\boldsymbol{\theta}^* \in \mathbb{R}^n$.

For each round, we focus on the error bound for $\pi_{t,a} - \mathbf{x}_{t,a_t}^{\mathrm{T}}\boldsymbol{\theta}^*$. By adding a term in m-dimension space, we have

$$\begin{aligned} r_t &= \pi_{t,a} - \mathbf{x}_{t,a_t}^{\mathrm{T}}\boldsymbol{\theta}^* \\ &= \pi_{t,a} - \mathbf{z}_{t,a_t}^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{z}}^* + \mathbf{z}_{t,a_t}^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{z}}^* - \mathbf{x}_{t,a_t}^{\mathrm{T}}\boldsymbol{\theta}^* \\ &\leq \|\pi_{t,a} - \mathbf{z}_{t,a_t}^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{z}}^* + \mathbf{z}_{t,a_t}^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{z}}^* - \mathbf{x}_{t,a_t}^{\mathrm{T}}\boldsymbol{\theta}^*\|_2 \\ &\leq \|\pi_{t,a} - \mathbf{z}_{t,a_t}^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{z}}^*\|_2 + \|\mathbf{z}_{t,a_t}^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{z}}^* - \mathbf{x}_{t,a_t}^{\mathrm{T}}\boldsymbol{\theta}^*\|_2, \end{aligned}$$

where we adopt the Cauchy-Schwarz inequality.

Now we investigate $\|\pi_{t,a} - \mathbf{z}_{t,a_t}^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{z}}^*\|_2$ and $\|\mathbf{z}_{t,a_t}^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{z}}^* - \mathbf{x}_{t,a_t}^{\mathrm{T}}\boldsymbol{\theta}^*\|_2$ separately.

Since $\boldsymbol{\theta}_{\mathbf{z}}^t$ is optimistic based on $\mathcal{F}_t'$, we have

$$\begin{aligned} &\pi_{t,a} - \mathbf{z}_{t,a_t}^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{z}}^* \\ &\leq \langle \mathbf{z}_{t,a_t}, \tilde{\boldsymbol{\theta}}_{\mathbf{z}}^t\rangle - \langle \mathbf{z}_{t,a_t}, \boldsymbol{\theta}_{\mathbf{z}}^*\rangle \\ &= \langle \mathbf{z}_{t,a_t}, \tilde{\boldsymbol{\theta}}_{\mathbf{z}}^t - \boldsymbol{\theta}_{\mathbf{z}}^*\rangle \\ &= \langle \mathbf{z}_{t,a_t}\hat{\boldsymbol{\theta}}_{\mathbf{z}}^{t-1} - \boldsymbol{\theta}_{\mathbf{z}}^*\rangle + \langle \mathbf{z}_{t,a_t}, \tilde{\boldsymbol{\theta}}_{\mathbf{z}}^t - \hat{\boldsymbol{\theta}}_{\mathbf{z}}^{t-1}\rangle \\ &\leq \|\mathbf{z}_{t,a_t}\|_{\mathbf{A}_t^{-1}}(\|\hat{\boldsymbol{\theta}}_{\mathbf{z}}^{t-1} - \boldsymbol{\theta}_{\mathbf{z}}^*\|_{\mathbf{A}_t^{-1}} + \|\tilde{\boldsymbol{\theta}}_{\mathbf{z}}^t - \hat{\boldsymbol{\theta}}_{\mathbf{z}}^{t-1}\|_{\mathbf{A}_t^{-1}}) \\ &\leq 2\beta_t(\delta)\|\mathbf{z}_{t,a_t}\|_{\mathbf{A}_t^{-1}}, \end{aligned}$$

where $\beta_t(\delta) = (\sqrt{4/m+R})\sqrt{m\log(\frac{1+tL_2^2/\gamma}{\delta})} + \gamma^{1/2}L_1$, $\tilde{\boldsymbol{\theta}}_{\mathbf{z}}^t$ is the optimal parameter for $\mathbf{z}_{t,a_t}$ with the condition of $\|\tilde{\boldsymbol{\theta}}_{\mathbf{z}}^t\|_2 \leq L_1$. Note that the first inequality follows the fact that $\langle \mathbf{z}_{t,a_t}, \tilde{\boldsymbol{\theta}}_{\mathbf{z}}^t\rangle$ is the optimal reward in round t. The last inequality follows Lemma 1.

For $\mathbf{z}_{t,a_t}^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{z}}^* - \mathbf{x}_{t,a_t}^{\mathrm{T}}\boldsymbol{\theta}^*$, based on (Kabán 2015), we have

$$\Pr\{\mathbf{z}^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{z}} < \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta} - \epsilon\} < \exp(-\frac{m\epsilon^2}{8}), \quad (17)$$

$$\Pr\{\mathbf{z}^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{z}} > \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta} + \epsilon\} < \exp(-\frac{m\epsilon^2}{8}), \quad (18)$$

which implies that, with high probability of $1 - 2\exp(-\frac{m\epsilon^2}{8})$,

$$\|\mathbf{z}^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{z}} - \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}\|_2 \leq \epsilon. \quad (19)$$

Thus, for each round t, with probability of $(1-\delta)(1 - 2\exp(-\frac{m\epsilon^2}{8}))$, we have

$$r_t \leq 2\beta_t(\delta)\|\mathbf{z}_{t,a_t}\|_{\mathbf{A}_t^{-1}} + \epsilon. \quad (20)$$

Finally, with probability of $(1-\delta)(1-2\exp(-\frac{m\epsilon^2}{8}))$,

$$\begin{aligned} Regret(T) &\leq \sqrt{T\sum_{t=1}^T r_t^2} \\ &\leq \sqrt{T\sum_{t=1}^T (2\beta_t(\delta)\|\mathbf{z}_{t,a_t}\|_{\mathbf{A}_t^{-1}} + \epsilon)^2} \\ &\leq \sqrt{4T\beta_T^2(\delta)\sum_{t=1}^T \|\mathbf{z}_{t,a_t}\|^2_{\mathbf{A}_t^{-1}} + T^2\epsilon^2 + 4T\epsilon\beta_T(\delta)\sum_{t=1}^T \|\mathbf{z}_{t,a_t}\|_{\mathbf{A}_t^{-1}}} \\ &\approx 2\sqrt{mT\log(1+\frac{TL_2^2}{\gamma m})}\sqrt{\beta_T^2(\delta) + \beta_T(\delta)\epsilon}, \end{aligned}$$

where $\beta_T(\delta) = (\sqrt{4/m+R})\sqrt{m\log(\frac{1+TL_2^2/\gamma}{\delta})} + \gamma^{1/2}L_1$. □

Remarks. Based on the work in (Kabán 2015), we know that the parameters L_1 and L_2 are associated with the reduced dimension m. Thus, the regret bound for `CBRAP` has non-liner relationship with m.

For the lower bound of `CBRAP`, based on (Chu et al. 2011), we can roughly obtain the relationship as

$$Regret(T) \geq \gamma\sqrt{mT},$$

where $\gamma > 0$ is a constant.