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

Misspecified Linear Bandits

Avishek Ghosh
University of California, Berkeley
California 94720 USA
avishek_ghosh@berkeley.edu

Sayak Ray Chowdhury
Indian Institute of Science
Bengaluru 560012 India
srchowdhury@ece.iisc.ernet.in

Aditya Gopalan
Indian Institute of Science
Bengaluru 560012 India
aditya@ece.iisc.ernet.in

Abstract

We consider the problem of online learning in misspecified linear stochastic multi-armed bandit problems. Regret guarantees for state-of-the-art linear bandit algorithms such as Optimism in the Face of Uncertainty Linear bandit (OFUL) hold under the assumption that the arms expected rewards are perfectly linear in their features. It is, however, of interest to investigate the impact of potential misspecification in linear bandit models, where the expected rewards are perturbed away from the linear subspace determined by the arms features. Although OFUL has recently been shown to be robust to relatively small deviations from linearity, we show that any linear bandit algorithm that enjoys optimal regret performance in the perfectly linear setting (e.g., OFUL) must suffer linear regret under a sparse additive perturbation of the linear model. In an attempt to overcome this negative result, we define a natural class of bandit models characterized by a non-sparse deviation from linearity. We argue that the OFUL algorithm can fail to achieve sublinear regret even under models that have non-sparse deviation. We finally develop a novel bandit algorithm, comprising a hypothesis test for linearity followed by a decision to use either the OFUL or Upper Confidence Bound (UCB) algorithm. For perfectly linear bandit models, the algorithm provably exhibits OFULs favorable regret performance, while for misspecified models satisfying the non-sparse deviation property, the algorithm avoids the linear regret phenomenon and falls back on UCBs sublinear regret scaling. Numerical experiments on synthetic data, and on recommendation data from the public Yahoo! Learning to Rank Challenge dataset, empirically support our findings.

1 Introduction

Stochastic multi-armed bandits have been used with significant success to model sequential decision making and optimization problems under uncertainty, due to their succinct expression of the exploration-exploitation tradeoff. Regret is one of the most widely studied performance measures for bandit problems, and it is well-known that the optimal regret that can be achieved in an iid stochastic bandit instance with N actions, $[0, 1]$-bounded rewards and T rounds, without any additional information about the reward distribution, is[1] $\tilde{O}(\sqrt{NT})$. This is achieved, for instance, by the celebrated Upper Confidence Bound (UCB) algorithm of (Auer, Cesa-Bianchi, and Fischer 2002).

The (polynomial) dependence of the regret in a standard stochastic bandit on the number of actions N can be rather prohibitive in settings with a very large number (and potentially infinite) of actions. Under the assumption that the rewards from playing arms are linear functions of known features or context vectors, linear bandit algorithms such as LinUCB (Li et al. 2010), **O**ptimism in the **F**ace of **U**ncertainty **L**inear bandit (OFUL) (Abbasi-Yadkori, Pál, and Szepesvári 2011) and Thompson sampling (Thompson 1933) give regret $\tilde{O}(d\sqrt{T})$ where d is the feature dimension. This is particularly attractive in practice where the feature dimension $d \ll N$ (for instance, news article recommendation data typically has d of the order of hundreds while N is 2 or 3 orders higher). The framework also extends to the more general contextual linear bandit model, where the features for arms are allowed to vary with time (Chu et al. 2011; Agrawal and Goyal 2013).

The design, and attractiveness, of linear bandit algorithms hinges on the assumption that the expected reward from playing arms are linear in their features, i.e., under a fixed ordering of the arms, the vector of expected rewards from all arms belongs to a known linear subspace, spanned by the arms' features. However, real-world environments may not necessarily conform perfectly to this linear reward model and in fact in most cases, have large deviation (Section 7 presents a case study using a real-world dataset to this effect). One possible reason for this is that features are often designed with careful domain expertise without explicit regard for linearity with respect to the utilities of actions. Another situation where linearity ma be violated is when there is feature noise or uncertainty (Hainmueller and Hazlett 2014) – even a small amount of noise in the assumed features shifts the expected reward vector out of the linear subspace. When the rewards need not be perfectly linear in terms of the features in hand, it becomes important to study how robust or fragile strategies for linear bandits can be to such misspecification.

The specific questions we address are: (a) With features available for arms with respect to which the arms' rewards need not necessarily be linear, how do deviations from linearity impact the performance of state-of-the-art linear bandit algorithms? (b) Is it possible to design bandit algorithms

[1]The notation $\tilde{O}$ hides polylogarithmic factors.

Deviation from linearity	OFUL	UCB	RLB (proposed)
Small	$O(d\sqrt{T})$	$O(\sqrt{NT})$	$O(d\sqrt{T})$
Large & non-sparse	$\Omega(T)$	$O(\sqrt{NT})$	$O(\sqrt{NT})$

Table 1: Regret of OFUL, UCB and the proposed algorithm (RLB) upto time horizon T under different deviations. We can see that RLB avoids linear regret of OFUL for large non-sparse deviations while enjoying the favorable regret of OFUL under very small deviations.

that control for deviations from linearity and still enjoy 'best-of-both-worlds' regret performance, i.e., regret that is sublinear in T and depends only on the feature dimension when the model is linear (or near-linear), and that falls back on the number of arms (as for UCB) when all bets are off (i.e., the model is far from linear)?

Overview of results. The paper makes the following contributions:

1. We first prove a negative result about the robustness of linear bandit algorithms to sparse deviations from linearity (Theorem 1): *Any* linear bandit algorithm that enjoys optimal regret guarantees on perfectly linear bandit problem instances (i.e., $O(d\sqrt{T})$ regret in dimension d), such as OFUL and LinUCB, must suffer linear regret on some misspecified linear bandit model. Furthermore our constructive argument shows that it is possible to find a misspecified model that differs only sparsely from a perfectly linear model – in fact, by a perturbation of the expected reward of only a single arm. We also rule out the possibility of using a state-of-the art bandit algorithm OFUL for handling instances with large non-sparse deviation (Theorem 2).
2. Towards overcoming this negative result, we propose and analyze a novel bandit algorithm (Algorithm 1) (abbreviated RLB in Table 1), which is not only robust to non-sparse deviations from linearity but also retains the order-wise optimal regret performance in the standard linear bandit model. The algorithm provably achieves OFUL's $\tilde{O}(d\sqrt{T})$ regret[2] in the ideal linear case, and UCB's $\tilde{O}(\sqrt{NT})$ regret for a broad class of reward models which are not linear but are well-separated from the feature subspace in a non-sparse sense, which we characterize (Theorem 3). The algorithm is comprised of a hypothesis test, followed by a decision to employ either OFUL or UCB. Numerical experiments on both synthetic as well as on the public Yahoo! Learning to Rank Challenge data [3], lend support to our theoretical results.

Related work. Many strategies have been devised and studied for stochastic multi-armed bandits for the general setting without structure – UCB (Auer, Cesa-Bianchi, and Fischer 2002), ϵ-greedy (Cesa-Bianchi and Fischer 1998), Boltzmann exploration (Sutton and Barto 1998), Bayes-UCB (Kaufmann, Garivier, and Cappe 2012), MOSS (Audibert and Bubeck 2009) and Thompson sampling (Thompson 1933; Agrawal and Goyal 2012; Kaufmann, Korda, and Munos 2012), to name a few. Linear stochastic bandits have been extensively investigated (Rusmevichientong and Tsitsiklis 2010; Dani, Hayes, and Kakade 2008; Abbasi-Yadkori, Pál, and Szepesvári 2011) under the well-specified or perfectly linear reward model, achieving (near) optimal problem-independent regret of $\tilde{O}(d\sqrt{T})$ if the features are of dimension d (note that the number of arms can in principle unbounded). Researchers have also considered extensions of linear-bandit algorithms for the case of rewards following a generalized linear model with a known, nonlinear link function (Filippi et al. 2008).

In contrast to the abundance of work on linear bandits, very little work, to the best of our knowledge, has dealt with the impact of misspecification on stochastic decision making with partial (bandit) feedback. A notable study is that of (Besbes and Zeevi 2015) who study misspecified models in a specific dynamic pricing setting. Working in a specialized 2-parameter linear reward setting, they arrive at the conclusion that, within a small range of perturbations of the model away from linearity, one can preserve the sublinear regret of a standard bandit algorithm. There has been significant work, in a different vein, on the effect of model misspecification for the classical linear regression problem (i.e., estimation) in statistics where the metric is overall distortion and not explicitly maximum reward – see for instance the work of (White 1981) and related references. Very recently (Gopalan, Maillard, and Zaki 2016) provides some results for the linear bandit algorithm OFUL when the devation fron linearity is small. We expect to contribute towards filling a much-needed gap in the study of sensitivity properties in linearly parameterized bandit decision-making in this work.

2 Setup & Preliminaries

Consider a multi-armed bandit problem with N arms, and a d-dimensional ($d \ll N$) context or feature vector $x_i \in \mathbb{R}^d$ associated with each arm i, $i = 1, \ldots, N$. An arm i, upon playing, yields a stochastic and independent reward with expectation μ_i. Let $\mu^* = \max_i \mu_i$ be the best expected reward, and let $\mathcal{X}$ be the matrix having the feature vectors for each arm as its columns: $\mathcal{X} = [x_1 \mid x_2 \mid \ldots \mid x_N] \in \mathbb{R}^{d \times N}$, with $\mathcal{X}^T$ assumed to have full column rank. Define $\mu = [\mu_1 \ \mu_2 \ \ldots \ \mu_N]^T \in \mathbb{R}^N$ to be the expected reward vector.

At each time instant $t = 1, 2, \ldots$, the learner chooses any one of the N arms and observes the reward collected from that arm. The action set for the player is $\mathcal{A} = \{1, 2, \ldots, N\}$. The regret after T rounds is defined to be the quantity $R(T) = T\mu^* - \sum_{t=1}^{T} \mu_{A_t}$. The goal of the player is to maximize the net reward, or equivalently, minimize the regret, over the course of T rounds. (If the learner has exact knowledge of θ^* and ϵ beforehand, the optimal choice is to play a best possible arm $i^* = \arg\max_i \mu_i$ at all time instances.)

Under a perfectly linear model, the observed reward Y_t at time t is modeled as the random variable, $Y_t = \langle x_{A_t}, \theta^* \rangle + \eta_t = \mu_{A_t} + \eta_t$, where A_t is the action chosen at time t, $\theta^* \in$

[2] Note that we concern ourselves with studying the gap-independent (worse-case over problem instances) regret; a similar exercise can be carried out in terms of the reward gap parameter.

[3] https://webscope.sandbox.yahoo.com/catalog.php?datatype=c

$\mathbb{R}^d$ is the unknown parameter vector, $\langle ., . \rangle$ denotes the inner product in $\mathbb{R}^d$ and η_t is zero-mean stochastic noise assumed to be conditionally R-sub-Gaussian given A_t. Thus, under a perfectly linear model, the mean reward for each arm is a linear function of its features: there exists a unique $\theta^* \in \mathbb{R}^d$ such that $\mu_i = x_i^T \theta^* \ \forall i \in \mathcal{A}$ (the uniqueness property follows from the full column rank of $\mathcal{X}^T$).

Consider now the case where a linear model for μ with respect to the features $\mathcal{X}$ may not be valid, resulting in a deviation from linearity or a *misspecified* linear bandit model. We model the reward in this case by

$$Y_t = \langle x_{A_t}, \theta \rangle + \epsilon_{A_t} + \eta_t = \mu_{A_t} + \epsilon_{A_t} + \eta_t,$$

where $\theta \in \mathbb{R}^d$ is a choice of weights, and $\epsilon := [\epsilon_1 \ \epsilon_2 \ \dots \ \epsilon_N]^T \in \mathbb{R}^N$ denotes the deviation in the expected rewards of arms. Note that (a) the model remains perfectly linear if[4] $\epsilon \in \text{span}(\mathcal{X}^T) \subseteq \mathbb{R}^N$), and (b) choice of θ satisfying the equation above is not unique if μ is separated from the subspace span($\mathcal{X}^T$), i.e., $\min_{\theta \in \mathbb{R}^d} \parallel \mathcal{X}^T \theta - \mu \parallel_2 > 0$.

3 Lower Bound for Linear Bandit Algorithms under Large Sparse Deviation

In this section, we present our first key result – a general lower bound on regret of any 'optimal' linear bandit algorithm on misspecified problem instances. Specifically, we show that any linear bandit algorithm that enjoys the optimal $O(d\sqrt{T})$ regret scaling, for linearly parameterized models of dimension d, must in fact suffer linear regret under a misspecified model in which only one arm has a mismatched expected reward.

Theorem 1. *Let $\mathbb{A}$ be an algorithm for the linear bandit problem, whose expected regret is $\tilde{O}(d\sqrt{T})$ on any linear problem instance with feature dimension d, time horizon T and expected rewards bounded in absolute value by 1. There exists an instance of a sparsely perturbed linear bandit, with the expected reward of one arm having been perturbed, for which $\mathbb{A}$ suffers linear, i.e., $\Omega(T)$, expected regret.*

The formal proof of Theorem 1 is deferred to the appendix, but we present the main ideas in the following.

Proof sketch. The argument starts by considering a perfectly linear bandit instance with order of $\sqrt{T}$ arms in dimension d. It follows from the regret hypothesis that number of suboptimal arm plays must be $O(\sqrt{T})$. By a pigeonhole argument, since there are order of $\sqrt{T}$ suboptimal arms, there must exist a suboptimal arm that is played no more than $O(1)$ times in expectation. Markov's inequality then gives that the event that both a) this suboptimal arm is played at most $O(1)$ times *and* b) overall regret is $O(d\sqrt{T})$, occurs with probability at least a constant, say $1/3$.

Having isolated a suboptimal arm that is played very rarely by the algorithm (note that the choice of such an arm may very well depend on the algorithm), the argument proceeds by adding a perturbation to this suboptimal arm's reward to make it the best arm in the problem instance. A change-of-measure argument is now used to reason that in the perturbed instance, the probability of the algorithm playing the arm in question does not change significantly as it was anyway played only a constant number of times in the pure linear model. But this must imply that the expected regret is linear due to neglecting the optimal arm in the perturbed problem instance.

[4]For a matrix M, span(M) denotes the subspace spanned by the columns of M.

4 Performance of OFUL Under Deviation

A state-of-the-art algorithm for the linear bandit problem is OFUL. We study the performance of OFUL[5] for various cases of deviations ϵ (suitably "small" and "large"). Specifically, we argue that OFUL is robust to small deviations, but for large deviations, the performance of OFUL is very poor leading to a linear regret scaling. The findings motivate us to propose a more robust algorithm to tackle linear bandit problems with significantly large deviations.

At time $t \geq 1$, based on previous actions and observations upto $t-1$, OFUL solves a regularized linear least squares problem to estimate the unknown parameter $\theta^* \in \mathbb{R}^d$ and constructs a high-confidence ellipsoid around the estimate using concentration-of-measure properties of the sampled rewards. Using the confidence set, the high probability regret of OFUL is $O(d\sqrt{T})$.

4.1 OFUL with Small Deviation

When the deviation from linearity is considerably small, it can be shown that OFUL performs similar to the perfect linear model in terms of regret scaling (see (Gopalan, Maillard, and Zaki 2016, Theorem 3) for details and a formal quantification of "small" deviation). Assuming $||\theta^*||_2 \leq S$, $||x_i||_2 \leq L$ and $|\mu_i| \leq 1$ for all $i \in \mathcal{A}$, with probability at least $1 - \tilde{\delta}$ ($\tilde{\delta} > 0$), the cumulative regret upto time T of OFUL is given by,

$$R_{OFUL}(T) \leq 8\rho' \sqrt{Td \log\left(1 + \frac{TL^2}{\lambda d}\right)} \Big(\lambda^{1/2} S + R\sqrt{2 \log \frac{1}{\tilde{\delta}} + d \log\left(1 + \frac{TL^2}{\lambda d}\right)} \Big)$$

where ρ' is a geometric constant that measures the "distortion" in the arms' actual rewards with respect to (linear) approximation and λ is a regularization parameter.

Remark: OFUL retains $O(d\sqrt{T})$ regret scaling even in the presence of "small" deviation.

4.2 OFUL with Large Sparse Deviation

The regret of OFUL under pure linear bandit instance is $O(d\sqrt{T})$. Therefore from Theorem 1, the cumulative expected regret under large sparse deviation will be $\Omega(T)$.

[5]We consider the OFUL algorithm in this work chiefly because it is known to be the most competitive in terms of regret scaling. It is conceivable that similar results can be shown for other, related, bandit strategies as well, such as ConfidenceBall (Dani, Hayes, and Kakade 2008), UncertaintyEllipsoid (Rusmevichientong and Tsitsiklis 2010), etc.

4.3 OFUL with Large Non-sparse Deviation

We need to identify a natural class of structured large deviations that we dub *non-sparse*. We impose the following structure in terms of sparsity on the expected rewards μ. Recall from Section 2 that $\mathcal{X}$ denotes the context matrix, μ the mean reward vector, θ a choice of weights, and ϵ the deviation from mean μ; thus, $\mu = \mathcal{X}^T\theta + \epsilon$.

Definition 1 (Non-sparse deviation). *Given a feature set $X^f = \{x_1, ..., x_N\} \subset R^d$ and constants $l > 0$, $\beta \in [0, 1]$, an expected reward vector $\mu \in R^N$ is said to have the (l, β) deviation property if,*

$$\mathbb{P}\big(|x_{i_{d+1}}^T [X_{i_1,\ldots,i_d}^f]^{-1} [\mu_{i_1,\ldots,i_d}] - \mu_{i_{d+1}}| \geq l\big) \geq 1 - \beta$$

for all $\{i_1, i_2, \ldots, i_d, i_{d+1}\} \subseteq \{1, 2, \ldots, N\}$, such that $\{x_{i_1}, x_{i_2}, \ldots, x_{i_d}\}$ linearly independent, where $X_{i_1,\ldots,i_d}^f = [x_{i_1}^T, \ldots, x_{i_d}^T]^T$ and $\mu_{i_1,\ldots,i_d} = [\mu_{i_1}, \ldots, \mu_{i_d}]^T$. The randomness is over the choice of $d + 1$ arms.

In other words, the deviation of reward μ is (l, β) non-sparse if, whenever one uses any d linearly independent features, with their corresponding rewards, to regress a $(d+1)$-th unknown reward linearly, then the magnitude of error is at least $l > 0$ (bounded away from 0) with probability at least $1 - \beta$. Typically, β is positive and close to 0.

For example, consider the problem instance of Theorem 1, i.e., only one arm is perturbed away from linearity. This is an example of sparse deviation. If the perturbed arm is picked as one of $d+1$ arms in Definition 1, l will be a large positive number, but when the perturbed arm is missed, l will be 0, which is inconsistent with Definition 1. Also, β can be chosen such that the probability of missing the perturbed arm is strictly greater than β.

We now argue, by counterexample (Theorem 2), that the regret of OFUL with large non-sparse deviation is $\Omega(T)$.

Theorem 2. *Consider a linear bandit problem with $\mathcal{A} = \{1, 2\}$, context matrix $\mathcal{X} = [1\ 2]$, mean reward vector $\mu = [\mu_1\ \mu_2]^T$ with $\mu_2 > \mu_1$ and $\mu_2 \neq 2\mu_1$. The deviation vector $\epsilon = [\epsilon_1\ \epsilon_2]^T$ is such that $|\epsilon_i| > c$ ($c > 0$) for $i = \{1, 2\}$ (with respect to Definition 1, $l = c$ and $\beta = 0$). There exists a problem instance for which the expected regret of OFUL until time T, $\mathbb{E}(R_{OFUL}) = \Omega(T)$.*

The description of problem instance with the formal proof of theorem is deferred to the supplementary material.

Summary: OFUL is robust to "small" deviation (irrespective of sparsity) but incurs linear regret under large deviation (for both sparse and non-sparse). Theorem 1 shows the futility of designing any linear bandit algorithm under sparse deviation. However the quest is still valid if the deviation is large but non-sparse. We will investigate this issue in rest of the paper. It is clear that under large deviation, context vectors do not contribute in reducing regret and thus a rational player should discard contexts under such circumstances. The player may choose any standard algorithm for basic multi-armed bandits (UCB for instance).

5 A Linear Bandit Algorithm Robust to Large, Non-sparse Deviations

This section accomplishes the objective of developing a new algorithm that maintains the sublinear regret property in a model with non-sparse, large deviations. Non-sparse deviations can be seen to naturally arise in the presence of stochastic measurement or estimation noise; e.g., let x_i and $\bar{x_i}$ be the measured and original context vector respectively for arm i with $x_i = \bar{x_i} + \zeta_i$. $\zeta_i^T\theta$ can be modeled as a Gaussian random variable with mean, $\mathbb{E}(\zeta_i^T\theta) = \epsilon_i$. Substituting, we get, $\mu = \mathcal{X}^T\theta + \epsilon$. It is possible to find suitable (l, β) pair (Definition 1) for this model and thus μ is non-sparse. The associated feature vectors corresponding to the mean reward vector satisfying Definition 1, are called "uniformly perturbed features".

We now define 2 hypotheses – $\mathcal{H}_\mathbf{0}$ and $\mathcal{H}_\mathbf{1}$, corresponding intuitively to "linear" and "not linear" – on $(\mathcal{X}, \mu)$, which will be used to quantify the performance of the algorithm developed in this section. We say that hypothesis $\mathcal{H}_\mathbf{0}$ holds if the separation of μ from span$(\mathcal{X}^T)$, i.e., the quantity $\min_{\theta \in \mathbb{R}^d} \parallel \mathcal{X}^T\theta - \mu \parallel_2$, is 0, i.e., the model is perfectly linear. On the other hand, we say that hypothesis $\mathcal{H}_\mathbf{1}$ holds if the separation is greater than 0 and μ satisfies the (l_1, β) deviation property of Definition 1 with $l_1 > 0$.

Remark: The definition of $\mathcal{H}_\mathbf{0}$ be generalized to handle small deviations in the $\| \cdot \|_2$ norm with distortion parameter $\rho' \geq 1$, in the sense of (Gopalan, Maillard, and Zaki 2016, Theorem 3).

5.1 A Robust Linear Bandit (RLB) Algorithm

The sequence of actions for the proposed novel bandit algorithm, namely Robust Linear Bandit (RLB) is summarized in Algorithm 1, mainly consisting of three steps. First, RLB executes an initial sampling phase, in which $d+1$ arms out of N are sampled. Based on these samples, it constructs a confidence ellipsoid for θ^* in the next phase. Finally, based on experimentation on the $(d + 1)$-th arm, it decides to play either OFUL or UCB for the remainder of the horizon. We will illustrate the necessity of non-sparse deviation as follows: consider a problem instance with $\epsilon = (0, \ldots, 0, c, 0, \ldots, 0)$, $|c| \gg 0$. As $N \gg d$, with high probability, the deviated arm can be missed in the sampling phase and according to Algorithm 1, the learner learns that the model is linear and decides to play OFUL which according to Theorem 1 incurs $\Omega(T)$ regret.

Step 1: Sampling of $d + 1$ arms

For non-sparse deviation, the choice of $d + 1$ among N arms may be arbitrary. Without loss of generality, we sample the arms indexed $\{1, 2, \ldots, d+1\}$, k times each (resulting $(d + 1) \times k$ ($:= \tau$) sampling instances). From Hoeffding's inequality, the sample mean estimate of $d + 1$-th arm, $\hat{\mu}_{d+1}$, satisfies $\mathbb{P}(|\hat{\mu}_{d+1} - \mu_{d+1}| > r_s) \leq \exp(-2r_s^2 k)$. With $\delta_s := \exp(-2r_s^2 k)$, the confidence interval around μ_{d+1} will be $[\mu_{d+1} - r_s, \mu_{d+1} + r_s]$ with probability at least $1 - \delta_s$.

Step 2: Construction of Confidence Ellipsoid

Based on the samples of first d arms, RLB constructs a confidence ellipsoid for θ^* assuming $\mathcal{H}_0$ is true. Under $\mathcal{H}_0$,

$$y_i^{(j)} = \langle x_i, \theta^* \rangle + \eta_{i,j} \ \forall i \in \{1, 2, \cdot, d\}, 1 \leq j \leq k.$$

In this setup, we re-define the reward vector $\mathbf{Y} = [y_1^{(1)}, \cdot, y_1^{(k)}, y_2^{(1)}, \cdot, y_2^{(k)}, \ldots, y_d^{(k)}]^T$, feature-matrix $\mathbf{X} = [x_1^T, x_1^T, \cdot, x_1^T, x_2^T, x_2^T, \ldots, x_d^T]^T$ and noise vector $\eta = [\eta_1, \eta_2, \ldots, \eta_{kd}]^T$ with $\mathbf{Y} = \mathbf{X}\theta^* + \eta$. Let $\hat{\theta}$ be the solution of ℓ^2 regularized least square, i.e., $\hat{\theta} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{Y}$, where $\lambda > 0$ is the regularization parameter.

Using the same line of argument as in (Abbasi-Yadkori, Pál, and Szepesvári 2011), it can be shown that for any $\bar{\delta} > 0$, with probability at least $1 - \bar{\delta}$, θ^* lies in the set,

$$C = \left\{\theta \in \mathbb{R}^d : \left\|\hat{\theta} - \theta\right\|_{\bar{V}} \leq R \times \sqrt{2\log(\frac{\det(\bar{V})^{1/2}\det(\lambda I)^{-1/2}}{\bar{\delta}})} + \lambda^{1/2}S\right\}$$

$\bar{V} = \lambda I + k\sum_{i=1}^{d} x_i x_i^T = \mathbf{X}^T\mathbf{X} + \lambda I$, $\|\theta^*\|_2 \leq S$.

Step 3: Hypothesis test for non-sparse deviation

We project the confidence ellipsoid onto the context of $d+1$-th arm. The projection, $\langle x_{d+1}, \theta \rangle, \theta \in C$ will result in an interval, I_e, centered at $x_{d+1}^T\hat{\theta}$ (Lemma **??**). We compare I_e with the interval obtained from sampling $d+1$-th arm, I_s. If $\mathcal{H}_0$ is true, Lemma **??** states that I_e and I_s overlap with high probability. Similarly, from Lemma **??**, under $\mathcal{H}_1$, I_e will not intersect with I_s with high probability, i.e., probability of choosing $\mathcal{H}_1$ when $\mathcal{H}_0$ is true (and vice versa), is significantly low. [6] Based on this experiment, the player adopts the following decision rule: if $I_e \cap I_s \neq \phi$, declare $\mathcal{H}_0$ and play OFUL, otherwise declare $\mathcal{H}_1$ and play UCB.

Algorithm 1 Robust Linear Bandit (RLB)

1: Sample the first d arms k times each.
2: Compute the ℓ^2-regularized least square estimate ($\hat{\theta}$) based on $d \times k$ samples assuming $\mathcal{H}_0$.
3: Construct a confidence ellipsoid C such that with high probability, $\theta^* \in C$.
4: Project the ellipsoid onto the context of $d+1$ th arm to obtain interval I_e.
5: Sample $d+1$ th arm k times, obtain mean estimate, $\hat{\mu}_{d+1}$, and confidence interval I_s.
6: If $I_e \cap I_s \neq \phi$, declare $\mathcal{H}_0$ and play OFUL for the remaining time instants, otherwise play UCB.

6 Regret Analysis

The objective of RLB is to learn the gap from linearity and play accordingly to obtain regret of Table 1. For zero deviation, RLB exploits linear reward structure and incur a regret of $O(d\sqrt{T})$. For large non-sparse deviation, RLB discards the contexts and avoids linear regret. During the initial sampling phase upto τ, regret will scale linearly as each step is either forced exploration or exploitation, i.e., $R_s(\tau) = O(\tau)$. After that, based on the player's decision, either OFUL or UCB is played. For $\mathcal{H}_0$, we use regret of OFUL as given in (Abbasi-Yadkori, Pál, and Szepesvári 2011). With N arms and time T, (Bubeck and Cesa-Bianchi 2012) provided $O(\sqrt{NT\log T})$ regret for standard UCB. Also, (Audibert and Bubeck 2009), gave an algorithm MOSS, inspired by UCB which incurs a regret upper bound of $49\sqrt{NT}$.

6.1 Regret of Algorithm 1

From Lemma (**??**) it can be seen that, if $\mathcal{H}_0$ is true, OFUL and UCB are played with a probability of $1 - \delta_1(k, \lambda)$ and $\delta_1(k, \lambda)$ respectively and accordingly regret is accumulated. By an appropriate choice of k and λ, $\delta_1(k, \lambda)$ can be made arbitrarily close to 0. Similarly, under $\mathcal{H}_1$, corresponding probabilities are $\delta_2(k, \lambda) + \beta$ and $1 - \delta_2(k, \lambda) - \beta$ respectively (Lemma **??**). β comes from the definition of non-sparse deviation. Therefore, under non-sparse deviation, probability of playing OFUL and incurring linear regret is $\delta_2(k, \lambda) + \beta$, which can be pushed to arbitrarily small value by proper choice of k and λ as typically, β is very small and close to 0. τ can be choosen as $\log(T)$, a sub-linear function of T. Now we are in a position to state our main result - an upper bound on regret of RLB.

Theorem 3 (Regret guarantees for RLB)**.** *The expected regret of RLB in T time steps satisfies the following: (a) Under hypothesis $\mathcal{H}_0$,*

$$\begin{aligned}\mathbb{E}(R_{RLB}(T)) &\leq c_1((d+1)k) + 4[(1 - \delta_1(k, \lambda)) \\ &\times\sqrt{(T - \log T)d\log(1 + \frac{(T - \log T)L^2}{\lambda d})}(\lambda^{1/2}S \\ + \quad &R\sqrt{2\log\frac{1}{\delta} + d\log(1 + \frac{(T - \log T)L^2}{\lambda d})})] \\ + \quad &49\delta_1(k, \lambda)\sqrt{N(T - \log T)}\end{aligned}$$

(b) Under $\mathcal{H}_1$,

$$\begin{aligned}\mathbb{E}(R_{RLB}(T)) &\leq c_1((d+1)k) + 49(1 - \delta_2(k, \lambda) - \beta) \\ &\times\sqrt{N(T - \log T)} + c_2(\delta_2(k, \lambda) + \beta)(T - \log T)\end{aligned}$$

where, a total of $d+1$ arms are sampled k times each, λ is regularization parameter, δ, L, c_1, c_2 are constants and,

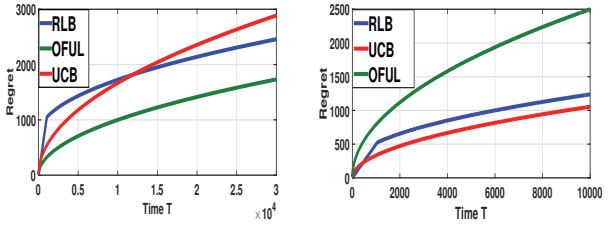
$$\delta_1(k, \lambda) := \exp(-\frac{k(r_s\sqrt{\log k} + r_p(\sqrt{\log k} - 1))^2}{2R^2})$$

$$\delta_2(k, \lambda) := \exp(-\frac{k(l_1 - r_p\sqrt{\log k} - r_s\sqrt{\log k})^2}{2R^2})$$

with $2r_s$ and $2r_p$ being the length of the intervals I_s and I_e respectively and l_1 comes from the definition of $\mathcal{H}_1$.

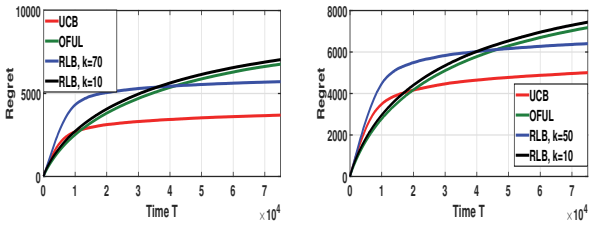
Implication. We see that if k increases, $\delta_1(k, \lambda)$ and $\delta_2(k, \lambda)$ goes to 0 exponentially. Under $\mathcal{H}_1$ and a given (l, β) pair, for RLB to decide in favor of $\mathcal{H}_1$ and hence ensuring sub-linear regret with probability greater than $1 - \delta_2(k, \lambda) - \beta$, we need, $\sqrt{\log k}(r_p + r_s) < l_1$, (shown in the proof of Lemma **??**). Since r_s and r_p are both $O(1/\sqrt{k})$, k satisfies, $k/\log k > b/l_1^2$ for some constant b

[6] Owing to space constraints, Lemma **??**, **??** and **??**, with their proofs are moved to supplementary material.



(a) (b)

Figure 1: Regret variations with synthetic data. Figure (a) represents the scenario with 0 deviation, thus regret of RLB follows that of OFUL. In (b), where deviation is non-sparse, RLB avoids the high regret of OFUL and follows UCB.



(a) (b)

Figure 2: RLB on Yahoo! data: In Figure (a), $N = 500$, $d = 50$. Regret of UCB, OFUL and RLB is plotted for different k. Figure (b) denotes similar plots, with $N = 800$, $d = 50$.

(> 0). Simulations show that a considerably small λ also pushes $\delta_1(k, \lambda)$ and $\delta_2(k, \lambda)$ close to 0. Therefore, with $\mathcal{H}_0$, $R_{RLB}(T) = O(\log T) + O(d\sqrt{T - \log T})$. Similarly, for $\mathcal{H}_1$, $R_{RLB}(T) = O(\log T) + O(\sqrt{N(T - \log T)})$, as shown in Table 1.

7 Simulation Results

7.1 Synthetic Data

In this setup, we assume, $N = 1000$, $d = 20$ and $k = 50$. λ and R are taken as 0.001 and 0.1 respectively. Context vectors and mean rewards are generated at random (in the range $[0, 1]$). All high probability events are simulated with an error probability of 0.001. The simulation is run for 1000 instances and cumulative regret is shown in Figure 1.

Under $\mathcal{H}_0$, RLB predicts correctly with a probability of false alarm 0.0001. Figure 1 shows the regret performance of RLB. In the sampling phase, regret is linear and thus greater than the perturbed OFUL and UCB algorithm. After the sampling phase, regret of RLB closely follows regret of OFUL with probability 0.9999. The false alarm probability can be further pushed if the value of k is increased. If we allow time horizon T to be very large, the deviation in terms of regret between UCB and RLB will be significantly large.

The same experiment is carried for $\mathcal{H}_1$ with $|\epsilon_i| > 2$ for all $i \in \{1, 2, \ldots, N\}$, and RLB verdicts in favor of UCB with an error (miss detection) of 0.0001. Figure 1 shows the variation of regret with time. Further, if k is increased, the error decreases but the regret from sampling phase increases.

7.2 Yahoo! Learning to Rank Data

The performance of RLB is evaluated on the Yahoo! dataset "Learning to Rank Challenge" (Chapelle and Chang 2011). Specifically, we use the file `set2.test.txt`. The dataset consists of query document instance pairs with 103174 rows and 702 columns. The first column lists rating given by user (which we take as reward) with entries $\{0, 1, 2, 3, 4\}$ and the second column captures user id. We treat the rest 700 columns as context vector corresponding to each user. We select $20,000$ rows and 50 columns at random (similar results were found for several random selections). We cluster the data using $\mathcal{K}$-means clustering with $\mathcal{K} = 500$. Each cluster can be treated as a bandit arm with mean reward equal to the empirical mean of the individual rating in the cluster and context (or feature) vector equals to the centroid of the cluster. Thus, we have a bandit setting with $N = 500$, $d = 50$.

To show that the obtained data does not fall in $\mathcal{H}_0$ (i.e., linear model), we fit a linear regression model. It is observed that, average value of residuals (error) is 0.15 (with a maximum value of 0.67), where average mean reward is 1.13. Therefore, we conclude that the data falls under $\mathcal{H}_1$. We run OFUL, UCB and RLB on the dataset and regret performance is shown in Figure 2. We consider the following cases:

1. $k = 70$: we conclude that all arms are sufficiently sampled and thus RLB avoids high regret of OFUL and plays UCB. But RLB suffers high regret upto 3570 rounds.
2. $k = 10$: arms are not properly sampled, leading to an increase in the radius I_s and violating the lower bound on k. Owing to this, RLB plays OFUL and incurs high regret.

We carry out the same experiment with $\mathcal{K} = 800$, i.e., $N = 800$, $d = 50$ and the observations are similar (Figure 2(b)). For a reasonable value of k (50 in this case), RLB properly identifies the optimal algorithm (UCB) to play, but with very low k (10), RLB suffers the high regret of OFUL. We omit the errorbars as over 1000 instances, regret values for different algorithms remain almost the same.

8 Conclusion and Future work

We addressed the problem of adapting to misspecification in linear bandits. We showed that a state-of-the art linear bandit algorithm like OFUL is not always robust to deviations away from linearity. To overcome this, we have proposed a robust bandit algorithm and provided a formal regret upper bound. Experiments on both synthetic and real world datasets support our reasoning that (a) feature-reward maps can often be far from linear in practice, and (b) employing a strategy that is aware of potential deviation from linearity and tests for it suitably does lead to performance gains. Moving forward, it would be interesting to explore other non-linearity structures than sparse deviations as was studied here, and to derive information-theoretic regret lower bounds for the class of general bandit problems with given feature sets. It is also intriguing to investigate the performance of Bayesian-inspired algorithms like Thompson Sampling on linear bandits in presence of deviations.

Acknowledgements

This work was partially supported by the DST INSPIRE faculty grant IFA13-ENG-69. The authors are grateful to anonymous reviewers for providing useful comments.

References

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

Agrawal, S., and Goyal, N. 2012. Analysis of Thompson sampling for the multi-armed bandit problem. *Journal of Machine Learning Research - Proceedings Track* 23:39.1–39.26.

Agrawal, S., and Goyal, N. 2013. Thompson sampling for contextual bandits with linear payoffs. In *ICML*.

Audibert, J.-Y., and Bubeck, S. 2009. Minimax policies for adversarial and stochastic bandits. In *COLT 2009*.

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

Besbes, O., and Zeevi, A. 2015. On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Science* 61(4):723–739.

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

Cesa-Bianchi, N., and Fischer, P. 1998. Finite-time regret bounds for the multiarmed bandit problem. In *In 5th International Conference on Machine Learning*, 100–108. Morgan Kaufmann.

Chapelle, O., and Chang, Y. 2011. Yahoo! learning to rank challenge overview. In *Yahoo! Learning to Rank Challenge*, 1–24.

Chu, W.; Li, L.; Reyzin, L.; and Schapire, R. E. 2011. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, 208–214.

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

Filippi, S.; Cappé, O.; Cérot, F.; and Moulines, E. 2008. A near optimal policy for channel allocation in cognitive radio. In Girgin, S.; Loth, M.; Munos, R.; Preux, P.; and Ryabko, D., eds., *Recent Advances in Reinforcement Learning*, volume 5323 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 69–81.

Gopalan, A.; Maillard, O.-A.; and Zaki, M. 2016. Low-rank bandits with latent mixtures. *arXiv preprint arXiv:1609.01508*.

Hainmueller, J., and Hazlett, C. 2014. Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis* 22(2):143–168.

Kaufmann, E.; Garivier, A.; and Cappe, O. 2012. On bayesian upper confidence bounds for bandit problems. In *AISTATS*.

Kaufmann, E.; Korda, N.; and Munos, R. 2012. Thompson sampling: an asymptotically optimal finite-time analysis. In *ALT*.

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

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

Sutton, R. S., and Barto, A. G. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25:285–294.

White, H. 1981. Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association* 76(374):419–433.