Randomized Exploration for Non-Stationary Stochastic Linear Bandits

Baekjin Kim
Department of Statistics
University of Michigan
Ann Arbor, MI 48109

Ambuj Tewari
Department of Statistics
University of Michigan
Ann Arbor, MI 48109

Abstract

We investigate two perturbation approaches to overcome conservatism that optimism based algorithms chronically suffer from in practice. The first approach replaces optimism with a simple randomization when using confidence sets. The second one adds random perturbations to its current estimate before maximizing the expected reward. For non-stationary linear bandits, where each action is associated with a d-dimensional feature and the unknown parameter is time-varying with total variation B_T, we propose two randomized algorithms, Discounted Randomized LinUCB (D-RandLinUCB) and Discounted Linear Thompson Sampling (D-LinTS) via the two perturbation approaches. We highlight the statistical optimality versus computational efficiency trade-off between them in that the former asymptotically achieves the optimal dynamic regret $\tilde{\mathcal{O}}(d^{2/3}B_T^{1/3}T^{2/3})$, but the latter is oracle-efficient with an extra logarithmic factor in the number of arms compared to minimax-optimal dynamic regret. In a simulation study, both algorithms show outstanding performance in tackling conservatism issue that Discounted LinUCB struggles with.

1 INTRODUCTION

A multi-armed bandit is the simplest model of decision making that involves the exploration versus exploitation trade-off [20]. Linear bandits are an extension of multi-armed bandits where the reward has linear structure with a finite-dimensional feature associated with each arm [2, 13]. Two standard exploration strategies in stochastic linear bandits are Upper Confidence Bound algorithm (LinUCB) [1] and Linear Thomson Sampling (LinTS) [8]. The former relies on optimism in face of uncertainty and is a deterministic algorithm built upon the construction of a high-probability confidence ellipsoid for the unknown parameter vector. The latter is a Bayesian solution that maximizes the expected rewards according to a parameter sampled from the posterior distribution. Chapelle and Li [10] showed that Linear Thompson Sampling empirically performs better and is more robust to corrupted or delayed feedback than LinUCB. From a theoretical perspective, it enjoys a regret bound that is a factor of $\sqrt{d}$ worse than minimax-optimal regret bound $\tilde{\Theta}(d\sqrt{T})$ that LinUCB enjoys. However, the minimax optimality of optimism comes at a cost: implementing UCB type algorithms can lead to NP-hard optimization problems even for convex action sets [7].

Random perturbation methods were originally proposed in the 1950s by Hannan [15] in the full information setting where losses of all actions are observed. Kalai and Vempala [16] showed Hannan's perturbation approach leads to efficient algorithms by making repeated calls to an offline optimization oracle. They also gave a new name to this family of randomized algorithms: Follow the Perturbed Leader (FTPL). Recent works [4, 5, 17] have studied the relationship between FTPL and Follow the Regularized Leader (FTRL) algorithms and also investigated whether FTPL algorithms achieve minimax-optimal regret in full and partial information settings.

Abeille et al. [3] viewed Linear Thompson Sampling as a perturbation based algorithm, characterized a family of perturbations whose regrets can be analyzed, and raised an open problem to find a minimax-optimal perturbation. In addition to its significant role in smartly balancing exploration with exploitation, a perturbation based approach to linear bandits also reduces the problem to one call to the offline optimization oracle in each round. Recent works [18, 19] have proposed randomized algorithms that use perturbation as a means to achieve oracle-efficient computation as well as better theoretical guaran-

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

tee than LinTS, but there is still a gap between their regret bounds and the lower bound of $\Omega(d\sqrt{T})$. This gap is logarithmic in the number of actions which can introduce extra dependence on dimension for large action spaces.

A new randomized exploration scheme was proposed in the recent work of Vaswani et al. [23]. In contrast to Hannan's perturbation approach that injects perturbation directly into an estimate, they replace optimism with random perturbation when using confidence sets for action selection in optimism based algorithms. This approach can be broadly applied to multi-armed bandit and structured bandit problems, and the resulting algorithms are theoretically optimal and empirically perform well since overall conservatism of optimism based algorithms can be tackled by randomizing the confidence level.

Linear bandit problems were originally motivated by applications such as online ad placement with features extracted from the ads and website users. However, users' preferences often evolve with time, which leads to interest in the non-stationary variant of linear bandits. Accordingly, adaptive algorithms that accommodate time-variation of environments have been studied in a rich line of works in both multi-armed bandit [9] and linear bandit. With prior information of total variation budget, SW-LinUCB [12] and D-LinUCB [22] were constructed on the basis of the optimism in face of uncertainty principle via sliding window and exponential discounting weights, respectively. Luo et al. [21] and Chen et al. [11] studied fully adaptive and oracle-efficient algorithms assuming access to an optimization oracle when total variation is unknown for the learner. It is still open problem to design a practically simple, oracle-efficient and statistically optimal algorithm for non-stationary linear bandits.

1.1 CONTRIBUTION

In Section 2, we explicate, in the simpler stationary setting, the role of two perturbation approaches in overcoming conservatism that UCB-type algorithms chronically suffer from in practice. In one approach, we replace optimism with a simple randomization when using confidence sets. In the other, we add random perturbations to the current estimate before maximizing the expected reward. These two approaches result in Randomized LinUCB and Gaussian Linear Thompson Sampling for stationary linear bandits. We highlight the statistical optimality versus oracle efficiency trade-off between them.

In Section 3, we study the non-stationary environment and present two randomized algorithms with exponential discounting weights, Discounted Randomized LinUCB (D-RandLinUCB) and Discounted Linear Thompson Sampling (D-LinTS) to gracefully adjust to the time-variation in the true parameter. We explain the trade-off between statistical optimality and oracle efficiency in that the former asymptotically achieves the optimal dynamic regret $\tilde{\mathcal{O}}(d^{2/3}B_T^{1/3}T^{2/3})$, but the latter enjoys computational efficiency due to sole reliance on an offline optimization oracle for large or infinite action set. However it incurs an extra $(\log K)^{1/3}$ gap in its dynamic regret bound, where K is the number of actions.

In Section 4, we run multiple simulation studies based on Criteo live traffic data [14] to evaluate the empirical performances of D-RandLinUCB and D-LinTS. We observe that the two show outstanding performance in tackling conservatism issue that the non-randomized D-LinUCB struggles with. When high dimension and a large set of actions are considered, in particular, D-LinTS performs as well as Linear Thompson Sampling with prior information on the change-point.

2 WARM-UP: STATIONARY STOCHASTIC LINEAR BANDIT

2.1 PRELIMINARIES

In stationary stochastic linear bandit, a learner chooses an action X_t from a given action set $\mathcal{X}_t \subset \mathbb{R}^d$ in every round t, and he subsequently observes a reward $Y_t = \langle X_t, \theta^\star \rangle + \eta_t$ where $\theta^\star \in \mathbb{R}^d$ is an unknown parameter and η_t is a conditionally 1-subGaussian random variable. For simplicity, assume that $\|\theta^\star\|_2 \leq 1$ and, for all $x \in \mathcal{X}_t$, $\|x\|_2 \leq 1$, and thus $|\langle x, \theta^\star \rangle|_2 \leq 1$.

As a measure of evaluating a learner, the regret is defined as the difference between rewards the learner would have received had it played the best in hindsight, and the rewards actually received. Therefore, minimizing the regret is equivalent to maximizing the expected cumulative reward. Denote the best action in a round t as $x_t^\star = \arg\max_{x \in \mathcal{X}_t} \langle x, \theta^\star \rangle$ and the expected regret as $E[R(T)] = \mathbb{E}\big[\sum_{t=1}^{T} [\langle x_t^\star, \theta^\star \rangle - \langle X_t, \theta^\star \rangle]\big]$.

To learn about unknown parameter $\theta^\star$ from history up to time $t-1$, $\mathcal{H}_{t-1} = \{(X_l, Y_l)_{1 \leq l \leq t-1}\}$, algorithms rely on l^2-regularized least-squares estimate of $\theta^\star$, $\hat{\theta}_t^{ls}$, and confidence ellipsoid centered from $\hat{\theta}_t^{ls}$. We define $\hat{\theta}_t^{ls} = V_{t,\lambda}^{-1} \sum_{l=1}^{t-1} X_l Y_l$, where $V_{t,\lambda} = \lambda I_d + \sum_{l=1}^{t-1} X_l X_l^T$ and λ is a positive regularization parameter.

2.2 RANDOMIZED EXPLORATION

The standard solutions in stationary stochastic linear bandit are optimism based algorithm (LinUCB, Abbasi-Yadkori et al. [1]) and Linear Thompson Sampling (LinTS, Agrawal and Goyal [8]). While the former obtains the theoretically optimal regret bound $\tilde{\mathcal{O}}(d\sqrt{T})$

Table 1: Algorithms in stationary stochastic linear bandits

ALGORITHM	REGRET BOUND	RANDOMNESS	ORACLE ACCESS
LinUCB [1]	$\tilde{\mathcal{O}}(d\sqrt{T})$	No	No
LinTS [8]	$\tilde{\mathcal{O}}(d^{3/2}\sqrt{T})$	Yes	Yes
Gaussian LinTS [19]	$\tilde{\mathcal{O}}(d\sqrt{T\log K})$	Yes	Yes
LinPHE [18]	$\tilde{\mathcal{O}}(d\sqrt{T\log K})$	Yes	Yes
RandLinUCB [23]	$\tilde{\mathcal{O}}(d\sqrt{T})$	Yes	No

matched to lower bound $\Omega(d\sqrt{T})$, the latter empirically performs better in spite of its regret bound $\sqrt{d}$ worse than LinUCB [10]. In finite-arm setting, the regret bound of Gaussian Linear Thompson Sampling (Gaussian-LinTS) is improved by $\sqrt{(\log K)/d}$ as a special case of Follow-the-Perturbed-Leader-GLM (FPL-GLM, Kveton et al. [19]). Also, a series of randomized algorithms for linear bandit were proposed in recent works: Linear Perturbed History Exploration (LinPHE, Kveton et al. [18]) and Randomized Linear UCB (RandLinUCB, Vaswani et al. [23]). They are categorized in terms of regret bounds, randomness, and oracle access in Table 1, where we denote $K = \max_{t\in[T]} |\mathcal{X}_t|$ in finite-arm setting.

There are two families of randomized algorithms according to the way perturbations are used. The first algorithm family is designed to choose an action by maximizing the expected rewards after adding the random perturbation to estimates. Gaussian-LinTS, LinPHE, and FPL-GLM are in this family. But they are limited in that their regret bounds, $\tilde{\mathcal{O}}(d\sqrt{T\log K})$, depend on the number of arms, and lead to $\tilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ regret bounds when the action set is infinite. The other family including RandLinUCB is constructed by replacing the optimism with simple randomization when choosing a confidence level to handle the chronic issue that UCB-type algorithms are too conservative. This randomized version of LinUCB matches optimal regret bounds of LinUCB as well as the empirical performance of LinTS.

Oracle point of view : We assume that the learner has access to an algorithm that returns a near-optimal solution to the offline problem, called an *offline optimization oracle*. It returns the optimal action that maximizes the expected reward from a given action space $\mathcal{X} \subset \mathbb{R}^d$ when a parameter $\theta \in \mathbb{R}^d$ is given as input.

Definition 1 (Offline Optimization Oracle)**.** *There exists an algorithm, $\mathcal{A.M.O.}$, which when given a pair of action space $\mathcal{X} \subset \mathbb{R}^d$, and a parameter $\theta \in \mathbb{R}^d$, computes $\mathcal{A.M.O.}(\mathcal{X}, \theta) = \arg\max_{x\in\mathcal{X}} \langle x, \theta\rangle$.*

Both the non-randomized LinUCB and RandLinUCB are required to compute spectral norms of all actions $\|x\|_{V_{t,\lambda}^{-1}}$ in every round so that they cannot be efficiently implemented with an infinite set of arms. The main advantage of the algorithms in the first family such as Gaussian-LinTS, LinPHE, and FPL-GLM is that they rely on an offline optimization oracle in every round t so that the optimal action can be efficiently obtained within polynomial times from large or even infinite action set.

Improved regret bound of Gaussian LinTS : In FTL-GLM, it is required to generate perturbations and save d-dimensional feature vectors $\{X_l\}_{l=1}^{t-1}$ in order to obtain perturbed estimate $\tilde{\theta}_t$ in every round t, which causes computation burden and memory issue for storage. However, once perturbations are Gaussian in the linear model, adding univariate Gaussian perturbations to historical rewards is the same as perturbing the estimate $\hat{\theta}_t$ by a multivariate Gaussian perturbation because of its linear invariance property, and the resulting algorithm is approximately equivalent to Gaussian Linear Thompson Sampling [8] as follows.

$$\begin{aligned}\tilde{\theta}_t &= \hat{\theta}_t + V_{t,\lambda}^{-1}\sum_{l=1}^{t-1} X_l Z_l^{(t)},\ Z_l^{(t)} \sim \mathcal{N}(0,a^2)\\ &\approx \hat{\theta}_t + V_{t,\lambda}^{-1/2} Z^{(t)},\ Z^{(t)} \sim \mathcal{N}(0,a^2 I_d)\\ &: \textbf{Gaussian-LinTS}.\end{aligned}$$

It naturally implies the regret bound of Gaussian-LinTS is improved by $\sqrt{(\log K)/d}$ with finite action sets [19].

Equivalence between Gaussian LinTS and RandLinUCB : Another perspective of Gaussian-LinTS algorithm is that it is equivalent to RandLinUCB with *decoupled* perturbations across arms due to linearly invariant property of Gaussian random variables:

$$\begin{aligned}\langle x, \tilde{\theta}_t\rangle &= \langle x, \hat{\theta}_t\rangle + x^T V_{t,\lambda}^{-1/2} Z^{(t)},\ Z^{(t)} \sim \mathcal{N}(0,a^2 I_d)\\ &= \langle x, \hat{\theta}_t\rangle + Z_{t,x}\|x\|_{V_{t,\lambda}^{-1}},\ Z_{t,x} \sim N(0,a^2)\\ &: \textbf{\textit{Decoupled} RandLinUCB}.\end{aligned}$$

If perturbations are coupled, we compute the perturbed expected rewards of all actions using randomly chosen confidence level $Z_t \sim N(0, a^2)$ instead of $Z_{t,x}$. In the decoupled RandLinUCB where each arm has its own random confidence level, more variations are generated so that its regret bound have extra logarithmic gap that depends on the number of decoupled actions. In other words, the standard (*coupled*) RandLinUCB enjoys minimax-optimal regret bound due to coupled perturbations. However, there is a cost to its theoretical optimality: it cannot just rely on an offline optimization oracle and thus loses computational efficiency. We thus have a trade-off between efficiency and optimality described in two design principles of perturbation based algorithms.

3 NON-STATIONARY STOCHASTIC LINEAR BANDIT

3.1 PRELIMINARIES

In each round $t \in [T]$, an action set $\mathcal{X}_t \in \mathbb{R}^d$ is given to the learner and it has to choose an action $X_t \in \mathcal{X}_t$. Then, the reward $Y_t = \langle X_t, \theta_t^\star \rangle + \eta_t$ is observed to the learner where $\theta_t^\star \in \mathbb{R}^d$ is an unknown time-varying parameter and η_t is a conditionally 1-subGaussian random variable. The non-stationary assumption allows unknown parameter $\theta_t^\star$ to be time-variant within total variation budget $B_T = \sum_{t=1}^{T-1} \|\theta_t^\star - \theta_{t+1}^\star\|_2$. It is a nice way of quantifying time-variations of $\theta_t^\star$ in that it covers both slowly-changing and abruptly-changing environments. For simplicity, assume $\|\theta_t^\star\|_2 \leq 1$, for all $x \in \mathcal{X}_t$, $\|x\|_2 \leq 1$, and thus $|\langle x, \theta_t^\star \rangle|_2 \leq 1$.

In a similar way to stationary setting, denote the best action in a round t as $x_t^\star = \arg\max_{x \in \mathcal{X}_t} \langle x, \theta_t^\star \rangle$ and denote the expected dynamic regret as $E[R(T)] = \mathbb{E}\big[\sum_{t=1}^T [\langle x_t^\star, \theta_t^\star \rangle - \langle X_t, \theta_t^\star \rangle]\big]$ where X_t is chosen action at time t. The goal of the learner is to minimize the expected dynamic regret.

In a stationary stochastic environment where the reward has a linear structure, Linear Upper Confidence Bound algorithm (LinUCB) follows a principle of optimism in the face of uncertainty (OFU). Under this OFU principle, two recent works of Wu et al. [24] and Russac et al. [22] proposed Sliding Window Linear UCB (SW-LinUCB) and Discounted Linear UCB (D-LinUCB), which are non-stationary variants of LinUCB to adapt to time-variation of $\theta_t^\star$. They rely on weighted least-squares estimators with equal weights only given to recent w observations where w is length of a sliding-window, and exponentially discounting weights, respectively.

Both SW-LinUCB and D-LinUCB achieve the minimax optimal dynamic regret bounds $\Theta(d^{2/3} B_T^{1/3} T^{2/3})$ when B_T is known to the learner, but share inefficiency of implementation with LinUCB [1] in that the computation of spectral norms of all actions are required. Furthermore, they are built upon the construction of a high-probability confidence ellipsoid for the unknown parameter, and thus they are deterministic and their confidence ellipsoids become too wide when high dimensional features are available. In this section, randomization exploration algorithms, Discounted randomized LinUCB (D-RandLinUCB) and Discounted Linear Thompson Sampling (D-LinTS), are proposed to handle computational inefficiency and conservatism that both optimism-based algorithms suffer from. The dynamic regret bound, randomness, and oracle access of algorithms are reported in Table 2.

3.2 WEIGHTED LEAST-SQUARES ESTIMATOR

First, we study the weighted least-squares estimator with discounting factor $0 < \gamma < 1$. In the round t, the weighted least-squares estimator is obtained in a closed form, $\hat{\theta}_t^{wls} = W_{t,\lambda}^{-1} \sum_{s=1}^{t-1} \gamma^{-l} X_l Y_l$ where $W_{t,\lambda} = \sum_{l=1}^{t-1} \gamma^{-l} X_l X_l^T + \lambda \gamma^{-(t-1)} I_d$. Additionally, we define $\tilde{W}_{t,\lambda} = \sum_{l=1}^{t-1} \gamma^{-2l} X_l X_l^T + \lambda \gamma^{-2(t-1)} I_d$. This form is closely connected with the covariance matrix of $\hat{\theta}_t^{wls}$. For simplicity, we denote $V_t = W_{t,\lambda} \tilde{W}_{t,\lambda}^{-1} W_{t,\lambda}$.

Lemma 2 (Weighted Least-Sqaures Confidence Ellipsoid, Theorem 1 [22])**.** *Assume the stationary setting where $\theta_t^\star = \theta^\star$. For any $\delta > 0$,*

$$P\big(\forall t \geq 1, \|\hat{\theta}_t^{wls} - \theta^\star\|_{W_{t,\lambda} \tilde{W}_{t,\lambda}^{-1} W_{t,\lambda}} \leq \beta_t\big) \geq 1 - \delta$$

where $\beta_t = \sqrt{\lambda} + \sqrt{2\log(1/\delta) + d\log(1 + \frac{(1-\gamma^{2t})}{\lambda d(1-\gamma^2)})}$.

While Lemma 2 states that the confidence ellipsoid $\mathcal{C}_t = \{\theta \in \mathbb{R}^d : \|\theta - \theta_t^{wls}\|_{W_{t,\lambda} \tilde{W}_{t,\lambda}^{-1} W_{t,\lambda}} \leq \beta_t\}$ contains true parameter $\theta_t^\star$ with high probability in stationary setting, the true parameter $\theta_t^\star$ is not necessarily inside the confidence ellipsoid $\mathcal{C}_t$ in the non-stationary setting because of variation in the parameters. We alternatively define a *surrogate parameter* $\bar{\theta}_t = W_{t,\lambda}^{-1}(\sum_{l=1}^{t-1} \gamma^{-l} X_l X_l^T \theta_l^\star + \lambda \gamma^{-(t-1)} \theta_t^\star)$, which belongs to $\mathcal{C}_t$ with probability at least $1 - \delta$, which is formally stated in Lemma 4.

3.3 RANDOMIZED EXPLORATION

In this section, we propose two randomized algorithms for non-stationary stochastic linear bandits, Discounted randomized LinUCB (D-RandLinUCB) and Discounted Linear Thompson Sampling (D-LinTS). To gracefully adapt to environmental variation, the weighted method with exponentially discounting factor is directly applied to both RandLinUCB and Gaussian-LinTS, respectively. The random perturbations are injected to D-RandLinUCB and D-LinTS in different fashions: either by replacing optimism with simple randomization in deciding the confidence level or perturbing estimates before maximizing the expected rewards.

Table 2: Algorithms in non-stationary stochastic linear bandits

ALGORITHM	REGRET BOUND	RANDOMNESS	ORACLE ACCESS
D-LinUCB [22]	$\mathcal{O}(d^{\frac{2}{3}} B_T^{\frac{1}{3}} T^{\frac{2}{3}})$	No	No
SW-LinUCB [12]	$\mathcal{O}(d^{\frac{2}{3}} B_T^{\frac{1}{3}} T^{\frac{2}{3}})$	No	No
D-RandLinUCB [This work]	$\mathcal{O}(d^{\frac{2}{3}} B_T^{\frac{1}{3}} T^{\frac{2}{3}})$	Yes	No
D-LinTS [This work]	$\mathcal{O}(d^{\frac{2}{3}} (\log K)^{\frac{1}{3}} B_T^{\frac{1}{3}} T^{\frac{2}{3}})$	Yes	Yes

3.3.1 Discounted Randomized Linear UCB

Following the optimism in face of uncertainty principle, D-LinUCB [22] chooses an action by maximizing the upper confidence bound of expected reward based on $\hat{\theta}_t^{wls}$ and confidence level a. Motivated by the recent work of Vaswani et al. [23], our first randomized algorithm in non-stationary linear bandit setting is constructed by replacing confidence level a with a random variable $Z_t \sim \mathcal{D}$ and this non-stationary variant of RandLinUCB algorithm is called Discounted Randomized LinUCB (D-RandLinUCB, Algorithm 1),

$$\text{D-LinUCB} : X_t = \arg\max_{x \in \mathcal{X}_t} \langle x, \hat{\theta}_t^{wls} \rangle + a \|x\|_{V_t^{-1}}$$

$$\text{D-RandLinUCB} : X_t = \arg\max_{x \in \mathcal{X}_t} \langle x, \hat{\theta}_t^{wls} \rangle + Z_t \|x\|_{V_t^{-1}}.$$

Algorithm 1 Discounted Randomized Linear UCB

Input: $\lambda \geq 1, 0 < \delta < 1, 0 < \gamma < 1$, and $a > 0$
Initialize $W = \lambda I_d$, $\tilde{W} = \lambda I_d$, $\bar{b} = 0$, and $\hat{\theta} = 0$.
for $t = 1$ **to** T **do**
  Randomly sample Z_t from a distribution $\mathcal{D}(\delta, a)$
  Obtain $UCB(x) = x^T \hat{\theta} + Z_t \sqrt{x^T W^{-1} \tilde{W} W^{-1} x}$
  $X_t = \arg\max_{x \in \mathcal{X}_t} UCB(x)$
  Play action X_t and receive reward Y_t
  Update $W = \gamma W + X_t X_t^T + (1-\gamma)\lambda I_d$,
  $\tilde{W} = \gamma^2 \tilde{W} + X_t X_t^T + (1-\gamma^2)\lambda I_d$,
  $\bar{b} = \gamma \bar{b} + X_t Y_t$, $\hat{\theta} = W^{-1}\bar{b}$.
end for

3.3.2 Discounted Linear Thompson Sampling

The idea of perturbing estimates via random perturbation in LinTS algorithm can be directly applied to non-stationary setting by replacing $\hat{\theta}_t^{ls}$ and Gram matrix $V_{t,\lambda}$ with the weighted least-squares estimator $\hat{\theta}_t^{wls}$ and its corresponding matrix $V_t = W_{t,\lambda} \tilde{W}_{t,\lambda}^{-1} W_{t,\lambda}$. We call it Discounted Linear Thompson Sampling (D-LinTS, Algorithm 2). The motivation of D-LinTS arises from its equivalence to D-RandLinUCB with *decoupled* perturbations $Z_{x,t}$ for all $x \in \mathcal{X}_t$ in round t as

$$\begin{aligned}\tilde{f}_t(x) = \langle x, \tilde{\theta}_t^{wls} \rangle &= \langle x, \hat{\theta}_t^{wls} \rangle + x^T W_{t,\lambda}^{-1} \tilde{W}_{t,\lambda}^{1/2} Z^{(t)} \\ &= \langle x, \hat{\theta}_t^{wls} \rangle + Z_{x,t} \|x\|_{V_t^{-1}}\end{aligned}$$

where $Z^{(t)} \sim \mathcal{N}(0_d, a^2 I_d), Z_{x,t} \sim \mathcal{N}(0, a^2)$. Perturbations above are decoupled in that random perturbation are not shared across every arm, and thus they obtain more variation and accordingly $(\log K)^{1/3}$ larger regret bound than that of D-RandLinUCB algorithm that is associated with *coupled* perturbations Z_t. By paying a logarithmic regret gap in terms of K at a cost, the innate perturbation of D-LinTS allows itself to have an offline optimization oracle access in contrast to D-LinUCB and D-RandLinUCB. Therefore, D-LinTS algorithm can be efficient in computation even with an infinite action set.

Algorithm 2 Discounted Linear Thompson Sampling

Input: $\lambda \geq 1, 0 < \gamma < 1$, and $a > 0$
Initialize $W = \lambda I_d$, $\tilde{W} = \lambda I_d$, $\bar{b} = 0$ and $\hat{\theta} = 0$.
for $t = 1$ **to** T **do**
  Obtain $\tilde{\theta} = \hat{\theta} + W^{-1} \tilde{W}^{1/2} Z$, $Z \sim \mathcal{N}(0, a^2 I_d)$
  Oracle : $X_t = \arg\max_{x \in \mathcal{X}_t} \langle x, \tilde{\theta} \rangle$
  Play action X_t and receive reward Y_t
  Update $W = \gamma W + X_t X_t^T + (1-\gamma)\lambda I_d$,
  $\tilde{W} = \gamma^2 \tilde{W} + X_t X_t^T + (1-\gamma^2)\lambda I_d$,
  $\bar{b} = \gamma \bar{b} + X_t Y_t$, $\hat{\theta} = W^{-1}\bar{b}$.
end for

3.4 ANALYSIS

We construct a general regret bound for linear bandit algorithm on the top of prior work of Kveton et al. [18]. The difference from their work is that an action set $\mathcal{X}_t$ varies from time t and can have infinite arms. Also, non-stationary environment is considered where true parameter $\theta_t^\star$ changes within total variation B_T. The expected dynamic regret is decomposed into surrogate regret and bias arising from total variation.

$$\begin{aligned}&E[R(T)] = \sum_{t=1}^{T} E[\langle x_t^\star - X_t, \theta_t^\star \rangle] \\ &= \sum_{t=1}^{T} E[\langle x_t^\star - X_t, \bar{\theta}_t \rangle] + \sum_{t=1}^{T} E[\langle x_t^\star - X_t, \theta_t^\star - \bar{\theta}_t \rangle] \\ &\leq \sum_{t=1}^{T} E[\langle x_t^\star - X_t, \bar{\theta}_t \rangle] + 2\sum_{t=1}^{T} \|\theta_t^\star - \bar{\theta}_t\|_2\end{aligned}$$

3.4.1 Surrogate Instantaneous Regret

To bound the surrogate instantaneous regret $E[\langle x_t^\star - X_t, \bar{\theta}_t\rangle]$, we newly define three events E^{wls}, E_t^{conc}, and E_t^{anti}:

$$
\begin{aligned}
E^{wls} &= \{\forall (x,t) \in \bar{\mathcal{X}}_T; |\langle x, \hat{\theta}_t^{wls} - \bar{\theta}_t\rangle| \le c_1 \|x\|_{V_t^{-1}}\},\\
E_t^{conc} &= \{\forall x \in \mathcal{X}_t; |\tilde{f}_t(x) - \langle x, \hat{\theta}_t^{wls}\rangle| \le c_2 \|x\|_{V_t^{-1}}\},\\
E_t^{anti} &= \{\tilde{f}_t(x_t^\star) - \langle x_t^\star, \hat{\theta}_t^{wls}\rangle > c_1 \|x_t^\star\|_{V_t^{-1}}\},
\end{aligned}
$$

where $\bar{\mathcal{X}}_T = \{(x,t) : x \in \mathcal{X}_t, t \in [T]\}$. The choice of $\tilde{f}_t(x)$ is made by algorithmic design, which decides choices on both c_1 and c_2 simultaneously. In round t, we consider the general algorithm which maximizes perturbed expected reward $\tilde{f}_t(x)$ over action space $\mathcal{X}_t$. The following theorem is a extension of Theorem 1 [18] to the time-evolving environment.

Theorem 3. *Assume we have $\lambda \ge 1$ and $c_1, c_2 \ge 1$ satisfying $P(E^{wls}) \ge 1 - p_1$, $P(E_t^{conc}) \ge 1 - p_2$, and $P(E_t^{anti}) \ge p_3$, and $c_3 = 2d\log(\frac{1}{\gamma}) + 2\frac{d}{T}\log(1 + \frac{1}{d\lambda(1-\gamma)})$. Let A be an algorithm that chooses arm $X_t = \arg\max_{\mathcal{X}_t} \tilde{f}_t(x)$ at time t. Then the expected surrogate instantaneous regret of A, $E[\langle x_t^\star - X_t, \bar{\theta}_t\rangle]$ is bounded by*

$$
p_2 + (c_1 + c_2)\big(1 + \frac{2}{p_3 - p_2}\big) E_t\big[\min(1, \|X_t\|_{V_t^{-1}})\big].
$$

Proof. Firstly, we newly define $\Delta_x = \langle x_t^\star - x, \bar{\theta}_t\rangle$ in round t. Given history $\mathcal{H}_{t-1}$, we assume that event E^{wls} holds and let $\bar{S}_t = \{x \in \mathcal{X}_t : (c_1 + c_2)\|x\|_{V_t^{-1}} \ge \Delta_x \text{ and } \Delta_x \ge 0\}$ be the set of arms that are under-sampled and worse than $x_t^\star$ given $\bar{\theta}_t$ in round t. Among them, let $U_t = \arg\min_{x\in\bar{S}_t} \|x\|_{V_t^{-1}}$ be the least uncertain under-sampled arm in round t. By definition of the optimal arm, $x_t^\star \in \bar{S}_t$. The set of sufficiently sampled arms is defined as $S_t = \{x \in \mathcal{X}_t : (c_1 + c_2)\|x\|_{V_t^{-1}} \le \Delta_x \text{ and } \Delta_x \ge 0\}$ and let $c = c_1 + c_2$. Note that any actions $x \in \mathcal{X}_t$ with $\Delta_x < 0$ can be neglected since the regret induced by these actions are always negative so that it is upper bounded by zero. Given history $\mathcal{H}_{t-1}$, U_t is deterministic term while X_t is random because of innate randomness in $\tilde{f}_t$. Thus surrogate instantaneous regret can be bounded as,

$$
\begin{aligned}
\Delta_{X_t} &= \Delta_{U_t} + \langle U_t, \bar{\theta}_t\rangle - \langle X_t, \bar{\theta}_t\rangle\\
&\le \Delta_{U_t} + \tilde{f}_t(U_t) - \tilde{f}_t(X_t) + c\|X_t\|_{V_t^{-1}} + c\|U_t\|_{V_t^{-1}}\\
&\le c\|X_t\|_{V_t^{-1}} + 2c\|U_t\|_{V_t^{-1}}.
\end{aligned}
$$

Thus, the expected surrogate instantaneous regret can be bounded as,

$$
\begin{aligned}
E_t[\Delta_{X_t}] &= E_t[\Delta_{X_t} I\{E_t^{conc}\}] + E_t[\Delta_{X_t} I\{\bar{E}_t^{conc}\}]\\
&\le cE_t[\|X_t\|_{V_t^{-1}}] + 2c\|U_t\|_{V_t^{-1}} + P_t(\bar{E}_t^{conc})\\
&\le cE_t[\|X_t\|_{V_t^{-1}}] + 2c\|U_t\|_{V_t^{-1}} + p_2\\
&\le cE_t[\|X_t\|_{V_t^{-1}}] + 2c\frac{E_t[\|X_t\|_{V_t^{-1}}]}{P_t(X_t \in \bar{S}_t)} + p_2\\
&= c(1 + \frac{2}{P_t(X_t \in \bar{S}_t)})E_t[\|X_t\|_{V_t^{-1}}] + p_2\\
&\le c(1 + \frac{2}{p_3 - p_2})E_t[\|X_t\|_{V_t^{-1}}] + p_2\\
&\le c(1 + \frac{2}{p_3 - p_2})E_t[\min(1, \|X_t\|_{V_t^{-1}})] + p_2.
\end{aligned}
$$

The third inequality holds because of definition of U_t that is the least uncertain in $\bar{S}_t$ and deterministic as follows,

$$
\begin{aligned}
E_t[\|X_t\|_{V_t^{-1}}] &\ge E_t[\|X_t\|_{V_t^{-1}} | X_t \in \bar{S}_t] \cdot P_t(X_t \in \bar{S}_t)\\
&\ge \|U_t\|_{V_t^{-1}} \cdot P_t(X_t \in \bar{S}_t).
\end{aligned}
$$

The last inequality works because $\lambda_{\min}(V_t) \ge 1$ implies $\|X_t\|_{V_t^{-1}} \le 1$.

The second last inequality holds since on event E_t^{ls},

$$
\begin{aligned}
P_t(X_t \in \bar{S}_t) &\ge P_t\big(\exists x \in \bar{S}_t : \tilde{f}_t(x) \ge \max_{y\in S_t} \tilde{f}_t(y)\big)\\
&\ge P_t\big(\tilde{f}_t(x_t^\star) \ge \max_{y\in S_t} \tilde{f}_t(y)\big)\\
&\ge P_t\big(\tilde{f}_t(x_t^\star) \ge \max_{y\in S_t} \tilde{f}_t(y), E_t^{conc}\big)\\
&\ge P_t\big(\tilde{f}_t(x_t^\star) \ge \langle x_t^\star, \bar{\theta}_t\rangle, E_t^{conc}\big)\\
&\ge P_t\big(\tilde{f}_t(x_t^\star) \ge \langle x_t^\star, \bar{\theta}_t\rangle\big) - P_t\big(\bar{E}_t^{conc}\big)\\
&\ge p_3 - p_2.
\end{aligned}
$$

The fourth inequality holds since for any $y \in S_t$, $\tilde{f}_t(y) \le \langle y, \bar{\theta}_t\rangle + c\|y\|_{V_t^{-1}} \le \langle y, \bar{\theta}_t\rangle + \Delta_y = \langle x_t^\star, \bar{\theta}_t\rangle$. $\square$

In the following three lemmas, the probability of events E^{wls}, E_t^{conc}, and E_t^{anti} can be controlled with optimal choices of c_1 and c_2 for D-RandLinUCB and D-LinTS algorithms.

Lemma 4 (Proposition 3, Russac et al. [22])**.** *For $\lambda > 0$, and $c_1 = \sqrt{2\log T + d\log(1 + \frac{1-\gamma^{2(T-1)}}{\lambda d(1-\gamma^2)})} + \lambda^{1/2}$, the event E^{wls} holds with probability at least $1 - 1/T$.*

Lemma 5 (Concentration)**.** *Given history $\mathcal{H}_{t-1}$,*
(a) D-RandLinUCB : $\tilde{f}_t(x) = \langle x, \hat{\theta}_t^{wls}\rangle + Z_t \cdot \|x\|_{V_t^{-1}}$ where $Z_t \sim \mathcal{N}(0, a^2)$, and $c_2 = a\sqrt{2\log(T/2)}$. Then, $P(\bar{E}_t^{conc}) \le 1/T$.
(b) D-LinTS : $\tilde{f}_t(x) = \langle x, \hat{\theta}_t^{wls}\rangle + x^T W_{t,\lambda}^{-1} \tilde{W}_{t,\lambda}^{1/2} Z^{(t)}$, where $Z^{(t)} \sim \mathcal{N}(0, a^2 I_d)$, and $c_2 = a\sqrt{2\log(KT/2)}$. Then, $P(\bar{E}_t^{conc}) \le 1/T$.

Proof. (a) We have $\tilde{f}_t(x) = \langle x, \hat{\theta}_t^{wls} \rangle + Z_t \|x\|_{V_t}^{-1}$ in D-RandLinUCB algorithm, and thus

$$\begin{aligned}
&P(\bar{E}_t^{conc}) = 1 - P(E_t^{conc}) \\
&= 1 - P(\forall x \in \mathcal{X}_t; |\tilde{f}_t(x) - \langle x, \hat{\theta}_t^{wls} \rangle| \le c_2 \|x\|_{V_t^{-1}}) \\
&= 1 - P(\forall x \in \mathcal{X}_t; |Z_t| \cdot \|x\|_{V_t^{-1}} \le c_2 \|x\|_{V_t^{-1}}) \\
&= 1 - P(|Z_t| \le c_2) \because \text{Lemma 10} \\
&\le 1/T, \text{ where } c_2 = a\sqrt{2\log(T/2)}.
\end{aligned}$$

(b) Given history $\mathcal{H}_{t-1}$, we have $\tilde{f}_t(x) = \langle x, \hat{\theta}_t^{wls} \rangle + x^T W_{t,\lambda}^{-1} \tilde{W}_{t,\lambda}^{1/2} Z^{(t)}$ is equivalent to $\tilde{f}_t(x) = \langle x, \hat{\theta}_t^{wls} \rangle + Z_{t,x} \cdot \|x\|_{V_t}^{-1}$ where $Z_{t,x} \sim \mathcal{N}(0, a^2)$ by the linear invariant property of Gaussian distributions. Thus,

$$\begin{aligned}
&P(\bar{E}_t^{conc}) = 1 - P(E_t^{conc}) \\
&= 1 - P(\forall x \in \mathcal{X}_t; |\tilde{f}_t(x) - \langle x, \hat{\theta}_t^{wls} \rangle| \le c_2 \|x\|_{V_t^{-1}}) \\
&= 1 - P(\forall x \in \mathcal{X}_t; |Z_{t,x}| \cdot \|x\|_{V_t^{-1}} \le c_2 \|x\|_{V_t^{-1}}) \\
&= 1 - P(\forall x \in \mathcal{X}_t; |Z_{t,x}| \le c_2) \because \text{Lemma 10} \\
&\le 1/T, \text{ where } c_2 = a\sqrt{2\log(KT/2)}.
\end{aligned}$$

□

Lemma 6 (Anti-concentration). *Given $\mathcal{H}_{t-1}$,*
(a) D-RandLinUCB : $\tilde{f}_t(x) = \langle x, \hat{\theta}_t^{wls} \rangle + Z_t \|x\|_{V_t^{-1}}$, where $Z_t \sim \mathcal{N}(0, a^2)$. Then, $P(E_t^{anti}) \ge e^{-1/4}/(8\sqrt{\pi})$ when we have $a^2 = 14c_1^2$.
(b) D-LinTS : $\tilde{f}_t(x) = \langle x, \hat{\theta}_t^{wls} \rangle + x^T W_{t,\lambda}^{-1} \tilde{W}_{t,\lambda}^{1/2} Z^{(t)}$ where $Z^{(t)} \sim \mathcal{N}(0, a^2 I_d)$. If we assume $a^2 = 14c_1^2$, then $P(E_t^{anti}) \ge e^{-1/4}/(8\sqrt{\pi})$.

Proof. (a) We denote perturbed expected reward as $\tilde{f}_t(x) = \langle x, \hat{\theta}_t^{wls} \rangle + Z_t \|x\|_{V_t}^{-1}$ for D-RandLinUCB. Thus,

$$\begin{aligned}
P(E_t^{anti}) &= P(\tilde{f}_t(x_t^\star) - \langle x_t^\star, \hat{\theta}_t^{wls} \rangle > c_1 \|x_t^\star\|_{V_t^{-1}}) \\
&= P(Z_t \ge c_1) \ge \exp\big(-7c_1^2/(2a^2)\big)/(8\sqrt{\pi}) \\
&= e^{-1/4}/(8\sqrt{\pi}) \quad \text{where } a^2 = 14c_1^2.
\end{aligned}$$

(b) In the same way as the proof of Lemma 5 (b), $\tilde{f}_t(x) = \langle x, \hat{\theta}_t^{wls} \rangle + x^T W_{t,\lambda}^{-1} \tilde{W}_{t,\lambda}^{1/2} Z^{(t)}$ is equivalent to $\tilde{f}_t(x) = \langle x, \hat{\theta}_t^{wls} \rangle + Z_{t,x} \cdot \|x\|_{V_t}^{-1}$ where $Z_{t,x} \sim \mathcal{N}(0, a^2)$. Thus,

$$\begin{aligned}
&P(E_t^{anti}) = P(\tilde{f}_t(x_t^\star) - \langle x_t^\star, \hat{\theta}_t^{wls} \rangle > c_1 \|x_t^\star\|_{V_t^{-1}}) \\
&= P(Z_{t,x_t^\star} \ge c_1) \ge \exp\big(-7c_1^2/(2a^2)\big)/(8\sqrt{\pi}) \\
&= e^{-1/4}/(8\sqrt{\pi}) \quad \text{where } a^2 = 14c_1^2.
\end{aligned}$$

□

3.4.2 Dynamic Regret

The dynamic regret bound of general randomized algorithm is stated below.

Theorem 7 (Dynamic Regret). *Assume we have $c_1, c_2 \ge 1$ satisfying $P(E^{wls}) \ge 1 - p_1$, $P(E_t^{conc}) \ge 1 - p_2$, and $P(E_t^{anti}) \ge p_3$, and $c_3 = 2d\log(\frac{1}{\gamma}) + 2\frac{d}{T}\log(1 + \frac{1}{d\lambda(1-\gamma)})$. Let A be an algorithm that chooses arm $X_t = \arg\max_{\mathcal{X}_t} \tilde{f}_t(x)$ at time t. The expected dynamic regret of A is bounded as for any integer $D > 0$,*

$$\begin{aligned}
E[R(T)] \le\ & (c_1 + c_2)\big(1 + \frac{2}{p_3 - p_2}\big)\sqrt{c_3 T} \\
& + T(p_1 + p_2) + d + 2DB_T + \frac{4}{\lambda}\frac{\gamma^D}{1-\gamma}T.
\end{aligned}$$

Proof. The dynamic regret bound is decomposed into two terms, (A) expected surrogate regret and (B) bias arising from time variation on true parameter,

$$E[R(T)] \le \sum_{t=1}^T E[\langle x_t^\star - X_t, \bar{\theta}_t \rangle] + 2\sum_{t=1}^T \|\theta_t^\star - \bar{\theta}_t\|_2.$$

The expected surrogate regret term (A) is bounded by

$$\begin{aligned}
&\sum_{t=d+1}^T E[\langle x_t^\star - X_t, \bar{\theta}_t \rangle I\{E^{wls}\}] + T \cdot P(\bar{E}^{wls}) + d \\
&\le (c_1 + c_2)\big(1 + \frac{2}{p_3 - p_2}\big)\sqrt{c_3 T} + T(p_1 + p_2) + d
\end{aligned}$$

The last inequality holds due to Theorem 3 and Lemma 11 in Appendix A.2. For any integer $D > 0$, the bias term (B) is bounded as

$$\begin{aligned}
&(B) = 2\sum_{t=1}^T \|W_{t,\lambda}^{-1} \sum_{l=1}^{t-1} \gamma^{-l} X_l X_l^T (\theta_l^\star - \theta_t^\star)\|_2 \\
&\le 2\sum_{t=1}^T \|W_{t,\lambda}^{-1} \sum_{l=t-D}^{t-1} \gamma^{-l} X_l X_l^T (\theta_l^\star - \theta_t^\star)\|_2 \\
&+ 2\sum_{t=1}^T \|W_{t,\lambda}^{-1} \sum_{l=1}^{t-D-1} \gamma^{-l} X_l X_l^T (\theta_l^\star - \theta_t^\star)\|_2 \\
&\le 2\sum_{t=1}^T \sum_{m=t-D}^{t-1} \|W_{t,\lambda}^{-1} \sum_{l=t-D}^{m} \gamma^{-l} X_l X_l^T (\theta_m^\star - \theta_{m+1}^\star)\|_2 \\
&+ \sum_{t=1}^T \frac{2}{\lambda} \| \sum_{l=1}^{t-D-1} \gamma^{t-l-1} X_l X_l^T (\theta_l^\star - \theta_t^\star)\|_2 \\
&\le 2\sum_{t=1}^T \sum_{m=t-D}^{t-1} \|\theta_m^\star - \theta_{m+1}^\star\|_2 + \frac{4}{\lambda}\frac{\gamma^D}{1-\gamma}T \\
&\le 2DB_T + \frac{4}{\lambda}\frac{\gamma^D}{1-\gamma}T.
\end{aligned}$$

The second inequality holds by interchanging the order of summations and $W_{t,\lambda}^{-2} \preccurlyeq (\frac{\gamma^{t-1}}{\lambda})^2 I_d$. The second last inequality is derived from the fact that for $t - D \le m \le t - 1$, $\lambda_{\max}\Big(W_{t,\lambda}^{-1} \sum_{l=t-D}^{m} \gamma^{-l} X_l X_l^T\Big) \le 1$. □

With the optimal choice of c_1, c_2 and a derived from Lemma 4-6, the dynamic regret bounds of D-RandLinUCB and D-LinTS are stated below.

Corollary 8 (Dynamic Regret of D-RandLinUCB)**.** *Suppose*

$$c_1 = \sqrt{2\log T + d\log(1 + \frac{1-\gamma^{2(T-1)}}{\lambda d(1-\gamma^2)})} + \lambda^{1/2},$$
$$c_2 = a\sqrt{2\log(T/2)}, \textit{and } a^2 = 14c_1^2.$$

Let A be D-RandLinUCB (Algorithm 1). By choosing $D = \log T/(1-\gamma)$ *and* $\gamma = 1 - (B_T/(dT))^{2/3}$*, the expected dynamic regret of A is asymptotically upper bounded by* $\mathcal{O}(d^{2/3}B_T^{1/3}T^{2/3})$ *as* $T \to \infty$*.*

Corollary 9 (Dynamic Regret of D-LinTS)**.** *Suppose*

$$c_1 = \sqrt{2\log T + d\log(1 + \frac{1-\gamma^{2(T-1)}}{\lambda d(1-\gamma^2)})} + \lambda^{1/2},$$
$$c_2 = a\sqrt{2\log(KT/2)}, \textit{and } a^2 = 14c_1^2$$

Let A be D-LinTS (Algorithm 2). By choosing $D = \log T/(1-\gamma)$ *and* $\gamma = 1 - (B_T/(dT\sqrt{\log K}))^{2/3}$*, the expected dynamic regret of A is asymptotically upper bounded by* $\mathcal{O}(d^{2/3}(\log K)^{1/3}B_T^{1/3}T^{2/3})$ *as* $T \to \infty$*.*

The detailed proof of Theorem 7 and Corollary 8 and 9 are deferred to Appendix A.2

Trade-off between Oracle Efficiency and Minimax Optimality : Corollary 8 shows that lower bound for dynamic regret, $\Omega(d^{2/3}B_T^{1/3}T^{2/3})$ is asymptotically matched by D-RandLinUCB, but it is computationally inefficient as D-LinUCB in large action space since the spectral norm of each action in terms of matrix V_t should be computed in every round t. In contrast, D-LinTS algorithm relies on offline optimization oracle access via perturbation and thus can be efficiently implemented in infinite-arm setting, and even contextual bandit setting. As a cost of its oracle efficiency, D-LinTS achieves the dynamic regret bound $(\log K)^{1/3}$ worse than that of D-RandLinUCB in finite-arm setting. There exist two variations in D-LinTS; algorithmic variation generated by perturbing an estimate $\hat{\theta}_t^{wls}$ and environmental variation induced by time-varying environments. Two variations are hard to distinguish from the learner's perspective, and thus the effect of algorithmic variation is alleviated by being partially absorbed in environmental variation. This is why D-LinTS and D-LinUCB produce $d^{1/3}$ gap of dynamic regret bounds with infinite set of arms which is less than $d^{1/2}$ gap between regret bounds of LinUCB and LinTS in the stationary environment.

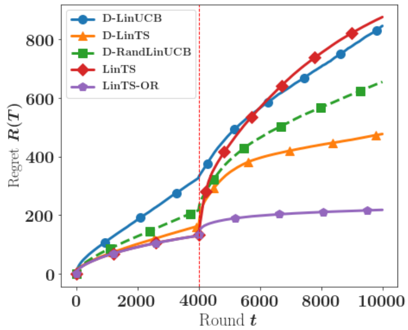
Note that exponentially discounting weights can be replaced by sliding window idea to accommodate to evolving environment so that Sliding-Window Linear UCB (SW-LinUCB) was proposed in the work of Cheung et al. [12]. We can construct Sliding-Window Randomized LinUCB (SW-RandLinUCB) and Sliding-Window Linear Thompson Sampling (SW-LinTS) via two perturbation approaches, and they maintain the trade-off between oracle efficiency and minimax optimality. With unknown total variation B_T, we can also design Bandit-over-Bandit (BOB) algorithm by applying the EXP3 algorithm over SW-RandLinUCB and SW-LinTS with different window sizes [12].

4 NUMERICAL EXPERIMENT

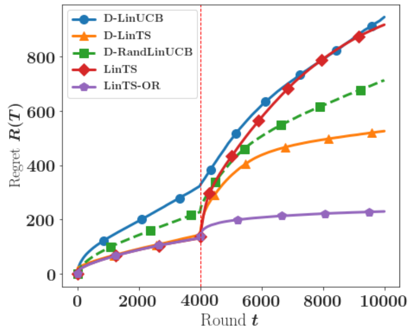
In simulation studies[1], we evaluate the empirical performance of D-RandLinUCB and D-LinTS. We use a sample of 30 days of Criteo live traffic data [14] by 10% downsampling without replacement. Each line corresponds to one impression that was displayed to a user with contextual variables as well as information of whether it was clicked or not. We kept *campaign* variable and categorical variables from *cat1* to *cat9* except for *cat7*. We experiment with several dimensions $d = 10, 20, 50$ and the number of arms $K = 10, 100$. Among all one-hot coded contextual variables, d feature variables were selected by Singular Value Decomposition for dimensionality reduction. We construct two linear models and the model switch occurs at time 4000. The parameter $\theta^\star$ in the initial model is obtained from linear regression model and we obtain true parameter $\theta^\star$ in the second model by switching the signs of 60% of the components of $\theta^\star$. In each round, K arms given to all algorithms are equally sampled from two separate pools of 10000 arms corresponding to clicked or not clicked impressions. The rewards are generated from linear model with additional Gaussian noise of variance $\sigma^2 = 0.15$.

We compare randomized algorithms D-RandLinUCB and D-LinTS to Discounted Linear UCB (D-LinUCB) as a benchmark. Also, we compare them to Linear Thompson Sampling (LinTS) and Oracle Restart LinTS (LinTS-OR). An Oracle Restart knows about the change-point and restarts the algorithm immediately after the change. In D-RandLinUCB, we use Truncated Normal distribution with zero mean and standard deviation $2/5$ over $[0, \infty)$ as $\mathcal{D}$ to ensure that its randomly chosen confidence bound belongs to that of D-LinUCB with high probability. Also, we use non-inflated version by setting $a = 1$ when implementing both LinTS and D-LinTS [23]. The regularization parameter is $\lambda = 1$, the time horizon is $T = 10000$ and the cumulative dynamic regret of algorithms are averaged over 100 independent replications in Figure 1.

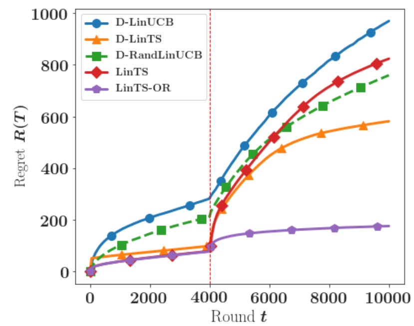
[1] https://github.com/baekjin-kim/NonstationaryLB



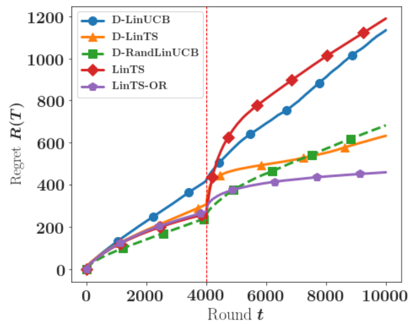
(a) $(d, K) = (10, 10)$



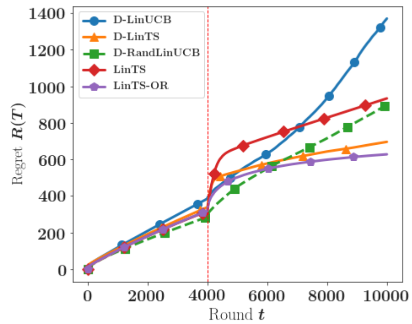
(b) $(d, K) = (20, 10)$



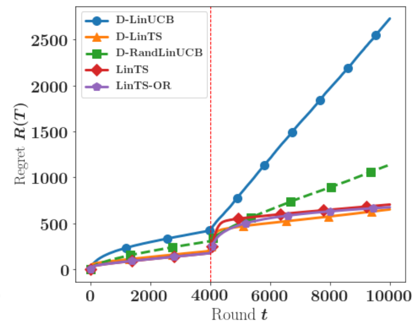
(c) $(d, K) = (50, 10)$



(d) $(d, K) = (10, 100)$



(e) $(d, K) = (20, 100)$



(f) $(d, K) = (50, 100)$

Figure 1: Plots of cumulative dynamic regret for $d = 10, 20$, and 50 and $K = 10$, and 100.

We observe the following patterns in Figure 1. First, two randomized algorithms, D-RandLinUCB and D-LinTS outperform the non-randomized one, D-LinUCB in all scenarios. D-LinTS not only works better than D-RandLinUCB in all scenarios, but also performs as well as LinTS-OR when the large action space and high dimension are considered as shown in Figure 1.(e)-(f).

Second, D-LinUCB produces cumulative dynamic regret less than LinTS does only if linear bandit has low dimension and small action space. Otherwise, it yields cumulative dynamic regret more than or equal to that of LinTS which is designed for stationary environment. The poor performance of D-LinUCB is due to its conservative confidence bound so that the issue regarding conservatism can be partially tackled by randomizing a confidence level in D-RandLinUCB.

Lastly, the interesting observation in Figure 1.(f) is that LinTS without prior information about the change-points performs as well as LinTS with Oracle Restart. This is because in high-dimensional setting, it takes a long time for LinTS-OR to recover a reliable estimate after restarting a naive algorithm at a change-point. On the other hand, the historical observations collected by LinTS are still meaningful though the environment has changed, and it also provides more information on true parameter as larger action space becomes available.

5 CONCLUSION

For non-stationary linear bandits, we propose two randomized algorithms, Discounted Randomized LinUCB and Discounted Linear Thompson Sampling which are the first of their kind by replacing optimism with a simple randomization in UCB-type algorithms, or by adding the random perturbations to estimates, respectively. We analyzed their dynamic regret bounds and evaluated their empirical performance in a simulation study.

The existence of a randomized algorithm that enjoys both theoretical optimality and oracle efficiency is still open in stationary and non-stationary stochastic linear bandits.

ACKNOWLEDGEMENT

We acknowledge the support of NSF CAREER grant IIS-1452099 and the UM-LSA Associate Professor Support Fund.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

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

[3] Marc Abeille, Alessandro Lazaric, et al. Linear thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

[4] Jacob Abernethy, Chansoo Lee, Abhinav Sinha, and Ambuj Tewari. Online linear optimization via smoothing. In *Conference on Learning Theory*, pages 807–823, 2014.

[5] Jacob D Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems*, pages 2197–2205, 2015.

[6] Milton Abramowitz and Irene A Stegun. Handbook of mathematical functions. 1965, 1964.

[7] Shipra Agrawal. Recent advances in multiarmed bandits for sequential decision making. In *Operations Research & Management Science in the Age of Analytics*, pages 167–188. INFORMS, 2019.

[8] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

[9] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems*, pages 199–207, 2014.

[10] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

[11] Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal, and parameter-free. In *Proceedings of the 32nd Annual Conference on Learning Theory*, 2019.

[12] Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Hedging the drift: Learning to optimize under non-stationarity. *arXiv preprint arXiv:1903.01461*, 2019.

[13] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory*, 2008.

[14] Eustache Diemert, Julien Meynet, Pierre Galland, and Damien Lefortier. Attribution modeling increases efficiency of bidding in display advertising. In *Proceedings of the ADKDD'17*, pages 1–6, 2017.

[15] James Hannan. Approximation to bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

[16] Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

[17] Baekjin Kim and Ambuj Tewari. On the optimality of perturbations in stochastic and adversarial multi-armed bandit problems. In *Advances in Neural Information Processing Systems*, pages 2691–2700, 2019.

[18] Branislav Kveton, Csaba Szepesvari, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic linear bandits. In *Proceedings of the 35th Annual Conference on Uncertainty in Artificial Intelligence*, 2019.

[19] Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 2020.

[20] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[21] Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *Proceedings of the 31st Annual Conference on Learning Theory*, 2018.

[22] Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, pages 12017–12026, 2019.

[23] Sharan Vaswani, Abbas Mehrabian, Audrey Durand, and Branislav Kveton. Old dog learns new tricks: Randomized ucb for bandit problems. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 2020.

[24] Qingyun Wu, Naveen Iyer, and Hongning Wang. Learning contextual bandits in a non-stationary environment. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 495–504. ACM, 2018.

A PROOF : NON-STATIONARY SETTING

A.1 LEMMA

Lemma 10 (Concentration and Anti-Concentration of Gaussian distribution [6])**.** *Let Z be the Gaussian random variable with mean μ and variance σ^2. For any $z > 0$,*

$$\frac{1}{4\sqrt{\pi}}\exp(-\frac{7z^2}{2}) \leq P(|Z-\mu| > z\sigma) \leq \frac{1}{2}\exp(-\frac{z^2}{2}).$$

A.2 PROOF OF THEOREM 7

Proof of Theorem 7. The dynamic regret bound is decomposed into two terms, (A) expected surrogate regret and (B) bias arising from variation on true parameter.

$$\begin{aligned} E[R(T)] &= \sum_{t=1}^{T} E[\langle x_t^\star - X_t, \theta_t^\star\rangle] = \sum_{t=1}^{T} E[\langle x_t^\star - X_t, \bar{\theta}_t\rangle] + \sum_{t=1}^{T} E[\langle x_t^\star - X_t, \theta_t^\star - \bar{\theta}_t\rangle] \\ &\leq \sum_{t=1}^{T} E[\langle x_t^\star - X_t, \bar{\theta}_t\rangle] + 2\sum_{t=1}^{T} \|\theta_t^\star - \bar{\theta}_t\|_2 = (A) + (B) \end{aligned}$$

The expected surrogate regret term (A) is bounded as,

$$\begin{aligned} (A) &= \sum_{t=1}^{T} E[\langle x_t^\star - X_t, \bar{\theta}_t\rangle] \leq \sum_{t=d+1}^{T} E[\langle x_t^\star - X_t, \bar{\theta}_t\rangle] + d \\ &\leq \sum_{t=d+1}^{T} E[\langle x_t^\star - X_t, \bar{\theta}_t\rangle I\{E^{wls}\}] + T\cdot P(\bar{E}^{wls}) + d \\ &\leq \sum_{t=d+1}^{T} E[\langle x_t^\star - X_t, \bar{\theta}_t\rangle I\{E^{wls}\}] + Tp_1 + d \\ &\leq (c_1+c_2)\big(1+\frac{2}{p_3-p_2}\big)E_t\big[\sum_{t=d+1}^{T} \min(1, \|X_t\|_{V_t^{-1}})\big] + T(p_1+p_2) + d \quad \because \text{Theorem 3} \\ &\leq (c_1+c_2)\big(1+\frac{2}{p_3-p_2}\big)\sqrt{c_3 T} + T(p_1+p_2) + d \quad \because \text{Cauchy-Schwarz inequality \& Lemma 11} \end{aligned}$$

Lemma 11 (Corollary 4, Russac et al. [22])**.** *For any $\lambda > 0$,*

$$\sum_{t=d+1}^{T} \min(1, \|X_t\|^2_{V_t^{-1}}) \leq c_3 T$$

where $c_3 = 2d\log(1/\gamma) + 2\frac{d}{T}\log(1+\frac{1}{d\lambda(1-\gamma)})$.

The bias term (B) is bounded in terms of total variation, B_T. We first bound the individual bias term at time t. For any integer $D > 0$,

$$\begin{aligned} &\|\theta_t^\star - \bar{\theta}_t\|_2 = \|W_{t,\lambda}^{-1}\sum_{l=1}^{t-1}\gamma^{-l}X_l X_l^T(\theta_l^\star - \theta_t^\star)\|_2 \\ &\leq \|W_{t,\lambda}^{-1}\sum_{l=t-D}^{t-1}\gamma^{-l}X_l X_l^T(\theta_l^\star - \theta_t^\star)\|_2 + \|W_{t,\lambda}^{-1}\sum_{l=1}^{t-D-1}\gamma^{-l}X_l X_l^T(\theta_l^\star - \theta_t^\star)\|_2 \\ &\leq \|W_{t,\lambda}^{-1}\sum_{l=t-D}^{t-1}\gamma^{-l}X_l X_l^T\sum_{m=l}^{t-1}(\theta_m^\star - \theta_{m+1}^\star)\|_2 + \|\sum_{l=1}^{t-D-1}\gamma^{-l}X_l X_l^T(\theta_l^\star - \theta_t^\star)\|_{W_{t,\lambda}^{-2}} \end{aligned}$$

$$
\begin{aligned}
&\leq \|W_{t,\lambda}^{-1}\sum_{m=t-D}^{t-1}\sum_{l=t-D}^{m}\gamma^{-l}X_lX_l^T(\theta_m^\star-\theta_{m+1}^\star)\|_2+\frac{1}{\lambda}\|\sum_{l=1}^{t-D-1}\gamma^{t-l-1}X_lX_l^T(\theta_l^\star-\theta_t^\star)\|_2\\
&\leq \sum_{m=t-D}^{t-1}\|W_{t,\lambda}^{-1}\sum_{l=t-D}^{m}\gamma^{-l}X_lX_l^T(\theta_m^\star-\theta_{m+1}^\star)\|_2+\frac{2}{\lambda}\frac{\gamma^D}{1-\gamma}\\
&\leq \sum_{m=t-D}^{t-1}\lambda_{\max}\Big(W_{t,\lambda}^{-1}\sum_{l=t-D}^{m}\gamma^{-l}X_lX_l^T\Big)\|\theta_m^\star-\theta_{m+1}^\star\|_2+\frac{2}{\lambda}\frac{\gamma^D}{1-\gamma}\leq \sum_{m=t-D}^{t-1}\|\theta_m^\star-\theta_{m+1}^\star\|_2+\frac{2}{\lambda}\frac{\gamma^D}{1-\gamma}
\end{aligned}
$$

The third inequality holds due to $W_{t,\lambda}^{-2} \preccurlyeq (\frac{\gamma^{t-1}}{\lambda})^2 I_d$. The last inequality works due to $\lambda_{\max}\Big(W_{t,\lambda}^{-1}\sum_{l=t-D}^{m}\gamma^{-l}X_lX_l^T\Big) \leq 1$ for $t-D \leq m \leq t-1$. By combining individual bias terms over T rounds, we can derive the upper bound of bias term (B) as,

$$
(B)=2\sum_{t=1}^{T}\|\theta_t^\star-\bar{\theta}_t\|_2\leq 2\sum_{t=1}^{T}\sum_{m=t-D}^{t-1}\|\theta_m^\star-\theta_{m+1}^\star\|_2+\frac{4}{\lambda}\frac{\gamma^D}{1-\gamma}T\leq 2DB_T+\frac{4}{\lambda}\frac{\gamma^D}{1-\gamma}T
$$

Therefore, the expected dynamic regret is bounded as,

$$
\begin{aligned}
E[R(T)] &\leq (A)+(B)\\
&\leq (c_1+c_2)\big(1+\frac{2}{p_3-p_2}\big)\sqrt{c_3T}+T(p_1+p_2)+d+2DB_T+\frac{4}{\lambda}\frac{\gamma^D}{1-\gamma}T
\end{aligned}
$$

In Corollary 8, the choices of a, c_1, c_2, and c_3 are

$$
\begin{aligned}
&a^2=14c_1^2,\ c_1=\sqrt{2\log T+d\log(1+\frac{1-\gamma^{2(T-1)}}{\lambda d(1-\gamma^2)})}+\lambda^{1/2}\\
&c_2=a\sqrt{2\log(T/2)}, \text{and } c_3=2d\log(1/\gamma)+2\frac{d}{T}\log(1+\frac{1}{d\lambda(1-\gamma)}).
\end{aligned}
$$

With optimal choice of $D=\frac{\log T}{1-\gamma}$ and $\gamma=1-d^{-2/3}B_T^{2/3}T^{-2/3}$, the regret of the D-RandLinUCB algorithm is asymptotically upper bounded by $\mathcal{O}(d^{2/3}B_T^{1/3}T^{2/3})$ as $T\to\infty$.

In Corollary 9, the choices of a, c_1, c_2, and c_3 are

$$
\begin{aligned}
&a^2=14c_1^2,\ c_1=\sqrt{2\log T+d\log(1+\frac{1-\gamma^{2(T-1)}}{\lambda d(1-\gamma^2)})}+\lambda^{1/2}\\
&c_2=a\sqrt{2\log(KT/2)}, \text{and } c_3=2d\log(1/\gamma)+2\frac{d}{T}\log(1+\frac{1}{d\lambda(1-\gamma)}).
\end{aligned}
$$

With optimal choice of $D=\frac{\log T}{1-\gamma}$ and $\gamma=1-d^{-2/3}(\log K)^{-1/3}B_T^{2/3}T^{-2/3}$, the regret of the D-LinTS algorithm is asymptotically upper bounded by $\mathcal{O}(d^{2/3}(\log K)^{1/3}B_T^{1/3}T^{2/3})$ as $T\to\infty$. $\square$