

Dynamic Global Sensitivity for Differentially Private Contextual Bandits

HUAZHENG WANG*, Oregon State University, USA

DAVID ZHAO, University of Virginia, USA

HONGNING WANG, University of Virginia, USA

Bandit algorithms have become a reference solution for interactive recommendation. However, as such algorithms directly interact with users for improved recommendations, serious privacy concerns have been raised regarding its practical use. In this work, we propose a differentially private linear contextual bandit algorithm, via a tree-based mechanism to add Laplace or Gaussian noise to model parameters. Our key insight is that as the model converges during online update, the global sensitivity of its parameters shrinks over time (thus named dynamic global sensitivity). Compared with existing solutions, our dynamic global sensitivity analysis allows us to inject less noise to obtain (ϵ, δ) -differential privacy with added regret caused by noise injection in $\tilde{O}(\log T \sqrt{T}/\epsilon)$. We provide a rigorous theoretical analysis over the amount of noise added via dynamic global sensitivity and the corresponding upper regret bound of our proposed algorithm. Experimental results on both synthetic and real-world datasets confirmed the algorithm’s advantage against existing solutions.

CCS Concepts: • **Security and privacy** → **Privacy protections**; • **Theory of computation** → **Sequential decision making**; **Online learning algorithms**.

Additional Key Words and Phrases: Differential privacy, contextual bandits

ACM Reference Format:

Huazheng Wang, David Zhao, and Hongning Wang. 2022. Dynamic Global Sensitivity for Differentially Private Contextual Bandits. In *Sixteenth ACM Conference on Recommender Systems (RecSys '22)*, September 18–23, 2022, Seattle, WA, USA. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3523227.3546781>

1 INTRODUCTION

Multi-armed bandit algorithms have become a reference solution for sequential decision-making; and they have been successfully applied to a wide variety of real-world applications such as recommendation [25], display advertisement [26], and clinical trials [13]. In each iteration of such problems, a learner selects among a set of recommendation candidates, often referred as arms, and receives the corresponding reward after each selection. The learner’s goal is to maximize the cumulative reward over time or equivalently to minimize regret. This requires the learner to both *exploit* the currently estimated best arm and *explore* among the arms to improve its estimation. Contextual bandits extend this setting to that the learner is given a context vector that encodes side information for reward estimation in each iteration.

As bandit algorithms oftentimes directly work with user feedback, e.g., treating user clicks as reward, serious privacy concerns have been raised [3, 33, 34, 37]. We use recommender systems equipped with bandit algorithms as an example

*Work was done while the first author was a PhD student at the University of Virginia.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

to illustrate its risk of privacy breach. In such a system, the algorithm chooses an item for a user, and the user decides whether to click the recommendation based on his/her true preference. Such click feedback is used to update the bandit model for improving its subsequent recommendations. As a result, any change in a user’s preference promptly leads to changes in the algorithm’s output, e.g., different sequences of recommended items. User’s private information, e.g., his/her preference over items in this example, is thus revealed even if the click feedback is kept private. The goal of privacy protection is to prevent the algorithm’s output from revealing a user’s private information, such as item preferences. Real-world privacy breaches have been reported in Amazon’s recommendation system [8] and Facebook’s advertisement system [23], where an adversary can learn considerable side information about a user solely based on the system’s recommendation sequence.

Differential privacy [14, 16] provides a rigorous guarantee that an algorithm’s output has little dependency on the change of input at a single data point, which limits the amount of sensitive information an adversary can infer from the algorithm’s output. It has been widely adopted in both industry and academia. The basic idea is to add a controlled level of noise to an algorithm’s output, such that the subsequent change in its output caused by input change is indistinguishable from that caused by the added noise. The key is to determine the scale of the noise, which depends on the *sensitivity* of the output given the change of input. However, the utility of a private algorithm decreases due to the injected noise. Under the same privacy budget, a good private algorithm adds less noise to its output, preserving more utility while achieving the same level of privacy.

Differential privacy has been studied in stochastic multi-armed bandits [27, 36], adversarial bandits [37], and linear contextual bandits [28, 33]. Because differential privacy is immune to post-processing [16], the output of a bandit algorithm (i.e., the sequence of selected arms) is differentially private once a private mechanism protects sensitive statistics at any internal stage of the algorithm, e.g., collected rewards or estimated bandit parameters. All existing differentially private bandit algorithms add noise to reward or context vectors [28, 33]; but they fail to realize the role of bandit model’s convergence in achieving privacy. Because the model parameter is converging during online update, the same reward change from the user (i.e., the input to the algorithm) will impose smaller impact to the model parameter in the later stage, and thus leads to less change in the algorithm’s output. This property suggests that the model is *less sensitive* to the change of input over time and thus less noise is needed for privacy protection in the later stage.

Based on this important observation, we first study the dynamic nature of sensitivity of the bandit model parameters over time, which we refer to as *dynamic global sensitivity* analysis. We apply the tree-based mechanism [10, 15] with dynamic global sensitivity to add Laplace noise to the model parameter of a linear UCB algorithm [25] to achieve the same level of differential privacy, but with much less added noise compared with previous approaches. We rigorously analyze the level of noise added by the tree-based mechanism using dynamic global sensitivity and prove the upper regret bound of our proposed algorithm. The main contributions of this paper can be summarized as follows:

- We quantify the dynamic global sensitivity of a linear bandit model based on its convergence property. As the model converges, it becomes less sensitive to the reward change over time, and thus less noise is needed to obtain the same level of privacy.
- We propose a tree-based mechanism with dynamic global sensitivity analysis. We show that this mechanism adds $O(\frac{1}{\epsilon} \log T \sqrt{\log T} \log \frac{1}{\zeta} / T)$ noise to the bandit model parameter at time T , and with probability $1 - \zeta$, it achieves (ϵ, δ) -differential privacy. As a result, the added regret caused by noise injection is $\tilde{O}(\log T \sqrt{T} / \epsilon)$ ¹.

¹ $\tilde{O}(\cdot)$ omits the logarithmic terms of $d, 1/\lambda, 1/\zeta, 1/\delta$

- We also empirically evaluate our algorithm on both synthetic and real-world datasets and validate the algorithm’s advantage against existing solutions.

2 RELATED WORK

Multi-armed bandit problem was first studied in [35] and [32]. Auer et al. [5] proposed the Upper Confidence Bound (UCB) method to solve the stochastic multi-armed bandit problem. We refer to [7, 24] for a comprehensive survey regarding stochastic multi-armed bandits and its variants. In this paper, we focus on the linear contextual bandit problem, where each arm is associated with a context feature vector that encodes side information. The reward is governed by a linear function of the context vector, characterized by the corresponding model parameters. UCB-style solution for linear contextual bandit has been popularly studied in [1, 4, 25].

Differential privacy [14] provides a formal notion to quantify the amount of information regarding to an algorithm’s input that an adversary could obtain by observing the algorithm’s output. A common technique is to add Laplace or Gaussian noise to the algorithm’s output. The scale of noise depends on both privacy level (ϵ, δ) and *sensitivity*, which is the change of algorithm’s output caused by the change of its input. Originally studied for static databases, differential privacy was first extended to online setting for stream data in [10, 15]. Differentially private online learning methods have been studied for online convex optimization [2, 3, 34] and bandit problems [6, 27, 28, 33, 37, 39]. The key component of these solutions is the *tree-based mechanism*, which was first proposed in [10, 15] for privately releasing sum statistics in stream data. There are some recent works studied local differential privacy for bandits [19, 31, 41] and DP for federated/distributed bandits [11, 12]. We follow the notion of global differential privacy, and *differential privacy* refers to the global notion in the rest of the paper.

In the bandit setting, an adversary can observe the selected arms (*not* the reward). The goal of a private bandit algorithm is to keep the sequence of reward private. In other words, a change in the reward sequence should not lead to any sensible change in the selected arm sequence. Regarding differentially private contextual bandit, Mishra and Thakurta [27] proposed private versions of contextual UCB and Thompson Sampling algorithms, but they did not provide any theoretical analysis of the resulting algorithms. Neel and Roth [28] provided regret analysis for a private LinUCB algorithm and proved the added regret introduced by their privacy mechanism is $\tilde{O}(\log^{2.5} T \sqrt{T}/\epsilon)$ to achieve ϵ -differential privacy. Shariff and Sheffet [33] studied the setting of making both context and reward private, and adopted a different privacy notion of joint differential privacy [22]. However, these solutions do not directly add noise to the bandit model itself but to the input of the model that has *constant* sensitivity. They thus cannot leverage the convergence property of a bandit model. We study the dynamic nature of sensitivity of bandit model during online update, and propose a tree-based mechanism with dynamic sensitivity to introduce less noise to the model parameter over time. Our proposed method adds additional regret to non-private LinUCB in a scale of $\tilde{O}(\log T \sqrt{T}/\epsilon)$ to achieve (ϵ, δ) -differential privacy. Pichapati et al. [30] proposed an adaptive differentially private SGD algorithm that has similar motivation of our work, i.e., analyzing the dynamics of sensitivity to preserve more utility. But it focuses on a very different perspective: it dynamically (adaptively) computes sensitivity on different dimensions giving the historical gradient, while we compute the dynamic sensitivity based on model convergence.

3 METHOD

We first provide a brief overview of contextual bandit algorithms and differential privacy. We then provide the dynamic global sensitivity analysis of bandit algorithms with a tree based mechanism. Based on it, we present our differentially private linear contextual bandit algorithm and prove its upper regret bound.

3.1 Preliminaries

Contextual Bandits. In a contextual bandit problem, an algorithm sequentially selects an arm a_t from a candidate pool $\mathcal{A}_t = \{a_{1,t}, a_{2,t}, \dots, a_{k,t}\}$, and receives the corresponding reward r_t afterwards. Each arm a is associated with a context feature vector \mathbf{x}_a and its reward is governed by a function of the context feature vector and an unknown bandit model parameter θ^* . We made the following assumption on context generation similar to [18].

ASSUMPTION 1. *Context vectors $C_t = \{\mathbf{x}_a | a \in \mathcal{A}_t\}$ are assumed to be generated i.i.d. conditioned on the past actions $\{a_1, \dots, a_{t-1}\}$ and contexts $\{C_1, \dots, C_{t-1}\}$ from a random process X such that $\mathbb{E}[XX^\top]$ is full rank, with minimal eigenvalue λ_0 ².*

The objective of a bandit algorithm is to maximize its cumulative reward (or equivalently minimize the regret) over a finite time horizon T . To simplify our discussion, we assume the observed reward is a linear function of feature vector \mathbf{x}_a and bandit model parameter θ^* with noise, i.e., $r_t = \mathbf{x}_{a_t}^\top \theta^* + \gamma_t$, where γ_t is a sub-Gaussian feedback noise. We evaluate a bandit algorithm by its pseudo-regret, which is defined as

$$\text{Regret}(T) = \sum_{t=1}^T (\mathbf{x}_{a_t}^\top \theta^* - \mathbf{x}_{a_t}^\top \hat{\theta}), \quad (1)$$

where a_t^* is the best arm according to θ^* . The LinUCB algorithm [1, 25] is a well-studied solution for linear bandit. It solves the bandit model parameter $\hat{\theta}$ using ridge regression, i.e., $\hat{\theta} = \mathbf{A}_t^{-1} \mathbf{b}_t$, where $\mathbf{A}_t = \sum_{i=1}^{t-1} \mathbf{x}_{a_i} \mathbf{x}_{a_i}^\top + \lambda \mathbf{I}$, $\mathbf{b}_t = \sum_{i=1}^{t-1} \mathbf{x}_{a_i} r_i$, λ is the L2-regularization coefficient; and Upper Confidence Bound is used to select an arm.

Differential Privacy. For a contextual bandit problem, denote $S = \{r_{1:T}\}$ as the reward sequence. S' is considered as an adjacent neighboring sequence of S , if it only differs from S at one point of reward r_i . The output O of a bandit algorithm is the sequence of its selected arms, i.e., $\{a_{1:T}\}$.

DEFINITION 1 (DIFFERENTIAL PRIVACY [14]). *A randomized algorithm \mathcal{M} is (ϵ, δ) -differentially private if for any adjacent neighboring sequences S, S' and output O ,*

$$\mathbb{P}(\mathcal{M}(S) \in O) \leq e^\epsilon \mathbb{P}(\mathcal{M}(S') \in O) + \delta$$

When $\delta = 0$, we say algorithm \mathcal{M} is ϵ -differentially private.

Differential privacy ensures the adversary observes almost the same output of a private algorithm in a probabilistic sense, if one input data point is changed, where the similarity between outputs is evaluated by ϵ and δ . Laplace and Gaussian noise are commonly used as additive noise to protect the output, where the noise scale is related to the privacy requirement (ϵ, δ) and the *global sensitivity* of algorithm's output caused by the change of input. We formally define global sensitivity below.

DEFINITION 2 (GLOBAL SENSITIVITY [14]). *For any adjacent neighboring sequences S, S' , global sensitivity of a function f is defined as,*

$$\Delta_f = \max_{S, S'} |f(S) - f(S')|$$

Since differential privacy is immune to post-processing [16], previous solutions of private linear bandits [27, 28] add noise to $\sum_t \mathbf{x}_{a_t} r_t$ to obtain privacy, which consequently makes the model parameter θ_t and the selected arms $\{a_t\}_{t=1}^T$

²We discuss the impact of this assumption in next section in detail.

Algorithm 1 Tree-based mechanism with Dynamic Global Sensitivity

```

1: Inputs:  $t, T, \epsilon, \delta, L, \lambda_0$ 
2: Initialize:  $\epsilon' = \epsilon/\log_2 T, \delta' = \delta/\log_2 T, \eta_t = 0$ 
3: Let  $b$  be the  $\lceil \log_2 T \rceil + 1$ -bit binary representation of  $t$ 
4: for  $i = 0$  to  $\lceil \log_2 T \rceil + 1$  do
5:   if  $b_i = 1$  then
6:      $\eta_t = \eta_t + \text{Lap}(\frac{\Delta_t - 2^{i+1}}{e^i})$ 
7:   end if
8: end for
9: Return noise  $\eta_t$ 

```

private. It is obvious that global sensitivity of $b_t = \sum_t \mathbf{x}_{a_t} r_t$ is a constant L for all round t , if we assume $\|\mathbf{x}\|_2 \leq L$ and $|r| \leq 1$, without loss of generality. In other words, constant amount of noise has to be added in each round to achieve differential privacy.

3.2 Dynamic Global Sensitivity Analysis for Tree-based Mechanism

Different from previous work, we directly add noise to the estimated bandit model parameter $\hat{\theta}_t$ after each round of model update. As $\hat{\theta}_t$ converges over time, the sensitivity of it decreases consequently, which we name as *dynamic global sensitivity*. We first quantify such sensitivity of $\hat{\theta}_t$ and then discuss how to combine it with the tree-based mechanism.

LEMMA 1 (SENSITIVITY OF ESTIMATED BANDIT MODEL PARAMETER $\hat{\theta}_t$). *Let $\hat{\theta}_t$ and $\hat{\theta}'_t$ be parameter estimations of adjacent neighboring reward sequences S and S' , assuming context vectors $\{\mathbf{x}_{1,t}, \dots, \mathbf{x}_{K,t}\}$ are generated according to Assumption 1. With probability at least $1 - \delta$, the dynamic global sensitivity of bandit model parameter $\hat{\theta}_t$ at time t is bounded as,*

$$\Delta_t = \max_{S, S'} \|\hat{\theta}_{t+1} - \hat{\theta}'_{t+1}\|_2 \leq \frac{2L}{\lambda'} \quad (2)$$

where $\lambda' = \lambda_0 t / 4 - 8 \log((t+3)/\delta) - 2\sqrt{t \log((t+3)/\delta)}$ is the lower bound of minimum eigenvalue of matrix $\mathbf{A}_t = \sum_t \mathbf{x}_{a_t} \mathbf{x}_{a_t}^\top + \lambda \mathbf{I}$.³ We use a simplified bound $\lambda' = \lambda_0 t / 16$ when $t > 32 \log(1/\delta) / \lambda_0$, which gives us $\Delta_t \leq 32L / \lambda_0$.

Proof Sketch. Since $\hat{\theta}_t = \mathbf{A}_t^{-1} \sum_t \mathbf{x}_{a_t} r_t$, the key idea is to quantify the minimum eigenvalue of matrix \mathbf{A}_t . Here we adopt the i.i.d. assumption of context vectors from Theorem 1 of [18] to analyze the corresponding eigen system and the key idea is to use a concentration inequality to bound the eigenvalue.

Note that we need to avoid negative sensitivity when t is small. It can be derived from the formula of λ' that by choosing probability $\delta > 2 \log(3) / (\sqrt{1 + 2\lambda_0} + 1)$, the sensitivity is guaranteed to be positive for small t . For large t , we have shown in the lemma that the sensitivity is guaranteed to be positive when $t > 32 \log(1/\delta) / \lambda_0$.

Lemma 1 provides the dynamic global sensitivity analysis of estimated bandit model parameter $\hat{\theta}_t$. The most important property of it is that it is monotonically shrinking over time. This indicates less noise is needed in later stage for privacy protection. We should emphasize that our analysis is for *global* sensitivity. A similar but quite different notion is *local* sensitivity (and smoothed sensitivity) proposed by [29], which studies the dynamic nature of sensitivity given different input S , i.e., $\max_{S'} |f(S) - f(S')|$. Our dynamic sensitivity analysis is not conditioned on any reward sequence, but based on the convergence property of bandit model parameter $\hat{\theta}_t$. And thus it holds for any reward sequence as input.

Remark Our Assumption 1 on the context vectors used in Lemma 1 follows [18] and is used to bound the rate of shrinkage of the bandit model's sensitivity. Similar assumptions are also discussed in [21] from the perspective of

³Note that λ is the coefficient for of L_2 regularization, and it not related with λ_0 .

perturbation on context vectors via a smoothed analysis. And in the appendix, we show that a similar sensitivity analysis can easily lead to a differentially private version of algorithm proposed in [21]. In the meanwhile, we also note that this environment assumption generally holds in practice, especially in a system with time-varying arm sets, as extensively discussed and studied in [21]. However, even if the context vectors are generated by an adversary instead of a random process, we can still achieve the same guarantee on the minimal eigenvalue λ_0 and the same result in Lemma 1 by perturbing the context vector \mathbf{x}_a with a Gaussian noise sampled from $\mathcal{N}(0, \sigma^2)$, such that $\lambda_0 \geq \Omega(\sigma^2)$, based on Lemma 3.7 in [21]. In the following, we assume Assumption 1 holds if not specified. And in Corollary 1, we discuss the impact on regret if the algorithm needs to perturb the context vector when the assumption does not hold.

The analyzed sensitivity provides us the level of noise needed for differential privacy. We present a tree-based mechanism with dynamic global sensitivity in Algorithm 1, which takes advantage of the sublinear property of exponential function to further reduce the amount of added noise. Basically, the idea of tree-based mechanism is to view the sum statistics as $\lceil \log_2 T \rceil + 1$ partial sums. Laplace noise is added to each partial sum to achieve (ϵ', δ') -differential privacy where $\epsilon' = \epsilon / \log_2 T$, $\delta' = \delta / \log_2 T$. Here we follow the notion of (ϵ', δ') -differential privacy because our sensitivity analysis in Lemma 1 is a high probability bound and holds with probability $1 - \delta'$. Based on composition theorem of differential privacy [17, 20], the final output, the sum of (ϵ', δ') -differentially private partial sums, is (ϵ, δ) -differentially private. The partial sums are segmented by the tree representation. We refer to [10, 15] for detailed discussion of tree-based mechanism.

However, previous differentially private bandit algorithms with tree-based mechanism only protect variables that have constant sensitivities, such as sum statistics of $\sum_t \mathbf{x}_a r_t$ for contextual bandits [27, 28, 33] or sum of rewards for non-contextual bandits [36, 37], instead of our dynamic global sensitivity. In Algorithm 1, we scale the noise of partial sum $[t - 2^{i+1}, t - 2^i + 1]$ with its dynamic sensitivity $\Delta_{t-2^{i+1}}$ as shown in line 6.

We first show the privacy guarantee of Algorithm 1 and then study its utility, i.e., the total amount of noise added by Algorithm 1.

THEOREM 1 (PRIVACY). *Algorithm 1 with dynamic global sensitivity Δ_i defined in Lemma 1 is (ϵ, δ) -differentially private.*

THEOREM 2 (UTILITY). *For a finite time horizon T , at time t , Algorithm 1 with dynamic global sensitivity of bandit model parameter $\hat{\theta}$ adds $O(\frac{1}{\epsilon} \log T \sqrt{\log t} \log \frac{1}{\zeta} / t)$ noise with probability $1 - \zeta$ to achieve (ϵ, δ) -differential privacy.*

Proof Sketch. We follow the definition of differential privacy and use a similar proof as Theorem 3 of [34] to get the privacy guarantee. The utility analysis is based on Lemma 1 and the property of sum of samples from Laplace distributions.

All existing tree-based mechanism for private bandit solutions add noise at a scale of $O(\frac{1}{\epsilon} \log T \sqrt{\log t} \log \frac{1}{\zeta})$ based on constant sensitivity [10]. By leveraging the dynamic global sensitivity analysis in Lemma 1, our solution adds less noise while achieving the same level of privacy.

Moreover, our finite-time analysis assumes the knowledge of time horizon T beforehand; but it can also be directly extended to the infinite/unknown time horizon by replacing tree-based mechanism with the hybrid mechanism [10]. Although in this paper our focus is the contextual bandit problem, our dynamic global sensitivity analysis can also be generalized to other settings such as online convex optimization, where we can leverage the algorithm's convergence property to add decreasing noise to the model parameter for better utility.

Algorithm 2 Private LinUCB with Dynamic Global Sensitivity

```

1: Inputs:  $\epsilon, \delta, \lambda, T, L, \lambda_0$ 
2: Initialize:  $\mathbf{A}_1 \leftarrow \lambda \mathbf{I}, \mathbf{b}_1 \leftarrow \mathbf{0}$ 
3: for  $t = 1$  to  $T$  do
4:   Observe context vectors  $C_t = \{\mathbf{x}_a | a \in \mathcal{A}_t\}$ 
5:   Take action  $a_t = \arg \max_{a \in \mathcal{A}_t} \mathbf{x}_a^\top \hat{\boldsymbol{\theta}}_t^p + \alpha_t \|\mathbf{x}_a\|_{\mathbf{A}_t^{-1}}$ 
6:   Observe reward  $r_{a_t}$ 
7:    $\mathbf{A}_{t+1} \leftarrow \mathbf{A}_t + \mathbf{x}_{a_t} \mathbf{x}_{a_t}^\top$ 
8:    $\mathbf{b}_{t+1} \leftarrow \mathbf{b}_t + \mathbf{x}_{a_t} r_{a_t}$ 
9:   Sample noise  $\eta_t \sim \text{TreeMechanism}(t, \epsilon, \delta)$ 
10:   $\hat{\boldsymbol{\theta}}_{t+1}^p \leftarrow \mathbf{A}_{t+1}^{-1} \mathbf{b}_{t+1} + \eta_t$ 
11: end for

```

3.3 Differentially Private LinUCB

We now provide a differentially private linear contextual algorithm built on top of the tree-based mechanism with dynamic global sensitivity. The details of this algorithm is described in Algorithm 2. At round t , the algorithm receives a set of arms, with each arm associated with a context vector \mathbf{x}_a . Different from previous solutions [27, 28] that add noise to $\mathbf{b}_t = \sum_t \mathbf{x}_{a_t} r_t$, our algorithm directly adds noise to the bandit parameter for privacy (i.e., line 10), and uses the private model parameter $\hat{\boldsymbol{\theta}}_t^p$ for arm selection (i.e., line 5). The selected sequence of arms are proved to be (ϵ, δ) -differentially private to the reward sequence.

CLAIM 1. *The sequence of selected arms $\{a_t : t \in [1..T]\}$ by Algorithm 2 is (ϵ, δ) -differentially private.*

PROOF. In line 9-10 we use Algorithm 1 to add noise and keep $\hat{\boldsymbol{\theta}}_{t+1}^p$ (ϵ, δ) -differentially private. Because differential privacy is post-processing invariant [16], the sequence of selected arms $\{a_t : t \in [1..T]\}$ produced by $\hat{\boldsymbol{\theta}}_t^p$ is thus also (ϵ, δ) -differentially private. \square

We now specify the confidence bound of reward estimation with $\hat{\boldsymbol{\theta}}_t^p$, which we use for arm selection in line 5 of Algorithm 2.

LEMMA 2 (CONFIDENCE BOUND). *Following Assumption 1 and assuming $\|\boldsymbol{\theta}^*\|_2 \leq S$, with probability at least $1 - 2\zeta$, confidence bound of the estimated reward $\mathbf{x}^\top \hat{\boldsymbol{\theta}}_t^p$ is*

$$CB_t(\mathbf{x}) = \|\mathbf{x}^\top \hat{\boldsymbol{\theta}}_t^p - \mathbf{x}^\top \boldsymbol{\theta}^*\| \leq \alpha_t \|\mathbf{x}\|_{\mathbf{A}_t^{-1}} \quad (3)$$

where we define α_t as the upper bound of $\|\hat{\boldsymbol{\theta}}_t^p - \boldsymbol{\theta}^*\|_{\mathbf{A}_t}$. We have

$$\|\hat{\boldsymbol{\theta}}_t^p - \boldsymbol{\theta}^*\|_{\mathbf{A}_t} \leq \frac{L}{\epsilon} \log T \sqrt{\log t \log(\frac{1}{\zeta})} / \sqrt{t} + \sqrt{d \log \frac{1 + tL^2\lambda}{\zeta}} + \sqrt{\lambda} S \quad (4)$$

PROOF. Eq. (3) is obtained by the Cauchy-Schwarz inequality. To bound α_t , we apply the triangle inequality: $\|\hat{\boldsymbol{\theta}}_t^p - \boldsymbol{\theta}^*\|_{\mathbf{A}_t} \leq \|\hat{\boldsymbol{\theta}}_t^p - \hat{\boldsymbol{\theta}}_t\|_{\mathbf{A}_t} + \|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_{\mathbf{A}_t}$, where $\hat{\boldsymbol{\theta}}_t$ is the non-private estimate of $\boldsymbol{\theta}^*$. The first term is bounded according to Theorem 2 under \mathbf{A}_t norm with probability at least $1 - \zeta$. The second term is the confidence ellipsoid of non-private LinUCB and it can be bounded by Theorem 2 of [1] with probability at least $1 - \zeta$. By taking a union bound, the inequality holds with probability at least $1 - 2\zeta$ \square

Lemma 2 gives a tight construction of uncertainty regarding reward estimation $\mathbf{x}^\top \hat{\boldsymbol{\theta}}_t^p$ and model parameter estimation $\|\hat{\boldsymbol{\theta}}_t^p - \boldsymbol{\theta}^*\|_{\mathbf{A}_t}$. Note that comparing to the non-private LinUCB, the confidence bound of model estimation α_t in our

algorithm is relaxed by an additional term $\frac{L}{\epsilon} \log T \sqrt{\log t} \log(\frac{1}{\zeta})/\sqrt{t}$, which captures the upper bound of uncertainty caused by the privacy-preserving noise η_t introduced by Algorithm 1.

Now we provide a gap-independent regret bound of Algorithm 2 in following theorem.

THEOREM 3 (REGRET BOUND). *Following Assumption 1, the pseudo-regret of Algorithm 2 up to time T can be bounded by,*

$$\begin{aligned} \text{Regret}(T) \leq & 2 \left(\sqrt{d \log \frac{1 + TL^2\lambda}{\zeta}} + \sqrt{\lambda} S \right) \sqrt{dT \log \left(\lambda + \frac{LT}{d} \right)} \\ & + \left(2 \frac{3^{1.5} L}{e^{1.5} \lambda_0 \epsilon} \log \left(\frac{1}{\zeta} \right) \sqrt{dT \log \left(\lambda + \frac{LT}{d} \right)} + 32 \log(1/\delta)/\lambda_0 \right) \end{aligned} \quad (5)$$

with probability at least $1 - 2\zeta$.

Proof Sketch. We rewrite the cumulative regret of LinUCB defined in Eq. (1) by $\sum_{t=1}^T 2\text{CB}_t(\mathbf{x}_{a_t})$, using the definition of confidence bound and UCB arm selection strategy in line 5 of Algorithm 2. We apply the Cauchy-Schwarz inequality to bound $\sum_t \alpha_t$ and $\sum_t \|\mathbf{x}_{a_t}\|_{\Lambda_t^{-1}}$. According to Eq. (4) in Lemma 2, we can separate the bound of α_t into two terms: one is the confidence bound of the original LinUCB and the other is the bound of injected noise. We use Lemma 11 of [1] to bound $\sum_t \|\mathbf{x}_{a_t}\|_{\Lambda_t^{-1}}$.

The first term of regret in Theorem 3 is the same as the regret of non-private LinUCB, which is caused by its parameter estimation uncertainty and exploration in arm selection. The second term is the added regret introduced for privacy based on dynamic global sensitivity. The private LinUCB algorithm in [28] adds noise to \mathbf{b}_t and incurs additional regret in the order of $\tilde{O}(\log^{2.5} T \sqrt{T}/\epsilon)$ to achieve ϵ -differential privacy, while our algorithm introduces additional regret in the order of $\tilde{O}(\log T \sqrt{T}/\epsilon)$ to achieve (ϵ, δ) -differential privacy ignoring logarithmic terms. The dependency of additional regret on δ is $O(\log(1/\delta))$. We note that our method leverages the convergence of parameter estimation to inject less noise; and since it is a high probability analysis, the relaxed (ϵ, δ) -differential privacy notion is required. Later, we also validate our theoretical analysis of the regret reduction in our empirical evaluations. We currently cannot prove if our regret is optimal with respect to T , since the lower bound of the differentially private linear bandit problem is still an open problem. We note the lower bound discussed in [33] is not applicable, because their privacy definition includes context and it is different from ours that focuses on the privacy of reward. We consider deriving the lower bound of this problem as an important future direction.

In general, the regret bound of contextual bandits could be categorized into gap-independent bound and gap-dependent bound. Our provided analysis is for a gap-independent bound and it is dominated by the term $O(\sqrt{T})$ after ignoring the logarithmic terms. The gap-dependent regret bound of a bandit algorithm is usually in the order of $O(\log T/\mu_{1,k})$, where $\mu_{1,k}$ is the minimal reward difference between the best arm and any sub-optimal arm k . We leave the analysis of gap-dependent bound of our proposed algorithm as future work.

Note that Algorithm 2 takes L and λ_0 as input. From a theoretical perspective, the knowledge of the maximum L2-norm of context features L is a commonly made assumption in linear bandits, which is required to derive important model parameters such as the size of confidence ellipsoid α_t in LinUCB [1]. From a practical view, one can normalize the context feature vectors or observe them ahead of the time to empirically estimate L and λ_0 .

When the context vectors are not generated from the required random process with minimal eigenvalue λ_0 , we can add an additional Gaussian perturbation on the context to achieve this condition similar to the idea in Kannan et al. [21]. This leads to the following auxiliary result.

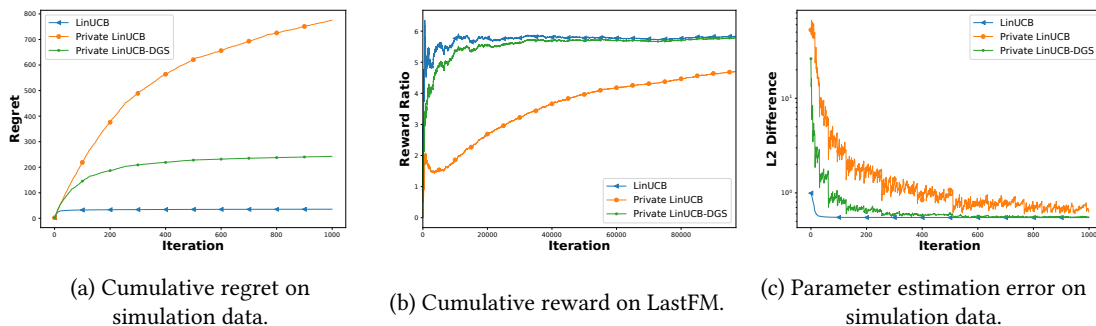


Fig. 1. Regret and parameter estimation quality on simulation and real-world dataset.

COROLLARY 1. *Following the assumptions in Lemma 1 and 2 about the context vectors, by perturbing every context vector \mathbf{x}_a with $z \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$, Algorithm 2 is (ϵ, δ) -differentially private. The pseudo-regret of Algorithm 2 up to time T can be bounded by*

$$\begin{aligned} \text{Regret}(T) \leq & 2 \frac{\sqrt{d \log \frac{1+TL^2\lambda}{\zeta}} + \sqrt{\lambda}S}{\sigma^2} \sqrt{dT \log \left(\lambda + \frac{LT}{d} \right)} \\ & + \left(2 \frac{3^{1.5}L}{e^{1.5}\sigma^2\epsilon} \log \left(\frac{1}{\zeta} \right) \sqrt{dT \log \left(\lambda + \frac{LT}{d} \right)} + 32 \log(1/\delta) / \sigma^2 \right) \end{aligned} \quad (6)$$

with probability at least $1 - 2\zeta$.

Here we notice that if we do not have the assumption on the generation of context vectors, while instead perturbing the context vectors to guarantee Lemma 1, the regret would be larger than the one in Theorem 3, especially when variance σ^2 is small. However note that the order of the regret in T is still the same. The proof of this corollary can be found in the appendix.

4 EXPERIMENT

We performed empirical evaluations of our proposed private LinUCB with dynamic global sensitivity (denoted as Private LinUCB-DGS) against two baseline algorithms:

- **LinUCB** [25]: it selects an arm based on its upper confidence bound of the estimated reward with given context vectors. As a non-private algorithm, LinUCB does not inject any noise to its parameter estimation.
- **Private LinUCB** [28]: it adds noise to \mathbf{b}_t using the tree-based mechanism with a constant sensitivity parameter, and incurs additional regret at the order of $\tilde{O}(\log^{2.5} T \sqrt{T} / \epsilon)$.
- **Experiment setup.** In our simulation-based study, we generate a size- K ($K = 1000$) arm pool \mathcal{A} , in which each arm a is associated with a d -dimensional ($d = 10$) feature vector $\mathbf{x}_a \in \mathbb{R}^d$ with $\|\mathbf{x}_a\|_2 \leq 1$. Similarly, we create a set of ground-truth bandit parameters $\theta^* \in \mathbb{R}^d$ with $\|\theta^*\|_2 \leq 1$, which are not disclosed to the learners. Each dimension of both \mathbf{x}_a and θ^* is drawn from a uniform distribution $U(0, 1)$. At each round t , a randomly-sampled decision set with 10 arms from \mathcal{A} are disclosed to the learners for selection. The ground-truth reward r_a is corrupted by Gaussian noise $\gamma \sim N(0, \sigma^2)$ before giving back to the learners and the standard deviation σ is set to 0.5.

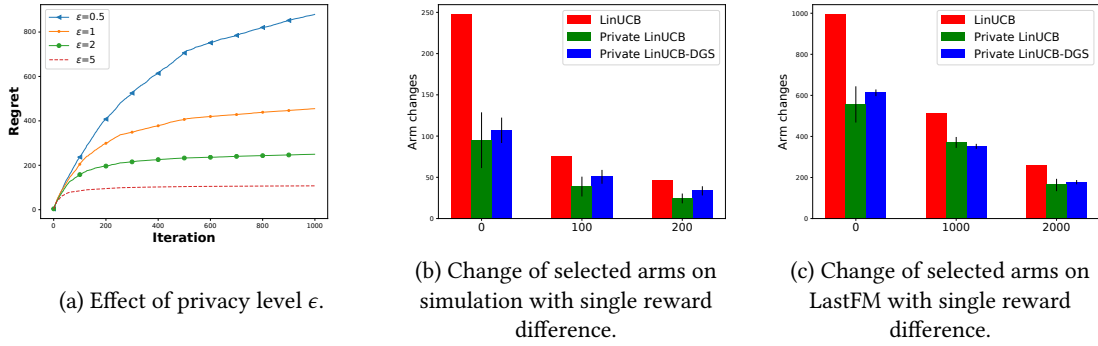


Fig. 2. Detailed empirical analysis of privacy-preserving mechanism for LinUCB.

We also experimented on a real-world dataset extracted from the music streaming service website Last.fm⁴. The LastFM datasets is published by the HetRec 2011 workshop⁵. This dataset contains 1,892 users and 17,632 items (artists). We used the information of “listened artists” of each user to create payoffs of recommendation candidates: if a user listened to an artist at least once, the payoff is 1, otherwise 0. Following the settings in [9, 38, 40], we extracted the context features of artists and pre-processed the dataset in order to fit them into a contextual bandit setting. Specifically, we create a TF-IDF feature vector using associated tags (6,036 tags in total), which uniquely represents the context of that artist. PCA is used to reduce the dimensionality of the feature vectors to $d = 25$. For a particular user i , we generate the candidate arm pool with size $|\mathcal{A}_t| = 25$ by first sampling one item from those non-zero reward candidate in user i 's past history, and then randomly fulfill the other 24 from those zero-reward candidates from this user.

Notice that both synthetic data and real-world data met Assumption 1. In synthetic data, the context features are sampled from a uniform distribution. For the LastFM dataset, the context vectors are generated by retaining the first 25 principal components of the original 6,036-dimension TD-IDF representation of items, which means the minimum eigenvalue of $\mathbb{E}[\mathbf{xx}^T]$ is lower bounded.

• **Regret comparison.** We first show the regret on simulation and real-world dataset in Figure 1 (a) and (b). As in the real-world dataset we do not have an oracle policy, we instead report each learning algorithm’s cumulative reward for comparison. We set the privacy level $\epsilon = 2$ and $\delta = 0.1$ in the following experiments as default. On the LastFM dataset, we report the reward ratio normalized by the reward collected from a random policy; and the resulting performance curve is thus the higher the better. We observe that the non-private LinUCB performs better than its two private variants, since no noise is injected to LinUCB’s model estimation. Our private LinUCB-DGS performs much better than private LinUCB on both datasets, as less noise is added. These results support our theoretical analysis that with dynamic global sensitivity, our algorithm adds additional regret in the order of $\tilde{O}(\log T\sqrt{T}/\epsilon)$. Note that because the privacy notion are different in private LinUCB and our private LinUCB-DGS, here we focus on comparing the utility of these algorithms with the same ϵ .

• **Parameter estimation quality.** Figure 1 (c) shows the parameter estimation quality of the three bandit algorithms, i.e., the L2 difference between the estimated bandit parameter $\hat{\theta}_t^P$ and the ground-truth parameter θ^* . Compared with private LinUCB, our model parameter $\hat{\theta}^P$ converges faster, which explains its improved performance in regret. We also notice that the convergence of private LinUCB oscillates more seriously. This is because private LinUCB injects noise

⁴<http://www.last.fm>⁵<http://grouplens.org/datasets/hetrec-2011>

with a larger variance (as we always introduce zero-mean Laplace noise) to model estimation. As a result, its quality of model parameters estimation varies significantly during online update, which directly leads to its worse regret.

- **Effect of privacy level ϵ .** In Figure 2 (a), we show the regret of our algorithm with different privacy parameter ϵ . We vary ϵ from 0.5 to 5. From the results, we notice a clear trade-off between the required privacy level ϵ and the resulting regret. Stronger privacy requirement (i.e., a smaller ϵ) requires the privacy mechanism to inject more noise, which results in larger regret. This result also supports our theoretical analysis that in our solution is in the order of $O(\frac{1}{\epsilon})$.

- **Sensitivity in arm selection.** We now evaluate the algorithm’s sensitivity to reward changes, which is exactly what differential privacy intends to protect in a bandit algorithm. Specifically, we fix the sequence of context and arms and only change reward feedback on one particular arm. Note that the feedback noise in simulation is also fixed for all arms ahead of time in this experiment. We compare how many arms are selected differently after the reward change by a bandit algorithm. In our simulation based study, we run each algorithm 500 iterations and change the reward at iteration $\{0, 100, 200\}$ by increasing the original reward by 0.5. For experiment on LastFM, we run 5,000 iterations, and flip reward on the selected item at iteration $\{0, 1000, 2000\}$. We run private LinUCB-DGS and private LinUCB-DGS for 5 times and report the mean and standard deviation. We run LinUCB just one time as it is a deterministic algorithm.

From Figure 2 (b) and (c) we can observe that the number of changed arms in our solution is clearly smaller than its non-private counterpart, which suggests that private LinUCB-DGS is less sensitive to reward change comparing to original LinUCB. Private LinUCB has a similar number of changed arms than ours in average but with a larger variance. This suggests less stable utility provided by the private LinUCB solution. We also notice that when the reward change happens later, the change in all algorithms’ arm selection gets smaller, which supports our motivation of dynamic sensitivity analysis that the algorithm’s output is less sensitive to its input in the later stage.

5 CONCLUSIONS & FUTURE WORK

In this paper, we first studied the dynamic nature of sensitivity of linear bandit models and presented a dynamic global sensitivity analysis of such an algorithm under the tree based mechanism, which leads to reduced noise addition for differential privacy. We then developed and analyzed a differentially private linear bandit algorithm based on the concept of dynamic global sensitivity. Our private linear bandit algorithm injects additional regret caused by privacy-preserving mechanism in $\tilde{O}(\log T \sqrt{T}/\epsilon)$ while guarantees (ϵ, δ) -differential privacy. Experimental results on both synthetic and real-world datasets demonstrated the advantage of our algorithm and supported our theoretical analysis.

One important property of bandit algorithms is their intrinsic noise/randomness introduced by exploration, such as the stochastic sampling of arms in Thompson Sampling and EXP3. Less explicit noise should be needed for differential privacy, because of such intrinsic randomness of an algorithm’s output, e.g., free privacy. We plan to take advantage of it for better balancing privacy and utility in bandit algorithms. In addition, as we mentioned before our proposed dynamic global sensitivity analysis is not only limited to linear bandit algorithms. It is meaningful to explore its application in a broader context, such as logistic bandits and online convex optimization. We also notice that the lower bound of the differentially private linear bandit problem that protects privacy of reward is currently still an open problem, and it is important to investigate this lower bound to show the optimality of a private linear bandit algorithm.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful comments. This work was supported by National Science Foundation Award IIS-2128019, IIS-2007492 and IIS-1553568, Google Faculty Research Award and Bloomberg Data Science Ph.D. Fellowship.

APPENDIX

Missing Proofs

In this section, we provide the detailed proofs of the lemma and theorems discussed in our paper.

PROOF OF LEMMA 1. For any neighbouring reward sequences $\{S, S'\}$ that only differ at the data point j , i.e., $r_i = r'_i$ when $i \neq j$, denote $\mathbf{A}_t = \lambda \mathbf{I} + \sum_{i=1}^t \mathbf{x}_{a_i} \mathbf{x}_{a_i}^\top$, we have,

$$\begin{aligned} & \|\hat{\boldsymbol{\theta}}_{t+1} - \hat{\boldsymbol{\theta}}'_{t+1}\|_2 \\ &= \left\| \left(\lambda \mathbf{I} + \sum_{i=1}^t \mathbf{x}_{a_i} \mathbf{x}_{a_i}^\top \right)^{-1} \sum_{i=1}^t \mathbf{x}_{a_i} r_i - \left(\lambda \mathbf{I} + \sum_{i=1}^t \mathbf{x}_{a_i} \mathbf{x}_{a_i}^\top \right)^{-1} \sum_{i=1}^t \mathbf{x}_{a_i} r'_i \right\|_2 \\ &= \left\| \mathbf{A}_t^{-1} \mathbf{x}_{a_j} (r_j - r'_j) \right\|_2 \\ &\leq 2L \lambda_{\max}(\mathbf{A}_t^{-1}) \\ &\leq \frac{2L}{\lambda_{\min}(\mathbf{A}_t)} \end{aligned}$$

$\lambda_{\min}(\cdot)$ denotes the minimal eigenvalue of the input matrix. The third step is because of the assumption $\|\mathbf{x}\|_2 \leq L$ and $\|r\| \leq 1$. To analyze the eigenvalue of \mathbf{A}_t , we adopt the assumption from Theorem 1 of Gentile et al. [18] that context vectors $\{\mathbf{x}_{1,t}, \dots, \mathbf{x}_{K,t}\}$ are i.i.d. conditioned on the algorithm's past actions and observed context. Based on Theorem 1 of [18], $\lambda' = \lambda_0 t / 4 - 8 \log((t+3)/\delta) - 2\sqrt{t \log((t+3)/\delta)}$ and $\lambda' = \lambda_0 t / 16$ when $t > 32 \log(1/\delta) / \lambda_0$. Thus we have $\lambda_{\min}(\mathbf{A}_t) \geq \lambda_0 t / 16$ when $t > 32 \log(1/\delta) / \lambda_0$. Substitute it back to the inequality and we finish the proof of Lemma 1. \square

PROOF OF THEOREM 1. We can rewrite the estimated model parameter $\hat{\boldsymbol{\theta}}_t$ as sum statistics using the Sherman–Morrison formula by

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{t+1} &= (\mathbf{A}_t + \mathbf{x}_{a_t} \mathbf{x}_{a_t}^\top)^{-1} (\mathbf{b}_t + \mathbf{x}_{a_t} r_{a_t}) \\ &= \mathbf{A}_t^{-1} \mathbf{b}_t + \Delta_{\mathbf{A}_t} \mathbf{b}_t + \mathbf{A}_t^{-1} \mathbf{x}_{a_t} r_{a_t} + \Delta_{\mathbf{A}_t} \mathbf{x}_{a_t} r_{a_t} \\ &= \hat{\boldsymbol{\theta}}_t + \Delta_{\mathbf{A}_t} \mathbf{b}_t + \mathbf{A}_t^{-1} \mathbf{x}_{a_t} r_{a_t} + \Delta_{\mathbf{A}_t} \mathbf{x}_{a_t} r_{a_t} \\ &:= \hat{\boldsymbol{\theta}}_t + \Delta(\hat{\boldsymbol{\theta}}_t) \end{aligned}$$

where $\Delta_{\mathbf{A}_t} = \frac{\mathbf{A}_t^{-1} \mathbf{x}_{a_t} \mathbf{x}_{a_t}^\top \mathbf{A}_t^{-1}}{1 + \mathbf{x}_{a_t}^\top \mathbf{A}_t^{-1} \mathbf{x}_{a_t}}$. Note that we do not actually need to store nor calculate the partial sum in the tree; but we only use the tree mechanism to sample noise for each partial sum, which makes our implementation efficient. We notice that not saving partial sums on the tree makes our algorithm fail to satisfy the definition of *pan privacy* [15]. However if needed, one can always explicitly maintain a tree for $\Delta(\hat{\boldsymbol{\theta}}_t)$ to achieve pan privacy, with a price of added computational cost.

For partial sum $\sum_{i=i}^j \Delta(\hat{\boldsymbol{\theta}}_i)$, i.e., $\hat{\boldsymbol{\theta}}_j - \hat{\boldsymbol{\theta}}_i$, its sensitivity is bounded by $\hat{\boldsymbol{\theta}}_j - \hat{\boldsymbol{\theta}}'_j$ and further bounded by Δ_j as discussed in Lemma 1. According to the definition of differential privacy, the partial sum is (ϵ', δ') -differentially private if we add $\text{Lap}(\frac{\Delta_j}{\epsilon'})$ noise with our sensitivity bound that holds with probability at least $1 - \delta'$.

We note that the incremental part $\Delta(\hat{\boldsymbol{\theta}}_t)$ depends on not only incoming $(\mathbf{x}_{a_t}, r_{a_t})$, but also the historical statistics summarized in \mathbf{A}_t . A similar problem also occurs in differentially private online convex optimization problem, where the current gradient depends on the choice of past gradients. We refer to Theorem 3 of [34], which uses the advanced

composition theorem, to prove that the composition of $(\epsilon' = \epsilon/(\log T), \delta' = \delta/(\log T),)$ -differentially private partial sums achieves (ϵ, δ) -differential privacy. \square

We clarify the analogy of problem structure between privacy of Online Convex Optimization (OCO) discussed in [34] and privacy of contextual bandits as follows: our $\Delta\theta$ could be viewed similarly as gradient in OCO, which is protected by private sum statistics releasing. In OCO, gradient depends on current loss and historical evaluated parameters $\{w_t\}_{t=1}^T$, while in contextual bandits $\Delta\theta$ depends on current reward and historical pulled arms $\{x_t\}_{t=1}^T$. An arm pulled in contextual bandits is analogous to parameter evaluated in OCO, and thus similar arguments and analysis apply.

PROOF OF THEOREM 2. To bound the total amount of noise η_t added by our tree mechanism, we first state the property of sum of independent Lapalace distributions (also stated in Lemma 2.8 of [10]):

LEMMA 3. *With probability $1 - \zeta$, sum of independent Lapalace distributions $\text{Lap}(a_i)$ is bounded by $O(\sqrt{\sum_i a_i^2} \log \frac{1}{\zeta})$*

Let b be the $\lceil \log_2 T \rceil + 1$ -bit binary representation of t . The noise introduced by tree-based mechanism with dynamic global sensitivity can be bounded by

$$\begin{aligned} \eta_t &= \sum_{b_i=1} \text{Lap}\left(\frac{\Delta_{t-2^i+1}}{\epsilon'}\right) \\ &= \sum_{b_i=1} \text{Lap}\left(\frac{\Delta_{t-2^i+1} \log T}{\epsilon}\right) \\ &\leq \log \frac{1}{\zeta} \sqrt{\frac{\log^2 T}{\epsilon^2} \left(\Delta_{\frac{1}{2}t'+1}^2 + \Delta_{\frac{3}{4}t'+1}^2 + \Delta_{\frac{7}{8}t'+1}^2 + \dots + \Delta_{t'}^2\right)} \\ &\leq \frac{32L}{\lambda_0} \log \frac{1}{\zeta} \frac{\log T}{t\epsilon} \sqrt{\left(\left(\frac{2}{1}\right)^2 + \left(\frac{4}{3}\right)^2 + \dots\right)} \\ &= \frac{32L}{\lambda_0} \log \frac{1}{\zeta} \frac{\log T}{t\epsilon} \sqrt{\left(\left(1 + \frac{1}{1}\right)^2 + \left(1 + \frac{1}{3}\right)^2 + \dots\right)} \\ &\leq \frac{32L}{\lambda_0} \log \frac{1}{\zeta} \frac{\log T}{t\epsilon} \sqrt{2 \log t + 2} \end{aligned}$$

where $t' > t$ is the smallest number that is the power of 2 after t . The last inequality holds because there are at most $\log t$ terms in the summation, and $(\frac{1}{1})^2 + (\frac{1}{3})^2 + \dots$ is upper bounded by 2. As a result, we conclude that the added noise η_t is in the order of $O(\frac{L}{\epsilon} \log T \sqrt{\log t} \log \frac{1}{\zeta} / t)$ with probability $1 - \zeta$. \square

PROOF OF THEOREM 3. We first bound the one-step regret at time t as,

$$\begin{aligned} \text{regret}(t) &= \mathbf{x}_{a_t^*}^\top \boldsymbol{\theta}^* - \mathbf{x}_{a_t}^\top \boldsymbol{\theta}^* \quad // \text{Definition of regret} \\ &\leq \mathbf{x}_{a_t^*}^\top \hat{\boldsymbol{\theta}}_t^p + \text{CB}_t(\mathbf{x}_{a_t^*}) - \mathbf{x}_{a_t}^\top \boldsymbol{\theta}^* \quad // \text{Definition of confidence bound} \\ &\leq \mathbf{x}_{a_t^*}^\top \hat{\boldsymbol{\theta}}_t^p + \text{CB}_t(\mathbf{x}_{a_t}) - \mathbf{x}_{a_t}^\top \boldsymbol{\theta}^* \quad // \text{Arm selection strategy} \\ &\leq 2\text{CB}_t(\mathbf{x}_{a_t}) \end{aligned}$$

The pseudo-regret up to time T is thus bounded by,

$$\begin{aligned}
\text{Regret}(T) &= \sum_{t=1}^T \text{regret}(t) \\
&\leq \sum_{t=1}^T 2\text{CB}_t(\mathbf{x}_{a_t}) \\
&= 2 \sum_{t=1}^T \alpha_t \|\mathbf{x}_{a_t}\|_{\mathbf{A}_t^{-1}} \\
&\leq 2 \sum_{t=1}^T \left(\|\hat{\theta}_t^p - \hat{\theta}_t\|_{\mathbf{A}_t} + \|\hat{\theta}_t - \theta^*\|_{\mathbf{A}_t} \right) \|\mathbf{x}_{a_t}\|_{\mathbf{A}_t^{-1}} \\
&\leq 2 \sqrt{\sum_{t=1}^T \|\hat{\theta}_t - \theta^*\|_{\mathbf{A}_t}^2} \sqrt{\sum_{t=1}^T \|\mathbf{x}_{a_t}\|_{\mathbf{A}_t^{-1}}^2} + 2 \sqrt{\sum_{t=1}^T \|\eta_t\|_{\mathbf{A}_t}^2} \sqrt{\sum_{t=1}^T \|\mathbf{x}_{a_t}\|_{\mathbf{A}_t^{-1}}^2} \\
&\leq 2 \sqrt{T \|\hat{\theta}_T - \theta^*\|_{\mathbf{A}_T}^2} \sqrt{\sum_{t=1}^T \|\mathbf{x}_{a_t}\|_{\mathbf{A}_t^{-1}}^2} + 2 \sqrt{T \max_{t'} \|\eta_{t'}\|_{\mathbf{A}_{t'}}^2} \sqrt{\sum_{t=1}^T \|\mathbf{x}_{a_t}\|_{\mathbf{A}_t^{-1}}^2}
\end{aligned}$$

In the fourth step, we separate the upper bound of α_t into two terms, among which one is the confidence bound of the original LinUCB and the other is the bound of injected noise. The fifth step follows the Cauchy–Schwarz inequality. Using Lemma 11 of [1], we have $\sum_t \|\mathbf{x}_{a_t}\|_{\mathbf{A}_t^{-1}}^2 \leq \sqrt{d \log(\lambda + \frac{LT}{d})}$. The bound of $\|\hat{\theta}_T - \theta^*\|_{\mathbf{A}_T}^2 \leq \sqrt{d \log \frac{1+TL^2\lambda}{\zeta}} + \sqrt{\lambda}S$ follows self-normalized bound for martingales of Lemma 9 in [1]. We bound $\max_{t'} \|\eta_{t'}\|_{\mathbf{A}_{t'}}^2$ when $\lambda \geq \lambda_0 t/16$, i.e., $t > 32 \log(1/\delta)/\lambda_0$, by taking the derivative of $\frac{\log^{1.5} t}{\sqrt{t}}$ and find its maximum value at $t = e^3$, since it is a concave function. Substitute it back and we get its bound of $\frac{3^{1.5}}{e^{1.5}}$. The regret for $t \leq 32 \log(1/\delta)/\lambda_0$ is at most $32L \log(1/\delta)/\lambda_0$. Combining all these terms together, we prove the regret bound of our developed private LinUCB based on global dynamic sensitivity analysis. \square

PROOF OF COROLLARY 1. When every context vector \mathbf{x} is perturbed by Gaussian noise $z \sim \mathcal{N}(0, \sigma^2)$, $\mathbb{E}[(\mathbf{x}+z)(\mathbf{x}+z)^\top]$ has minimal eigenvalue λ_0 where $\lambda_0 \geq \Omega(\sigma^2)$. Thus Lemma 1 and 2 hold, and the privacy guarantee directly follows Theorem 1. Following Lemma 3.3 in [21], we have $\|\hat{\theta}_T - \theta^*\|_{\mathbf{A}_T}^2 \leq \frac{\sqrt{d \log \frac{1+TL^2\lambda}{\zeta}} + \sqrt{\lambda}S}{\sigma^2}$. Plug the result into the analysis of Theorem 3 and we have the regret bound. \square

REFERENCES

- [1] Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. 2011. Improved Algorithms for Linear Stochastic Bandits. In *NIPS*.
- [2] Jacob Abernethy, Chansoo Lee, Audra McMillan, and Ambuj Tewari. 2017. Online learning via differential privacy. *arXiv preprint arXiv:1711.10019* (2017).
- [3] Naman Agarwal and Karan Singh. 2017. The price of differential privacy for online learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 32–40.
- [4] Peter Auer. 2002. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research* 3 (2002), 397–422.
- [5] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Mach. Learn.* 47, 2-3 (May 2002), 235–256. <https://doi.org/10.1023/A:1013689704352>
- [6] Debabrota Basu, Christos Dimitrakakis, and Aristide Tossou. 2019. Differential privacy for multi-armed bandits: What is it and what is its cost? *arXiv preprint arXiv:1905.12298* (2019).
- [7] Sébastien Bubeck, Nicolò Cesa-Bianchi, et al. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5, 1 (2012), 1–122.
- [8] Joseph A Calandrino, Ann Kilzer, Arvind Narayanan, Edward W Felten, and Vitaly Shmatikov. 2011. " You might also like:" Privacy risks of collaborative filtering. In *2011 IEEE Symposium on Security and Privacy*. IEEE, 231–246.
- [9] Nicolò Cesa-Bianchi, Claudio Gentile, and Giovanni Zappella. 2013. A Gang of Bandits. In *Pro. NIPS* (2013).
- [10] T-H Hubert Chan, Elaine Shi, and Dawn Song. 2011. Private and continual release of statistics. *ACM Transactions on Information and System Security (TISSEC)* 14, 3 (2011), 26.
- [11] Abhimanyu Dubey and AlexSandy' Pentland. 2020. Differentially-private federated linear bandits. *Advances in Neural Information Processing Systems* 33 (2020), 6003–6014.
- [12] Abhimanyu Dubey and Alex Pentland. 2020. Private and Byzantine-Proof Cooperative Decision-Making. In *AAMAS*.
- [13] Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D Mitsis, and Joelle Pineau. 2018. Contextual Bandits for Adapting Treatment in a Mouse Model of de Novo Carcinogenesis. In *Machine Learning for Healthcare Conference*. 67–82.
- [14] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*. Springer, 265–284.
- [15] Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N Rothblum. 2010. Differential privacy under continual observation. In *Proceedings of the forty-second ACM symposium on Theory of computing*. ACM, 715–724.
- [16] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.
- [17] Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. 2010. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE, 51–60.
- [18] Claudio Gentile, Shuai Li, and Giovanni Zappella. 2014. Online clustering of bandits. In *International Conference on Machine Learning*. 757–765.
- [19] Yuxuan Han, Zhipeng Liang, Yang Wang, and Jiheng Zhang. 2021. Generalized linear bandits with local differential privacy. *Advances in Neural Information Processing Systems* 34 (2021).
- [20] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2017. The composition theorem for differential privacy. *IEEE Transactions on Information Theory* 63, 6 (2017), 4037–4049.
- [21] Sampath Kannan, Jamie H Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. 2018. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Advances in Neural Information Processing Systems*. 2227–2236.
- [22] Michael Kearns, Malleh Pai, Aaron Roth, and Jonathan Ullman. 2014. Mechanism design in large games: Incentives and privacy. In *Proceedings of the 5th conference on Innovations in theoretical computer science*. ACM, 403–410.
- [23] Aleksandra Korolova. 2010. Privacy violations using microtargeted ads: A case study. In *2010 IEEE International Conference on Data Mining Workshops*. IEEE, 474–482.
- [24] Tor Lattimore and Csaba Szepesvári. Bandit algorithms. (????).
- [25] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of 19th WWW*. ACM, 661–670.
- [26] Wei Li, Xuerui Wang, Ruofei Zhang, Ying Cui, Jianchang Mao, and Rong Jin. 2010. Exploitation and exploration in a performance based contextual advertising system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 27–36.
- [27] Nikita Mishra and Abhradeep Thakurta. 2015. (Nearly) optimal differentially private stochastic multi-arm bandits. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 592–601.
- [28] Seth Neel and Aaron Roth. 2018. Mitigating bias in adaptive data gathering via differential privacy. *arXiv preprint arXiv:1806.02329* (2018).
- [29] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2007. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM, 75–84.
- [30] Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X Yu, Sashank J Reddi, and Sanjiv Kumar. 2019. AdaClip: Adaptive clipping for private SGD. *arXiv preprint arXiv:1908.07643* (2019).

- [31] Wenbo Ren, Xingyu Zhou, Jia Liu, and Ness B Shroff. 2020. Multi-armed bandits with local differential privacy. *arXiv preprint arXiv:2007.03121* (2020).
- [32] Herbert Robbins. 1952. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* 58, 5 (1952), 527–535.
- [33] Roshan Shariff and Or Sheffet. 2018. Differentially Private Contextual Linear Bandits. In *Advances in Neural Information Processing Systems*. 4296–4306.
- [34] Abhradeep Guha Thakurta and Adam Smith. 2013. (Nearly) optimal algorithms for private online learning in full-information and bandit settings. In *Advances in Neural Information Processing Systems*. 2733–2741.
- [35] William R Thompson. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25, 3/4 (1933), 285–294.
- [36] Aristide CY Tossou and Christos Dimitrakakis. 2016. Algorithms for differentially private multi-armed bandits. In *Thirtieth AAAI Conference on Artificial Intelligence*.
- [37] Aristide Charles Yedia Tossou and Christos Dimitrakakis. 2017. Achieving privacy in the adversarial multi-armed bandit. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- [38] Huazheng Wang, Qingyun Wu, and Hongning Wang. 2016. Learning hidden features for contextual bandits. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 1633–1642.
- [39] Huazheng Wang, Qian Zhao, Qingyun Wu, Shubham Chopra, Abhinav Khaitan, and Hongning Wang. 2020. Global and local differential privacy for collaborative bandits. In *Fourteenth ACM Conference on Recommender Systems*. 150–159.
- [40] Qingyun Wu, Huazheng Wang, Quanquan Gu, and Hongning Wang. 2016. Contextual Bandits in a Collaborative Environment. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 529–538.
- [41] Kai Zheng, Tianle Cai, Weiran Huang, Zhenguo Li, and Liwei Wang. 2020. Locally differentially private (contextual) bandits learning. *Advances in Neural Information Processing Systems* 33 (2020), 12300–12310.