Proceedings of Machine Learning Research vol 247:1–6, 2024 37th Annual Conference on Learning Theory

Open Problem: What is the Complexity of Joint Differential Privacy in Linear Contextual Bandits?

Achraf Azize ACHRAF.AZIZE@INRIA.FR
Debabrota Basu DEBABROTA.BASU@INRIA.FR
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France

Editors: Shipra Agrawal and Aaron Roth

Abstract

Contextual bandits serve as a theoretical framework to design recommender systems, which often rely on user-sensitive data, making privacy a critical concern. However, a significant gap remains between the known upper and lower bounds on the regret achievable in linear contextual bandits under Joint Differential Privacy (JDP), which is a popular privacy definition used in this setting. In particular, the best regret upper bound is known to be $O\left(d\sqrt{T}\log(T) + d^{3/4}\sqrt{T\log(1/\delta)}/\sqrt{\epsilon}\right)$, while the lower bound is $\Omega\left(\sqrt{dT\log(K)} + d/(\epsilon+\delta)\right)$. We discuss the recent progress on this problem, both from the algorithm design and lower bound techniques, and posit the open questions.

Keywords: Differential Privacy, Contextual Bandits, Regret Analysis

1. Introduction

We consider the setting of contextual bandits, where a policy (aka learner/agent) observes at each step t a context $c_t \in \mathcal{C}$, which may be random or not. Having observed the context, the policy chooses an action $a_t \in [K]$ and observes a reward r_t. For the linear contextual bandits, the reward r_t depends on both the arm a_t and the context c_t in terms of a linear structural equation:

$$r_t \triangleq \langle \theta^\star, \psi(a_t, c_t)\rangle + \eta_t. \tag{1}$$

Here, $\psi : [K] \times \mathcal{C} \to \mathbb{R}^d$ is the feature map, $\theta^\star \in \mathbb{R}^d$ is the unknown parameter, and η_t is the noise, which may be assumed to be conditionally 1-subgaussian. While decision-making with Equation (1), all that matters is the value of the feature vector. Thus, the bandit literature often studies a reduced model (Lattimore and Szepesvári, 2020), where in round t, the policy is served with the decision set $\mathcal{A}_t \subset \mathbb{R}^d$, from which it chooses an action $a_t \in \mathcal{A}_t$ and receives a reward $r_t \triangleq \langle \theta^\star, a_t\rangle + \eta_t$, where η_t is 1-subgaussian given $\mathcal{A}_1, a_1, R_1, \dots, \mathcal{A}_{t-1}, a_{t-1}, R_{t-1}, \mathcal{A}_t$, and A_t. Different choices of $\mathcal{A}_t$ lead to different settings. For example, if $\mathcal{A}_t \triangleq \{\psi(c_t, a) : a \in [K]\}$, then we have a contextual linear bandit, or if $\mathcal{A}_t \triangleq \{e_1, \dots, e_d\}$, where $(e_i)_i$ are the unit vectors of $\mathbb{R}^d$ then the resulting bandit problem reduces to a d-finite armed bandit. For the contextual bandit setting, the contexts can be either generated stochastically, i.e. sampled from some distribution (Gentile et al., 2014), or assumed to be generated arbitrarily, i.e. adversarial contexts (Abbasi-Yadkori et al., 2011). Impacts of further assumptions on the context generation, like the margin condition (Goldenshluger and Zeevi, 2013) or diversity conditions (Bastani et al., 2021), have also been studied.

Contextual bandits are increasingly used in a wide range of sequential decision-making tasks under uncertainty, such as recommender systems (Silva et al., 2022), strategic pricing (Bergemann and Välimäki, 1996), clinical trials (Thompson, 1933). These applications often involve individuals' sensitive data, such as personal preferences, financial situation, and health conditions. Thus, these applications naturally invoke data privacy concerns in contextual bandits. For example, let us consider a contextual bandit algorithm π recommending one of K medicines. On the t-th day, a new patient u_t arrives, and π recommends medicine $a_t \in [K]$. To recommend a medicine a_t, the policy considers the specific medical conditions (or context) of patient u_t, i.e. c_t. Then, the patient's reaction to the medicine is observed. If the medicine cures the patient, the observed reward $r_t = 1$, otherwise $r_t = 0$. Both the observed reward and the context can reveal sensitive information about the health condition of patient u_t. Thus, *the goal of a privacy-preserving bandit algorithm is to recommend a sequence of medicines (actions) that cures the maximum number of patients while protecting the privacy of these patients*. Since both rewards and contexts are considered private information, a variation of Differential Privacy (DP) (Dwork et al., 2014), i.e. Joint Differential Privacy under continuous observations is proposed for contextual linear bandits (Shariff and Sheffet, 2018) and reinforcement learning (Vietri et al., 2020).

Definition 1 (t-neighbouring context-reward sequences) *Let $S \triangleq \{(\mathcal{A}_1, r_1), \dots, (\mathcal{A}_T, r_T)\}$ and $S' \triangleq \{(\mathcal{A}'_1, r'_1), \dots, (\mathcal{A}'_T, r'_T)\}$ be two context-reward sequences. S and S' are said to be t-*neighbours *if for all $s \neq t$ it holds that $(\mathcal{A}_s, r_s) = (\mathcal{A}'_s, r'_s)$.*

Definition 2 (JDP, Shariff and Sheffet (2018)) *A randomised policy π for the contextual bandit problem is (ϵ, δ)-*Jointly Differentially Private *(JDP) if for any t and any pair of t-neighbouring context-reward sequences S and S', and any subset $E_{>t} \subset \mathcal{A}_{t+1} \times \mathcal{A}_{t+2} \times \cdots \times \mathcal{A}_T$ of sequence of actions ranging from step $t+1$ to the end of the sequence, it holds that*

$$\Pr\{\pi(S) \in E_{>t}\} \leq e^{\epsilon} \Pr\{\pi(S') \in E_{>t}\} + \delta. \tag{2}$$

where $\pi(S)$ represents the sequence of actions recommended by the policy π when interacting with S, and $\Pr$ accounts only for randomness due to the policy.

JDP requires that changing the context at step t does not affect the actions chosen *only in the future rounds* ($> t$), i.e. $(a_{t+1}, \dots, a_T)$. In contrast, the standard notion of DP would require that the change does not have any effect on the full sequence of actions $(a_1, \dots, a_T)$, including the one chosen at step t. Claim 13 of Shariff and Sheffet (2018) shows that the standard notion of DP for linear contextual bandits, where both the reward and contexts are private, and the *full* sequence of actions is published, always leads to linear regret. In addition, in the reduced model based on decision sets $\mathcal{A}_t$, the standard notion of DP is ill-defined, as it requires the full sequence of actions to remain unchanged under any change in context-reward. This is true because two t-neighbouring context-reward sequences might yield different sets $\mathcal{A}_t$ and $\mathcal{A}'_t$. Since the action a_t should be an element of the decision set at step t, i.e. $\mathcal{A}_t$ or $\mathcal{A}'_t$, then it is impossible to expect that a_t is unchanged between the two neighbouring cases.

The goal is to design an (ϵ, δ)-JDP policy that minimises the regret, which is defined as

$$R_T \triangleq \mathbb{E}[\textstyle\sum_{t=1}^T \max_{a \in \mathcal{A}_t} \langle \theta^\star, a - a_t \rangle].$$

Now, we ask the questions:

Q1. Is it possible to derive matching upper and lower bounds on regret, for linear contextual bandits, under *JDP and adversarial contexts*?

Q2. Is JDP achievable for *free* in this setting?

Q3. If not, what is the minimal set of assumptions on context generation to achieve JDP for *free*?

In the following, we first revisit the settings of finite-armed and linear bandits, where the complexity of privacy is well-studied. Then, we present the gap between regret upper and lower bounds in contextual linear bandits under JDP.

Table 1: The known complexity of ρ-zCDP for finite and linear bandits (Azize and Basu, 2024).

Bandit Setting	Regret Upper Bound	Regret Lower Bound
Finite-armed bandits	$\mathcal{O}\left(\sqrt{KT\log(T)}\right) + \mathcal{O}\left(\frac{K}{\sqrt{\rho}}\sqrt{\log(T)}\right)$	$\Omega\left(\max\left(\sqrt{KT}, \sqrt{\frac{K}{\rho}}\right)\right)$
Linear bandits	$\mathcal{O}\left(\sqrt{dT\log(KT)}\right) + \mathcal{O}\left(\frac{d}{\sqrt{\rho}}\log^{\frac{3}{2}}(KT)\right)$	$\Omega\left(\max\left(d\sqrt{T}, \frac{d}{\sqrt{\rho}}\right)\right)$

2. Warm-up: Finite-armed bandits and linear bandits under DP

For a K-armed bandits, the policy chooses at each step t an action $a_t \in \{1, \dots, K\}$, and observes a reward $r_t \sim P_{a_t}$, where $\nu = (P_a : a \in [K])$ is a bandit instance with K reward distributions $(P_a)_{a\in[K]}$ with unknown means $(\mu_a)_{a\in[K]}$. For linear bandits, a fixed set of actions $\mathcal{A} \subset \mathbb{R}^d$ is available at each round, such that $|\mathcal{A}| = K$. The rewards are generated by a linear structural equation. Specifically, at step t, the observed reward is $r_t \triangleq \langle \theta^\star, a_t \rangle + \eta_t$, where $\theta^\star \in \mathbb{R}^d$ is the unknown parameter, and η_t is a conditionally 1-subgaussian noise. The main difference between linear bandits, and the reduced model of linear contextual bandits, is that *the decision set $\mathcal{A}$ is fixed for linear bandits, while $\mathcal{A}_t$ is allowed to change from step to step*. In both settings, the policy *does not have access to any side information, i.e. contexts*. The only private quantity is the reward.

Here, a policy is perceived as a randomised algorithm, that yields a sequence of actions given a sequence of rewards. Thus, DP is extended to bandits w.r.t. the full sequence of actions (Thakurta and Smith, 2013), in contrast to only the future rounds in JDP. Also, Azize and Basu (2024) discusses different ways to extend DP to bandits, against interactive and non-interactive adversaries, and the effect of partial information on these definitions. The complexity of finite and linear bandits under DP is well understood for both regret (Azize and Basu, 2022, 2024), and best-arm identification settings (Azize et al., 2023). For *pure* ϵ-DP, Azize and Basu (2022) provides regret lower bounds for finite-armed and linear bandits, and also algorithm design techniques to match them. Azize and Basu (2024) complete the picture for a "relaxation" of pure DP, by providing regret upper and lower bounds for finite and linear bandits, under ρ-zCDP. We present these bounds in Table 1 that shows the price of ρ-zCDP (presented in blue) to be asymptotically negligible. Specifically, the additional cost due to ρ-zCDP is in poly $\log(T)$, while the non-private regret is in $\sqrt{T}$. The upper and lower bounds in Table 1 match up to logarithmic factors in T.

Azize and Basu (2022, 2024) share similar techniques to provide matching regret upper and lower bounds. First, the algorithmic design follows the same blueprint: the algorithms run in adaptive and non-overlapping phases. This helps avoid the use of sequential composition, and thus *the tree-based mechanism* (aka binary mechanism) (Dwork et al., 2010; Chan et al., 2011). By running in adaptive and non-overlapping phases, the algorithms utilise the "parallel composition" property of DP to add less noise, and thus, yield less regret. This has been first used by Sajed and Sheffet (2019) to provide a DP version of the Successive Elimination algorithm, and then used in multiple works to create DP versions of Thompson sampling (Hu and Hegde, 2022), UCB (Hu et al., 2021; Azize and Basu, 2022) and more (Hanna et al., 2022; Li et al., 2022). Then, coupling ideas have been adapted to the sequential setting to provide lower bounds. Specifically, the main technical result used is upper bounding the divergence between sequences of actions, under a change of bandit environment, which are generalisations of the coupling techniques of Karwa and Vadhan (2017). All these techniques could be seen as a "stochastic" generalisation of the group property of DP, adapted to the sequential setting.

3. The gap between upper and lower bounds in contextual bandits under JDP

Existing upper bound. To solve linear contextual bandits under JDP, Shariff and Sheffet (2018) propose a variant of LinUCB (Abbasi-Yadkori et al., 2011). The LinUCB algorithm applies the "optimism in the face of uncertainty principle" to the contextual linear bandit setting, i.e. act in each round as if the environment is as nice as probably possible. At round t, LinUCB first computes a regularised least-squares estimator $\hat{\theta}_t \triangleq V_t^{-1} u_t$, where $V_t \triangleq \lambda I_d + \sum_{s=1}^{t-1} a_s a_s^T$ is the Gram matrix, $u_t \triangleq \sum_{s=1}^{t-1} a_s r_s$, and $\lambda > 0$. Then, the algorithm builds an elliptic confidence ball $\mathcal{C}_t \triangleq \{\theta \in \mathbb{R}^d : \|\theta - \theta_t\|_{V_t}^2 \leq \beta_t\}$ around the estimation $\hat{\theta}_t$ and chooses the arm $a_t = \arg\max_{a \in \mathcal{A}_t} \max_{\theta \in \mathcal{C}_t} \langle \theta, a \rangle = \arg\max_{a \in \mathcal{A}_t} \left\langle \hat{\theta}_t, a \right\rangle + \sqrt{\beta_t} \|a\|_{V_t^{-1}}$ with the highest optimistic reward estimation.

To design a JDP variant of LinUCB, Shariff and Sheffet (2018) estimate the quantities V_t and u_t privately using the *tree-based mechanism*, since these two quantities are written as sums. Then, the algorithm accounts for the noise addition by adapting the confidence intervals using the concentration properties of Gaussian and Wishart noise. We refer to the private estimations of V_t and u_t as $\tilde{V}_t$ and $\tilde{u}_t$. Thus, $\tilde{\theta}_t = \tilde{V}_t^{-1} \tilde{u}_t$ and $a_t = \arg\max_{a \in \mathcal{A}_t} \left\langle \tilde{\theta}_t, a \right\rangle + \sqrt{\tilde{\beta}_t} \|a\|_{\tilde{V}_t^{-1}}$. Shariff and Sheffet (2018) analyse the corresponding algorithm, for adversarial contexts, and show that it yields a regret upper bound of $O\left(d\sqrt{T}\log(T) + d^{3/4}\sqrt{T\log(1/\delta)}/\sqrt{\epsilon}\right)$, where the price of JDP is non-negligible even asymptotically in T. Given the advancements in other settings, we wonder: *is it possible to propose a JDP variant of LinUCB, such that the price of JDP is negligible for adversarial contexts?*

Existing lower bound. He et al. (2022) propose a reduction technique to go from *regret* lower bounds to *estimation* lower bounds in contextual bandits. Then, they instantiate this technique for contextual bandits under JDP. First, He et al. (2022) uses the coupling technique of Karwa and Vadhan (2017) to get a lower bound on the estimation error. Then, in Theorem 4.3 of He et al. (2022), they plug this estimation lower bound in their generic method, to get a regret lower bound of $\Omega\left(\sqrt{dT\log(K)} + d/(\epsilon + \delta)\right)$. It is also worth noting that the lower bounds are established for stochastic contexts, which are still valid lower bounds for adversarial contexts. The mismatch is significant with the upper bound of Shariff and Sheffet (2018), both in the horizon T and the privacy parameter δ. Thus, we wonder whether *the JDP lower bound of He et al. (2022) can be improved by plugging sharper estimation lower bounds under JDP? Is it possible to provide a regret lower bound under JDP using the KL techniques of* Azize and Basu (2022, 2024)?

The vantage point. Finally, Azize and Basu (2024) propose AdaC-OFUL, a ρ-zCDP extension of the Rarely Switching OFUL (Abbasi-Yadkori et al., 2011). AdaC-OFUL shares the same blueprint of the algorithms discussed in Section 2. It runs in adaptive phases. At the beginning of each episode, the least square estimate and the confidence ellipsoid are privately computed. Then, for the whole episode, the same estimate and confidence ellipsoid are used to choose the optimistic action. An episode ends, i.e. we update the estimates, when we accumulate enough "useful information" in the Gram matrix, i.e. when its determinant doubles. However, Azize and Basu (2024) only analyse AdaC-OFUL for *public* and *stochastically* generated contexts, and show that the price of privacy is negligible in this setting. This means that the private context estimation part, i.e. estimating the Gram matrix privately, remains the bottleneck. This raises the question: *Can* AdaC-OFUL *be adapted to private and adversarial contexts with a negligible cost in the regret? If not, then under which conditions on the context generation, i.e. stochastically generated, and with or without margin/diversity conditions, is it possible to achieve JDP for free?*

Acknowledgments

This work has been partially supported by the THIA ANR program "AI_PhD@Lille". D. Basu acknowledges the Inria-Kyoto University Associate Team "RELIANT" for supporting the project, the ANR JCJC for the REPUBLIC project (ANR-22-CE23-0003-01), and the PEPR project FOUNDRY (ANR23-PEIA-0003). We thank Philippe Preux for his support.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Achraf Azize and Debabrota Basu. When privacy meets partial information: A refined analysis of differentially private bandits. *Advances in Neural Information Processing Systems*, 35:32199–32210, 2022.

Achraf Azize and Debabrota Basu. Concentrated differential privacy for bandits. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 78–109. IEEE, 2024.

Achraf Azize, Marc Jourdan, Aymen Al Marjani, and Debabrota Basu. On the complexity of differentially private best-arm identification with fixed confidence. *arXiv preprint arXiv:2309.02202*, 2023.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.

Dirk Bergemann and Juuso Välimäki. Learning and strategic pricing. *Econometrica: Journal of the Econometric Society*, pages 1125–1149, 1996.

T.-H. Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. *ACM Trans. Inf. Syst. Secur.*, 14(3), nov 2011. ISSN 1094-9224. doi: 10.1145/2043621.2043626. URL https://doi.org/10.1145/2043621.2043626.

Cynthia Dwork, Moni Naor, Toniann Pitassi, Guy N Rothblum, and Sergey Yekhanin. Pan-private streaming algorithms. In *ICS*, pages 66–80, 2010.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *International Conference on Machine Learning*, pages 757–765. PMLR, 2014.

Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.

Osama A Hanna, Antonious M Girgis, Christina Fragouli, and Suhas Diggavi. Differentially private stochastic linear bandits:(almost) for free. *arXiv preprint arXiv:2207.03445*, 2022.

Jiahao He, Jiheng Zhang, and Rachel Zhang. A reduction from linear contextual bandit lower bounds to estimation lower bounds. In *International Conference on Machine Learning*, pages 8660–8677. PMLR, 2022.

Bingshan Hu and Nidhi Hegde. Near-optimal thompson sampling-based algorithms for differentially private stochastic bandits. In *Uncertainty in Artificial Intelligence*, pages 844–852. PMLR, 2022.

Bingshan Hu, Zhiming Huang, and Nishant A Mehta. Near-optimal algorithms for private online learning in a stochastic environment. *arXiv preprint arXiv:2102.07929*, 2021.

Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals, 2017. URL https://arxiv.org/abs/1711.03908.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Fengjiao Li, Xingyu Zhou, and Bo Ji. Differentially private linear bandits with partial distributed feedback. In *2022 20th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt)*, pages 41–48. IEEE, 2022.

Touqir Sajed and Or Sheffet. An optimal private stochastic-mab algorithm based on optimal private stopping rule. In *International Conference on Machine Learning*, pages 5579–5588. PMLR, 2019.

Roshan Shariff and Or Sheffet. Differentially private contextual linear bandits. In *Advances in Neural Information Processing Systems*, pages 4296–4306, 2018.

Nicollas Silva, Heitor Werneck, Thiago Silva, Adriano CM Pereira, and Leonardo Rocha. Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions. *Expert Systems with Applications*, 197:116669, 2022.

Abhradeep Guha Thakurta and Adam Smith. (Nearly) optimal algorithms for private online learning in full-information and bandit settings. *Advances in Neural Information Processing Systems*, 26, 2013.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Giuseppe Vietri, Borja Balle, Akshay Krishnamurthy, and Steven Wu. Private reinforcement learning with PAC and regret guarantees. In *International Conference on Machine Learning*, pages 9754–9764. PMLR, 2020.