
Differentially Private Contextual Linear Bandits

Roshan Shariff

Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada
`roshan.shariff@ualberta.ca`

Or Sheffet

Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada
`osheffet@ualberta.ca`

Abstract

We study the contextual linear bandit problem, a version of the standard stochastic multi-armed bandit (MAB) problem where a learner sequentially selects actions to maximize a reward which depends also on a user provided per-round *context*. Though the context is chosen arbitrarily or adversarially, the reward is assumed to be a stochastic function of a feature vector that encodes the context and selected action. Our goal is to devise private learners for the contextual linear bandit problem.

We first show that using the standard definition of differential privacy results in linear regret. So instead, we adopt the notion of *joint* differential privacy, where we assume that the action chosen on day t is only revealed to user t and thus needn't be kept private that day, only on following days. We give a general scheme converting the classic linear-UCB algorithm into a joint differentially private algorithm using the tree-based algorithm [10, 18]. We then apply either Gaussian noise or Wishart noise to achieve joint-differentially private algorithms and bound the resulting algorithms' regrets. In addition, we give the first lower bound on the *additional* regret any private algorithms for the MAB problem must incur.

1 Introduction

The well-known *stochastic multi-armed bandit* (MAB) is a sequential decision-making task in which a learner repeatedly chooses an action (or arm) and receives a noisy reward. The objective is to maximize cumulative reward by *exploring* the actions to discover optimal ones (having the best expected reward), balanced with *exploiting* them. The *contextual* bandit problem is an extension of the MAB problem, where the learner also receives a *context* in each round, and the expected reward depends on *both* the context and the selected action.

As a motivating example, consider online shopping: the user provides a context (composed of query words, past purchases, etc.), and the website responds with a suggested product and receives a reward if the user buys it. Ignoring the context and modeling the problem as a standard MAB (with an action for each possible product) suffers from the drawback of ignoring the variety of users' preferences; whereas separately learning each user's preferences doesn't allow us to generalize between users. Therefore it is common to model the task as a contextual *linear bandit* problem: Based on the user-given context, each action is mapped to a feature vector; the reward probability is then assumed to depend on the *same* unknown linear function of the feature vector across all users.

The above example motivates the need for privacy in the contextual bandit setting: users' past purchases and search queries are sensitive personal information, yet they strongly predict future purchases. In this work, we give upper and lower bounds for the problem of (joint) *differentially private* contextual linear bandits. Differential privacy is the *de facto* gold standard of privacy-preserving data analysis in both academia and industry, requiring that an algorithm's output have very limited dependency on any single user interaction (one context and reward). However, as we later illustrate, adhering to the standard notion of differential privacy (under event-level continual observation) in the contextual bandit requires us to essentially ignore the context and thus incur linear regret. We Preprint. Work in progress.

therefore adopt the more relaxed notion of *joint differential privacy* [23] which, intuitively, allows us to present the t -th user with products corresponding to her preferences, while guaranteeing that all interactions with all users at times $t' > t$ have very limited dependence on user t 's preferences. The guarantee of differential privacy under continuous observation assures us that even if all later users collude in an effort to learn user t 's context or preference, they still have very limited advantage over a random guess.

1.1 Problem Formulation

Stochastic Contextual Linear Bandits. In the classic MAB, in every round t a learner selects an *action* a_t from a fixed set \mathcal{A} and receives a *reward* y_t . In the (stationary) *stochastic* MAB, the reward is noisy with a fixed but unknown expectation $\mathbb{E}[y_t | a_t]$ that depends only on the selected action. In the stochastic contextual bandit problem, before each round the learner also receives a *context* $c_t \in C$ — the expected reward $\mathbb{E}[y_t | c_t, a_t]$ depends on both c_t and a_t . It is common to assume that the context affects the reward in a linear way: map every context-action pair to a *feature vector* $\phi(c, a) \in \mathbb{R}^d$ (where ϕ is an arbitrary but known function) and assume that $\mathbb{E}[y_t | c_t, a_t] = \langle \theta^*, \phi(c_t, a_t) \rangle$. The vector $\theta^* \in \mathbb{R}^d$ is the key unknown parameter of the environment which the learner must discover to maximize reward. Alternatively, we say that on every round the learner is given a *decision set* $\mathcal{D}_t := \{\phi(c_t, a) | a \in \mathcal{A}\}$ of all the pre-computed feature vectors: choosing $x_t \in \mathcal{D}_t$ effectively determines the action $a_t \in \mathcal{A}$. Thus, the contextual stochastic linear bandit framework consists of repeated rounds in which the learner: (i) receives a *decision set* $\mathcal{D}_t \subset \mathbb{R}^d$; (ii) chooses an *action* $x_t \in \mathcal{D}_t$; and (iii) receives a stochastic *reward* $y_t = \langle \theta^*, x_t \rangle + \eta_t$. When all the \mathcal{D}_t are identical and consist of the standard basis vectors, the problem reduces to standard MAB.

The learner's objective is to maximize cumulative reward, which is equivalent to minimizing *regret*: the extra reward a learner would have received by always choosing the best available action. In other words, the regret characterizes the cost of having to *learn* the optimal action over just *knowing* it beforehand. For stochastic problems, we are usually interested in a related quantity called *pseudo-regret*, which is the extra *expected* reward that the learner could have earned if it had known θ^* in advance. In our setting, the cumulative pseudo-regret after n rounds is $\widehat{R}_n := \sum_{t=1}^n \max_{x \in \mathcal{D}_t} \langle \theta^*, x - x_t \rangle$.¹

Joint Differential Privacy. As discussed above, the context and reward may be considered private information about the users which we wish to keep private from all *other* users. We thus introduce the notion of jointly differentially private learners under continuous observation, a combination of two definitions [given in 23, 18]. First, we say two sequences $S = \langle (\mathcal{D}_1, y_1), (\mathcal{D}_2, y_2), \dots, (\mathcal{D}_n, y_n) \rangle$ and $S' = \langle (\mathcal{D}'_1, y'_1), \dots, (\mathcal{D}'_n, y'_n) \rangle$ are *t-neighbors* if for all $t' \neq t$ it holds that $(\mathcal{D}_{t'}, y_{t'}) = (\mathcal{D}'_{t'}, y'_{t'})$.

Definition 1. A randomized algorithm A for the contextual bandit problem is (ϵ, δ) -*jointly differentially private* (JDP) under continual observation if for any t and any pair of t -neighboring sequences S and S' , and any subset $\mathcal{S}_{>t} \subset \mathcal{D}_{t+1} \times \mathcal{D}_{t+2} \times \dots \times \mathcal{D}_n$ of sequence of actions ranging from day $t+1$ to the end of the sequence, it holds that $\mathbb{P}(A(S) \in \mathcal{S}_{>t}) \leq e^\epsilon \mathbb{P}(A(S') \in \mathcal{S}'_{>t}) + \delta$.

The standard notion of differential privacy under continual observation would require that changing the context c_t cannot have much effect on the probability of choosing action a_t — even for round t itself (not just for future rounds as with JDP). In our problem formulation, however, changing c_t to c'_t may change the decision set \mathcal{D}_t to a possibly disjoint \mathcal{D}'_t , making that notion ill-defined. Therefore, when we discuss the impossibility of regret-minimization under standard differential privacy in Section 5, we revert back to a fixed action set \mathcal{A} with an explicit per-round context c_t .

1.2 Our Contributions and Paper Organization

In this work, in addition to formulating the definition of JDP under continual observation, we also present a framework for implementing JDP algorithms for the contextual linear bandit problem. Not surprisingly, our framework combines a tree-based privacy algorithm [10, 18] with a linear upper confidence bound (LinUCB) algorithm [13]. For modularity, in Section 3 we analyze a family of linear UCB algorithms that use different regularizers in every round, under the premise that the regularizers

¹The pseudo-regret ignores the stochasticity of the reward but not the resulting randomness in the learner's choice of actions. It equals the regret in expectation, but is more amenable to high-probability bounds such as ours. In particular, in some cases we can achieve $\text{polylog}(n)$ bounds on pseudo-regret because, unlike regret, it doesn't have added regret noise of variance $\Omega(n)$.

are PSD with bounded singular values. Moreover, we repeat our analysis twice — first we obtain a general $\tilde{O}(\sqrt{n})$ upper bound on regret; then, for problem instances that maintain a Δ reward gap separating the optimal and sub-optimal actions, we obtain a $\text{polylog}(n)/\Delta$ regret upper bound. Our leading application of course is privacy, though one could postulate other reasons where such changing regularizers would be useful (e.g., if parameter estimates turn out to be wrong and have to be updated). We then plug two particular regularizers into our scheme: the first is a privacy-preserving mechanism that uses additive Wishart noise [30] (which is always PSD); the second uses additive Gaussian noise [19] (shifted to make it PSD w.h.p. over all rounds). The main term in the two regret bounds obtained by both algorithms is $\tilde{O}(\sqrt{n} \cdot d^{3/4}/\sqrt{\varepsilon})$ (the bound itself depends on numerous parameters, a notation-list given in Section 2). Details of both techniques appear in Section 4. Experiments with a few variants of our algorithms are detailed in Section D of the supplementary material. In Section 5 we also give a lower bound for the ε -differentially private MAB problem. Whereas all previous work on the private MAB problem uses standard (non-private) bounds, we show that *any* private algorithm must incur *an additional* regret of $\Omega(k \log(n)/\varepsilon)$. While the result resembles the lower bound in the adversarial setting, the proof technique cannot rely on standard packing arguments [e.g. 20] since the input for the problem is stochastic rather than adversarial. Instead, we rely on a recent coupling argument [22] to prove any private algorithm must substantially explore suboptimal arms.

Future Directions. The linear UCB algorithm we adapt in this work is a canonical approach to the linear bandit problem, using the principle of “optimism in the face of uncertainty.” However, recent work [24] shows that all such “optimistic” algorithms are sub-optimal, and instead proposes adapting to the decision set in a particular way by solving an intricate optimization problem. It remains an open question to devise a private version of this algorithm which interpolates between UCB and fine-tuning to the specific action set.

1.3 Related Work

MAB and the Contextual Bandit Problem. The MAB dates to the seminal work of Robbins [28], with the UCB approach developed in a series of works [8, 4] culminating in [6]. Stochastic linear bandits were formally first introduced in [3], and [5] was the first paper to consider UCB-style algorithms. An algorithm that is based on a confidence ellipsoid is described by [13], with a variant based on ridge regression given in [12], or explore-then-commit variant in [29], and a variant related to a sparse setting appears in [2]. Abbasi-Yadkori et al. [1] gives an instance dependent bound for linear bandits, which we convert to the contextual setting.

Differential Privacy. Differential privacy, first introduced by Dwork et al. [17, 16], is a rigorous mathematical notion of privacy that requires the probability of any observable output to change very little when any single datum changes. (We omit the formal definition, having already defined JDP.) Among its many elegant traits is the notion of group privacy: should k datums change then the change in the probability of any event is still limited by (roughly) k times the change when a single datum was changed. Differential privacy also composes: the combination of k (ε, δ) -differentially private algorithms is $(O(k\varepsilon^2 + 2\sqrt{k \log(1/\delta')}, k\delta + \delta'))$ -differentially private for any $\delta' > 0$ [14].

The notion of differential privacy under continual observation was first defined by Dwork et al. [18] using the **tree-based algorithm** [originally appearing in 10]. This algorithm maintains a binary tree whose n leaves correspond to the n entries in the input sequence. Each node in the tree maintains a noisy (privacy-preserving) sum of the input entries in its subtree — the cumulative sums of the inputs can thus be obtained by combining at most $\log(n)$ noisy sums. This algorithm is the key ingredient of a variety of works that deal with privacy in an online setting, including counts [18], online convex optimization [21], and regret minimization in both the adversarial [31, 34] and stochastic [26, 33] settings. We comment that Mishra and Thakurta [26] proposed an algorithm similar to our own for the contextual bandit setting, however (i) without maintaining PSD, (ii) without any analysis, only empirical evidence, and (iii) without presenting lower bounds. A partial utility analysis of this algorithm, in the reward-privacy model (where the context’s privacy is not guaranteed), appears in the recent work of Neel and Roth [27]. Further details about achieving differential privacy via additive noise and the tree-based algorithm appear in Section A of the supplementary material. The related problem of private linear regression has also been extensively studied in the offline setting [11, 7].

2 Preliminaries and Notation

We use ***bold*** letters to denote vectors and bold ***CAPITALS*** for matrices. Given a d -column matrix \mathbf{M} , its *Gram matrix* is the $(d \times d)$ -matrix $\mathbf{M}^\top \mathbf{M}$. A symmetric matrix \mathbf{M} is positive-semidefinite (PSD, denoted $\mathbf{M} \succeq 0$) if $\mathbf{x}^\top \mathbf{M} \mathbf{x} \geq 0$ for any vector \mathbf{x} . Any such \mathbf{M} defines a norm on vectors, so we define $\|\mathbf{x}\|_{\mathbf{M}}^2 = \mathbf{x}^\top \mathbf{M} \mathbf{x}$. We use $\mathbf{M} \succeq \mathbf{N}$ to mean $\mathbf{M} - \mathbf{N} \succeq 0$. The Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ is defined by the density function $(2\pi\sigma^2)^{-1/2} \exp(-(x-\mu)^2/2\sigma^2)$. The squared L_2 -norm of a d -dimensional vector whose coordinates are drawn i.i.d. from $\mathcal{N}(0, 1)$ is given by the $\chi^2(d)$ distribution, which is tightly concentrated around d . Given two distributions \mathbb{P} and \mathbb{Q} we denote their *total variation distance* by $d_{\text{TV}}(\mathbb{P}, \mathbb{Q}) = \max_{\text{event } E} |\mathbb{P}(E) - \mathbb{Q}(E)|$.

Notation. Our bound depends on many parameters of the problem, specified below. Additional parameters (bounds) are specified in the assumptions stated below.

n	horizon, i.e. number of rounds	$m := \lceil \log_2(n) + 1 \rceil$
s, t	indices of rounds	$X_{<t} \in \mathbb{R}^{(t-1) \times d}$, with $X_{<t,s} = \mathbf{x}_s^\top$ for $s < t$
d	dimensionality of action space	\mathbf{G}_t Gram matrix of the actions: $\mathbf{X}_{<t}^\top \mathbf{X}_{<t}$
$\mathcal{D}_t \subset \mathbb{R}^d$	decision set at round t	\mathbf{H}_t regularizer at round t
$\mathbf{x}_t \in \mathcal{D}_t$	action at round t	$\mathbf{y}_{<t}$ vector of rewards up to round $t - 1$
$y_t \in \mathbb{R}$; reward at round t		\mathbf{u}_t action-reward product: $\mathbf{X}_{<t}^\top \mathbf{y}_{<t}$
$\theta^* \in \mathbb{R}^d$; unknown parameter vector		\mathbf{h}_t perturbation of \mathbf{u}_t

3 Linear UCB with Changing Regularizers

In this section we introduce and analyze a variation of the well-studied LinUCB algorithm, an application of the Upper Confidence Bound (UCB) idea to stochastic linear bandits [13, 29, 1]. At every round t , LinUCB constructs a *confidence set* \mathcal{E}_t that contains the unknown parameter vector θ^* with high probability. It then computes an upper confidence bound on the reward of each action in the decision set \mathcal{D}_t , and “optimistically” chooses the action with the highest UCB: $\mathbf{x}_t \leftarrow \arg \max_{\mathbf{x} \in \mathcal{D}_t} \text{UCB}_t(\mathbf{x})$, where $\text{UCB}_t(\mathbf{x}) := \max_{\theta \in \mathcal{E}_t} \langle \theta, \mathbf{x} \rangle$. We assume the rewards are linear with added subgaussian noise (i.e., $y_s = \langle \theta^*, \mathbf{x}_s \rangle + \eta_s$ for $s < t$), so it is natural to center the confidence set \mathcal{E}_t on the (regularized) linear regression estimate:

$$\hat{\theta}_t := \arg \min_{\theta \in \mathbb{R}^d} \|X_{<t} \hat{\theta} - \mathbf{y}_{<t}\|^2 + \|\hat{\theta}\|_{\mathbf{H}_t}^2 = (\mathbf{G}_t + \mathbf{H}_t)^{-1} X_{<t}^\top \mathbf{y}_{<t}. \quad \text{where } \mathbf{G}_t := X_{<t}^\top X_{<t}$$

The matrix $\mathbf{V}_t := \mathbf{G}_t + \mathbf{H}_t \in \mathbb{R}^{d \times d}$ is a regularized version of the Gram matrix \mathbf{G}_t . Whenever the learner chooses an action vector $\mathbf{x} \in \mathcal{D}_t$, the corresponding reward gives it some information about the projection of θ^* onto \mathbf{x} . In other words, the estimate $\hat{\theta}$ is probably closer to θ^* along the directions where many actions have been taken. This motivates the use of ellipsoidal confidence sets that are smaller in such directions, inversely corresponding to the eigenvalues of \mathbf{G}_t (or \mathbf{V}_t). The ellipsoid is uniformly scaled by β_t to achieve the desired confidence level, as prescribed by Proposition 4.

$$\mathcal{E}_t := \{\theta \in \mathbb{R}^d \mid \|\theta - \hat{\theta}_t\|_{\mathbf{V}_t} \leq \beta_t\}, \quad \text{for which } \text{UCB}_t(\mathbf{x}) = \langle \hat{\theta}_t, \mathbf{x} \rangle + \beta_t \|\mathbf{x}\|_{\mathbf{V}_t^{-1}}. \quad (1)$$

Just as the changing regularizer \mathbf{H}_t perturbs the Gram matrix \mathbf{G}_t , our algorithm allows for the vector $\mathbf{u}_t := X_{<t}^\top \mathbf{y}_{<t}$ to be perturbed by \mathbf{h}_t to get $\tilde{\mathbf{u}}_t := \mathbf{u}_t + \mathbf{h}_t$. The estimate $\hat{\theta}_t$ is replaced by $\tilde{\theta}_t := \mathbf{V}_t^{-1} \tilde{\mathbf{u}}_t$.

Algorithm 1 Linear UCB with Changing Perturbations

```

Initialize:  $\mathbf{G}_1 \leftarrow \mathbf{0}_{d \times d}$ ,  $\mathbf{u}_1 \leftarrow \mathbf{0}_d$ .
for each round  $t = 1, 2, \dots, n$  do
    Receive  $\mathcal{D}_t \leftarrow$  decision set  $\subset \mathbb{R}^d$ .
    Receive regularized  $\mathbf{V}_t \leftarrow \mathbf{G}_t + \mathbf{H}_t$  and perturbed  $\tilde{\mathbf{u}}_t \leftarrow \mathbf{u}_t + \mathbf{h}_t$ 
    Compute regressor  $\tilde{\theta}_t \leftarrow \mathbf{V}_t^{-1} \tilde{\mathbf{u}}_t$ 
    Compute confidence-set bound  $\beta_t$  based on Proposition 4.
    Pick action  $\mathbf{x}_t \leftarrow \arg \max_{\mathbf{x} \in \mathcal{D}_t} \langle \tilde{\theta}_t, \mathbf{x} \rangle + \beta_t \|\mathbf{x}\|_{\mathbf{V}_t^{-1}}$ .
    Observe  $y_t \leftarrow$  reward for action  $\mathbf{x}_t$ 
    Update:  $\mathbf{G}_{t+1} \leftarrow \mathbf{G}_t + \mathbf{x}_t \mathbf{x}_t^\top$ ,  $\mathbf{u}_{t+1} \leftarrow \mathbf{u}_t + \mathbf{x}_t y_t$ 
end for

```

Our analysis relies on the following assumptions about the environment and algorithm:

Assumptions (Algorithm 1). For all rounds $t = 1, \dots, n$ and actions $\mathbf{x} \in \mathcal{D}_t$:

1. Bounded action set: $\|\mathbf{x}\| \leq L$.
2. Bounded *mean* reward: $|\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle| \leq B$ with $B \geq 1$.²
3. Bounded target parameter: $\|\boldsymbol{\theta}^*\| \leq S$.
4. All regularizers are PSD $\mathbf{H}_t \succeq 0$.
5. $y_t = \langle \boldsymbol{\theta}^*, \mathbf{x}_t \rangle + \eta_t$ where η_t is σ^2 -conditionally subgaussian on previous actions and rewards, i.e.: $\mathbb{E}[\exp(\lambda\eta_t) | \mathbf{x}_1, y_1, \dots, \mathbf{x}_{t-1}, y_{t-1}, \mathbf{x}_t] \leq \exp(\lambda^2\sigma^2/2)$, for all $\lambda \in \mathbb{R}$.
6. Strongest bound on both the action and the reward: $\|\mathbf{x}_t\|^2 + y_t^2 \leq \tilde{L}^2$ for all rounds t . In particular, if $\|\mathbf{x}_t\| \leq L$ (Assumption 1) and the rewards are bounded: $|y_t| \leq \tilde{B}$ (not just their means as in Assumption 2), we can set $\tilde{L}^2 = L^2 + \tilde{B}^2$.

Assumptions 1 and 2 aren't required to have good pseudo-regret bounds, they merely simplify the bounds on the confidence set (see Proposition 4 and Section 3.1). Assumption 6 is not required at all for now, it is only used for joint differential privacy in Section 4.

To fully describe Algorithm 1 we need to specify how to compute the confidence-set bounds $(\beta_t)_t$. On the one hand, these bounds have to be accurate — the confidence set \mathcal{E}_t should contain the unknown $\boldsymbol{\theta}^*$; on the other hand, the larger they are, the larger the regret bounds we obtain. In other words, the β_t should be as small as possible subject to being accurate.

Definition 2 (Accurate $(\beta_t)_t$). A sequence $(\beta_t)_{t=1}^n$ is (α, n) -accurate for $(\mathbf{H}_t)_{t=1}^n$ and $(\mathbf{h}_t)_{t=1}^n$ if, with probability at least $1 - \alpha$, it satisfies $\|\boldsymbol{\theta}^* - \tilde{\boldsymbol{\theta}}_t\|_{V_t} \leq \beta_t$ for all rounds $t = 1, \dots, n$ simultaneously.

We now argue that three parameters are the key to establishing accurate confidence-set bounds $(\beta_t)_{t=1}^n$ — taking into account the noise in the setting *and* the noise added by a changing \mathbf{H}_t and \mathbf{h}_t .

Definition 3 (Accurate ρ_{\min} , ρ_{\max} , and γ). The bounds $0 < \rho_{\min} \leq \rho_{\max}$ and γ are $(\alpha/2n)$ -accurate for $(\mathbf{H}_t)_{t=1}^n$ and $(\mathbf{h}_t)_{t=1}^n$ if for each round t :

$$\|\mathbf{H}_t\| \leq \rho_{\max}, \quad \|\mathbf{H}_t^{-1}\| \leq 1/\rho_{\min}, \quad \|\mathbf{h}_t\|_{\mathbf{H}_t^{-1}} \leq \gamma; \quad \text{with probability at least } 1 - \alpha/2n.$$

Proposition 4 (Calculating β_t). Suppose Assumptions 3 to 5 hold and let ρ_{\min} , ρ_{\max} , and γ be $(\alpha/2n)$ -accurate for some $\alpha \in (0, 1)$ and horizon n . Then $(\beta_t)_{t=1}^n$ is (α, n) -accurate where

$$\begin{aligned} \beta_t &:= \sigma \sqrt{2 \log(2/\alpha) + \log(\det V_t) - d \log(\rho_{\min})} + S \sqrt{\rho_{\max}} + \gamma \\ &\leq \sigma \sqrt{2 \log(2/\alpha) + d \log\left(\frac{\rho_{\max}}{\rho_{\min}} + \frac{tL^2}{d\rho_{\min}}\right)} + S \sqrt{\rho_{\max}} + \gamma. \end{aligned} \quad (\text{if Assumption 1 also holds})$$

3.1 Regret Bounds

We now present bounds on the maximum regret of Algorithm 1. However, due to space constraints, we defer an extensive discussion of the proof techniques used and the significance of the results to Section B in the supplementary material. The proofs are based on those of Abbasi-Yadkori et al. [1], who analyzed LinUCB with constant regularizers. On one hand, our changes are mostly technical; however, it turns out that various parts of the proof diverge and now depend on ρ_{\max} and ρ_{\min} ; tracing them all is somewhat involved. It is an interesting question to establish similar bounds using known results only as a black box; we were not able to accomplish this.

Theorem 5 (Regret of Algorithm 1). Suppose Assumptions 1 to 5 hold and the β_t are as given by Proposition 4. Then with probability at least $1 - \alpha$ the pseudo-regret of Algorithm 1 is bounded by

$$\widehat{R}_n \leq B \sqrt{8n} \left[\sigma \left(2 \log\left(\frac{2}{\alpha}\right) + d \log\left(\frac{\rho_{\max}}{\rho_{\min}} + \frac{nL^2}{d\rho_{\min}}\right) \right) + (S \sqrt{\rho_{\max}} + \gamma) \sqrt{d \log\left(1 + \frac{nL^2}{d\rho_{\min}}\right)} \right] \quad (2)$$

Theorem 6 (Gap-Dependent Regret of Algorithm 1). Suppose Assumptions 1 to 5 hold and the β_t are as given by Proposition 4. If the optimal actions in every decision set \mathcal{D}_t are separated from the sub-optimal actions by a reward gap of at least Δ , then with probability at least $1 - \alpha$ the pseudo-regret of Algorithm 1 satisfies

$$\widehat{R}_n \leq \frac{8B}{\Delta} \left[\sigma \left(2 \log\left(\frac{2}{\alpha}\right) + d \log\left(\frac{\rho_{\max}}{\rho_{\min}} + \frac{nL^2}{d\rho_{\min}}\right) \right) + (S \sqrt{\rho_{\max}} + \gamma) \sqrt{d \log\left(1 + \frac{nL^2}{d\rho_{\min}}\right)} \right]^2 \quad (3)$$

²See Remark 2 preceding the proof of Lemma 21 in Section B of the supplementary material for a discussion as to bounding B by 1 from below.

4 Linear UCB with Joint Differential Privacy

Notice that Algorithm 1 uses its history of actions and rewards up to round t only via the confidence set \mathcal{E}_t , which is to say via \mathbf{V}_t and $\tilde{\mathbf{u}}_t$, which are perturbations of the Gram matrix \mathbf{G}_t and the vector $\mathbf{u}_t := \mathbf{X}_{\leq t}^\top \mathbf{y}_{\leq t}$, respectively; these also determine β_t . By recording this history with differential privacy, we obtain a Linear UCB algorithm that is jointly differentially private (Definition 1) because it simply post-processes \mathbf{G}_t and \mathbf{u}_t .

Claim 7 (see Dwork and Roth [15, Proposition 2.1]). *If the sequence $(\mathbf{V}_t, \tilde{\mathbf{u}}_t)_{t=1}^{n-1}$ is (ε, δ) -differentially private with respect to $(\mathbf{x}_t, y_t)_{t=1}^{n-1}$, then Algorithm 1 is (ε, δ) -jointly differentially private.*

Remark 1. Algorithm 1 is only jointly differentially private even though the history maintains full differential privacy — its action choice depends not only on the past contexts c_s ($s < t$, via the differentially private $\mathbf{X}_{\leq t}$) but also on the current context c_t via the decision set \mathcal{D}_t . This use of c_t is *not* differentially private, as it is revealed by the algorithm’s chosen \mathbf{x}_t .

Rather than applying the tree-based algorithm separately to \mathbf{G}_t and \mathbf{u}_t , we aggregate both into the single matrix $\mathbf{M}_t \in \mathbb{R}^{(d+1) \times (d+1)}$, which we now construct. Define $\mathbf{A} := [\mathbf{X}_{1:n} \quad \mathbf{y}_{1:n}] \in \mathbb{R}^{n \times (d+1)}$, with \mathbf{A}_t holding the top $t - 1$ rows of \mathbf{A} (and $\mathbf{A}_1 = \mathbf{0}_{1 \times (d+1)}$). Now let $\mathbf{M}_t := \mathbf{A}_t^\top \mathbf{A}_t$ — then the top-left $d \times d$ sub-matrix of \mathbf{M}_t is the Gram matrix \mathbf{G}_t and the first d entries of its last column are \mathbf{u}_t . Furthermore, since $\mathbf{M}_{t+1} = \mathbf{M}_t + [\mathbf{x}_t^\top \quad y_t]^\top [\mathbf{x}_t^\top \quad y_t]$, the tree-based algorithm for private cumulative sums can be used to maintain a private estimation of \mathbf{M}_t using additive noise, releasing $\mathbf{M}_t + \mathbf{N}_t$. The top-left $d \times d$ sub-matrix of \mathbf{N}_t becomes \mathbf{H}_t and the first d entries of its last column become \mathbf{h}_t . Lastly, to have a private estimation of \mathbf{M}_t , Assumption 6 must hold.

Below we present two techniques for maintaining (and updating) the private estimations of \mathbf{M}_t . As mentioned in Section 1.3, the key component of our technique is the tree-based algorithm, allowing us to estimate \mathbf{M}_t using at most $m := 1 + \lceil \log_2 n \rceil$ noisy counters. In order for the entire tree-based algorithm to be (ε, δ) -differentially private, we add noise to each node in the tree so that each noisy count on its own preserves $(\varepsilon/\sqrt{8m \ln(2/\delta)}, \delta/2m)$ -differential privacy. Thus in each day, the noise \mathbf{N}_t that we add to \mathbf{M}_t comes from the sum of at most m such counters.

4.1 Differential Privacy via Wishart Noise

First, we instantiate the tree-based algorithm with noise from a suitably chosen Wishart distribution $\mathcal{W}_{d+1}(\mathbf{V}, k)$, which is the result of sampling k independent $(d + 1)$ -dimensional Gaussians from $\mathcal{N}(\mathbf{0}_{d+1}, \mathbf{V})$ and computing their Gram matrix.

Theorem 8 (Theorem 4.1 [30]). *Fix positive ε_0 and δ_0 . If the L_2 -norm of each row in the input is bounded by \tilde{L} then releasing the input’s Gram matrix with added noise sampled from $\mathcal{W}_{d+1}(\tilde{L}^2 \mathbf{I}, k_0)$ is $(\varepsilon_0, \delta_0)$ -differentially private, provided $k_0 \geq d + 1 + 28\varepsilon_0^{-2} \ln(4/\delta_0)$.*

Applying this guarantee to our setting, where each count needs to preserve $(\varepsilon/\sqrt{8m \ln(2/\delta)}, \delta/2m)$ -differential privacy, it suffices to sample a matrix from $\mathcal{W}_{d+1}(\tilde{L}\mathbf{I}, k)$ with $k := d + 1 + \lceil 224m\varepsilon^{-2} \ln(8m/\delta) \ln(2/\delta) \rceil$. Moreover, the sum of m independent samples from the Wishart distribution is a noise matrix $\mathbf{N}_t \sim \mathcal{W}_{d+1}(\tilde{L}^2 \mathbf{I}, mk)$.³ Furthermore, consider the regularizers \mathbf{H}_t and \mathbf{h}_t derived from \mathbf{N}_t (the top-left submatrix and the right-most subcolumn resp.) — \mathbf{H}_t has distribution $\mathcal{W}_d(\tilde{L}^2 \mathbf{I}, mk)$, and each entry of \mathbf{h}_t is the dot-product of two multivariate Gaussians. Knowing their distribution, we can infer the accurate bounds required for our regret bounds. Furthermore, since the Wishart noise has eigenvalues that are fairly large, we consider a post-processing of the noise matrix — shifting it by $-c\mathbf{I}$ with

$$c := \tilde{L}^2(\sqrt{mk} - \sqrt{d} - \sqrt{2 \ln(8n/\alpha)})^2 - 4\tilde{L}^2\sqrt{mk}(\sqrt{d} + \sqrt{2 \ln(8n/\alpha)}) \quad (4)$$

making the bounds we require smaller than without the shift. The derivations are deferred to Section C of the supplementary material.

³Intuitively, we merely concatenate the m batches of multivariate Gaussians sampled in the generation of each of the m Wishart noises.

Proposition 9. Fix any $\alpha > 0$. If for each t the \mathbf{H}_t and \mathbf{h}_t are generated by the tree-based algorithm with Wishart noise $\mathcal{W}_{d+1}(\tilde{L}^2 \mathbf{I}, k)$, then the following are $(\alpha/2n)$ -accurate bounds:

$$\begin{aligned}\rho_{\min} &= \tilde{L}^2 (\sqrt{mk} - \sqrt{d} - \sqrt{2 \ln(8n/\alpha)})^2, \\ \rho_{\max} &= \tilde{L}^2 (\sqrt{mk} + \sqrt{d} + \sqrt{2 \ln(8n/\alpha)})^2, \\ \gamma &= \tilde{L} (\sqrt{d} + \sqrt{2 \ln(2n/\alpha)}).\end{aligned}$$

Moreover, if we use the shifted regularizer $\mathbf{H}'_t := \mathbf{H}_t - c\mathbf{I}$ with c as given in Eq. (4), then the following are $(\alpha/2n)$ -accurate bounds:

$$\begin{aligned}\rho'_{\min} &= 4\tilde{L}^2 \sqrt{mk} (\sqrt{d} + \sqrt{2 \ln(8n/\alpha)}), \\ \rho'_{\max} &= 8\tilde{L}^2 \sqrt{mk} (\sqrt{d} + \sqrt{2 \ln(8n/\alpha)}), \\ \gamma' &= \tilde{L} \sqrt{\sqrt{mk} (\sqrt{d} + \sqrt{2 \ln(2n/\alpha)})}.\end{aligned}$$

Plugging these into Theorems 5 and 6 gives us the following upper bounds on pseudo-regret.

Corollary 10. Algorithm 1 with \mathbf{H}_t and \mathbf{h}_t generated by the tree-based mechanism with each node adding noise independently from $\mathcal{W}_{d+1}((L^2 + \tilde{B}^2)\mathbf{I}, k)$ and then subtracting $c\mathbf{I}$ using Eq. (4), we get a pseudo-regret bound of

$$O\left(B\sqrt{n} \left[\sigma(\log(1/\alpha) + d \log(n)) + S\tilde{L} \sqrt{d} \log(n)^{3/4} (d^{1/4} + \varepsilon^{-1/2} \log(1/\delta)^{1/4}) (d^{1/4} + \log(n/\alpha)^{1/4}) \right] \right)$$

in general, and a gap-dependent pseudo-regret bound of

$$O\left(\frac{B}{\Delta} \left[\sigma(\log(1/\alpha) + d \log(n)) + S\tilde{L} \sqrt{d} \log(n)^{3/4} (d^{1/4} + \varepsilon^{-1/2} \log(1/\delta)^{1/4}) (d^{1/4} + \log(n/\alpha)^{1/4}) \right]^2 \right)$$

4.2 Differential Privacy via Additive Gaussian Noise

Our second alternative is to instantiate the tree-based algorithm with symmetric Gaussian noise: sample $\mathbf{Z}' \in \mathbb{R}^{(d+1) \times (d+1)}$ with each $\mathbf{Z}'_{i,j} \sim \mathcal{N}(0, \sigma_{\text{noise}}^2)$ i.i.d. and symmetrize to get $\mathbf{Z} = (\mathbf{Z}' + \mathbf{Z}'^\top)/\sqrt{2}$.⁴

Recall that each datum has a bounded L_2 -norm of \tilde{L} , hence a change to a single datum may alter the Frobenius norm of \mathbf{M}_t by \tilde{L}^2 . It follows that in order to make sure each node in the tree-based algorithm preserves $(\varepsilon/\sqrt{8m \ln(2/\delta)}, \delta/2)$ -differential privacy,⁵ the variance in each coordinate must be $\sigma_{\text{noise}}^2 = 16m\tilde{L}^4 \ln(4/\delta)^2/\varepsilon^2$. When all entries on \mathbf{Z} are sampled from $\mathcal{N}(0, 1)$, known concentration results [32] on the top singular value of \mathbf{Z} give that $\mathbb{P}[\|\mathbf{Z}\| > (4\sqrt{d+1} + 2 \ln(2n/\alpha))] < \alpha/2n$. Note however that in each day t the noise N_t is the sum of $\leq m$ such matrices, thus the variance of each coordinate is $m\sigma_{\text{noise}}^2$. The top-left $(d \times d)$ -submatrix of N_t has operator norm of at most

$$\Upsilon := \sigma_{\text{noise}} \sqrt{2m} (4\sqrt{d} + 2 \ln(2n/\alpha)) = \sqrt{32m\tilde{L}^2 \ln(4/\delta)} (4\sqrt{d} + 2 \ln(2n/\alpha))/\varepsilon.$$

However, it is important to note that the result of adding Gaussian noise may not preserve the PSD property of the noisy Gram matrix. To that end, we ought to shift N_t by some $c\mathbf{I}$ in order to make sure that we maintain strictly positive eigenvalues throughout the execution of the algorithm. Since the bounds in Theorems 5 and 6 mainly depend on $\sqrt{\rho_{\max}} + \gamma$, we choose the shift-magnitude to be $2\Upsilon\mathbf{I}$. This makes $\rho_{\max} = 3\Upsilon$ and $\rho_{\min} = \Upsilon$ and as a result $\|\mathbf{h}_t\|_{\mathbf{H}_t^{-1}} \leq \sqrt{\Upsilon^{-1}} \|\mathbf{h}_t\|$, which we can bound using standard concentration bounds on the χ^2 -distribution (see Claim 17). This culminates in the following bounds.

Proposition 11. Fix any $\alpha > 0$. Given that for each t the regularizers $\mathbf{H}_t, \mathbf{h}_t$ are taken by applying the tree-based algorithm with symmetrized shifted Gaussian noise whose entries are sampled i.i.d. from $\mathcal{N}(0, \sigma_{\text{noise}}^2)$, then the following ρ_{\min} , ρ_{\max} , and γ are $(\alpha/2n)$ -accurate bounds:

$$\rho_{\min} = \Upsilon, \quad \rho_{\max} = 3\Upsilon, \quad \gamma = \sigma_{\text{noise}} \sqrt{\Upsilon^{-1} m} (\sqrt{d} + \sqrt{2 \ln(2n/\alpha)}) \leq \sqrt{m\tilde{L}^2 (\sqrt{d} + 2 \ln(2n/\alpha)) / (\sqrt{2}\varepsilon)}$$

⁴This increases the variance along the diagonal entries beyond the noise magnitude required to preserve privacy, but only by a constant factor of 2.

⁵We use here the slightly better bounds for the composition of Gaussian noise based on zero-Concentrated DP [9].

Note how this choice of shift indeed makes both ρ_{\max} and γ^2 roughly on the order of $O(\Upsilon)$.

The end result is that for each day t , \mathbf{h}_t is given by summing at most m d -dimensional vectors whose entries are sampled i.i.d. from $\mathcal{N}(0, \sigma_{\text{noise}}^2)$; the symmetrization doesn't change the distribution of each coordinate. The matrix \mathbf{H}_t is given by: (i) summing at most m matrices whose entries are sampled i.i.d. from $\mathcal{N}(0, \sigma_{\text{noise}}^2)$; (ii) symmetrizing the result as shown above; and (iii) adding $2\Upsilon\mathbf{I}$. This leads to a bound on the regret of Algorithm 1 with the tree-based algorithm using Gaussian noise.

Corollary 12. *Applying Algorithm 1 where the regularizers \mathbf{H}_t and \mathbf{h}_t are derived by applying the tree-based algorithm where each node holds a symmetrized matrix whose entries are sampled i.i.d. from $\mathcal{N}(0, \sigma_{\text{noise}}^2)$ and adding $2\Upsilon\mathbf{I}$, we get a regret bound of*

$$O\left(B\sqrt{n}\left(\sigma(d\log(n) + \log(1/\alpha)) + S\tilde{L}\log(n)\sqrt{d(\sqrt{d} + \ln(n/\alpha))\ln(1/\delta)/\varepsilon}\right)\right)$$

in general, and a gap-dependent pseudo-regret bound of

$$O\left(\frac{B}{\Delta}\left(\sigma(d\log(n) + \log(1/\alpha)) + S\tilde{L}\log(n)\sqrt{d(\sqrt{d} + \ln(n/\alpha))\ln(1/\delta)/\varepsilon}\right)^2\right)$$

5 Lower Bounds

In this section, we present lower bounds for two versions of the problem we deal with in this work. The first, and probably the more obvious of the two, deals with the impossibility of obtaining sub-linear regret for the contextual bandit problem with the standard notion of differential privacy (under continual observation). Here, we assume user t provides a context c_t which actually determines the mapping of the arms into feature vectors: $\phi(c_t, a) \in \mathbb{R}^d$. The sequence of choice thus made by the learner is $a_1, \dots, a_n \in \mathcal{A}^n$ which we aim to keep private. The next claim, whose proof is deferred to Section C in the supplementary material, shows that this is impossible without effectively losing any reasonable notion of utility.

Claim 13. *For any $\varepsilon < \ln(2)$ and $\delta < 0.25$, any (ε, δ) -differentially private algorithm A for the contextual bandit problem must incur pseudo-regret of $\Omega(n)$.*

The second lower bound we show is more challenging. We show that any ε -differentially private algorithm for the classic MAB problem must incur *an additional* pseudo-regret of $\Omega(k\log(n)/\varepsilon)$ on top of the standard (non-private) regret bounds. We consider an instance of the MAB where the leading arm is a^1 , the rewards are drawn from a distribution over $\{-1, 1\}$, and the gap between the means of arm a^1 and arm $a \neq a^1$ is Δ_a . Simple calculation shows that for such distributions, the total-variation distance between two distributions whose means are μ and $\mu - \Delta$ is $\Delta/2$. Fix $\Delta_2, \Delta_3, \dots, \Delta_k$ as some small constants, and we now argue the following.

Claim 14. *Let A be any ε -differentially private algorithm for the MAB problems with k arms whose expected regret is at most $n^{3/4}$. Fix any arm $a \neq a^1$, whose difference between it and the optimal arm a^1 is Δ_a . Then, for sufficiently large ns, A pulls arm a at least $\log(n)/100\varepsilon\Delta_a$ many times w.p. $\geq 1/2$.*

We comment that the bound $n^{3/4}$ was chosen arbitrarily, and we only require a regret upper bound of n^{1-c} for some $c > 0$. Of course, we could have used standard assumptions, where the regret is asymptotically smaller than *any* polynomial; or discuss algorithms of regret $\tilde{O}(\sqrt{n})$ (best minimax regret). Aiming to separate the standard lower-bounds on regret from the private bounds, we decided to use $n^{3/4}$. As an immediate corollary we obtain the following *private* regret bound:

Corollary 15. *The expected pseudo-regret of any ε -differentially private algorithm for the MAB is $\Omega(k\log(n)/\varepsilon)$. Combined with the non-private bound of $\Omega(\sum_{a \neq a^1} \log(n)/\Delta_a)$ we get that the private regret bound is the max of the two terms, i.e.: $\Omega(k\log(n)/\varepsilon + \sum_{a \neq a^1} \log(n)/\Delta_a)$.*

Proof. Based on Claim 14, the expected pseudo-regret is at least $\sum_{a \neq a^1} \frac{\Delta_a \log(n)}{200\varepsilon\Delta_a} = \frac{(k-1)\log(n)}{200\varepsilon}$. \square

Proof of Claim 14. Fix arm a . Let \bar{P} be the vector of the k -probability distributions associated with the k arms. Denote E as the event that arm a is pulled $< \log(n)/100\varepsilon\Delta_a := t_a$ many times. Our goal is to show that $\mathbb{P}_{A; \text{rewards} \sim \bar{P}}[E] < 1/2$.

To that end, we postulate a different distribution for the rewards of arm a — a new distribution whose mean is *greater* by Δ_a than the mean reward of arm a^1 . The total-variation distance between the given distribution and the postulated distribution is Δ_a . Denote \bar{Q} as the vector of distributions of

arm-rewards (where only $P_a \neq Q_a$). We now argue that should the rewards be drawn from \bar{Q} , then the event E is very unlikely: $\mathbb{P}_{A; \text{rewards} \sim \bar{Q}}[E] \leq 2n^{-1/4}/\Delta_a$. Indeed, the argument is based on a standard Markov-like argument: the expected pseudo-regret of A is at most $n^{3/4}$, yet it is at least $\mathbb{P}_{A; \text{rewards} \sim \bar{Q}}[E] \cdot (n - t_a)\Delta_a \geq (n\Delta_a/2)\mathbb{P}_{A; \text{rewards} \sim \bar{Q}}[E]$, for sufficiently large n .

We now apply a beautiful result of Karwa and Vadhan [22, Lemma 6.1], stating that the “effective” group privacy between the case where the n datums of the inputs are drawn i.i.d. from either distribution P or from distribution Q is proportional to $\epsilon n \cdot d_{\text{TV}}(P, Q)$. In our case, the key point is that we only consider this change *under the event E* , thus the number of samples we need to redraw from the distribution P_a rather than Q_a is strictly smaller than t_a , and the elegant coupling argument of [22] reduces it to $6\Delta_a \cdot t_a$. To better illustrate the argument, consider the coupling argument of [22] as an oracle O . The oracle generates a collection of precisely t_a pairs of points, the left ones are i.i.d. samples from P_a and the right ones are i.i.d. samples from Q_a , and, in expectation, in $(1 - \Delta_a)$ fraction of the pairs the right- and the left-samples are identical. Whenever the learner A pulls arm a it makes an oracle call to O , and depending on the environment (whether the distribution of rewards is \bar{P} or \bar{Q}) O provides either a fresh left-sample or a right-sample. Moreover, suppose there exists a counter C that stands between A and O , and in case O runs out of examples then C routes A ’s oracle calls to a different oracle. Now, Karwa and Vadhan [22, Lemma 6.1] assures that the probability of the event “ C never re-routes the requests” happens with similar probability under P or under Q , different only up to a multiplicative factor of $\exp(\epsilon\Delta_a t_a)$. And seeing as the event “ C never re-routes the requests” is quite unlikely when O only provides right-samples (from \bar{Q}), it is also fairly unlikely when O only provides left-samples (from \bar{P}).

Formally, we conclude the proof by applying the result of [22] to infer that $\mathbb{P}_{A; \text{rewards} \sim \bar{P}}[E] \leq \exp(6\epsilon\Delta_a t_a)\mathbb{P}_{A; \text{rewards} \sim \bar{Q}}[E] \leq \exp(0.06 \log(n)) \cdot \frac{2}{\Delta_a} n^{-1/4} = n^{-0.19} \frac{2}{\Delta_a} \leq 1/2$ for sufficiently large n , proving the required claim. \square

Acknowledgements

We gratefully acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC) for supporting R.S. with the Alexander Graham Bell Canada Graduate Scholarship and O.S. with grant #2017-06701. R.S. was also supported by Alberta Innovates and O.S. is also an unpaid collaborator on NSF grant #1565387.

References

- [1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24, pages 2312–2320. Curran Associates, Inc., 2011.
- [2] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *AISTATS*, pages 1–9, 2012.
- [3] Naoki Abe, Alan W. Biermann, and Philip M. Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003.
- [4] Rajeev Agrawal. *Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem.*, volume 27, pages 1054–1078. Applied Probability Trust, 1995.
- [5] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *JMLR*, 3:397–422, 2003.
- [6] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *JMLR*, 47(2-3):235–256, 2002.
- [7] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science, FOCS ’14*, pages 464–473, Washington, DC, USA, 2014. IEEE Computer Society. ISBN 978-1-4799-6517-5. doi: 10.1109/FOCS.2014.56.
- [8] Donald A Berry and Bert Fristedt. *Bandit problems: sequential allocation of experiments*. London ; New York : Chapman and Hall, 1985.

- [9] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography*, Lecture Notes in Computer Science, pages 635–658. Springer, Berlin, Heidelberg, November 2016. ISBN 978-3-662-53640-7 978-3-662-53641-4. doi: 10.1007/978-3-662-53641-4_24.
- [10] T.-H. Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. In *Automata, Languages and Programming*, Lecture Notes in Computer Science, pages 405–417. Springer, Berlin, Heidelberg, July 2010. ISBN 978-3-642-14161-4 978-3-642-14162-1. doi: 10.1007/978-3-642-14162-1_34.
- [11] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 12:1069–1109, July 2011. ISSN 1532-4435.
- [12] Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *AISTATS*, volume 15 of *JMLR Proceedings*, pages 208–214, 2011.
- [13] Varsha Dani, Thomas Hayes, and Sham Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, pages 355–366, January 2008.
- [14] C. Dwork, G. N. Rothblum, and S. Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60, October 2010. doi: 10.1109/FOCS.2010.12.
- [15] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, August 2014. ISSN 1551-305X, 1551-3068. doi: 10.1561/0400000042.
- [16] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology - EUROCRYPT 2006*, Lecture Notes in Computer Science, pages 486–503. Springer, Berlin, Heidelberg, May 2006. ISBN 978-3-540-34546-6 978-3-540-34547-3. doi: 10.1007/11761679_29.
- [17] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, Lecture Notes in Computer Science, pages 265–284. Springer, Berlin, Heidelberg, March 2006. ISBN 978-3-540-32731-8 978-3-540-32732-5. doi: 10.1007/11681878_14.
- [18] Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N. Rothblum. Differential privacy under continual observation. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, STOC ’10, pages 715–724, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0050-6. doi: 10.1145/1806689.1806787.
- [19] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze Gauss: Optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, STOC ’14, pages 11–20, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2710-7. doi: 10.1145/2591796.2591883.
- [20] Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, STOC ’10, pages 705–714, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0050-6. doi: 10.1145/1806689.1806786.
- [21] Prateek Jain, Pravesh Kothari, and Abhradeep Thakurta. Differentially private online learning. In *Conference on Learning Theory*, pages 24.1–24.34, June 2012.
- [22] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. *arXiv:1711.03908 [cs, math, stat]*, November 2017.
- [23] Michael Kearns, Mallesh Pai, Aaron Roth, and Jonathan Ullman. Mechanism design in large games: Incentives and privacy. pages 403–410. ACM Press, 2014. ISBN 978-1-4503-2698-8. doi: 10.1145/2554797.2554834.
- [24] Tor Lattimore and Csaba Szepesvári. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *AISTATS*, pages 728–737, 2017.
- [25] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, October 2000. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1015957395.
- [26] Nikita Mishra and Abhradeep Thakurta. (Nearly) optimal differentially private stochastic multi-arm bandits. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI’15, pages 592–601, Arlington, Virginia, United States, 2015. AUAI Press. ISBN 978-0-9966431-0-8.

- [27] Seth Neel and Aaron Roth. Mitigating bias in adaptive data gathering via differential privacy. In *ICML*, pages 3717–3726, 2018.
- [28] Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58(5):527–535, 09 1952.
- [29] Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, April 2010. ISSN 0364-765X. doi: 10.1287/moor.1100.0446.
- [30] Or Sheffet. Private approximations of the 2nd-moment matrix using existing techniques in linear regression. *arXiv:1507.00056 [cs]*, June 2015.
- [31] Adam Smith and Abhradeep Thakurta. (Nearly) optimal algorithms for private online learning in full-information and bandit settings. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2733–2741. Curran Associates, Inc., 2013.
- [32] Terence Tao. *Topics in Random Matrix Theory*, volume 132. American Mathematical Society Providence, RI, 2012.
- [33] Aristide C. Y. Tossou and Christos Dimitrakakis. Algorithms for differentially private multi-armed bandits. In *Thirtieth AAAI Conference on Artificial Intelligence*, March 2016.
- [34] Aristide C. Y. Tossou and Christos Dimitrakakis. Achieving privacy in the adversarial multi-armed bandit. *arXiv:1701.04222 [cs]*, January 2017.
- [35] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027 [cs, math]*, November 2010.
- [36] Fuzhen Zhang. *Matrix Theory: Basic Results and Techniques*. Universitext. Springer, New York, 2nd edition, 2011. ISBN 978-1-4614-1098-0.

Supplementary Material

A Additional Background Information

A.1 Differential Privacy

In the offline setting, a dataset D is a n -tuple of elements from some universe \mathcal{U} . Two datasets are called neighbors if they differ just on a single element. An algorithm A is said to be (ε, δ) -differentially private if for any pair of neighboring datasets D and D' and any subset of possible outputs \mathcal{S} we have that $\mathbb{P}[A(D) \in \mathcal{S}] \leq e^\varepsilon \mathbb{P}[A(D') \in \mathcal{S}] + \delta$. A common technique [16] for approximating the value of a query f on dataset D is to first find its L_2 -sensitivity, $GS_2 := \max_{D, D' \text{ neighboring}} \|f(D) - f(D')\|_2$, and then add Gaussian noise of 0-mean and variance $\frac{2GS_2^2 \ln(2/\delta)}{\varepsilon^2}$.

A.2 The Tree-Based Mechanism

Assume for simplicity that $n = 2^i$ for some positive integer i . Let T be a complete binary tree with its leaf nodes being l_1, \dots, l_n . Each internal node $x \in T$ stores the sum of all the leaf nodes in the tree rooted at x . First notice that one can compute any partial sum $\sum_{j=1}^i l_i$ using at most $m := \lceil \log(n) + 1 \rceil$ nodes of T . Second, notice that for any two neighboring data sequences D and D' the partial sums stored at no more than m nodes in T are different. Thus, if the count in each node preserves $(\varepsilon_0, \delta_0)$ -differential privacy, using the advanced composition of [14] we get that the entire algorithm is $(O(m\varepsilon_0^2 + \varepsilon_0 \sqrt{2m \ln(1/\delta')}), m\delta_0 + \delta')$ -differentially private. Alternatively, to make sure the entire tree is (ε, δ) -differentially private, it suffices to set $\varepsilon_0 = \varepsilon / \sqrt{8m \ln(2/\delta)}$ and $\delta_0 = \frac{\delta}{2m}$ (with $\delta' = \delta/2$).

A.3 Useful Facts.

In this work, we repeatedly apply the following facts about PSD matrices, the Gaussian distribution, the χ^2 -distribution and the Wishart-distribution.

Claim 16 (36, Theorem 7.8). *If $A \succeq B \geq 0$, then*

1. $\text{rank}(A) \geq \text{rank}(B)$
2. $\det A \geq \det B$
3. $B^{-1} \succeq A^{-1}$ if A and B are nonsingular.

Claim 17 (Corollary to Lemma 1, 25, p. 1325). *If $U \sim \chi^2(d)$ and $\alpha \in (0, 1)$,*

$$\mathbb{P}\left(U \geq d + 2\sqrt{d \ln \frac{1}{\alpha}} + 2 \ln \frac{1}{\alpha}\right) \leq \alpha, \quad \mathbb{P}\left(U \leq d - 2\sqrt{d \ln \frac{1}{\alpha}}\right) \leq \alpha.$$

Claim 18 (Adaptation of 35, Corollary 5.35). *Let A be an $n \times d$ matrix whose entries are independent standard normal variables. Then for every $\alpha \in (0, 1)$, with probability at least $1 - \alpha$ it holds that*

$$\sigma_{\min}(A), \sigma_{\max}(A) \in \sqrt{n} \pm (\sqrt{d} + \sqrt{2 \ln(2/\alpha)})$$

with $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ denoting the smallest- and largest singular values of A resp.

Claim 19 (30, Lemma A.3). *Fix $\alpha \in (0, 1/e)$ and let $W \sim \mathcal{W}_d(V, k)$ with $\sqrt{m} > \sqrt{d} + \sqrt{2 \ln(2/\alpha)}$. Then, denoting the j -th largest eigenvalue of W as $\sigma_j(W)$, with probability at least $1 - \alpha$ it holds that for every $j = 1, 2, \dots, d$:*

$$\sigma_j(W) \in \left(\sqrt{m} \pm \left(\sqrt{d} + \sqrt{2 \ln(2/\alpha)}\right)\right)^2 \sigma_j(V).$$

Claim 20. *For any matrix $H \succeq \rho I > 0$, vector v , and constant $c \geq 0$ satisfying $H - cI \succ 0$,*

$$\|v\|_{(H - cI)^{-1}} \leq \|v\|_{H^{-1}} \sqrt{\frac{\rho}{\rho - c}}.$$

Proof. Since $\mathbf{0} < \rho\mathbf{I} \leq \mathbf{H}$, an application of Claim 16 gives

$$c\mathbf{I} \leq (c/\rho)\mathbf{H}$$

multiplying by $c/\rho \geq 0$

$$\begin{aligned}\mathbf{H} - c\mathbf{I} &\geq \mathbf{H} - (c/\rho)\mathbf{H} = \left(\frac{\rho - c}{\rho}\right)\mathbf{H} \\ (\mathbf{H} - c\mathbf{I})^{-1} &\leq \left(\frac{\rho}{\rho - c}\right)\mathbf{H}^{-1}.\end{aligned}$$

The result follows from the definition of $\|\mathbf{v}\|_{\mathbf{H}^{-1}}$. \square

B Discussion and Proofs from Section 3

The proofs in this section are based on those of Abbasi-Yadkori et al. [1], who analyze the LinUCB algorithm with a constant regularizer. The main difference from that work is that, in our case, quantities involving the regularizer must be bounded above or below (as appropriate) by the constants ρ_{\max} and ρ_{\min} , respectively. We will make extensive use of Claim 16, which shows that for any two matrices $\mathbf{0} \leq \mathbf{V} \leq \mathbf{U}$, we have $\det \mathbf{V} \leq \det \mathbf{U}$ and $\mathbf{V}^{-1} \geq \mathbf{U}^{-1}$. We start by proving the following proposition about the sizes of the confidence ellipsoids, which illustrates this general idea.

Proposition 4 (Calculating β_t). *Suppose Assumptions 3 to 5 hold and let ρ_{\min} , ρ_{\max} , and γ be $(\alpha/2n)$ -accurate for some $\alpha \in (0, 1)$ and horizon n . Then $(\beta_t)_{t=1}^n$ is (α, n) -accurate where*

$$\begin{aligned}\beta_t &:= \sigma\sqrt{2\log(2/\alpha) + \log(\det \mathbf{V}_t) - d\log(\rho_{\min})} + S\sqrt{\rho_{\max}} + \gamma \\ &\leq \sigma\sqrt{2\log(2/\alpha) + d\log\left(\frac{\rho_{\max}}{\rho_{\min}} + \frac{tL^2}{d\rho_{\min}}\right)} + S\sqrt{\rho_{\max}} + \gamma. \quad (\text{if Assumption 1 also holds})\end{aligned}$$

Proof. By definition, $\tilde{\theta}_t = \mathbf{V}_t^{-1}\tilde{\mathbf{u}}_t$, $\tilde{\mathbf{u}}_t = \mathbf{u}_t + \mathbf{h}_t$, and $\mathbf{u}_t = \mathbf{X}_{<t}^\top \mathbf{y}_{<t}$, so that

$$\begin{aligned}\theta^* - \tilde{\theta}_t &= \theta^* - \mathbf{V}_t^{-1}(\mathbf{X}_{<t}^\top \mathbf{y}_{<t} + \mathbf{h}_t) \\ &= \theta^* - \mathbf{V}_t^{-1}(\mathbf{X}_{<t}^\top \mathbf{X}_{<t} \theta^* + \mathbf{X}_{<t}^\top \boldsymbol{\eta}_{t-1} + \mathbf{h}_t) \quad \text{since } \mathbf{y}_{<t} = \mathbf{X}_{<t} \theta^* + \boldsymbol{\eta}_{t-1} \\ &= \theta^* - \mathbf{V}_t^{-1}(\mathbf{V}_t \theta^* - \mathbf{H}_t \theta^* + \mathbf{z}_t + \mathbf{h}_t) \quad \text{defining } \mathbf{z}_t := \mathbf{X}_{<t}^\top \boldsymbol{\eta}_{t-1} \\ &= \mathbf{V}_t^{-1}(\mathbf{H}_t \theta^* - \mathbf{z}_t - \mathbf{h}_t)\end{aligned}$$

Multiplying both sides by $\mathbf{V}_t^{1/2}$ gives

$$\begin{aligned}\mathbf{V}_t^{1/2}(\theta^* - \tilde{\theta}_t) &= \mathbf{V}_t^{-1/2}(\mathbf{H}_t \theta^* - \mathbf{z}_t - \mathbf{h}_t) \\ \|\theta^* - \tilde{\theta}_t\|_{\mathbf{V}_t} &= \|\mathbf{H}_t \theta^* - \mathbf{z}_t - \mathbf{h}_t\|_{\mathbf{V}_t^{-1}} \quad \text{applying } \|\cdot\| \text{ to both sides} \\ &\leq \|\mathbf{H}_t \theta^*\|_{\mathbf{V}_t^{-1}} + \|\mathbf{z}_t\|_{\mathbf{V}_t^{-1}} + \|\mathbf{h}_t\|_{\mathbf{V}_t^{-1}} \quad \text{triangle inequality} \\ &\leq \|\mathbf{z}_t\|_{\mathbf{V}_t^{-1}} + \|\mathbf{H}_t \theta^*\|_{\mathbf{H}_t^{-1}} + \|\mathbf{h}_t\|_{\mathbf{H}_t^{-1}} \quad \text{by Claim 16 since } \mathbf{V}_t \geq \mathbf{H}_t \\ &= \|\mathbf{z}_t\|_{(\mathbf{G}_t + \rho_{\min}\mathbf{I})^{-1}} + \|\theta^*\|_{\mathbf{H}_t} + \|\mathbf{h}_t\|_{\mathbf{H}_t^{-1}}, \quad \text{since } \mathbf{V}_t \geq \mathbf{G}_t + \rho_{\min}\mathbf{I}.\end{aligned}$$

We use a union bound over all n rounds to bound $\|\theta^*\|_{\mathbf{H}_t} \leq \sqrt{\|\mathbf{H}_t\|} \|\theta^*\| \leq S\sqrt{\rho_{\max}}$ and $\|\mathbf{h}_t\|_{\mathbf{H}_t^{-1}} \leq \gamma$ with probability at least $1 - \alpha/2$. Finally, by the ‘‘self-normalized bound for vector-valued martingales’’ of Abbasi-Yadkori et al. [1, Theorem 1], with probability $1 - \alpha/2$ for all rounds simultaneously

$$\|\mathbf{z}_t\|_{(\mathbf{G}_t + \rho_{\min}\mathbf{I})^{-1}} \leq \sigma\sqrt{2\log\frac{2}{\alpha} + \log\frac{\det(\mathbf{G}_t + \rho_{\min}\mathbf{I})}{\det \rho_{\min}\mathbf{I}}} \leq \sigma\sqrt{2\log\frac{2}{\alpha} + \log \det \mathbf{V}_t - d\log \rho_{\min}}.$$

It only remains to show the upper-bound on each β_t . By Claim 16, we have $\det \mathbf{V}_t = \det(\mathbf{G}_t + \mathbf{H}_t) \leq \det(\mathbf{G}_t + \rho_{\max}\mathbf{I})$ and

$$\log \det \mathbf{V}_t \leq \log \det(\mathbf{G}_t + \rho_{\max}\mathbf{I}) \leq d\log(\rho_{\max} + tL^2/d).$$

using the trace-determinant inequality as in the proof of Lemma 22. All the β_t are therefore bounded by the constants

$$\bar{\beta}_t := \sigma\sqrt{2\log\frac{2}{\alpha} + d\log\left(\frac{\rho_{\max}}{\rho_{\min}} + \frac{tL^2}{d\rho_{\min}}\right)} + S\sqrt{\rho_{\max}} + \gamma. \quad \square$$

We now take our first steps towards a regret bound by giving a “generic” version that depends only on LinUCB taking “optimistic” actions, the sizes of the confidence sets, and the rewards being bounded. We rely upon the upper bound for each β_t shown in the previous proposition. We use the Cauchy-Schwarz inequality to bound the sum of per-round regrets r_t by $\sum r_t^2$; this results in the leading $O(\sqrt{n})$ factor in the regret bound. Our gap-dependent analysis later avoids this, but has other trade-offs.

Lemma 21 (Generic LinUCB Regret). *Suppose Assumptions 2 and 4 hold (i.e. $|\langle \theta^*, \mathbf{x} \rangle| \leq B$ and all $\mathbf{H}_t \succeq 0$) and $\bar{\beta}_n \geq \max\{\beta_1, \dots, \beta_n, 1\}$; also assume that $B = 1$. If all the confidence sets \mathcal{E}_t contain θ^* (i.e., $\|\theta^* - \bar{\theta}_t\|_{V_t^{-1}} \leq \beta_t$), then the pseudo-regret of Algorithm 1 is bounded by*

$$\widehat{R}_n \leq \bar{\beta}_n \sqrt{4n \sum_{t=1}^n \min\{1, \|\mathbf{x}_t\|_{V_t^{-1}}^2\}}.$$

Remark 2 (On the quantity B appearing in Assumption 2). For the following proofs, we assume (as in this lemma) that Assumption 2 holds with $B = 1$. Eventually, however, our regret bounds end up with a factor B ; we now explain how. First note that B is trivially at most LS by Cauchy-Schwarz: $|\langle \theta^*, \mathbf{x}_t \rangle| \leq \|\theta^*\| \|\mathbf{x}_t\| \leq LS$. The case where $B < 1$ yet is some constant is trivial: clearly we can take $B = 1$ without violating the assumption. The case where $B = o(1)$ is actually quite intricate and somewhat “unnatural”: while a-priori we know the mean-reward can be as large as LS , it is in fact *much* smaller. This means we have to scale down actions, and shrink the entire problem by a sub-constant; and as a result the noise σ is actually now *far* larger (it is like σ/B in the original setting). While this can be a mere technicality in general, since our leading application is privacy this also means that the bounds on the actual reward we use in Section 4 needs to be scaled by a very large factor. Thus, allowing for ridiculously small B turns into an unnecessary nuisance, and we simply assume that our upper-bound B is not tiny — namely, we assume $B \geq 1$.

It remains to deal with the situation where $B > 1$. In this case, we can pre-process the rewards to the algorithm, scaling them down by a factor of B . If we also scale down all the actions $\mathbf{x} \in \mathcal{D}_t$, then the rest of the assumptions remain inviolate and the regret bound for $B = 1$ applies. However, this bounds the regret in the scaled problem: in the original problem the rewards and hence regret must be scaled up by a factor of B . Note that by only scaling the \mathcal{D}_t , we are not modifying any quantities used in the actual algorithm, just the regret bound; this would not be true if we scaled θ^* (whose maximum norm appears in the β_t), which is the other possibility to maintain the linearity of rewards.

Indeed, in the scaled-down problem the regret is somewhat lower than the bound because both L and σ can be scaled down by B (the noise variance scales proportionally to the reward). For simplicity, however, we refrain from replacing L and σ in the upper bound with L/B and σ/B , respectively.

Proof of Lemma 21. At every round t , Algorithm 1 selects an “optimistic” action \mathbf{x}_t satisfying

$$(\mathbf{x}_t, \bar{\theta}_t) \in \arg \max_{(\mathbf{x}, \theta) \in \mathcal{D}_t \times \mathcal{E}_t} \langle \theta, \mathbf{x} \rangle. \quad (5)$$

Let $\mathbf{x}_t^* \in \arg \max_{\mathbf{x} \in \mathcal{D}_t} \langle \theta^*, \mathbf{x} \rangle$ be an optimal action and $r_t = \langle \theta^*, \mathbf{x}_t^* - \mathbf{x}_t \rangle$ be the immediate pseudo-regret suffered for round t :

$$\begin{aligned} r_t &= \langle \theta^*, \mathbf{x}_t^* \rangle - \langle \theta^*, \mathbf{x}_t \rangle \\ &\leq \langle \bar{\theta}_t, \mathbf{x}_t \rangle - \langle \theta^*, \mathbf{x}_t \rangle && \text{from (5) since } (\mathbf{x}_t^*, \theta^*) \in \mathcal{D}_t \times \mathcal{E}_t \\ &= \langle \bar{\theta}_t - \theta^*, \mathbf{x}_t \rangle \\ &= \langle V_t^{1/2}(\bar{\theta}_t - \theta^*), V_t^{-1/2}\mathbf{x}_t \rangle && \text{since } V_t \succeq \mathbf{H}_t \succeq 0 \\ &\leq \|\bar{\theta}_t - \theta^*\|_{V_t} \|\mathbf{x}_t\|_{V_t^{-1}} && \text{by Cauchy-Schwarz} \\ &\leq (\|\bar{\theta}_t - \theta^*\|_{V_t} + \|\theta^* - \bar{\theta}_t\|_{V_t}) \|\mathbf{x}_t\|_{V_t^{-1}} && \text{by the triangle inequality} \\ &\leq 2\beta_t \|\mathbf{x}_t\|_{V_t^{-1}} && \text{since } \bar{\theta}_t, \theta^* \in \mathcal{E}_t \\ &\leq 2\bar{\beta}_n \|\mathbf{x}_t\|_{V_t^{-1}} && \text{since } \bar{\beta}_n \geq \beta_t. \end{aligned}$$

From our assumptions that the mean absolute reward is at most 1 and $\bar{\beta}_n \geq 1$, we also get that $r_t \leq 2 \leq 2\bar{\beta}_n$. Putting these together,

$$r_t \leq 2\bar{\beta}_n \min\{1, \|\mathbf{x}_t\|_{V_t^{-1}}\} \quad (6)$$

Now we apply the Cauchy-Schwarz inequality, since $\widehat{R}_n = \langle \mathbf{1}_n, \mathbf{r}/n \rangle$, where $\mathbf{1}_n$ is the all-ones vector and \mathbf{r} is the vector of per-round regrets:

$$\widehat{R}_n^2 = n^2 \left(\sum_{t=1}^n \frac{r_t}{n} \right)^2 \leq n^2 \sum_{t=1}^n \frac{r_t^2}{n} = n \sum_{t=1}^n r_t^2 \leq 4n\bar{\beta}_n^2 \sum_{t=1}^n \min\{1, \|\mathbf{x}_t\|_{V_t^{-1}}^2\}.$$

Taking square roots completes the proof. \square

The following technical lemma relates the quantity from the previous result to the volume (i.e. determinant) of the V_n matrix. We will see shortly that the U_t are all lower bounds on the V_t .

Lemma 22 (Elliptical Potential). *Let $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ be vectors with each $\|\mathbf{x}_t\| \leq L$. Given a positive definite matrix $U_1 \in \mathbb{R}^{d \times d}$, define $U_{t+1} := U_t + \mathbf{x}_t \mathbf{x}_t^\top$ for all t . Then*

$$\sum_{t=1}^n \min\{1, \|\mathbf{x}_t\|_{U_t^{-1}}^2\} \leq 2 \log \frac{\det U_{n+1}}{\det U_1} \leq 2d \log \frac{\text{tr } U_1 + nL^2}{d \det^{1/d} U_1}.$$

Proof. We use the fact that $\min\{1, u\} \leq 2 \log(1 + u)$ for any $u \geq 0$:

$$\sum_{t=1}^n \min\{1, \|\mathbf{x}_t\|_{U_t^{-1}}^2\} \leq 2 \sum_{t=1}^n \log(1 + \|\mathbf{x}_t\|_{U_t^{-1}}^2).$$

We will show that this last summation is $2 \log(\det U_{n+1}/\det U_n)$. For all t , we have

$$\begin{aligned} U_{t+1} &= U_t + \mathbf{x}_t \mathbf{x}_t^\top = U_t^{1/2} (I + U_t^{-1/2} \mathbf{x}_t \mathbf{x}_t^\top U_t^{-1/2}) U_t^{1/2} \\ \det U_{t+1} &= \det U_t \det(I + U_t^{-1/2} \mathbf{x}_t \mathbf{x}_t^\top U_t^{-1/2}). \end{aligned}$$

Consider the eigenvectors of the matrix $I + \mathbf{y}\mathbf{y}^\top$ for an arbitrary vector $\mathbf{y} \in \mathbb{R}^d$. We know that \mathbf{y} itself is an eigenvector with eigenvalue $1 + \|\mathbf{y}\|^2$:

$$(I + \mathbf{y}\mathbf{y}^\top)\mathbf{y} = \mathbf{y} + \mathbf{y}\langle\mathbf{y}, \mathbf{y}\rangle = (1 + \|\mathbf{y}\|^2)\mathbf{y}.$$

Moreover, since $I + \mathbf{y}\mathbf{y}^\top$ is symmetric, every other eigenvector \mathbf{u} is orthogonal to \mathbf{y} , so that

$$(I + \mathbf{y}\mathbf{y}^\top)\mathbf{u} = \mathbf{u} + \mathbf{y}\langle\mathbf{y}, \mathbf{u}\rangle = \mathbf{u}.$$

Therefore the only eigenvalues of $I + \mathbf{y}\mathbf{y}^\top$ are $1 + \|\mathbf{y}\|^2$ (with eigenvector \mathbf{y}) and 1. In our case $\mathbf{y} = U_t^{-1/2} \mathbf{x}_t$ and $\|\mathbf{y}\|^2 = \mathbf{x}_t^\top U_t^{-1} \mathbf{x}_t = \|\mathbf{x}_t\|_{U_t^{-1}}^2$, so we get our first inequality:

$$\begin{aligned} \det U_{n+1} &= \det U_1 \prod_{t=1}^n (1 + \|\mathbf{x}_t\|_{U_t^{-1}}^2) \\ 2 \log \frac{\det U_{n+1}}{\det U_1} &= 2 \sum_{t=1}^n \log(1 + \|\mathbf{x}_t\|_{U_t^{-1}}^2). \end{aligned}$$

To get the second inequality, we apply the arithmetic-geometric mean inequality to the eigenvalues λ_i of U_n :

$$\begin{aligned} \det U_n &= \prod_{i=1}^d \lambda_i \leq \left(\frac{1}{d} \sum_{i=1}^d \lambda_i \right)^d = ((1/d) \text{tr } U_n)^d \leq ((\text{tr } U_1 + nL^2)/d)^d \\ 2 \log \frac{\det U_n}{\det U_1} &\leq 2d \log \frac{\text{tr } U_1 + nL^2}{d \det^{1/d} U_1} \end{aligned}$$

\square

We are finally in a position to prove the main regret theorem. The proof is straightforward and essentially comes down to plugging in our preceding results.

Theorem 5 (Regret of Algorithm 1). *Suppose Assumptions 1 to 5 hold and the β_t are as given by Proposition 4. Then with probability at least $1 - \alpha$ the pseudo-regret of Algorithm 1 is bounded by*

$$\widehat{R}_n \leq B\sqrt{8n} \left[\sigma \left(2 \log \left(\frac{2}{\alpha} \right) + d \log \left(\frac{\rho_{\max}}{\rho_{\min}} + \frac{nL^2}{d\rho_{\min}} \right) \right) + (S\sqrt{\rho_{\max}} + \gamma) \sqrt{d \log \left(1 + \frac{nL^2}{d\rho_{\min}} \right)} \right] \quad (2)$$

Proof. We restrict ourselves to the event that all the confidence ellipsoids contain θ^* and all $\rho_{\min} \mathbf{I} \preceq H_t \preceq \rho_{\max} \mathbf{I}$. Proposition 4 assures us that this happens with probability at least $1 - \alpha$, and

furthermore gives us the bound $\beta_t \leq \bar{\beta}_n$:

$$\bar{\beta}_n := \sigma \sqrt{2 \log \frac{2}{\alpha} + d \log \left(\frac{\rho_{\max}}{\rho_{\min}} + \frac{nL^2}{d\rho_{\min}} \right)} + S\sqrt{\rho_{\max}} + \gamma.$$

Next, we have $\|\mathbf{x}_t\|_{V_t^{-1}} \leq \|\mathbf{x}_t\|_{(\mathbf{G}_t + \rho_{\min}\mathbf{I})^{-1}}$, which applied to the result of Lemma 21 gives, using Lemma 22

$$\begin{aligned} \widehat{R}_n &\leq \bar{\beta}_n \sqrt{8dn \log \left(1 + \frac{nL^2}{d\rho_{\min}} \right)} \\ &\leq \sqrt{8n} \left[\sigma \left(2 \log \frac{2}{\alpha} + d \log \left(\frac{\rho_{\max}}{\rho_{\min}} + \frac{nL^2}{d\rho_{\min}} \right) \right) + (S\sqrt{\rho_{\max}} + \gamma) \sqrt{d \log \left(1 + \frac{nL^2}{d\rho_{\min}} \right)} \right]. \end{aligned}$$

The argument outlined in Remark 2 preceding the proof of Lemma 21 tells us how to reintroduce the missing factor of B in this regret bound. \square

The proof of the gap-dependent regret bound diverges from the previous proof in only one major way: the gap is used to bound each r_t by r_t^2/Δ . Then the sum of r_t^2 is bounded as before; this avoids the \sqrt{n} factor introduced by the use of the Cauchy-Schwarz inequality.

Theorem 6 (Gap-Dependent Regret of Algorithm 1). *Suppose Assumptions 1 to 5 hold and the β_t are as given by Proposition 4. If the optimal actions in every decision set \mathcal{D}_t are separated from the sub-optimal actions by a reward gap of at least Δ , then with probability at least $1 - \alpha$ the pseudo-regret of Algorithm 1 satisfies*

$$\widehat{R}_n \leq \frac{8B}{\Delta} \left[\sigma \left(2 \log \left(\frac{2}{\alpha} \right) + d \log \left(\frac{\rho_{\max}}{\rho_{\min}} + \frac{nL^2}{d\rho_{\min}} \right) \right) + (S\sqrt{\rho_{\max}} + \gamma) \sqrt{d \log \left(1 + \frac{nL^2}{d\rho_{\min}} \right)} \right]^2 \quad (3)$$

Proof. Because of the gap assumption, for every round t if the per-round pseudo-regret $r_t \neq 0$ then $r_t \geq \Delta$. We use this fact to decompose the regret in a different way than we did in Lemma 21. The rest of the proof is similar to that of Theorem 5. As before, see Remark 2 preceding the proof of Lemma 21 to introduce the missing B factor.

$$\begin{aligned} \widehat{R}_n &= \sum_{t \in B_n} r_t \leq \sum_{t \in B_n} \frac{r_t^2}{\Delta} \\ &\leq \frac{4}{\Delta} \bar{\beta}_n^2 \sum_{t \in B_n} \min \{1, \|\mathbf{x}_t\|_{V_t^{-1}}^2\} \quad \text{from (6)} \\ &\leq \frac{8}{\Delta} \bar{\beta}_n^2 d \log \left(1 + \frac{nL^2}{d\rho_{\min}} \right) \\ &\leq \frac{8}{\Delta} \left[\sigma \left(2 \log \frac{2}{\alpha} + d \log \left(\frac{\rho_{\max}}{\rho_{\min}} + \frac{nL^2}{d\rho_{\min}} \right) \right) + (S\sqrt{\rho_{\max}} + \gamma) \sqrt{d \log \left(1 + \frac{nL^2}{d\rho_{\min}} \right)} \right]^2 \quad \square \end{aligned}$$

B.1 Regret Bounds Open Problem

The first conclusion from these regret bounds is that allowing changing regularizers does not incur significant additional regret, as long as they are bounded both above and below. Broadly speaking, these bounds for contextual linear bandits match those for standard MAB algorithms in terms of their dependence on n and Δ — just like with UCB, for example, the minimax bound is $O(\sqrt{n})$ and the gap-dependent bound is $O(\log(n)/\Delta)$. However, the dependence on d (which corresponds to the number of arms for the MAB) is much worse, with $O(d)$ in the minimax case and $O(d^2)$ in the gap-dependent case.

It is an interesting open question whether the $O(d^2)$ dependence on d is necessary to achieve $O(\log n)$ gap-dependent regret bounds. As we were unable to prove a lower bound of $\Omega(d^2)$, we resorted to empirically checking the performance of the (non-private) LinUCB on such Δ -gap instances; the results can be found in Section D.1.

C Privacy Proofs

We now provide the missing privacy proofs from the main body of the paper. First, we give the omitted proof from Section 4.1.

Proposition 9. *Fix any $\alpha > 0$. If for each t the \mathbf{H}_t and \mathbf{h}_t are generated by the tree-based algorithm with Wishart noise $\mathcal{W}_{d+1}(\tilde{L}^2 \mathbf{I}, k)$, then the following are $(\alpha/2n)$ -accurate bounds:*

$$\begin{aligned}\rho_{\min} &= \tilde{L}^2 (\sqrt{mk} - \sqrt{d} - \sqrt{2 \ln(8n/\alpha)})^2, \\ \rho_{\max} &= \tilde{L}^2 (\sqrt{mk} + \sqrt{d} + \sqrt{2 \ln(8n/\alpha)})^2, \\ \gamma &= \tilde{L} (\sqrt{d} + \sqrt{2 \ln(2n/\alpha)}).\end{aligned}$$

Moreover, if we use the shifted regularizer $\mathbf{H}'_t := \mathbf{H}_t - c\mathbf{I}$ with c as given in Eq. (4), then the following are $(\alpha/2n)$ -accurate bounds:

$$\begin{aligned}\rho'_{\min} &= 4\tilde{L}^2 \sqrt{mk} (\sqrt{d} + \sqrt{2 \ln(8n/\alpha)}), \\ \rho'_{\max} &= 8\tilde{L}^2 \sqrt{mk} (\sqrt{d} + \sqrt{2 \ln(8n/\alpha)}), \\ \gamma' &= \tilde{L} \sqrt{\sqrt{mk} (\sqrt{d} + \sqrt{2 \ln(2n/\alpha)})}.\end{aligned}$$

Proof. Seeing as $\mathbf{H}_t \sim \mathcal{W}_d(\tilde{L}^2 \mathbf{I}, mk)$, straight-forward bounds on the eigenvalues of the Wishart distribution (e.g. [30], Lemma A.3) give that w.p. $\geq 1 - \alpha/2n$ all of the eigenvalues of \mathbf{H}_t lie in the interval $\tilde{L}^2 (\sqrt{mk} \pm (\sqrt{d} + \sqrt{2 \ln(8n/\alpha)}))^2$. To bound $\|\mathbf{h}_t\|_{\mathbf{H}_t^{-1}}$ we draw back to the definition of the Wishart distribution as the Gram matrix of samples from a multivariate Gaussian $\mathcal{N}(\mathbf{0}, \tilde{L}^2 \mathbf{I})$. Denote this matrix of Gaussians as $[\mathbf{Z}; \mathbf{z}]$ where $\mathbf{Z} \in \mathbb{R}^{mk \times d}$ and $\mathbf{z} \in \mathbb{R}^{mk}$, and we have that $\mathbf{H}_t = \mathbf{Z}^\top \mathbf{Z}$ and $\mathbf{h}_t = \mathbf{Z}^\top \mathbf{z}$, thus $\|\mathbf{h}_t\|_{\mathbf{H}_t^{-1}} = \sqrt{\mathbf{z}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{z}}$. The matrix $\mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top$ is a projection matrix onto a d -dimensional space, and projecting the spherical Gaussian \mathbf{z} onto this subspace results in a d -dimensional spherical Gaussian. Using concentration bounds on the χ^2 -distribution (Claim 17) we have that w.p. $\geq 1 - \alpha/2n$ it holds that $\|\mathbf{h}_t\|_{\mathbf{H}_t^{-1}} \leq \gamma := \tilde{L} (\sqrt{d} + \sqrt{2 \ln(2n/\alpha)})$.

It is straightforward to modify these bounds for the shifted regularizer matrix $\mathbf{H}'_t := \mathbf{H}_t - c\mathbf{I}$; the minimum and maximum eigenvalues are bounded as $\rho'_{\min} = \rho_{\min} - c$ and $\rho'_{\max} = \rho_{\max} - c$, respectively. The value of c in Eq. (4) is chosen so that $\rho'_{\min} = \rho_{\min} - c = \rho_{\max} - \rho_{\min} = 4\tilde{L}^2 \sqrt{mk} (\sqrt{d} + \sqrt{2 \ln(8n/\alpha)})$. It follows that $\rho'_{\max} = \rho_{\max} - c = \rho'_{\min} + \rho_{\max} - \rho_{\min} = 2\rho'_{\min}$. Finally, Claim 20 gives

$$\begin{aligned}\|\mathbf{h}_t\|_{\mathbf{H}'_t^{-1}} &\leq \|\mathbf{h}_t\|_{\mathbf{H}_t^{-1}} \sqrt{\rho_{\min}/\rho'_{\min}} \leq \gamma \sqrt{\rho_{\min}/\rho'_{\min}} \\ &= \tilde{L} (\sqrt{d} + \sqrt{2 \ln(2n/\alpha)}) \sqrt{\frac{\tilde{L}^2 (\sqrt{mk} - \sqrt{d} - \sqrt{2 \ln(8n/\alpha)})^2}{4\tilde{L}^2 \sqrt{mk} (\sqrt{d} + \sqrt{2 \ln(8n/\alpha)})}} \\ &\leq \tilde{L} \frac{(\sqrt{mk} - \sqrt{d} - \sqrt{2 \ln(8n/\alpha)}) \sqrt{\sqrt{d} + \sqrt{2 \ln(2n/\alpha)}}}{4(mk)^{1/4}} \\ &\leq \tilde{L} \sqrt{\sqrt{mk} (\sqrt{d} + \sqrt{2 \ln(2n/\alpha)})} =: \gamma'\end{aligned}\quad \square$$

Theorem 23 (30, Theorem 4.1). *Fix $\varepsilon \in (0, 1)$ and $\delta \in (0, 1/e)$. Let $A \in \mathbb{R}^{n \times p}$ be a matrix whose rows have l_2 -norm bounded by \tilde{L} . Let W be a matrix sampled from the d -dimensional Wishart distribution with k degrees of freedom using the scale matrix $\tilde{L}^2 \mathbf{I}p$ (i.e. $W \sim \mathcal{W}_p(\tilde{L}^2 \mathbf{I}p, k)$) for $k \geq p + \lfloor \frac{14}{\varepsilon^2} \cdot 2 \log(4/\delta) \rfloor$. Then outputting $A^\top A + N$ is (ε, δ) -differentially private with respect to changing a single row of A .*

We now give the proof of the lower bound of any private algorithm under the standard notion of differential privacy under continual observation, as discussed in Section 5. First, of course, we need to define this notion. Formally, two sequences $S = \langle (c_1, y_1), \dots, (c_n, y_n) \rangle$ and $S' = \langle (c'_1, y'_1), \dots, (c'_n, y'_n) \rangle$ are called neighbors if there exists a single t such that for any $t' \neq t$ we have $(c_{t'}, y_{t'}) = (c'_{t'}, y'_{t'})$; and an algorithm A is (ε, δ) -differentially private if for any two neighboring sequences S and S' and any subsets of sequences of actions $\mathcal{S} \subset \mathcal{A}^n$ it holds that $\mathbb{P}[A(S) \in \mathcal{S}] \leq e^\varepsilon \mathbb{P}[A(S') \in \mathcal{S}] + \delta$. We now prove the following.

Claim 13. For any $\varepsilon < \ln(2)$ and $\delta < 0.25$, any (ε, δ) -differentially private algorithm A for the contextual bandit problem must incur pseudo-regret of $\Omega(n)$.

Proof. We consider a setting with two arms $\mathcal{A} = \{a^1, a^2\}$ and two possible contexts: c^1 which maps $a^1 \mapsto \theta^*$ and $a^2 \mapsto -\theta^*$; and c^2 which flips the mapping. Assuming $\|\theta^*\| = 1$ it is evident we incur a pseudo-regret of 2 when pulling arm a^1 is under context c^2 or pulling arm a^2 under c^1 . Fix a day t and a history of previous inputs and arm pulls H_{t-1} . Consider a pair of neighboring sequences that agree on the history H_{t-1} and differ just on day t — in S the context $c_t = c^1$ whereas in S' it is set as $c_t = c^2$. Denote \mathcal{S} as the subset of action sequences that are fixed on the first $t-1$ days according to H_{t-1} , have the t -th action be a^1 and on days $> t$ may have any action. Thus, applying the guarantee of differential privacy w.r.t to \mathcal{S} we get that $\mathbb{P}[a_t = a^1 | S] = \mathbb{P}[A(S) \in \mathcal{S}] \leq e^\varepsilon \mathbb{P}[a_t = a^1 | S'] + \delta$. Consider an adversary that sets the context of day t to be either c^1 or c^2 uniformly at random and independently of other days. The pseudo-regret incurred on day t is thus: $2 \cdot \frac{1}{2} (\mathbb{P}[a_t = a^2 | S] + \mathbb{P}[a_t = a^1 | S']) \geq (1 - \mathbb{P}[a_t = a^1 | S]) + e^{-\varepsilon}(\mathbb{P}[a_t = a^1 | S] - \delta) = 1 + (e^{-\varepsilon} - 1)\mathbb{P}[a_t = a^2 | S] - \delta > 1 - 1 \cdot \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$. As the above applies to any day t , the algorithm's pseudo-regret is $\geq \frac{n}{4}$ against such random adversary. \square

D Experiments

We performed some experiments with synthetic data to characterize the performance of the algorithms in this paper.

Setting. We first describe the common setting used for all the experiments: Given a dimension d , we first select θ^* to be a random unit vector in \mathbb{R}^d (distributed uniformly on the hyper-sphere, so that $S = 1$). Then we construct decision sets of size K ($K = d^2$ in our experiments), consisting of one *optimal action* and $K - 1$ *suboptimal actions*, all of unit length (so $L = 1$). The optimal action is chosen uniformly at random from the $(d-2)$ -dimensional set $\{\mathbf{x} \in \mathbb{R}^d | \|\mathbf{x}\| = 1, \langle \mathbf{x}, \theta^* \rangle = 0.75\}$. The suboptimal actions are chosen independently and uniformly at random from the $(d-1)$ -dimensional set $\{\mathbf{x} \in \mathbb{R}^d | \|\mathbf{x}\| = 1, \langle \mathbf{x}, \theta^* \rangle \in [-0.75, 0.65]\}$. This results in a suboptimality gap of $\Delta = 0.1$, since the optimal arm has mean reward 0.75 and the suboptimal arms have mean rewards in the $[-0.75, 0.65]$ interval. To simulate the contextual bandit setting, a new decision set is sampled before each round.

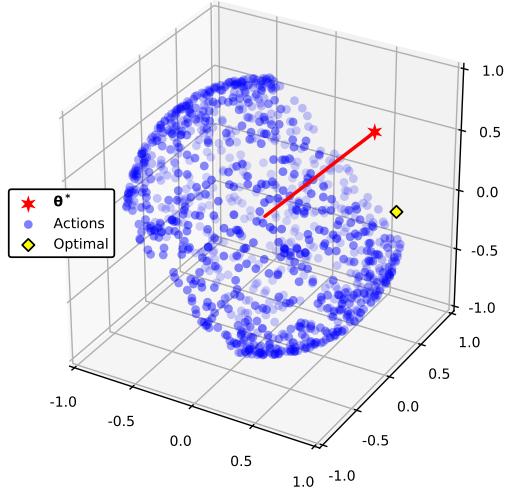


Figure 1: An example decision set in \mathbb{R}^3 with $K = 1000$ actions.

The rewards are either -1 or $+1$, with probabilities chosen so that $\mathbb{E}[y_t] = \langle \mathbf{x}_t, \theta^* \rangle$. Therefore $\tilde{B} = 1$ and, being bounded in the $[-1, 1]$ interval, the reward distribution is subgaussian with $\sigma^2 = 1$.

The experiments below measure the expected regret in each case; the confidence parameter is $\alpha = 1/n$, which is the usual choice when one wishes to minimize expected regret.

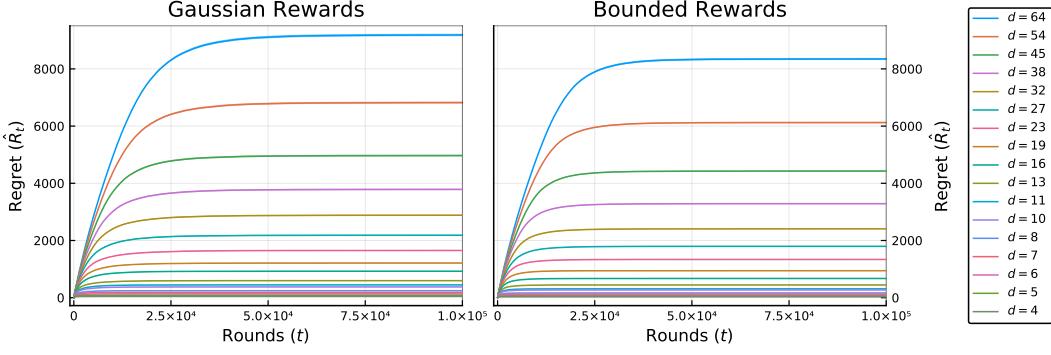


Figure 2: Experiment 1 — Regret over time for varying dimensions.

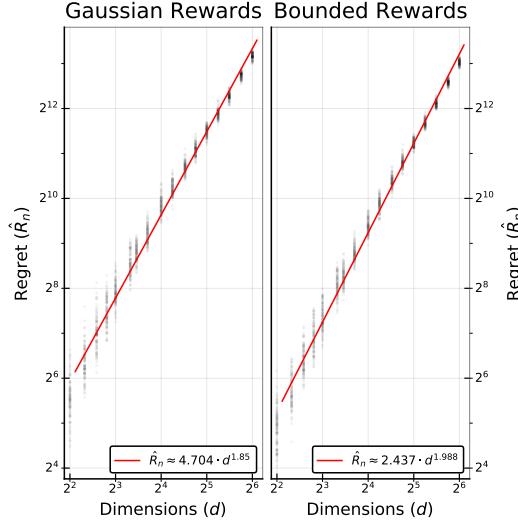


Figure 3: Experiment 1 — Regret vs. dimension with log–log axes and best-fit line.

D.1 The Dependency of the Pseudo-Regret on the Dimension for Gap Instances

The first experiment was aimed at the open question of Section B.1, namely whether the gap-dependent regret is $\Omega(d^2)$ in the dimension of the problem. Thus privacy wasn't a concern in this particular setting; rather, our goal was to determine the performance of our general recipe algorithm in a contextual setting with a clear-cut gap. We measured the pseudo-regret of the non-private LinUCB algorithm as a function of the dimension over $n = 10^5$ rounds with the regularizer $\rho = I_{d \times d}$ and $K = d^2$ arms. The values of d were logarithmically spaced in the interval $[4, 64]$. The results of the experiment are plotted in Fig. 2. The two sub-experiments differ only in the reward noise distribution used. In the first, the reward noise is truly a Gaussian with $\sigma^2 = 1$, whereas in the second the reward is ± 1 as described above (subgaussian with $\sigma^2 = 1$). In the latter case, the actual variance in the reward depends on its expectation, and is somewhat lower than 1. This is perhaps why the regret is somewhat lower than with gaussian reward noise.

Figure 3 shows the same results with total accumulated regret plotted against dimension using a log–log scale. The best-fit line on this plot has a slope of roughly 2, clearly pointing to a super-linear dependency on d . We conjecture that, in general, the dependency on d is indeed quadratic.

D.2 Empirical Performance of the Privacy-Preserving Algorithms over Time

This experiment compares the expected regret of the various algorithm variants presented in this paper. The two major privacy-preserving algorithms are based on Wishart noise (Section 4.1) and Gaussian noise (Section 4.2); both were run with privacy parameters $\varepsilon = 1.0$ and $\delta = 0.1$ over a horizon of $n = 5 \times 10^7$ rounds and dimension $d = 5$ and $K = d^2 = 25$. The results are shown in Fig. 4; the curves are truncated after 2×10^7 rounds because they are essentially flat after this point.

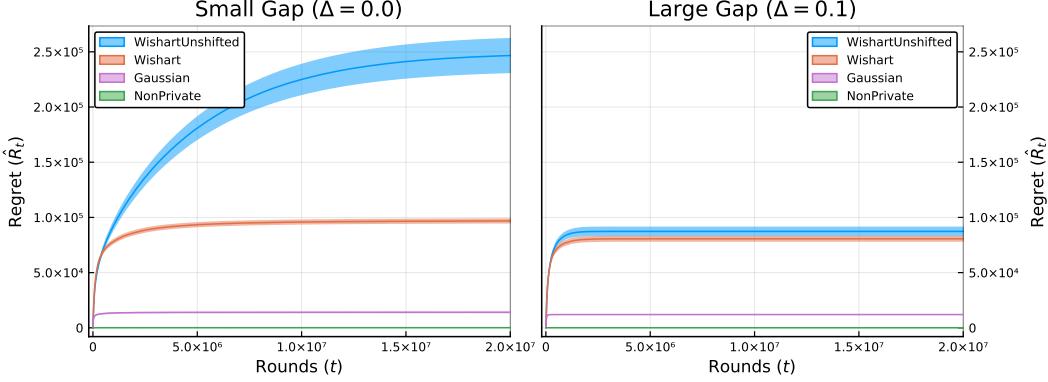


Figure 4: Experiment 2 — Regret over time, with and without forced sub-optimality gaps.

The sub-figures of Fig. 4 show two settings that differ in the sub-optimality gap Δ between the rewards of the optimal and sub-optimal arms. In the left sub-figure, the algorithms are run in a setting without a structured gap ($\Delta = 0$), where we have not forced all arms to be strictly separated from the optimal arm by a large reward gap. Here, all sub-optimal arms are distributed uniformly on the set $\{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\| = 1, \langle \mathbf{x}, \theta^* \rangle \in [-0.75, 0.75]\}$ (and *not* from $[-0.75, 0.65]$ as in the previous experiment). Note that while we cannot guarantee that in *all* rounds there exists a gap between the optimal and sub-optimal arms, it is still true that *in expectation* we should observe a gap of $\Theta(1/K)$ between the optimal arm and the second-best arm (and as $K = 25$ this expected gap is, still, a constant in comparison to n). In the right sub-figure, however, the sub-optimal arms are indeed separated by a gap of $\Delta = 0.1$ from the optimal arms; their rewards lie in the interval $[-0.75, 0.65]$ as in the previous experiment. In both cases there is always an optimal arm with reward 0.75.

The figures show the following algorithm variants:

- The NonPrivate algorithm is LinUCB with regularizer $\rho = 1.0$. Its regret is too small to be distinguished from the x-axis in this plot.
- The Gaussian variant is described in Section 4.2.
- The Wishart variant is described in Section 4.1 with the shift given in Eq. (4).
- The WishartUnshifted variant is that of Section 4.1 but with no shift.

Results. It is apparent that, at least for this setting, the Gaussian noise algorithm outperforms Wishart noise. This shows that while the asymptotic performance of the two algorithms is fairly close, the constants in the Gaussian version of the algorithm are far better than the ones in the Wishart-noise based algorithm.

Furthermore, the performance of the WishartUnshifted variant changes significantly between the two cases — it has the worst regret in the no-gap setting ($\Delta = 0$) but, surprisingly, it is statistically indistinguishable from the shifted Wishart variant in the large gap instance ($\Delta = 0.1$). We investigate this relationship between the sub-optimality gap and shifted regularizers in the next experiment.

D.3 Empirical Performance of Shifted Regularizers for Different Suboptimality Gaps

In both the Wishart and Gaussian variants of our algorithm, we use a *shifted* regularization matrix $\mathbf{H}_t \pm c\mathbf{I}$, choosing the shift parameter c to approximately optimize our regret bound in each case. This optimal shift parameter turns out not to depend on the sub-optimality gap Δ of the problem instance. The previous experiment showed, however, that in practice the relative performance of the shifted and unshifted Wishart variants changes drastically depending on the gap. In this experiment, we investigate the impact of varying the shift parameter for the Wishart and Gaussian mechanisms under different sub-optimality gaps Δ .

All the parameters are the same as the previous experiment — the only difference is the shift parameter; the results are shown in Fig. 5. The two sub-figures show the performance of the Wishart and Gaussian variants, respectively. The x-axis is a logarithmic scale indicating ρ_{\min} , the high-probability lower bound on the minimum eigenvalue of the shifted regularizer matrix. ρ_{\min} serves as a good proxy for

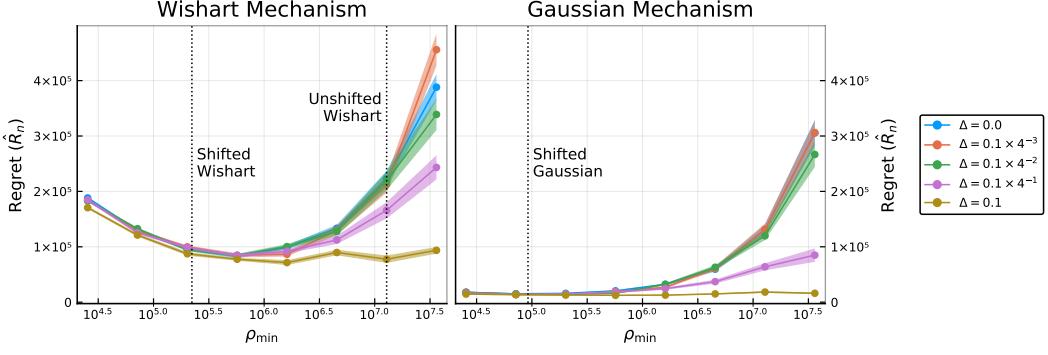


Figure 5: Experiment 3 — Varying shift parameters with different sub-optimality gaps.

the shift parameter because changing one has the effect of shifting the other by the same amount; it has the added benefit of being meaningfully comparable amongst the different algorithm variants. The vertical dotted lines indicate the ρ_{\min} values corresponding to the algorithms from Sections 4.1 and 4.2 for the Wishart and Gaussian variants, respectively; these are also the algorithms examined in the previous experiment. The Gaussian mechanism does not have an unshifted variant.

Results. Tuning the shift parameter appears to significantly affect the performance only for problem instances with relatively small or zero sub-optimality gaps. In the large-gap settings, on the other hand, having too much regularization does not seem to increase regret appreciably. The small-gap settings are exactly those in which exploration is crucial, so we conjecture that large regularizers inhibit exploration and thereby incur increased regret.