Differential Privacy for Multi-armed Bandits: What Is It and What Is Its Cost?

Debabrota Basu, Christos Dimitrakakis, Aristide Tossou
Department of Computer Science and Engineering
Chalmers University of Technology
Göteborg, Sweden
(basud,chrdimi,tossou)@chalmers.se

Abstract

We introduce a number of privacy definitions for the multi-armed bandit problem, based on differential privacy. We relate them through a unifying graphical model representation and connect them to existing definitions. We then derive and contrast lower bounds on the regret of bandit algorithms satisfying these definitions. We show that for all of them, the learner's regret is increased by a multiplicative factor dependent on the privacy level ϵ, but that the dependency is weaker when we do not require local differential privacy for the rewards.

1 Introduction

Multi-armed Bandits. The stochastic K-armed bandit problem (Bellman, 1956; Lattimore and Szepesvári, 2018) involves a learner sequentially choosing among K different arms so as to maximise her expected cumulative reward. More precisely, At time t, the algorithm *draws an arm* $A_t = a \in [K]$ and the arms *generate rewards* $\mathbf{r}_t = [r_{t,j}]_{j=1}^{K} \in \mathbb{R}^K$. Then the learner *obtains reward* $X_t = r_{t,a}$, without observing the other rewards.

In the stochastic bandit problem, the environment ν where the learner is acting consists of a set of reward distributions $\{f_1, \ldots, f_K\}$ with means $\mu_a \triangleq \mathbb{E}(f_a)$, and optimal expected reward $\mu^* \triangleq \max_a \mu_a$ so that the reward distribution for each arm is $\mathbb{P}_\nu(X_t = r_{t,a}) = f_a$ for all t.

The learner's policy π for selecting actions is generally a stochastic mapping $\pi : \mathcal{H} \to \mathbf{\Delta}([K])$. Here $\mathcal{H}$ is the *observed history*, i.e. the sequence of actions taken and rewards obtained by the learner. The objective is to policy *maximise the expected cumulative reward*,

$$S(\pi, \nu, T) \triangleq \sum_{t=1}^{T} \mathbb{E}_\pi^\nu[X_t] = \sum_{a=1}^{K} \mathbb{E}_\pi^\nu \left[N_a(T)\right] \mu_a,$$

where $N_a(T)$ denotes the number of time arm a is pulled till time step T and μ_a is the expected reward of the arm a. The quality of a learning algorithm π is best summarised by its *(expected cumulative) regret*, i.e. its loss in total reward relative to an oracle that knows ν:

$$\mathrm{Reg}(\pi, \nu, T) \triangleq \sum_{a=1}^{K} \mathbb{E}_\pi^\nu \left[N_a(T)\right] (\mu^* - \mu_a).$$

This is the cost incurred by the algorithm due to the incomplete information, as it has to play the sub-optimal arms to gain information about the suboptimal arms. This process decreases the uncertainty in decision making and facilitates maximisation of the expected cumulative reward.

Lower Bounds on Regret. Lower bounds illustrate the inherent hardness of bandit problems. Lai and Robbins (1985) proved that any consistent bandit policy π must incur at least logarithmic

Preprint. Under review.

growth in expected cumulative regret. This lower bound is problem-dependent as it has a multiplicative factor dependent on the given environment ν. Vogel (1960) proved an environment-independent lower bound of $\Omega(\sqrt{KT})$. This also called a *problem-independent minimax lower bound* as the minimax regret is as $\mathrm{Reg}_{\mathrm{minimax}}(T) \triangleq \min_\pi \max_\nu \mathrm{Reg}(\pi, \nu, T)$. A similar lower bound of $\Omega(\sqrt{KT})$ is established for the *Bayesian regret* under any prior (Lattimore and Szepesvári, 2018). A detailed description of the existing lower bounds is provided in Section 4.

Differential Privacy is a rigorous and highly successful definition of algorithmic privacy introduced by Dwork et al. (2006). Given a definition of neighbourhood between possible inputs to an algorithm, it differential privacy allows us to design algorithms which maintain data privacy by ensuring that the algorithm's output renders neighbouring inputs indistinguishable. Informally, if an algorithm is ϵ-differentially private, then the amount of information that inferred by an adversary about the algorithm's input is bounded by ϵ.

One specific way to implement differential privacy is to ensure that the input to the algorithm is already a differentially private version of the original data. This notion of privacy is called *local differential privacy* (Duchi et al., 2013), and it allows the algorithm to be agnostic about privacy. For this reason, this notion is presently adapted by Apple and Google for their large-scale systems.

Our contributions. In Section 2, we define, discuss, and unify different notions of differential privacy. In particular, we examine the effect of considering different notions of private input, observable output and neighbourhoods on the regret of multi-armed bandits that are operating under a differential privacy constraint. We illustrate the differences between those definitions using graphical models. This graphical model definition also invokes a new notion of privacy in bandits.

In Section 3, we provide a unified framework to prove minimax lower bounds for both differentially private multi-armed bandits. This is based on a generalised KL-divergence decomposition lemma adapted for local and standard differential privacy definitions. Though the literature consists of problem-dependent regret bounds, these are the first minimax and Bayesian regret bounds for both differentially private bandits. We show that both in general and when differential privacy is achieved using a local mechanism, the regret scales as a multiplicative factor of ϵ. As expected, local privacy mechanisms have a slightly worse performance. Note that our results do not contradict the upper bound of Tossou and Dimitrakakis (2016), since they achieve an additive regret loss with respect to instantaneous privacy, rather than the stricter sequential privacy definition that a constant privacy loss can be achieved with only an additive term on the regret cannot be true.

In Section 4, we elaborate that the proposed lower bounds pose several open problems of designing optimal bandit algorithms that satisfy different notions of privacy.

2 Differential Privacy for Bandits

Local Differential Privacy. There is a stronger notion of privacy proposed by Duchi et al. (2013), called local privacy, where the user does not trust the authority of the algorithm. Instead, she uses a privacy-preserving mechanism $\mathcal{M}$ to send locally differentially private inputs to the algorithm.

Definition 1 ϵ-Local Differential Privacy (Duchi et al., 2013). *A local privacy-preserving mechanism $\mathcal{M}$ is ϵ-local differentially private if for all privatised inputs $Z \subseteq$ Range$(\mathcal{M})$ and neighbouring input datasets $D, D' \in \mathcal{D}$ with Hamming distance $d_H(D, D') = 1$:*

$$\log\left(\left|\frac{\mathbb{P}_{\mathcal{M}}(Z \mid D)}{\mathbb{P}_{\mathcal{M}}(Z \mid D')}\right|\right) \leq \epsilon.$$

Local Differential Privacy for Bandits. If the bandit algorithm only observes a private version of the reward sequence, then the algorithm's output is differentially private with respect to the *generated reward sequence* $\{\mathbf{r}_1, \ldots, \mathbf{r}_T\} = \{\mathbf{r}_i\}_{i=1}^T$.

Definition 2 ϵ-Local Private Reward Sequence for Bandits. *A mechanism $\mathcal{M}$ preserves local privacy for rewards if*

$$\log\left(\left|\frac{\mathbb{P}_{\mathcal{M}}(\mathbf{Z} \mid \{\mathbf{r}_i\}_{i=1}^T)}{\mathbb{P}_{\mathcal{M}}(\mathbf{Z} \mid \{\mathbf{r}'_i\}_{i=1}^T)}\right|\right) \leq \epsilon,$$

for all privatised reward sequences $\mathbf{Z} \in \mathbb{R}^{KT}$ and generated reward sequences $\{\mathbf{r}_i\}_{i=1}^T, \{\mathbf{r}'_i\}_{i=1}^T \in \mathbb{R}^{KT}$ with Hamming distance $d_H\left(\{\mathbf{r}_i\}_{i=1}^T, \{\mathbf{r}'_i\}_{i=1}^T\right) = 1$.

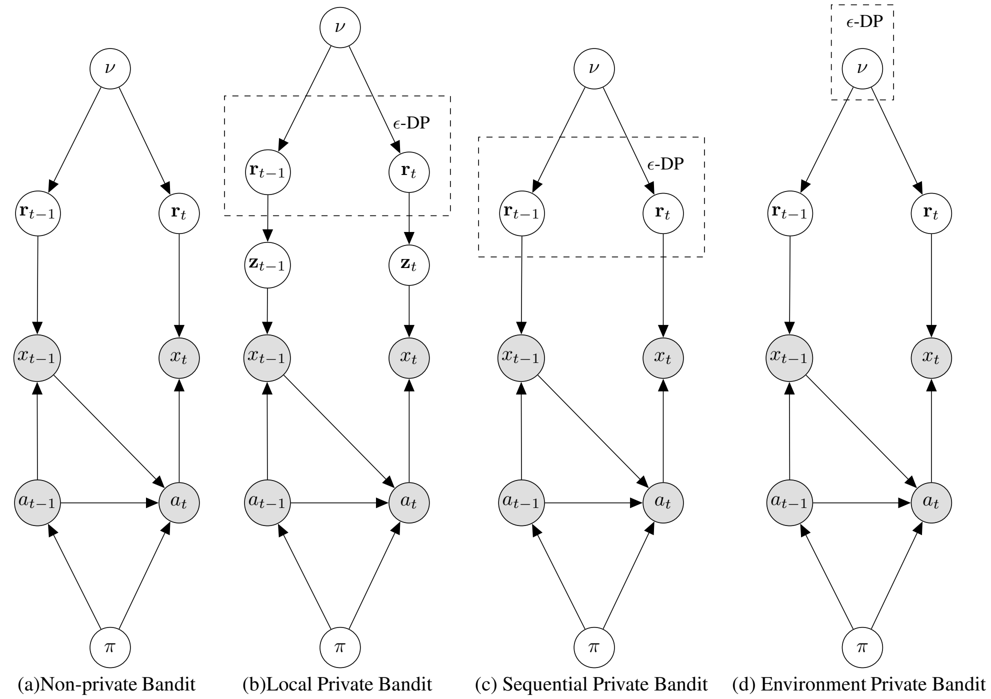

Figure 1: Graphical models for non-private, local private, sequential private, and environment private multi-armed bandits.

Definition 2 is analogous to the local privacy definition in the corrupted bandit setting of Gajane et al. (2017) but generalises the notion of privacy from obtained rewards to the generated rewards.

Differential Privacy. Dwork et al. (2014) proposed the notion of ϵ-differential privacy for a given dataset D belonging to a corpus $\mathcal{D}$ and an algorithm π equipped with a privacy preserving mechanism $\mathcal{M}$ on top of it. This definition of differential privacy aims to keep the input dataset D private from an adversary that can access only the privatised outputs Z of the privacy-preserving algorithm $\mathcal{M} \circ \pi$.

Definition 3 ϵ-Differential privacy (Dwork et al., 2014). *A privacy-preserving mechanism $\mathcal{M}$ composed with an algorithm π is ϵ-differentially private if for all $Z \subseteq$ Range$(\mathcal{M})$ and $D, D' \in \mathcal{D}$ such that $D \sim D'$:*

$$\log\left(\left|\frac{\mathbb{P}((\mathcal{M} \circ \pi)(D) \in Z)}{\mathbb{P}((\mathcal{M} \circ \pi)(D') \in Z)}\right|\right) \leq \epsilon.$$

In the notion of differential privacy, the algorithm π gets the input accurately and the privacy preserving mechanism is applied inside it or on its output. A privacy-preserving algorithm $\mathcal{M} \circ \pi$ provides perfect privacy ($\epsilon = 0$) if it yields indistinguishable outputs for all neighbouring input datasets. The privacy level ϵ quantifies the privacy guarantee provided by ϵ-differential privacy. A smaller value of privacy level ϵ indicates higher privacy.

Differential Privacy for Bandits. In case of sequential decision making problems like bandit, the notion of privacy and neighbouring dataset can be defined in several ways. We discuss them in the following section.

Case 1: We consider *the generated history* h_T of a privacy-preserving bandit algorithm π at time T as *the input dataset*. Here, generated history $g_T \triangleq \{(a_i, \mathbf{r}_i)\}_{i=1}^T$ is defined as the set of actions chosen and all the rewards generated till time T. Now, we can define a neighbouring input dataset i.e. a neighbouring generated history in two ways:

1.1. In the case 1.1., a neighbouring generated history differs from a given generated history by one of the actions but all the generated rewards are the same such that $g_T^A \triangleq \{(a_1', \mathbf{r}_1)\} \cup \{(a_i, \mathbf{r}_i)\}_{i=2}^T$.

1.2. In the case 1.2., a neighbouring generated history differs from a given generated history by one of the generated rewards but all the actions are the same such that $g_T^R \triangleq \{(a_1, \mathbf{r}_1')\} \cup \{(a_i, \mathbf{r}_i)\}_{i=2}^T$.

We consider *the action selected at time A_{T+1} as the output*. These notions of input dataset, neighbouring datasets, and output allow us to formulate two definitions of differential privacy.

Definition 4 Privacy for Bandits in Case 1.1. (ϵ-Pan-privacy). *A privacy-preserving bandit algorithm π is ϵ-pan private if for all actions taken $A_{T+1} \in [K]$ and neighbouring generated histories $g_T, g_T^A \in ([K] \times \mathbb{R}^K)^T$:*

$$\log \left(\left| \frac{\mathbb{P}_\pi \left(A_{T+1} \in [K] \mid g_T \right)}{\mathbb{P}_\pi \left(A_{T+1} \in [K] \mid g_T^A \right)} \right| \right) \leq \epsilon.$$

Definition 5 Privacy for Bandits in Case 1.2. (ϵ-Instantaneous Privacy). *A privacy-preserving bandit algorithm π is ϵ-instantaneous private if for all actions taken $A_{T+1} \in [K]$ and neighbouring generated histories $g_T, g_T^R \in ([K] \times \mathbb{R}^K)^T$:*

$$\log \left(\left| \frac{\mathbb{P}_\pi \left(A_{T+1} \in [K] \mid g_T \right)}{\mathbb{P}_\pi \left(A_{T+1} \in [K] \mid g_T^R \right)} \right| \right) \leq \epsilon.$$

Definition 4 is analogous to the definition of pan-privacy (Dwork et al., 2010) as the notion of neighbouring histories in Case 1.1 and neighbouring input datasets in pan-privacy (Definition 3 in (Mir et al., 2010)) are identical. Our notion also generalises the pan-privacy to bandit setup as the pan-privacy considers only the obtained reward sequence as the input and secures it but we consider the generated reward sequence as the input which is a superset of the obtained reward. Hereafter, we refer to Definition 4 as the *pan-privacy for bandits*. Definition 5 is analogous to the definition of differential privacy used in (Tossou and Dimitrakakis, 2016). We refer to Definition 5 as the *instantaneous privacy for bandits*. The joint differential privacy definition proposed by Shariff and Sheffet (2018) is reducible to either Definition 4 or 5 depending on the notion of neighbouring dataset. Our definitions generalise all these definitions as they depend only on the obtained rewards.

Case 2: We consider *the sequence of generated rewards till time T*, i.e. $\{\mathbf{r}_i\}_{i=1}^T \in \mathbb{R}^{KT}$ as *the input dataset* and *the sequence of actions taken till time T*, i.e. $\{a_i\}_{i=1}^T \in [K]^T$ as *the output dataset*. Thus, analogous to the local differential privacy for bandits, we define two generated reward sequences $\{\mathbf{r}_i\}_{i=1}^T$ and $\{\mathbf{r}'_i\}_{i=1}^T$ to be neighbouring if their Hamming distance is 1. Now, we define the corresponding global privacy and refer to it as the *sequential privacy*.

Definition 6 ϵ-Sequential Privacy for Bandits. *A privacy-preserving bandit algorithm π preserves ϵ-sequential differential privacy if for all action sequences $\{a_i\}_{i=1}^T \in [K]^T$ and neighbouring generated reward sequences $\{\mathbf{r}_i\}_{i=1}^T, \{\mathbf{r}'_i\}_{i=1}^T \in \mathbb{R}^{KT}$,*

$$\log \left(\left| \frac{\mathbb{P}_\pi \left(\{a_i\}_{i=1}^T \in [K]^T \mid \{\mathbf{r}_i\}_{i=1}^T \right)}{\mathbb{P}_\pi \left(\{a_i\}_{i=1}^T \in [K]^T \mid \{\mathbf{r}'_i\}_{i=1}^T \right)} \right| \right) \leq \epsilon.$$

We limit ourselves to the study of sequential privacy here, which is equivalent to the definition given in (Tossou and Dimitrakakis, 2017). This is because an algorithm satisfying sequential privacy at level ϵ also satisfies instantaneous privacy (Definition 5) with privacy level 2ϵ. More precisely:

Lemma 1. *If a policy π satisfies sequential privacy (Definition 6) with privacy level ϵ, π will also satisfy instantaneous privacy (Definition 5) with privacy level 2ϵ. Conversely, if a policy satisfies ϵ instantaneous privacy, it only achieves $t\epsilon$ sequential privacy after t steps.*

Discussion. *We obtain a unified framework for privacy in bandits as sequential differential privacy (Definition 6) and the local differential privacy (Definition 2) adopt the same notion of input dataset and neighbouring dataset.* We observe that sequential differential privacy also ensures the instantaneous differential privacy (Definition 5). The other definition of privacy (Definition 4) being the pan-privacy for bandits does not ensure differential privacy for more than one observation by the adversary (Dwork et al., 2010). On contrary, sequential differential privacy (Definition 6) ensures differential privacy under continuous observation. This property is desired in the sequential setting of bandits. Additionally, Definition 4 of differential privacy imposes a constraint on the actions taken by the bandit algorithm. This may lead to a narrower space of feasible bandit algorithms which is not desired from the algorithm design perspective.

Unified Graphical Model Representation of Privacy for Bandits. Figure 2 provides a unified graphical model perspective of non-private, private (locally and not) multi-armed bandits. The shaded nodes represent the observed variable. The clear nodes represent the hidden variables. The dashed-rectangle covers the input quantities with respect to which the privacy has to be maintained. All of these representations treat the generated rewards as input and all possible action sequences as output with local and sequential privacy-mechanisms acting at two different levels ensuring local and sequential privacies respectively. This representation allows us to define a new notion of privacy for bandits, namely *environment privacy*.

Environment Privacy for Bandits. For a bandit with a stationary environment ν, the reward generation mechanism can be represented as a distribution with an environment-dependent parameter ν. The user may consider this *environment parameter to be the input* and *the generated histories* g_T, i.e. the sequence of generated rewards $\{\mathbf{r}_i\}_{i=1}^T \in \mathbb{R}^{KT}$ and actions taken $\{a_i\}_{i=1}^T \in [K]^T$ by environment parameter ν and the policy π as *the output*.

Definition 7 ϵ-Environment Privacy for Bandits. *A privacy-preserving mechanism $\mathcal{M}$ preserves ϵ-environment privacy if for all generated histories $g_T \triangleq \{(a_i, \mathbf{r}_i)\}_{i=1}^T \in ([K] \times \mathbb{R}^K)^T$, and environment parameters $\nu, \nu' \in \mathbb{R}^d$,*

$$\log\left(\left|\frac{\mathbb{P}_{\pi\nu}\left(\{(a_i,\mathbf{r}_i)\}_{i=1}^T \mid \nu\right)}{\mathbb{P}_{\pi\nu'}\left(\{(a_i,\mathbf{r}_i)\}_{i=1}^T \mid \nu'\right)}\right|\right) \leq \epsilon\rho(\nu,\nu'),$$

where ρ is a distance metric defined in the space of ν.

Example. In order to understand how each definition affects privacy, consider the example of web advertising for a specific individual. In this example, at time t the individual is presented with some set of advertisements a_t. These advertisements generate potential responses $\mathbf{r}_t$. Out of these generated responses of the user, we only see the clicked response x_t. Let us assume that we use a bandit algorithm π in order to perform adaptive web adverstising. If π is ϵ-sequential private (Definition 6) with respect to $\{\mathbf{r}_t\}$, an adversary cannot distinguish similar responses between individuals. If we are locally ϵ-sequential-DP (Definition 2), even the authority of the algorithm would not be able to distinguish between $\{\mathbf{r}_t\}$. Thus indistinguishability of individuals can be achieved for both the adversary and the authority of the algorithm. Finally, if we are ϵ-envrionment private with respect to ν (Definition 7), no adversary can infer the inherent preferences of the individual. For all of our definitions, the privacy loss is bounded by a constant privacy level ϵ independent of the length of interactions.

3 Regret Lower Bounds for Private Bandits

Lower bounds on the performance measure of a problem provides us insight about the intrinsic hardness of the problem and sets a target for optimal algorithm design. In this section, we prove minimax and Bayesian lower bounds for local and sequential private bandits respectively. We also prove a problem-dependent lower bound for local privacy.

In order to prove the lower bounds, we adopt the general canonical bandit model (Lattimore and Szepesvári, 2018) that is general enough not to impose additional constraints on the bandit algorithm and the environment. A privacy-preserving bandit algorithm π and an environment ν interacts up to a given time horizon T to produce *observed history* $\mathcal{H}_T \triangleq \{(A_i, X_i)\}_{i=1}^T$. Thus, an observed history $\mathcal{H}_T$ is a random variable sampled from the measurable space $\left(([K]\times\mathbb{R})^T, \mathcal{B}([K]\times\mathbb{R})^T\right)$ and a probability measure $\mathbb{P}_{\pi\nu}$. Here, $\mathcal{B}([K]\times\mathbb{R})^T$ is the Borel set on $([K]\times\mathbb{R})^T$. $\mathbb{P}_{\pi\nu}$ is the probability measure induced by the algorithm π and environment ν such that,

1. the probability of choosing an action $A_t = a$ at time t is dictated only by the algorithm $\pi(a|\mathcal{H}_{t-1})$,
2. the distribution of reward X_t is f_{A_t} and is conditionally independent of the previous observed history $\mathcal{H}_{t-1}$.

Hence, we get for any observed history $\mathcal{H}_T$,

$$\mathbb{P}_{\pi\nu}^T \triangleq \mathbb{P}_{\pi\nu}(\mathcal{H}_T) = \prod_{t=1}^T \pi(A_t|\mathcal{H}_{t-1}) f_{A_t}(X_t). \tag{1}$$

This canonical bandit framework allows us to state Lemma 2 that further leads to Lemma 3 and 4 for local and sequential privacy respectively.

Lemma 2 KL-divergence Decomposition. *Given a bandit algorithm π, two environments ν_1 and ν_2, and a probability measure $\mathbb{P}_{\pi\nu}$ satisfying Equation 1,*

$$D\left(\mathbb{P}^T_{\pi\nu_1} \,\|\, \mathbb{P}^T_{\pi\nu_2}\right) = \sum_{t=1}^{T} D\left(\pi(A_t|\mathcal{H}_t,\nu_1) \,\|\, \pi(A_t|\mathcal{H}_t,\nu_2)\right) + \sum_{a=1}^{K} \mathbb{E}_{\nu_1}\left[N_a(T)\right] D\left(f_a \in \nu_1 \,\|\, f_a \in \nu_2\right).$$

For non-private and locally-private algorithms, the first term vanishes and the rest remains. The corresponding equality for non-private bandit algorithms was first proposed in (Garivier et al., 2018). The non-private decomposition of (Garivier et al., 2018) is also used in (Gajane et al., 2017) to derive a regret lower bound for locally private bandits.

3.1 Lower Bounds for Local Privacy

Lemma 3 Local Private KL-divergence Decomposition. *If the reward generation process is ϵ-local differentially private for both the environments ν_1 and ν_2,*

$$D\left(\mathbb{P}^T_{\nu_1\pi} \,\|\, \mathbb{P}^T_{\nu_2\pi}\right) \leq 2\min\{4, e^{2\epsilon}\}(e^{\epsilon}-1)^2 \sum_{a=1}^{K} \mathbb{E}_{\nu_1}\left[N_a(T)\right] D\left(f_a \in \nu_1 \,\|\, f_a \in \nu_2\right). \quad (2)$$

Theorem 1 Local Private Minimax Regret Bound. *Given an ϵ-locally private reward generation mechanism with $\epsilon \in \mathbb{R}$, and a time horizon $T \geq g(K,\epsilon)$, then for any environment with finite variance, any algorithm π satisfies*

$$\mathrm{Reg}_{\mathrm{minimax}}(T) \geq \frac{c}{\min\{2, e^{\epsilon}\}(e^{\epsilon}-1)}\sqrt{(K-1)T}. \quad (3)$$

For small ϵ, $e^{\epsilon}-1 \approx \epsilon$. Thus, for small ϵ, the minimax regret bound for local privacy worsens by a multiplicative factor $\frac{1}{\epsilon e^{\epsilon}}$. If the $\epsilon = 0$ which means the rewards obtained are completely randomised, the arms would not be separable any more and would lead to unbounded minimax regret.

Following this, we establish a lower bound for Bayesian regret of local private bandits. In the Bayesian setup, the bandit algorithm assumes a prior distribution Q_0 over the possible environments $\nu \in \mathcal{E}$. As the algorithm π plays further and observe corresponding rewards at each time t, it updates the prior over the possible environments to a posterior distribution Q_t. This framework is adopted for efficient algorithms like Thompson sampling (Thompson, 1933) and Gittins indices (Gittins et al., 1989). In the Bayesian setting, we define the Bayesian regret as $\mathrm{Reg}_{\mathrm{Bayes}}(\pi,T,Q) \triangleq \int_{\nu^T} \mathrm{Reg}(\pi,\nu,T)dQ(\nu)$. Bayesian regret is a weaker measure than minimax regret as Bayesian regret is the average regret over believed environments whereas minimax regret is the worst-case regret. Bayesian minimax regret Lattimore and Szepesvári (2018) is defined as the worst possible Bayesian regret for any prior: $\mathrm{Reg}^*_{\mathrm{Bayes}}(T) \triangleq \min_{\pi}\max_{Q}\mathrm{Reg}_{\mathrm{Bayes}}(\pi,T,Q)$.

Corollary 1 Local Private Bayesian Minimax Regret Bound. *Given an ϵ-locally private reward generation mechanism with $\epsilon \in \mathbb{R}$, and a finite time horizon $T \geq g(K,\epsilon)$, then for any environment with bounded rewards $\mathbf{r} \in [0,1]^K$, any algorithm π satisfies*

$$\mathrm{Reg}^*_{\mathrm{Bayes}}(T) \geq \frac{c}{\min\{2, e^{\epsilon}\}(e^{\epsilon}-1)}\sqrt{(K-1)T}. \quad (4)$$

Both minimax and Bayesian minimax regret bounds are problem independent. They represent the worst-case regret for any environment and any prior over environments respectively. Someone may want to design algorithms that is optimal for a given environment ν and the minimax and Bayesian minimax bounds are too pessimistic for them. Thus, researchers study the problem-dependent lower bounds of regret involving environment dependent quantities. Lai and Robbins (1985) proved that a bandit algorithm achieves $\Omega(\log T)$ problem-dependent lower bound. We prove that for ϵ-local privacy this lower bound worsens by a multiplicative factor $\frac{1}{e^{2\epsilon}(e^{\epsilon}-1)^2}$.

Theorem 2 Problem-dependent Local-Private Regret Bound. *For any asymptotically consistent bandit algorithm π, an environment ν with optimal reward distribution f^*, and an ϵ-locally private reward generation mechanism, the expected cumulative regret*

$$\liminf_{T\to\infty} \frac{\mathrm{Reg}(\pi,\nu,T)}{\log T} \geq \sum_{a\neq a^*} \frac{\Delta_a}{2\min\{4, e^{2\epsilon}\}(e^{\epsilon}-1)^2 D\left(f_a \,\|\, f^*\right)}. \quad (5)$$

3.2 Lower Bounds for Sequential Privacy

Lemma 4 Sequential Private KL-divergence Decomposition. *For a sequentially private bandit algorithm π satisfying $l(T) \leq \mathbb{E}_{\pi\nu}[N_a(T)]$ for any arm a, and two environments ν_1 and ν_2,*

$$D\left(\mathbb{P}^T_{\pi\nu_1} \,\|\, \mathbb{P}^T_{\pi\nu_2}\right) \leq 2\epsilon(e^{2\epsilon}-1)\frac{1-2e^{-\frac{T}{l(T)}}}{1-e^{-\frac{T}{l(T)}}} + \sum_{a=1}^{K} \mathbb{E}_{\nu_1}\left[N_a(T)\right]\left(D\left(f_a \in \nu_1 \,\|\, f_a \in \nu_2\right)\right). \quad (6)$$

Theorem 3 Sequential Private Minimax Regret Bound. *Given a finite privacy level $\epsilon \leq a/2$, and a time horizon $T \geq h(K,\epsilon)$, then for any environment with finite variance, any algorithm π that is ϵ-sequential private satisfies*

$$\mathrm{Reg}_{\mathrm{minimax}}(T) \geq c(a)\sqrt{\frac{(K-1)T}{2\epsilon(e^{2\epsilon}-1)}}. \tag{7}$$

This implies that for small ϵ, the minimax regret bound for sequential privacy worsens by a multiplicative factor $\frac{1}{\epsilon}$. Thus, the lower bound of minimax regret for sequentially private bandit is better than the locally private bandit by factors $e^{\epsilon/2}$ and $\sqrt{\frac{2}{\epsilon(e^{\epsilon}+1)}}$ for small and large ϵ's respectively.

Corollary 2 Sequential Private Bayesian Minimax Regret Bound. *Given a finite privacy level $\epsilon \in \mathbb{R}$, and a finite time horizon $T \geq h(K,\epsilon)$, then for any environment with finite variance and bounded reward $\mathbf{r} \in [0,1]^K$, any algorithm π that is ϵ-sequential private satisfies*

$$\mathrm{Reg}^*_{\mathrm{Bayes}}(T) \geq c(\epsilon)\sqrt{\frac{(K-1)T}{2\epsilon(e^{2\epsilon}-1)}}. \tag{8}$$

Discussion. This shows that for both local and sequential privacy for bandits the regret lower bound changes by a multiplicative factor dependent on the privacy level ϵ. The bounds also show that the lower bound for local privacy is worse than that for sequential privacy. This shows learning from randomised reward in local privacy is inherently harder to learn than to use sequential privacy inducing policy. In Section 4, we discuss that our lower bounds falsify the conjecture of additive factor in the lower bound by Tossou and Dimitrakakis (2016) and discovers existing gaps in optimal algorithm design for local and sequential private bandits.

4 Existing Lower Bounds for Non-private and Private Bandits

Problem-independent Non-private Lower Bounds: Minimax regret is the worst case regret that a bandit algorithm can incur if the environment is unknown. Thus, it is often referred as the problem-independent regret. Vogel (1960) performed the first minimax analysis of two-armed Bernoulli bandits. Auer et al. (2002) generalised it to K-arm Bernoulli distributions. Gerchinovitz and Lattimore (2016) provided a novel technique to establish high probability regret lower bounds for adversarial bandits with bounded reward. For any bandit algorithm, the minimax regret is lower bounded by $\mathrm{Reg}_{\mathrm{minimax}}(T) \geq c\sqrt{(K-1)T}$. A bandit algorithm π is called *minimax optimal* if its minimax regret is upper bounded by $C\sqrt{(K-1)T}$. In the Bayesian setup, the bandit algorithm assumes a prior distribution Q_0 over the possible environments $\nu \in \mathcal{E}$. Lattimore and Szepesvári (2018) proved that for any bandit algorithm π there exists a prior distribution Q_0 that the Bayesian regret $\mathrm{Reg}_{\mathrm{Bayes}}(T) \geq C\sqrt{KT}$. This indicates that the minimax regret and the Bayesian regret lower bounds are identical for non-private bandits. This also holds for private bandits. Lattimore and Szepesvári (2019) provides the reasoning behind this connection using a modified minimax theorem.

Existing Lower Bounds for Differentially Private Bandits. Mishra and Thakurta (2015) proposed differentially private variants of UCB and Thompson sampling algorithms. Tossou and Dimitrakakis (2016) improved the differentially private variant of UCB algorithm to obtain regret upper bound of $\Omega(\sum_{a\neq a^*} \frac{\Delta_a}{D(f_a \,\|\, f^*)} \log T + \frac{1}{\epsilon})$ for instantaneous privacy with time varying privacy loss. Our results show that the stricter sequential privacy definition that a constant privacy loss can be achieved with only an additive term on the regret cannot be true. Shariff and Sheffet (2018) proves a finite-time problem-dependent lower bound for contextual bandits. It indicates that the finite-time lower bound of regret is $\Omega(\log T)$ like the non-private lower bounds but with a modified multiplicative factor $(\sum_{a\neq a^*} \frac{\Delta_a}{D(f_a \,\|\, f^*)} + \frac{K}{\epsilon})$. Though their definition of privacy is a bit ambiguous as it can be reduced to either of Definition 4 and 5. Among them, the first being unsuitable for continual observation defeats the purpose for bandits. *We try to clarify at this point with the pan, instantaneous, and sequential privacy definitions.*

Gajane et al. (2017) uses an analogous local privacy definition. They proved a finite-time problem-dependent lower bound of regret for locally private multi-armed bandits where the local privacy is induced by the corrupt bandit mechanism. In this case also, the $\Omega(\log T)$ regret bound of non-private bandits is maintaines with a modified multiplicative factor $\sum_{a\neq a^*} \frac{\Delta_a}{D(g_a \,\|\, g^*)}$. Here, g_a and g^* are the corrupt versions of the reward distributions f_a and f^* ensuring ϵ-local privacy. Theorem 2 shows that the algorithms they have proposed, namely kl-UCB-CF and TS-CF, are suboptimal by at least by a factor $(1 + e^{-\epsilon})^2$. *This opens up the problem of designing optimal local-private bandit algorithms.*

We are not aware of any problem-independent minimax lower bounds for (locally and standard) differentially private bandits, before this paper. Tossou and Dimitrakakis (2017) proposed a differentially private variant of EXP3 algorithm that achieves privacy for adversarial bandits with regret upper bound $O(\frac{\sqrt{T}\log T}{\epsilon})$. *The lower bound of Theorem 3 shows that designing an optimal private algorithm for adversarial bandits is an open problem.*

5 Discussion and Future Work

Our definitions of sequential differential privacy provide a framework to look into differentially private bandit algorithms. This also resolves the eclectic definitions available in the literature and discusses their applicability. The KL-divergence decomposition based method provides a unified proving mechanism for lower bounds for bandit. This allows us to propose the minimax lower bounds for both locally and standard sequentially private multi-armed bandits which was absent in literature. These bounds also pose design of optimal local and sequential private bandit algorithms as open problems since the existing algorithms are suboptimal.

We are now working on deriving the problem-dependent lower bounds for sequentially and environment private bandits and problem-independent bounds for environment private bandits. In future, researchers can utilise these bounds to design optimal private bandit algorithms for real-life applications, such as recommender systems and web advertising.

References

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77.

Bellman, R. (1956). A problem in the sequential design of experiments. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 16(3/4):221–229.

Bretagnolle, J. and Huber, C. (1979). Estimation des densités: risque minimax. *Probability Theory and Related Fields*, 47(2):119–137.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. (2013). Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

Dwork, C., Naor, M., Pitassi, T., Rothblum, G. N., and Yekhanin, S. (2010). Pan-private streaming algorithms. In *ICS*, pages 66–80.

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

Gajane, P., Urvoy, T., and Kaufmann, E. (2017). Corrupt bandits for preserving local privacy. *arXiv preprint arXiv:1708.05033*.

Garivier, A., Ménard, P., and Stoltz, G. (2018). Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*.

Gerchinovitz, S. and Lattimore, T. (2016). Refined lower bounds for adversarial bandits. In *Advances in Neural Information Processing Systems*, pages 1198–1206.

Gittins, J. C., Glazebrook, K. D., Weber, R., and Weber, R. (1989). *Multi-armed bandit allocation indices*, volume 25. Wiley Online Library.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Lattimore, T. and Szepesvári, C. (2018). Bandit algorithms. *preprint*.

Lattimore, T. and Szepesvári, C. (2019). An information-theoretic approach to minimax regret in partial monitoring. *arXiv preprint arXiv:1902.00470*.

Mir, D., Muthukrishnan, S., Nikolov, A., and Wright, R. N. (2010). Pan-private algorithms: When memory does not help. *arXiv preprint arXiv:1009.1544*.

Mishra, N. and Thakurta, A. (2015). (nearly) optimal differentially private stochastic multi-arm bandits. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 592–601. AUAI Press.

Shariff, R. and Sheffet, O. (2018). Differentially private contextual linear bandits. In *Advances in Neural Information Processing Systems*, pages 4296–4306.

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

Tossou, A. C. and Dimitrakakis, C. (2016). Algorithms for differentially private multi-armed bandits. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Tossou, A. C. Y. and Dimitrakakis, C. (2017). Achieving privacy in the adversarial multi-armed bandit. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Vogel, W. (1960). An asymptotic minimax theorem for the two armed bandit problem. *The Annals of Mathematical Statistics*, 31(2):444–451.

A Proofs for Section 2 (Differential Privacy for Bandits)

For simplicity, we write $a^t = \{a_1, \dots, a_t\}$ and $\mathbf{r}^t = \{\mathbf{r}_1, \dots, \mathbf{r}_t\}$ for reward and action sequences respectively.

Lemma 1. *If a policy π satisfies sequential privacy (Definition 6) with privacy level ϵ, π will also satisfy instantaneous privacy (Definition 5) with privacy level 2ϵ. Conversely, if a policy satisfies ϵ instantaneous privacy, it only achieves $t\epsilon$ sequential privacy after t steps.*

Proof of Lemma 1. If π is 2ϵ-instantaneous private then (by definition) the following ratio must be bounded from above by $e^{2\epsilon}$ for any two neighbouring reward sequences $\mathbf{r}^t, \hat{\mathbf{r}}^t$:

$$\frac{\pi(a_t \mid a^{t-1}, \mathbf{r}^{t-1})}{\pi(a_t \mid a^{t-1}, \hat{\mathbf{r}}^{t-1})} = \frac{\pi(a^t \mid \mathbf{r}^{t-1})}{\pi(a^t \mid \hat{\mathbf{r}}^{t-1})} \frac{\pi(a^{t-1} \mid \hat{\mathbf{r}}^{t-1})}{\pi(a^{t-1} \mid \mathbf{r}^{t-1})} = \frac{\pi(a^t \mid \mathbf{r}^{t-1})}{\pi(a^t \mid \hat{\mathbf{r}}^{t-1})} \frac{\pi(a^{t-1} \mid \hat{\mathbf{r}}^{t-2})}{\pi(a^{t-1} \mid \mathbf{r}^{t-2})} \leq e^{2\epsilon},$$

where the first equality is obtained through the definition of conditional probability, the second through independence of actions on current rewards, while the final inequality is through assumption of ϵ-sequential privacy and that $\mathbf{r}, \hat{\mathbf{r}}$ are neighbours. The converse follows from composition of differential privacy. □

B Proofs for Section 3 (Regret Lower Bounds for Private Bandits)

First, let us remind ourselves of the chain rule of KL divergence for two probability measures P, Q on a product space $\mathcal{X}^T$ for a given T:

$$\begin{aligned}
D\left(P \,\|\, Q\right) &\triangleq \int\limits_{\mathcal{X}^T} \ln \frac{\mathrm{d}P(x^T)}{\mathrm{d}Q(x^T)} \,\mathrm{d}P(x^T) \\
&= \int\limits_{\mathcal{X}^T} \ln \frac{\mathrm{d}[P(x_T \mid x^{T-1}) P(x^{T-1})]}{\mathrm{d}[Q(x_T \mid x^{x-1}) Q(x^{T-1})]} \,\mathrm{d}[P(x_T \mid x^{T-1}) P(x^{T-1})] \\
&= \int\limits_{\mathcal{X}^T} \left[\ln \frac{\mathrm{d}P(x_T \mid x^{T-1})}{\mathrm{d}Q(x_T \mid x^{x-1})} + \ln \frac{\mathrm{d}P(x^{T-1})}{\mathrm{d}Q(x^{T-1})}\right] \mathrm{d}[P(x_T \mid x^{T-1}) P(x^{T-1})] \\
&= \int\limits_{\mathcal{X}^T} \ln \frac{\mathrm{d}P(x_T \mid x^{T-1})}{\mathrm{d}Q(x_T \mid x^{x-1})} \,\mathrm{d}P(x^T) + \int\limits_{\mathcal{X}^{T-1}} \ln \frac{\mathrm{d}P(x^{T-1})}{\mathrm{d}Q(x^{T-1})} \,\mathrm{d}P(x^{T-1}) \\
&= \sum_{t=1}^{T} \int\limits_{\mathcal{X}^t} \ln \frac{\mathrm{d}P(x_t \mid x^{t-1})}{\mathrm{d}Q(x_t \mid x^{t-1})} \,\mathrm{d}P(x^t) = \sum_{t=1}^{T} \mathbb{E}_{P(x^{t-1})} \left[D\left(P(x_t \mid x^{t-1}) \,\|\, Q(x_t \mid x^{t-1})\right)\right].
\end{aligned}$$

Here, x_t denotes the reward at time t and x^t denotes the sequence of rewards obtained from the beginning to time t, i.e. $\{x_1, \dots, x_t\}$. Now, the conditional KL-divergence is defined here as

$$D\left(P(x \mid y) \,\|\, Q(x \mid y)\right) \triangleq \int_{\mathcal{X} \times \mathcal{Y}} \ln \frac{\mathrm{d}P(x \mid y)}{\mathrm{d}Q(x \mid y)} \,\mathrm{d}P(x, y).$$

Thus, we get the chain rule of KL-divergence

$$D\left(P \,\|\, Q\right) \triangleq \int\limits_{x^T \in \mathcal{X}^T} \ln \frac{P(x^T)}{Q(x^T)} \,\mathrm{d}P(x^T) = \sum_{t=1}^{T} D\left(P(x_t \mid x^{t-1}) \,\|\, Q(x_t \mid x^{t-1})\right). \tag{9}$$

Lemma 2 KL-divergence Decomposition. *Given a bandit algorithm π, two environments ν_1 and ν_2, and a probability measure $\mathbb{P}_{\pi\nu}$ satisfying Equation 1,*

$$D\left(\mathbb{P}^T_{\pi\nu_1} \,\|\, \mathbb{P}^T_{\pi\nu_2}\right) = \sum_{t=1}^{T} \mathbb{E}_{\pi\nu_1}\left[D\left(\pi(A_t | \mathcal{H}_t, \nu_1) \,\|\, \pi(A_t | \mathcal{H}_t, \nu_2)\right)\right] + \sum_{a=1}^{K} \mathbb{E}_{\pi\nu_1}\left[N_a(T)\right] D\left(f_a \in \nu_1 \,\|\, f_a \in \nu_2\right). \tag{10}$$

This style of KL-divergence decomposition appeared in proofs of Auer et al. (2002); Garivier et al. (2018); Lattimore and Szepesvári (2018). We adopt the proof in our context and notations with enough generality to proof the differentially private versions of it later.

Proof.

$$
\begin{aligned}
&D\left(\mathbb{P}^T_{\pi\nu_1} \,\|\, \mathbb{P}^T_{\pi\nu_2}\right)\\
=&\sum_{t=1}^{T} \mathbb{E}_{\mathbb{P}^T_{\pi\nu_1}}\left[D\left(\pi(A_t|\mathcal{H}_t,\nu_1) \,\|\, \pi(A_t|\mathcal{H}_t,\nu_2)\right) + D\left(f(X_t|A_t,\mathcal{H}_t,\nu_1) \,\|\, f(X_t|A_t,\mathcal{H}_t,\nu_2)\right)\right]\\
=&\sum_{t=1}^{T} \mathbb{E}_{\mathbb{P}^T_{\pi\nu_1}}\left[D\left(\pi(A_t|\mathcal{H}_t,\nu_1) \,\|\, \pi(A_t|\mathcal{H}_t,\nu_2)\right)\right] + \sum_{t=1}^{T} \mathbb{E}_{\mathbb{P}^T_{\pi\nu_1}}\left[\sum_{a=1}^{K} \mathbb{1}_{A_t=a} D\left(f_a(X_t)\in\nu_1 \,\|\, f_a(X_t)\in\nu_2\right)\right]\\
=&\sum_{t=1}^{T} \mathbb{E}_{\mathbb{P}^T_{\pi\nu_1}}\left[D\left(\pi(A_t|\mathcal{H}_t,\nu_1) \,\|\, \pi(A_t|\mathcal{H}_t,\nu_2)\right)\right] + \sum_{a=1}^{K}\left[\sum_{t=1}^{T} \mathbb{E}_{\mathbb{P}^T_{\pi\nu_1}}[\mathbb{1}_{A_t=a}] D\left(f_a(X_t)\in\nu_1 \,\|\, f_a(X_t)\in\nu_2\right)\right]\\
=&\sum_{t=1}^{T} \mathbb{E}_{\mathbb{P}^T_{\pi\nu_1}}\left[D\left(\pi(A_t|\mathcal{H}_t,\nu_1) \,\|\, \pi(A_t|\mathcal{H}_t,\nu_2)\right)\right] + \sum_{a=1}^{K} \mathbb{E}_{\mathbb{P}^T_{\pi\nu_1}}\left[N_a(T)\right] D\left(f_a\in\nu_1 \,\|\, f_a\in\nu_2\right).
\end{aligned}
$$

The first equality is followed by the chain rule of KL-divergence and Equation 1. The second equality is from the conditioning. The third equality is obtained from linearity of expectation. The fourth equality is from the fact that expectation of an indicator function of an event returns its probability of occurrence. □

B.1 Proofs for Local Privacy

Lemma 3 Locally Private KL-divergence Decomposition. *If the reward generation process is ϵ-local differentially private for both the environments ν_1 and ν_2,*

$$
D\left(\mathbb{P}^T_{\nu_1\pi} \,\|\, \mathbb{P}^T_{\nu_2\pi}\right) \le 2\min\{4, e^{2\epsilon}\}(e^{\epsilon}-1)^2 \sum_{a=1}^{K} \mathbb{E}_{\nu_1}\left[N_a(T)\right] D\left(f_a\in\nu_1 \,\|\, f_a\in\nu_2\right). \tag{11}
$$

Proof. In case of local privacy, the reward observed by the algorithm π is obtained at time t, $\mathcal{X}_t$, from a privatised version of generated rewards $\mathbf{Z}_t$. Thus, $x_t = z_{t,a}$, where a refers to the action selected at time t. We denote the distribution over the privatised generated reward of arm a as $g_a(z)$. $g_a(z)$ is obtained by imposing a local privacy mechanism $\mathcal{M}$ on the original reward distribution $f_a(z)$.

We note that the KL-divergence decomposition of Lemma 2 is obtained on the probability space over observed histories. Since the observed rewards are now coming from $g_a(z)$ rather than $f_a(x)$, we begin our derivations of local-private KL-divergence decomposition by substituting $g_a(z)$ in Equation 10. For brevity, we denote g^1_a and g^2_a to represent the privatised reward distributions for arm a in environments ν_1 and ν_2 respectively. Thus,

$$
\begin{aligned}
D\left(\mathbb{P}^T_{\pi\nu_1} \,\|\, \mathbb{P}^T_{\pi\nu_2}\right) =&\sum_{t=1}^{T} \mathbb{E}_{\pi\nu_1}\left[D\left(\pi(A_t|\mathcal{H}_t,\nu_1) \,\|\, \pi(A_t|\mathcal{H}_t,\nu_2)\right)\right] + \sum_{a=1}^{K} \mathbb{E}_{\pi\nu_1}\left[N_a(T)\right] D\left(g^1_a(Z) \,\|\, g^2_a(Z)\right)\\
=&\sum_{a=1}^{K} \mathbb{E}_{\pi\nu_1}\left[N_a(T)\right] D\left(g^1_a(Z) \,\|\, g^2_a(Z)\right)\\
\le&\sum_{a=1}^{K} \mathbb{E}_{\pi\nu_1}\left[N_a(T)\right] \left[D\left(g^1_a(Z) \,\|\, g^2_a(Z)\right) + D\left(g^2_a(Z) \,\|\, g^1_a(Z)\right)\right]\\
\le&\min\{4, e^{2\epsilon}\}(e^{\epsilon}-1)^2 \sum_{a=1}^{K} \mathbb{E}_{\nu_1}\left[N_a(T)\right] \,\| f^1_a(X) - f^2_a(X) \|^2_{TV}\\
\le&2\min\{4, e^{2\epsilon}\}(e^{\epsilon}-1)^2 \sum_{a=1}^{K} \mathbb{E}_{\nu_1}\left[N_a(T)\right] D\left(f^1_a(X) \,\|\, f^2_a(X)\right)
\end{aligned}
$$

The first step is due to the fact that given the same history, $\pi(A_t|\mathcal{H}_t,\nu_1) = \pi(A_t|\mathcal{H}_t,\nu_2)$ as they do not vary with the model and depends only on the internal randomisation of the algorithm π.

The inequality in the second step is derived from non-negativity of KL-divergence (Cover and Thomas, 2012) and the fact that for two non-negative numbers a and b, $a \leq a + b$. The inequality in the third step is obtained from Theorem 1 in (Duchi et al., 2013). The last inequality is obtained by applying Pinsker's inequality (Cover and Thomas, 2012). Pinsker's inequality states that for any two distributions P and Q, square of their total variance distance is upper bounded by 2 times their Kl-divergence: $\| P - Q \|_{TV}^2 \leq 2D\left(P \,\|\, Q\right)$. □

Theorem 1 Locally Private Minimax Regret Bound. *Given an ϵ-locally private reward generation mechanism with $\epsilon \in \mathbb{R}$, and a time horizon $T \geq g(K,\epsilon)$, then for any environment with finite variance, any algorithm π satisfies*

$$\mathrm{Reg}_{\mathrm{minimax}}(T) \geq \frac{c}{\min\{2, e^{\epsilon}\}(e^{\epsilon}-1)}\sqrt{(K-1)T}. \tag{12}$$

Proof. Step 1: Fix two environments ν_1 and ν_2 such that drawing arm 1 in ν_1 for more than $T/2$ times is good but doing the same is bad for ν_2.

We define ν_1 to be a set of K-reward distribution with mean reward $\{\Delta, 0, \ldots, 0\}$ and finite Fisher information I. Similarly, we define ν_2 to be to be a set of K-reward distribution with mean reward $\{\Delta, \ldots, 0, 2\Delta\}$ and finite Fisher information I. Drawing arm 1 is the optimal choice in ν_1 whereas drawing arm K is the optimal choice in ν_2.

Step 2: We get the lower bounds of expected cumulative regret for a policy π, and the environments ν_1 and ν_2 as follows:

$$\begin{aligned}\mathrm{Reg}(\pi,\nu_1,T) &\geq \mathbb{P}_{\pi\nu_1}\left(N_1(T) \leq T/2\right)\frac{T\Delta}{2},\\ \mathrm{Reg}(\pi,\nu_2,T) &> \mathbb{P}_{\pi\nu_2}\left(N_1(T) > T/2\right)\frac{T\Delta}{2}.\end{aligned}$$

Let us denote the event $N_1(T) \leq T/2$ as E. Hence, we get

$$\begin{aligned}\mathrm{Reg}(\pi,\nu_1,T) + \mathrm{Reg}(\pi,\nu_2,T) &\geq \frac{T\Delta}{2}\left(\mathbb{P}_{\pi\nu_1}(E) + \mathbb{P}_{\pi\nu_2}(E^C)\right)\\ &\geq \frac{T\Delta}{4}\exp(-D\left(\mathbb{P}_{\pi\nu_1} \,\|\, \mathbb{P}_{\pi\nu_2}\right)).\end{aligned}$$

We obtain the last inequality from the Lemma 2.1 in (Bretagnolle and Huber, 1979). This is called Bretagnolle-Huber inequality or probabilistic Pinsker's inequality (Auer et al., 2002; Lattimore and Szepesvári, 2019) and used for several proofs of bandit algorithms. This states that for any two distributions P and Q defined on the same measurable space and an event E, $P(E) + Q(E^C) \geq \exp(-D\left(P \,\|\, Q\right))$.

Step 3: From Lemma 3, we get

$$\begin{aligned}D\left(\mathbb{P}_{\pi\nu_1} \,\|\, \mathbb{P}_{\pi\nu_2}\right) &= \mathbb{E}_{\pi\nu_1}[N_K(T)]D\left(\mathcal{M}(f_K(0,I)) \,\|\, \mathcal{M}(f_K(2\Delta,I))\right)\\ &\leq 2\min\{4, e^{2\epsilon}\}(e^{\epsilon}-1)^2\mathbb{E}_{\pi\nu_1}\left[N_K(T)\right]D\left(f_K(0,I) \,\|\, f_K(2\Delta,I)\right)\\ &\leq 2\min\{4, e^{2\epsilon}\}(e^{\epsilon}-1)^2\mathbb{E}_{\pi\nu_1}\left[N_K(T)\right](2\Delta^2)\\ &\leq 2\min\{4, e^{2\epsilon}\}(e^{\epsilon}-1)^2\frac{2T\Delta^2}{K-1}.\end{aligned}$$

Hence, we get the regret bound to be

$$\begin{aligned}\max\{\mathrm{Reg}(\pi,\nu_1,T), \mathrm{Reg}(\pi,\nu_2,T)\} &\geq \frac{1}{2}\left(\mathrm{Reg}(\pi,\nu_1,T) + \mathrm{Reg}(\pi,\nu_2,T)\right)\\ &\geq \frac{T\Delta}{4}\exp\left[-\min\{2, e^{\epsilon}\}(e^{\epsilon}-1)^2\frac{2T\Delta^2}{K-1}\right].\end{aligned}$$

Step 4: Let us choose the optimality gap Δ to be $\sqrt{\frac{(K-1)}{\min\{4,e^{2\epsilon}\}(e^{\epsilon}-1)^2T}} \leq \frac{1}{2}$. This holds for any for $T \geq \frac{(K-1)}{\min\{4,e^{2\epsilon}\}(e^{\epsilon}-1)^2} \triangleq g(K,\epsilon)$. Hence, by using the results of Step 3, we obtain:

$$\mathrm{Reg}_{\mathrm{minimax}}(T) \geq \frac{1}{24}\sqrt{\frac{(K-1)T}{\min\{4,e^{2\epsilon}\}(e^{\epsilon}-1)^2}}.$$

□

Corollary 1 Locally Private Bayesian Minimax Regret Bound. *Given an ϵ-locally private reward generation mechanism with $\epsilon \in \mathbb{R}$, and a finite time horizon $T \geq g(K,\epsilon)$, then for any environment with bounded rewards $\mathbf{r} \in [0,1]^K$, any algorithm π satisfies*

$$\mathrm{Reg}^*_{\mathrm{Bayes}}(T) \geq \frac{c}{\min\{2,e^{\epsilon}\}(e^{\epsilon}-1)}\sqrt{(K-1)T}. \tag{13}$$

Proof. Let us denote the space of all plausible priors for a given bandit problem to be $\mathcal{L} \triangleq \{\mu\}$. Lattimore and Szepesvári (2019) proves in their recent paper that if $\mathcal{L}$ is the space of all finitely supported probability measures on $[0,1]^{KT}$, the Bayesian regret and the minimax regret would be the same.

Fact 1 Theorem 1 in (Lattimore and Szepesvári, 2019). *Let $\mathcal{L}$ be the space of all finitely supported probability measures on $\mathcal{R}^T$, where $\mathcal{R} \triangleq [0,1]^K$. Then*

$$\mathrm{Reg}_{\mathrm{minimax}}(T) = \mathrm{Reg}^*_{\mathrm{Bayes}}(T).$$

Since variance of bounded random variable in $[0,1]$ is less than $\frac{1}{4}$, the bounded reward assumption satisfies the finite variance requirement of Theorems 1. Thus, the results of Theorems 1 and 1 prove the statement of Corollary 1 for bounded rewards. □

Theorem 2 Problem-dependent Local-Private Regret Bound. *For any asymptotically consistent bandit algorithm π, an environment ν with optimal reward distribution f^*, and an ϵ-locally private reward generation mechanism, the expected cumulative regret*

$$\liminf_{T\to\infty} \frac{\mathrm{Reg}(\pi,\nu,T)}{\log T} \geq \sum_{a\neq a^*} \frac{\Delta_a}{2\min\{4,e^{2\epsilon}\}(e^{\epsilon}-1)^2 D\left(f_a \parallel f^*\right)}.$$

Step 1 and Step 2 of this proof are similar in essence as that of Theorem 1. Step 3 differs from the fact that rather than substituting the KL-divergence and the expected number of draws using the problem-independent terms, we keep the problem dependent terms. This leads to a problem-dependent bound in Step 4.

Proof. Step 1: Fix two environments ν_1 and ν_2 such that ν_1 contains K reward distributions $\{f_1,\ldots,f_K\}$ and ν_2 contains $K-1$ same reward distributions while the reward distribution i-th arm f_i is replaced by f_i' such that $D\left(f_i \parallel f_i'\right) \leq D\left(f_i \parallel f^*\right) + \delta$ for some $\delta > 0$. Here, f^* represents the privatised reward distribution obtained for the optimal arm a^*. After imposing the ϵ-local private mechanism, we obtain privatised reward distribution $\{g_1,\ldots,g_i,\ldots,g_K\}$ and $\{g_1,\ldots,g_i',\ldots,g_K\}$. Let us denote the expected privatised rewards corresponding to the distributions as $\{\mu_1,\ldots,\mu_K\}$ and μ_i'.

Step 2: We get the lower bounds of expected cumulative regret for a policy π, and the environments ν_1 and ν_2 as follows:

$$\mathrm{Reg}(\pi,\nu_1,T) \geq \mathbb{P}_{\pi\nu_1}\left(N_1(T) \leq T/2\right)\frac{T}{2}(\mu_i - \mu^*),$$
$$\mathrm{Reg}(\pi,\nu_2,T) > \mathbb{P}_{\pi\nu_2}\left(N_1(T) > T/2\right)\frac{T}{2}(\mu_i' - \mu^*).$$

Let us denote the event $N_1(T) \leq T/2$ as E. Hence, we get

$$\begin{aligned}\mathrm{Reg}(\pi,\nu_1,T)+\mathrm{Reg}(\pi,\nu_2,T) &\geq \frac{T}{2}\left(\mathbb{P}_{\pi\nu_1}(E)(\mu_i-\mu^*)+\mathbb{P}_{\pi\nu_2}(E^C)(\mu_i'-\mu^*)\right)\\ &\geq \frac{T}{2}\left(\mathbb{P}_{\pi\nu_1}(E)+\mathbb{P}_{\pi\nu_2}(E^C)\right)\min\{(\mu_i-\mu^*),(\mu_i'-\mu^*)\}\\ &\underset{(a)}{\geq} \frac{T}{4}\exp(-D\left(\mathbb{P}_{\pi\nu_1}\,\|\,\mathbb{P}_{\pi\nu_2}\right))\min\{(\mu_i-\mu^*),(\mu_i'-\mu^*)\}.\end{aligned}$$

We obtain the inequality (a) from the Lemma 2.1 in (Bretagnolle and Huber, 1979) as mentioned in the proof of Theorem 1.

Step 3: From Lemma 3, we get

$$\begin{aligned}D\left(\mathbb{P}_{\pi\nu_1}\,\|\,\mathbb{P}_{\pi\nu_2}\right) &= \mathbb{E}_{\pi\nu_1}[N_i(T)]D\left(g_i\,\|\,g_i'\right)\\ &\leq 2\min\{4,e^{2\epsilon}\}(e^\epsilon-1)^2\mathbb{E}_{\pi\nu_1}\left[N_i(T)\right]D\left(f_i\,\|\,f_i'\right)\\ &\leq 2\min\{4,e^{2\epsilon}\}(e^\epsilon-1)^2\mathbb{E}_{\pi\nu_1}\left[N_i(T)\right]\left(D\left(f_i\,\|\,f^*\right)+\delta\right)\end{aligned}$$

Hence, we get the regret bound to be

$$\begin{aligned}&\mathrm{Reg}(\pi,\nu_1,T)+\mathrm{Reg}(\pi,\nu_2,T)\\ &\geq \frac{T}{4}\min\{(\mu_i-\mu^*),(\mu_i'-\mu^*)\}\exp\left[-2\min\{4,e^{2\epsilon}\}(e^\epsilon-1)^2\mathbb{E}_{\pi\nu_1}\left[N_i(T)\right]\left(D\left(f_i\,\|\,f^*\right)+\delta\right)\right].\end{aligned}$$

Taking logarithm on both sides and simplifying, we get

$$\begin{aligned}\frac{\mathbb{E}_{\pi\nu_1}\left[N_i(T)\right]}{\log T} &\geq \frac{1}{2\min\{4,e^{2\epsilon}\}(e^\epsilon-1)^2(D\left(f_i\,\|\,f^*\right)+\delta)}\frac{\log\left(\frac{T\min\{(\mu_i-\mu^*),(\mu_i'-\mu^*)\}}{4(\mathrm{Reg}(\pi,\nu_1,T)+\mathrm{Reg}(\pi,\nu_2,T))}\right)}{\log T}\\ &\geq \frac{1+\frac{\log\left(0.25\min\{(\mu_i-\mu^*),(\mu_i'-\mu^*)\}\right)}{\log T}-\frac{\log(\mathrm{Reg}(\pi,\nu_1,T)+\mathrm{Reg}(\pi,\nu_2,T))}{\log T}}{2\min\{4,e^{2\epsilon}\}(e^\epsilon-1)^2(D\left(f_i\,\|\,f^*\right)+\delta)}.\end{aligned}$$

Step 4: We obtain the asymptotic lower bound by taking limit inferior on both sides

$$\begin{aligned}\liminf_{T\to\infty}\frac{\mathbb{E}_{\pi\nu_1}\left[N_i(T)\right]}{\log T} &\geq \liminf_{T\to\infty}\frac{1+\frac{\log\left(0.25\min\{(\mu_i-\mu^*),(\mu_i'-\mu^*)\}\right)}{\log T}-\frac{\log(\mathrm{Reg}(\pi,\nu_1,T)+\mathrm{Reg}(\pi,\nu_2,T))}{\log T}}{2\min\{4,e^{2\epsilon}\}(e^\epsilon-1)^2(D\left(f_i\,\|\,f^*\right)+\delta)}\\ &= \frac{1+\liminf\limits_{T\to\infty}\frac{\log\left(0.25\min\{(\mu_i-\mu^*),(\mu_i'-\mu^*)\}\right)}{\log T}-\limsup\limits_{T\to\infty}\frac{\log(\mathrm{Reg}(\pi,\nu_1,T)+\mathrm{Reg}(\pi,\nu_2,T))}{\log T}}{2\min\{4,e^{2\epsilon}\}(e^\epsilon-1)^2(D\left(f_i\,\|\,f^*\right)+\delta)}\\ &\underset{(b)}{\geq} \frac{1-p}{2\min\{4,e^{2\epsilon}\}(e^\epsilon-1)^2(D\left(f_i\,\|\,f^*\right)+\delta)}\\ &\underset{(c)}{\geq} \frac{1}{2\min\{4,e^{2\epsilon}\}(e^\epsilon-1)^2D\left(f_i\,\|\,f^*\right)}.\end{aligned}$$

We obtain inequality (b) because:

1. The first limit contains a constant in the numerator. Thus the limit goes to 0 as $T\to\infty$.

2. In order to obtain the other limit. We use the asymptotic consistency property of π. Since π is assumed to be asymptotically consistent, $\mathrm{Reg}(\pi,\nu_1,T)+\mathrm{Reg}(\pi,\nu_2,T)\leq CT^p$ for some constant $p\in(0,1)$ and large enough time horizon T. Thus, $\limsup\limits_{T\to\infty}\frac{\log(\mathrm{Reg}(\pi,\nu_1,T)+\mathrm{Reg}(\pi,\nu_2,T))}{\log T}\leq \limsup\limits_{T\to\infty}\frac{p\log T+\log C}{\log T}=p$.

We obtain the inequality (c) from the fact that $p>0$ and $\delta>0$.

Step 5: Using the definition of regret and the resulting inequality of Step 4, we obtain

$$\begin{aligned}\liminf_{T\to\infty}\frac{\mathrm{Reg}(\pi,\nu_1,T)}{\log T} &= \liminf_{T\to\infty}\sum_{a\neq a^*}\frac{\mathbb{E}_{\pi\nu_1}\left[N_a(T)\right](\mu_a-\mu^*)}{\log T}\\ &\geq \sum_{a\neq a^*}\frac{(\mu_a-\mu^*)}{2\min\{4,e^{2\epsilon}\}(e^{\epsilon}-1)^2 D\left(f_a \,\|\, f^*\right)}\\ &= \sum_{a\neq a^*}\frac{\Delta_a}{2\min\{4,e^{2\epsilon}\}(e^{\epsilon}-1)^2 D\left(f_a \,\|\, f^*\right)}.\end{aligned}$$

□

These results establish that for both minimax and Bayesian minimax regret the lower bounds degrade by a multiplicative factor $\frac{1}{\min\{2,e^{\epsilon}\}(e^{\epsilon}-1)}$ whereas for problem-dependent lower bound degrades by a multiplicative factor $\frac{1}{2\min\{4,e^{2\epsilon}\}(e^{\epsilon}-1)^2}$.

B.2 Proofs for Sequential Privacy

Lemma 4 Sequential Private KL-divergence Decomposition. *For a sequentially private bandit algorithm π satisfying $l(T)\leq\mathbb{E}_{\pi\nu}[N_a(T)]$ for any arm a, and two environments ν_1 and ν_2,*

$$D\left(\mathbb{P}^T_{\pi\nu_1}\,\|\,\mathbb{P}^T_{\pi\nu_2}\right)\leq 2\epsilon(e^{2\epsilon}-1)\frac{1-2e^{-\frac{T}{l(T)}}}{1-e^{-\frac{T}{l(T)}}}+\sum_{a=1}^{K}\mathbb{E}_{\nu_1}\left[N_a(T)\right]\left(D\left(f_a\in\nu_1\,\|\,f_a\in\nu_2\right)\right).$$

Proof Sketch. The second term of Lemma 2 remains the same. Whereas the first term is bounded as follows:

$$\begin{aligned}\sum_{t=1}^{T}D\left(\pi(A_t|\mathcal{H}^t,\nu_1)\,\|\,\pi(A_t|\mathcal{H}^t,\nu_2)\right) &\leq \sum_{t=1}^{T}\max_{A_t,\mathcal{H}_t}\left|\pi(A_t|\mathcal{H}^t,\nu_1)\right|\left|\log\frac{\pi(A_t|\mathcal{H}^t,\nu_1)}{\pi(A_t|\mathcal{H}^t,\nu_2)}\right|\\ &\leq 2\epsilon(e^{2\epsilon}-1)\max_{A_t,\mathcal{H}_t}\sum_{t=1}^{T}\left|\pi(A_t|\mathcal{H}^t,\nu_1)\right|\\ &\leq 2\epsilon(e^{2\epsilon}-1)\frac{1-2e^{-\frac{T}{l(T)}}}{1-e^{-\frac{T}{l(T)}}}.\end{aligned}$$

for $\mathbb{E}[N_a(T)]\geq l(T)$ for all arms a. □

Theorem 3 Sequential Private Minimax Regret Bound. *Given a finite privacy level $\epsilon\leq a/2$, and a time horizon $T\geq h(K,\epsilon)$, then for any environment with finite variance, any algorithm π that is ϵ-sequential private satisfies*

$$\mathrm{Reg}_{\mathrm{minimax}}(T)\geq c(a)\sqrt{\frac{(K-1)T}{2\epsilon(e^{2\epsilon}-1)}}.$$

Proof Sketch. Repeat the Steps 1 and 2 described in the proof sketch of Theorem 1.

Step 3: Use Lemma 4 for KL-divergence decomposition under sequential privacy to obtain

$$D\left(\mathbb{P}_{\pi\nu_1}\,\|\,\mathbb{P}_{\pi\nu_2}\right)\leq 2\epsilon(e^{2\epsilon}-1)+\frac{2T\Delta^2}{K-1}.$$

Hence, we get the regret bound to be

$$\begin{aligned}\max\{\mathrm{Reg}_{\nu_1}(\pi,T),\mathrm{Reg}_{\nu_2}(\pi,T)\} &\geq \frac{1}{2}\left(\mathrm{Reg}_{\nu_1}(\pi,T)+\mathrm{Reg}_{\nu_2}(\pi,T)\right)\\ &\geq \frac{T\Delta}{4}\exp\left[-2\epsilon(e^{2\epsilon}-1)-\frac{2T\Delta^2}{K-1}\right].\end{aligned}$$

Step 4: Let us choose the optimality gap Δ to be

$$\Delta = \sqrt{\frac{(K-1)C(\epsilon)}{4\epsilon(e^{2\epsilon}-1)T}} \leq \frac{1}{2}.$$

Here, we choose

$$C(\epsilon) = -2\epsilon(e^{2\epsilon}-1)W\left(-\frac{e^{-\delta(a)+2\epsilon(e^{2\epsilon}-1)}}{2\epsilon(e^{2\epsilon}-1)}\right),$$

where W is the Lambert function or the product-log function and $\delta(a)$ is a function of a such that $2\epsilon(e^{2\epsilon}-1) \leq \delta(a)$ for $\epsilon \leq a$. In the given range of ϵ, $C(\epsilon) \leq 1$.

Δ being less than $\frac{1}{2}$ holds for any for $T \geq \frac{2(K-1)C(\epsilon)}{4\epsilon(e^{2\epsilon}-1)} \triangleq h(K,\epsilon)$. Hence, by combining the results of Step 3 and the upper bound on privacy level $\epsilon \leq a$ and $T \geq h(K,\epsilon)$, we obtain:

$$\mathrm{Reg}_{minimax}(T) \geq \frac{e^{\delta(a)}}{4\sqrt{2}}\sqrt{\frac{(K-1)T}{\min\{2, e^{\epsilon}\}(e^{2\epsilon}-1)}}. \tag{14}$$

□

Corollary 2 Sequential Private Bayesian Minimax Regret Bound. *Given a finite privacy level $\epsilon \in \mathbb{R}$, and a finite time horizon T, then for any environment with bounded reward $\mathbf{r} \in [0,1]^K$, any algorithm π that is ϵ-sequential private satisfies*

$$\mathrm{Reg}^*_{\mathrm{Bayes}}(T) \geq c(\epsilon)\sqrt{\frac{(K-1)T}{2\epsilon(e^{2\epsilon}-1)}}.$$

Proof. Similar to the proof of Corollary 1, the results of Theorem 3 and Fact 1 prove the statement of Corollary 2. □