DP-NCB: Privacy Preserving Fair Bandits

Dhruv Sarkar[1], **Nishant Pandey**[2], **Sayak Ray Chowdhury**[2]

[1] Indian Institute of Technology Kharagpur
[2] Indian Institute of Technology Kanpur
dhruv.sarkar@kgpian.iitkgp.ac.in, nishantp22@iitk.ac.in, sayakrc@iitk.ac.in

Abstract

Multi-armed bandit algorithms are fundamental tools for sequential decision-making under uncertainty, with widespread applications across domains such as clinical trials and personalized decision-making. As bandit algorithms are increasingly deployed in these socially sensitive settings, it becomes critical to protect user data privacy and ensure fair treatment across decision rounds. While prior work has independently addressed privacy and fairness in bandit settings, the question of whether both objectives can be achieved simultaneously has remained largely open. Existing privacy-preserving bandit algorithms typically optimize average regret, a utilitarian measure, whereas fairness-aware approaches focus on minimizing Nash regret, which penalizes inequitable reward distributions, but often disregard privacy concerns.

To bridge this gap, we introduce Differentially Private Nash Confidence Bound (DP-NCB)—a novel and unified algorithmic framework that simultaneously ensures ϵ-differential privacy and achieves order-optimal Nash regret, matching known lower bounds up to logarithmic factors. The framework is sufficiently general to operate under both global and local differential privacy models, and is anytime, requiring no prior knowledge of the time horizon. We support our theoretical guarantees with simulations on synthetic bandit instances, showing that DP-NCB incurs substantially lower Nash regret than state-of-the-art baselines. Our results offer a principled foundation for designing bandit algorithms that are both privacy-preserving and fair, making them suitable for high-stakes, socially impactful applications.

Code — https://github.com/NP-Hardest/DP-NCB

Introduction

The multi-armed bandit framework (Bubeck, Cesa-Bianchi et al. 2012; Lattimore and Szepesv´ari 2020) is a foundational model for sequential decision-making under uncertainty. It has been widely applied across domains such as online advertising and product recommendations (Li et al. 2010; Schwartz, Bradlow, and Fader 2017), education and tutoring systems (Clement et al. 2015), and healthcare and clinical trials (Tewari and Murphy 2017; Villar, Bowden, and Wason 2015). Given its growing use in socially sensitive areas, there is increasing interest in integrating principles of privacy and fairness into bandit algorithms. Privacy concerns arise from the reliance on user feedback, which can inadvertently reveal sensitive information (Pan et al. 2019). Fairness is motivated by social welfare considerations — ensuring that all users perceive the system as treating them equitably (Moulin 2004).

To highlight the importance of privacy and fairness in bandit problems, consider a clinical trial scenario originally proposed by Thompson (1933): given k drugs and T patients, the decision maker selects one drug to administer to the t-th patient in each round $t \leq T$. The treatment outcome is inherently private and sensitive i.e. patients may not wish to disclose their medical conditions publicly post-trial. However, since future drug assignments are based on prior outcomes, an individual patient's response can influence subsequent decisions. This means that even if outcomes are not explicitly revealed, they may still be inferred from changes in the drug selection policy. Therefore, the learning algorithm must ensure that individual patient data (e.g., treatment outcomes) remains private while taking future decisions (e.g., drug choices). At the same time, it is crucial to ensure that no patient is ex-ante disadvantaged by the learner, i.e., fairness must be maintained across rounds. A fair learning algorithm should guarantee that patients participating in different rounds of the trial benefit from progressively better drug choices, while still permitting sufficient exploration to accurately identify the most effective drug.

Achieving privacy in isolation is straightforward, for example, the learner could ignore treatment outcomes entirely and always administer a fixed, publicly announced drug. However, such an approach would likely lead to poor treatment efficacy. This highlights the need to carefully balance privacy with utility. Differential Privacy (DP) offers a principled way to do so by providing strong guarantees even against adversaries with access to arbitrary auxiliary information. It achieves this by adding calibrated random noise to obscure any individual's influence on the algorithm's output, ensuring that no single user's data significantly alters the outcome (Dwork and Roth 2014). Importantly, DP includes a tunable parameter $\epsilon > 0$ that allows explicit control over the privacy-utility trade-off: smaller ϵ yields stronger privacy but reduced utility, and vice versa. In recent years, DP has been effectively incorporated into the design of privacy-preserving bandit algorithms (Mishra and Thakurta 2015; Tossou and Dimitrakakis 2016; Azize and Basu 2022), typ-

ically using average reward (e.g., treatment outcome) or, equivalently, regret, as the measure of utility.

Minimizing **average regret**, which is defined as the difference between the (a priori unknown) highest expected reward and the *arithmetic mean* of the expected rewards obtained by the algorithm—is equivalent to maximizing social welfare in a utilitarian sense (Moulin 2004). While this objective promotes overall efficiency, it may inadvertently favor short-term reward maximization, potentially resulting in inequities across decision rounds. For example, achieving a high average treatment efficacy does not preclude some patients from receiving highly ineffective treatments. To address this limitation, Nash social welfare (NSW) considers the *geometric mean* of expected rewards, promoting a more equitable distribution across rounds in an ex-ante sense (Moulin 2004). The associated metric, **Nash regret**—the gap between the highest expected reward and the NSW offers a fairness-aware alternative to average regret. It penalizes disparities in individual rewards and has gained recent attention as a foundation for designing fair bandit algorithms (Barman et al. 2023; Sawarni, Pal, and Barman 2023; Krishna et al. 2025).

While prior work has extensively explored privacy and fairness in multi-armed bandits, these concerns have been addressed in isolation. The fundamental question of whether these concerns can be simultaneously addressed in bandit settings remains largely unresolved. In this work, we address this gap by introducing a novel differentially private Nash regret minimization framework that combines controlled noise addition of DP mechanisms with fairness-aware reward aggregation. This ensures a learning process that is both private and fair, making it well-suited for high stakes, sensitive applications like healthcare and personalized decision-making.

Our Contributions

- We propose a general framework — Differentially Private Nash Confidence Bound (DP-NCB) — that guarantees ϵ-differential privacy for any ϵ, while achieving order-optimal Nash regret that vanishes as the number of rounds $T \to \infty$.
- In the global model of differential privacy, where the learner is trusted and has access to raw user data, Algorithm 1 (GDP-NCB) achieves a Nash regret[1] of $\widetilde{O}\Big(\sqrt{\frac{k}{T}} + \frac{k}{\epsilon T}\Big)$. In the stronger local privacy model, where the learner is not trusted and has access to only randomized user data, Algorithm 2 (LDP-NCB) achieves the Nash regret $\widetilde{O}\Big(\sqrt{\frac{k}{T}} + \frac{1}{\epsilon}\sqrt{\frac{k}{T}}\Big)$.
- Since Nash regret is a stricter benchmark than average regret (see Remark 1), existing lower bounds for average regret also apply to Nash regret. In particular, the known minimax lower bounds—$\Omega\Big(\max\big\{\sqrt{\frac{k}{T}}, \frac{k}{\epsilon T}\big\}\Big)$ in the global privacy model and $\Omega\Big(\frac{1}{\epsilon}\sqrt{\frac{k}{T}}\Big)$ in the local model—serve as fundamental performance limits. This implies that our upper bound on Nash regret is nearly optimal (up to logarithmic factors), as it matches the best-known lower bounds even for the easier benchmark of average regret.
- We further extend our algorithms to the anytime setting, where prior knowledge of the time horizon T is not required. We show that this generalization incurs only an $O(\log T)$ multiplicative increase in Nash regret in both the global and local privacy models.
- To empirically validate fairness across rounds, we evaluate the Nash regret of our algorithms under both the global and local privacy models. Our results show that the proposed algorithms incur significantly lower Nash regret compared to state-of-the-art private bandit algorithms designed for average regret, thereby supporting our theoretical claims.

[1] $\widetilde{O}$ hides constant and poly-logarithmic factors in T and k.

Preliminaries

Multi-armed Bandit (MAB) In the stochastic multi-armed bandit problem, the learner has sample access to k probability distributions (called arms) supported on the interval $[0, 1]$. Each arm $i \in [k] := \{1, \ldots, k\}$ has its mean $\mu_i \in [0, 1]$ a-priori unknown to the learner. Let $\mu^* = \max_{i\in[k]} \mu_i$ denote the highest mean. We consider settings where the learning algorithm operates over a population of $T \geq 1$ unique users, one for each round. At each round $t \in [T] := \{1, 2, \ldots T\}$, the algorithms selects an arm $I_t \in [k]$ and observes a random reward X_t drawn independently from the distribution with mean μ_{I_t}. I_t is selected based on past history of draws and rewards $\{I_1, X_1, \ldots, I_{t-1}, X_{t-1}\}$.

Typically, the learner seeks to maximize the social welfare (arithmetic mean of expected rewards) $\sum_{t=1}^{T} \mathbb{E}[\mu_{I_t}]$. The average regret AR_T measures the algorithm's shortfall relative to the optimal social welfare μ^*, that is

$$\mathrm{AR}_T := \mu^* - \frac{1}{T}\sum\nolimits_{t=1}^{T} \mathbb{E}[\mu_{I_t}]\,. \tag{1}$$

The average regret is a utilitarian metric and ignores fairness across rounds (e.g, users). The Nash social welfare (geometric mean of expected rewards)$\big(\prod_{t=1}^{T} \mathbb{E}[\mu_{I_t}]\big)^{\frac{1}{T}}$ achieves per-round fairness since for the geometric mean to be large, the expected reward at each round should be large enough. Note that μ^* is also the optimal Nash social welfare, which yields the Nash regret

$$\mathrm{NR}_T := \mu^* - \big(\prod\nolimits_{t=1}^{T} \mathbb{E}[\mu_{I_t}]\big)^{\frac{1}{T}}\,. \tag{2}$$

Remark 1 ($\mathrm{NR}_T \geq \mathrm{AR}_T$ (Barman et al. 2023))**.** *The AM-GM inequality yields that* $\mathrm{NR}_T \geq \mathrm{AR}_T$*, implying that Nash regret is a stricter metric than average regret. This further implies that any upper bound on* NR_T *is also an upper bound on* AR_T*, and any lower bound on* AR_T *is also a lower bound on* NR_T*.*

Differential Privacy (DP) DP is a rigorous framework that ensures an algorithm's output remains almost the same under a change in one input datum, thereby protecting the

sensitive information about any individual data point. Let $\mathcal{D}$ be the data universe. Two datasets $D, D' \in \mathcal{D}$ are called neighboring if they differ only in a single datum. The standard definition of DP is as follows (Dwork et al. 2009).

Definition 1 (Differential Privacy (DP)). *For any $\epsilon > 0$, a randomized mechanism $\mathcal{M}$ satisfies ϵ-differential privacy if for all neighboring datasets D, D', and for all measurable subsets $S \subseteq \mathrm{Range}(\mathcal{M})$,*

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(D') \in S] \,.$$

Here ϵ controls the level of privacy; smaller values of ϵ imply stronger privacy and vice versa. Differential privacy in bandits is generally studied under two models: *global differential privacy* and *local differential privacy* to protect users' private and sensitive rewards.

In the global model, a trusted server (learning agent) has access to the user's rewards $X_{1:T} = \{X_1, \ldots, X_T\}$. It needs to ensure that its decisions $I_{1:T} = \{I_1, \ldots I_T\}$ are not distinguishable for two neighboring reward sequences. Formally, we call a bandit algorithm over T rounds ϵ-global DP if

$$\Pr[I_{1:T} | X_{1:T}] \leq e^{\epsilon} \Pr[I_{1:T} | X'_{1:T}]$$

for every arm sequence $I_{1:T} \in [k]^T$ and every pair of neighboring reward sequences $X_{1:T}, X'_{1:T} \in [0,1]^T$ such that $X_s \neq X'_s$ for some $s \in [T]$ and $X_t = X'_t$ for all $t \in [T] \backslash \{s\}$.

In contrast, the local model assumes no trusted aggregator or learning agent: each user t perturbs her reward X_t locally using a mechanism $\mathcal{M}$ before sending it to the agent. The agent decides on which arm to pull based on the perturbed reward $\mathcal{M}(X_t)$ and the history. Formally, we call a bandit algorithm over T rounds ϵ-local DP if

$$\forall t \in [T] \,, \Pr[\mathcal{M}(X_t) = z] \leq e^{\epsilon} \Pr[\mathcal{M}(X'_t) = z]$$

for every $z \in \mathbb{R}$ and every pair of rewards $X_t, X'_t \in [0,1]$ such that $X_t \neq X'_t$.

Local DP (LDP) offers stronger individual-level privacy than global DP (GDP) but often at the cost of slower learning due to higher noise and lack of centralized access to rewards, leading to worse regret performance. Differential privacy in bandits is a well-explored area and has been studied under both global (Azize and Basu 2022; Tossou and Dimitrakakis 2016) and local (Basu, Dimitrakakis, and Tossou 2019; Wenbo Ren and Shroff 2020) models. For more details on prior work in DP bandits, see the appendix.

A Primer on Algorithm Design

Before presenting our differentially private algorithms under the global and local privacy models, we first introduce a common framework that underlines both. At the core of our approach is a Nash regret-minimizing dynamic, which guides the learning process. This section outlines a high-level blueprint for how privacy-preserving estimators are integrated into this dynamic. Understanding this foundation will help clarify the algorithmic design choices specific to the GDP and LDP variants. Our algorithms are structured in two main phases:

Phase I (Uniform Exploration) In this phase, we repeatedly sample each arm uniformly until the number of pulls n_i times the private and randomized empirical mean $\widetilde{\mu}_i$ for some arm i exceeds a threshold that depends on the horizon T and the privacy level ϵ. Specifically, for the GDP setting (Algorithm 1), we stop Phase I when

$$n_i \, \widetilde{\mu}_i \;\geq\; 1600\Big(c^2 \log T + \tfrac{(\log T)^2}{\epsilon}\Big) \text{ for some arm } i \,.$$

Similarly, for the LDP setting (Algorithm 2), Phase I is terminated when the following holds for some arm i:

$$n_i \widetilde{\mu}_i \geq \max \Bigg\{ \frac{1}{\epsilon} \sqrt{8 n_i \alpha \log T}, \;\; \frac{\sqrt{8 n_i \alpha \log T}}{\epsilon} + 1600\Big(c^2 \log T + \frac{n_i (\log T)^2}{\big(n_i \widetilde{\mu}_i - \frac{1}{\epsilon}\sqrt{8\, n_i \alpha \log T}\big)\, \epsilon^2}\Big) \Bigg\} \,.$$

The stopping conditions are carefully designed to ensure that the number of exploration rounds—while dependent on the underlying bandit instance—does not grow excessively large with high probability, thereby keeping the (Nash) regret due to random exploration in control.

Phase II (Private Adaptive Exploitation) In this phase, the algorithm switches to an upper confidence bound (UCB)-style arm selection. To minimize Nash regret, the standard UCB computation for average regret minimization (Bubeck, Cesa-Bianchi et al. 2012) is modified in Barman et al. (2023) to define the Nash confidence bound at round t:

$$\mathrm{NCB}(i,t) \coloneqq \widehat{\mu}_i + 4\sqrt{\frac{\widehat{\mu}_i \log T}{n_i}} \,, \tag{3}$$

where $\widehat{\mu}_i$ is the non-private empirical mean of arm i at round t computed from n_i reward samples.

To ensure privacy and then to account for different amounts of noise added for privacy, we adjust the NCB differently for GDP and LDP settings. Since the observed rewards lie in $[0,1]$, to ensure GDP, we add Laplace noise with scale $\frac{\log T}{\epsilon n_i}$ to each empirical mean $\widehat{\mu}_i$ and obtain the private mean $\widetilde{\mu}_i$. A detailed proof of privacy is given in the appendix. Similarly, to ensure LDP, Laplace noise with scale $1/\epsilon$ is directly added to each observed reward, which leads to a higher amount of overall noise in the private mean $\widetilde{\mu}_i$.

For the global setup, we follow the batch arm selection idea of Azize and Basu (2022), where a single arm is selected for a fixed duration (denoted by episode ℓ). To account for the added noise, the modified NCB is computed using the private mean estimate ($\widetilde{\mu}_i$) made at the end of the previous episode (c, α are constants):

$$\mathrm{NCB}_{\mathrm{GDP}}(i,t) \;=\; \widetilde{\mu}_i \;+\; 2c\sqrt{\frac{2\,\widetilde{\mu}_i \,\log T}{n_i}} \;+\; \frac{\alpha (\log T)^2}{\epsilon\, n_i} \;+ 4\sqrt{\frac{2\alpha}{\epsilon}}\, \frac{(\log T)^{3/2}}{n_i} \,, \tag{4}$$

We compute $\mathrm{NCB}_{\mathrm{GDP}}(i,t)$ for each arm i at the beginning of each episode ℓ, and the arm with the highest NCB is the

one that is pulled throughout the episode. At the end of the episode, the number of pulls and the empirical mean are reset to those obtained in Phase I.

For the local setup, we follow the standard sequential arm selection rule, where the modified NCB is computed using the private mean $\widetilde{\mu}_i$ at round t:

$$\mathrm{NCB}_{\mathrm{LDP}}(i,t) \;=\; \widetilde{\mu}_i \;+\; 2c\sqrt{\frac{2\,\widetilde{\mu}_i\,\log T}{n_i}} + \frac{1}{\epsilon}\sqrt{\frac{8\alpha\log T}{n_i}} \;+ 4c\frac{(2\alpha)^{\frac{1}{4}}(\log T)^{\frac{3}{4}}}{\sqrt{\epsilon}(n_i)^{\frac{3}{4}}}\,, \quad (5)$$

and the arm with the highest $\mathrm{NCB}_{\mathrm{LDP}}(i,t)$ is pulled. Finally, as a post-processing step, we clip the private means to $[0,1]$ to keep them in the desired range for both algorithms.

The first two terms in our modified indices arise from substituting the empirical mean in the original NCB expression with its randomized and private counterpart. The remaining two terms are carefully constructed to ensure that, with high probability, the overall index remains a valid upper bound on the true mean, despite the added noise. This design is essential to uphold the core principle of the Upper Confidence Bound (UCB) framework, which relies on optimistic estimates—that is, selecting arms based on high-probability overestimates of their expected rewards. In effect, these additional terms compensate for the injected noise to maintain optimism: they bound the potential deviation due to privacy-induced randomness with probability at least $1-1/T^{\alpha}$, and are instrumental in the transition from private to non-private mean estimates during the regret analysis.

GDP-NCB: Nash Confidence Bound in Global DP

In this section, we present our globally differentially private algorithm: GDP-NCB. To privatize the mean estimates, we adopt the hybrid mechanism proposed by Chan, Shi, and Song (2011). Notably, the overall privacy guarantee of GDP-NCB aligns with that of the mechanism used to compute the private mean for each arm. Moreover, the privacy budget is consumed only when a new private mean is released.

To preserve privacy more effectively, we release private means less frequently using an episodic structure. We issue a new private mean only once per episode instead of every round. This strengthens privacy but introduces a trade-off, as the estimate may not capture recent rewards, potentially affecting performance. Each private mean is computed solely from rewards within its episode, keeping sensitivity local and preventing cumulative privacy leakage over time.

Algorithm 1 incorporates these design choices. Specifically, $N_{1,i}$ denotes the number of times arm i is pulled during Phase I, while $N_{2,i}$ counts the number of pulls for arm i within an episode during Phase II. The empirical and private means in Phase I are denoted by $\widehat{\mu}$ and $\widetilde{\mu}$, respectively, and the empirical mean in episode ℓ of Phase II is denoted by $\widehat{\mu}^{\ell}$. The following theorem provides a bound on the Nash regret.

Theorem 1 (Nash regret in global model)**.** *Fix time horizon $T \in \mathbb{N}$ and privacy budget $\epsilon > 0$. Then, for the k-armed bandit problem, GDP-NCB enjoys the Nash regret*

$$\mathrm{NR}_T \;=\; O\left(\sqrt{\frac{k\log T}{T}} \;+\; \underbrace{\frac{k(\log T)^2}{\epsilon T}}_{\textit{Privacy Cost}}\right).$$

Remark 2 (Privacy cost)**.** *The cost for privacy appears in the additive term $\frac{k\,(\log T)^2}{\epsilon T}$. Notably, it only becomes dominant when $\epsilon \le \widetilde{O}\left(\sqrt{\frac{k}{T}}\right)$. This indicates that in the low-privacy regime (i.e., when ϵ is relatively large), the impact of privacy on the Nash regret is minimal. Moreover, since the privacy cost goes down as $1/\epsilon T$, GDP-NCB maintains strong performance even in the high-privacy regime.*

Remark 3 (Comparison with lower bound)**.** *Azize and Basu (2022) shows that the minimax average regret of stochastic bandits with ϵ-global DP is $\mathrm{AR}_T = \Omega\Big(\max\big\{\sqrt{\frac{k}{T}}, \frac{k}{\epsilon T}\big\}\Big)$. From Remark 1, this lower bound applies to Nash regret NR_T also. Therefore, our upper bound on Nash regret in Theorem 1 is order optimal up to poly-logarithmic factors.*

The next theorem states the privacy guarantee.

Theorem 2. *(Privacy guarantee) GDP-NCB satisfies ϵ-global differential privacy for any $\epsilon > 0$.*

Algorithm 1: GDP-NCB

Input: Number of arms k, horizon T, privacy budget ϵ

Initialization: $\widehat{\mu}_i = 0$, $\widetilde{\mu}_i = 0$, $N_{1,i} = 0 \;\forall\, i \in [k]$; $t = 1$, $c = 3$, $\alpha = 3.1$

Phase I **- Uniform Exploration**

while $\max_i n_i\,\widetilde{\mu}_i \le 1600\left(c^2\log T + \frac{(\log T)^2}{\epsilon}\right)$ and $t \le T$ **do**

  Pull arm $I_t \sim \mathrm{Unif}([k])$, observe X_t

  Update $N_{1,I_t} \leftarrow N_{1,I_t}+1$, $\widehat{\mu}_{I_t} \leftarrow \frac{N_{1,I_t}-1}{N_{1,I_t}}\widehat{\mu}_{I_t} + \frac{X_t}{N_{1,I_t}}$

  Update $\widetilde{\mu}_{I_t} \leftarrow \widehat{\mu}_{I_t} + \mathrm{Lap}\left(\frac{\log T}{\epsilon N_{1,I_t}}\right)$

  Update $t \leftarrow t+1$

end while

Phase II **- Private Adaptive Exploration**

Set $N_{2,i}(t) = 1\;\forall\, i \in [k]$

for episode $\ell = 1, 2, \ldots$ **do**

  Let $t_\ell = t+1$

  Compute $A = \mathrm{argmax}_{i\in[k]}\,\mathrm{NCB}_{\mathrm{GDP}}(i, t_\ell - 1)$

  Compute $n_s = N_{2,A}(t_\ell - 1)$

  Update $\widehat{\mu}^{\ell}_A \leftarrow \widehat{\mu}_A$, $N_{2,A} \leftarrow 0$

  for $m = 1$ to $2n_s$ **do**

    Pull arm A, observe X_t

    Update $N_{2,A} \leftarrow N_{2,A} + 1$

    Compute $n_A = N_{2,A} + N_{1,A}$

    Update $\widehat{\mu}^{\ell}_A \leftarrow \frac{n_A-1}{n_A}\widehat{\mu}^{\ell}_A + \frac{X_t}{n_A}$

  end for

  Update $\widetilde{\mu}_A \leftarrow \widehat{\mu}^{\ell}_A + \mathrm{Lap}\left(\frac{\log T}{\epsilon n_A}\right)$

  Clip private mean $\widetilde{\mu}_A = \min\{1, \max\{0, \widetilde{\mu}_A\}\}$

end for

Proof Sketch. The proof proceeds via a case analysis, a standard approach in ϵ-differential privacy (DP) proofs, where we consider the perturbation of a single reward and analyze the effect based on the phase in which the perturbed reward appears. In the first case, we assume the perturbed reward belongs to Phase II. Here, we leverage the fact that rewards are non-overlapping across episodes, which ensures that the Laplace noise added in Line (20) of Algorithm 1 maintains differential privacy within each episode. In the second case, we consider a reward perturbed in Phase I. While Phase I itself involves no privacy leakage (as no private outputs are released during this phase), the same reward may influence multiple episodes in Phase II. This could, in principle, lead to a composition of privacy loss. However, we exploit the doubling schedule used in the algorithm, which bounds the number of episodes by $O(\log T)$. Consequently, the overall privacy cost remains controlled and does not become prohibitive. A detailed proof is provided in the appendix. □

LDP-NCB: Nash Confidence Bound in Local DP

Local differential privacy (LDP) presents a compelling alternative for decentralized or federated environments where no trusted curator exists. In the LDP model, each user independently perturbs their reward using the Laplace mechanism, i.e., by adding noise drawn from $\mathrm{Lap}(1/\epsilon)$ before reporting, thereby ensuring ϵ-DP at the individual level for rewards in the range $[0, 1]$. Although LDP typically incurs higher regret than GDP due to increased noise, it is often more suitable for real-world scenarios involving privacy-sensitive or distributed data collection. Building on the work of Wenbo Ren and Shroff (2020), we propose LDP-NCB, a locally private variant of the NCB algorithm. Analogous to our global model result, we also provide formal guarantees for Nash Regret in the LDP setting, as stated in the next theorem.

Theorem 3 (Nash regret in local model)**.** *Fix time horizon $T \in \mathbb{N}$ and privacy budget $\epsilon > 0$. Then, for the k-armed bandit problem, LDP-NCB enjoys the Nash regret*

$$\mathrm{NR}_T \;=\; O\left(\sqrt{\frac{k\log T}{T}} \;+\; \underbrace{\frac{\sqrt{k}(\log T)^2}{\epsilon\sqrt{T}}}_{\textit{Privacy Cost}}\right).$$

Remark 4 (Privacy cost)**.** *In the local-DP setting, the privacy penalty is significantly more severe due to the multiplicative dependence on $1/\epsilon$. Unlike the global-DP setting—where the impact of privacy can be negligible for moderately high values of ϵ—LDP does not admit a regime where performance is unaffected by privacy constraints. This is expected, as LDP imposes a stricter privacy requirement, with the algorithm observing only locally privatized data. As a result, the regret scales as $\widetilde{O}\left(\frac{1}{\epsilon}\sqrt{\frac{k}{T}}\right)$, deteriorating rapidly as stronger privacy (i.e., smaller ϵ) is enforced.*

Remark 5 (Lower bound comparison)**.** *Our bound is order-optimal up to logarithmic factors. This follows from Remark 1 and the $\Omega\left(\frac{1}{\epsilon}\sqrt{\frac{k}{T}}\right)$ minimax lower bound on average regret for ϵ-LDP algorithms in high-privacy (sufficiently small ϵ) regime (Basu, Dimitrakakis, and Tossou 2019).*

Algorithm 2: LDP-NCB

Input: Number of arms k, horizon T and privacy budget ϵ

Initialization: $\widetilde{\mu}_i = 0$, $N_i = 0 \;\forall\, i \in [k]$; $t = 1, c = 3$, $\alpha = 3.1$

Phase I **- Uniform Exploration**

while $\widetilde{\mu}_i \le \frac{1}{\epsilon}\sqrt{\frac{8\alpha\log T}{n_i}}$ or $n_i\,\widetilde{\mu}_i \;\le\; 1600\left(c^2\log T + \frac{(\log T)^2}{\left(\widetilde{\mu}_i - \frac{1}{\epsilon}\sqrt{\frac{8\alpha\log T}{n_i}}\right)\epsilon^2}\right) + \frac{\sqrt{8n_i\alpha\log T}}{\epsilon}\;\forall\, i \in [k]$ and $t \le T$ **do**

  Pull arm $I_t \sim \mathrm{Unif}([k])$

  Observe private reward $\widetilde{X}_t = X_t + \mathrm{Lap}(1/\epsilon)$

  Update $N_{I_t} \leftarrow N_{I_t} + 1$, $\widetilde{\mu}_{I_t} \leftarrow \frac{N_{I_t}-1}{N_{I_t}}\widetilde{\mu}_{I_t} + \frac{\widetilde{X}_t}{N_{I_t}}$

  Update $t \leftarrow t + 1$.

end while

Phase II **- Private Adaptive Exploration**

while $t \le T$ **do**

  Pull arm $I_t = \mathrm{argmax}_{i\in[k]}\,\mathrm{NCB}_{\mathrm{LDP}}(i, t)$

  Observe private reward $\widetilde{X}_t = X_t + \mathrm{Lap}(1/\epsilon)$

  Update $N_{I_t} \leftarrow N_{I_t} + 1$, $\widetilde{\mu}_{I_t} \leftarrow \frac{N_{I_t}-1}{N_{I_t}}\widetilde{\mu}_{I_t} + \frac{\widetilde{X}_t}{N_{I_t}}$

  Clip private mean $\widetilde{\mu}_{I_t} = \min\{1, \max\{0, \widetilde{\mu}_{I_t}\}\}$

  Update $t \leftarrow t + 1$.

end while

Analysis and Proof Sketch

We now provide a proof sketch for Theorems 1 and 3. The sketch is presented in a unified manner, as the key ideas underlying both proofs are largely similar, differing only in a few crucial aspects. A complete and rigorous treatment is deferred to the Appendix.

Denote by $\widehat{\mu}_{i,s}$ the empirical mean reward of arm i after its first s pulls. For the global setup, define

$$S := \frac{c^2\,\log T}{\mu^*} + \frac{(\log T)^2}{\mu^*\,\epsilon}\,, \tag{6}$$

and for the local setup, define

$$S = \frac{c^2\log T}{\mu^*} + \left(\frac{\log T}{\mu^*\epsilon}\right)^2\,. \tag{7}$$

Next, we define a *good event* $\mathcal{E}$, which holds with high probability. All subsequent arguments are conditioned on the event $\mathcal{E}$ holding. Let $\mathcal{E}$ be the intersection of four sub-events: $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$, and $\mathcal{E}_4$, defined as follows:

$\mathcal{E}_1$: During the initial r rounds of uniform sampling, for every arm $i \in [k]$ and any $r \ge 512k\,S$, the number of pulls of arm i lies between $r/(2k)$ and $3r/(2k)$.

$\mathcal{E}_2$: For each arm $i \in [k]$ with mean $\mu_i > \mu^*/256$, and for all sample sizes s satisfying $256\,S \le s \le T$, the empirical estimate obeys $\left|\widehat{\mu}_{i,s} - \mu_i\right| \;\le\; c\sqrt{\frac{\mu_i\,\log T}{s}}$.

$\mathcal{E}_3$: For each arm $j \in [k]$ with $\mu_j \leq \mu^*/256$, and for all s in the range $256\, S \leq s \leq T$, we have $\widehat{\mu}_{j,s} < \frac{\mu^*}{128}$.

$\mathcal{E}_4$: For each arm $i \in [k]$, the difference between private and non-private empirical mean $|\widetilde{\mu}_i - \widehat{\mu}_i|$ is δ, where $\delta = \frac{\alpha(\log T)^2}{\epsilon\, n_i}$ for GDP and $\delta = \frac{1}{\epsilon}\sqrt{\frac{8\alpha \log T}{n_i}}$ for LDP.

We now state a lemma that provides a high-probability guarantee for the event $\mathcal{E}$ with its proof deferred to the appendix.

Lemma 2 (Good event)**.** *For any $T \in \mathbb{N}$, $\Pr[\mathcal{E}] \geq 1 - \frac{6}{T}$ for both GDP-NCB and LDP-NCB algorithms.*

Next, we state the following informal lemma, which helps us bound the total length of Phase I in our algorithms. Rigorous statements are given in lemmas 10-12 (for GDP) and 17-19 (for LDP) in the appendix.

Lemma (Informal)**.** *Under event $\mathcal{E}$, Phase I in both algorithms runs for $\Theta(kS)$ rounds.*

We now turn to three crucial properties of our algorithms that hold under the "good" event $\mathcal{E}$.

Lemma (Informal)**.** *Throughout Phase II, the NCB of the optimal arm i^* never falls below its true mean μ^*.*

Lemma (Informal)**.** *Any arm j with mean $\mu_j \leq \mu^*/256$ is never selected during Phase II.*

Lemma (Informal)**.** *If an arm i is pulled in Phase II, then μ_i must be close to μ^*, so its contribution to the Nash regret is negligible.*

More rigorous treatment for the above informal lemmata is provided in lemmas 13-15 (for GDP) and 20-22 (for LDP) in the appendix. The above three lemmas help us derive a bound on the Nash Social Welfare (NSW) for our algorithms, which is captured below informally.

Lemma (Informal)**.** *For $\mu^* = \widetilde{\Omega}(\frac{1}{\sqrt{T}} + \frac{1}{\epsilon T})$, the NSW for the GDP-NCB upto T rounds satisfies $(\prod_{t=1}^{T} \mathbb{E}[\mu_{I_t}])^{\frac{1}{T}} \approx \mu^* - \sqrt{\frac{\mu^*}{T}} - \frac{1}{\epsilon T}$.*

This bound is stated in detail in Lemma 16 (for GDP). Finally, we use this lemma to derive the Nash regret bound in Theorem 1). Observe that the conclusion in Theorem 1 holds trivially when $\mu^* = \widetilde{O}(\frac{1}{\sqrt{T}} + \frac{1}{\epsilon T})$ Thus, we need to prove the bound for $\mu^* = \widetilde{\Omega}(\frac{1}{\sqrt{T}} + \frac{1}{\epsilon T})$, which follows from the NSW identity stated above. Detailed proof of Theorem 1 is given in the appendix.

A similar analysis for NSW works for the LDP-NCB algorithm (Lemma 23) and its Nash regret (Theorem 3).

Remark 6 (Clipping)**.** *The concentration bounds remain valid even after clipping of the private mean to the interval $[0, 1]$. Specifically, if $\tilde{\mu}$ lies within an interval $[L, U]$, then the clipped value $\tilde{\mu}_{clip} = \min\{1, \max\{0, \tilde{\mu}\}\}$ continues to satisfy $L \leq \tilde{\mu}_{clip} \leq U$. If $L \leq 0$, clipping only increases $\tilde{\mu}$ (if needed) up to 0, thereby preserving the lower bound. If $L > 0$, then $\tilde{\mu} \geq L > 0$, so clipping does not alter the value. Similarly, the upper bound is preserved since $\tilde{\mu}_{clip} \leq \tilde{\mu} \leq U$. Hence, all upper and lower bounds used in regret and confidence interval analyses remain valid under clipping.*

Algorithm 3: Anytime Algorithm for Nash Regret

Input: Number of arms k, privacy budget ϵ

Initialize $W = 1$.
while the MAB process continues **do**
  With probability $\frac{1}{W^2}$ set flag = UNIFORM , otherwise, with probability $\left(1 - \frac{1}{W^2}\right)$, set flag = DP-NCB
  if flag = UNIFORM **then**
    for $t = 1$ to W **do**
      Select I_t uniformly at random from $[k]$. Pull arm I_t and observe reward X_t.
    end for
  else if flag = DP-NCB **then**
    Execute DP-NCB(k, W, ϵ).
  end if
  Update $W \leftarrow 2 \times W$.
end while

Anytime Private and Fair Algorithm

We now present the *anytime* version of our DP-NCB framework, described in Algorithm 3. An *anytime algorithm* is designed to operate without prior knowledge of the total time horizon. Following the approach of Barman et al. (2023), our algorithm employs the *doubling trick*. It maintains a current window length $W \in \mathbb{Z}_+$ as a guess for the horizon. During each epoch of W rounds, the algorithm proceeds as follows:

1. With probability $1/W^2$, it performs uniform exploration;
2. With probability $1 - 1/W^2$, it invokes either Algorithm 1 or Algorithm 2, depending on the privacy setting, over the remaining rounds of the window.

At the end of each epoch, the window length is doubled, and the process is repeated until a termination signal is received.

Note that in Algorithm 3, DP-NCB can refer to either GDP-NCB or LDP-NCB, depending on the privacy setting. Algorithm 3 leads to Theorem 4 and Theorem 5, which share similar proof techniques. A proof sketch is provided below, with complete proofs deferred to the appendix.

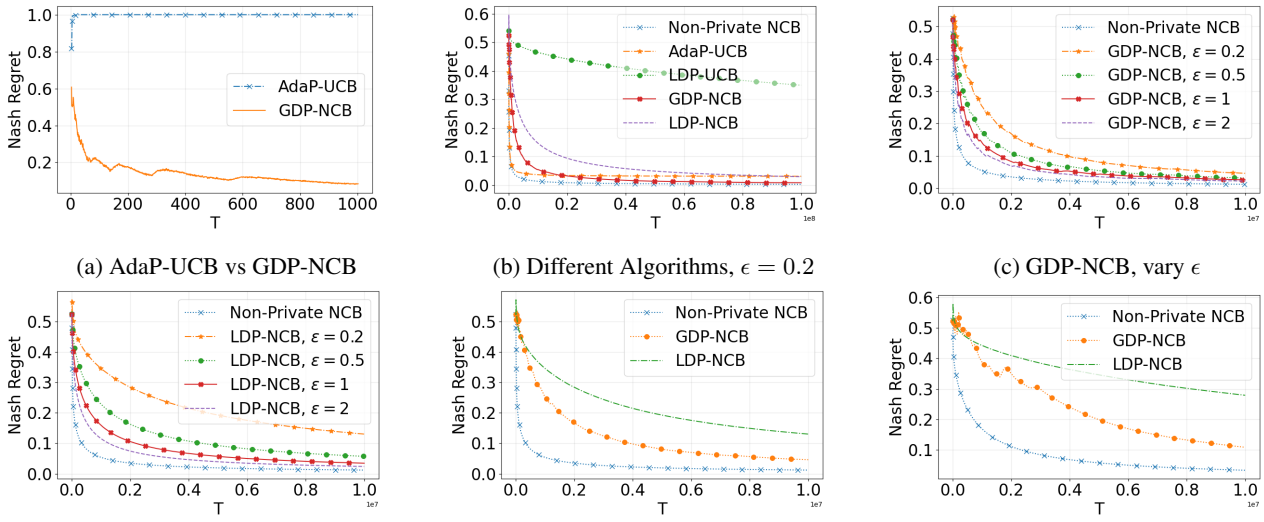
Theorem 4. *For any $\epsilon > 0$ and a sufficiently large T, Algorithm 3, instantiated with GDP-NCB, satisfies ϵ-global DP and guarantees Nash regret*

$$\mathrm{NR}_T = O\left(\sqrt{\frac{k \log T}{T}} \log T + \frac{k\,(\log T)^3}{\epsilon T}\right).$$

Theorem 5. *For any $\epsilon > 0$ and a sufficiently large T, Algorithm 3, instantiated with LDP-NCB, satisfies ϵ-local DP and guarantees Nash regret*

$$\mathrm{NR}_T = O\left(\sqrt{\frac{k \log T}{T}} \log T + \frac{\sqrt{k}(\log T)^2}{\epsilon\sqrt{T}}\right).$$

Proof Sketch. The proof uses the doubling trick to create an anytime version of the DP-NCB algorithm. In each epoch of guessed length W, the algorithm either performs uniform exploration with probability $\frac{1}{W^2}$ or runs the base DP-NCB algorithm (GDP-NCB or LDP-NCB) on the epoch. Since the probability of choosing uniform exploration is low and decays with W, most epochs run DP-NCB, ensuring that the



(a) AdaP-UCB vs GDP-NCB

(b) Different Algorithms, $\epsilon = 0.2$

(c) GDP-NCB, vary ϵ

(d) LDP-NCB, vary ϵ

(e) Bernoulli arms, $\epsilon = 0.2$

(f) Mixed types of arms, $\epsilon = 0.2$.

Figure 1: Numerical results for GDP-NCB and LDP-NCB. (a) shows that GDP-NCB significantly outperforms AdaP-UCB by avoiding blown-up Nash regret in extreme instances. (b) illustrates the comparison between our algorithms and existing algorithms. (c) and (d) show that as ϵ decreases, Nash regret for DP-NCB increases, aligning with theoretical predictions. (e) shows that GDP-NCB's regret decays slower than NCB, and LDP-NCB follows the slowest decay, under Bernoulli arms. (f) shows that GDP-NCB remains robust and maintains Nash regret trends even under heterogeneous reward distributions.

Nash regret remains close to that of the fixed-horizon case. The overall regret is then obtained by summing the regrets (calculated from Theorem 1 and Theorem 3 for each epoch) over all epochs, using the fact that the number of epochs is $O(\log T)$. This introduces only an $O(\log T)$ multiplicative overhead in regret, yielding the final bound stated in Theorem 4. Theorem 5 can be proved along similar lines. □

Experiments

Here we present numerical results comparing our methods with NCB (Barman et al. 2023) and AdaP-UCB (Azize and Basu 2022) algorithms. All experiments were conducted on an Apple M4 CPU with 16GB RAM. All experiments report the average reward over 50 runs to estimate $\mathbb{E}[\mu_{I_T}]$, using a privacy parameter of $\epsilon = 0.2$ wherever applicable.

First, we demonstrate how the Nash regret NR_T of AdaP-UCB can grow rapidly for certain bandit instances. Specifically, we consider a setting with $k = 2$ Bernoulli arms having means $\mu_1 = (2e)^{-T}$ and $\mu_2 = 1$, for $T \in (0, 1000)$, following Barman et al. (2023). In this case, AdaP-UCB exhibits significantly inflated Nash regret, while our private algorithm maintains controlled regret, as shown in Figure (a).

Next, we compare the Nash regret of Non-Private NCB, AdaP-UCB, LDP-UCB (Wenbo Ren and Shroff 2020), GDP-NCB, and LDP-NCB on a problem instance with $k = 50$ Bernoulli arms, where the means are sampled uniformly from the interval $(0.005, 1)$. We use parameters $c = 3$ and $\alpha = 3.1$, following the experimental setup of Krishna et al. (2025). As shown in Figure (b), GDP-NCB successfully reduces Nash regret compared to AdaP-UCB for larger values of T. Similarly, in the LDP setting, LDP-NCB achieves significantly lower regret compared to LDP-UCB.

Further, we evaluate Nash Regret for GDP-NCB and LDP-NCB for varying ϵ (Figure (c) and (d)). As predicted by our analysis, regret increases as ϵ decreases, with non-private NCB consistently achieving the lowest Nash Regret. We then compare Non-Private NCB, GDP-NCB, and LDP-NCB on $k = 50$ Bernoulli arms with similar setup as the previous experiment. Figure (e) shows that GDP-NCB's regret decays more slowly than Non-Private NCB (due to Laplace noise), and LDP-NCB decays the slowest, following an $O(1/\sqrt{T})$ rather than $O(1/T)$ rate.

Finally, we evaluate performance with mixed arms, where the 50 arms have means sampled uniformly from $(0.005, 1)$, but differ in reward distributions: arms with $\mu_i \geq 0.75$ are Bernoulli; arms with $\mu_i \in [0.5, 0.75)$ follow $\mathrm{Beta}(4, 1)$; arms with $\mu_i \in [0.25, 0.5)$ follow a two-point distribution $\{0.4, 1\}$ and arms with $\mu_i < 0.25$ follow $\mathrm{Unif}(0, 1)$. As shown in Figure (f), Nash regret trends are similar to Figure (e) as T increases, highlighting the algorithms' robustness to heterogeneous reward distributions.

Conclusion

This work bridges the gap between privacy-preserving algorithms and welfarist concerns by introducing a unified framework—DP-NCB. Our algorithms provide a versatile solution to the MAB problem with applications in sensitive domains such as personalized recommendations and healthcare. Future directions include extending the framework to other reinforcement learning settings and exploring alternative fairness metrics tailored to specific domains.

Acknowledgments

SRC would like to thank Siddharth Barman for the discussion on the Nash welfare metric and acknowledges an ANRF Early-Career Research Grant.

References

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

Barman, S.; Khan, A.; Maiti, A.; and Sawarni, A. 2023. Fairness and welfare quantification for regret in multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 6762–6769.

Basu, D.; Dimitrakakis, C.; and Tossou, A. 2019. Differential privacy for multi-armed bandits: What is it and what is its cost? *arXiv preprint arXiv:1905.12298*.

Bubeck, S.; Cesa-Bianchi, N.; et al. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1): 1–122.

Chan, T.-H. H.; Shi, E.; and Song, D. 2011. Private and continual release of statistics. *ACM Transactions on Information and System Security (TISSEC)*, 14(3): 1–24.

Clement, B.; Roy, D.; Oudeyer, P.-Y.; and Lopes, M. 2015. Multi-Armed Bandits for Intelligent Tutoring Systems. *Journal of Educational Data Mining*, 7(2).

Dwork, C.; Naor, M.; Reingold, O.; Rothblum, G. N.; and Vadhan, S. 2009. On the complexity of differentially private data release: efficient algorithms and hardness results. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, 381–390.

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

Krishna, A.; John, P. G.; Barik, A.; and Tan, V. Y. 2025. p-Mean Regret for Stochastic Bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 17966–17973.

Lattimore, T.; and Szepesv´ari, C. 2020. *Bandit Algorithms*. Cambridge University Press.

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, 661–670.

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*, 592–601.

Moulin, H. 2004. *Fair division and collective welfare*. MIT press.

Pan, X.; Wang, W.; Zhang, X.; Li, B.; Yi, J.; and Song, D. 2019. How You Act Tells a Lot: Privacy-Leaking Attack on Deep Reinforcement Learning. In *Aamas*, volume 19, 368–376.

Sawarni, A.; Pal, S.; and Barman, S. 2023. Nash regret guarantees for linear bandits. *Advances in Neural Information Processing Systems*, 36: 33288–33318.

Schwartz, E. M.; Bradlow, E. T.; and Fader, P. S. 2017. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4): 500–522.

Tewari, A.; and Murphy, S. A. 2017. From ads to interventions: Contextual bandits in mobile health. In *Mobile health: sensors, analytic methods, and applications*, 495–517. Springer.

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.; and Dimitrakakis, C. 2016. Algorithms for differentially private multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Villar, S. S.; Bowden, J.; and Wason, J. 2015. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2): 199.

Wenbo Ren, J. L., Xingyu Zhou; and Shroff, N. B. 2020. Multi-Armed Bandits with Local Differential Privacy. arXiv:2007.03121.