rithm takes as input: the window size w, a non-decreasing function f, the horizon T and the corruptions functions $g_1, \ldots, g_K$. We assume that the horizon T is known; an unknown T can be handled using the doubling trick (**?**). We use $d(x, y)$ to denote the KullbackLeibler divergence between two Bernoulli distributions with mean x and y. We also use a shorthand of $x \wedge y$ to denote $\min(x, y)$.

At each time time step t, the algorithm computes an $\text{Index}_a(t)$, which is an upper-confidence bound on $\mu_{a,t}$ built from a confidence interval on $\lambda_{a,t}$ based on the KL-divergence. The quantity $N_a(t, w)$ denotes the number of times arm a was chosen in the last w time steps until time t. Correspondingly, $\hat{\lambda}_a(t, w)$ denotes the empirical mean of the feedback observed from arm a in the last w time steps until time t: $\hat{\lambda}_a(t, w) \coloneqq \frac{1}{N_a(t,w)} \sum_{s=\min\{1,t-w+1\}}^{t} F_s \cdot \mathbb{1}_{(\hat{a}_s = a)}$.

Theorem 1 gives an upper bound on the regret of SW-KLUCB-CF. A more explicit bound is proved in the Appendix.

Theorem 1 *The regret of* SW-KLUCB-CF *using* $f(x) \coloneqq \log(x) + 3\log(\log(x))$ *and* $w = \sqrt{\frac{4eT}{L_T+4}}$ *on a Bernoulli non-stationary stochastic corrupt bandits problem with strictly monotonic and continuous corruption functions* $\{g_a\}_{a \in A}$ *at time* T *is upper-bounded by* [2]

$$\tilde{O}\left(\sum_{a \in A} \sqrt{L_T T} + \sum_{i=1}^{L_T} \sum_{a \neq a_*(i)} \frac{\log\left(\sqrt{\frac{T}{L_T}}\right)}{d(\lambda_a(i), g_a(\mu_*(i))}\right),$$

where $a_*(i)$ *and* $\mu_*(i)$ *are the optimum arm and the corresponding optimal mean respectively after* i^{th} *change and before the subsequent change.*

The lower bound on regret in terms T for classical non-stationary stochastic bandits is $\Omega(\sqrt{T})$ (**?**). Theorem 1 matches the lower bound up to logarithmic factors, so SW-KLUCB-CF has near-optimal regret guarantees in terms of the time horizon T. The best known regret upper bounds for classical non-stationary stochastic bandits (e.g., **?**) also feature logarithmic terms besides the lower bound, hence our regret bound is in line with the best known results for analogous problems. Moreover, the bound in Theorem 1 also matches the best known regret bound in terms of L_T for classical non-stationary stochastic bandits which is $O\sqrt{L_T}$.

We can use SW-KLUCB-CF on non-stationary stochastic corrupts bandits where the corruption is done via randomized response. The following corollary bounds the resulting regret.

Corollary 1 *The regret of* SW-KLUCB-CF *on a Bernoulli non-stationary stochastic corrupt bandits problem with randomized response using corruption matrices* $\{\mathbb{M}\}_{a \in A}$ *at time* T *is upper-bounded by*

$$\tilde{O}\left(\sum_{a \in A} \sqrt{L_T T} + \sum_{i=1}^{L_T} \sum_{a \neq a_*(i)} \frac{\log\left(\sqrt{\frac{T}{L_T}}\right)}{(p_{00}(a) + p_{11}(a) - 1)^2}\right).$$

[2] $\tilde{O}$ ignores logarithmic factors and constants.

Algorithm 1: Sliding Window KLUCB for Non-Stationary Stochastic Corrupt Bandits (SW-KLUCB-CF)

Input: Window size w, a non-decreasing function $f : \mathbb{N} \to \mathbb{R}$, T, monotonic and continuous corruption functions $g_1, \ldots, g_K$ and $d(x, y) \coloneqq \text{KL}(\mathcal{B}(x), \mathcal{B}(y))$,

1. **Initialization:** Pull each arm once.
2. **for** time $t = K, \ldots, T-1$ **do**
   (a) Compute for each arm $a \in A$ the quantity

$$\text{Index}_a(t) \coloneqq \max\left\{q : \; N_a(t, w) \cdot d(\hat{\lambda}_a(t, w), g_a(q)) \leq f\left(t \wedge w\right)\right\}$$

   (b) Pull arm $\hat{a}_{t+1} \coloneqq \underset{a \in A}{\text{argmax}}\, \text{Index}_a(t)$ and observe the feedback F_{t+1}.

end for

This corollary follows from Theorem 1 and Pinsker's inequality: $d(x, y) > 2(x-y)^2$. The term $(p_{00}(a)+p_{11}(a)-1)$ can be understood as the slope of the corruption function g_a.

Corruption Mechanism to Preserve Local Privacy in Non-Stationary Environment

First, let us formally define local differential privacy.

Definition 1 *(Locally differentially private mechanism) Any randomized mechanism* $\mathcal{M}$ *is* ϵ*-locally differentially private for* $\epsilon \geq 0$*, if for all* $d_1, d_2 \in Domain(\mathcal{M})$ *and for all* $S \subset Range(\mathcal{M})$*,*

$$\mathbb{P}[\mathcal{M}(d_1) \in S] \leq e^{\epsilon} \cdot \mathbb{P}[\mathcal{M}(d_2) \in S].$$

As done in **?**, a straightforward approach to achieve local differential privacy using corrupt bandits is to employ a corruption scheme on the user feedback. This is similar to how randomized response is used in data collection by **?**.

Definition 2 *(ϵ-locally differentially private bandit feedback corruption scheme) A bandit feedback corruption scheme* $\tilde{g}$ *is* ϵ*-locally differentially private for* $\epsilon \geq 0$*, if for all reward sequences* $R_{t1}, \ldots, R_{t2}$ *and* $R'_{t1} \ldots, R'_{t2}$*, and for all* $\mathcal{S} \subset Range(\tilde{g})$*,*

$$\mathbb{P}[\tilde{g}(R_{t1}, \ldots, R_{t2}) \in \mathcal{S}] \leq e^{\epsilon} \cdot \mathbb{P}[\tilde{g}(R'_{t1}, \ldots, R'_{t2}) \in \mathcal{S}].$$

When corruption is done by randomized response, local differential privacy requires that $\max_{1 \leq a \leq K}\left(\frac{p_{00}(a)}{1-p_{11}(a)}, \frac{p_{11}(a)}{1-p_{00}(a)}\right) \leq e^{\epsilon}$. From Corollary 1, we can see that to achieve lower regret, $p_{00}(a) + p_{11}$(a) is to be maximized for all $a \in A$. Using **?**, Result 1, we can state that, in order to achieve ϵ-local differential privacy while maximizing $p_{00}(a) + p_{11}(a)$,

$$\mathbb{M}_a = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} \frac{e^{\epsilon}}{1+e^{\epsilon}} & \frac{1}{1+e^{\epsilon}} \\ \frac{1}{1+e^{\epsilon}} & \frac{e^{\epsilon}}{1+e^{\epsilon}} \end{bmatrix} \end{array}. \tag{2}$$

As it turns out, this is equivalent to the *staircase* mechanism for local privacy which is the optimal local differential privacy mechanism for low privacy regime (**?**, Theorem 14). The trade-off between utility and privacy is controlled by ϵ. Using the corruption parameters from Eq. (2) with Corollary 1, we arrive at the following upper bound.

Corollary 2 *At time T, the regret of* SW-KLUCB-CF *with ϵ-locally differentially private bandit feedback corruption scheme given by Eq.* (2) *is* $\tilde{O}\left(\sum_{a\in A}\sqrt{L_T T} + \sum_{i=1}^{L_T}\sum_{a\neq a_*(i)} \frac{\log\left(\sqrt{\frac{T}{L_T}}\right)}{\left(\frac{e^\epsilon-1}{e^\epsilon+1}\right)^2}\right)$.

The term $\left(\frac{e^\epsilon-1}{e^\epsilon+1}\right)^2$ in the above expression conveys the relationship of the regret with the level of local differential privacy symbolized by ϵ. For low values of ϵ, $\left(\frac{e^\epsilon-1}{e^\epsilon+1}\right) \approx \epsilon/2$. This is in line with other bandit algorithms providing differential privacy (e.g., **?**).

Elements of Mathematical Analysis

Here, we provide a proof outline for Theorem 1. Please refer to the Appendix for the complete proof.

We start by bounding the expected number of times a suboptimal arm (i.e., an arm other than the optimal arm at the time of selection) is pulled by the algorithm till horizon T. Recall that, at any time step t, SW-KLUCB-CF pulls an arm maximizing an index defined as

$$\begin{aligned}&\mathrm{Index}_a(t)\\ &\coloneqq \max\left\{q:\ N_a(t,w)\cdot d\left(\hat{\lambda}_a(t,w), g_a(q)\right) \le f\left(t\wedge w\right)\right\}\\ &= \max g_a^{-1}\left(\left\{q: N_a(t,w)\cdot d\left(\hat{\lambda}_a(t,w), q\right) \le f\left(t\wedge w\right)\right\}\right).\end{aligned}$$

We further decompose the computation of index as follows,

$$\mathrm{Index}_a(t) \coloneqq \begin{cases} g_a^{-1}(\ell_a(t)) & \text{if } g_a \text{ is decreasing,}\\ g_a^{-1}(u_a(t)) & \text{if } g_a \text{ is increasing}\end{cases}$$

where,

$$\ell_a(t) \coloneqq \min\left\{q: N_a(t,w)\cdot d\left(\hat{\lambda}_a(t,w), q\right) \le f\left(t\wedge w\right)\right\},$$

$$u_a(t) \coloneqq \max\left\{q: N_a(t,w)\cdot d\left(\hat{\lambda}_a(t,w), q\right) \le f\left(t\wedge w\right)\right\}.$$

The interval $[\ell_a(t), u_a(t)]$ is a KL-based confidence interval on the mean feedback $\lambda_{a,t}$ of arm a at time t. This is in contrast to kl-UCB (**?**) where a confidence interval is placed on the mean reward. Furthermore, This differs from kl-UCB-CF (**?**) where the mean feedback of an arm remains the same for all the time steps and f does not feature w.

In our analysis, we use the fact that when an arm a is picked at time $t+1$ by SW-KLUCB-CF, one of the following is true: Either the mean feedback of the optimal arm $a_{*,t}$ with mean reward $\mu_{*,t}$ is outside its confidence interval (i.e., $g_{a_{*,t}}(\mu_{*,t}) < \ell_{a_{*,t}}(t)$ or $g_{a_{*,t}}(\mu_{*,t}) > u_{a_{*,t}}(t)$) which is unlikely. Or, the mean feedback of the optimal arm is where it should be, and then the fact that arm a is selected indicates that the confidence interval on λ_a cannot be too small as either $(u_a(t) \ge g_a(\mu_{*,t}))$ or $(\ell_a(t) \le g_a(\mu_{*,t}))$. The previous statement follows from considering various cases depending on whether the corruption functions g_a and $g_{a_{*,t}}$ are increasing or decreasing. We then need to control the two terms in the decomposition of the expected number of draws of arm a. The term regarding the "unlikely" event, is bounded using the same technique as in the kl-UCB analysis, however with some added challenges due to the use of a sliding window. In particular, the analysis of a typical upper confidence bound algorithm for bandits relies on the fact that the confidence interval for any arm is always non-increasing, however this is not true while using a sliding window. To control the second term, depending on the monotonicity of the corruption functions g_a and $g_{a_{*,t}}$, we need to meticulously adapt the arguments in **?** to control the number of draws of a suboptimal arm, as can be seen in the Appendix.

Concluding Remarks

In this work, we proposed the setting of non-stationary stochastic corrupt bandits for preserving privacy while still maintaining high utility in sequential decision making in a changing environment. We devised an algorithm called SW-KLUCB-CF and proved its regret upper bound which is near-optimal in the number of time steps and matches the best known bound for analogous problems in terms of the number of time steps and the number of changes. Moreover, we provided an optimal corruption scheme to be used with our algorithm in order to attain the dual goal of achieving high utility while maintaining the desired level of privacy.

Interesting directions for future work include:

1. Complete an empirical evaluation of the proposed algorithm on simulated as well as real-life data. 2. Characterize the changes in the environment by a variation budget (as done in **?** for classical bandits) instead of the number of changes. 3. Incorporate contextual information in the learning process. 4. Propose a Bayesian algorithm for non-stationary stochastic corrupt bandits. 5. Propose a (near-)optimal differentially private algorithm which does not need to know the number of changes.

Quo tenetur repudiandae incidunt minus, consequuntur esse pariatur fuga facilis neque, quaerat accusantium quidem eligendi amet alias mollitia, dolor esse at deleniti quas voluptas officiis reprehenderit ut hic.Cupiditate cumque recusandae sequi incidunt nemo similique sit reiciendis, consequuntur ut labore possimus vero eveniet cumque at aliquid laboriosam eligendi atque.Illo vel omnis debitis aliquam optio quisquam a dignissimos ipsa repellendus nemo, ad sint nobis ea a rerum accusamus consequatur placeat tempore, facere ipsam iusto et eum fuga, natus doloribus quaerat doloremque temporibus cupiditate hic?Illo veniam cumque porro, fuga quas quo pariatur repellat, laboriosam impedit inventore ab architecto officiis neque iste cumque aperiam iusto culpa, eaque quia ab repellendus sapiente.Quibusdam ipsam tempore dolorum, accusamus error animi iusto quam nisi maxime, eaque beatae doloribus, ratione hic animi consequatur sit recusandae eius quo?Corrupti facere numquam omnis illum nam eius nobis, in esse dolorum, praesentium veniam ab sint atque accusantium tempore?Cum esse vero alias obcaecati exercitationem corporis sed