Self-Concordant Analysis of Generalized Linear Bandits with Forgetting

Yoan Russac* DI ENS, CNRS, Inria, ENS, Université PSL

Louis Faury* Criteo AI Lab, LTCI TélécomParis

Olivier Cappé DI ENS, CNRS, Inria, ENS, Université PSL

Aurélien Garivier UMPA, CNRS, Inria, ENS Lyon

Abstract

Contextual sequential decision problems with categorical or numerical observations are ubiquitous and Generalized Linear Bandits (GLB) offer a solid theoretical framework to address them. In contrast to the case of linear bandits, existing algorithms for GLB have two drawbacks undermining their applicability. First, they rely on excessively pessimistic concentration bounds due to the non-linear nature of the model. Second, they require either non-convex projection steps or burn-in phases to enforce boundedness of the estimators. Both of these issues are worsened when considering non-stationary models, in which the GLB parameter may vary with time. In this work, we focus on self-concordant GLB (which include logistic and Poisson regression) with forgetting achieved either by the use of a sliding window or exponential weights. We propose a novel confidence-based algorithm for the maximum-likehood estimator with forgetting and analyze its perfomance in abruptly changing environments. These results as well as the accompanying numerical simulations highlight the potential of the proposed approach to address non-stationarity in GLB.

1 INTRODUCTION

In recent years, linear bandits (Abbasi-Yadkori et al., 2011; Chu et al., 2011; Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2010) have become the go-to paradigm to balance exploration and exploitation in contextual sequential decision making problems. Linear bandits have typically found applications for content-based recommendations (Li et al., 2010; Valko et al., 2014), real-time bidding (Flajolet and Jaillet, 2017) and even mobile-health interventions (Tewari and Murphy, 2017). Concurrently, Generalized linear bandits (GLB) have been introduced as a generalization of linear bandits, able to describe broader reward models of considerable practical relevance, in particular binary or categorical rewards (Filippi et al., 2010; Li et al., 2017). GLB are for instance a natural option in online advertising applications where the rewards take the form of clicks (Chapelle and Li, 2011). In this work, we focus on deterministic algorithms and refer to (Chapelle and Li, 2011; Kveton et al., 2020) for randomized algorithms applicable to GLB. Compared to the linear bandits case, there are two distinctive drawbacks of GLB algorithms. The first is **(1)** the presence of a problem-dependent constant, imposed by the non-linear nature of the model, that is possibly *prohibitively large* and has a negative impact both on the design of algorithms and on their analysis. The second is **(2)** the need to modify the Maximum Likelihood Estimator (MLE) to ensure that it has a bounded norm. Usually this is achieved by resorting to an additional *non-convex* projection program applied to the MLE (Filippi et al., 2010). These distinctions correspond to a fundamental difference between the models, and explain why methods developed for linear bandits may fail in the case of GLB.

The first drawback **(1)** was recently addressed by Faury et al. (2020), in the specific case of logistic bandits. They showed that in this particular setting, the regret bounds of carefully designed algorithms could be significantly improved only at the cost of minor algorithmic modifications. Their analysis tightens the gap with the linear case, and takes a significant step towards the development of efficient GLB algorithms.

The second drawback **(2)** has seen little treatment in the literature, except for the work of Li et al. (2017) who proved that the projection step of Filippi et al. (2010) could be avoided by resorting to random initialization phases. However, a careful examination of the required conditions shows that these initialization phases can be prohibitively long to be deployed in scenarios of

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. *Equal contribution.

practical interest.

The aforementioned improvements to the original GLB algorithm of Filippi et al. (2010) were developed under a stationarity assumption. However, non-stationary environments are ubiquitous in real-world applications of contextual bandits. In the linear bandits literature, this has motivated the development of adequate algorithms, able to handle changes in the structure of the reward signal (Cheung et al., 2019b; Russac et al., 2019; Zhao et al., 2020). Russac et al. (2020) generalized such approaches to GLB, but without addressing neither **(1)** nor **(2)**. As a result, the practical relevance of their approach remains questionable and the development of *efficient* and *non-stationary* GLB algorithms stands incomplete.

This paper aims at closing this gap. We study a broad family of GLB, known as *self-concordant* (which includes for instance the logistic and Poisson bandits), in environments where the parameter is allowed to switch arbitrarily over time. Under this setting, we answer **(1)** by providing a non-trivial extension of the concentration results from Faury et al. (2020). We also leverage the self-concordance property to *remove* the projection step, henceforth overcoming **(2)**. This is made possible by an improved characterization of the, possibly weighted, MLE in (self-concordant) generalized linear models. Combined together, these two contributions lead to the design of *efficient* GLB algorithms, with improved regret bounds and which do not require to solve hard (i.e. non-convex) optimization programs. In doing so, we also answer the long-standing issue of providing proper confidence regions centered around the pristine MLE in GLB.

2 BACKGROUND

2.1 Setting and Assumptions

At each time step, the environment provides a time-dependent action set $\mathcal{A}_t$ and the agent plays a d-dimensional action $a_t \in \mathcal{A}_t$. We will assume that the reward's distribution belongs to a *canonical exponential family* with respect to a reference measure ν, such that $d\mathbb{P}_\theta(r|a) = \exp(ra^\top\theta - b(a^\top\theta) + c(r))d\nu(r)$. Here, the function $c(\cdot)$ is real-valued and $b(\cdot)$ is assumed to be twice continuously differentiable. Thanks to the properties of exponential families, b is convex and can be related to the function $\mu = \dot{b}$, itself referred to as the *inverse link* or *mean* function. A key feature of this description is that given a ground-truth parameter $\theta^\star$, selecting an action a_t at time t yields a reward r_{t+1} conditionally independent on the past and such that $\mathbb{E}[r_{t+1}|a_t] = \mu(a_t^\top\theta^\star)$.

The non-stationary nature of the considered environments is characterized as follows: the bandit parameter $\theta^\star$ is allowed to change in an arbitrary fashion up to Γ_T times within the horizon T. In the following, $\theta^\star$ will be indexed by t to clearly exhibit its dependency w.r.t round t, and the reward signal will follow

$$\mathbb{E}[r_{t+1}|a_t] = \mu(a_t^\top\theta_t^\star) \ .$$

The focus of this paper is the *dynamic regret* defined as

$$R_T = \sum_{t=1}^{T} \max_{a\in\mathcal{A}_t} \mu\big(a^\top\theta_t^\star\big) - \mu\big(a_t^\top\theta_t^\star\big) \ .$$

Note that in this setting, there is no fixed best arm, both due to the non-stationarity of the environment and to the fact that the action set $\mathcal{A}_t$ may vary with time. We will work under the following assumptions.

Assumption 1 (Bounded actions and bandit parameters)**.**

$$\forall t \geq 1, \|\theta_t^\star\|_2 \leq S \quad \textit{and} \quad \forall a \in \mathcal{A}_t, \|a\|_2 \leq 1 \ .$$

We define the admissible parameter space $\Theta = \big\{\theta \in \mathbb{R}^d, \|\theta\|_2 \leq S\big\}$.

Assumption 2 (Bounded rewards)**.**

$$\exists m \in \mathbb{R}^+ \textit{such that} \, \forall t \geq 1, 0 \leq r_t \leq m \ .$$

Assumption 3. *The mean function $\mu : \mathbb{R} \mapsto \mathbb{R}$ is continuously differentiable, Lipschitz with constant k_μ and such that*

$$c_\mu = \inf_{\theta\in\Theta, \|a\|_2\leq 1} \dot{\mu}\big(a^\top\theta\big) > 0 \ .$$

The quantity c_μ is crucial in the analysis, as it represents the (worst case) sensitivity of the mean function. Our last assumption differs from most of existing works as we focus here on *self-concordant* GLMs. This assumption on the curvature of the mean function is rather mild, and covers for instance the logistic and Poisson models.

Assumption 4 (Generalized self-concordance)**.** *The mean function verifies $|\ddot{\mu}| \leq \dot{\mu}$.*

In order to estimate the unknown bandit parameter $\theta_t^\star$, we will adopt a *weighted* regularized maximum-likelihood principle. Formally, we define $\hat{\theta}_t$ for $\lambda > 0$ and $\gamma \in (0, 1]$ as the solution of the strictly convex program

$$\hat{\theta}_t = \underset{\theta\in\mathbb{R}^d}{\arg\min} - \sum_{s=1}^{t-1} \gamma^{t-1-s} \log \mathbb{P}_\theta(r_{s+1}|a_s) + \frac{\lambda}{2}\|\theta\|_2^2 \ . \tag{1}$$

Equivalently, $\hat{\theta}_t$ may be defined as the minimizer of $-\sum_{s=1}^{t-1}\gamma^{-s}\log\mathbb{P}_\theta(r_{s+1}|a_s) + \frac{\lambda\gamma^{-(t-1)}}{2}\|\theta\|_2^2$, with time-independent increasing weights γ^{-s} and time-varying regularization $\lambda\gamma^{-(t-1)}$, which is more handy for analysis purposes, see (Russac et al., 2019).

2.2 Stationary GLB

GLB were first considered in the seminal work of Filippi et al. (2010) who proposed `GLM-UCB`, an optimistic algorithm with a regret upper bound of the form $\tilde{\mathcal{O}}(c_\mu^{-1} d\sqrt{T})$. A key characteristic of `GLM-UCB` is a *projection step*, used to map the MLE onto the set of admissible parameters Θ. Formally, when the MLE $\hat{\theta}_t$ is not in Θ, it needs to be replaced by

$$\tilde{\theta}_t = \arg\min_{\theta \in \Theta} \left\| \sum_{s=1}^{t-1} \left[\mu\left(a_s^\top \theta\right) - \mu\left(a_s^\top \hat{\theta}_t\right) \right] a_s \right\|_{\mathbf{V_t}^{-1}} \tag{2}$$

where $\mathbf{V_t}$ is an invertible $d \times d$ square matrix.

With `GLM-UCB`, both the size of the confidence set (thus the exploration bonus) and the regret bound scale as c_μ^{-1}. However, this constant can be prohibitively large. In the cases of the logistic and Poisson bandits, one has $c_\mu^{-1} \geq e^S$, revealing an *exponential* dependency on S. If we consider the example of click prediction in online advertising with the logistic GLB, c_μ^{-1} is of the order 10^3, corresponding to typical click rates of less than a percent.

This critical dependency was addressed by Faury et al. (2020) for the logistic bandit. They introduce `LogUCB1` and `LogUCB2` for which they respectively prove $\widetilde{\mathcal{O}}(c_\mu^{-1/2} d\sqrt{T})$ and $\widetilde{\mathcal{O}}(d\sqrt{T} + c_\mu^{-1})$ regret upper bounds. Their analysis relies on the self-concordance property of the logistic log-likelihood. Self-concordance offers a refined way to control the curvature of the log-likelihood, and has been used in batch statistical learning (Bach, 2010) and online optimization (Bach and Moulines, 2013) (see also (Boyd and Vandenberghe, 2004, Section 9.6) for a broader picture). However, the analysis of Faury et al. (2020) does not use the self-concordance to its fullest and a projection step is still required, as detailed in Section 5.

Since the mean function μ can be non-convex (as for example in the case of logistic regression), the projection step defined in Equation (2) generally involves the minimization of a non-convex function. Solving this program can be arduous and finding ways to bypass it is desirable. This was achieved by Li et al. (2017) using a *burn-in phase* corresponding to an initial number of rounds during which the agent plays randomly. This ensures that $\hat{\theta}_t$ stays in Θ for subsequent rounds and therefore avoids the projection step. This technique was re-used in other recent works, such as (Kveton et al., 2020; Zhou et al., 2019). A major drawback of this approach however is the length of this burn-in phase, which typically grows with c_μ^{-2} (Kveton et al., 2020, Section 4.5). In the previously cited example of click-prediction, this would lead the agent to act randomly for approximately 10^6 rounds.

2.3 Forgetting in Non-Stationary Environments

Motivated by the non-stationary nature of most real-life applications of contextual bandits, a consequent theory for linear bandits in non-stationary environments has been recently developed (Cheung et al., 2019a; Russac et al., 2019; Zhao et al., 2020). We focus here on forgetting policies, a broader perspective is discussed in Section 5. In (Cheung et al., 2019a), a sliding window is used and the estimator is constructed based on the most recent observations only. In (Russac et al., 2019) exponentially increasing weights are used to give more importance to most recent observations. In (Zhao et al., 2020) the algorithm is restarted on a regular basis. These contributions were generalized to GLB by Russac et al. (2020); Cheung et al. (2019a); Zhao et al. (2020). However, the approach of Russac et al. (2020) still suffers from the aforementioned limitations (dependency w.r.t. c_μ and need for a projection step) while the analysis of both Cheung et al. (2019a) and Zhao et al. (2020) are missing key features of the problem at hand (see (Russac et al., 2020, Section 1)).

The non-stationary nature of the problem rules out the use of burning phases as changes in the GLB parameter can lead $\hat{\theta}_t$ to leave Θ, even when well initialized. This also accentuates the inconveniences brought by the projection step, as $\hat{\theta}_t$ leaving Θ is more likely to happen. This is why finding alternatives without projection is even more attractive in this particular setting. Furthermore, a generalization of the improvements brought by Faury et al. (2020) to non-stationary world is missing, and it is unclear if the dependency in c_μ can still be reduced in this harder setting.

2.4 Contributions

The present paper addresses these challenges, focusing on the use of exponential weights to adapt to changes in the model. First, we extend in Theorem 3 the Bernstein-like tail-inequality of (Faury et al., 2020, Theorem 1) to *weighted* self-normalized martingales. We then leverage the self-concordance property (Assumption 4) to provide an improved characterization of the maximum-likelihood estimator (Proposition 1). This allows to provide concentration guarantees *without* projecting $\hat{\theta}_t$ back to Θ. Combining these results leads to the `SC-D-GLUCB` strategy (Algorithm 1), which does not resort to a non-convex projection step and enjoys an $\tilde{\mathcal{O}}(c_\mu^{-1/3} d^{2/3} \Gamma_T^{1/3} T^{2/3})$ worst case regret upper bound (Theorem 2). A $\mathcal{O}(c_\mu^{-1/2} \Delta^{-1} d\sqrt{\Gamma_T T})$ regret bound is also obtained (Theorem 1) under an additional minimal gap $\Delta > 0$ assumption (Assumption 5). A summary of our contributions and comparison with prior work are given in Table 1.

Algorithm	Setting	Projection	Regret Upper Bound
GLM-UCB Filippi et al. (2010)	Stationary GLM	Non-convex	$\widetilde{\mathcal{O}}\left(\boldsymbol{c_\mu^{-1}} \cdot d \cdot \sqrt{T}\right)$
LogUCB1 Faury et al. (2020)	Stationary Logistic	Non-convex	$\widetilde{\mathcal{O}}\left(\boldsymbol{c_\mu^{-1/2}} \cdot d \cdot \sqrt{T}\right)$
D-GLUCB Russac et al. (2020)	Non-Stationary GLM	Non-convex	$\widetilde{\mathcal{O}}\left(\boldsymbol{c_\mu^{-1}} \cdot d^{2/3} \cdot \Gamma_T^{1/3} \cdot T^{2/3}\right)$
SC-D-GLUCB **(this paper)**	Non-Stationary GLM + SC + Ass. 5	**No projection**	$\widetilde{\mathcal{O}}\left(\boldsymbol{c_\mu^{-1/2}} \cdot d \cdot \sqrt{\Gamma_T T}\right)$
SC-D-GLUCB **(this paper)**	Non-Stationary GLM + SC	**No projection**	$\widetilde{\mathcal{O}}\left(\boldsymbol{c_\mu^{-1/3}} \cdot d^{2/3} \cdot \Gamma_T^{1/3} \cdot T^{2/3}\right)$

Table 1: Comparison of regret guarantees for different algorithms in the GLM setting with respect to the degree of non-linearity c_μ, the dimension d, the horizon T and the number Γ_T of abrupt changes. In the table SC stands for self-concordant. Regret guarantees for SC-SW-GLUCB are the same than for SC-D-GLUCB.

3 ALGORITHM AND RESULTS

3.1 Algorithms

In this section, we consider the abruptly changing environments defined in Section 2. We propose two algorithms: SC-D-GLUCB, which is based on discount factors, and SC-SW-GLUCB using a sliding window. Due to space limitation constraints, the pseudo-code of SC-SW-GLUCB and the corresponding theoretical results are reported in Appendix C. Associated with the weighed MLE defined in Equation (1), define the weighted design matrix as

$$\mathbf{V}_t = \sum_{s=1}^{t-1} \gamma^{t-1-s} a_s a_s^\top + \frac{\lambda}{c_\mu} \mathbf{I}_d \,. \tag{3}$$

The SC-D-GLUCB algorithm proceeds as follows. First, based on the previous rewards and actions, $\hat{\theta}_t$ is computed. After receiving the action set $\mathcal{A}_t$, the action a_t is chosen optimistically as the maximizer of the current estimate $\mu(a^\top \hat{\theta}_t)$ of each arm's reward inflated by the confidence bonus $c_\mu^{-1/2} \beta_T^\delta \|a\|_{\mathbf{V}_t^{-1}}$. Finally, the reward r_{t+1} is received and the matrix $\mathbf{V}_t$ is updated. The expression of β_T^δ is a consequence of our novel concentration result and is defined in Equation (4). A pseudo-code of the algorithm is presented in Algorithm 1.

There are two differences between SC-D-GLUCB and the algorithm proposed in Russac et al. (2020). First, we directly use $\hat{\theta}_t$ to make predictions about the arms' performances, whether it belongs to Θ or not. Second, the exploration term scales as $c_\mu^{-1/2}$ (instead of c_μ^{-1}), as in Faury et al. (2020). The latter has a direct impact on the regret-bound of SC-D-GLUCB, to be stated below.

Algorithm 1 SC-D-GLUCB

Input: Probability δ, dimension d, regularization λ, upper bound for bandit parameters S, discount factor γ.
Initialize: $\mathbf{V}_0 = (\lambda/c_\mu)\mathbf{I}_d$, $\hat{\theta}_0 = 0_{\mathbb{R}^d}$.
for $t = 1$ **to** T **do**
  Receive $\mathcal{A}_t$, compute $\hat{\theta}_t$ according to (1)
  Play $a_t = \arg\max_{a \in \mathcal{A}_t} \mu(a^\top \hat{\theta}_t) + \frac{\beta_T^\delta}{\sqrt{c_\mu}} \|a\|_{\mathbf{V}_t^{-1}}$ with β_T^δ defined in Equation (4)
  Receive reward r_{t+1}
  Update: $\mathbf{V}_{t+1} \leftarrow a_t a_t^\top + \gamma \mathbf{V}_t + \frac{\lambda}{c_\mu}(1-\gamma)\mathbf{I}_d$
end for

3.2 Regret Upper Bounds

We detail in this section the performance guarantees for SC-D-GLUCB. Define

$$\beta_T^\delta = k_\mu \sqrt{\lambda} \left(1 + \bar{S} + \sqrt{\frac{1+\bar{S}}{\lambda}} \rho_T^\delta + \left(\frac{\rho_T^\delta}{\sqrt{\lambda}}\right)^2\right)^{3/2} \tag{4}$$

with

$$\bar{S} = S + \frac{2Sk_\mu + m}{T\lambda(1-\gamma)} \,, \tag{5}$$

and where

$$\rho_T^\delta = \frac{\sqrt{\lambda}}{2m} + \frac{2m}{\sqrt{\lambda}} \log\left(\frac{T}{\delta}\right) + \frac{2m}{\sqrt{\lambda}} d \log(2) + \frac{dm}{\sqrt{\lambda}} \log\left(1 + \frac{k_\mu(1-T^{-2})}{d\lambda(1-\gamma^2)}\right) \,.$$

The latter expression is a direct consequence of the concentration result presented in Theorem 3 below. The difference between $\bar{S}$ and S is a bias term due to the non-stationarity.

Before stating our first theorem, we add an additional assumption on the minimal gap. This assumption is discussed in Section 5 and is only used in Theorem 1.

Assumption 5. *The reward gaps* $\Delta_t = \min_{a \in \mathcal{A}_t, \mu(a^\top \theta_t^\star) < \mu(a_\star^\top \theta_t^\star)} \mu(a_\star^\top \theta_t^\star) - \mu(a^\top \theta_t^\star)$ *satisfies*

$$\forall t \leq T, \Delta_t \geq \Delta > 0 \, .$$

Theorem 1. *Under Assumption 5, the regret of the* SC-D-GLUCB *algorithm is bounded for all* $\gamma \in (1/2, 1)$ *with probability at least* $1 - \delta$ *by*

$$R_T \leq C_1 \frac{\Gamma_T}{1-\gamma} + C_2 \frac{1}{T(1-\gamma)^2 \Delta} + C_3 \frac{\beta_T^\delta \sqrt{dT}}{\sqrt{c_\mu}\Delta} \sqrt{T \log(1/\gamma) + \log\left(1 + \frac{1}{d\lambda(1-\gamma)}\right)} + C_4 \frac{d(\beta_T^\delta)^2}{c_\mu \Delta} \left(T \log(1/\gamma) + \log\left(1 + \frac{1}{d\lambda(1-\gamma)}\right)\right) ,$$

where C_1, C_2, C_3, C_4 *are universal constants independent of* c_μ, γ *with only logarithmic terms in* T.

In particular, setting $\gamma = 1 - \frac{\sqrt{c_\mu \Gamma_T}}{d\sqrt{T}}$ *and* $\lambda = d \log(T)$ *leads to*

$$R_T = \widetilde{\mathcal{O}}\big(\Delta^{-1} c_\mu^{-1/2} d \sqrt{\Gamma_T T}\big) \, .$$

There is a strong link between the cost of non-stationarity in the K-arm setting and the one observed in the more general GLB setting. In the K-arm setting, any sub-optimal arm i is played at most $\mathcal{O}(\Delta_i^{-2} \log(T))$ times (e.g (Munos, 2014, Proposition 1.1)), whereas in any abruptly changing environment, forgetting policies play a sub-optimal arm i at most $\widetilde{\mathcal{O}}((\Delta_T(i))^{-2}\sqrt{\Gamma_T T})$ (Garivier and Moulines, 2011). $\Delta_T(i)$ is the minimum distance between the mean of the optimal arm and the mean of the suboptimal arm i over the entire time horizon. For GLBs, in the stationary case Filippi et al. (2010, Theorem 1) give a gap-dependent bound on the regret scaling as $\mathcal{O}(\Delta^{-1} c_\mu^{-2} d^2 \log(T))$. Here, the bound of Theorem 1 is of order $\mathcal{O}(\Delta^{-1} c_\mu^{-1/2} d\sqrt{\Gamma_T T})$. The reduced dependency in c_μ in the latter bound is a direct consequence of the use of self-concordance. Also note that when the inverse link function is the identity and the action set is the canonical basis, our analysis recovers the results of Garivier and Moulines (2011).

We give an upper bound for the worst case regret of Algorithm 1 in the following theorem; its proof is deferred to the appendix.

Theorem 2. *The regret of the* SC-D-GLUCB *algorithm is bounded for all* $\gamma \in (1/2, 1)$ *with probability at least* $1 - \delta$ *by*

$$R_T \leq C_1 \frac{\Gamma_T}{1-\gamma} + C_2 \frac{\beta_T^\delta \sqrt{dT}}{\sqrt{c_\mu}} \sqrt{T \log\left(\frac{1}{\gamma}\right) + \log\left(1 + \frac{1}{d\lambda(1-\gamma)}\right)} ,$$

where C_1 *and* C_2 *are universal constants independent of* c_μ *and* γ *with only logarithmic terms in* T.

In particular, setting $\gamma = 1 - \left(\frac{c_\mu^{1/2}\Gamma_T}{dT}\right)^{2/3}$ *and* $\lambda = d \log(T)$ *leads to*

$$R_T = \widetilde{\mathcal{O}}\big(c_\mu^{-1/3} d^{2/3} \Gamma_T^{1/3} T^{2/3}\big) \, .$$

As in the linear case, this regret bound highlights the existence of two mechanisms of different nature. The first term is due to non-stationarity, the number of changes Γ_T being multiplied by $1/(1-\gamma)$, which is a rough measure of the forgetting time induced by the exponential weights. The second term characterizes the rate at which the weighted MLE $\hat{\theta}_t$ approaches $\theta_t^\star$. By balancing both terms, we can characterize the asymptotic behavior of the regret bound.

In Theorem 2, optimally tuning γ yields the asymptotic worst case rate of $T^{2/3}$. This is similar to the asymptotic rate achievable in the linear case with a different measure of non-stationarity (Russac et al., 2019) and the same dependency is attained with a sliding window for MDPs in abruptly changing environments (Gajane et al., 2018) and with restart factors (Auer et al., 2008).

Remark 1. *The proof of Theorem 2 reveals that for rounds* t *where* $\hat{\theta}_t$ *lies in* Θ, *it is possible to obtain a (usually) tighter concentration result (depending on the values of* λ *and* S*) by replacing* β_T^δ *with* $k_\mu \sqrt{1 + 2S}(\sqrt{\lambda} S + \rho_T^\delta)$. *This cannot be used to improve the result of Theorem 2, as one doesn't know in advance for which rounds the condition will be satisfied, but this minor modification of Algorithm 1 is most often advisable in practice. See Section B.4 in Appendix for more details.*

4 KEY ARGUMENTS

In this section, we detail some key elements of our analysis. First, we describe the concentration result in its most generic form. Then, we explain the main steps to derive the upper bound of the regret of SC-D-GLUCB.

4.1 A Tail-Inequality for Self-Normalized Weighted Martingales

To reduce the dependency in c_μ, it is essential to take into account the actual conditional variance of the generalized linear model (Faury et al., 2020). With exponentially increasing weights, we also need time-dependent regularization parameters to avoid a vanishing effect of the regularization (Russac et al., 2019). Carefully combining these two elements yields the following concentration result.

Theorem 3. *Let t be a fixed time instant. Let $\{\mathcal{F}_u\}_{u=1}^t$ be a filtration. Let $\{a_u\}_{u=1}^t$ be a stochastic process on $\mathbb{R}^d$ such that a_u is $\mathcal{F}_u$ measurable and $\|a_u\|_2 \leq 1$. Let $\{\epsilon_u\}_{u=2}^t$ be a martingale difference sequence such that ϵ_{u+1} is $\mathcal{F}_{u+1}$ measurable. Assume that the weights are non-decreasing, strictly positive and the time horizon is known. Furthermore, assume that conditionally on $\mathcal{F}_u$ we have $|\epsilon_{u+1}| \leq m$ a.s. Let $\{\lambda_u\}_{u=1}^t$ be a deterministic sequence of regularization terms and denote $\sigma_t^2 = \mathbb{E}\left[\epsilon_{t+1}^2 | \mathcal{F}_t\right]$.*

Let $\widetilde{\mathbf{H}}_t = \sum_{s=1}^{t-1} w_s^2 \sigma_s^2 a_s a_s^\top + \lambda_{t-1}\mathbf{I}_d$ and $S_t = \sum_{s=1}^{t-1} w_s \epsilon_{s+1} a_s$, then for any $\delta \in (0,1]$,

$$\|S_t\|_{\widetilde{\mathbf{H}}_t^{-1}} \geq \frac{\sqrt{\lambda_{t-1}}}{2mw_{t-1}} + \frac{2mw_{t-1}}{\sqrt{\lambda_{t-1}}} \log\left(\frac{\det(\widetilde{\mathbf{H}}_t)^{1/2}}{\delta \lambda_t^{d/2}}\right) + \frac{2mw_{t-1}}{\sqrt{\lambda_{t-1}}} d\log(2)$$

with probability smaller than δ.

4.2 Upper Bounding the Regret of SC-D-GLUCB

In a non-stationary environment, each change in the parameter will necessarily result in a number of rounds where the bias of the weighted MLE estimator cannot be controlled. This gives rise to the first term in the upper bound in Theorem 2. To make this observation more explicit, for $D \geq 1$, define $\mathcal{T}(\gamma) = \{1 \leq t \leq T,$ such that $\theta_s^\star = \theta_t^\star$ for $t - D \leq s \leq t-1\}$ the set of time instants that are at least D steps away from the previous closest breakpoint. Central in the analysis of weighted GLBs is the matrix

$$\mathbf{G}_t(\hat{\theta}_t, \theta_t^\star) = \sum_{s=1}^{t-1} \gamma^{t-1-s} \alpha(a_s, \hat{\theta}_t, \theta_t^\star) a_s a_s^\top + \lambda \mathbf{I}_d,$$

where

$$\alpha(a_s, \hat{\theta}_t, \theta_t^\star) = \int_0^1 \dot{\mu}(a_s^\top((1-v)\theta_t^\star + v\hat{\theta}_t))dv.$$

As in the linear case, we define its analogue with squared exponential weights,

$$\widetilde{\mathbf{G}}_t(\hat{\theta}_t, \theta_t^\star) = \sum_{s=1}^{t-1} \gamma^{2(t-1-s)} \alpha(a_s, \hat{\theta}_t, \theta_t^\star) a_s a_s^\top + \lambda \mathbf{I}_d\,.$$

We add the subscript $t-D:t$ to a quantity when the sum is for time instants between $t-D$ and $t-1$. In this subsection, for space constraints, we will denote equivalently $\widetilde{\mathbf{G}}_t(\hat{\theta}_t, \theta_t^\star)$ (resp. $\mathbf{G}_t(\hat{\theta}_t, \theta_t^\star)$) by $\widetilde{\mathbf{G}}_t$ (resp. $\mathbf{G}_t$). As for linear bandits, the exploration bonus is designed to mitigate the impact of prediction errors. We focus below on upper bounding the prediction error in $\hat{\theta}_t$ defined as $\Delta_t(a, \hat{\theta}_t) = |\mu(a^\top \hat{\theta}_t) - \mu(a^\top \theta_t^\star)|$. The exact link between the regret and this quantity is made explicit in Proposition 9 in the appendix. By defining $g_t(\theta) = \sum_{s=t-D}^{t-1} \gamma^{t-1-s} \mu(a_s^\top \theta) a_s + \lambda\theta$, when $t \in \mathcal{T}(\gamma)$ one can upper bound the prediction error in $\hat{\theta}_t$.

$$\Delta_t(a, \hat{\theta}_t) \leq \frac{c\gamma^D}{1-\gamma} + k_\mu \underbrace{\|g_t(\hat{\theta}_t) - g_t(\theta_t^\star)\|_{\widetilde{\mathbf{G}}_{t-D:t}^{-1}}}_{①} \underbrace{\|a\|_{\mathbf{G}_t^{-1}}}_{②}$$

The first term corresponds to the bias due to non-stationarity. ① is a measure of the deviation of $\hat{\theta}_t$ from $\theta_t^\star$ adapted to the non-linear nature of the problem. Note that $g_t(\hat{\theta}_t) - g_t(\theta_t^\star)$ involves a martingale difference sequence (thanks to the optimality condition of the MLE) that can be controlled using Theorem 3. However, to bound ① using Theorem 3 one needs to link the matrix $\widetilde{\mathbf{G}}_{t-D:t}$ with $\widetilde{\mathbf{H}}_{t-D:t}$, the self-concordance allows exactly to do this.

Self-Concordance More precisely, the use of self-concordance offers a sharp relation (independent of c_μ) between the first derivative of the mean function evaluated at different points. Using Lemma 4 reported in Appendix D, standard calculations yield:

$$\widetilde{\mathbf{G}}_{t-D:t} \geq \left(1 + C + \frac{1}{\sqrt{\lambda}} \|g_t(\hat{\theta}_t) - g_t(\theta_t^\star)\|_{\widetilde{\mathbf{G}}_{t-D:t}^{-1}}\right) \widetilde{\mathbf{H}}_{t-D:t} \tag{6}$$

Note that Equation (6) involves the deviation term that we want to control. Here, C is a residual bias due to the non-stationarity of the environment.

Better Characterization of the MLE By leveraging Equation (6) to bound the deviation $g_t(\hat{\theta}_t) - g_t(\theta_t^\star)$ in the $\widetilde{\mathbf{G}}_{t-D:t}^{-1}$-norm, one obtains an implicit equation. Solving it leads to the following proposition.

Proposition 1. *When $t \in \mathcal{T}(\gamma)$, the following holds,*

$$\|g_t(\hat{\theta}_t) - g_t(\theta_t^\star)\|_{\widetilde{\mathbf{G}}_{t-D:t}^{-1}(\hat{\theta}_t, \theta_t^\star)} \leq \sqrt{1+C}\rho_T^\delta + \frac{1}{\sqrt{\lambda}}\left(\rho_T^\delta\right)^2\,,$$

where C is a residual term due to non-stationarity.

Remark. *In stark contrast with previously existing works (see (Filippi et al., 2010, Proposition 1)), deviations from the true parameter $\theta_t^\star$ are characterized uniquely by the MLE (and not by its projected counterpart). This can be done whether $\hat{\theta}_t$ belongs to Θ or not and without any projection. This is not specific to the non-stationary nature of the problem but fundamentally relies on an improved analysis of the MLE. Similar guarantees can be obtained in any stationary environment. See Section 5 for a more detailed comparison of the possible uses of the self-concordance property.*

① can be upper bounded using Proposition 1. To

upper bound ② we use the following inequality.

$$\mathbf{G}_t \geq \left(1 + C + \frac{1}{\sqrt{\lambda}}\|g_t(\hat{\theta}_t) - g_t(\theta_t^\star)\|_{\widetilde{\mathbf{G}}_{t-D:t}^{-1}}\right)^{-1} c_\mu \mathbf{V}_t \ . \tag{7}$$

Combining Proposition 1 with Equation (7) gives the upper bound for ②. Putting everything together, we obtain the form of β_T^δ given in Equation (4). The regret bound is then obtained by summing the exploration bonus for the different time instants. Applying the so-called elliptical lemma (see (Lattimore and Szepesvári, 2019, Chap. 19)) and letting $D = \log(T)/\log(1/\gamma)$ completes the proof.

5 DISCUSSION

Assumption on the Gaps. Assumptions similar to our Assumption 5 requiring a minimum gap are frequent in non-stationary bandits. First, note that Δ is not required for the algorithm but only for the theoretical analysis. Second, similar assumptions can be found for K-arm bandits in several works to obtain the optimal $\widetilde{\mathcal{O}}(\sqrt{\Gamma_T T})$ regret bound. This is in particular the case for change-points detection methods: (Cao et al., 2019, Corollary 1) and (Zhou et al., 2020, Corollary 4.3) is proved under an assumption on the minimal gap. This remains true for forgetting strategies: the bound of Garivier and Moulines (2011) is gap-dependent, Trovo et al. (2020) achieve a $\mathcal{O}(\Delta^{-1}\sqrt{T\Gamma_T})$ regret. More demanding, the LM-DSEE and SW-UCB# algorithms from Wei and Srivatsva (2018) require the minimum gap as an input of the algorithm. Generally speaking, none of those works provide an analysis when the minimum gap can depend on the time horizon T and when the mean of different arms can be arbitrarily close. We suspect that forgetting policies would obtain a $\mathcal{O}(\Gamma_T^{1/3}T^{2/3})$ worst case dependency as in Theorem 2 and that changepoint detection methods are likely to fail in such a case.

Tightness of the Bound. For problems with a finite number of actions, Auer et al. (2018) have developed an algorithm that does not require the knowledge of the number of breakpoints nor assumption on the gaps. This was extended to the K-arm setting by Auer et al. (2019) and to the more general contextual bandits by Chen et al. (2019). Both works (Auer et al. (2019); Chen et al. (2019)) achieve the optimal $\widetilde{\mathcal{O}}(\sqrt{\Gamma_T T})$ regret bound. Yet, their analysis does not apply to the GLB framework. Furthermore, both works rely on replaying phases that are incompatible with time-dependent action sets as considered here. Additionally, in (Chen et al., 2019) the regret is defined with respect to the best policy in some finite class, whereas our results apply to the general setting where actions can change over time and the regret benchmark is the ground-truth of the environment. The best lower-bound for forgetting policies in abruptly changing environments with time-dependent action sets remains unknown. While it is known that forgetting policies are minimax optimal when non-stationarity is measured through the so-called variational budget (see Cheung et al. (2019b); Russac et al. (2019)), whether such methods are optimal in abruptly changing environments is unclear. Nonetheless, the bound obtained by Garivier and Moulines (2011) in the K-arm setting yields a worst case regret bound that can be shown to be of order $\mathcal{O}(\Gamma_T^{1/3}T^{2/3})$ (see Appendix E).

Knowledge of Γ_T Optimizing the choice of the forgetting parameter γ (w.r.t. the regret bound) requires the knowledge of Γ_T. The Bandit over Bandit (BOB) framework introduced by Cheung et al. (2019b) can be used to circumvent this requirement. When the assumption 5 is satisfied, following the proof from Cheung et al. (2019a) one would obtain a regret bound of order $\widetilde{\mathcal{O}}(\Delta^{-1}dc_\mu^{-1/2}\sqrt{T\max(\Gamma_T, T^{1/2})})$ (see (Auer et al., 2019, Remark 2)). Similarly, in the absence of Assumption 5 an upper bound of order $\widetilde{\mathcal{O}}(c_\mu^{-1/3}d^{2/3}T^{2/3}\max(\Gamma_T, d^{-1/2}T^{1/4})^{1/3})$ can be achieved (see (Zhao et al., 2020, Theorem 4)).

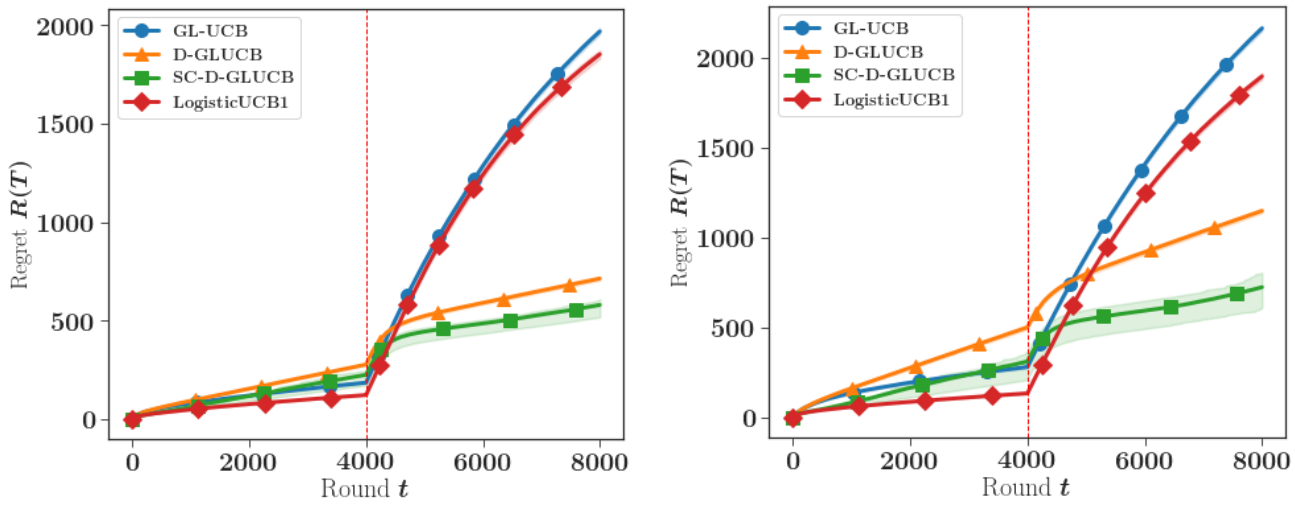
Self-Concordance The analysis of Faury et al. (2020) does not use self-concordance to its fullest. We present an improved analysis valid in any stationary time frame, proving that a better treatment of the self-concordance removes the need for the inconvenient projection. Informally, the self-concordance links $\mu(x^\top\hat{\theta}_t)$ to $\mu(x^\top\theta^\star)$ without resorting to global bounds on $\dot{\mu}$ (e.g k_μ and c_μ). In Faury et al. (2020), this takes the form of a Taylor-like expansion:

$$\mu(x^\top\theta_t) \leq \mu(x^\top\theta^\star) + \frac{|x^\top(\theta^\star - \theta_t)|}{1+2S}\dot{\mu}(x^\top\theta^\star) \ ,$$

where θ_t is a projected version of $\hat{\theta}_t$ in Θ. The denominator of the r.h.s. is reminiscent of this projection step. We show here that a finer analysis yields the following, more implicit but powerful bound:

$$\mu(x^\top\hat{\theta}_t) \leq \mu(x^\top\theta^\star) + \frac{|x^\top(\theta^\star - \hat{\theta}_t)|}{1+|x^\top(\theta^\star - \hat{\theta}_t)|}\dot{\mu}(x^\top\theta^\star) \ .$$

Note that when $\hat{\theta}_t \in \Theta$ (i.e there is no need for a projection), our bound implies the one of Faury et al. (2020). The kind of relationship displayed in the above equation allows us to derive a tail inequality for the deviation from $\hat{\theta}_t$ to $\theta^\star$ without projecting $\hat{\theta}_t$, by solving an implicit equation. We believe that this new approach is of interest in other settings involving self-concordant



(a) $c_\mu^{-1} = 400$ (b) $c_\mu^{-1} = 1000$

Figure 1: Regret of the different algorithms in a 2D abruptly changing environment averaged on 200 independent experiments and the 25% associated quantiles.

GLBs. The self-concordance assumption (Assumption 4) is not particularly restrictive and goes beyond logistic functions. Under the classical Assumption 1 (i.e. bounded features) all GLMs are self-concordant (cf. Sec. 2 of Bach (2014)) with constants that depend on the link function.

6 EXPERIMENTS

In this section, we illustrate the empirical performance of `SC-D-GLUCB` in a simulated, abruptly changing environment with a logistic link function $\mu(x) = 1/(1 + \exp(-x))$. In this two-dimensional problem, there is a switch in the reward distribution at $t = 4000$ (red dashed line on Figure 1).

`SC-D-GLUCB` (Algorithm 1) is compared with `GLM-UCB` from Filippi et al. (2010), `LogUCB1` from Faury et al. (2020) and with `D-GLUCB` from Russac et al. (2020). `SC-D-GLUCB` (resp. `D-GLUCB`) is related with `LogUCB1` (resp. `GLM-UCB`) in the sense that the exploration terms have the same scaling but the former incorporate the exponential weights making it possible to adapt to changes. The average regret of the different policies together with their central 50% quantiles, averaged on 200 independent runs, are reported in Figure 1 for two different parameter values.

In Fig. 1a, $\theta^\star$ starts on the circle of radius $S = 6$ (corresponding to $c_\mu^{-1} = \exp(S) \approx 400$) with an angle of $2\pi/3$ and jumps at $t = 4000$ to an angle of $4\pi/3$. The experiment reported on Fig. 1b is identical with a radius $S = 7$ corresponding to a $c_\mu^{-1} \approx 1000$. As previously discussed, using such values of S is required in situation where the actions return binary rewards with expected values in the range 10^{-3} – 10^{-2}, which is typically the case in web advertising or recommendation applications.

For both experiments, at every time steps, 50 randomly generated actions in the unit circle are proposed to the learner. For `SC-D-GLUCB` and `D-GLUCB` the asymptotically optimal choice of the discount factors is used: $\gamma = 1 - (\Gamma_T/(d \times T))^{2/3}$ with $d = 2$, $\Gamma_T = 2$ and $T = 8000$. To speed up the learning that is hard with those values of c_μ, all the algorithms have their exploration bonus divided by 5.

As expected, the algorithms tuned for non stationary situations (`SC-D-GLUCB`, `D-GLUCB`) perform worse than their stationary counterparts (`LogUCB1` and `GLM-UCB`) during the first stationary phase. More precisely, with the choice made for γ the estimation of $\hat{\theta}_t$ for algorithms that use exponential weights is roughly based on the $1/(1-\gamma) \approx 400$ most recent observations. In contrast, `LogUCB1` and `GLM-UCB` use all the observations from the start to compute the MLE, which eventually leads to a more precise estimation. Right after the change, the bias caused by the non-stationarity results in a significant increase in regret. Unweighted algorithms are affected much more deeply by this phenomenon that will eventually cause large losses in performance due to the persistence of obsolete information.

The theoretical analysis of Section 3.2 suggests that the advantage of `SC-D-GLUCB` is all the more significant in strongly non-linear (large c_μ^{-1}) non-stationary environments. This is obvious in Figure 1, particularly when comparing Fig. 1a and Fig. 1b, which differ by the range on which the logistic function is used for

making reward predictions. Note that, on average, for these two simulated scenarios the fact that the MLE $\hat{\theta}_t$ does not belong to Θ happens for several hundred of rounds. All the algorithms except SC-D-GLUCB would require non convex projection steps at these instants, or equivalently, one should inflate S (and thus c_μ^{-1}) to ensure the compliance of these algorithms with the associated theory. In producing Figure 1, this projection step was simply bypassed, which provides an optimistic evaluation of the performance of the competitors of SC-D-GLUCB. Interestingly, the observation that the dispersion of performance of SC-D-GLUCB is slightly higher than that of D-GLUCB can be traced back to the use of Remark 1 in these simulations: SC-D-GLUCB adapts to the events $\{\hat{\theta}_t \notin \Theta\}$ (rather than pretending that these did not happen) and thus its performance is made somewhat dependent on the actual occurrence of these events.

7 CONCLUSION

In this paper, we design GLB algorithms for piecewise stationary environments by resorting to forgetting mechanisms. We improve existing solutions by circumventing important drawbacks affecting their applicability in real-life scenarios. More precisely, under a generic self-concordance assumption, we remove the need for burdensome non-convex projections and leverage refined and exponentially deflated confidence regions. At the heart of our approach are a refined characterization of the maximum-likelihood estimator and an extension of a Bernstein-like tail inequality to *weighted* self-normalized martingales. We believe that both can be of independent interest and leveraged in other settings (e.g drifting environments). We can see two natural extensions of our work; the first involves achieving similar success for other nature of non-stationarity such as drifting environments. The second could try to replicate the recent progress of Abeille et al. (2020) in the stationary logistic case, and provide minimax rates w.r.t c_μ in non-stationary self-concordant GLBs.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems, NeurIPS 2011*, pages 2312–2320, 2011.

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. *arXiv preprint arXiv:2010.12642*, 2020.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in neural information processing systems, NeurIPS*, pages 89–96, 2008.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best arm with an unknown number of distribution changes. In *European Workshop on Reinforcement Learning, EWRL 2018*, 2018.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory, COLT 2019*, pages 138–158, 2019.

Francis Bach. Self-concordant analysis for logistic regression. *Electron. J. Statist.*, 4:384–414, 2010. doi: 10.1214/09-EJS521.

Francis Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(19):595–627, 2014.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in Neural Information Processing Systems, NeurIPS 2013*, pages 773–781, 2013.

S.P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge Univ Press, 2004.

Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, 2019.

O. Chapelle and L. Li. An empirical evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems, NeurIPS 2011*, 2011.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal, and parameter-free. *Proceedings of the 32nd Conference on Learning Theory, COLT 2019*, 2019.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Hedging the drift: Learning to optimize under non-stationarity. *arXiv preprint arXiv:1903.01461*, 2019a.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, 2019b.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011*, pages 208–214, 2011.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feed-

back. In *21st Annual Conference on Learning Theory, COLT 2008*, pages 355–366, 2008.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, 2020.

Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems, NeurIPS 2010*, pages 586–594, 2010.

Arthur Flajolet and Patrick Jaillet. Real-time bidding with side information. In *Advances in Neural Information Processing Systems, NeurIPS 2017*, pages 5168–5178, 2017.

Pratik Gajane, Ronald Ortner, and Peter Auer. A sliding-window algorithm for markov decision processes with arbitrarily changing rewards and transitions. *arXiv preprint arXiv:1805.10066*, 2018.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory, ALT 2011*, pages 174–188, 2011.

Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. *Proceedings of the 23nd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*, 2020.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2019.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *WWW*, 2010.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, pages 2071–2080, 2017.

Rémi Munos. *From bandits to Monte-Carlo Tree Search: The optimistic principle applied to optimization and planning*. 2014.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, pages 395–411, 2010.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems, NeurIPS 2019*, pages 12017–12026, 2019.

Yoan Russac, Olivier Cappé, and Aurélien Garivier. Algorithms for non-stationary generalized linear bandits. *arXiv preprint arXiv:2003.10113*, 2020.

Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer, 2017.

Francesco Trovo, Stefano Paladino, Marcello Restelli, and Nicola Gatti. Sliding-window thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research*, 68:311–364, 2020.

Michal Valko, Rémi Munos, Branislav Kveton, and Tomáš Kocák. Spectral bandits for smooth graph functions. In *Proceedings of the 31st International Conference on Machine Learning, ICML 2014*, pages 46–54, 2014.

Lai Wei and Vaibhav Srivatsva. On abruptly-changing and slowly-varying multiarmed bandit problems. In *2018 Annual American Control Conference (ACC)*, pages 6291–6296. IEEE, 2018.

Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*, 2020.

Huozhi Zhou, Lingda Wang, Lav R Varshney, and Ee-Peng Lim. A near-optimal change-detection based algorithm for piecewise-stationary combinatorial semi-bandits. *AAAI*, 2020.

Zhengyuan Zhou, Renyuan Xu, and Jose Blanchet. Learning in generalized linear contextual bandits with stochastic delays. In *Advances in Neural Information Processing Systems, NeurIPS 2019*, 2019.