Algorithms for Non-Stationary Generalized Linear Bandits

Yoan Russac, Olivier Cappé, Aurélien Garivier

Algorithms for Non-Stationary Generalized Linear Bandits

Yoan Russac[1], Olivier Cappé[1], Aurélien Garivier[2]
[1] DI ENS, CNRS, Inria, ENS, Universit PSL; [2] UMPA, CNRS, Inria, ENS Lyon

Abstract

The statistical framework of Generalized Linear Models (GLM) can be applied to sequential problems involving categorical or ordinal rewards associated, for instance, with clicks, likes or ratings. In the example of binary rewards, logistic regression is well-known to be preferable to the use of standard linear modeling. Previous works have shown how to deal with GLMs in contextual online learning with bandit feedback when the environment is assumed to be stationary. In this paper, we relax this latter assumption and propose two upper confidence bound based algorithms that make use of either a sliding window or a discounted maximum-likelihood estimator. We provide theoretical guarantees on the behavior of these algorithms for general context sequences and in the presence of abrupt changes. These results take the form of high probability upper bounds for the dynamic regret that are of order $d^{2/3}\Gamma_T^{1/3}T^{2/3}$, where d, T and Γ_T are respectively the dimension of the unknown parameter, the number of rounds and the number of breakpoints up to time T. The empirical performance of the algorithms is illustrated in simulated environments.

1 Introduction

The multi-armed bandit model is a well-known abstraction of the exploration-exploitation dilemma that occurs whenever predictions need to be made while learning a parameter of interest. When contextual information is available, a popular framework is the stochastic linear model (Dani et al., 2008; Li et al., 2010; Rusmevichientong and Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011), where the reward observed at each round is a noisy version of a linear combination of the contextual features that describe the selected action.

More precisely, we assume that at time t a set of contextual actions $\mathcal{A}_t \subset \mathbb{R}^d$ is available. Based on previous choices and rewards, the learner selects one of them and observes the associated reward. The learner's goal is to maximize the accumulated rewards. The particularity of the bandit setting is that the learner does not know the reward she would have obtained by selecting another action. In a recommendation setting, the contextualized actions may for instance combine information about both the users and the products to be recommended. By selecting the action A_t, a noisy version of $A_t^\top \theta^\star$ is observed, where $\theta^\star$ is an unknown parameter associated with the environment.

The Generalized Linear Model (GLM) setting (Filippi et al., 2010; Li et al., 2017) extends this model by assuming that conditionally on A_t, the learner observes a noisy version of $\mu(A_t^\top \theta^\star)$, where μ is a non-linear mapping, referred to as the inverse link or mean function. More details on the probabilistic structure of GLMs are given in Section 2. A particular case of great practical interest occurs when μ is the logistic function, which is the dominant approach for regression modeling with binary outcomes.

The classical bandit framework assumes stationarity of the environment parameter $\theta^\star$. This is clearly unrealistic in many potential applications. In news recommendation for instance, as considered by (Li et al., 2010), it has been consistently observed that the intrinsic interest in news stories is a decreasing function of time to original publication date. But, on the other hand, infrequent increases in interest for older items can also be triggered by the publication of fresh news. Regularly restarting the learning algorithm is a (frequent) basic approach to mitigate this issue. However, there is also a strong interest for developing bandit approaches that are inherently robust to possible changes in the environment. The aim of this work is to propose and analyze methods that achieve this goal in contextual bandits based on GLMs (which we shall refer to as "generalized linear bandits").

Related Work. Two types of approaches are generally adopted to deal with non-stationarity. The first one consists in detecting changes in distribution (Auer et al., 2018; Besson and Kaufmann, 2019) and restarting the algorithm whenever a change is detected. The second one builds progressively forgetting policies (Garivier and Moulines, 2011) based either on the use of a sliding window –computing the estimator only on the most recent observations–, or, on the use of exponentially increasing weights to reduce the influence of past observations. Both approaches have been studied in the K-armed and linear settings. In linear bandits, Wu et al. (2018) build a pool of plausible models to make recommendations. When no model satisfies a given statistical test, a change point is declared and a new model is added to the pool. In (Cheung et al., 2019b) the sliding window approach is used to build the least squares estimator. In (Russac et al., 2019) the past is progressively forgotten with the use of a discount factor that gives more weights to recent observations and the estimator is defined through

weighted least squares.

Assessing the performance of these methods, requires quantified measures of non-stationarity and here again there are several options. The notion of *variation budget* that includes both slowly and abruptly changing environments was considered in the K-armed bandit setting by Besbes et al. (2014) and in the linear setting by Cheung et al. (2019b) for example. In this work, as in –among others– (Garivier and Moulines, 2011; Liu et al., 2018; Cao et al., 2019), we focus on abruptly changing environments, and measure non-stationarity by the number of breakpoints up to time T.

GLMs with bandit feedback were studied by Filippi et al. (2010) with a fixed actions set; the authors proposed a first UCB algorithm in this setting. Our work extends theirs to the case where the preference parameters $\theta^\star$ can dynamically evolve over time. We stress that their analysis assumed static actions whereas ours also works with time dependent actions sets. No regularization term was used in Filippi et al. (2010) implying unsatisfactory initialization assumptions that we were able to remove by considering a penalized estimator.

Another analysis was proposed by Li et al. (2017), where, in contrast to our work, statistical assumptions are made on the distributions of the contextual vectors, allowing the use of results from random matrix theory for establishing concentration inequalities. We work in the more general framework where the available actions at each round can even be chosen by an adversary.

Randomized algorithms have also been developed to study generalized linear bandits. The extension of Thompson Sampling to this setting was analyzed by Abeille et al. (2017) and a $O(d^{3/2}\sqrt{T\log(K)})$ regret bound valid for infinite actions sets was derived. In (Kveton et al., 2020) two others randomized algorithms are proposed. One method consists in fitting a GLM on a randomly perturbed history of the past rewards to guarantee sufficient exploration. The second method consists in sampling a GLM from the Laplace approximation to the posterior distribution. In a d-dimensional problem with K fixed actions the T rounds regret of those methods is of order $O(d\sqrt{\log(K)T})$. Both methods in (Kveton et al., 2020) assume a static actions set and have a logarithmic dependence in the number of actions. In contrast, the upper-bounds on the regret that we obtain do not depend on the number of available actions.

The regret of most existing algorithms for generalized linear bandits is inversely proportional to the minimum value of the derivative of the inverse link function. This quantity can be large (as in the logistic model), and hence designing policies that do not depend on this quantity is of particular interest. In the particular case of logistic bandits under strong assumptions on the features and $\theta^\star$, Dong et al. (2019) propose a first Bayesian analysis that does not depend on this quantity. However, the analysis of Dong et al. (2019) relies on specifics of the logistic model and cannot be directly extended to the broader class of GLMs or to control the (stronger) notion of frequentist regret.

Non-Stationary GLM have been studied in recent works that consider both abruptly changing and smoothly changing environments (Cheung et al., 2019a; Zhao et al., 2020). However the analysis in both of these works have gaps: (Zhao et al., 2020) define c_μ (see our Assumption 4) as the minimum value of $\dot{\mu}(a^\top\theta)$ for $\theta \in \mathbb{R}^d$, which for the logistic regression model would be zero; (Cheung et al., 2019a) implicitly assume that the maximum likelihood estimator at all time instants belongs to $\Theta = \{\theta \in \mathbb{R}^d, \|\theta\|_2 \leq S\}$, which may not be true in general.

Main Contributions. In this paper, we propose the first upper confidence bound algorithms designed for non-stationary environments in generalized linear bandits. The algorithms are extensions of the SW-LinUCB (Cheung et al., 2019b) and the D-LinUCB (Russac et al., 2019) algorithms and can achieve a dynamic regret over T rounds of order $O(d^{2/3}\Gamma_T^{1/3}T^{2/3})$, where Γ_T denotes the number of breakpoints up to time T. This rate is known to be optimal up to logarithmic terms. We propose an original and simplified analysis that is valid with time-dependent actions sets and does not required statistical assumption on the distribution of the contextual vectors. In the two algorithms, we make use of (possibly weighted) penalized maximum likelihood estimation. To the best of our knowledge, the analysis of penalized MLE in generalized linear bandits is also an original contribution. Note that in non-stationary environments there is no simple way to circumvent the need for regularization by using a proper initialization for the algorithms (as is done by Filippi et al. (2010)): when using the sliding window for instance the initialization procedure would need to be repeated regularly, resulting in a large drop in performance.

2 Problem Setting

Extending the generalized linear bandit framework introduced by Filippi et al. (2010), we consider a structured bandit model where the number of arms at each round is upper-bounded by a finite K, the action set $\mathcal{A}_t$ is time-dependent and at each step an action $A_t \in \mathcal{A}_t \subset \mathbb{R}^d$ is chosen. The conditional distribution of the rewards belongs to a *canonical exponential family* wrt a reference measure μ: $d\mathbb{P}_{A^\top\theta}(x) = d\mathbb{P}_\theta(x|A) = \exp(xA^\top\theta - b(A^\top\theta) + c(x))d\mu(x)$, where $c(.)$ is a real-valued function an $b(.)$ is assumed to be twice continuously differentiable.

A random variable X with the above density verifies $\mathbb{E}(X) = \dot{b}(A^\top\theta)$ and $\text{var}(X) = \ddot{b}(A^\top\theta)$, showing that $b(.)$ is strictly convex. The inverse link function is $\mu = \dot{b}$.

At time t, when the action A_t is chosen, the received reward X_t is conditionally independent of the past actions and satisfies $\mathbb{E}(X_t|A_t) = \mu(A_t^\top\theta^\star)$. In the non-stationary framework, the difference is that at time t the conditional expectation is equal to $\mu(A_t^\top\theta_t^\star)$ rather than $\mu(A_t^\top\theta^\star)$ after selecting an action A_t.

We first assume that the L2-norms of the available actions

and the admissible parameters $(\theta_t^\star)_{t\geq 1}$ are bounded,

Assumption 1. $\forall t \geq 1, \forall a \in \mathcal{A}_t,\ \|a\|_2 \leq L$.

Assumption 2. $\forall t \geq 1,\ \|\theta_t^\star\|_2 \leq S$.

The following assumption is also useful to derive concentration bounds.

Assumption 3. *There exists $m > 0$ such that for any $t \geq 1$, $0 < X_t < m$.*

Remark: We define the noise term as $\eta_t = X_t - \mu(A_t^\top \theta_t^\star)$, so that $\mathbb{E}[\eta_t | X_{t-1}, ..., X_1] = 0$. As explained in Lemma 1 of Appendix C.1, η_t is $m/2$-subgaussian conditionally on the past.

The maximum likelihood estimator $\hat{\theta}_t$ based on the rewards $X_1, ..., X_{t-1}$ and the selected actions $A_1, ..., A_{t-1}$ is defined as the maximizer of

$$\sum_{s=1}^{t-1} \log(\mathbb{P}_\theta(X_s|A_s)) = \sum_{s=1}^{t-1} X_s A_s^\top \theta - b(A_s^\top \theta) + c(X_s)\,. \tag{1}$$

By convexity of b, the rhs of the previous equation is concave in θ. After differentiating the log-likelihood, $\hat{\theta}_t$ appears as the solution of the equation

$$\sum_{s=1}^{t-1} (X_s - \mu(A_s^\top \theta)) A_s = 0\,. \tag{2}$$

Extra assumptions on the link function are also necessary for the theoretical analysis, in particular:

Assumption 4. *The inverse link function $\mu : \mathbb{R} \mapsto \mathbb{R}$ is a continuously differentiable Lipschitz function, with Lipschitz constant k_μ, such that*

$$c_\mu = \inf_{\|\theta\|_2 \leq S, \|a\|_2 \leq L} \dot{\mu}(a^\top \theta) > 0\,.$$

Assumption 4 could be relaxed by only considering the θ parameters in a neighborhood of the true unknown parameter $\theta^\star$ as in (Li et al., 2017). However, doing so would require assuming that the actions are drawn from a distribution verifying particular conditions. In a non-stationary environment, even this extra assumption is not always sufficient as $\theta^\star$ evolves over time.

In the non-stationary environment, the goal of the learner is to minimize the expected *dynamic regret* defined as

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

Note that in contrast to the settings considered by Filippi et al. (2010) or Kveton et al. (2020), the available actions sets $\mathcal{A}_t$ are time-dependent. Hence, in the above definition of regret, the best action can differ between rounds and it is no more possible to control the regret by upper-bounding the number of times each sub-optimal arm is played.

3 Algorithms

In this section, we describe two estimators together with the corresponding algorithms. The first estimator is based on a sliding window where only the τ most recent rewards and actions are considered. The second one uses a discount factor γ and gives more weight to the most recent actions and rewards. Both estimators rely on a penalization of the log-likelihood that has a regularizing effect and avoids the need of specific initialization procedures.

3.1 Sliding Window and Penalized MLE

The first estimator we consider is a truncated version of the penalized MLE. Equation (1) is replaced by

$$\sum_{s=\max(t-\tau,1)}^{t-1} \log(\mathbb{P}_\theta(X_s|A_s)) - \frac{\lambda}{2}\|\theta\|_2^2\,. \tag{3}$$

By differentiating the (strictly concave) penalized log-likelihood, $\hat{\theta}_t^{\texttt{SW}}$ appears as the unique solution of

$$\sum_{s=\max(t-\tau,1)}^{t-1} (X_s - \mu(A_s^\top \theta))A_s - \lambda\theta = 0\,. \tag{4}$$

We introduce

$$V_{t-1} = \sum_{s=\max(1,t-\tau)}^{t-1} A_s A_s^\top + \frac{\lambda}{c_\mu} I_d\,, \tag{5}$$

and we define $g_t(\theta) = \sum_{s=\max(1,t+1-\tau)}^{t} \mu(A_s^\top \theta)A_s + \lambda\theta$ and $\tilde{\theta}_t^{\texttt{SW}}$ by

$$\tilde{\theta}_t^{\texttt{SW}} = \underset{\|\theta\|_2 \leq S}{\arg\min} \|g_{t-1}(\hat{\theta}_t^{\texttt{SW}}) - g_{t-1}(\theta)\|_{V_{t-1}^{-1}}\,, \tag{6}$$

where V_{t-1} is defined in Equation (5). We need to consider $\tilde{\theta}_t^{\texttt{SW}}$ because $\hat{\theta}_t$ is not guaranteed to satisfy $\|\hat{\theta}_t\|_2 \leq S$ and the lower bound on $\dot{\mu}$ with c_μ is only valid for parameters whose L2 norm is smaller than S. $\tilde{\theta}_t^{\texttt{SW}}$ should be understood as a "projection" on the admissible parameters.

Using this notation, we can now present our first algorithm for generalized linear bandits in non-stationary environments. `SW-GLUCB` (Sliding Window Generalized Linear Upper Confidence Bound) uses a sliding window to focus on the most recent events. The `SW-GLUCB` algorithm uses a confidence bonus $\rho^{\texttt{SW}}$ that will defined in Section 4.1 devoted to the analysis of the algorithms (see Equation (13) for the definition of $\rho^{\texttt{SW}}$).

3.2 Discounting Factors and Penalized MLE

The second estimator we construct is based on a weighted penalized log-likelihood. Rather than using Equation (1), $\hat{\theta}_t^{\texttt{D}}$

Algorithm 1 SW-GLUCB

Input: Probability δ, dimension d, regularization λ, upper bound for actions L, upper bound for parameters S, sliding window τ.
Initialize: $V_0 = \lambda/c_\mu I_d$, $\hat{\theta}_0^{\mathtt{SW}} = 0_{\mathbb{R}^d}$.
for $t = 1$ **to** T **do**
  Receive $\mathcal{A}_t$, compute $\hat{\theta}_t^{\mathtt{SW}}$ according to (4)
  if $\|\hat{\theta}_t^{\mathtt{SW}}\|_2 \leq S$ **let** $\widetilde{\theta}_t^{\mathtt{SW}} = \hat{\theta}_t^{\mathtt{SW}}$ **else** compute $\tilde{\theta}_t^{\mathtt{SW}}$ with (6)
  Play $A_t = \arg\max_{a \in \mathcal{A}_t} \left(\mu(a^\top \widetilde{\theta}_t^{\mathtt{SW}}) + \rho_t^{\mathtt{SW}}(\delta) \|a\|_{V_{t-1}^{-1}} \right)$
  Receive reward X_t
  Update:
  if $t \leq \tau$ **then**
    $V_t \leftarrow V_{t-1} + A_t A_t^\top$
  else
    $V_t \leftarrow V_{t-1} + A_t A_t^\top - A_{t-\tau} A_{t-\tau}^\top$
  end if
end for

Algorithm 2 D-GLUCB

Input: Probability δ, dimension d, regularization λ, upper bound for actions L, upper bound for parameters S, discount factor γ.
Initialize: $W_0 = \lambda/c_\mu I_d$, $\hat{\theta}_0^{\mathtt{D}} = 0_{\mathbb{R}^d}$.
for $t = 1$ **to** T **do**
  Receive $\mathcal{A}_t$, compute $\hat{\theta}_t^{\mathtt{D}}$ according to (8)
  if $\|\hat{\theta}_t^{\mathtt{D}}\|_2 \leq S$ **let** $\tilde{\theta}_t^{\mathtt{D}} = \hat{\theta}_t^{\mathtt{D}}$ **else** compute $\tilde{\theta}_t^{\mathtt{D}}$ with (11)
  Play $A_t = \arg\max_{a \in \mathcal{A}_t} \left(\mu(a^\top \tilde{\theta}_t^{\mathtt{D}}) + \rho_t^{\mathtt{D}}(\delta) \|a\|_{W_{t-1}^{-1}} \right)$
  Receive reward X_t
  Update: $W_t \leftarrow A_t A_t^\top + \gamma W_{t-1} + \frac{\lambda}{c_\mu}(1-\gamma) I_d$
end for

is defined as the unique maximum of

$$\sum_{s=1}^{t-1} \gamma^{t-1-s} \log(\mathbb{P}_\theta(X_s|A_s)) - \frac{\lambda}{2}\|\theta\|_2^2 \,. \tag{7}$$

As before, thanks to the concavity in θ, $\hat{\theta}_t^{\mathtt{D}}$ is also the solution of

$$\sum_{s=1}^{t-1} \gamma^{t-1-s} (X_s - \mu(A_s^\top \theta)) A_s - \lambda\theta = 0 \,. \tag{8}$$

We introduce

$$W_t = \sum_{s=1}^{t} \gamma^{t-s} A_s A_s^\top + \frac{\lambda}{c_\mu} I_d \tag{9}$$

and

$$\widetilde{W}_t = \sum_{s=1}^{t} \gamma^{2(t-s)} A_s A_s^\top + \frac{\lambda}{c_\mu} I_d \,. \tag{10}$$

As in the linear setting, there is a need to introduce a covariance matrix containing the squares of the weights because the stochastic term can only be controlled in $\widetilde{W}_t^{-1}$ norm (Russac et al., 2019).

Let $g_t : \mathbb{R}^d \mapsto \mathbb{R}^d$ denote the following function

$$g_t(\theta) = \sum_{s=1}^{t} \gamma^{t-s} \mu(A_s^\top \theta) A_s + \lambda\theta \,.$$

Finally, let $\tilde{\theta}_t^{\mathtt{D}}$ be defined as

$$\tilde{\theta}_t^{\mathtt{D}} = \underset{\|\theta\|_2 \leq S}{\arg\min} \|g_{t-1}(\hat{\theta}_t^{\mathtt{D}}) - g_{t-1}(\theta)\|_{\widetilde{W}_{t-1}^{-1}} \,. \tag{11}$$

The second algorithm that we propose is D-GLUCB: exponentially increasing weights are used to progressively forget the past. The theoretical aspects of this algorithm are detailed in Section 4.2

The parameter $\rho^{\mathtt{D}}$ (line 8 above) will be defined below in Equation (14).

Remark: In the linear setting, the form of the upper confidence bound is a direct consequence of the high probability confidence ellipsoid that can be built around the estimate of the unknown parameter (Abbasi-Yadkori et al., 2011). There is no such confidence ellipsoid for generalized linear bandits. Therefore, the upper confidence bound has a different form. A possible approach that is chosen here is to consider $\text{UCB}_t(a) = \mathbb{E}_{\widetilde{\theta}_t}[X_t|A_t = a] + \rho(t)\|a\|_{M_{t-1}^{-1}}$, where $\mathbb{E}_{\widetilde{\theta}_t}[X_t|A_t = a]$ is equal to $\mu(a^\top \widetilde{\theta}_t)$ under a GLM. For SW-GLUCB, $\rho = \rho^{\mathtt{SW}}$ and $M_{t-1} = V_{t-1}$, as defined in Equation (5). Similarly, for D-GLUCB, $\rho = \rho^{\mathtt{D}}$ and $M_{t-1} = W_{t-1}$ is defined in Equation (9).

4 Concentration Bounds and Regret Analysis

In this section we give concentration results for the two estimators that we propose. Based on these concentration results, high probability upper-bounds for the dynamic regret of both algorithms are given. We show that we obtain results comparable to the ones in the linear setting. The main difference is that our analysis is valid only for abruptly changing environments. Proposing an algorithm that can be analyzed in both slowly drifting and abruptly changing environments under a generalized linear bandit remains an open question.

4.1 Analysis of SW-GLUCB

To obtain concentration inequalities, we need to restrict ourselves to segments of observations that are sufficiently far away from the changepoints. More precisely, let

$$\mathcal{T}(\tau) = \{t \leq T \text{ s. t. } \forall\, t - \tau \leq s \leq t, \theta_s^\star = \theta_t^\star\} \,. \tag{12}$$

$\mathcal{T}(\tau)$ contains all the time instants that are at least τ steps away from the closest previous breakpoint. At time instants in $\mathcal{T}(\tau)$, there is no bias due to non-stationarity of the environment as the sliding window of length τ is fully included in a stationary segment.

Proposition 1. *Let $0 < \delta < 1$ and $t \in \mathcal{T}(\tau)$. Let $\tilde{A}_t$ be any $\mathcal{A}_t$-valued random variable. Let*

$$\mathrm{c}_t^{\mathtt{SW}}(\delta) = \frac{m}{2}\sqrt{2\log(T/\delta) + d\log\left(1 + \frac{c_\mu L^2 \min(t,\tau)}{d\lambda}\right)}$$

$$\text{and} \quad \rho_t^{\mathtt{SW}}(\delta) = \frac{2k_\mu}{c_\mu}\left(\mathrm{c}_t^{\mathtt{SW}}(\delta) + \sqrt{c_\mu \lambda} S\right). \tag{13}$$

Then, simultaneously for all $t \in \mathcal{T}(\tau)$,

$$\left|\mu(\tilde{A}_t^\top \theta_t^\star) - \mu(\tilde{A}_t^\top \widetilde{\theta}_t^{\mathtt{SW}})\right| \leq \rho_t^{\mathtt{SW}}(\delta)\|\tilde{A}_t\|_{V_{t-1}^{-1}},$$

holds with probability higher than $1-\delta$.

Proof Sketch: Only a proof sketch is given here: the complete proof is to be found in Appendix A.1. The big picture is to use the assumption on the inverse link function and on the MLE to relate the deviations of the regression estimate to those of the martingale $S_{t-1} = \sum_{s=\max(1,t-\tau)}^{t-1} A_s \eta_s$. For $t \in \mathcal{T}(\tau)$, this can be done by upper bounding $|\mu(\tilde{A}_t^\top \theta_t^\star) - \mu(\tilde{A}_t^\top \tilde{\theta}_t)|$ by the quantity $2k_\mu/c_\mu \|a\|_{V_{t-1}^{-1}} (\|S_{t-1}\|_{V_{t-1}^{-1}} + \|\lambda\theta_t^\star\|_{V_{t-1}^{-1}})$. Then, the concentration result is established by upper-bounding the self-normalized quantity $\|S_{t-1}\|_{V_{t-1}^{-1}}$.

The concentration result of Proposition 1 is a prerequisite to give a high probability upper-bound on the instantaneous regret $\max_{a\in\mathcal{A}_t} \mu(a^\top\theta_t^\star) - \mu(A_t^\top\theta_t^\star)$.

Corollary 1. *Let $0 < \delta < 1$ and $A_{t,\star} = \arg\max_{a\in\mathcal{A}_t} \mu(a^\top\theta_t^\star)$. Then, simultaneously for all $t \in \mathcal{T}(\tau)$*

$$\mu(A_{t,\star}^\top\theta_t^\star) - \mu(A_t^\top\theta_t^\star) \leq 2\rho_t^{\mathtt{SW}}(\delta)\|A_t\|_{V_{t-1}^{-1}}$$

holds with probability at least $1-2\delta$.

The proof of this result is available in Appendix A.2. Corollary 1 allows us to give a high probability upper bound on the instantaneous regret for all time instants far enough from any breakpoints $t \in \mathcal{T}(\tau)$.

Based on those two concentration results, we can establish the following theorem for the regret of SW-GLUCB.

Theorem 1 (Regret of SW-GLUCB)**.** *The regret of the* SW-GLUCB *policy is upper-bounded with probability $\geq 1-2\delta$ by*

$$R_T \leq 2\sqrt{2}\rho_T^{\mathtt{SW}}(\delta)\sqrt{T}\sqrt{d\lceil T/\tau\rceil \log\left(1 + \frac{c_\mu L^2 \tau}{d\lambda}\right)} + m\Gamma_T\tau\,,$$

where $\rho^{\mathtt{SW}}$ is defined in Equation (13) *and Γ_T is the number of changes up to time T.*

Proof. In the following proof let ρ denote $\rho^{\mathtt{SW}}$.

$$\begin{aligned} R_T &= \sum_{t=1}^{T}(\mu(A_{t,\star}^\top\theta_t^\star) - \mu(A_t^\top\theta_t^\star)) \\ &\leq m\Gamma_T\tau + \sum_{t\in\mathcal{T}(\tau)} \min\{m, \mu(A_{t,\star}^\top\theta_t^\star) - \mu(A_t^\top\theta_t^\star)\} \end{aligned}$$

where in the last inequality the instantaneous regret $\forall t \notin \mathcal{T}(\tau)$ was upper-bounded by m. Using Corollary 1 with probability $\geq 1-2\delta$

$$\begin{aligned} R_T &\leq m\Gamma_T\tau + \sum_{t\in\mathcal{T}(\tau)} \min\{m, 2\rho_t(\delta)\|A_t\|_{V_{t-1}^{-1}}\} \\ &\leq m\Gamma_T\tau + 2\rho_T(\delta)\sum_{t\in\mathcal{T}(\tau)} \min\{1, \|A_t\|_{V_{t-1}^{-1}}\} \\ &\leq m\Gamma_T\tau + 2\rho_T(\delta)\sum_{t=1}^{T} \min\{1, \|A_t\|_{V_{t-1}^{-1}}\} \\ &\leq m\Gamma_T\tau + 2\rho_T(\delta)\sqrt{T}\sqrt{\sum_{t=1}^{T} \min\{1, \|A_t\|_{V_{t-1}^{-1}}^2\}}\,. \end{aligned}$$

The second inequality holds thanks to $m \leq 2\rho(T)$ and the last inequality is CauchySchwarz. Proposition 9 in Appendix C of Russac et al. (2019) yields

$$\sum_{t=1}^{T} \min\{1, \|A_t\|_{V_{t-1}^{-1}}^2\} \leq 2d\lceil T/\tau\rceil \log\left(1 + \frac{c_\mu L^2\tau}{\lambda d}\right),$$

which concludes the proof. □

In the following corollary, we denote $\tilde{O}$ the function growth when omitting the logarithmic terms.

Corollary 2. *If Γ_T is known, by choosing $\tau = \lceil (\frac{dT}{\Gamma_T})^{2/3}\rceil$, the regret of the* SW-GLUCB *algorithm is asymptotically upper bounded with high probability by a term $\tilde{O}(d^{2/3}\Gamma_T^{1/3}T^{2/3})$.*

If Γ_T is unknown, by choosing $\tau = \lceil d^{2/3}T^{2/3}\rceil$, the regret of the SW-GLUCB *algorithm is asymptotically upper bounded with high probability by a term $\tilde{O}(d^{2/3}\Gamma_T T^{2/3})$.*

This corollary is proved in Appendix A.3.

4.2 Analysis of D-GLUCB

The main difference when establishing concentration results in the weighted setting is the need to control the bias term which was avoided with the sliding window thanks to the condition $t \in \mathcal{T}(\tau)$. In the weighted setting, with a discount factor γ, we introduce $\mathcal{T}(\gamma)$ defined as

$$\mathcal{T}(\gamma) = \{t \leq T \text{ s. t. } \forall\, t - D(\gamma) < s \leq t, \theta_s^\star = \theta_t^\star\}\,,$$

where $D(\gamma)$ is an analysis parameter that will be specified later. The main reason for introducing this parameter is to

control the bias. Basically, as in the linear setting, the bias for time instants far enough from a breakpoint can be upper bounded more roughly than for the others.

Proposition 2. *Let $0 < \delta < 1$ and. Let $\tilde{A}_t$ be any $\mathcal{A}_t$-valued random variable. Let*

$$\mathrm{c}_t^{\mathtt{D}}(\delta) = \frac{m}{2}\sqrt{2\log(1/\delta) + d\log\left(1 + \frac{c_\mu L^2(1-\gamma^{2t})}{d\lambda(1-\gamma^2)}\right)},$$

$$\rho_t^{\mathtt{D}}(\delta) = \frac{2k_\mu}{c_\mu}\left(\mathrm{c}_t^{\mathtt{D}}(\delta) + \sqrt{c_\mu\lambda}S + 2L^2 S k_\mu \sqrt{\frac{c_\mu}{\lambda}}\frac{\gamma^{D(\gamma)}}{1-\gamma}\right). \tag{14}$$

Then simultaneously for all $t \in \mathcal{T}(\gamma)$

$$|\mu(\tilde{A}_t^\top \theta_t^\star) - \mu(\tilde{A}_t^\top \tilde{\theta}_t^{\mathtt{D}})| \le \rho_t^{\mathtt{D}}(\delta)\|\tilde{A}_t\|_{W_{t-1}^{-1}},$$

holds with a probability higher than $1-\delta$.

Proof Sketch:

As with the sliding window, we would like to use the concentration results established in the linear setting and extend the analysis to GLMs. The first step consists in upper bounding $|\mu(\tilde{A}_t^\top \theta_t^\star) - \mu(\tilde{A}_t^\top \tilde{\theta}_t^{\mathtt{D}})|$ with assumption 4. The upper-bound is a sum of two main terms. The first one is related to the weighted martingale $S_{t-1} = \sum_{s=1}^{t-1}\gamma^{-s}A_s\eta_s$. The self-normalized quantity $\|S_{t-1}\|_{\gamma^{2(t-1)}\widetilde{W}_{t-1}^{-1}}$ can be upper-bounded with high probability and we use Corollary 5 to do so. The next step consists in controlling the bias $\|\sum_{s=1}^{t-1-D(\gamma)}\gamma^{t-s}(\mu(A_s^\top\theta_t^\star) - \mu(A_s^\top\theta_s^\star))A_s\|_{\widetilde{W}_{t-1}^{-1}}$. The assumption $t \in \mathcal{T}(\gamma)$ is required at this step to have a proper control on this term. By combining the Lipschitz assumption (Assumption 4) on the inverse link function and a triangle inequality, the bias term can be upper-bounded by $2L^2Sk_\mu\sqrt{c_\mu/\lambda}\gamma^{D(\gamma)}/(1-\gamma)$. A detailed proof is available in Appendix B.2

Remark: In the linear setting, the bias can be controlled independently from the stochastic term. For example, Russac et al. (2019) consider a confidence ellipsoid centered around $\bar{\theta}_t$ (Proposition 3 of Russac et al. (2019)) to separate the two terms. With the particular geometry of the GLMs this is not achievable with the estimator we considered and the bias appears explicitly in the confidence bound as an additive term.

Proposition 2 can now be used to obtain a high probability upper bound for the instantaneous regret for all time instants $t \in \mathcal{T}(\gamma)$. We have the following corollary.

Corollary 3. *Let $0 < \delta < 1$, and $A_{t,\star} = \arg\max_{a\in\mathcal{A}_t}\mu(a^\top\theta_t^\star)$. Then, simultaneously for all $t \in \mathcal{T}(\gamma)$*

$$\mu(A_{t,\star}^\top\theta_t^\star) - \mu(A_t^\top\theta_t^\star) \le 2\rho_t^{\mathtt{D}}(\delta)\|A_t\|_{W_{t-1}^{-1}}$$

holds with probability at least $1-2\delta$.

The proof of this corollary essentially follows the ideas of the proof of Corollary 1. The main difference is the term $\log(1/\delta)$ in the high probability upper-bound (in $c_t^{\mathtt{D}}(\delta)$) instead of $\log(T/\delta)$ (in $c_t^{\mathtt{SW}}(\delta)$). This is because in the weighted setting an anytime deviation bound can be obtained (Corollary 5 in Appendix). On the contrary, with the sliding window, we cannot avoid the union bound argument to obtain the concentration result valid for all $t \in \mathcal{T}(\tau)$ which gives the extra T term.

The reader familiar with the analysis in the weighted linear setting may be surprised by the presence of the term $\|a\|_{W_{t-1}^{-1}}$ in the exploration bonus for D-GLUCB. In fact, one of the conclusion of Russac et al. (2019) was to prove that the exploration term in the upper confidence bound must contain the $W_{t-1}^{-1}\widetilde{W}_{t-1}W_{t-1}^{-1}$ norm of A_t (with $c_\mu = 1$ in the linear setting). However, knowing that $0 < \gamma < 1$, we have $\gamma^{2(t-s)} \le \gamma^{t-s}$ for $s \le t$, implying that $\widetilde{W}_{t-1} \le W_{t-1}$. Consequently, $\|a\|_{W_{t-1}^{-1}\widetilde{W}_{t-1}W_{t-1}^{-1}} \le \|a\|_{W_{t-1}^{-1}}$. The take home message is that it is possible to obtain a tighter bound in the linear case with a control in the $W_{t-1}\widetilde{W}_{t-1}^{-1}W_{t-1}$ norm for the confidence ellipsoid (Theorem 1 of Russac et al. (2019)), while the exploration term features the W_{t-1}^{-1} norm in the GLM.

Theorem 2 (Regret of D-GLUCB)**.** *The regret of the* D-GLUCB *policy is upper-bounded with probability $\ge 1-2\delta$ by*

$$R_T \le 2\rho_T^{\mathtt{D}}(\delta)\sqrt{2dT}\sqrt{T\log\left(\frac{1}{\gamma}\right) + \log\left(1 + \frac{c_\mu L^2}{d\lambda(1-\gamma)}\right)} + m\Gamma_T D(\gamma),$$

where $\rho^{\mathtt{D}}$ is defined in Equation (14) *and Γ_T is the number of changes up to time T.*

The proof essentially follows the arguments presented in Theorem 1 and is reported in Appendix B.3

Corollary 4. *By taking $D(\gamma) = \frac{\log(1/(1-\gamma))}{1-\gamma}$,*

1. *If Γ_T is known, by choosing $\gamma = 1 - (\frac{\Gamma_T}{dT})^{2/3}$, the regret of the* D-GLUCB *algorithm is asymptotically upper bounded with high probability by a term $\tilde{O}(d^{2/3}\Gamma_T^{1/3}T^{2/3})$.*
2. *If Γ_T is unknown, by choosing $\gamma = 1 - (\frac{1}{dT})^{2/3}$, the regret of the* D-GLUCB *algorithm is asymptotically upper bounded with high probability by a term $\tilde{O}(d^{2/3}\Gamma_T T^{2/3})$.*

This corollary is proved in Appendix B.4.

5 Experiments

In this section, we evaluate the empirical performance of the two proposed algorithms. In a first part, we reproduce the

simulation proposed in an abruptly changing environment in Russac et al. (2019). It consists in a two-dimensional problem with 3 different breakpoints. The theoretical aspects developed in the previous sections suggest that SW-GLUCB and D-GLUCB should have better performance than generalized linear bandit algorithms that do not take into account the non-stationarity. In a second part, we use a real world dataset to test the performances of the algorithms on a 9-dimensional problem where non-stationarity is artificially created.

5.1 Simulated environment

In this simulated environment, we compare different generalized linear bandits algorithms and linear bandits algorithms when the inverse link function is the sigmoid $\mu(x) = 1/(1+\exp(-x))$: the SW-GLUCB algorithm using a sliding window, the D-GLUCB algorithm based on the use of exponentially increasing weights and the stationary algorithm, where the maximum likelihood estimator is solution of Equation (1). Additionally to those three algorithms, we add their linear counterpart, LinUCB as in (Abbasi-Yadkori et al., 2011), SW-LinUCB as in (Cheung et al., 2019b) and D-LinUCB as presented in (Russac et al., 2019). Those three algorithms do not assume that the rewards are generated by a logistic function and use a misspecified linear model; we expect them to have higher regrets.

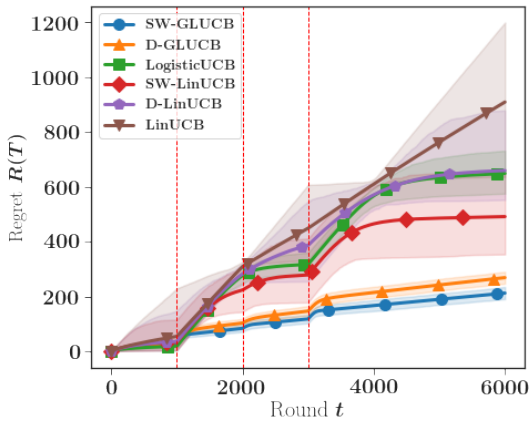

Figure 1: Regret of the different algorithms in a 2D abruptly changing environment and the 5% quantiles averaged on 500 independent runs

In this experiment the number of rounds is set to $T = 6000$. $\theta_t^\star$ the parameter in the logistic function is evolving over time: before $t = 1000$, $\theta_t^\star = (1, 0)$; for $1001 \leq t \leq 2000$, $\theta_t^\star = (-1, 0)$; for $2001 \leq t \leq 3000$, $\theta_t^\star = (0, 1)$ and for $t > 3000$, $\theta_t^\star = (0, -1)$. The position of $\theta^\star$ at the different periods are represented by the light blue triangles in the scatter plot in Figure 2. The locations of the changepoints are also represented on Figure 1 by the red dashed vertical lines. In this problem, $\theta^\star$ is widely spread over the 2 dimensional unit ball. At each round $K = 6$ actions randomly generated in the unit ball are presented to the different algorithms. The instantaneous regret in round t is defined as $r_t = \max_{a \in \{A_{t,1}, \dots A_{t,6}\}} \mu(a^\top \theta_t^\star) - \mu(A_t^\top \theta_t^\star)$, where A_t is the action chosen by the algorithm. In Figure 1 the cumulative dynamic regret of the different algorithms averaged on 500 independent runs is represented. The shaded region correspond to the 5% and the 95% quantiles for the cumulative regrets of the different algorithms. We can see that the variation of the performance is much larger for linear bandits algorithms than for the generalized linear bandits algorithms, a potential reason for this is that the confidence ellipsoid for the linear algorithms do not hold if the linear assumption of the rewards is not satisfied.

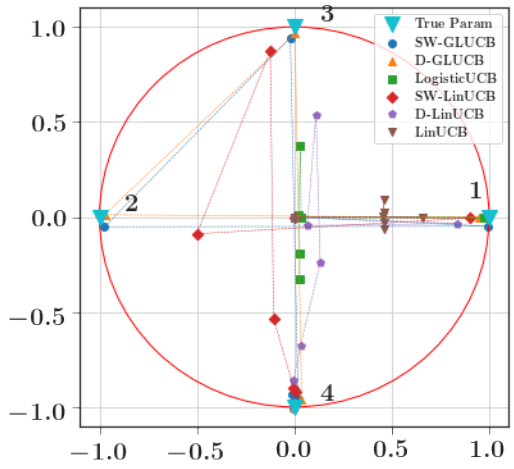

Figure 2: Estimated parameter ($\hat{\theta}_t$) every 1000 steps for the different algorithms in a 2D abruptly changing environment averaged on 500 independent runs.

After the different changepoints a 3000 rounds stationary period is added to check if the estimators of the different algorithms converge to the true parameter. In Figure 2, the estimator $\hat{\theta}_t$ is plotted every 1000 rounds for the different algorithms. We expect well-performing algorithms to approach the ground truth $\theta^\star$. The evolution of $\theta_t^\star$ requires the different algorithms to adapt to the changes. LogisticUCB and LinUCB fail in doing so. The failure is even worse for LinUCB because the algorithm does not leverage the logistic function information and does not converge, even after the stationary period corresponding to the second half of the experiment. On the scatter plot, the estimator for LinUCB never approaches the ground truth which explains the important regret. The LogisticUCB estimator catches the ground truth in the first stationary period but is not able to adapt to the changes in $\theta^\star$ and fails in estimating the evolving parameter. If the final stationary period is longer, it will eventually build a better estimator and converge.

The best performing policies are SW-GLUCB and D-GLUCB.

The estimators built in those algorithms track the evolving parameter accurately as can be seen on the scatter plot on Figure 2. SW-LinUCB performs surprisingly well. By progressively forgetting the past, the algorithm builds quite precise estimate of $\theta^\star$. Of course, the algorithm is not as precise as SW-GLUCB because it doesn't rely on the additional logistic assumption on the rewards, which implies a slower convergence to the true unknown parameter.

5.2 Simulation with a real-world dataset

In this section, we illustrate the performance of the generalized linear bandits algorithms with a real dataset. In contrast with the previous simulated environment, the rewards here are not generated by a logistic function but are the target variable of the dataset. We use the Pima Indian Diabetes Database [1] where the aim is to predict if a patient has diabetes or not. The predictions are based on 8 variables characterizing the different patients: number of pregnancies, the glucose level, the blood pressure, the thickness of the skin, insulin, the body mass index, the diabetes pedigree function and the age.

All the variables are numerical and the processing step consists in centering and standardizing the different variables. The outcome variable is binary and has the value 1 if the patient has diabetes. We run a 2000 steps experiment designed as follows: at each round, a patient without diabetes and a patient with diabetes are randomly selected and proposed to the different algorithms. The reward is $+1$ if the patient with diabetes was selected by the algorithm. We artificially create non-stationarity by inverting the population of diabetic and non-diabetic patients at time $t = 1000$. This change corresponds to a large perturbation but the algorithms that progressively forget the past should be able to adapt to the change and progressively recover a classification performance comparable to the level attained in the first segment.

We report in Figure 3 the proportion of diabetic patients detected averaged on 500 independent runs. Here, contrarily to the simulated environment, the rewards are not generated with a logistic function but the taken from the original dataset. Hence, we cannot directly evaluate the regret and the learning is more complex because the model misspecified. Nevertheless, even in this setting SW-GLUCB and D-GLUCB are able to learn continuously. After the changepoint, the diabetic patients are harder to detect and the averaged cumulative sum decreases for all the algorithms. The recovery is much faster for SW-GLUCB and D-GLUCB than for the stationary logistic bandit model. Although very simplistic, this experiment suggests that the proposed algorithms are robust enough to be successfully used for online bandit learning in realistic non-stationary binary regression tasks.

[1] The dataset can be downloaded here.

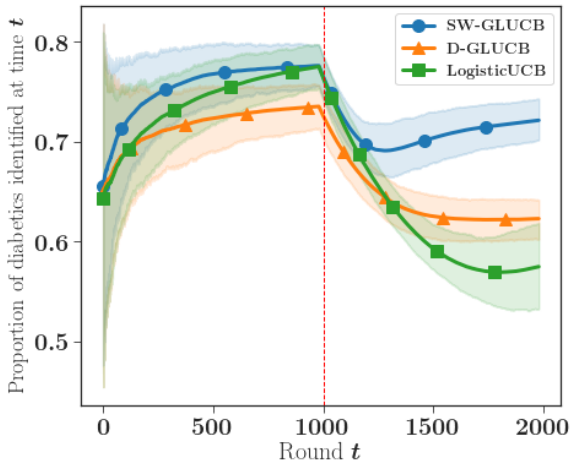

Figure 3: Proportion of diabetic patients detected at time t in an artificially created non-stationary environment averaged on 500 independent runs.

References

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

M. Abeille, A. Lazaric, et al. Linear thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

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

O. Besbes, Y. Gur, and A. Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems, NeurIPS 2014*, pages 199–207, 2014.

L. Besson and E. Kaufmann. The generalized likelihood ratio test meets klucb: an improved algorithm for piece-wise non-stationary bandits. *arXiv preprint arXiv:1902.01575*, 2019.

Y. Cao, Z. Wen, B. Kveton, and Y. 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.

W. C. Cheung, D. Simchi-Levi, and R. Zhu. Hedging the drift: Learning to optimize under non-stationarity. *arXiv preprint arXiv:1903.01461*, 2019a.

W. C. Cheung, D. Simchi-Levi, and R. Zhu. Learning to optimize under non-stationarity. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, 2019b.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory, COLT 2008*, pages 355–366, 2008.

S. Dong, T. Ma, and B. Van Roy. On the performance of thompson sampling on logistic bandits. In *32nd Annual Conference on Learning Theory, COLT 2019*, 2019.

S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems, NeurIPS 2010*, pages 586–594, 2010.

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

B. Kveton, M. Zaheer, C. Szepesvari, L. Li, M. Ghavamzadeh, and C. Boutilier. Randomized exploration in generalized linear bandits. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*, 2020.

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

L. Li, Y. Lu, and D. 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.

F. Liu, J. Lee, and N. Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Thirty-Second AAAI Conference on Artificial Intelligence, AAAI-2018*, 2018.

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

Y. Russac, C. Vernade, and O. Cappé. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems 32, NeurIPS 2019*, pages 12017–12026, 2019.

Q. Wu, N. Iyer, and H. Wang. Learning contextual bandits in a non-stationary environment. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 495–504. ACM, 2018.

P. Zhao, L. Zhang, Y. Jiang, and Z.-H. Zhou. A simple approach for non-stationary linear bandits. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*, 2020.

Supplementary for

Algorithms for Non-Stationary Generalized Linear Bandits

A Proof for the sliding window GLM

A.1 Proof of Proposition 1

Proposition 1. *Let $0 < \delta < 1$ and $t \in \mathcal{T}(\tau)$. Let $\tilde{A}_t$ be any $\mathcal{A}_t$-valued random variable. Let*

$$\mathrm{c}_t^{\mathtt{SW}}(\delta) = \frac{m}{2}\sqrt{2\log(T/\delta) + d\log\left(1 + \frac{c_\mu L^2 \min(t,\tau)}{d\lambda}\right)}$$

$$\textit{and} \quad \rho_t^{\mathtt{SW}}(\delta) = \frac{2k_\mu}{c_\mu}\left(\mathrm{c}_t^{\mathtt{SW}}(\delta) + \sqrt{c_\mu \lambda} S\right) .$$

Then, simultaneously for all $t \in \mathcal{T}(\tau)$,

$$\left|\mu(\tilde{A}_t^\top \theta_t^\star) - \mu(\tilde{A}_t^\top \widetilde{\theta}_t^{\mathtt{SW}})\right| \leq \rho_t^{\mathtt{SW}}(\delta) \|\tilde{A}_t\|_{V_{t-1}^{-1}}$$

holds with probability higher than $1-\delta$.

Proof. We define $g_{t-1} : \mathbb{R}^d \mapsto \mathbb{R}^d$ by $g_{t-1}(\theta) = \sum_{s=\max(1,t-\tau)}^{t-1} \mu(A_s^\top \theta) A_s + \lambda\theta$. Let J_{t-1} denotes the Jacobian matrix of g_{t-1}. We have $J_{t-1}(\theta) = \sum_{s=\max(1,t-\tau)}^{t-1} \dot{\mu}(A_s^\top \theta) A_s A_s^\top + \lambda I_d$.

Thanks to the definition of the estimator $\hat{\theta}_t^{\mathtt{SW}}$ defined in Equation (4), we have $g_{t-1}(\hat{\theta}_t^{\mathtt{SW}}) = \sum_{s=\max(1,t-\tau)}^{t-1} A_s X_s$. We also introduce the martingale $S_{t-1} = \sum_{s=\max(1,t-\tau)}^{t-1} A_s \eta_s$. In the following proof, we use $\tilde{\theta}_t$ instead of $\tilde{\theta}_t^{\mathtt{SW}}$.

We define the $G_{t-1}(\theta_t^\star, \tilde{\theta}_t)$ matrix as follows,

$$G_{t-1}(\theta_t^\star, \tilde{\theta}_t) = \int_0^1 J_{t-1}(u\theta_t^\star + (1-u)\tilde{\theta}_t)\, du \ .$$

The Fundamental Theorem of Calculus gives

$$g_{t-1}(\theta_t^\star) - g_{t-1}(\tilde{\theta}_t) = G_{t-1}(\theta_t^\star, \tilde{\theta}_t)(\theta_t^\star - \tilde{\theta}_t) \ . \tag{15}$$

Knowing that both $\theta_t^\star$ and $\tilde{\theta}_t$ have an L2-norm smaller than S, $\forall u \in [0,1], \|u\theta_t^\star + (1-u)\tilde{\theta}_t)\|_2 \leq S$. This implies in particular that

$$G_{t-1}(\theta_t^\star, \tilde{\theta}_t) \geq c_\mu \left(\sum_{s=\max(1,t-\tau)}^{t-1} A_s A_s^\top + \frac{\lambda}{c_\mu} I_d\right) = c_\mu V_{t-1} \ , \tag{16}$$

which in turn ensures $G_{t-1}(\theta_t^\star, \tilde{\theta}_t)$ is invertible.

Let $\tilde{A}_t$ be any $\mathcal{A}_t$ valued random variable and t be a fixed time instant,

$$
\begin{aligned}
|\mu(\tilde{A}_t^\top \theta_t^\star) - \mu(\tilde{A}_t^\top \tilde{\theta}_t)| &\leq k_\mu |\tilde{A}_t^\top (\theta_t^\star - \tilde{\theta}_t)| \quad \text{(Assumption 4)} \\
&= k_\mu |\tilde{A}_t^\top G_{t-1}^{-1}(\theta_t^\star, \tilde{\theta}_t)(g_{t-1}(\theta_t^\star) - g_{t-1}(\tilde{\theta}_t))| \quad \text{(Equation ((15)))} \\
&\leq k_\mu \|\tilde{A}_t\|_{G_{t-1}^{-1}(\theta_t^\star, \tilde{\theta}_t)} \|g_{t-1}(\theta_t^\star) - g_{t-1}(\tilde{\theta}_t)\|_{G_{t-1}^{-1}(\theta_t^\star, \tilde{\theta}_t)} \quad \text{(C-S)} \\
&\leq \frac{k_\mu}{c_\mu} \|\tilde{A}_t\|_{V_{t-1}^{-1}} \|g_{t-1}(\theta_t^\star) - g_{t-1}(\tilde{\theta}_t)\|_{V_{t-1}^{-1}} \quad \text{(Equation ((16)))} \\
&\leq \frac{2k_\mu}{c_\mu} \|\tilde{A}_t\|_{V_{t-1}^{-1}} \|g_{t-1}(\theta_t^\star) - g_{t-1}(\hat{\theta}_t^{\mathtt{SW}})\|_{V_{t-1}^{-1}} \quad \text{(Definition of } \tilde{\theta}_t) \\
&\leq \frac{2k_\mu}{c_\mu} \|\tilde{A}_t\|_{V_{t-1}^{-1}} \| \sum_{s=\max(1,t-\tau)}^{t-1} \mu(A_s^\top \theta_t^\star) A_s + \lambda \theta_t^\star - \sum_{s=\max(1,t-\tau)}^{t-1} A_s X_s \|_{V_{t-1}^{-1}} \\
&\leq \frac{2k_\mu}{c_\mu} \|\tilde{A}_t\|_{V_{t-1}^{-1}} \| \sum_{s=\max(1,t-\tau)}^{t-1} (\mu(A_s^\top \theta_t^\star) - \mu(A_s^\top \theta_s^\star)) A_s - \sum_{s=\max(1,t-\tau)}^{t-1} A_s \eta_s + \lambda \theta_t^\star \|_{V_{t-1}^{-1}} \\
&\leq \frac{2k_\mu}{c_\mu} \|\tilde{A}_t\|_{V_{t-1}^{-1}} \| -S_{t-1} + \lambda \theta_t^\star \|_{V_{t-1}^{-1}} \quad (\, t \in \mathcal{T}(\tau)\,) \\
&\leq \frac{2k_\mu}{c_\mu} \|\tilde{A}_t\|_{V_{t-1}^{-1}} \left(\|S_{t-1}\|_{V_{t-1}^{-1}} + \|\lambda \theta_t^\star\|_{V_{t-1}^{-1}} \right) \quad \text{(Triangle inequality)} \\
&\leq \frac{2k_\mu}{c_\mu} \|\tilde{A}_t\|_{V_{t-1}^{-1}} \left(\|S_{t-1}\|_{V_{t-1}^{-1}} + \sqrt{\lambda c_\mu} \|\theta_t^\star\|_2 \right) \quad (V_{t-1} \geq \frac{\lambda}{c_\mu} I_d) \\
&\leq \frac{2k_\mu}{c_\mu} \|\tilde{A}_t\|_{V_{t-1}^{-1}} \left(\frac{m}{2} \sqrt{2\log(1/\delta) + d\log\left(1 + \frac{c_\mu L^2 \min(t,\tau)}{d\lambda}\right)} + \sqrt{\lambda c_\mu} S \right) \quad \text{(with h.p.)} \,.
\end{aligned}
$$

In the last inequality we have used the concentration result established in the Proposition 5 of Russac et al. (2019) for the self-normalized quantity $\|S_{t-1}\|_{V_{t-1}^{-1}}$, and the assumption $\forall t, \|\theta_t^\star\|_2 \leq S$. To obtain the concentration result for all $t \in \mathcal{T}(\tau)$ we use a union bound. The final statement holds with probability $\geq 1 - \delta$. □

A.2 Proof of Corollary 1

Corollary 1. *Let $0 < \delta < 1$, and $A_{t,\star} = \arg\max_{a \in \mathcal{A}_t} \mu(a^\top \theta_t^\star)$. Then, simultaneously for all $t \in \mathcal{T}(\tau)$*

$$\mu(A_{t,\star}^\top \theta_t^\star) - \mu(A_t^\top \theta_t^\star) \leq 2\rho_t^{\mathtt{SW}}(\delta) \|A_t\|_{V_{t-1}^{-1}}$$

holds with probability at least $1 - 2\delta$.

Proof. In the following proof, we abbreviate $\tilde{\theta}_t^{\mathtt{SW}}$ to $\tilde{\theta}_t$.

$$\mu(A_{t,\star}^\top \theta_t^\star) - \mu(A_t^\top \theta_t^\star) = \underbrace{\mu(A_{t,\star}^\top \theta_t^\star) - \mu(A_{t,\star}^\top \tilde{\theta}_t)}_{A1} + \underbrace{\mu(A_{t,\star}^\top \tilde{\theta}_t) - \mu(A_t^\top \tilde{\theta}_t)}_{A2} + \underbrace{\mu(A_t^\top \tilde{\theta}_t) - \mu(A_t^\top \theta_t^\star)}_{A3}$$

Thanks to Proposition 1, we can give an upper bound for the term $A1$ and for the term $A3$. Upper bounding $A2$ with high probability requires extra-work.

With a union bound, we can simultaneously upper bound $A1$ and $A3$ for all $t \in \mathcal{T}(\tau)$ and the following holds

$$\mathbb{P}\left(\forall t \in \mathcal{T}(\tau), \mu(A_{t,\star}^\top \theta_t^\star) - \mu(A_{t,\star}^\top \tilde{\theta}_t) \leq \rho_t^{\mathtt{SW}}(\delta) \|A_{t,\star}\|_{V_{t-1}^{-1}} \cap \mu(A_t^\top \tilde{\theta}_t) - \mu(A_t^\top \theta_t^\star) \leq \rho_t^{\mathtt{SW}}(\delta) \|A_t\|_{V_{t-1}^{-1}} \right) \geq 1 - 2\delta \tag{17}$$

Let E denote this event. The upper confidence at time t for an action a is defined by,

$$\mathrm{UCB}_t(a) = \mu(a^\top \tilde{\theta}_t) + \rho_t^{\mathtt{SW}}(\delta) \|a\|_{V_{t-1}^{-1}},$$

The action chosen at time t, A_t is the action maximizing $\mathrm{UCB}_t(a)$ for $a \in \mathcal{A}_t$.

$$
\begin{aligned}
A2 &= \mu(A_{t,\star}^\top \tilde{\theta}_t) - \mu(A_t^\top \tilde{\theta}_t) \\
&= \mu(A_{t,\star}^\top \tilde{\theta}_t) + \rho_t^{\mathtt{SW}}(\delta)\|A_{t,\star}\|_{V_{t-1}^{-1}} - \rho_t^{\mathtt{SW}}(\delta)\|A_{t,\star}\|_{V_{t-1}^{-1}} - \mu(A_t^\top \tilde{\theta}_t) \\
&\leq \mu(A_t^\top \tilde{\theta}_t) + \rho_t^{\mathtt{SW}}(\delta)\|A_t\|_{V_{t-1}^{-1}} - \rho_t^{\mathtt{SW}}(\delta)\|A_{t,\star}\|_{V_{t-1}^{-1}} - \mu(A_t^\top \tilde{\theta}_t) \quad \text{(Definition of } A_t\text{)} \\
&\leq \rho_t^{\mathtt{SW}}(\delta)\|A_t\|_{V_{t-1}^{-1}} - \rho_t^{\mathtt{SW}}(\delta)\|A_{t,\star}\|_{V_{t-1}^{-1}}.
\end{aligned}
$$

Under the event E, that occurs with a probability higher than $1-\delta$,

$$
\begin{aligned}
\mu(A_{t,\star}^\top \theta_t^\star) - \mu(A_t^\top \theta_t^\star) &\leq \underbrace{\rho_t^{\mathtt{SW}}(\delta)\|A_{t,\star}\|_{V_{t-1}^{-1}}}_{\text{coming from} A1} + \underbrace{\rho_t^{\mathtt{SW}}(\delta)\|A_t\|_{V_{t-1}^{-1}} - \rho_t^{\mathtt{SW}}(\delta)\|A_{t,\star}\|_{V_{t-1}^{-1}}}_{\text{coming from} A2} + \underbrace{\rho_t^{\mathtt{SW}}(\delta)\|A_t\|_{V_{t-1}^{-1}}}_{\text{coming from} A3} \\
&\leq 2\rho_t^{\mathtt{SW}}(\delta)\|A_t\|_{V_{t-1}^{-1}}.
\end{aligned}
$$

□

A.3 Proof of Corollary 2

Corollary 2. *If Γ_T is known, by choosing $\tau = \lceil (\frac{dT}{\Gamma_T})^{2/3} \rceil$, the regret of the* SW-GLUCB *algorithm is asymptotically upper bounded with high probability by a term $\tilde{O}(d^{2/3}\Gamma_T^{1/3}T^{2/3})$.*

If Γ_T is unknown, by choosing $\tau = \lceil d^{2/3}T^{2/3} \rceil$, the regret of the SW-GLUCB *algorithm is asymptotically upper bounded with high probability by a term $\tilde{O}(d^{2/3}\Gamma_T T^{2/3})$.*

Proof. With this particular choice of τ we have:

$$\tau\Gamma_T \sim d^{2/3}T^{2/3}\Gamma_T^{1/3}$$

$$\rho_T^{\mathtt{SW}}(\delta) \sim \sqrt{d\log(T)}$$

$$\sqrt{T}\sqrt{\lceil T/\tau \rceil} \sim d^{-1/3}T^{1-1/3}\Gamma_T^{1/3}$$

Therefore the behavior of $\rho_T^{\mathtt{SW}}(\delta)\sqrt{dT}\sqrt{\lceil T/\tau \rceil}\sqrt{\log\left(1+\frac{\tau L^2}{\lambda d}\right)}$ is similar to $d^{2/3}\Gamma_T^{1/3}T^{2/3}\sqrt{\log(T)}\sqrt{\log(T/\Gamma_T)}$. By neglecting the logarithmic term, we have with high probability,

$$R_T = \widetilde{O}_{T\to\infty}(d^{2/3}\Gamma_T^{1/3}T^{2/3}) \ .$$

□

B Proof for the discounted GLM

B.1 Self-normalized concentration result

Corollary 5 (Corollary 3 of Russac et al. (2019))**.** *$\forall \delta > 0$, with $S_t = \sum_{s=1}^t \gamma^{-s} A_s \eta_s$, $\widetilde{V}_t = \sum_{s=1}^t \gamma^{-2s} A_s A_s^\top + \frac{\lambda\gamma^{-2t}}{c_\mu} I_d$ and when $(\eta_s)_{s\geq 1}$ are σ-subgaussian conditionally on the past, we have*

$$\mathbb{P}\left(\exists t \geq 0, \|S_t\|_{\widetilde{V}_t^{-1}} \geq \sigma\sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{c_\mu L^2(1-\gamma^{2t})}{d\lambda(1-\gamma^2)}\right)}\right) \leq \delta \ .$$

Proof. The proof is exactly the same than the one proposed in Russac et al. (2019), except that $\tilde{\lambda} = \lambda/c_\mu$ is used rather than λ, which explains the slight difference in the formula proposed in Corollary 5 compared to the original lemma. □

B.2 Proof of Proposition 2

Proposition 2. *Let $0 < \delta < 1$ and. Let $\tilde{A}_t$ be any $\mathcal{A}_t$-valued random variable. Let*

$$\mathrm{c}_t^{\mathtt{D}}(\delta) = \frac{m}{2}\sqrt{2\log(1/\delta) + d\log\left(1 + \frac{c_\mu L^2(1-\gamma^{2t})}{d\lambda(1-\gamma^2)}\right)}$$

$$\textit{and} \quad \rho_t^{\mathtt{D}}(\delta) = \frac{2k_\mu}{c_\mu}\left(\mathrm{c}_t^{\mathtt{D}}(\delta) + \sqrt{c_\mu\lambda}S + 2L^2 S k_\mu\sqrt{\frac{c_\mu}{\lambda}}\frac{\gamma^{D(\gamma)}}{1-\gamma}\right).$$

Then simultaneously for all $t \in \mathcal{T}(\gamma)$

$$|\mu(\tilde{A}_t^\top\theta_t^\star) - \mu(\tilde{A}_t^\top\tilde{\theta}_t^{\mathtt{D}})| \le \rho_t^{\mathtt{D}}(\delta)\|\tilde{A}_t\|_{W_{t-1}^{-1}},$$

holds with a probability higher than $1-\delta$.

Proof. During the proof, when no confusion is possible, we will forget the upper-script for the terms $\widetilde{\theta}_t^{\mathtt{D}}$ and $\hat{\theta}_t^{\mathtt{D}}$. In the weighted setting, $g_{t-1} : \mathbb{R}^d \mapsto \mathbb{R}^d$ is defined by $g_{t-1}(\theta) = \sum_{s=1}^{t-1}\gamma^{t-1-s}\mu(A_s^\top\theta)A_s + \lambda\theta$. The associated Jacobian matrix denoted by J_{t-1} verifies $J_{t-1}(\theta) = \sum_{s=1}^{t-1}\gamma^{t-1-s}\dot{\mu}(A_s^\top\theta)A_sA_s^\top + \lambda I_d$. $\hat{\theta}_t^{\mathtt{D}}$ verifies $g_{t-1}(\hat{\theta}_t^{\mathtt{D}}) = \sum_{s=1}^{t-1}\gamma^{t-1-s}A_sX_s$.

We also need to introduce two more matrices,

$$V_t = \gamma^{-t}W_t = \sum_{s=1}^{t}\gamma^{-s}A_sA_s^\top + \frac{\lambda\gamma^{-t}}{c_\mu}I_d \tag{18}$$

and

$$\widetilde{V}_t = \gamma^{-2t}\widetilde{W}_t = \sum_{s=1}^{t}\gamma^{-2s}A_sA_s^\top + \frac{\lambda\gamma^{-2t}}{c_\mu}I_d\,. \tag{19}$$

In the previous equations W_t and $\widetilde{W}_t$ are defined in Equation (9) and (10) respectively. Thanks to the fundamental Theorem of Calculus with $G_{t-1}(\theta_t^\star, \tilde{\theta}_t) = \int_0^1 J_{t-1}(u\theta_t^\star + (1-u)\tilde{\theta}_t)du$, the following holds

$$g_{t-1}(\theta_t^\star) - g_{t-1}(\tilde{\theta}_t) = G_{t-1}(\theta_t^\star, \tilde{\theta}_t)(\theta_t^\star - \tilde{\theta}_t). \tag{20}$$

Using the same argument than for Proposition 1, we have G_t is an invertible matrix and $G_{t-1}(\theta_t^\star, \tilde{\theta}_t) \ge c_\mu W_{t-1}$. Knowing that $0 < \gamma < 1$, it ensures $\widetilde{W}_{t-1} \le W_{t-1}$. Combining both inequalities gives,

$$\widetilde{W}_{t-1} \le W_{t-1} \le \frac{1}{c_\mu}G_{t-1}(\theta_t^\star, \tilde{\theta}_t)\,, \tag{21}$$

$$G_{t-1}^{-1}(\theta_t^\star, \tilde{\theta}_t) \le \frac{1}{c_\mu}W_{t-1}^{-1}\,. \tag{22}$$

We introduce the martingale $S_t = \sum_{s=1}^{t}\gamma^{-s}A_s\eta_s$. Let $B_t = \sum_{s=1}^{t-D(\gamma)-1}\gamma^{-s}(\mu(A_s^\top\theta_t^\star) - \mu(A_s^\top\theta_s^\star))A_s$ and let us

abbreviate $G_t(\theta_t^\star, \tilde{\theta}_t)$ as G_t, then

$$
\begin{aligned}
|\mu(\tilde{A}_t^\top \theta_t^\star) - \mu(\tilde{A}_t^\top \tilde{\theta}_t)| &\leq k_\mu |\tilde{A}_t^\top (\theta_t^\star - \tilde{\theta}_t)| = k_\mu |\tilde{A}_t^\top G_{t-1}^{-1}(g_{t-1}(\theta_t^\star) - g_{t-1}(\tilde{\theta}_t))| \quad \text{(Equation (20))} \\
&= k_\mu |\tilde{A}_t^\top G_{t-1}^{-1} \widetilde{W}_{t-1}^{1/2} \widetilde{W}_{t-1}^{-1/2}(g_{t-1}(\theta_t^\star) - g_{t-1}(\tilde{\theta}_t))| \\
&\leq k_\mu \|\tilde{A}_t\|_{G_{t-1}^{-1}\widetilde{W}_{t-1}G_{t-1}^{-1}} \|g_{t-1}(\theta_t^\star) - g_{t-1}(\tilde{\theta}_t)\|_{\widetilde{W}_{t-1}^{-1}} \quad \text{(C-S)} \\
&\leq \frac{k_\mu}{\sqrt{c_\mu}} \|\tilde{A}_t\|_{G_{t-1}^{-1}} \|g_{t-1}(\theta_t^\star) - g_{t-1}(\tilde{\theta}_t)\|_{\widetilde{W}_{t-1}^{-1}} \quad \text{(Inequality 21)} \\
&\leq \frac{k_\mu}{c_\mu} \|\tilde{A}_t\|_{W_{t-1}^{-1}} \|g_{t-1}(\theta_t^\star) - g_{t-1}(\tilde{\theta}_t)\|_{\widetilde{W}_{t-1}^{-1}} \quad \text{(Inequality 22)} \\
&\leq 2\frac{k_\mu}{c_\mu} \|\tilde{A}_t\|_{W_{t-1}^{-1}} \|g_{t-1}(\theta_t^\star) - g_{t-1}(\hat{\theta}_t)\|_{\widetilde{W}_{t-1}^{-1}} \quad \text{(Definition of } \tilde{\theta}_t) \\
&\leq \frac{2k_\mu}{c_\mu} \|\tilde{A}_t\|_{W_{t-1}^{-1}} \left(\|\sum_{s=1}^{t-1} \gamma^{t-1-s} \mu(A_s^\top \theta_t^\star) A_s - \sum_{s=1}^{t-1} \gamma^{t-1-s} A_s X_s\|_{\widetilde{W}_{t-1}^{-1}} + \|\lambda \theta_t^\star\|_{\widetilde{W}_{t-1}^{-1}} \right) \\
&\leq \frac{2k_\mu}{c_\mu} \|\tilde{A}_t\|_{W_{t-1}^{-1}} \left(\|\sum_{s=1}^{t-1} \gamma^{-s} (\mu(A_s^\top \theta_t^\star) - \mu(A_s^\top \theta_s^\star)) A_s - \sum_{s=1}^{t-1} \gamma^{-s} A_s \eta_s\|_{\widetilde{V}_{t-1}^{-1}} + \sqrt{\lambda c_\mu} S \right) \\
&\leq \frac{2k_\mu}{c_\mu} \|\tilde{A}_t\|_{W_{t-1}^{-1}} \left(\|B_t - \sum_{s=1}^{t-1} \gamma^{-s} A_s \eta_s\|_{\widetilde{V}_{t-1}^{-1}} + \sqrt{\lambda c_\mu} S \right) \quad \text{(Thanks to } t \in \mathcal{T}(\gamma)) \\
&\leq \frac{2k_\mu}{c_\mu} \|\tilde{A}_t\|_{W_{t-1}^{-1}} \left(\|B_t\|_{\widetilde{V}_{t-1}^{-1}} + \|S_{t-1}\|_{\widetilde{V}_{t-1}^{-1}} + \sqrt{c_\mu \lambda} S \right) \quad \text{(Triangle Inequality).}
\end{aligned}
$$

By using the results of Corollary 5 and the fact that $(\eta_s)_{s\geq 1}$ are conditionally $m/2$-subgaussian, with probability $\geq 1-\delta$ it holds that

$$
\forall t \geq 1, \|S_t\|_{\widetilde{V}_t^{-1}} \leq \frac{m}{2} \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{c_\mu L^2 (1-\gamma^{2t})}{\lambda d (1-\gamma^2)}\right)}\,.
$$

The next step consists in upper-bounding the bias term B_t.

$$
\begin{aligned}
\|B_t\|_{\widetilde{V}_{t-1}^{-1}} &= \| \sum_{s=1}^{t-D(\gamma)-1} \gamma^{-s} (\mu(A_s^\top \theta_t^\star) - \mu(A_s^\top \theta_s^\star)) A_s \|_{\widetilde{V}_{t-1}^{-1}} \\
&\leq \sqrt{\frac{c_\mu}{\lambda \gamma^{-2(t-1)}}} \left\| \sum_{s=1}^{t-D(\gamma)-1} \gamma^{-s} (\mu(A_s^\top \theta_t^\star) - \mu(A_s^\top \theta_s^\star)) A_s \right\|_2 \quad (\widetilde{V}_{t-1} \geq \frac{\lambda \gamma^{-2(t-1)}}{c_\mu}) \\
&\leq \sqrt{\frac{c_\mu}{\lambda \gamma^{-2(t-1)}}} \sum_{s=1}^{t-D(\gamma)-1} \gamma^{-s} |(\mu(A_s^\top \theta_t^\star) - \mu(A_s^\top \theta_s^\star))| \|A_s\|_2 \quad \text{(Triangle Inequality)} \\
&\leq L\sqrt{\frac{c_\mu}{\lambda}} \sum_{s=1}^{t-D(\gamma)-1} \gamma^{t-1-s} |(\mu(A_s^\top \theta_t^\star) - \mu(A_s^\top \theta_s^\star))| \\
&\leq L\sqrt{\frac{c_\mu}{\lambda}} \sum_{s=1}^{t-D(\gamma)-1} \gamma^{t-1-s} k_\mu |A_s^\top (\theta_t^\star - \theta_s^\star)| \quad \text{(Assumption 4)} \\
&\leq 2L^2 S k_\mu \sqrt{\frac{c_\mu}{\lambda}} \sum_{s=1}^{t-D(\gamma)-1} \gamma^{t-1-s} \quad \text{(C-S + Assumption 1 + Assumption 2)} \\
&\leq 2L^2 S k_\mu \sqrt{\frac{c_\mu}{\lambda}} \frac{\gamma^{D(\gamma)}}{1-\gamma}\,.
\end{aligned}
$$

The result is obtained by combining the inequalities. □

B.3 Proof of Theorem 2

Theorem 2 (Regret of D-GLUCB)**.** *The regret of the* D-GLUCB *policy is upper-bounded with probability* $\geq 1-2\delta$ *by*

$$R_T \leq 2\rho_T^{\mathtt{D}}(\delta)\sqrt{2dT}\sqrt{T\log\left(\frac{1}{\gamma}\right)+\log\left(1+\frac{c_\mu L^2}{d\lambda(1-\gamma)}\right)}+m\Gamma_T D(\gamma)\ ,$$

where $\rho^{\mathtt{D}}$ *is defined in Equation* (14) *and* Γ_T *is the number of changes up to time* T*.*

Proof. The regret is defined in the following way.

$$\begin{aligned}R_T &= \sum_{t\notin\mathcal{T}(\gamma)}(\mu(A_{t,\star}^\top\theta_t^\star)-\mu(A_t^\top\theta_t^\star))+\sum_{t\in\mathcal{T}(\gamma)}(\mu(A_{t,\star}^\top\theta_t^\star)-\mu(A_t^\top\theta_t^\star))\\ &\leq m\Gamma_T D(\gamma)+\sum_{t\in\mathcal{T}(\gamma)}\min\{m,\mu(A_{t,\star}^\top\theta_t^\star)-\mu(A_t^\top\theta_t^\star)\}\ .\end{aligned}$$

By using the result of Corollary 3, it holds that with probability $\geq 1-2\delta$

$$\begin{aligned}R_T &\leq m\Gamma_T D(\gamma)+\sum_{t\in\mathcal{T}(\gamma)}\min\{m,2\rho_t^{\mathtt{D}}(\delta)\|A_t\|_{W_{t-1}^{-1}}\}\\ &\leq m\Gamma_T D(\gamma)+2\rho_T^{\mathtt{D}}(\delta)\sum_{t\in\mathcal{T}(\gamma)}\min\{1,\|A_t\|_{W_{t-1}^{-1}}\}\\ &\leq m\Gamma_T D(\gamma)+2\rho_T^{\mathtt{D}}(\delta)\sqrt{T}\sqrt{\sum_{t=1}^{T}\min\{1,\|A_t\|^2_{W_{t-1}^{-1}}\}}\quad\text{(C-S)}\ .\end{aligned}$$

Based on the proof of Proposition 4 in Appendix B of (Russac et al., 2019), we have

$$\sqrt{\sum_{t=1}^{T}\min\{1,\|A_t\|^2_{W_{t-1}^{-1}}\}}\leq\sqrt{2d}\sqrt{T\log\left(\frac{1}{\gamma}\right)+\log\left(1+\frac{c_\mu L^2}{d\lambda(1-\gamma)}\right)}\ .$$

Therefore, with probability greater than $1-2\delta$,

$$R_T\leq 2\rho_T^{\mathtt{D}}(\delta)\sqrt{2d}\sqrt{T}\sqrt{T\log\left(\frac{1}{\gamma}\right)+\log\left(1+\frac{c_\mu L^2}{d\lambda(1-\gamma)}\right)}+m\Gamma_T D(\gamma)\ .$$

□

B.4 Proof of Corollary 4

Corollary 4. *By taking* $D(\gamma)=\frac{\log(1/(1-\gamma))}{1-\gamma}$*,*

1. *If* Γ_T *is known, by choosing* $\gamma=1-(\frac{\Gamma_T}{dT})^{2/3}$*, the regret of the* D-GLUCB *algorithm is asymptotically upper bounded with high probability by a term* $\tilde{O}(d^{2/3}\Gamma_T^{1/3}T^{2/3})$*.*
2. *If* Γ_T *is unknown, by choosing* $\gamma=1-\frac{1}{d^{2/3}T^{2/3}}$*, the regret of the* D-GLUCB *algorithm is asymptotically upper bounded with high probability by a term* $\tilde{O}(d^{2/3}\Gamma_T T^{2/3})$*.*

Proof. Let γ be defined as $\gamma=1-(\frac{\Gamma_T}{dT})^{2/3}$ and $D(\gamma)=\frac{\log(1/(1-\gamma))}{(1-\gamma)}$. With this choice of γ, $D(\gamma)$ is equivalent to $d^{2/3}\Gamma_T^{-2/3}T^{2/3}\log(T)$. Thus, $D(\gamma)\Gamma_T$ is equivalent to $d^{2/3}\Gamma_T^{1/3}T^{2/3}\log(T/\Gamma_T)$.

In addition,

$$\gamma^{D(\gamma)}=\exp(D(\gamma)\log(\gamma))=\exp\left(-\frac{\log(\gamma)}{1-\gamma}\log(1-\gamma)\right)\sim 1-\gamma\ .$$

Hence, when omitting the logarithmic terms, $\rho_T^{\mathtt{D}}(\delta)$ behaves as $\sqrt{d}$.
Furthermore, $\log(1/\gamma) \sim d^{-2/3}\Gamma_T^{2/3}T^{-2/3}$, implying that $T\log(1/\gamma) \sim d^{-2/3}\Gamma_T^{2/3}T^{1/3}$.
As a result, it holds that when neglecting the log terms,

$$\rho_T^{\mathtt{D}}(\delta)\sqrt{dT}\sqrt{T\log(1/\gamma) + \log\left(1 + \frac{c_\mu L^2}{d\lambda(1-\gamma)}\right)} \approx dT^{1/2}\sqrt{d^{-2/3}\Gamma_T^{2/3}T^{1/3}} = d^{2/3}\Gamma_T^{1/3}T^{2/3} \ .$$

We obtain the desired result. □

C Subgaussianity of the noise term

C.1 Conditional Hoeffding lemma

Lemma 1 (Conditional Hoeffding lemma)**.** *Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ be a probability space where $(\mathcal{F}_t)_{t\geq 0}$ is a filtration and $(X_t)_{t\geq 0}$ is a sequence of adapted random variables. Under the assumptions:*

1. *G_t is $(\mathcal{F}_{t-1})$-measurable*
2. *$G_t + a_t \leq X_t \leq G_t + b_t \quad a.s$*
3. *$\mathbb{E}\left[X_t|\mathcal{F}_{t-1}\right] = 0$*

Then,

$$\forall \lambda \in \mathbb{R}, \mathbb{E}\left[e^{\lambda X_t}|\mathcal{F}_{t-1}\right] \leq e^{\frac{\lambda^2(b_t-a_t)^2}{8}}, \quad a.s.$$

This means that under the assumption of Lemma 1, X_n is $(b_n - a_n)/2$-subgaussian conditionally on the past.

C.2 Consequence on the noise term in GLMs

In our bandit setting, the filtration associated with the random observations is denoted $\mathcal{F}_t = \sigma(X_1, ..., X_t)$ and is such that A_t is $\mathcal{F}_{t-1}$-measurable and η_t is $\mathcal{F}_t$-measurable. Under assumption 3, $\eta_t = X_t - \mu(A_t^\top \theta_t^\star)$ satisfies:

1. $-\mu(A_t^\top \theta_t^\star) \leq \eta_t \leq m - \mu(A_t^\top \theta_t^\star) \quad$ a.s
2. $\mu(A_t^\top \theta_t^\star)$ is $\mathcal{F}_{t-1}$-measurable
3. $\mathbb{E}\left[\eta_t|\mathcal{F}_{t-1}\right] = 0$

Lemma 1 implies that η_t is $m/2$-subgaussian conditionally on the past.