Proceedings of Machine Learning Research vol 132:1–37, 2021 32nd International Conference on Algorithmic Learning Theory

Optimal Regret Bounds for Generalized Linear Bandits under Parameter Drift

Louis Faury [*] L.FAURY@CRITEO.COM
LTCI TélécomParis, Criteo AI Lab

Yoan Russac [*] YOAN.RUSSAC@ENS.FR
ENS Paris, Université PSL, CNRS, Inria

Marc Abeille M.ABEILLE@CRITEO.COM
Criteo AI Lab

Clément Calauzènes C.CALAUZENES@CRITEO.COM
Criteo AI Lab

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

Generalized Linear Bandits (GLBs) are powerful extensions to the Linear Bandit (LB) setting, broadening the benefits of reward parametrization beyond linearity. In this paper we study GLBs in non-stationary environments, characterized by a general metric of non-stationarity known as the variation-budget or *parameter-drift*, denoted B_T. While previous attempts have been made to extend LB algorithms to this setting, they overlook a salient feature of GLBs which flaws their results. In this work, we introduce a new algorithm that addresses this difficulty. We prove that it enjoys a $\tilde{\mathcal{O}}(d^{2/3}B_T^{1/3}T^{2/3})$ regret-bound, matching (up to logarithmic factors) the minimax lower-bound established for LB. At the core of our contribution is a generalization of the projection step introduced in Filippi et al. (2010), adapted to the non-stationary nature of the problem. Our analysis sheds light on central mechanisms inherited from the setting by explicitly splitting the treatment of the learning and tracking aspects of the problem.

Keywords: Stochastic Bandits, Generalized Linear Model, Non-Stationarity.

1. Introduction

Linear Bandits and non-stationarity. The Linear Bandit (LB) framework has proven to be an important paradigm for sequential decision making under uncertainty. It notably extends the Multi-Arm Bandit (MAB) framework to address the exploration-exploitation dilemma when the arm-set is large (potentially infinite) or changing over time. While the LB has now been extensively studied (Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011; Abeille and Lazaric, 2017) in its original formulation, a recent strand of research studies its adaptation to non-stationary environments. Notable are the contributions of Cheung et al. (2019b); Russac et al. (2019); Zhao et al. (2020) which prove that under appropriate algorithmic changes, existing LB concepts can be leveraged to handle a drift of the reward model. Aside their theoretical interests, these results further anchor the spectrum of potential applications of the LB framework to real-world problems, where non-stationarity is commonplace.

[*] Equal contribution.

Extensions to Generalized Linear Bandits. Perhaps the main limitation of LB resides in its inability to model specific (e.g binary, discrete) rewards. One axis of research to operate beyond linearity was initiated with the introduction of Generalized Linear Bandit (GLBs) by Filippi et al. (2010). This framework allows to handle rewards which (in expectation) can be expressed as a generalized linear model. Notable members of this family are the logistic and Poisson models. Given the remarkable importance and widespread use of such models in practice, ensuring their resilience to non-stationarity stands as a crucial missing piece. At first glance, as the analysis of GLBs mainly relies on tools from the LB literature, one could expect this demonstration to be straight-forward, and almost anecdotal. As a matter of fact, the treatment of GLBs in non-stationary environments was already proposed as a direct extension of non-stationary LB algorithms ((Cheung et al., 2019a, Section 8.3) and (Zhao et al., 2020, Section 5.2)). However, as recently pointed out by Russac et al. (2020a), some crucial subtleties of the GLBs flaw the analysis and negates the validity of such extensions. An answer to this issue was brought by Russac et al. (2020a) who proposed a valid analysis for GLBs in non-stationary environments. However, their investigation is restricted to a specific kind of non-stationarity known as *abrupt changes*, leaving the treatment of the superior *parameter-drift* case for future work. To the best of our knowledge, a correct derivation of GLBs' behavior under this more general description of non-stationarity is still missing.

Scope and contributions. We focus in this paper on closing this gap. Our main contribution is **(1)** the design of `BVD-GLM-UCB` (Algorithm 1), the first GLB algorithm resilient to parameter-drift and matching the minimax rates of non-stationary LB (Theorem 1). This result relies on **(2)** a generalization of the projection step of Filippi et al. (2010) to non-stationary environments, of similar complexity than its stationary counterpart (Proposition 1). Our analysis **(3)** sheds light on some salient mechanisms of non-stationary bandits.

2. Preliminaries

We consider in this work the stochastic contextual bandit setting under parameter-drift. The environment starts by picking a sequence of parameters $\{\theta_\star^t\}_{t=1}^\infty$. A repeated game then begins between the environment and an agent. At each round t, the environment presents the agent with a set of actions $\mathcal{X}_t$ (potentially contextual, large or even infinite). The agent selects an action $x_t \in \mathcal{X}_t$ and receives a (stochastic) reward r_{t+1}. In this paper we work under the fundamental assumption that there exists a structural relationship between actions and their associated reward in the form of:

$$\mathbb{E}\left[r_{t+1} \,|\, \mathcal{F}_t, x_t\right] = \mu\left(\langle x_t, \theta_\star^t\rangle\right). \tag{1}$$

The filtration $\mathcal{F}_t := \sigma(\{x_s, r_{s+1}\}_{s=1}^{t-1})$ represents the information acquired at round t, and μ is a strictly increasing, continuously differentiable real-valued function most often referred to as the inverse link function. Notable instances of such a problem include the logistic bandit and the Poisson bandit. The goal of the agent is to minimize the cumulative pseudo-regret:

$$R_T := \sum_{t=1}^{T} \mu(\langle x_\star^t, \theta_\star^t\rangle) - \mu(\langle x_t, \theta_\star^t\rangle) \text{ where } x_\star^t = \arg\max_{x\in\mathcal{X}_t} \mu(\langle x, \theta_\star^t\rangle)\,.$$

We make the following assumption common in the study of parametric bandits:

Assumption 1 (Bounded decision set) *For all $t \geq 1$, the following holds true: $\left\|\theta_\star^t\right\|_2 \leq S$. Further, the actions have bounded norms: $\|x\|_2 \leq L$ for all $x \in \mathcal{X}_t$.*

Assumption 2 (Bounded reward) *There exists $\sigma > 0$ s.t $0 \leq r_t \leq 2\sigma$ holds almost surely.*

We will denote $\Theta = \{\theta, \|\theta\|_2 \leq S\}$ the set of admissible parameters and $\mathcal{X} = \{x, \|x\|_2 \leq L\}$. We assume that the quantities L, S and σ are known to the agent. The true parameters $\{\theta_\star^t\}_{t=1}^\infty$ are unknown, and their drift is quantified by the variation *variation-budget*, which characterizes the magnitude of the non-stationarity in the environment:

$$B_{T,\star} := \sum_{t=1}^{T-1} \left\|\theta_\star^{t+1} - \theta_\star^t\right\|_2 .$$

Naturally $B_{T,\star}$ is unknown. For the sake of simplicity and to isolate the main contribution of this paper (*i.e* minimax-optimality in non-stationary GLBs), we will make the following assumption.

Assumption 3 (Variation-budget upper-bound) *B_T is a* known *quantity such that $B_T \geq B_{T,\star}$.*

This assumption is common in non-stationary bandits (Besbes et al., 2014; Cheung et al., 2019a; Zhao et al., 2020). We will show in Section 4.4 how to bypass it with little to no impact on the regret. For a given inverse link function μ, we will follow the notation from Filippi et al. (2010) and denote:

$$k_\mu = \sup_{x\in\mathcal{X},\theta\in\Theta} \dot{\mu}\left(\langle x, \theta\rangle\right), \qquad c_\mu = \inf_{x\in\mathcal{X},\theta\in\Theta} \dot{\mu}\left(\langle x, \theta\rangle\right), \qquad R_\mu = k_\mu/c_\mu \ .$$

As in the stationary setting, learning can be canonically performed through the *quasi-maximum likelihood* principle, albeit with adequate modifications. Let b be a primitive of μ. Thanks to the strict increasing nature of the latter, b is a strictly convex function. Let $\lambda > 0$ and for $\gamma \in (0, 1)$ define[1]:

$$\hat{\theta}_t = \underset{\theta\in\mathbb{R}^d}{\arg\min} \sum_{s=1}^{t-1} \gamma^{t-1-s} \left[b(\langle x_s, \theta\rangle) - r_{s+1}\langle x_s, \theta\rangle\right] + \frac{\lambda c_\mu}{2} \|\theta\|_2^2 , \tag{2}$$

which is well-defined and unique as the minimizer of a strictly convex and coercive function. Further:

$$g_t(\theta) := \sum_{s=1}^{t-1} \gamma^{t-1-s} \mu(\langle x_s, \theta\rangle) x_s + \lambda c_\mu \theta.$$

Finally, we will use $\mathbf{V_t} := \sum_{s=1}^{t-1} \gamma^{t-1-s} x_s x_s^\mathsf{T} + \lambda \mathbf{I_d}$ and $\widetilde{\mathbf{V}}_\mathbf{t} := \sum_{s=1}^{t-1} \gamma^{2(t-1-s)} x_s x_s^\mathsf{T} + \lambda \mathbf{I_d}$.

3. Related work: limitations and challenges

3.1. GLBs and non-stationary LB

GLBs were first introduced by Filippi et al. (2010) who studied optimistic algorithms which enjoy a $\tilde{\mathcal{O}}(R_\mu d\sqrt{T})$ regret upper-bound, later refined for K-arms problem to $\tilde{\mathcal{O}}(R_\mu\sqrt{d\log(K)T})$ (Li et al., 2017). These findings were extended to randomized algorithms, both in the frequentist (Abeille and Lazaric, 2017) and Bayesian setting (Russo and Van Roy, 2014; Dong and Van Roy, 2018). GLBs also received an increasing attention targeted at improving their practical implementations (Jun et al., 2017; Dumitrascu et al., 2018).

1. We follow Russac et al. (2019) and use an exponential moving-average strategy. Our contribution is not specific to this approach and can easily be extended to other alternatives, e.g the sliding window.

The effects of parameter-drift were first studied in the MAB setting by Besbes et al. (2014) who for K-arm MAB achieved a dynamic regret bound of $\tilde{\mathcal{O}}(K^{1/3}B_T^{1/3}T^{2/3})$. Such results were recently extended to the stochastic LB: Cheung et al. (2019b) developed dynamic policies by resorting to a sliding-window, Russac et al. (2019) introduced a similar approach based on an exponential moving average, and Zhao et al. (2020) advocated for a simpler restart-based solution. All three aforementioned approaches enjoy regret bounds of the form $\tilde{\mathcal{O}}(d^{2/3}B_T^{1/3}T^{2/3})$, matching the lower-bound of Cheung et al. (2019a) up to logarithmic factors. Under the more specific assumption of abruptly-changing environments (also known as *switching* or *piece-wise stationary* bandits), regret bounds have been refined to $\tilde{\mathcal{O}}\left(\sqrt{\Gamma_T T}\right)$ in the MAB setting (Garivier and Moulines, 2011), where Γ_T is an upper bound on the number of switches.

3.2. Toward non-stationary GLBs: limitations

On the limits of piece-wise stationarity. To the best of our knowledge, the first valid analysis of non-stationary GLBs was conducted by Russac et al. (2020a). However, their work is restricted to piece-wise stationary environments, characterized by the number Γ_T of switches of the reward signal. On the practical side, this drastically narrows down the non-stationary scenarios that can be efficiently addressed, as the measure Γ_T can grossly overestimate the importance of the non-stationarity. In such case, any algorithm based on this measure will be sub-optimal and discard too fast previous data, quickly judged uninformative since the level of non-stationarity is expected to be high. This is typically the case in environments with many switches of small amplitude, characteristic of smooth drifts (e.g user-fatigue in recommender systems). On the theoretical side, this approach tells us little about the difficulties and challenges brought by the non-stationarity, as it relies on the fact that far enough from a switch, the environment is stationary. On the contrary, the variation-budget metric B_T introduced and discussed in Besbes et al. (2014, Section 2), allows for much finer considerations. It stands as a powerful characterization of the non-stationarity, measuring the number of switches and their amplitude *jointly*. As a result, it can efficiently cover different scenarios, from drifting to piece-wise stationary environments. An adequate treatment of GLBs under this superior metric is therefore a crucial missing piece, and requires a sensibly different analysis and an appropriate algorithmic design.

Parameter-drift and GLBs: flaws of previous approaches. Most of the existing non-stationary LB algorithms address the parameter-drift setting and their extension to GLBs was at first considered as relatively straight-forward (Cheung et al., 2019a; Zhao et al., 2020). Unfortunately, existing analyses suffer from important caveats because they overlook a crucial feature of GLBs. Following Filippi et al. (2010), they rely on a linearization of the reward function around $\hat{\theta}_t$. Naturally, the linear approximation must accurately describe the *effective* behavior of the reward signal (characterized by the ground-truth $\theta_\star^t$). From Assumption 2, this translates in the structural constraint $\hat{\theta}_t \in \Theta$, which is implicitly assumed to hold in previous attempts. Unfortunately, there exists no proof guaranteeing that $\hat{\theta}_t \in \Theta$ could hold. Even worse, existing deviation bounds (Abbasi-Yadkori et al., 2011, Theorem 1) rather suggest that in some directions, *even in the stationary case*, $\hat{\theta}_t$ can grow to be $\sqrt{\log(t)}$ far from Θ! The situation is even worse under non-stationarity since, as we shall see, $\hat{\theta}_t$ can be B_t far from Θ. This flaw in the analysis is critical and cannot be easily fixed without severely degrading the regret guarantee. When $\hat{\theta}_t \notin \Theta$, this impacts the ratio R_μ which captures the degree of non-linearity of the inverse link function. For the highly non-linear logistic function, easy computations show that $R_\mu \geq e^{SL}$. If we were to inflate the radius of the admissible set Θ from S to $S + \delta_S$ (so that it

contains $\hat{\theta}_t$), the estimated non-linearity of the reward function would be even stronger and R_μ would be multiplied by a factor $e^{L\delta_S}$! Because the regret bound scales linearly with R_μ, this exponential growth would lead to prohibitively deficient performance guarantees.

Remark 1 *The fact that $\hat{\theta}_t$ can leave the admissible set Θ is not merely a theoretical construction inherited from potentially loose deviation bounds. As highlighted in Figure 2(b), we can see in our numerical simulations that this often happens in practice when the environment is non-stationary.*

3.3. Non-stationary GLBs: challenges

In their seminal work, Filippi et al. (2010) countered the aforementioned difficulty by introducing a *projection* step, mapping $\hat{\theta}_t$ back to an admissible parameter $\tilde{\theta}_t \in \Theta$. Formally, they compute:

$$\tilde{\theta}_t = \arg\min_{\theta \in \Theta} \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{V_t^{-1}}} \tag{P0}$$

and use $\tilde{\theta}_t$ to predict the performance of the available actions. The projection step (**P0**) essentially incorporates the prior knowledge $\theta_\star \in \Theta$ (Assumption 2) without degrading the learning guarantees of the maximum likelihood estimator. This strategy was also leveraged by Russac et al. (2020a), which was made possible thanks to their piece-wise stationarity assumption.

The situation is different in our setting, as the parameter-drift framework allows the sequence $\{\theta_\star^t\}$ to change *at every round*. This introduces **(1)** the need to characterize two phenomenons of different nature that we will designate as *learning* and *tracking*. The former (learning) is linked to the deviation of the maximum-likelihood estimator $\hat{\theta}_t$ from its noiseless counterpart $\bar{\theta}_t$ (the estimator that one would have obtained if one could have averaged an infinite number of realization of the trajectory). The later (tracking) measures the deviation of $\bar{\theta}_t$ from the current $\theta_\star^t$, due to an incompressible error inherited from the drifting nature of the sequence $\{\theta_\star^s\}_{s=1}^t$. The learning and tracking mechanisms are both sources of deviation of $\hat{\theta}_t$ away from Θ, each under a different metric. This leads to **(2)** a tension in the design of the projection as this requires to incorporate the knowledge $\{\theta_\star^t\} \in \Theta$, without degrading neither the learning nor the tracking guarantees. This rules out the projection step (**P0**), oblivious to the tracking aspect of the problem and which needs to be generalized to adapt to the two sources of deviation (i.e learning and tracking).

4. Algorithm and regret bound

4.1. Algorithm

This section is dedicated to the description of the design of our new algorithm `BVD-GLM-UCB`. It operates in two steps: **(Step 1)** the computation of an appropriate admissible parameter $\tilde{\theta}_t \in \Theta$ (to be used for predicting the rewards associated with the actions $x \in \mathcal{X}_t$ available at round t) and **(Step 2)** the construction of a suitable exploration bonus to compensate for prediction errors.

The first step builds on the following set, linked to the deviation incurred through the learning process:

$$\mathcal{E}_t^\delta(\theta) := \left\{ \theta' \in \mathbb{R}^d \text{ s.t } \left\| g_t(\theta') - g_t(\theta) \right\|_{\widetilde{\mathbf{V}}_\mathbf{t}^{-1}} \leq \beta_t(\delta) \right\}, \tag{3}$$

where $\beta_t(\delta)$ is a slowly-increasing function of time (to be defined later) and $\delta \in (0, 1]$.

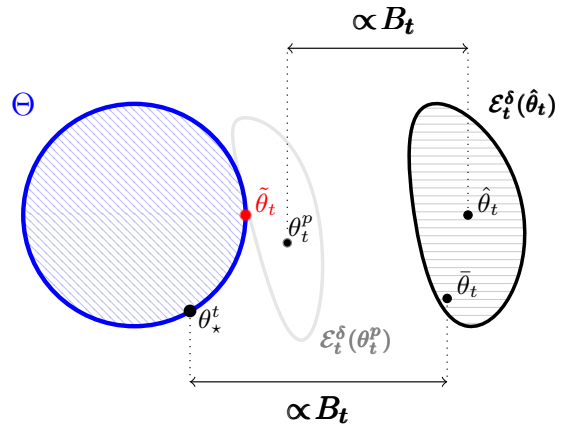


Figure 1: Illustration of the different parameters of interest. As stated by Lemma 2 and Lemma 4, the deviations $(\theta_t^p \leftrightarrow \hat{\theta}_t)$ and $(\bar{\theta}_t \leftrightarrow \theta_\star^t)$ are linked to the parameter-drift B_t. On the other hand, the deviations $(\hat{\theta}_t \leftrightarrow \bar{\theta}_t)$ and $(\tilde{\theta}_t \leftrightarrow \theta_t^p)$ are characterized by the stochastic nature of the problem.

Step 1. We start by identifying an intermediary parameter θ_t^p, solution of the following constrained optimization program (ties can be broken arbitrarily):

$$\theta_t^p \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\mathbf{V_t^{-2}}} \text{ s.t } \Theta \cap \mathcal{E}_t^\delta(\theta) \neq \emptyset \right\}. \tag{P1}$$

The optimization program (**P1**) is well-posed as it consists in minimizing a smooth function over a non-empty compact set[2]. Once θ_t^p is computed, the algorithm simply chooses any parameter $\tilde{\theta}_t \in \Theta \cap \mathcal{E}_t^\delta(\theta_t^p)$. An efficient procedure to find such a parameter is detailed in Section 4.3. The different parameters of interest for `BVD-GLM-UCB` are illustrated in Figure 1.

Remark 2 *Notice the difference with the projection step used in the stationary case. In our case it is possible that $\mathcal{E}_t^\delta(\hat{\theta}_t)$ (which is the confidence set centered at $\hat{\theta}_t$) does not intersect the admissible set Θ. Our strategy for finding $\tilde{\theta}_t$ is then to compute an appropriate **vibration** $\mathcal{E}_t^\delta(\theta_t^p)$ of $\mathcal{E}_t^\delta(\hat{\theta}_t)$ which does intersect Θ, while minimizing the deviation between θ_t^p and $\hat{\theta}_t$ according to a metric related to the tracking error (through the map g_t and the squared inverse of the design matrix).*

Step 2. The exploration bonus at round t for a given arm $x \in \mathcal{X}_t$ is defined as $b_t(x) = 2R_\mu \beta_t(\delta) \|x\|_{\mathbf{V_t^{-1}}}$, where $\delta \in (0, 1]$ and:

$$\beta_t(\delta) = \sqrt{\lambda} c_\mu S + \sigma \sqrt{2\log(1/\delta) + d\log\left(1 + \frac{L^2(1-\gamma^{2t})}{\lambda d(1-\gamma^2)}\right)}.$$

`BVD-GLM-UCB` then follows an optimistic strategy, boosting the predicted reward associated with $\tilde{\theta}_t$ by b_t and plays $x_t \in \arg\max_{x \in \mathcal{X}_t} \mu(\langle x, \tilde{\theta}_t \rangle) + b_t(x)$. The pseudo-code is summarized in Algorithm 1.

2. Notice that $\{\theta \text{ s.t } \Theta \cap \mathcal{E}_t^\delta(\theta) \neq \emptyset\}$ always contains 0_d, while the compactness is inherited from Θ.

Algorithm 1 BVD-GLM-UCB

Input. regularization λ, confidence δ, inverse link function μ, weight γ, constants S, L and σ.
Initialization. Compute R_μ, let $\mathbf{V_1} \leftarrow \lambda \mathbf{I}_d$ and $\hat{\theta}_1 \leftarrow 0_d$.
for $t \geq 1$ **do**
  Find θ_t^p by solving (**P1**) and select $\tilde{\theta}_t \in \Theta \cap \mathcal{E}_t^\delta(\theta_t^p)$.
  Play $x_t \leftarrow \arg\max_{x \in \mathcal{X}_t} \mu(\langle x, \tilde{\theta}_t \rangle) + 2R_\mu \beta_t(\delta) \|x\|_{\mathbf{V_t^{-1}}}$.
  Observe reward r_{t+1}, update $\hat{\theta}_{t+1}$ by solving Equation (2).
  Update design matrix: $\mathbf{V_{t+1}} \leftarrow \gamma \mathbf{V}_t + x_t x_t^\mathsf{T} + (1-\gamma)\lambda \mathbf{I_d}$.
end for

4.2. Regret bound

We provide in Theorem 1 a high-probability bound on the regret of BVD-GLM-UCB.

Theorem 1 *Under Assumptions 1-2-3, setting* $\gamma = 1 - (B_T/(dT))^{2/3}$ *ensures that the regret of* BVD-GLM-UCB *satisfies:*

$$R_T = \tilde{\mathcal{O}}\left(R_\mu d^{2/3} B_T^{1/3} T^{2/3}\right) \qquad \textit{w.h.p}$$

A few comments are in order. First, we note that the upper-bound on R_T matches the asymptotical rates of the LB lower-bound under parameter drift (Cheung et al., 2019a, Theorem 1). Second, one can notice the presence in the bound of the ratio R_μ, typical of the linearization approach performed to analyze GLBs. The bound presented in Theorem 1 is therefore quite natural and extends the work of Filippi et al. (2010) to non-stationary worlds. We emphasize that if the result seems unsurprising, it required a substantially different machinery, both for the design of the algorithm and its analysis. We highlight this last point in Section 5, dedicated at providing a comprehensive sketch of proof for Theorem 1. The complete and detailed proof is deferred to Section B in the supplementary material.

4.3. Solving the projection step

The optimization program (**P1**) and the subsequent search of a valid parameter $\tilde{\theta}_t$ can raise some legitimate concerns regarding the ease of practical implementation. Indeed, the feasible set of (**P1**) is given by $\{\theta \text{ s.t. } \Theta \cap \mathcal{E}_t^\delta(\theta) \neq \emptyset\}$, where $\mathcal{E}_t^\delta(\theta)$ is defined in (3). Hence, the associated constraint is *implicit* as it involves an additional *non-convex* minimization program. As a result, it makes the constraint uneasy to manipulate and even hard to check. The same difficulty arises when searching for $\tilde{\theta}_t \in \Theta \cap \mathcal{E}_t^\delta(\theta_t^p)$ where θ_t^p is a solution of (**P1**), due to the non-convexity of the set $\mathcal{E}_t^\delta(\theta_t^p)$. The following proposition provides an alternative that avoids those difficulties.

Proposition 1 *Let* $\tilde{\theta}_t$ *be such that:*

$$\begin{pmatrix} \tilde{\theta}_t \\ \eta_t^p \end{pmatrix} \in \underset{\theta' \in \mathbb{R}^d, \eta \in \mathbb{R}^d}{\arg\min} \left\{ \left\| g_t(\theta') + \beta_t(\delta) \widetilde{\mathbf{V}}_\mathbf{t}^{\mathbf{1/2}} \eta - g_t(\hat{\theta}_t) \right\|_{\mathbf{V_t^{-2}}} \textit{ s.t } \left\|\theta'\right\|_2 \leq S, \|\eta\|_2 \leq 1 \right\}. \qquad \textbf{(P2)}$$

It exists θ_t^p *solution of* (**P1**) *such that* $\tilde{\theta}_t \in \Theta \cap \mathcal{E}_t^\delta(\theta_t^p)$.

Proposition 1 shows that a valid $\tilde{\theta}_t$ can be found by solving (**P2**), bypassing the need to compute θ_t^p. Essentially, the initial two-steps procedure to find $\tilde{\theta}_t$ (through the intermediary program (**P1**)) is

replaced by a single minimization program augmented with a slack variable η. The attentive reader may notice that (**P2**) is now similar to (**P0**), the projection step employed in Filippi et al. (2010). As a result, BVD-GLM-UCB is comparable to the original algorithm GLM-UCB in terms of computational burden. The proof of Proposition 1 is given in Section C in the appendix.

4.4. Online estimation of the variation-budget

Motivation. The attentive reader may notice that the minimax-optimality of BVD-GLM-UCB is conditioned on the knowledge of an upper-bound B_T for the true parameter-drift $B_{T,\star}$. Naturally, the tighter this upper-bound, the better the performance. Yet, whether such a knowledge is available in real-life problems is, to say the least, questionable. This issue is not specific to our approach but is shared with all non-stationary parametric bandit methods - see for instance (Cheung et al., 2019b; Zhao et al., 2020). For linear bandits, previous approaches circumvented this drawback with a Bandit-over-Bandit strategy (Cheung et al., 2019a, Section 7), where $B_{T,\star}$ is learned online by a *master* algorithm. This guarantees an expected regret scaling as $\tilde{\mathcal{O}}\left(d^{2/3}B_{T,\star}^{1/3}T^{2/3} + d^{1/2}T^{3/4}\right)$ (Cheung et al., 2019a, Theorem 4) without having the knowledge of $B_{T,\star}$. We however note that this technique was specialized for linear bandits and for the sliding-window strategy. As hinted in the introduction one could easily design a sliding-window approach of BVD-GLM-UCB (using very similar arguments as the ones displayed in this paper) and extend the Bandit-over-Bandit of Cheung et al. (2019a) to the GLB framework. Here, we follow a different path and introduce an equivalent method for the exponential-weighting strategy. To the best of our knowledge, this technique was missing in the non-stationary parametric bandit literature. It notably proves that the online learning of $B_{T,\star}$ can be efficiently performed under discounted strategies.

Bandit-over-Bandit for discounted strategies. Notice that naive bounding gives $B_{T,\star} \in (0, 2ST]$. The main idea for learning $B_{T,\star}$ online is to grid on a log-scale the interval $(0, 2ST]$ with N values $\{B_{T,j}\}_{j=1}^{N}$. We then create N instances of BVD-GLM-UCB, each set with a different discount factor:

$$\gamma_j = 1 - \left(\frac{B_{T,j}}{dT}\right)^{2/3} = 1 - \frac{2^{j-1}}{2^{5/3}d^{2/3}TS^{2/3}} \ .$$

These instances will be our *experts*. We then deploy a *master* algorithm - a version of EXP3 (Auer et al., 2002), which acts repeatedly as follows: **1.** it chooses an expert j (*i.e* a new instance of BVD-GLM-UCB with parameter γ_j) to interact with the environment during a time frame of length H (H is a positive integer). **2.** The master algorithm then observes the cumulative reward (aggregated on the time frame) of the expert j. We give the pseudo-algorithm of this procedure in Algorithm 2.

Informally, the idea is that EXP3 will learn to select the best performing γ_j associated with the best estimate $B_{T,j}$ of $B_{T,\star}$. Intuitively, this should guarantee small regret as EXP3 will mostly play instances of BVD-GLM-UCB which nearly capture the true magnitude of the non-stationarity. This intuition is made rigorous in Theorem 2, whose proof is deferred to Section E in the appendix.

Theorem 2 *Under Assumptions 1-2, the regret of BOB-BVD-GLM-UCB when setting* $H = \lfloor d\sqrt{T} \rfloor$ *satisfies:*

$$\mathbb{E}[R_T] = \tilde{\mathcal{O}}\left(R_\mu d^{2/3}T^{2/3}\max\left(B_{T,\star}, d^{-1/2}T^{1/4}\right)^{1/3}\right) \ .$$

Algorithm 2 `BOB-BVD-GLM-UCB` (a more detailed version is deferred to Appendix E.2).

Input. Length H, time horizon T, regularization λ, confidence δ, inverse link function μ, constants S, L and σ.
Initialization. Let $N \leftarrow \lceil 2\log_2(2ST^{3/2})\rceil$ and $\mathcal{H} \leftarrow \{\gamma_j = 1 - \frac{2^{j-1}}{2^{5/3}d^{2/3}TS^{2/3}}\}_{j=1}^{N}$, initialize `EXP3` with action set indexed by $\mathcal{H}$.
for $i = 1, \dots, \lceil T/H \rceil$ **do**
  $j \leftarrow$ action selected by `EXP3`.
  Initialize a sub-routine `BVD-GLM-UCB` with parameter γ_j.
  for $t = 1, \dots, H$ **do**
    Play with `BVD-GLM-UCB` with parameter γ_j, observe reward r_{t+1}.
  end for
  Update `EXP3` with reward $\sum_{t=1}^{H} r_{t+1}$.
end for

Essentially, we obtain a regret bound which is identical to the ones of the Bandit-over-Bandit algorithms of Cheung et al. (2019a) and Zhao et al. (2020). The conclusions are therefore of similar nature: namely, when $B_{T,\star} \geq d^{-1/2}T^{1/4}$ we obtain a minimax rate, *without* knowing $B_{T,\star}$. Again, note here the presence of the problem-dependant constant R_μ, inherited from the non-linear reward structure imposed in GLBs.

5. Proof sketch

In this section, we detail the key steps of the proof of Theorem 1. In particular, we shed light on the tension between the learning and tracking aspects of the problem and their role in the choice of the estimator $\tilde{\theta}_t$, through the use of an appropriate projection step.

Learning versus tracking. A crucial feature of non-stationary GLBs lies in the singular nature of the deviation of $\hat{\theta}_t$ from $\theta_\star^t$. This arises from two fundamentally different mechanisms: learning and tracking. We introduce the following estimator, which allows for a clean-cut distinction between the two phenomenons:

$$\bar{\theta}_t := \underset{\theta\in\mathbb{R}^d}{\arg\min} \left\{ \sum_{s=1}^{t-1} \gamma^{t-1-s} \left[b(\langle x_s, \theta\rangle) - \mu\left(\langle x_s, \theta_\star^s\rangle\right) \langle x_s, \theta\rangle \right] + \frac{\lambda c_\mu}{2} \left\|\theta - \theta_\star^t\right\|_2^2 \right\}. \tag{4}$$

The parameter $\bar{\theta}_t$ is the minimizer of a strictly convex and coercive function, thus is well-defined and unique. Intuitively, $\bar{\theta}_t$ would be the estimator obtained under a perfect (e.g noiseless) observation of the reward[3]. As a result, the deviation between $\hat{\theta}_t$ and $\bar{\theta}_t$ is solely due to the stochastic nature of the problem (*learning*). On the other hand, the deviation between $\bar{\theta}_t$ and $\theta_\star^t$ is a consequence of the unpredictable changes of the sequence $\{\theta_\star^s\}_s$ (*tracking*). The introduction of the reference point $\bar{\theta}_t$ allows us to characterize both deviations separately in Lemma 1 and Lemma 2.

Lemma 1 *[Learning] Let $\delta \in (0, 1]$. With probability at least $1 - \delta$:*

$$\textit{for all } t \geq 1, \quad \bar{\theta}_t \in \mathcal{E}_t^\delta(\hat{\theta}_t) = \left\{ \theta \in \mathbb{R}^d \textit{ s.t } \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\tilde{\mathbf{V}}_t^{-1}} \leq \beta_t(\delta) \right\}.$$

3. Note the difference between $\hat{\theta}_t$ and $\bar{\theta}_t$, where the rewards r_{t+1} are replaced by their conditional expected values $\mu\left(\langle x_s, \theta_\star^s\rangle\right)$

Lemma 1 ensures that with high probability the set $\mathcal{E}_t^{\delta}(\hat{\theta}_t)$ is a *confidence set* for $\bar{\theta}_t$. A complete proof of this result is deferred to Section A.1 in the supplementary material.

Lemma 2 *[Tracking] Let $D \in \mathbb{N}^*$. The following holds:*

$$\|g_t(\bar{\theta}_t) - g_t(\theta_\star^t)\|_{\mathbf{V_t^{-2}}} \leq \frac{2k_\mu L^2 S}{\lambda} \frac{\gamma^D}{1-\gamma} + k_\mu \sum_{s=t-D}^{t-1} \left\|\theta_\star^s - \theta_\star^{s+1}\right\|_2 .$$

Lemma 2 effectively links the deviation of $\bar{\theta}_t$ from $\theta_\star^t$ to the variation-budget B_T through the drift $\sum_{s=t-D}^{t-1} \left\|\theta_\star^s - \theta_\star^{s+1}\right\|_2$. The proof of this result borrows tools from Russac et al. (2019) and is deferred to Section A.3 in the supplementary material. The integer D appearing in Lemma 2 is introduced for the sake of the analysis only. It allows to treat separately old and recent observations. We provide its optimal value later in this section.

Remark 3 *Behind the statement of Lemma 1 and Lemma 2 hides the main reason why the projection step of Filippi et al. (2010) needs to be generalized. Indeed, it appears that the deviations $(\hat{\theta}_t \leftrightarrow \bar{\theta}_t)$ and $(\bar{\theta}_t \leftrightarrow \theta_\star^t)$ are controlled through different metrics ($\widetilde{\mathbf{V}}_\mathbf{t}^{-\mathbf{1}}$ and $\mathbf{V_t^{-2}}$, respectively). Projecting according to the first metric would corrupt the control of the second deviation, and conversely.*

Regret decomposition and prediction error. To bound the instantaneous regret at round t, we rely on the prediction error Δ_t defined as follows for any arm $x \in \mathcal{X}_t$:

$$\Delta_t(x) := \left|\mu\left(\langle x, \tilde{\theta}_t\rangle\right) - \mu\Big(\langle x, \theta_\star^t\rangle\Big)\right| .$$

The next Lemma ties the cumulative pseudo-regret to the sum of prediction errors. This derivation is classical and the proof is deferred to Section B.1 in the supplementary material.

Lemma 3 *The following holds:*

$$R_T \leq 2R_\mu \sum_{t=1}^{T} \beta_t(\delta) \left[\|x_t\|_{\mathbf{V_t^{-1}}} - \|x_\star^t\|_{\mathbf{V_t^{-1}}}\right] + \sum_{t=1}^{T} \left[\Delta_t(x_t) + \Delta_t(x_\star^t)\right] .$$

Thanks to Lemma 3 we are left to characterize the prediction error $\Delta_t(x)$ for any $x \in \mathcal{X}_t$. Following Filippi et al. (2010), we rely on the mean-value theorem to ensure that it exists $\mathring{\theta}_t \in [\tilde{\theta}_t, \theta_\star^t]$ such that[4]:

$$\Delta_t(x) \leq k_\mu \left\langle x, \mathbf{H_t}(\mathring{\theta}_t)\left(g_t(\tilde{\theta}_t) - g_t(\theta_\star^t)\right)\right\rangle , \tag{5}$$

where $\mathbf{H_t}(\theta) := \sum_{s=1}^{t-1} \dot{\mu}(\langle x_s, \theta\rangle) x_s x_s^\mathsf{T} + \lambda c_\mu \mathbf{I_d}$. Since $\tilde{\theta}_t, \theta_\star^t \in \Theta$, we obtain by convexity that $\mathring{\theta}_t \in \Theta$ and we can use the lower bound $\mathbf{H_t}(\mathring{\theta}_t) \succeq c_\mu \mathbf{V_t}$.

4. Formally, $\mathring{\theta}_t \in [\tilde{\theta}_t, \theta_\star^t]$ means that there exists $v \in [0,1]$ such that $\mathring{\theta}_t = v\tilde{\theta}_t + (1-v)\theta_\star^t$.

Remark 4 *In this last inequality resides the mistake that was made in previous extension of Filippi et al. (2010) to the non-stationary setting (Cheung et al., 2019a; Zhao et al., 2020). Indeed, if the prediction error is measured at $\hat{\theta}_t$, we are left with $\mathring{\theta}_t \in [\theta_\star^t, \hat{\theta}_t]$, and $\mathring{\theta}_t$ can lie outside of the admissible set Θ (since $\hat{\theta}_t$ can). The lower-bound linking $\mathbf{H_t}(\mathring{\theta}_t)$ and $\mathbf{V_t}$ would therefore not hold. More precisely, and as detailed in Section 3.2, when $\mathring{\theta}_t \in [\theta_\star^t, \hat{\theta}_t]$ not much can be said on the link between $\mathbf{H_t}(\mathring{\theta}_t)$ and $\mathbf{V_t}$ without severely degrading the final regret guarantees.*

Adding and removing $g_t(\hat{\theta}_t) + g_t(\theta_t^p) + g_t(\bar{\theta}_t)$ inside the inner-product in Equation (5), followed by easy manipulations yields:

$$\Delta_t(x) \leq \underbrace{R_\mu \|x\|_{\mathbf{V_t^{-1}}} \left(\left\| g_t(\tilde{\theta}_t) - g_t(\theta_t^p) \right\|_{\tilde{\mathbf{V}}_\mathbf{t}^{-1}} + \left\| g_t(\bar{\theta}_t) - g_t(\hat{\theta}_t) \right\|_{\tilde{\mathbf{V}}_\mathbf{t}^{-1}} \right)}_{:=\Delta_t^{\text{learn}}(x)}$$
$$+ \underbrace{R_\mu \|x\|_2 \left(\left\| g_t(\theta_t^p) - g_t(\hat{\theta}_t) \right\|_{\mathbf{V_t^{-2}}} + \left\| g_t(\bar{\theta}_t) - g_t(\theta_\star^t) \right\|_{\mathbf{V_t^{-2}}} \right)}_{:=\Delta_t^{\text{track}}(x)} .$$

Leveraging the projection step We can now bound the terms $\Delta_t^{\text{learn}}(x)$ and $\Delta_t^{\text{track}}(x)$ separately. Lemma 1 along with the design $\tilde{\theta}_t \in \mathcal{E}_t^\delta(\theta_t^p)$ leads to:

$$\Delta_t^{\text{learn}}(x) \leq 2R_\mu \|x\|_{\mathbf{V_t^{-1}}} \beta_t(\delta) \quad \text{w.h.p} \tag{6}$$

The first term in $\Delta_t^{\text{track}}(x)$ is kept under control by the specific design of the projection step (**P1**). This is formalized in the following Lemma, whose proof is deferred to Section A.4 in the appendix.

Lemma 4 *Under the event $\{\bar{\theta}_t \in \mathcal{E}_t^\delta(\hat{\theta}_t)\}$ the following holds:*

$$\|g_t(\theta_t^p) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \leq \|g_t(\bar{\theta}_t) - g_t(\theta_\star^t)\|_{\mathbf{V_t^{-2}}} .$$

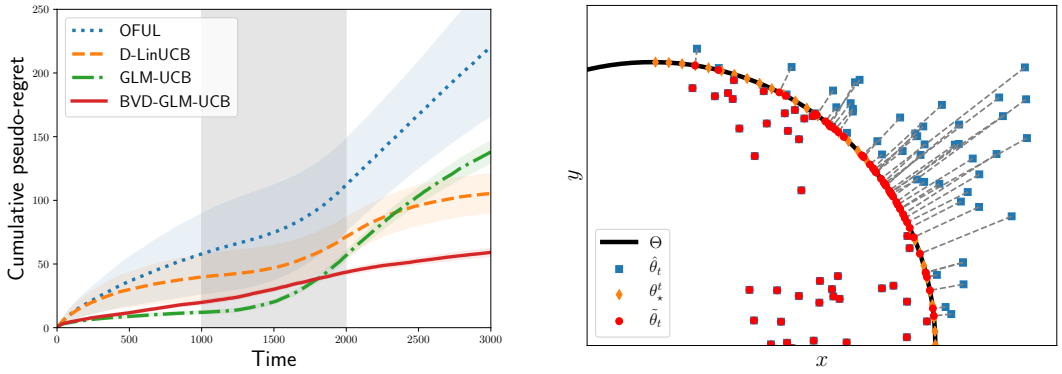
As a result, bounding $\Delta_t^{\text{track}}(x)$ reduces to bounding $\|g_t(\bar{\theta}_t) - g_t(\theta_\star^t)\|_{\mathbf{V_t^{-2}}}$. Combined with Lemma 2, this result states that the deviation between θ_t^p and $\hat{\theta}_t$ is characterized by B_t, the parameter-drift up to round t, as illustrated in Figure 1. This leads to:

$$\Delta_t^{\text{track}}(x) \leq 2R_\mu \|x\|_2 \left(\frac{2k_\mu L^2 S}{\lambda} \frac{\gamma^D}{1-\gamma} + k_\mu \sum_{s=t-D}^{t-1} \left\| \theta_\star^s - \theta_\star^{s+1} \right\|_2 \right) \quad \text{w.h.p} \tag{7}$$

Putting everything together. Combining Equations (6) and (7) with Lemma 3 and the Elliptical Lemma (Lemma 6 in the supplementary material) yields:

$$R_T \leq C_1 R_\mu dT \log(1/\gamma) + C_2 R_\mu \gamma^D T/(1-\gamma) + C_3 R_\mu D B_T \qquad \text{w.h.p}$$

where the constants C_1, C_2 and C_3 hide $\log(T)$ multiplicative dependencies. A detailed proof of this result is deferred to Section B.2 in the supplementary material. Setting the hyper-parameters $D = \log(T)/(1-\gamma)$ and $\gamma = 1 - (\frac{B_T}{dT})^{2/3}$ concludes the proof of Theorem 1.



(*a*) Regret bounds of different stochastic bandit algorithms under parameter-drift. The grey region indicates a smooth drift of $\theta_\star^t$.

(*b*) Evolution of the parameters of interest $(\theta_\star^t, \hat{\theta}_t, \tilde{\theta}_t)$ for `BVD-GLM-UCB`. Note that in this non-stationary setting $\hat{\theta}_t \notin \Theta$ is frequent.

Figure 2: Numerical simulations in a non-stationary logistic setting. For the first figure, results are average over 50 independent runs and shaded areas represent one standard-deviation variation.

6. Experiments

We illustrate in Figure 2 the behavior and performance of `BVD-GLM-UCB` with numerical simulations in a two-dimensional non-stationary logistic environment. Formally, we let $r_{t+1} \sim \text{Bernoulli}(\mu(\langle x_t, \theta_\star^t \rangle))$ where $\mu(z) = (1+e^{-z})^{-1}$ is the logistic function. The sequence $\{\theta_\star^t\}_{t\geq 1}$ evolves as follows: we let $\theta_\star^t = (0,1)$ for $t \in [1, T/3]$. Between $t = T/3$ and $t = 2T/3$ we smoothly rotate $\theta_\star^t$ from $(0,1)$ to $(1,0)$. Finally we let $\theta_\star^t = (0,1)$ for $t \in [2T/3, T]$. A thorough description of the experimental setting can be found in Appendix F. We compare in Figure 2(*a*) the four following algorithms: `OFUL` (Abbasi-Yadkori et al., 2011) (stationary, here mispecified), `GLM-UCB` (Filippi et al., 2010) (stationary, here well-specified), `D-LinUCB` (Russac et al., 2019) (an exponentially weighted LB algorithm, non-stationary but here mispecified) and `BVD-GLM-UCB` (non-stationary, well-specified). For `D-LinUCB` and `BVD-GLM-UCB` we use the value of γ recommended by the asymptotic analysis. This figure highlights the necessity to employ algorithms that are well-specified; both `GLM-UCB` and `BVD-GLM-UCB` outperform their linear counterparts (`OFUL` and `D-LinUCB`, respectively). Note that an appropriate treatment of non-stationarity is also crucial to obtain small regret as for the considered horizon the two best performing algorithms are `D-LinUCB` and `BVD-GLM-UCB`. The latter being well-specified and resilient to non-stationary, it naturally performs best. In Figure 2(*b*) we highlight the fact that the projection step is necessary as, in this non-stationary setting, $\hat{\theta}_t$ regularly leaves the admissible set Θ.

7. Conclusion and future work

We highlight in this paper a central difficulty in the theoretical treatment of non-stationary GLBs, overlooked in existing approaches and intimately linked to the non-linear nature of the reward function.

To overcome this difficulty, we introduce a generalization of the projection step from (Filippi et al., 2010), which allows to simultaneously *track* the non-stationary ground-truth while preserving the *learning* guarantees of weighted maximum-likelihood strategies. This novel algorithmic design along with a careful analysis proves that an order-optimal (w.r.t d, T and B_T) regret-bound can be achieved for GLBs under parameter-drift.

We underlined in Section 3.2 the problematic scaling of the problem-dependent constant R_μ. Consequent research efforts have recently been deployed to reduce its impact on regret-bounds, both in the stationary (Faury et al., 2020; Abeille et al., 2020; Jun et al., 2020) and piece-wise stationary (Russac et al., 2020b) settings. What is the optimal dependency w.r.t R_μ in the more general parameter-drift setting, and how it can be achieved are exciting open questions that we here leave for future work.

Acknowledgments

The authors would like to thank the anonymous referees which insightful and constructive comments helped improving the clarity of this manuscript. LF also thanks Olivier Fercoq for his helpful observations regarding the derivation of the simplified projection step. YR thanks Arnaud Buisson and Jianjun Yuan for fruitful discussion on the BOB framework and its extension using discount factors.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Marc Abeille and Alessandro Lazaric. Linear Thompson Sampling Revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-Wise Minimax-Optimal Algorithms for Logistic Bandits. *arXiv preprint arXiv:2010.12642*, 2020.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic Multi-Armed-Bandit problem with Non-Stationary Rewards. In *Advances in Neural Information Processing Systems*, pages 199–207, 2014.

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 22rd International Conference on Artificial Intelligence and Statistics*, pages 1079–1087, 2019b.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic Linear Optimization under Bandit Feedback. In *COLT*, 2008.

Shi Dong and Benjamin Van Roy. An Information-Theoretic Analysis for Thompson Sampling with Many Actions. In *Advances in Neural Information Processing Systems*, pages 4157–4165, 2018.

Bianca Dumitrascu, Karen Feng, and Barbara Engelhardt. PG-TS: Improved Thompson Sampling for Logistic Contextual Bandits. In *Advances in Neural Information Processing Systems*, pages 4624–4633, 2018.

Louis Faury, Marc Abeille, Clement Calauzenes, and Olivier Fercoq. Improved Optimistic Algorithms for Logistic Bandits. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3052–3060, Virtual, 13–18 Jul 2020. PMLR.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric Bandits: The Generalized Linear Case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

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

Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable Generalized Linear Bandits: Online Computation and Hashing. In *Advances in Neural Information Processing Systems*, pages 99–109, 2017.

Kwang-Sung Jun, Lalit Jain, and Houssam Nassif. Improved Confidence Bounds for the Linear Logistic Model and Applications to Linear Bandits. *arXiv preprint arXiv:2011.11222*, 2020.

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-Volume 70*, pages 2071–2080. JMLR. org, 2017.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted Linear Bandits for Non-Stationary Environments. In *Advances in Neural Information Processing Systems*, pages 12017–12026, 2019.

Yoan Russac, Olivier Cappé, and Aurélien Garivier. Algorithms for Non-Stationary Generalized Linear Bandits. *arXiv preprint arXiv:2003.10113*, 2020a.

Yoan Russac, Louis Faury, Olivier Cappé, and Aurélien Garivier. Self-Concordant Analysis of Generalized Linear Bandits with Forgetting. *arXiv preprint arXiv:2011.00819*, 2020b.

Daniel Russo and Benjamin Van Roy. Learning to Optimize via Posterior Sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

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*, volume 2020, 2020.

Organization of the appendix

The appendix is organized as follows:

- In Section A we provide some concentration results, along with a bound on the prediction error Δ_t inherited from the design of the projection step.
- In Section B we link the prediction error Δ_t to the regret R_T of BVD-GLM-UCB. We then proceed to prove the bound on R_T announced in Theorem 1.
- In Section C we provide a proof for the equivalence of the optimization programs (**P1**) (along with the computation of $\tilde{\theta}_t$) and (**P2**).
- Section D contains some secondary lemmas needed for the analysis, such as a version of the Elliptical Lemma for weighted matrices.
- In Section E we provide a proof for the regret upper-bound of BOB-BVD-GLM-UCB claimed in Theorem 2.
- Finally, in Section F we provide some details on our numerical simulations.

Appendix A. Concentration and predictions bound

A.1. Confidence sets

Lemma 1 *[Learning] Let $\delta \in (0, 1]$. With probability at least $1 - \delta$:*

$$\textit{for all } t \geq 1, \quad \bar{\theta}_t \in \mathcal{E}_t^{\delta}(\hat{\theta}_t) = \left\{ \theta \in \mathbb{R}^d \; s.t \; \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{\widetilde{\mathbf{V}}_t^{-1}} \leq \beta_t(\delta) \right\}.$$

Proof Recall that:

$$\mathcal{E}_t^{\delta}(\hat{\theta}_t) = \left\{ \theta \in \mathbb{R}^d \text{ s.t } \|g_t(\theta) - g_t(\hat{\theta}_t)\|_{\widetilde{\mathbf{V}}_t^{-1}} \leq \beta_t(\delta) \right\} ,$$

where

$$\beta_t(\delta) = \sqrt{\lambda} c_\mu S + \sigma \sqrt{2 \log(1/\delta) + d \log \left(1 + \frac{L^2(1 - \gamma^{2t})}{\lambda d (1 - \gamma^2)} \right)} .$$

Also, from the definition of $\bar{\theta}_t$ in Equation (4), by setting to 0 the differential of the convex objective minimized by $\bar{\theta}_t$ we obtain that:

$$g_t(\bar{\theta}_t) = \sum_{s=1}^{t-1} \gamma^{t-1-s} \mu \left(\langle \theta_\star^s, x_s \rangle \right) x_s + \lambda c_\mu \theta_\star^t . \tag{8}$$

Further, for all $s \geq 1$, define

$$\epsilon_{s+1} = r_{s+1} - \mu(\langle \theta_\star^s, x_s \rangle) . \tag{9}$$

Let $\tilde{\mathcal{F}}_s = \sigma(x_1, r_2, .., x_{s-1}, r_s, x_s)$, which compared to $\mathcal{F}_s$ includes the arm x_s. Note that:

$$\begin{cases} \mathbb{E}\left[\epsilon_{s+1}\middle|\tilde{\mathcal{F}}_s\right] = 0 & \text{(Equation (1))} \\ -\mu(\langle\theta^s_\star, x_s\rangle) \leq \epsilon_{s+1} \leq 2\sigma + \mu(\langle\theta^s_\star, x_s\rangle) \quad \text{a.s} & \text{(Assumption 2)} \end{cases}$$

Therefore ϵ_{s+1} is σ-subGaussian conditionally on $\tilde{\mathcal{F}}_s$. Furthermore, by optimality of $\hat{\theta}_t$, differentiating the objective function in Equation (2) yields:

$$\begin{aligned} &\sum_{s=1}^{t-1} \gamma^{t-1-s}\left[\mu(\langle\hat{\theta}_t, x_s\rangle) - r_{s+1}\right] x_s + \lambda c_\mu \hat{\theta}_t = 0 \\ &\Leftrightarrow g_t(\hat{\theta}_t) = \sum_{s=1}^{t-1} \gamma^{t-1-s}\mu(\langle\theta^s_\star, x_s\rangle)x_s + \sum_{s=1}^{t-1}\gamma^{t-1-s}\epsilon_{s+1}x_s && \text{(Equation (9))} \\ &\Leftrightarrow g_t(\hat{\theta}_t) = g_t(\bar{\theta}_t) + \sum_{s=1}^{t-1}\gamma^{t-1-s}\epsilon_{s+1}x_s - \lambda c_\mu \theta^t_\star && \text{(Equation (8))} \qquad (10) \\ &\Leftrightarrow \|g_t(\bar{\theta}_t) - g_t(\hat{\theta}_t)\|_{\widetilde{\mathbf{V}}_\mathbf{t}^{-1}} = \left\|\sum_{s=1}^{t-1}\gamma^{t-1-s}\epsilon_{s+1}x_s - \lambda c_\mu \theta^t_\star\right\|_{\widetilde{\mathbf{V}}_\mathbf{t}^{-1}} . \end{aligned}$$

Therefore since $\theta^t_\star \in \Theta$ and $\widetilde{\mathbf{V}}_\mathbf{t} \succeq \lambda \mathbf{I_d}$ we obtain:

$$\|g_t(\bar{\theta}_t) - g_t(\hat{\theta}_t)\|_{\widetilde{\mathbf{V}}_\mathbf{t}^{-1}} \leq \sqrt{\lambda} c_\mu S + \left\|\sum_{s=1}^{t-1}\gamma^{t-1-s}\epsilon_{s+1}x_s\right\|_{\widetilde{\mathbf{V}}_\mathbf{t}^{-1}} .$$

Simplifying the factors γ^{t-1} in the most right term and applying Proposition 1 of Russac et al. (2019) proves that with probability at least $1-\delta$, for all $t \geq 1$:

$$\|g_t(\bar{\theta}_t) - g_t(\hat{\theta}_t)\|_{\widetilde{\mathbf{V}}_\mathbf{t}^{-1}} \leq \sqrt{\lambda} c_\mu S + \sigma\sqrt{2\log(1/\delta) + d\log\left(1 + \frac{L^2(1-\gamma^{2t})}{\lambda d(1-\gamma^2)}\right)} = \beta_t(\delta) ,$$

hence proving the desired result. ■

A.2. Bounding the prediction error

Lemma 5 *Let $\delta \in (0,1]$ and $D \in \mathbb{N}^*$. With probability at least $1-\delta$: for all $t \geq 1$, for all $x \in \mathcal{X}_t$ the following holds.*

$$\Delta_t(x) \leq \frac{2k_\mu}{c_\mu}\beta_t(\delta)\|x\|_{\mathbf{V}_\mathbf{t}^{-1}} + \frac{4k_\mu^2 L^3 S}{c_\mu \lambda}\frac{\gamma^D}{(1-\gamma)} + \frac{2k_\mu^2 L}{c_\mu}\sum_{s=t-D}^{t-1}\left\|\theta^s_\star - \theta^{s+1}_\star\right\|_2 .$$

Proof In the following, we assume that the event $E_\delta = \{\bar{\theta}_t \in \mathcal{E}_t^\delta(\hat{\theta}_t) \text{ for all } t \geq 1\}$ holds, which happens with probability at least $1-\delta$ (Lemma 1). From the definition of the prediction error:

$$\begin{aligned}
\Delta_t(x) &= \left|\mu(\langle x, \tilde{\theta}_t\rangle) - \mu(\langle x, \theta_\star^t\rangle)\right| \\
&\leq \left(\sup_{x\in\mathcal{X},\theta\in\Theta} \dot{\mu}\left(\langle x,\theta\rangle\right)\right)\left|\langle x, \tilde{\theta}_t - \theta_\star^t\rangle\right| \qquad (x\in\mathcal{X}_t, \theta_\star^t\in\Theta, \tilde{\theta}_t\in\Theta) \\
&\leq k_\mu \left|\langle x, \tilde{\theta}_t - \theta_\star^t\rangle\right| \ . \qquad \text{(by definition of } k_\mu) \qquad (11)
\end{aligned}$$

Further, thanks to the mean value theorem:

$$\begin{aligned}
g_t(\tilde{\theta}_t) - g_t(\theta_\star^t) &= \sum_{s=1}^{t-1}\gamma^{t-1-s}\left[\mu(\langle\tilde{\theta}_t, x_s\rangle) - \mu(\langle\theta_\star^t, x_s\rangle)\right] + \lambda c_\mu(\tilde{\theta}_t - \theta_\star^t) \\
&= \sum_{s=1}^{t-1}\gamma^{t-1-s}\left[\int_{v=0}^1 \dot{\mu}\left(\langle x_s, (1-v)\theta_\star^t + v\tilde{\theta}_t\rangle\right) dv\right] x_s x_s^\mathsf{T}(\tilde{\theta}_t - \theta_\star^t) + \lambda c_\mu(\tilde{\theta}_t - \theta_\star^t) \\
&= \mathbf{G_t}\cdot(\tilde{\theta}_t - \theta_\star^t) \ , \qquad (12)
\end{aligned}$$

where:

$$\mathbf{G_t} := \sum_{s=1}^{t-1}\gamma^{t-1-s}\left[\int_{v=0}^1 \dot{\mu}\left(\langle x_s, (1-v)\theta_\star^t + v\tilde{\theta}_t\rangle\right) dv\right] x_s x_s^\mathsf{T} + \lambda c_\mu \mathbf{I_d} \succeq c_\mu \mathbf{V_t} \ .$$

Note that because $x_s \in \mathcal{X}$ for all $s \in [t-1]$ and $\tilde{\theta}_t, \theta_\star^t \in \Theta$ we have $\mathbf{G_t} \geq c_\mu \mathbf{V_t}$. Assembling together Equations (11) and (12) we get:

$$\begin{aligned}
\Delta_t(x) &\leq k_\mu \left|\left\langle x, \mathbf{G_t^{-1}}(g_t(\tilde{\theta}_t) - g_t(\theta_\star^t))\right\rangle\right| \\
&\leq k_\mu \left|\left\langle x, \mathbf{G_t^{-1}}(g_t(\tilde{\theta}_t) - g_t(\theta_t^p) + g_t(\theta_t^p) - g_t(\hat{\theta}_t) + g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta_\star^t))\right\rangle\right| \\
&\leq \underbrace{k_\mu \left|\left\langle x, \mathbf{G_t^{-1}}(g_t(\tilde{\theta}_t) - g_t(\theta_t^p) + g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t))\right\rangle\right|}_{:=\Delta_t^{\text{learn}}(x)} \\
&\quad + \underbrace{k_\mu \left|\left\langle x, \mathbf{G_t^{-1}}(g_t(\theta_t^p) - g_t(\hat{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta_\star^t))\right\rangle\right|}_{:=\Delta_t^{\text{track}}(x)} \\
&\leq \Delta_t^{\text{learn}}(x) + \Delta_t^{\text{track}}(x) \ . \qquad (13)
\end{aligned}$$

This decomposition brings out the contribution of two different phenomenons (*learning* and *tracking*) which will be handled separately. Starting with the learning:

$$
\begin{aligned}
\Delta_t^{\text{learn}}(x) &= k_\mu \left| \left\langle x, \mathbf{G_t^{-1}}(g_t(\tilde{\theta}_t) - g_t(\theta_t^p) + g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t)) \right\rangle \right| \\
&= k_\mu \left| \left\langle \widetilde{\mathbf{V}}_\mathbf{t}^{\mathbf{1/2}} \mathbf{G_t^{-1}} x, \widetilde{\mathbf{V}}_\mathbf{t}^{\mathbf{-1/2}}(g_t(\tilde{\theta}_t) - g_t(\theta_t^p) + g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t)) \right\rangle \right| \\
&\leq k_\mu \|x\|_{\mathbf{G_t^{-1}}\widetilde{\mathbf{V}}_\mathbf{t}\mathbf{G_t^{-1}}} \left(\|g_t(\tilde{\theta}_t) - g_t(\theta_t^p)\|_{\widetilde{\mathbf{V}}_\mathbf{t}^{-1}} + \|g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t)\|_{\widetilde{\mathbf{V}}_\mathbf{t}^{-1}} \right) && \text{(Cauchy-Schwarz)} \\
&\leq k_\mu \|x\|_{\mathbf{G_t^{-1}}\mathbf{V_t}\mathbf{G_t^{-1}}} \left(\|g_t(\tilde{\theta}_t) - g_t(\theta_t^p)\|_{\widetilde{\mathbf{V}}_\mathbf{t}^{-1}} + \|g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t)\|_{\widetilde{\mathbf{V}}_\mathbf{t}^{-1}} \right) && (\widetilde{\mathbf{V}}_\mathbf{t} \leq \mathbf{V}_t) \\
&\leq \frac{k_\mu}{\sqrt{c_\mu}} \|x\|_{\mathbf{G_t^{-1}}} \left(\|g_t(\tilde{\theta}_t) - g_t(\theta_t^p)\|_{\widetilde{\mathbf{V}}_\mathbf{t}^{-1}} + \|g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t)\|_{\widetilde{\mathbf{V}}_\mathbf{t}^{-1}} \right) && (\mathbf{V_t} \leq c_\mu^{-1}\mathbf{G_t}) \\
&\leq \frac{k_\mu}{c_\mu} \|x\|_{\mathbf{V_t^{-1}}} \left(\|g_t(\tilde{\theta}_t) - g_t(\theta_t^p)\|_{\widetilde{\mathbf{V}}_\mathbf{t}^{-1}} + \|g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t)\|_{\widetilde{\mathbf{V}}_\mathbf{t}^{-1}} \right) && (\mathbf{G_t^{-1}} \leq c_\mu^{-1}\mathbf{V_t^{-1}}) \\
&\leq \frac{k_\mu}{c_\mu} \|x\|_{\mathbf{V_t^{-1}}} \left(\beta_t(\delta) + \|g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t)\|_{\widetilde{\mathbf{V}}_\mathbf{t}^{-1}} \right) && (\tilde{\theta}_t \in \mathcal{E}_t^\delta(\theta_t^p)) \\
&\leq \frac{k_\mu}{c_\mu} \|x\|_{\mathbf{V_t^{-1}}} \left(\beta_t(\delta) + \beta_t(\delta) \right) . && (E_\delta \text{ holds})
\end{aligned}
$$

We used $\widetilde{\mathbf{V}}_\mathbf{t} \leq \mathbf{V}_t$ which is a consequence of $\gamma \in (0, 1)$. As a result:

$$
\Delta_t^{\text{learn}}(x) \leq \frac{2k_\mu}{c_\mu} \beta_t(\delta) \|x\|_{\mathbf{V_t^{-1}}} . \tag{14}
$$

Before bounding the tracking term, we state two technical lemmas that will be proven in Section A.3 and A.4, respectively.

Lemma 2 *[Tracking] Let $D \in \mathbb{N}^*$. The following holds:*

$$
\|g_t(\bar{\theta}_t) - g_t(\theta_\star^t)\|_{\mathbf{V_t^{-2}}} \leq \frac{2k_\mu L^2 S}{\lambda} \frac{\gamma^D}{1-\gamma} + k_\mu \sum_{s=t-D}^{t-1} \left\| \theta_\star^s - \theta_\star^{s+1} \right\|_2 .
$$

Lemma 4 *Under the event $\{\bar{\theta}_t \in \mathcal{E}_t^\delta(\hat{\theta}_t)\}$ the following holds:*

$$
\|g_t(\theta_t^p) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \leq \|g_t(\bar{\theta}_t) - g_t(\theta_\star^t)\|_{\mathbf{V_t^{-2}}} .
$$

We now bound the tracking term:

$$
\begin{aligned}
\Delta_t^{\text{track}}(x) &= k_\mu \left| \left\langle x, \mathbf{G_t}^{-1}(g_t(\theta_t^p) - g_t(\hat{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta_\star^t)) \right\rangle \right| \\
&\leq k_\mu \|x\|_2 \left\| g_t(\theta_t^p) - g_t(\hat{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta_\star^t) \right\|_{\mathbf{G}_t^{-2}} && \text{(Cauchy-Schwarz)} \\
&\leq \frac{k_\mu L}{c_\mu} \left\| g_t(\theta_t^p) - g_t(\hat{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta_\star^t) \right\|_{\mathbf{V}_t^{-2}} && (\|x\|_2 \leq L, \mathbf{G_t} \succeq c_\mu \mathbf{V_t}) \\
&\leq \frac{k_\mu L}{c_\mu} \left(\left\| g_t(\theta_t^p) - g_t(\hat{\theta}_t) \right\|_{\mathbf{V}_t^{-2}} + \left\| g_t(\bar{\theta}_t) - g_t(\theta_\star^t) \right\|_{\mathbf{V}_t^{-2}} \right) && \text{(Triangle inequality)} \\
&\leq \frac{2k_\mu L}{c_\mu} \left\| g_t(\bar{\theta}_t) - g_t(\theta_\star^t) \right\|_{\mathbf{V}_t^{-2}} && \text{(Lemma 4)}
\end{aligned}
$$

Thanks to Lemma 2, we finally obtain,

$$\Delta_t^{\text{track}}(x) \leq \frac{4k_\mu^2 L^3 S}{c_\mu \lambda} \frac{\gamma^D}{(1-\gamma)} + \frac{2k_\mu^2 L}{c_\mu} \sum_{s=t-D}^{t-1} \left\|\theta_\star^s - \theta_\star^{s+1}\right\|_2 . \tag{15}$$

Assembling Equations (13), (14) and (15) finishes the proof. ■

A.3. Proof of Lemma 2

Lemma 2 *[Tracking] Let $D \in \mathbb{N}^*$. The following holds:*

$$\|g_t(\bar{\theta}_t) - g_t(\theta_\star^t)\|_{\mathbf{V_t^{-2}}} \leq \frac{2k_\mu L^2 S}{\lambda} \frac{\gamma^D}{1-\gamma} + k_\mu \sum_{s=t-D}^{t-1} \left\|\theta_\star^s - \theta_\star^{s+1}\right\|_2 .$$

Proof Thanks to Equation (8) we have:

$$\begin{aligned}
&g_t(\bar{\theta}_t) = \sum_{s=1}^{t-1} \gamma^{t-1-s} \mu(\langle x_s, \theta_\star^s \rangle) x_s + \lambda c_\mu \theta_\star^t \\
&\Leftrightarrow g_t(\bar{\theta}_t) - g_t(\theta_\star^t) = \sum_{s=1}^{t-1} \gamma^{t-1-s} \left[\mu(\langle x_s, \theta_\star^s \rangle) - \mu(\langle x_s, \theta_\star^t \rangle)\right] x_s \\
&\Leftrightarrow g_t(\bar{\theta}_t) - g_t(\theta_\star^t) = \sum_{s=1}^{t-1} \gamma^{t-1-s} \left[\int_{v=0}^{1} \dot{\mu}\left(\left\langle x_s, v\theta_\star^t + (1-v)\theta_\star^s \right\rangle\right) dv\right] x_s x_s^\mathsf{T} (\theta_\star^s - \theta_\star^t) \quad \text{(mean-value theorem)} \\
&\Leftrightarrow g_t(\bar{\theta}_t) - g_t(\theta_\star^t) = \sum_{s=1}^{t-1} \gamma^{t-1-s} \alpha_s x_s x_s^\mathsf{T} (\theta_\star^s - \theta_\star^t) ,
\end{aligned}$$

where we defined:

$$\alpha_s := \int_{v=0}^{1} \dot{\mu}\left(\left\langle x_s, v\theta_\star^t + (1-v)\theta_\star^s \right\rangle\right) dv \in [c_\mu, k_\mu] .$$

Therefore:

$$\|g_t(\bar{\theta}_t) - g_t(\theta_\star^t)\|_{\mathbf{V_t^{-2}}} = \left\| \sum_{s=1}^{t-1} \gamma^{t-1-s} \alpha_s x_s x_s^\mathsf{T} (\theta_\star^s - \theta_\star^t) \right\|_{\mathbf{V_t^{-2}}} .$$

The rest of the proof follows the strategy of Russac et al. (2019) to yield the announced result. Let $D \in \mathbb{N}^*$ and notice that:

$$\begin{aligned}
\left\| \sum_{s=1}^{t-1} \gamma^{t-1-s} \alpha_s x_s x_s^\mathsf{T} (\theta_\star^s - \theta_\star^t) \right\|_{\mathbf{V_t^{-2}}} \leq & \underbrace{\left\| \sum_{s=1}^{t-D-1} \gamma^{t-1-s} \alpha_s x_s x_s^\mathsf{T} (\theta_\star^s - \theta_\star^t) \right\|_{\mathbf{V_t^{-2}}}}_{:=d_1} \\
& + \underbrace{\left\| \sum_{s=t-D}^{t-1} \gamma^{t-1-s} \alpha_s x_s x_s^\mathsf{T} (\theta_\star^s - \theta_\star^t) \right\|_{\mathbf{V_t^{-2}}}}_{:=d_2} .
\end{aligned}$$

Both terms are bounded separately; starting with d_1:

$$\begin{aligned}
d_1 &\leq \lambda^{-1} \left\| \sum_{s=1}^{t-D-1} \gamma^{t-1-s} \alpha_s x_s x_s^\mathsf{T} (\theta_\star^s - \theta_\star^t) \right\| && (\mathbf{V_t} \geq \lambda \mathbf{I_d}) \\
&\leq \lambda^{-1} \sum_{s=1}^{t-D-1} \gamma^{t-1-s} |\alpha_s| \left\| x_s x_s^\mathsf{T} (\theta_\star^s - \theta_\star^t) \right\| && \text{(Triangle inequality)} \\
&\leq 2k_\mu \lambda^{-1} S L^2 \sum_{s=1}^{t-D-1} \gamma^{t-1-s} && (\|x_s\|_2 \leq L,\, \theta_\star^s, \theta_\star^t \in \Theta, |\alpha_s| \leq k_\mu) \\
&\leq 2k_\mu \lambda^{-1} S L^2 \gamma^D (1-\gamma)^{-1} \,.
\end{aligned}$$

And for d_2:

$$\begin{aligned}
d_2 &= \left\| \mathbf{V_t^{-1}} \sum_{s=t-D}^{t-1} \gamma^{t-1-s} \alpha_s x_s x_s^\mathsf{T} (\theta_\star^s - \theta_\star^t) \right\| \\
&= \left\| \sum_{s=t-D}^{t-1} \mathbf{V_t^{-1}} \gamma^{t-1-s} \alpha_s x_s x_s^\mathsf{T} (\theta_\star^s - \theta_\star^t) \right\| \\
&= \left\| \sum_{s=t-D}^{t-1} \mathbf{V_t^{-1}} \gamma^{t-1-s} \alpha_s x_s x_s^\mathsf{T} \sum_{p=s}^{t-1} \left(\theta_\star^p - \theta_\star^{p+1} \right) \right\| && \text{(Telescopic sum)} \\
&\leq \left\| \sum_{p=t-D}^{t-1} \mathbf{V_t^{-1}} \sum_{s=t-D}^{p} \gamma^{t-1-s} \alpha_s x_s x_s^\mathsf{T} \left(\theta_\star^p - \theta_\star^{p+1} \right) \right\| && \text{(Re-arranging)} \\
&\leq \sum_{p=t-D}^{t-1} \left\| \mathbf{V_t^{-1}} \sum_{s=t-D}^{p} \gamma^{t-1-s} \alpha_s x_s x_s^\mathsf{T} \left(\theta_\star^p - \theta_\star^{p+1} \right) \right\| && \text{(Triangle inequality)} \\
&\leq \sum_{p=t-D}^{t-1} \lambda_{\max} \left(\mathbf{V_t^{-1}} \sum_{s=t-D}^{p} \gamma^{t-1-s} \alpha_s x_s x_s^\mathsf{T} \right) \|\theta_\star^p - \theta_\star^{p+1}\| \,.
\end{aligned}$$

Finishing the bound:

$$\begin{aligned}
\lambda_{\max}\left(\mathbf{V_t^{-1}}\sum_{s=t-D}^{p}\gamma^{t-1-s}\alpha_s x_s x_s^\mathsf{T}\right) &= \lambda_{\max}\left(\mathbf{V_t^{-1/2}}\left(\sum_{s=t-D}^{p}\gamma^{t-1-s}\alpha_s x_s x_s^\mathsf{T}\right)\mathbf{V_t^{-1/2}}\right)\\
&= \max_{\|x\|_2\le 1}\left\{x^T\left(\mathbf{V_t^{-1/2}}\sum_{s=t-D}^{p}\gamma^{t-1-s}\alpha_s x_s x_s^\mathsf{T}\mathbf{V_t^{-1/2}}\right)x\right\}\\
&= \max_{\|x\|_2\le 1}\left\{\sum_{s=t-D}^{p}\gamma^{t-1-s}\alpha_s\left(x_s^\mathsf{T}\mathbf{V_t^{-1/2}}x\right)^2\right\}\\
&\le k_\mu \max_{\|x\|_2\le 1}\left\{\sum_{s=t-D}^{p}\gamma^{t-1-s}\left(x_s^\mathsf{T}\mathbf{V_t^{-1/2}}x\right)^2\right\}\\
&= k_\mu\lambda_{\max}\left(\mathbf{V_t^{-1}}\sum_{s=t-D}^{p}\gamma^{t-1-s}x_s x_s^\mathsf{T}\right)\ .
\end{aligned}$$

Easy computations show that $\lambda_{\max}\left(\mathbf{V_t^{-1}}\sum_{s=t-D}^{p}\gamma^{t-1-s}x_s x_s^\mathsf{T}\right)\le 1$, which concludes the proof. ■

A.4. Proof of Lemma 4

Lemma 4 *Under the event* $\{\bar{\theta}_t \in \mathcal{E}_t^\delta(\hat{\theta}_t)\}$ *the following holds:*

$$\|g_t(\theta_t^p) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \le \|g_t(\bar{\theta}_t) - g_t(\theta_\star^t)\|_{\mathbf{V_t^{-2}}}\ .$$

Proof We prove this result by contradiction. Assume that:

$$\|g_t(\theta_t^p) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} > \|g_t(\bar{\theta}_t) - g_t(\theta_\star^t)\|_{\mathbf{V_t^{-2}}}, \tag{16}$$

For all $s \ge 1$ define:

$$\tilde{r}_{s+1} := \mu(\langle x_s, \theta_\star^t\rangle) + \epsilon_{s+1}\ , \tag{17}$$

where $\{\epsilon_s\}_s$ is defined in Equation (9). Further, let:

$$\theta_c := \arg\min_{\theta\in\mathbb{R}^d}\sum_{s=1}^{t-1}\gamma^{t-1-s}\left[b(\langle\theta, x_s\rangle) - \tilde{r}_{s+1}\langle\theta, x_s\rangle\right] + \frac{\lambda c_\mu}{2}\|\theta\|_2^2\ ,$$

which is well-defined as the minimizer of a strictly convex, coercive function. Upon differentiating we get:

$$\begin{aligned}
g_t(\theta_c) &= \sum_{s=1}^{t-1}\gamma^{t-1-s}\tilde{r}_{s+1}x_s\\
&= \sum_{s=1}^{t-1}\gamma^{t-1-s}\epsilon_{s+1}x_s + \sum_{s=1}^{t-1}\gamma^{t-1-s}\mu(\langle x_s, \theta_\star^t\rangle)x_s && \text{(Equation (17))}\\
&= g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t) + \lambda c_\mu\theta_\star^t + \sum_{s=1}^{t-1}\gamma^{t-1-s}\mu(\langle x_s, \theta_\star^t\rangle)x_s && \text{(Equation (10))}\\
&= g_t(\hat{\theta}_t) - g_t(\bar{\theta}_t) + g_t(\theta_\star^t)\ . && (18)
\end{aligned}$$

Therefore:

$$\begin{aligned}\|g_t(\theta_c) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} &= \|g_t(\bar{\theta}_t) - g_t(\theta_\star^t)\|_{\mathbf{V_t^{-2}}} \\ &< \|g_t(\theta_t^p) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \ . \qquad \text{(Equation 16)}\end{aligned}$$

Further from Equation (18) we get:

$$\begin{aligned}\|g_t(\theta_c) - g_t(\theta_\star^t)\|_{\mathbf{V_t^{-2}}} &= \|g_t(\bar{\theta}_t) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \\ &\leq \beta_t(\delta) \qquad (\bar{\theta}_t \in \mathcal{E}_t^\delta(\hat{\theta}_t)) \\ &\Leftrightarrow \theta_\star^t \in \mathcal{E}_t^\delta(\theta_c) \ .\end{aligned}$$

To sum-up, we have $\|g_t(\theta_c) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} < \|g_t(\theta_t^p) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}}$ and $\mathcal{E}_t^\delta(\theta_c) \cap \Theta \neq \emptyset$ since $\theta_\star^t \in \Theta \cap \mathcal{E}_t^\delta(\theta_c)$. This contradicts the definition of θ_t^p (in (**P1**)) and therefore Equation (16) must be wrong, which proves the announced result. ■

Appendix B. Regret bound

B.1. Regret decomposition

Lemma 3 *The following holds:*

$$R_T \leq \frac{2k_\mu}{c_\mu} \sum_{t=1}^T \beta_t(\delta) \left[\|x_t\|_{\mathbf{V_t^{-1}}} - \|x_\star^t\|_{\mathbf{V_t^{-1}}}\right] + \sum_{t=1}^T \left[\Delta_t(x_t) + \Delta_t(x_\star^t)\right] .$$

Proof We recall that $x_\star^t = \arg\max_{x \in \mathcal{X}_t} \mu(\langle \theta_\star^t, x\rangle)$. Note that:

$$\begin{aligned}R_T &= \sum_{t=1}^T \mu(\langle x_\star^t, \theta_\star^t\rangle) - \mu(\langle x_t, \theta_\star^t\rangle) \\ &= \sum_{t=1}^T \mu(\langle x_\star^t, \theta_\star^t\rangle) - \mu(\langle x_\star^t, \tilde{\theta}_t\rangle) + \mu(\langle x_\star^t, \tilde{\theta}_t\rangle) - \mu(\langle x_t, \tilde{\theta}_t\rangle) + \mu(\langle x_t, \tilde{\theta}_t\rangle) - \mu(\langle x_t, \theta_\star^t\rangle) \\ &= \sum_{t=1}^T \left[\mu(\langle x_\star^t, \tilde{\theta}_t\rangle) - \mu(\langle x_t, \tilde{\theta}_t\rangle)\right] + \sum_{t=1}^T \left[\mu(\langle x_\star^t, \theta_\star^t\rangle) - \mu(\langle x_\star^t, \tilde{\theta}_t\rangle)\right] + \sum_{t=1}^T \left[\mu(\langle x_t, \tilde{\theta}_t\rangle) - \mu(\langle x_t, \theta_\star^t\rangle)\right] \\ &\leq \frac{2k_\mu}{c_\mu} \sum_{t=1}^T \beta_t(\delta) \left[\|x_t\|_{\mathbf{V_t^{-1}}} - \|x_\star^t\|_{\mathbf{V_t^{-1}}}\right] \\ &\quad + \sum_{t=1}^T \left[\mu(\langle x_\star^t, \theta_\star^t\rangle) - \mu(\langle x_\star^t, \tilde{\theta}_t\rangle)\right] + \sum_{t=1}^T \left[\mu(\langle x_t, \tilde{\theta}_t\rangle) - \mu(\langle x_t, \theta_\star^t\rangle)\right] \ .\end{aligned}$$

In the last inequality, we used the fact that $x_t = \arg\max_{x \in \mathcal{X}} \left\{\mu(\langle x, \tilde{\theta}_t\rangle) + \frac{2k_\mu}{c_\mu}\beta_t(\delta)\|x\|_{\mathbf{V_t^{-1}}}\right\}$. Using the definition of $\Delta_t(x)$ we conclude that:

$$R_T \leq \frac{2k_\mu}{c_\mu} \sum_{t=1}^T \beta_t(\delta) \left[\|x_t\|_{\mathbf{V_t^{-1}}} - \|x_\star^t\|_{\mathbf{V_t^{-1}}}\right] + \sum_{t=1}^T \left[\Delta_t(x_t) + \Delta_t(x_\star^t)\right] \ .$$

■

B.2. Regret bound

We now claim Theorem 1, bounding the regret of BVD-GLM-UCB.

Theorem 1 *Let $\delta \in (0,1]$ and $D \in \mathbb{N}^*$. Under Assumptions 1 -2-3, with probability at least $1-\delta$:*

$$R_T \leq C_1 R_\mu \beta_T(\delta)\sqrt{dT}\sqrt{T\log(1/\gamma) + \log\left(1 + \frac{L^2(1-\gamma^T)}{\lambda d(1-\gamma)}\right)} + C_2 R_\mu \frac{\gamma^D}{1-\gamma}T + C_3 R_\mu D B_T$$

Further, setting $\gamma = 1 - (B_T/(dT))^{2/3}$ ensures:

$$R_T = \tilde{\mathcal{O}}\left(\frac{k_\mu}{c_\mu} d^{2/3} B_T^{1/3} T^{2/3}\right) \qquad \textit{w.h.p}$$

Proof In the following, we assume that the event $\{\bar{\theta}_t \in \mathcal{E}_t^\delta(\hat{\theta}_t), \forall t \geq 1\}$ holds, which happens with probability at least $1-\delta$ (Lemma 1). Thanks to Lemma 5, the following holds:

$$\Delta_t(x_t) + \frac{2k_\mu}{c_\mu}\beta_t(\delta)\|x_t\|_{\mathbf{V_t^{-1}}} \leq \frac{4k_\mu}{c_\mu}\beta_t(\delta)\|x_t\|_{\mathbf{V_t^{-1}}} + \frac{4k_\mu^2 L^3 S}{c_\mu\lambda(1-\gamma)}\gamma^D + \frac{2k_\mu^2 L}{c_\mu}\sum_{s=t-D}^{t-1}\left\|\theta_\star^s - \theta_\star^{s+1}\right\|_2$$

$$\Delta_t(x_\star^t) - \frac{2k_\mu}{c_\mu}\beta_t(\delta)\|x_\star^t\|_{\mathbf{V_t^{-1}}} \leq \frac{4k_\mu^2 L^3 S}{c_\mu\lambda(1-\gamma)}\gamma^D + \frac{2k_\mu^2 L}{c_\mu}\sum_{s=t-D}^{t-1}\left\|\theta_\star^s - \theta_\star^{s+1}\right\|_2$$

Assembling this result with Lemma 3 yields:

$$R_T \leq \underbrace{\sum_{t=1}^{T}\frac{4k_\mu}{c_\mu}\beta_t(\delta)\|x_t\|_{\mathbf{V_t^{-1}}}}_{R_T^{\text{learn}}} + \underbrace{\sum_{t=1}^{T}\left[\frac{8k_\mu^2 L^3 S}{c_\mu\lambda(1-\gamma)}\gamma^D + \frac{4k_\mu^2 L}{c_\mu}\sum_{s=t-D}^{t-1}\left\|\theta_\star^s - \theta_\star^{s+1}\right\|_2\right]}_{R_T^{\text{track}}} .$$

We now bound each term separately. Starting with R_T^{learn}:

$$\begin{aligned}
R_T^{\text{learn}} &\leq \frac{4k_\mu}{c_\mu}\beta_T(\delta)\sum_{t=1}^{T}\|x_t\|_{\mathbf{V_t^{-1}}} && (t \to \beta_t(\delta) \text{ increasing})\\
&\leq \frac{4k_\mu}{c_\mu}\beta_T(\delta)\sqrt{T}\sqrt{\sum_{t=1}^{T}\|x_t\|^2_{\mathbf{V_t^{-1}}}} && \text{(Cauchy-Schwarz)}\\
&\leq \frac{4k_\mu}{c_\mu}\beta_T(\delta)\sqrt{2T\max(1,L^2/\lambda)}\sqrt{dT\log(1/\gamma) + \log\left(\frac{\det \mathbf{V_{T+1}}}{\lambda^d}\right)} && \text{(Lemma 6)}\\
&\leq \frac{4k_\mu}{c_\mu}\beta_T(\delta)\sqrt{2dT\max(1,L^2/\lambda)}\sqrt{T\log(1/\gamma) + \log\left(1 + \frac{L^2(1-\gamma^T)}{\lambda d(1-\gamma)}\right)} . && \text{(Lemma 7)}
\end{aligned}$$

The bounding of the tracking term is straight-forward:

$$\begin{aligned}
R_T^{\text{track}} &= \frac{8k_\mu^2 L^3 S}{c_\mu\lambda(1-\gamma)}\gamma^D T + \frac{4k_\mu^2 L}{c_\mu}\sum_{t=1}^{T}\sum_{s=t-D}^{t-1}\left\|\theta_\star^s - \theta_\star^{s+1}\right\|_2\\
&\leq \frac{8k_\mu^2 L^3 S}{c_\mu\lambda(1-\gamma)}\gamma^D T + \frac{4k_\mu^2 L}{c_\mu} D B_T .
\end{aligned}$$

Assembling this two bounds (R_T^{learn} and R_T^{track}) yields the first announced result, with the following constants:

$$C_1 = \sqrt{32\max(1, L^2/\lambda)}\ .$$
$$C_2 = \frac{8k_\mu L^3 S}{\lambda}\ .$$
$$C_3 = 4k_\mu L\ .$$

The last part of the proof follows the asymptotic argument of Russac et al. (2019). We assume that B_T is sub-linear and let:

$$D = \frac{\log T}{1-\gamma}\ , \qquad \gamma = 1 - \left(\frac{B_T}{dT}\right)^{2/3}\ .$$

We therefore have the following asymptotic equivalences (omitting logarithmic dependencies):

$$\beta_T(\delta)\sqrt{dT}\sqrt{T\log(1/\gamma)} \sim dT \cdot \left(\frac{B_T}{dT}\right)^{1/3} = d^{2/3}B_T^{1/3}T^{2/3}$$
$$\gamma^D T/(1-\gamma) \sim \exp(-\log T)T\left(\frac{B_T}{dT}\right)^{-2/3} = d^{2/3}B_T^{-2/3}T^{2/3}$$
$$DB_T \sim B_T\left(\frac{B_T}{dT}\right)^{-2/3} = d^{2/3}B_T^{1/3}T^{2/3}$$

Merged with the regret-bound we just proved, this yields the second announced result. ■

Appendix C. On the projection step

C.1. Equivalent minimization program

Recall the original minimization program for finding θ_t^p:

$$\theta_t^p \in \underset{\theta\in\mathbb{R}^d}{\arg\min}\left\{\left\|g_t(\theta) - g_t(\hat{\theta}_t)\right\|_{\mathbf{V_t^{-2}}} \text{ s.t } \Theta \cap \mathcal{E}_t^\delta(\theta) \neq \emptyset\right\}\ . \tag{P1}$$

Note that this minimum exists (0_d is feasible) and is indeed attained (the feasible set is compact and the objective smooth). The following reformulation is motivated by the fact that only $\tilde{\theta}_t \in \Theta \cap \mathcal{E}_t^\delta(\theta_t^p)$ is needed for the algorithm. To this end, we explicitly introduce $\tilde{\theta}_t$ in the program via a slack variable. Formally, we study:

$$\begin{pmatrix}\tilde{\theta}_t \\ \theta_t^p\end{pmatrix} \in \underset{\theta'\in\mathbb{R}^d,\theta\in\mathbb{R}^d}{\arg\min}\left\{\left\|g_t(\theta) - g_t(\hat{\theta}_t)\right\|_{\mathbf{V_t^{-2}}} \text{ s.t } \theta' \in \mathcal{E}_t^\delta(\theta) \cap \Theta\right\}\ . \tag{P1'}$$

We also introduce the following program:

$$\begin{pmatrix}\tilde{\theta}_t \\ \eta\end{pmatrix} \in \underset{\theta'\in\mathbb{R}^d,\eta\in\mathbb{R}^d}{\arg\min}\left\{\left\|g_t(\theta') + \beta_t(\delta)\mathbf{\tilde{V}_t^{1/2}}\eta - g_t(\hat{\theta}_t)\right\|_{\mathbf{V_t^{-2}}} \text{ s.t } \left\|\theta'\right\|_2 \leq S, \|\eta\|_2 \leq 1\right\}\ . \tag{P2}$$

We claim and prove the following result, which is an equivalent reformulation of Proposition 1.

Proposition 2 *The programs* (**P1'**) *and* (**P2**) *are equivalent.*

Proof The proof consists in building a bijection between the solutions of (**P1'**) and (**P2**). Let us introduce the mapping:

$$\begin{aligned} f \,:\, \Theta \times \mathbb{R}^d &\to \Theta \times \mathbb{R}^d \\ \begin{pmatrix} x \\ y \end{pmatrix} &\to \begin{pmatrix} f_1(x) \\ f_2(x,y) \end{pmatrix} = \begin{pmatrix} x \\ \beta_t^{-1}(\delta)\widetilde{\mathbf{V}}_{\mathbf{t}}^{-\mathbf{1/2}}(g_t(y) - g_t(x)) \end{pmatrix} \end{aligned}$$

We now claim the following Lemma, which proof is deferred to Section C.2.

Lemma 1 *The function:*

$$\begin{aligned} g_t \,:& \mathbb{R}^d \to \mathbb{R}^d \\ & \theta \to \sum_{s=1}^{t-1} \gamma^{t-1-s} \mu(\langle \theta, x_s \rangle) x_s + \lambda c_\mu \theta \end{aligned}$$

is a bijection.

A straight-forward implication of this Lemma is the bijectivity of f. Let $(\tilde{\theta}^1, \theta^p)$ be a solution of (**P1'**) and let:

$$\begin{pmatrix} \tilde{\theta}^2 \\ \eta^p \end{pmatrix} = f \begin{pmatrix} \tilde{\theta}^1 \\ \theta^p \end{pmatrix} .$$

We are going to show that $(\tilde{\theta}^2, \eta^p)$ is a solution of (**P2**). Because $(\tilde{\theta}^1, \theta^p)$ is optimal for (**P1'**), we have that:

$$\begin{aligned} &\|g_t(\theta^p) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \leq \|g_t(\theta) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \\ &\qquad\qquad \forall (\theta', \theta) \in \Theta \times \mathbb{R}^d \text{ s.t } \theta' \in \mathcal{E}_t^\delta(\theta) \\ \Leftrightarrow\ &\|g_t(\theta^p) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \leq \|g_t(\theta) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \qquad \text{(definition of } \mathcal{E}_t^\delta(\theta)) \\ &\qquad\qquad \forall (\theta', \theta) \in \Theta \times \mathbb{R}^d \text{ s.t } \|g_t(\theta') - g_t(\theta)\|_{\widetilde{\mathbf{V}}_{\mathbf{t}}^{-\mathbf{1}}} \leq \beta_t(\delta) \\ \Leftrightarrow\ &\|g_t(\theta^p) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \leq \|g_t(\theta) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \\ &\qquad\qquad \forall (\theta', \theta) \in \Theta \times \mathbb{R}^d \text{ s.t } \left\| f_2(\theta', \theta) \right\|_2 \leq 1 \end{aligned}$$

Noticing that for all $(x, y) \in \Theta \times \mathbb{R}^d$ we have $g_t(y) = g_t(x) + \beta_t(\delta)\mathbf{V_t^{1/2}} f_2(x, y)$ we therefore obtain:

$$\begin{aligned}
&\|g_t(\tilde{\theta}^1) + \beta_t(\delta)\widetilde{\mathbf{V}}_\mathbf{t}^{\mathbf{1/2}} f_2(\tilde{\theta}^1, \theta^p) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \leq \|g_t(\theta') + \beta_t(\delta)\widetilde{\mathbf{V}}_\mathbf{t}^{\mathbf{1/2}} f_2(\theta', \theta) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \\
&\qquad\qquad \forall (\theta', \theta) \in \Theta \times \mathbb{R}^d \text{ s.t } \left\|f_2(\theta', \theta)\right\|_2 \leq 1 \\
\Leftrightarrow\ &\|g_t(\tilde{\theta}^1) + \beta_t(\delta)\widetilde{\mathbf{V}}_\mathbf{t}^{\mathbf{1/2}} \eta^p - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \leq \|g_t(\theta') + \beta_t(\delta)\widetilde{\mathbf{V}}_\mathbf{t}^{\mathbf{1/2}} f_2(\theta', \theta) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \\
&\qquad\qquad \forall (\theta', \theta) \in \Theta \times \mathbb{R}^d \text{ s.t } \left\|f_2(\theta', \theta)\right\|_2 \leq 1 \\
\Leftrightarrow\ &\|g_t(\tilde{\theta}^2) + \beta_t(\delta)\widetilde{\mathbf{V}}_\mathbf{t}^{\mathbf{1/2}} \eta^p - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \leq \|g_t(\theta') + \beta_t(\delta)\widetilde{\mathbf{V}}_\mathbf{t}^{\mathbf{1/2}} f_2(\theta', \theta) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \\
&\qquad\qquad \forall (\theta', \theta) \in \Theta \times \mathbb{R}^d \text{ s.t } \left\|f_2(\theta', \theta)\right\|_2 \leq 1 \qquad (\tilde{\theta}^1 = \tilde{\theta}^2) \\
\Leftrightarrow\ &\|g_t(\tilde{\theta}^2) + \beta_t(\delta)\widetilde{\mathbf{V}}_\mathbf{t}^{\mathbf{1/2}} \eta^p - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \leq \|g_t(\theta') + \beta_t(\delta)\widetilde{\mathbf{V}}_\mathbf{t}^{\mathbf{1/2}} f_2(\theta', \theta) - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \\
&\qquad\qquad \forall (\theta', \theta) \text{ s.t } \left\|f_2(\theta', \theta)\right\|_2 \leq 1, \left\|\theta'\right\|_2 \leq S \\
\Leftrightarrow\ &\|g_t(\tilde{\theta}^2) + \beta_t(\delta)\widetilde{\mathbf{V}}_\mathbf{t}^{\mathbf{1/2}} \eta^p - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \leq \|g_t(\theta') + \beta_t(\delta)\widetilde{\mathbf{V}}_\mathbf{t}^{\mathbf{1/2}} \eta - g_t(\hat{\theta}_t)\|_{\mathbf{V_t^{-2}}} \\
&\qquad\qquad \forall (\theta', \eta) \text{ s.t } \|\eta\|_2 \leq 1, \left\|\theta'\right\|_2 \leq S
\end{aligned}$$

where we last used the fact that f_2 spans $\mathbb{R}^d$ (surjectivity). Finally, we have that:

$$\begin{aligned}
&\left\|\tilde{\theta}^2\right\|_2 \leq S && (\tilde{\theta}^2 = \tilde{\theta}^1 \in \Theta) \\
&\|\eta^p\|_2 = \beta_t^{-1}(\delta) \left\|g_t(\theta^p) - g_t(\tilde{\theta}^1)\right\|_{\mathbf{V_t^{-1}}} \leq 1 && (\tilde{\theta}^1 \in \mathcal{E}_t^\delta(\theta^p))
\end{aligned}$$

Combining the last two results proves that $(\tilde{\theta}^2, \eta^p)$ is feasible for (**P2**), and optimal within the feasible set. As a consequence, $(\tilde{\theta}^2, \eta^p)$ is a solution of (**P2**). Therefore, f is a bijection between the minimizers of (**P1'**) and (**P2**), which concludes the proof. ■

C.2. Bijectivity of g_t

Lemma 2 *The function:*

$$\begin{aligned}
g_t\ &: \mathbb{R}^d \to \mathbb{R}^d \\
&\theta \to \sum_{s=1}^{t-1} \gamma^{t-1-s} \mu(\langle \theta, x_s \rangle) x_s + \lambda c_\mu \theta
\end{aligned}$$

is a bijection.

Proof Injectivity. Notice that $\forall \theta \in \mathbb{R}^d$:

$$\nabla_\theta g(\theta) = \sum_{s=1}^{t-1} \gamma^{t-1-s} \dot{\mu}(\langle \theta, x_s \rangle) x_s x_s^\mathsf{T} + \lambda c_\mu \mathbf{I_d} \succ 0 \,.$$

Hence $\nabla_\theta g$ is P.S.D, and a simple integral Taylor expansion is enough to prove injectivity.

Surjectivity Let $z \in \mathbb{R}^d$. Let $A = \text{Span}(x_1, .., x_{t-1})$ be the vectorial space spanned by $\{x_s\}_{s=1}^{t-1}$. Let $z_\perp$ be the orthogonal projection of z on A and $z_\parallel = z - z_\perp$. Since $z_\perp \in A$, there exists $\{\alpha_s\}_{s=1}^{t-1} \in \mathbb{R}^{t-1}$ such that:

$$z_\perp = \sum_{s=1}^{t-1} \alpha_s x_s \ .$$

Recall that $b(\cdot)$ is a primitive of μ, which is convex since μ is strictly increasing. Define:

$$L(\theta) = \sum_{s=1}^{t-1} \gamma^{t-1-s} \left[b(\langle \theta, x_s \rangle) - \frac{\alpha_s}{\gamma^{t-1-s}} \langle \theta, x_s \rangle \right] + \frac{\lambda c_\mu}{2} \left\| \theta - \frac{z_\parallel}{\lambda c_\mu} \right\|^2 \ .$$

which is a strictly convex, coercive function. Its minimum θ_z (which therefore exists and is uniquely defined) checks:

$$\begin{aligned}
&\nabla_\theta L(\theta_z) = 0 \\
&\Leftrightarrow \sum_{s=1}^{t-1} \gamma^{t-1-s} \left[\mu(\langle \theta_z, x_s \rangle) - \frac{\alpha_s}{\gamma^{t-1-s}} \right] x_s + \lambda c_\mu \left(\theta_z - \frac{z_\parallel}{\lambda c_\mu} \right) = 0 \\
&\Leftrightarrow g(\theta_z) = \sum_{s=1}^{t-1} \alpha_s x_s + z_\parallel \\
&\Leftrightarrow g(\theta_z) = z_\perp + z_\parallel = z \ .
\end{aligned}$$

which proves surjectivity. ■

Appendix D. Useful lemmas

The following Lemma is a version of the Elliptical Potential Lemma for weighted sums, similar to Proposition 4 of Russac et al. (2019).

Lemma 6 *Let $\{x_s\}_{s=1}^{\infty}$ a sequence in $\mathbb{R}^d$ such that $\|x_s\|_2 \leq L$ for all $s \in \mathbb{N}^*$, and let λ be a non-negative scalar. For $t \geq 1$ define $\mathbf{V_t} := \sum_{s=1}^{t-1} \gamma^{t-1-s} x_s x_s^T + \lambda \mathbf{I_d}$. The following inequality holds:*

$$\sum_{t=1}^{T} \|x_t\|^2_{\mathbf{V_t^{-1}}} \leq 2 \max(1, L^2/\lambda) \left(dT \log(1/\gamma) + \log \left(\frac{\det \mathbf{V_{T+1}}}{\lambda^d} \right) \right) \ .$$

Proof For all $t \geq 1$, by definition:

$$\begin{aligned}
\mathbf{V_{t+1}} &= \sum_{s=1}^{t} \gamma^{t-s} x_s x_s^\mathsf{T} + \lambda \mathbf{I_d} \\
&= \gamma \sum_{s=1}^{t-1} \gamma^{t-1-s} x_s x_s^T + x_t x_t^\mathsf{T} + \lambda \mathbf{I_d} \\
&\succeq \gamma \left(\sum_{s=1}^{t-1} \gamma^{t-1-s} x_s x_s^T + x_t x_t^\mathsf{T} + \lambda \mathbf{I_d} \right) \qquad (\gamma \leq 1) \\
&\succeq \gamma \left(\mathbf{V_t} + x_t x_t^\mathsf{T} \right) \\
&\succeq \gamma \mathbf{V_t} \left(\mathbf{I_d} + \mathbf{V_t^{-1/2}} x_t x_t^\mathsf{T} \mathbf{V_t^{-1/2}} \right) ,
\end{aligned}$$

which after some easy manipulations yields:

$$d \log(1/\gamma) + \log \det \mathbf{V_{t+1}} - \log \det \mathbf{V_t} \geq \log \left(1 + \|x_t\|^2_{\mathbf{V_t^{-1}}} \right) .$$

After summing from $t = 1$ to $t = T$ and telescoping we obtain:

$$\begin{aligned}
dT \log(1/\gamma) + \log \left(\frac{\det \mathbf{V_{T+1}}}{\lambda^d} \right) &\geq \sum_{t=1}^{T} \log \left(1 + \|x_t\|^2_{\mathbf{V_t^{-1}}} \right) \\
&\geq \sum_{t=1}^{T} \log \left(1 + \frac{1}{\max(1, L^2/\lambda)} \|x_t\|^2_{\mathbf{V_t^{-1}}} \right) .
\end{aligned}$$

Finally, noticing that $\frac{1}{\max(1,L^2/\lambda)} \|x_t\|^2_{\mathbf{V_t^{-1}}} \leq 1$ and using the fact that for all $x \in (0, 1]$ we have $\log(1 + x) \geq x/2$ we obtain:

$$dT \log(1/\gamma) + \log \left(\frac{\det \mathbf{V_{T+1}}}{\lambda^d} \right) \geq \frac{1}{2 \max(1, L^2/\lambda)} \sum_{t=1}^{T} \|x_t\|^2_{\mathbf{V_t^{-1}}} ,$$

which in turn yields:

$$\sum_{t=1}^{T} \|x_t\|^2_{\mathbf{V_t^{-1}}} \leq 2 \max(1, L^2/\lambda) \left(dT \log(1/\gamma) + \log \left(\frac{\det \mathbf{V_{T+1}}}{\lambda^d} \right) \right) ,$$

which is the announced result. ■

We also remind here the determinant-trace inequality for the weighted design matrix which can be extracted from Proposition 2 of Russac et al. (2019).

Lemma 7 *Let $\{x_s\}_{s=1}^{\infty}$ a sequence in $\mathbb{R}^d$ such that $\|x_s\|_2 \leq L$ for all $s \in \mathbb{N}^*$, and let λ be a non-negative scalar. For $t \geq 1$ define $\mathbf{V_t} := \sum_{s=1}^{t-1} \gamma^{t-1-s} x_s x_s^T + \lambda \mathbf{I_d}$. The following inequality holds:*

$$\det(\mathbf{V_{t+1}}) \leq \left(\lambda + \frac{L^2(1 - \gamma^t)}{d(1 - \gamma)} \right)^d .$$

Appendix E. BVD-GLM-UCB algorithm

E.1. High-level ideas

In this part of the appendix, we denote $\gamma^\star$ as follows:

$$\gamma^\star = 1 - \frac{1}{2}\left(\frac{B_{T,\star}}{dT2S}\right)^{2/3} . \tag{19}$$

Remark 3 *$\gamma^\star$ as defined in Equation (19) has a different expression than the discount factor proposed in Theorem 1. This slight modification is to ensure that $\gamma^\star$ is larger than 1/2 and simplifies the finite time analysis of the regret. Yet, it has no consequence on the asymptotic bound.*

$B_{T,\star}$ being unknown, we cannot compute the optimal discount factor that depends on the parameter drift. The general idea is to use a set of different values for the discount factor (respectively the $B_{T,\star}$ values) called $\mathcal{H}$, covering the $[1/2, 1)$ space (respectively the $[0, 2ST)$ space). Then, we divide the time horizon T into different blocks of length H. Every H steps, we create **a new instance** of BVD-GLM-UCB with a γ that is chosen by a *master* algorithm: the EXP3 algorithm from Auer et al. (2002). At the end of each block, this *master* algorithm receives the cumulative rewards from the instantiated *worker* and updates its probability distribution over the set $\mathcal{H}$. The objective of the master algorithm is to learn the most suitable value of γ so as to maximise the cumulative rewards in accordance with the dynamics of the environment. On the other side, the different *workers* algorithms act exactly as if the BVD-GLM-UCB algorithm was launched on a H-steps experiment. This setting is similar to the one presented in Cheung et al. (2019a) (respectively Zhao et al. (2020)) with discount factors instead of sliding windows (respectively restart parameters). This framework is called Bandit-over-Bandit (BOB) precisely because of this two-stage structure between the *master* and the *workers* algorithms.

E.2. Algorithm

The coverage $\mathcal{H}$ with the different discount factors is defined in the following way:

$$\mathcal{H} = \{\gamma_i = 1 - \mu_i | i = 1, \dots, N\} \tag{20}$$

$$\text{with } N = \left\lceil \frac{2}{3}\log_2\left(2ST^{3/2}\right)\right\rceil + 1 \text{ and } \mu_i = \frac{1}{2}\frac{2^{i-1}}{d^{2/3}T(2S)^{2/3}} . \tag{21}$$

The *main* algorithm is an instance of the EXP3 algorithm from Auer et al. (2002) where the different arms correspond to the different discount factors. Following EXP3 analysis (Auer et al., 2002), the probability of drawing γ_j for the block i is

$$p_i^{\gamma_j} = (1-\alpha)\frac{s_i^{\gamma_j}}{\sum_j s_i^{\gamma_j}} + \frac{\alpha}{N},\ \forall j = 1, 2, \dots, N , \tag{22}$$

where α is defined as

$$\alpha = \min\left\{1, \sqrt{\frac{N\log(N)}{(e-1)\lceil T/H\rceil}}\right\} \tag{23}$$

and $s_i^{\gamma_j}$ is initialised at 1 and is updated at the end of each block **when selected** with

$$s_{i+1}^{\gamma_j} = s_i^{\gamma_j} \exp\left(\frac{\alpha}{N p_i^{\gamma_j}} \frac{\sum_{t=(i-1)H+1}^{\min\{iH,T\}} r_{t+1}}{2\sigma H}\right). \tag{24}$$

Note that in Equation (24), r_{t+1} is the noisy reward obtained when the action x_t is selected with the BVD-GLM-UCB algorithm with parameter γ_j. Equation (22), (23) and (24) are the same as in Auer et al. (2002) except for the rescaling of the cumulative rewards on a block that is required to ensure that they lie in $[0, 1]$. Details on this rescaling part can be found in Proposition 4.

Algorithm 3 BOB-BVD-GLM-UCB(detailed)

Input. Length H, time horizon T, regularization λ, confidence δ, inverse link function μ, constants S, L and σ.
Initialization. Create the covering space $\mathcal{H}$ as defined in Eq. (20), set $s_1^{\gamma_i} = 1, \forall \gamma_i \in \mathcal{H}$.
for $i = 1, \dots, \lceil T/H \rceil$ **do**
  $\gamma_j \sim p_i^{\gamma}$, the probability vector defined in Eq. (22).
  Start a BVD-GLM-UCB subroutine with parameter γ_j
  for $t = (i-1)H + 1, \dots, \min\{iH, T\}$ **do**
    Receive the action set $\mathcal{X}_t$.
    Select $x_t(\gamma_j) \in \mathcal{X}_t$ with BVD-GLM-UCB.
    Observe reward r_{t+1}.
  end for
  Update $s_{i+1}^{\gamma_j}$ according to Equation (24).
  Update $s_{i+1}^{\gamma} = s_i^{\gamma}, \forall \gamma \neq \gamma_j$.
end for

Remark 4 *We denote $x_t(\gamma)$ the action chosen with the BVD-GLM-UCB algorithm with a discount factor γ.*

E.3. Regret guarantees

In this section, we give an upper-bound for the expected dynamic regret of BOB-BVD-GLM-UCB. By construction, it is natural to decompose the regret into two sources of errors. First the *master* error committed by the EXP3 algorithm by not choosing the best possible discount factor. Second the *worker* error inherent to the BVD-GLM-UCB algorithm. Note that there are two independent sources of randomness: the stochasticity of the rewards (whose expectation is denoted $\mathbb{E}_N$) and the randomness of the EXP3 algorithm (denoted $\mathbb{E}_{\texttt{EXP3}}$). Bringing things together,

$$\begin{aligned}\mathbb{E}[R_T] &= \mathbb{E}_N\left[\sum_{t=1}^{T}\mu(\langle x_\star^t,\theta_\star^t\rangle) - \mathbb{E}_{\texttt{EXP3}}[r_{t+1}]\right] \\ &= \underbrace{\mathbb{E}_N\left[\sum_{t=1}^{T}\mu(\langle x_\star^t,\theta_\star^t\rangle) - \sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{\min\{iH,T\}}\mu(\langle x_t(\widehat{\gamma}),\theta_\star^t\rangle)\right]}_{worker} \\ &+ \underbrace{\mathbb{E}_N\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{\min\{iH,T\}}\mu(\langle x_t(\widehat{\gamma}),\theta_\star^t\rangle) - \mathbb{E}_{\texttt{EXP3}}\left[r_{t+1}\right]\right]}_{master} .\end{aligned} \tag{25}$$

The next step consists in upper-bounding the $worker$ error and the $master$ error from Eq. (25) respectively.

Lemma 8 *With pavement $\mathcal{H}$ defined in Equation (20) for any unknown $B_{T,\star} > 0$, setting $k = \lfloor \frac{2}{3}\log_2(B_{T,\star}T^{1/2})\rfloor + 1$ yields*

$$\gamma_{k+1} \leq \gamma^\star \leq \gamma_k \,.$$

Proof With assumption 1, we have $B_{T,\star} \leq 2ST$. Using this, k (as defined in the statement of the lemma) is smaller than N. We have,

$$\begin{aligned} & k-1 \leq \frac{2}{3}\log_2(B_{T,\star}T^{1/2}) \leq k \\ \Leftrightarrow & -\frac{1}{2}\frac{2^{k-1}}{d^{2/3}T(2S)^{2/3}} \geq -\frac{1}{2}\left(\frac{B_{T,\star}}{dT2S}\right)^{2/3} \geq -\frac{1}{2}\frac{2^k}{d^{2/3}T(2S)^{2/3}} \,. \end{aligned}$$

Adding one for the different terms gives the result. ■

For the rest of the section, we set $\widehat{\gamma} = \gamma_k$ with k defined in Lemma 8. We denote $B_{i,\star} = \sum_{t=(i-1)H+1}^{iH-1}\|\theta_\star^{t+1} - \theta_\star^t\|_2$ and

$$\beta_H^\star = \sqrt{\lambda}S + \sigma\sqrt{2\log(T) + d\log\left(1 + \frac{2L^2}{\lambda d(1-\gamma^{\star 2})}\right)} \,. \tag{26}$$

Proposition 3 *The $worker$ error can be upper-bounded in the following way:*

$$\begin{aligned} worker \leq\ & 2\sigma\frac{T}{H} + C_1R_\mu\beta_H^\star\sqrt{dT}\sqrt{2T(1-\gamma^\star) + \frac{T}{H}\log\left(1 + \frac{2L^2}{d\lambda(1-\gamma^\star)}\right)} \\ & + 2C_2R_\mu\frac{1}{\sqrt{T}}\frac{1}{1-\gamma^\star} + \frac{3C_3R_\mu}{\log(2)}\frac{B_{T,\star}\log(T)}{1-\gamma^\star} \,, \end{aligned}$$

with C_1, C_2, C_3 constant terms from Theorem 1 and $\beta_H^\star$ defined in Equation (26).

Proof First, note that our objective here is to bound the expected regret whereas Theorem 1 bounds the pseudo-regret and gives a high probability upper-bound. We denote $E_\delta^i = \{\bar{\theta}_t \in \mathcal{E}_t^\delta(\hat{\theta}_t)$ for t s.t $(i-$

$1)H+1 \leq t \leq \min\{iH, T\}\}$. This event holds with probability higher than $1-\delta$. When E_δ^i does not hold, the maximum regret could theoretically be suffered for all time instants.

As explained in the algorithm mechanism, a new instance of BVD-GLM-UCB will be launched every H steps with a discount factor selected by the EXP3 algorithm. Restarting a new algorithm and forgetting previous information comes at a cost in terms of regret. This is made explicit in the following decomposition of $worker$.

$$
\begin{aligned}
worker &= \mathbb{E}_N \left[\sum_{i=1}^{\lceil T/H \rceil} \sum_{t=(i-1)H+1}^{\min\{iH,T\}} \mu(\langle x_\star^t, \theta_\star^t \rangle) - \mu(\langle x_t(\widehat{\gamma}), \theta_\star^t \rangle) \right] \\
&= \underbrace{\mathbb{E}_N \left[\sum_{i=1}^{\lceil T/H \rceil} \sum_{t=(i-1)H+1}^{\min\{iH,T\}} \langle \mu(\langle x_\star^t, \theta_\star^t \rangle) - \mu(\langle x_t(\widehat{\gamma}), \theta_\star^t \rangle) \Big| \{\cap_{i=1}^{\lceil T/H \rceil} E_\delta^i\} \right] \mathbb{P}\left(\cap_{i=1}^{\lceil T/H \rceil} E_\delta^i \right)}_{worker_1} \\
&+ \underbrace{\mathbb{E}_N \left[\sum_{i=1}^{\lceil T/H \rceil} \sum_{t=(i-1)H+1}^{\min\{iH,T\}} \mu(\langle x_\star^t, \theta_\star^t \rangle) - \mu(\langle x_t(\widehat{\gamma}), \theta_\star^t \rangle) \Big| \{\cap_{i=1}^{\lceil T/H \rceil} E_\delta^i\}^c \right] \mathbb{P}\left(\{\cap_{i=1}^{\lceil T/H \rceil} E_\delta^i\}^c \right)}_{worker_2}
\end{aligned}
$$

Thanks to Lemma 1, E_δ^i holds with probability higher than 1-δ. By setting $\delta = 1/T$, we have

$$
\mathbb{P}\left(\cup_{i=1}^{\lceil T/H \rceil} (E_\delta^i)^c \right) \leq \lceil T/H \rceil 1/T \; . \tag{27}
$$

Under the event $\{\cup_{i=1}^{\lceil T/H \rceil} (E_\delta^i)^c\}$ not much can be said. The maximum regret $r_{\max} = 2\sigma$ can be suffered at every time step. Therefore, using the upper-bound from Eq. (27), we obtain

$$
\begin{aligned}
worker_2 &= \mathbb{E}_N \left[\sum_{i=1}^{\lceil T/H \rceil} \sum_{t=(i-1)H+1}^{\min\{iH,T\}} \mu(\langle x_\star^t, \theta_\star^t \rangle) - \mu(\langle x_t(\widehat{\gamma}), \theta_\star^t \rangle) \Big| \{\cup_{i=1}^{\lceil T/H \rceil} (E_\delta^i)^c\} \right] \mathbb{P}\left(\cup_{i=1}^{\lceil T/H \rceil} (E_\delta^i)^c \right) \\
&\leq r_{\max} \lceil T/H \rceil \; .
\end{aligned}
$$

This term is related to the number of restarts of the algorithm. In the BOB framework, whatever the worker algorithm (sliding window, restart factor) a cost of order T/H will be paid due to the restarting of the *worker* at the beginning of each block.

On the contrary, under the event $\{\cap_{i=1}^{\lceil T/H \rceil} E_\delta^i\}$, using the assumption that the blocks are independent, we can follow the line of proof from Lemma 3 and Theorem 1 for every block. We introduce,

$$
\beta_H = \sqrt{\lambda} S + \sigma \sqrt{2 \log(T) + d \log\left(1 + \frac{L^2 (1 - \gamma_k^{2H})}{\lambda d (1 - \gamma_k^2)} \right)} \; . \tag{28}
$$

$$
\begin{aligned}
worker_1 &= \mathbb{E}_N\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{\min\{iH,T\}}\mu(\langle x_\star^t,\theta_\star^t\rangle)-\mu(\langle x_t(\widehat{\gamma}),\theta_\star^t\rangle)\Big|\{\cap_{i=1}^{\lceil T/H\rceil}E_\delta^i\}\right]\mathbb{P}\left(\cap_{i=1}^{\lceil T/H\rceil}E_\delta^i\right)\\
&\leq \mathbb{E}_N\left[\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=(i-1)H+1}^{\min\{iH,T\}}\mu(\langle x_\star^t,\theta_\star^t\rangle)-\mu(\langle x_t(\widehat{\gamma}),\theta_\star^t\rangle)\Big|\{\cap_{i=1}^{\lceil T/H\rceil}E_\delta^i\}\right]\\
&\leq \sum_{i=1}^{\lceil T/H\rceil}\left(C_1\beta_H\sqrt{dH}\sqrt{H\log(1/\widehat{\gamma})+\log\left(1+\frac{L^2}{d\lambda(1-\widehat{\gamma})}\right)}+C_2\frac{\widehat{\gamma}^D}{1-\widehat{\gamma}}H+C_3B_{i,\star}D\right)\\
&\leq C_1\beta_H\sqrt{dT}\sqrt{T\log(1/\widehat{\gamma})+\frac{T}{H}\log\left(1+\frac{L^2}{d\lambda(1-\widehat{\gamma})}\right)}+C_2\frac{\widehat{\gamma}^D}{1-\widehat{\gamma}}T+C_3B_{T,\star}D\,,
\end{aligned}
$$

where the second inequality is a consequence of Theorem 1. We set,

$$
D=\frac{3/2\log(T)}{\log(1/\widehat{\gamma})}\,. \tag{29}
$$

Hence,

$$
\begin{aligned}
C_3B_{T,\star}D &\leq \frac{3}{2}\frac{C_3B_{T,\star}\log(T)}{\log(1/\widehat{\gamma})}\\
&\leq \frac{3C_3}{2\log(2)}B_{T,\star}\log(T)\frac{\widehat{\gamma}}{1-\widehat{\gamma}} \quad (\text{Using } \log(x)\geq\log(2)(x-1) \text{ for } x\in[1,2])\\
&\leq \frac{3C_3}{2\log(2)}\frac{B_{T,\star}\log(T)}{1-\gamma_k} \quad (\widehat{\gamma}\leq 1)\\
&\leq \frac{3C_3}{\log(2)}\frac{B_{T,\star}\log(T)}{1-\gamma_{k+1}} \quad (\text{Definition of } \mathcal{H})\\
&\leq \frac{3C_3}{\log(2)}\frac{B_{T,\star}\log(T)}{1-\gamma^\star} \quad (\text{Lemma8})\,.
\end{aligned}
$$

We also have,

$$
\begin{aligned}
C_2\frac{\widehat{\gamma}^D}{1-\widehat{\gamma}}T &\leq C_2\frac{1}{\sqrt{T}}\frac{1}{1-\widehat{\gamma}} \quad (\text{Equation (29)})\\
&\leq 2C_2\frac{1}{\sqrt{T}}\frac{2}{1-\gamma_{k+1}} \quad (\text{Definition of } \mathcal{H})\\
&\leq 2C_2\frac{1}{\sqrt{T}}\frac{1}{1-\gamma^\star} \quad (\text{Lemma 8})\,.
\end{aligned}
$$

Finally, using $x\mapsto\log(x)\leq x-1$ for $x>1$ and Lemma 8, one has:

$$
\begin{aligned}
T\log(1/\widehat{\gamma})+\frac{T}{H}\log\left(1+\frac{L^2}{d\lambda(1-\widehat{\gamma})}\right) &\leq T\frac{1-\widehat{\gamma}}{\widehat{\gamma}}+\frac{T}{H}\log\left(1+\frac{2L^2}{d\lambda(1-\gamma^\star)}\right)\\
&\leq 2T(1-\gamma^\star)+\frac{T}{H}\log\left(1+\frac{2L^2}{d\lambda(1-\gamma^\star)}\right)\,.
\end{aligned}
$$

Following similar steps, we can upper-bound β_H from Equation (28) by

$$\beta_H \leq \beta_H^\star \ .$$

Bringing things together, we have shown that under the event $\{\cap_{i=1}^{\lceil T/H \rceil} \mathcal{E}_i\}$ all the terms depending on $\widehat{\gamma}$ can be replaced by terms depending only on $\gamma^\star$ at the cost of multiplicative constant independent of T. Finally, one has

$$\begin{aligned} worker \leq 2\sigma \frac{T}{H} + C_1 R_\mu \beta_H^\star \sqrt{dT} \sqrt{2T(1-\gamma^\star) + \frac{T}{H} \log\left(1 + \frac{2L^2}{d\lambda(1-\gamma^\star)}\right)} \\ + \ 2C_2 R_\mu \frac{1}{\sqrt{T}} \frac{1}{1-\gamma^\star} + \frac{3C_3 R_\mu}{\log(2)} \frac{B_{T,\star} \log(T)}{1-\gamma^\star} \ . \end{aligned}$$

■

The above proposition bounds the regret incurred if the same discount factor $\widehat{\gamma}$ is used for each block. To successfully upper bound BVD-GLM-UCB's regret, we need to upper bound the second part $master$ which is the error due to the use of the EXP3 algorithm. This part can be controlled thanks to the analysis proposed in Auer et al. (2002). Yet, two issues need to be overcome. (1) The rewards received at the end of a block does not lie in $[0, 1]$ which is required to use the result from Auer et al. (2002). (2) We are in a stochastic environment with noisy rewards.

In the next proposition, we upper-bound the term of interest and explain how to deal with the two issues. The big picture is the following: using the assumption on the bounded rewards we can obtain an upper-bound for the maximum reward on a single block.

Proposition 4 *The regret due to the master algorithm can be bounded in the following way,*

$$\mathbb{E}_N \left[\sum_{i=1}^{\lceil T/H \rceil} \sum_{t=(i-1)H+1}^{\min\{iH,T\}} \mu(\langle x_t(\widehat{\gamma}), \theta_\star^t \rangle) - \mathbb{E}_{\texttt{EXP3}}\left[r_{t+1}\right] \right] \leq 4\sigma H \sqrt{e-1} \sqrt{\frac{T}{H} card(\mathcal{H}) \log(card(\mathcal{H}))}$$

Proof We denote γ_i the discount factor chosen by the EXP3 algorithm in the i-th block. The regret due to the use of the EXP3 *main* algorithm can be written as follows:

$$master = \mathbb{E}_N \left[\sum_{i=1}^{\lceil T/H \rceil} \sum_{t=(i-1)H+1}^{\min\{iH,T\}} \mu(\langle x_t(\widehat{\gamma}), \theta_\star^t \rangle) - \mathbb{E}_{\texttt{EXP3}} \left[\sum_{i=1}^{\lceil T/H \rceil} \sum_{t=(i-1)H+1}^{\min\{iH,T\}} r_{t+1} \right] \right] \ .$$

We introduce $Q_i(\gamma_j) = \sum_{t=(i-1)H+1}^{\min\{iH,T\}} r_{t+1}(\gamma_j) = \sum_{t=(i-1)H+1}^{\min\{iH,T\}} \mu(\langle x_t(\gamma_j), \theta_\star^t \rangle) + \epsilon_{t+1}$, using Equation (9). This quantity corresponds to the reward obtained on the i-th block when using BVD-GLM-UCB with the discount factor γ_j. We also use $Q_i = \max_{\gamma \in \mathcal{H}} Q_i(\gamma)$.

Contrarily to existing works in the linear setting (e.g (Cheung et al., 2019b, Lemma3)) our assumption on the bounded rewards is sufficient to solve both problems. We have, $|Q_i| \leq 2\sigma H$ almost surely using $r_t \leq 2\sigma$ for all time instants.

Let $\mathcal{U} = \{\forall t \leq T, 0 \leq r_t \leq 2\sigma\}$. Thanks to assumption 2, we have $\mathbb{P}(\mathcal{U}) = 1$.

One has,

$$\begin{aligned} master &\leq \mathbb{E}_N \left[\sum_{i=1}^{\lceil T/H \rceil} Q_i(\gamma_k) - \max_{\gamma \in \mathcal{H}} \sum_{i=1}^{\lceil T/H \rceil} Q_i(\gamma) + \max_{\gamma \in \mathcal{H}} \sum_{i=1}^{\lceil T/H \rceil} Q_i(\gamma) - \mathbb{E}_{\texttt{EXP3}} \left[\sum_{i=1}^{\lceil T/H \rceil} Q_i(\gamma_i) \right] \right] \\ &\leq \mathbb{E}_N \left[\max_{\gamma \in \mathcal{H}} \sum_{i=1}^{\lceil T/H \rceil} Q_i(\gamma) - \mathbb{E}_{\texttt{EXP3}} \left[\sum_{i=1}^{\lceil T/H \rceil} Q_i(\gamma_i) \right] \right] \\ &\leq \mathbb{E}_N \left[\max_{\gamma \in \mathcal{H}} \sum_{i=1}^{\lceil T/H \rceil} Q_i(\gamma) - \mathbb{E}_{\texttt{EXP3}} \left[\sum_{i=1}^{\lceil T/H \rceil} Q_i(\gamma_i) \right] \Big| \mathcal{U} \right] \mathbb{P}(\mathcal{U}) \, . \end{aligned}$$

We introduce

$$Y_i(\gamma_j) = \frac{Q_i(\gamma_j)}{2\sigma H} \, .$$

For all γ in $\mathcal{H}$, $Y_i(\gamma)$ lies in $[0, 1]$. Therefore,

$$master \leq 2\sigma H \mathbb{E}_N \left[\max_{\gamma \in \mathcal{H}} \sum_{i=1}^{\lceil T/H \rceil} Y_i(\gamma) - \mathbb{E}_{\texttt{EXP3}} \left[\sum_{i=1}^{\lceil T/H \rceil} Y_i(\gamma_i) \right] \Big| \mathcal{U} \right] \, .$$

The last step consists in using (Auer et al., 2002, Corollary 3.2). We have,

$$\max_{\gamma \in \mathcal{H}} \sum_{i=1}^{\lceil T/H \rceil} Y_i(\gamma) \leq \frac{T}{H} \, .$$

All the conditions of Corollary 3.2 in Auer et al. (2002) are met and we obtain:

$$master \leq 4\sigma H \sqrt{e-1} \sqrt{\frac{T}{H} \text{card}(\mathcal{H}) \log(\text{card}(\mathcal{H}))} \, .$$

■

The two parts of regret in Equation (25) are bounded in Proposition 3 and Proposition 4 respectively. Combining them, we get our main result below:

Theorem 2 *Under Assumptions 1-2, the regret of* `BOB-BVD-GLM-UCB` *when setting* $H = \lfloor d\sqrt{T} \rfloor$ *satisfies:*

$$\mathbb{E}[R_T] = \tilde{\mathcal{O}} \left(R_\mu d^{2/3} T^{2/3} \max \left(B_{T,\star}, d^{-1/2} T^{1/4} \right)^{1/3} \right) \, .$$

Remark 5 *This theorem establishes an upper-bound for the expected regret in the Generalized Linear Bandits framework when the variational budget is unknown. When $B_{T,\star}$ is sufficiently large ($B_{T,\star} \geq d^{-1/2} T^{1/4}$) the obtained bound can not be improved. Yet, there is still a gap with the lower bound when the variation budget is small. This can be explained by the frequent restarts in the BOB framework.*

Proof Using Proposition 4 and Proposition 3, we obtain:

$$\mathbb{E}\left[R_T\right] \leq 2\sigma\frac{T}{H} + C_1 R_\mu \beta_H^\star \sqrt{dT}\sqrt{2T(1-\gamma^\star) + \frac{T}{H}\log\left(1+\frac{2L^2}{d\lambda(1-\gamma^\star)}\right)}$$
$$+\ C_2 R_\mu \frac{2}{\sqrt{T}}\frac{1}{1-\gamma^\star} + \frac{3C_3 R_\mu}{\log(2)}\frac{B_{T,\star}\log(T)}{1-\gamma^\star} + 4\sigma H\sqrt{e-1}\sqrt{\frac{T}{H}\text{card}(\mathcal{H})\log(\text{card}(\mathcal{H}))}$$

First note that $\text{card}(\mathcal{H}) = N$ defined in Equation (21) scales as $\log(T)$ and $\beta_H^\star$ scales as $\sqrt{d\log(T)}$. By plugging $H = \lfloor d\sqrt{T} \rfloor$ in the upper-bound we obtain:

$$\frac{T}{H} = \mathcal{O}(d^{-1/2}\sqrt{T})\ .$$

$$\beta_H^\star\sqrt{dT}\sqrt{2T(1-\gamma^\star) + \frac{T}{H}\log\left(1+\frac{2L^2}{d\lambda(1-\gamma^\star)}\right)} = \widetilde{\mathcal{O}}\left(d\sqrt{T}\sqrt{\max\left(\frac{TB_{T,\star}^{2/3}}{d^{2/3}T^{2/3}}, \frac{T}{d\sqrt{T}}\right)}\right)$$
$$= d^{2/3}T^{2/3}\max(B_{T,\star}^{1/3}, d^{-1/6}T^{1/12})$$
$$= d^{2/3}T^{2/3}(\max(B_{T,\star}, d^{-1/2}T^{1/4}))^{1/3}\ .$$

$$\frac{1}{\sqrt{T}}\frac{1}{1-\gamma^\star} = \mathcal{O}\left(\frac{T^{1/6}}{d^{2/3}B_{T,\star}^{2/3}}\right)\ .$$

$$\frac{B_{T,\star}}{1-\gamma^\star} = \mathcal{O}\left(d^{2/3}B_{T,\star}^{1/3}T^{2/3}\right)\ .$$

$$H\sqrt{\frac{T}{H}\text{card}(\mathcal{H})\log(\text{card}(\mathcal{H}))} = \widetilde{\mathcal{O}}\left(d^{1/2}T^{3/4}\right)\ .$$

To conclude we notice that when $B_{T,\star} \leq d^{-1/2}T^{1/4}$,

$$d^{1/2}T^{3/4} = d^{2/3}T^{2/3}(\max(B_{T,\star}, d^{-1/2}T^{1/4}))^{1/3}\ .$$

On the contrary, when $B_{T,\star} \geq d^{-1/2}T^{1/4}$,

$$d^{1/2}T^{3/4} \leq d^{2/3}T^{2/3}(\max(B_{T,\star}, d^{-1/2}T^{1/4}))^{1/3}\ .$$

Finally, keeping the highest order term yields the announced result. ■

Appendix F. Experimental set-up

This section is dedicated at providing useful details about the illustrative experiments presented in Section 6. The logistic setting at hand is characterized by the constants $S = L = 1$. At each round, the environment randomly draws 10 news arms, presented to the agent. All algorithms use the same ℓ_2 regularization parameter $\lambda = 1$. The sequence $\theta_\star^t$ evolves as follows: we let $\theta_\star^t = (0, 1)$ for $t \in [1, T/3]$. Between $t = T/3$ and $t = 2T/3$ we smoothly rotate $\theta_\star^t$ from $(0, 1)$ to $(1, 0)$. Finally we let $\theta_\star^t = (0, 1)$ for $t \in [2T/3, T]$. Easy computations show that the total variation budget is

$$B_T = (2T/3) \sin\left(\frac{3\pi}{4T}\right) \simeq 1.5 \, .$$

We used:

$$\gamma = 1 - \left(\frac{B_T}{dT}\right)^{2/3} \simeq 0.995,$$

recommended by the asymptotic analysis for `D-LinUCB` and `BVD-GLM-UCB`. We solve the projection step of `GLM-UCB` and `BVD-GLM-UCB` by (constrained) gradient-based methods, thanks to the SLSQP solver of `scipy`.

Remark *In our experiments, we did not report performances of the algorithms from Russac et al. (2020a) (which use a similar projection step as in Filippi et al. (2010)). Because such algorithms are based on* discrete *switches of the reward signal, their behavior in this slowly-varying environment is largely sub-optimal. Indeed, in our experiment the number of abrupt-changes is* $\Gamma_T = 1000$*. For exponentially weighted algorithms, the recommended asymptotic value for the weights becomes* $\gamma \simeq 0.70$*, which in turns leads to algorithms that over-estimate the non-stationary nature of the problem, and perform poorly in practice.*