Perturbed History Exploration in Stochastic Subgaussian Generalized Linear Bandits

by

Shuai Liu

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

We consider stochastic generalized linear bandit (GLB) problems when the reward distributions are log-concave and subgaussian. We consider for this problem the perturbed history exploration (PHE) algorithmIn each round of its operation, PHE perturbs the observed rewards by adding fresh noise to them, fits a model to this perturbed data and selects the arm that has the highest reward according to the fitted model. The appeal of PHE is that it is efficient whenever model fitting and best arm selection enjoy efficient implementations. In this thesis, we present a refinement of the basic perturbed history exploration (PHE) algorithm, whereas the perturbations are adapted to the structure of GLBs. Our main result is a novel bound on the regret of the resulting algorithm. Building on an idea that was worked out for stochastic logistic bandits, a special case of GLBs, we prove that the negative log-likelihood function on the observed data is a generalized self-concordant function. This allows us to obtain regret bounds that extend previous state-of-the-art results from special GLBs to our setting, achieving a new state-of-the-art. Finally, to reduce the computation cost, we present a rarely-switching variant of PHE. The resulting method is shown to suffer a small constant-factor multiplicative increase of the regret. To the best of the author's knowledge, this is the first result that shows that randomized algorithms can also be sped up by reducing the frequency with which they update what action should be played.

Preface

A preface is required if you need to describe how parts of your thesis were published or co-authored, and what your contributions to these sections were. Also mention if you intend to publish parts of your thesis, or have submitted them for publication. It is also required if ethics approval was needed for any part of the thesis.

Otherwise it is optional.

See the FGSR requirements for examples of how this can look.

We must know - we will know!

– David Hilbert.

Acknowledgements

Put any acknowledgements here, such as to your supervisor, and supervisory committee. Remember to list funding bodies, and external scholarships. The acknowledgements can't be more than 2 pages in length.

Acknowledgements are optional, but are recommended by the FGSR.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Effective exploration is key to the success of algorithms that learn to optimize long term reward while interacting with their environments [LS18]. Algorithms based on perturbed history exploration (PHE) follow the optimal policy given a model fitted to data where the past observed rewards are randomly perturbed [Kve+19b]. The appeal of PHE is that it promises to **reduce** exploration to the widely studied problems of efficient **model fitting** and **efficient policy optimization** against a fixed model. To apply PHE, the main question is whether an appropriate reward perturbation can be designed, which results in an efficient and effective exploration method.

PHE is only one of the many possible ways of **using randomization to induce exploration**. As such, it has a few alternatives even if we restrict to methods that use randomization. We claim, however, that it stands out among these due to its simplicity and universality. The first documented case of using randomization for inducing exploration is due to Thompson [Tho33], who suggested a Bayesian approach, where a model is sampled from the posterior is used for action selection. There are numerous challenges with this approach, which is often dubbed as **Thompson sampling** after its inventor, also known as **posterior sampling**.

The challenges are the following: *(i)* exact sampling from the posterior may be intractable, *(ii)* approximate sampling can be costly and *(iii)* can ruin performance [PAD19], and, finally, *(iv)* sampling from the posterior can in fact be insufficient to achieve good, robust performance, e.g., in linear bandits

[HB20]. Another early approach is to **directly randomize action selection**, as in ϵ-greedy, or Boltzmann exploration. The weaknesses of these approaches is that it is not obvious how to tune them for robust performance [LS18; OVW14; Osb+17]. Yet another approach is to **randomize parameters** of models that are used to predict rewards, or, more generally, values [OVW14; Osb+16; Osb+17]. With nonlinear models, the choice of the parameter of the noise distribution becomes nontrivial [Osb+16].

For generalized linear bandits, Kveton et al. [Kve+19c] proposed a noise distribution and proved that the resulting algorithm achieves state-of-the-art regret among randomized methods. However, it remains unclear to what degree this result can generalize to more complicated setting. The challenge here is that if parameters are transformed in a complicated way by a nonlinear model, an appropriate noise distribution will be hard to choose.

While PHE is not completely immune to this issue, it may be true that good reward perturbations are easier to find than good parameter perturbations. In particular, in this work, we put forward the following conjecture:

Conjecture 1. *Good reward perturbations are those that set the standard deviation of the noise to be added to the reward associated with a data point to match the uncertainty level associated with the data point: The more uncertain the reward prediction would be at a data point, the more noise is to be added to it.*

The **main contributions** in this thesis are the following: *(i)* It is proved that subgaussian generalized linear bandits (sGLBs) satisfy a self-concordance property, a result that was previously unnoticed and which immediately gives great improvements over the state-of-the-art for sGLBs; *(ii)* The approach of Kveton et al. [Kve+19c] is modified by adding a new initialization method and by adapting the reward perturbations to match Conjecture 1 stated earlier; *(iii)* It is shown that the resulting method achieves regret with dominant term $O(d^{\frac{3}{2}}\sqrt{n})$ where d is the number of free parameters and n is the number of rounds; and *(iv)* It is shown that one can apply rarely switching to the proposed randomized method, with almost no loss in performance. This is the first time that rarely switching is shown to work together with a randomized method. In the case of PHE, this means that model-fitting and policy optimization only need to be performed $O(\log n)$ times for a horizon of length n, which is an exponential improvement in the runtime. As expected from previous results on linear bandits, there is a tradeoff between the dependence of the regret on the dimension d and whether the method is applicable for large action sets where iterating through all actions is not an option, but an efficient linear optimization oracle is available.

The contributions of this work is: for chapter 3, the proofs were done with collaboration with Alex Ayoub, David Janz and Csaba Szepesvári. For chapter 4, the proofs were done by the author.

Chapter 2

Related Work

Ensembling is a close relative to parameter perturbation and also to PHE: Ensembling methods use perturbed data to create a number of alternative models and make decisions by following in each step a model that is randomly selected from the so-created ensemble [LR17]. While ensemble sampling has seen much empirical success, for example, in reinforcement learning [OAC18; Osb+16; Osb+17], it tends to be costly. For example, the current state of the art that guarantees robust performance requires a poly(n)-sized ensemble [Qin+22]. Even if this is brought down to a constant, the method will still be expensive to run if model training is costly, though the same applies to PHE. Various **bootstrapping** based approaches have also been explored. Bootstrapping alone is insufficient, as can be easily seen by thinking of Bernoulli bandits when an optimal arm starts in an unlucky manner and generates only zero rewards [Kve+19a; OV15]. This realization is what originally led to the idea of **perturbed history exploration** (PHE) [Kve+19a; Kve+19b; Kve+19c; Kve+20]. PHE is known to achieve state-of-the-art results in unstructured multi-armed, and linear bandits [Kve+19b; Kve+20], but little is known beyond these cases about PHE. In particular, the only result available for PHE for generalized linear bandits assumes that the arms are all a multiple of some basis vector of the standard Euclidean basis [Kve+19c]. Needless to say, this greatly limits the scope of their result and because of this, up to this work, it remained unclear whether PHE can also work for generalized linear bandits. Resolving this open question was one of the main motivations behind this work. Indirect support for the strength of PHE is that PHE can be seen as **follow-the-perturbed-leader**, which is known as a robust method in online learning in adversarial environments [Han57; KV05]. As noted earlier, the appeal of PHE is its simplicity and that it reduces exploration to the well-studied problems of efficient model fitting and policy optimization. The compute cost of PHE can often be further reduced by performing model fitting and policy optimization in an incremental fashion (e.g., [Kve+19c; Kve+20]).

Besides randomized methods, effective exploration can be induced by **optimistic algorithms** [ACF02; Fau+20; JOA10; LR85; RV13], or by following the **exploration by optimization** approach [Fos+21; FGH23; LS20; RV18]. As opposed to PHE, where in the presence of simple perturbation distributions compute efficiency mainly depends on whether efficient model fitting and optimization methods are available, the efficiency of optimistic and exploration by optimization approaches depends on whether complex optimization problems admit efficient solutions. Finding such efficient methods is highly nontrivial and is usually done on a case by case basis, which limits the applicability of these methods, and can lead to algorithms that are overspecialized, which increases complexity and which is a problem for practicioners who are looking for simple, and generally applicable solutions.

Besides PHE, there is a rich line of work in GLB as well as its special case: logistic bandit. In general, algorithms in GLB assume subgaussian reward while some of the works in GLB assume bounded reward and works in logistic bandits, which is a special case of GLB, have to assume binary reward. In the remaining of this section, n is the number of rounds, d is the dimension of the unknown parameter θ_*, which is assumed to lie in a set $\Theta \subset \mathbb{R}^d$, which is in the ℓ_2 ball of radius $S > 0$. The arm set $\mathcal{X} \subset \mathbb{R}^d$ is assumed to lie in the ℓ_2-ball of radius 1 (Filippi et al. [Fil+10] allows arbitrary radius and pays a logarithmic price for the radius). We denote the cardinality of the arm set by K, which is allowed to be infinite. The regret bounds ignore logarithmic factors other than those that depend on the number of arms. We use $y_{\max}$ to denote the bound on the reward. We use $L \geq 1$ to denote the upper bound on the Lipschitz constant of the reward function μ, which is $1/4$ for logistic bandits, κ is the worst-case, inverse sensitivity of the reward function. The GLB framework was first introduced by Filippi et al. [Fil+10] proposes GLM-UCB algorithm and the analysis of GLM-UCB which shows that GLM-UCB achieves a performance of order $\tilde{\mathcal{O}}(y_{\max}L\kappa d\sqrt{n})$. The reward in Filippi et al. [Fil+10] is assumed to be bounded. Another algorithm SupCB-GLM achieving a regret of order $\tilde{\mathcal{O}}(L\kappa\sqrt{dn\log(K)})$ was proposed by Li et al. [LLZ17] in which finite number of arms are assumed, improving upon Filippi et al.[Fil+10] by a factor of $\sqrt{d}$. GLB-TSL is proposed by Abeille et al. [AL+17] to solve GLB that enjoys a regret upper bound of order $\tilde{\mathcal{O}}(Ld^{3/2}\kappa\sqrt{n})$. Later Kveton et al. [Kve+19c] improved the analysis that tightens the regret upper bound of GLM-TSL to the order of $\tilde{\mathcal{O}}(L^{3/2}\kappa^{3/2}d\sqrt{n\log K}+\kappa d^2)$ which also assumes finite number of arms. Apart from the improvement on GLM-TSL, Kveton et al. [Kve+19c] proposes GLM-FPL which enjoys a regret bound of order $\tilde{\mathcal{O}}(L^2\kappa^2 d\sqrt{n\log K} + \kappa d^2)$ on the assumption that the number of non-zero elements in each feature vector is **at most** 1. In consideration of the space complexity and time complexity of the existing GLB algorithms, Kwang et al. [Jun+17] proposes GLOC that enjoys a constant space and time complexity. GLOC itself achieves a regret bound of order $\tilde{\mathcal{O}}(L^2\kappa d\sqrt{n})$. GLOC-TS, the posterior sampling extention of GLOC, enjoys a regret bound of order $\tilde{\mathcal{O}}(L^2\kappa d^{3/2}\sqrt{n})$ that scales linearly with $d^{3/2}$, being far from d of GLOC. Towards closing the gap, QGLOC was proposed, enjoying a regret bound of order $\tilde{\mathcal{O}}(L^2\kappa d^{5/4}\sqrt{n})$.

Logsitic bandit is a special case of GLM that has been extensively investigated on. Faury et al. [Fau+20] improved the guarantee on UCB algorithm, by proposing LogUCB-2, a variant of GLM-UCB on logistic bandit, that uses the properties of self-concordant functions and enjoys a regret bound of order $\tilde{\mathcal{O}}(\sqrt{L}d\sqrt{n}+\kappa)$, pushing κ to the second-order term for the first time for logistic bandits. Russac et al. [Rus+20] proposes SCD-GLUCB that generalizes the approach from Faury et al. [Fau+20] to GLBs, enjoying a regret bound of order $\tilde{\mathcal{O}}(y_{\max}\sqrt{L\kappa}d\sqrt{n})$ with the bounded reward assumption. Following Faury et al. [Fau+20], Abeille et al. [AFC21] tighten the regret bound of UCB-type algorithm to minimax optimal by proposing OFULog, another variant of GLM-UCB as well as proving a lower bound that matches the regret upper bound of OFULog. The OFULog algorithm enjoys a regret upper bound of order $\tilde{\mathcal{O}}(d\sqrt{\dot{\mu}(x_*^\top \theta_*)} + \kappa)$. However, the problem of logistic bandit is closed by Abeille et al. [AFC21] statistically but not computationally: OFULog is computational intractable. Faury et al. [Fau+22] solves this problem by designing the OFU-ECOLog algorithm, where an efficient local learning procedure is added on top of the OFULog algorithm, reducing the total computational cost of the whole algorithm to $O(d^2 \log(1/\varepsilon))$ many operations. A warm-up procedure is needed in OFU-ECOLog algorithm whose number of steps required may exceed the total number of rounds n in practice. To tackle this problem, an adaptive procedure that rejects the collected data on-the-fly and adapt the optimization constraint based on the rejected data is introduced to make the algorithm practical to use. Adding the efficient learning procedure to posterior sampling results in the TS-ECOLog procedure is also given by Faury et al. [Fau+22], which enjoys a regret upper bound of order $\tilde{\mathcal{O}}(d^{3/2}\sqrt{\dot{\mu}(x_*^\top \theta_*)} + \kappa L)$.

Chapter 3

Preliminaries

In this chapter, we first introduce some notations that will be used throughout this thesis. Then we review two most renowned sequential decision making frameworks: stochastic *multi-armed bandits* (MABs) with finite arms and stochastic *linear bandits* (LBs). These two frameworks will be helpful to understanding the setting that this thesis is working on, the stochastic *Generalized Linear Bandits* (GLBs) setting, which as the name implies, is a generalization of stochastic linear bandits. The exploration and exploitation tradeoff in sequential decision making is then briefly summed up and explained in the multi-armed bandit framework. With all of these preparations we are ready to dive into the problem setting of this thesis: stochastic subgaussian generalized linear bandits. We introduce the problem setting of this thesis, including the main assumptions, the optimization targets, as well as the notations that will be used. As one can infer from the title, subgaussianity is one of the most critical assumptions of this work so we use a section to review its definition as well as some of its useful properties. The generalized self-concordant functions, one of the key tools we use to tackle the challenges in GLBs, is introduced right after the section of subgaussian distributions. Finally two methods to deal with the exploration and exploitation dilemma as well as their implementations in stochastic linear bandits framework and the stochastic generalized linear bandits framework are reviewed: posterior sampling and perturbed history exploration.

3.1 Notation

The following notations are used throughout this thesis: For a positive integer n, we let $[n]$ denote the set $\{1, ..., n\}$. All vectors (unless transposed) are column vectors. For any *positive semi-definite (PSD)* matrix $M \in \mathbb{R}^{d \times d}$ and $x \in \mathbb{R}^d$, we define $\|x\|_M = \sqrt{x^\top M x}$. For two $d \times d$ PSD matrices M_1 and M_2, we write $M_1 \succeq M_2$ if $x^\top M_1 x \geq x^\top M_2 x$ for all $x \in \mathbb{R}^d$. $\mathbb{I}\{\cdot\}$ is used to denote the indicator function of an event and $\tilde{\mathcal{O}}$ is used for big-O notation up to logarithmic factors. Denote $B_2(d)$ as a d-dimensional ball with radius 1, $B_2(d) = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$. $\text{poly}(a_1, a_2, \dots)$ is used to denote a function that is polynomial in the scalar inputs $a_1, a_2, \dots$. For a twice differentiable function $f : \mathbb{R} \to \mathbb{R}$, $\dot{f}(\cdot)$ and $\ddot{f}(\cdot)$ denote its first and second order derivative with respect to their own arguments.

3.2 Stochastic Multi-armed Bandits with Finitely Many Arms

A stochastic multi-armed bandit with k arms is a sequential decision making framework where an *learner* makes decisions in an environment in which there are k choices from a space $\mathcal{A}$ and each of them is associated with a distribution ν_i with mean μ_i. Since the number of choices is finite, we can establish a bijection between $\mathcal{A}$ and $[k]$, which allows us to refer to $[k]$ as $\mathcal{A}$ in this section. The distributions $\{\nu_i\}_{i=1}^k$ are not disclosed to the learner. As the analogy between bandit[1] and the decision environment, each choice is called an *arm* and $\mathcal{A}$ is called the arm space.

3.2.1 The interaction protocol

The learner and the environment interacts sequentially over n rounds. In each round $t \in [n]$, the learner makes a decision $a_t \in [k]$ from the k arms according to a policy, which is characterized by a probability kernel $\pi_t(\cdot | a_1, y_1, \dots, a_{t-1}, y_{t-1})$,

[1] The name comes from the slot machines in casinos which are also known as one-armed bandit

Algorithm 1 The interaction protocol of a multi-armed bandit with k arms

```
for each round $t = 1, 2, \dots, n$ do
    The agent chooses an action $a_t \in [k]$
    The environment samples a reward $y_t \in \mathbb{R}$ from the distribution $\nu_{a_t}$
    The agent potentially updates its decision strategy after receving the reward
end for
```

and the decision is fed into the environment. After receiving the learner's decision, the environment samples a *reward* $y_t \in \mathbb{R}$ from the distribution ν_{a_t},

$$y_t = \mu_{a_t} + \eta_t,$$

where η_t is a zero-mean noise. The reward is then revealed to the learner. Based on the observed reward, the learner may update its policy π_{t+1}. The interaction protocol is summarized in Algorithm 1. Stochastic MABs are also known as *unstructured* bandits which means by pulling an arm a, only information about the reward sampled from ν_a is disclosed to the learner, or the learner cannot infer anything about arms other than the arm a.

3.3 Stochastic Linear Bandits

Linear bandit is a sequential decision making framework where the learner interacts with an environment in which there are potentially *infinitely* many arms from the arm space $\mathcal{A}$. Each arm $a \in \mathcal{A}$ is associated, by a bijection $\phi : \mathcal{A} \to \mathbb{R}^d$, to a d-dimensional *feature vector* $x \in \mathcal{X}$ where $\mathcal{X} = \{\phi(a) | a \in \mathcal{A}\} \subset \mathbb{R}^d$ is the set of all feature vectors. The reward mean of pulling arm a is:

$$\mu_a = \phi(a)^\top \theta_*,$$

where $\theta_* \in \mathbb{R}^d$ is known as the model parameter.

The set of all feature vectors $\mathcal{X}$ and the bijection ϕ is accessible to the learner while θ_* is not. It is common in literature to assume the l_2-norm of the feature vectors for all arms are bounded by 1 and the l_2 norm of the model parameter is bounded:

Algorithm 2 The interaction protocol with a linear bandit

```
for each round $t = 1, 2, \dots, n$ do
    The agent chooses an action $x_t \in \mathcal{X}$
    The environment samples a reward $y_t = x_t^\top \theta_* + \eta_t$
    The agent potentially update the decision strategy after receving the reward
end for
```

Assumption 3.3.1. *There exists $S > 0$ for which $\|\theta_*\|_2 \leq S$.*

Assumption 3.3.2. *For all $x \in \mathcal{X}$, $\|x\|_2 \leq 1$.*

3.3.1 The interaction protocol

Since ϕ is a bijection between $\mathcal{A}$ and $\mathcal{X}$, in the context of LBs, we represent an arm a with its feature vector $x = \phi(a)$. The learner and the environment interacts sequentially over n rounds. In each round $t \in [n]$, the learner makes a decision $x_t \in \mathcal{X}$ according to a policy, characterized by a probability kernel $\pi_t(\cdot|x_1, y_1, \dots, x_{t-1}, y_{t-1})$, and the decision is fed into the environment. After receiving the learner's decision, the environment samples a *reward* $y_t \in \mathbb{R}$ from the distribution ν_{a_t},

$$y_t = x_t^\top \theta_* + \eta_t,$$

where η_t is a zero-mean noise. Typically we assume the noise is σ-subgaussian:

Assumption 3.3.3. *For $\sigma > 0$, for all $a \in \mathcal{A}$, the noise η_t satisfies*

$$\mathbb{E}[\exp(\lambda \eta_t)|x_1, y_1, ..., x_{t-1}, y_{t-1}] \leq \exp(-\lambda^2 \sigma^2 / 2), \forall \lambda \in \mathbb{R}.$$

The reward is then revealed to the learner. Based on the observed reward, the learner may update its policy π_{t+1}. The interaction protocol is summarized in Algorithm 2. One of the main differences between MABs and LBs is that LBs are *structured* bandits, which is defined to be bandits that are not unstructured. When pulling an arm x, the learner gets some information about the model parameter θ_*, which can be used to infer the means of other arms since $\mathcal{X}$ is revealed to the learner.

3.4 The Optimization Objectives

In this section, we review one of the most frequently used optimization objectives: cumulative (pseudo) regret minimization. One natural thought is that the learner would like to maximize the pay-off, i.e., the sum of the rewards collected, during the sequential interaction with the environment. Maximizing the cumulative reward collected is equivalent to minimizing the cumulative regret, which is defined to be the expectation of the difference between the maximum reward that could have been achieved if the learner knew the best arm, and the reward of the arms pulled. For MAB, the cumulative regret is

$$\bar{R}_n^{\mathrm{MAB}} = n\mu_* - \mathbb{E}_{\nu\pi}\left[\sum_{t=1}^{n} y_t\right],$$

where $\mu_* = \max_{i\in[k]} \mu_i$ and $\mathbb{E}_{\nu\pi}$ is the expectation under the probability measure $\mathbb{P}_{\nu\pi}$ induced by the interconnection between the learner and the environment. Note that the randomness of the reward y_t in each round t does not only come from the reward distribution ν_{a_t}, but also comes from the policy π used by the learner.

For LBs, the cumulative regret is

$$\bar{R}_n^{\mathrm{LB}} = nx_*^\top\theta_* - \mathbb{E}_{\nu\pi}\left[\sum_{t=1}^{n} x_t^\top\theta_*\right],$$

where $x_* \in \arg\max_{x\in\mathcal{X}} x^\top\theta_*$. The definition of the cumulative regret removes the randomness by taking expectation. Cumulative pseudo regret incorporate this randomness by removing the expectation in the definition of cumulative regret, that is, the difference between the maximum reward that could have been achieved if the learner knew the best arm, and the reward of the arms pulled. For MAB, the cumulative pseudo regret is

$$R_n^{\mathrm{MAB}} = n\mu_* - \sum_{t=1}^{n} \mu_{a_t},$$

and for LBs, we the cumulative pseudo-regret is

$$R_n^{\mathrm{LB}} = nx_*^\top\theta_* - \sum_{t=1}^{n} x_t^\top\theta_*,$$

If a learner is able to achieve a small cumulative pseudo-regret with probability at least $1-\delta$ for $\delta \in (0,1)$, it can achieve a small cumulative regret by carefully setting δ. The target of a learner is to incur a sublinear (pseudo) regret, that is, $R_n = o(n)$ as $n \to \infty$. In that case, asymptotically, the mean (pseudo) regret will be 0 as the number of interaction rounds n goes to infinity. Otherwise, if the regret is linear, it indicates that the learner did not learn enough information about the optimal arm(s) from the interaction, which causes it making incorrect choices in every round on average. We mainly focus on the pseudo-regret in this thesis so the term "regret" refers to the cumulative pseudo-regret for MABs, LBs and the forthcoming GLBs.

3.5 Exploration and Exploitation Trade-off

One of the most renowned challenge in multi-armed bandit framework is the exploration and exploitation tradeoff where a learner needs to balance between *exploiting* the current knowledge by selecting the best known arm and *exploring* arms to improve the current knowledge. It is a challenge in the sense that if one does not explore enough, chances are that the true optimal (or near-optimal) arms are not identified, hence they are not pulled much, resulting into sub-optimal decisions. For example, if a learner stops exploring after learning only about arms that are at most Δ-optimal in the first cn rounds, for some $c \in (0,1)$, that is, $\mu_* - \max_{t \leq cn} \mu_{a_t} \geq \Delta$ for some $\Delta > 0$ and exploits its current knowledge, that is, pulling the arm with the highest mean so far for the remaining $(1-c)n$ rounds. Then the regret in this case is at least $\Delta(1-c)n$, which is at least linear in n.

On the other hand, if one explores too much, or does not exploit enough, resources are wasted by making meaningless exploration decisions, resulting into sub-optimal decisions as well. For example, if a learner has already learned about the optimal arm a_* and still keeps exploring, meaning it is keeping pulling suboptimal arms even if with knowledge of the optimal arm, unnecessary increment in regret is therefore incurred and it can be harmful to the final regret. For example, the learner still explores $c'n$ rounds in total after knowing the optimal arm for some $c' \in (0, 1)$. Let $\Delta' = \mu_* - \max_{\mu_i \neq \mu_*, i \in [k]} \mu_i$ be the gap between the arm with highest mean μ_* and the arm with second-highest mean. The regret in this case is at least $\Delta' c' n$, which is also at least linear in n.

Judiciously balancing exploration and exploitation is thus key to keep the regret low.

3.6 Problem Setting

We consider the stochastic generalized linear bandit problem [Fil+10]. At the beginning of each round $t = 1, 2, \ldots$, the learner chooses an action x_t from the set of arms $\mathcal{X} \subset \mathbb{R}^d$, which is potentially infinite. Next, a scalar reward y_t is incurred. Conditionally on past observations, $h_{t-1} := (x_1, y_1, \ldots, x_{t-1}, y_{t-1})$, the distribution of the reward y_t follows a distribution that is a member of a **single parameter natural exponential family**, where the exponential family parameter $u_t \in \mathbb{R}$ at time step t is $u_t = x_t^\top \theta_*$, with $\theta_* \in \Theta \subset \mathbb{R}^d$ being an unknown parameter:

$$y_t \sim p(y; u_t)\, \rho(dy),$$

where $p(y; u) = \exp(yu - \psi(u))$, $y, u \in \mathbb{R}$ and $\psi, h : \mathbb{R} \to \mathbb{R}$ are suitable normalizing mappings and ρ is a reference distribution over the reals, which we call the base distribution. It is assumed that u_t is so that $(y; u_t)$ is well-defined. (Properties of natural exponential family distributions will be explored in Section 3.7, where the reader will find details about ψ, h and other quantities.) After the reward is observed, the process repeats. In particular, in step $t + 1$, the learner can choose x_{t+1} based on h_t.

The goal of the learner is to pick the best possible action, denoted by $x_* = \arg\max_{x \in \mathcal{X}} \mu(x^\top \theta_*)$ so as to maximize the total reward collected over n steps of interaction. This goal is equivalent to minimizing the cumulative pseudo regret,

$$R(n) = \sum_{t=1}^{n} \mu(x_*^\top \theta_*) - \mu(x_t^\top \theta_*)\,,$$

where

$$\mu(u) = \int p(y; u)\rho(dy)$$

is the mean of the reward y with parameter u. Similar to MAB and LB, pseudo-regret is defined to be the difference between the maximum reward that could have been achieved if the learner knew the best action, and the expected reward of the arms taken. If a learner achieves sublinear regret (that is, $R(n) = o(n)$), or, the regret per time step converges to zero as n gets large, then learner can be seen as having learned the best possible action.

Let $\mathbb{P}$ be the probability measure induced by the interconnection between the learner and the environment. Denote $\mathbb{P}_t(\cdot) := \mathbb{P}(\cdot|h_{t-1})$ and $\mathbb{E}_t(\cdot) := \mathbb{P}(\cdot|h_{t-1})$. We make the following additional assumptions, which are borrowed from the work that introduced the generalized linear bandit framework [Fil+10]:

Assumption 3.6.1. *The noise in the reward, $\epsilon_t = y_t - \mu(x_t^\top \theta_*)$, is conditionally subgaussian with a parameter $\sigma > 0$:*

$$\mathbb{E}_t[\exp\{\lambda\varepsilon_t\}] \leq \exp\{\lambda^2\sigma^2/2\}, \qquad \textit{for all } \lambda \in \mathbb{R}\,.$$

Assumption 3.6.2. *We have $\theta_* \in \Theta$ where Θ is compact, and a value S is known such that Θ is included in the ℓ^2-ball of radius S with center zero: For any $\theta \in \Theta$, $\|\theta\|_2 \leq S$.*

Assumption 3.6.3. *For all $x \in \mathcal{X}$, $\|x\|_2 \leq 1$.*

The subgaussian assumption, as the name suggests, expresses that the tails of the distribution decay at least as fast as that of the normal (or, Gaussian) distribution. Naturally, this holds for normal distributions, but also holds, for example, when y_t belongs to a bounded interval such as the beta or hypergeometric distribution. We include more details about subgaussianity in Section 3.9. In the presence of the subgaussian assumption, we refine the problem that is discussed in this thesis to be subgaussian generalized linear bandit (sGLB) problem as it is a subset of all generalized linear bandit problems. As long as the action set is bounded, the assumption that the action set is a subset of the unit ball can be always met by increasing the radius of the ball containing the parameter set—it serves to simplify the presentation. Finally using $\dot{\mu}$ to denote the derivative of μ with respect it argument (which is guaranteed to exist, cf. Section 3.7), we let

$$L := \sup_{x\in\mathcal{X},\theta\in\Theta} \dot{\mu}(x^\top\theta) \quad \text{and} \quad \kappa := \sup_{x\in\mathcal{X},\theta\in\Theta} \frac{1}{\dot{\mu}(x^\top\theta)}$$

be the largest growth rate of μ and a parameter that characterizes how flat μ becomes over its effective domain, respectively. Note that knowing the model allows one to obtain the exact numerical values of κ and L, which can thus be used in the algorithms. We impose that κ cannot be unbounded above:

Assumption 3.6.4. *The reciprocal of the first order derivative of the mean function $\mu(\cdot)$ has a finite least upper bound on its domain $\{x^\top\theta | x \in \mathcal{X}, \theta \in \Theta\} \subset \mathbb{R}$:*

$$\kappa = \sup_{x\in\mathcal{X},\theta\in\Theta} \frac{1}{\dot{\mu}(x^\top\theta)} < \infty.$$

As we shall see later, for every round t and for $\theta \in \Theta$, the first order derivative $\dot{\mu}(x_t^\top\theta)$ equals to the conditional variance of the reward distribution y_t associated to x_t, the arm pulled in round t. Therefore, what Assumption 3.6.4 imposes in fact is that the reward distribution does not degenerate to a Dirac distribution over its domain.

3.7 Generalized Linear Models

We start with the definition of generalized linear models (GLMs) [McC19]. All unattributed results in this section can be found either in the above reference, or in [Bro86].

Definition 1 (Natural exponential family)**.** *Let ρ be a probability distribution over the reals. The natural exponential family $(P_\theta)_{\theta\in D}$ with base ρ is a family of probability distributions over the reals with the following properties:*

1. *$D = \{\theta \in \mathbb{R} : S(\theta) < \infty\}$ where we let $S(\theta) = \int e^{\theta y}\rho(dy)$.*

2. *$P_\theta(dy) = \frac{e^{\theta y}}{S(\theta)}\rho(dy)$ for $\theta \in D$.*

Proposition 3.1. *The set $D \subset \mathbb{R}$ is a convex subset of the reals.*

From Definition 1, if a random variable $Y \in \mathbb{R}$ has a distribution from exponential family with parameter $\theta \in \mathbb{R}$, then its density function (d.f.) with respect to ρ can be written in the form

$$p_\theta(y) = \exp(y\theta - \psi(\theta)) \,, \tag{3.1}$$

where

$$\psi(\theta) = \log S(\theta) \,.$$

Recall that for a random variable Y, its moment generating function is defined via $M_Y(t) = \mathbb{E}_\theta[\exp(tY)]$ for all values of $t \in \mathbb{R}$ where the expectation on the right-hand side exist, while for the same values, the cumulant generation function of Y is $K_Y(t) = \log M_Y(t)$.

Fix $\theta \in D$ and let $Y \sim P_\theta$. By abusing notation, we let $M_\theta(t) := M_Y(t)$ and $K_\theta(t) := M_Y(t)$. Then,

$$\begin{aligned} M_\theta(t) &= \int \exp(y\theta - \psi(\theta)) \exp(ty)\rho(dy) \\ &= \frac{1}{\exp(\psi(\theta))} \int \exp((\theta + t)y)\rho(dy) \\ &= \exp(\psi(t + \theta) - \psi(\theta)) \,, \end{aligned}$$

where the last equality comes from the definition of ψ. It follows that

$$K_\theta(t) = \psi(t + \theta) - \psi(\theta). \tag{3.2}$$

It follows from the standard result of GLMs (page 38, Brown [Bro86]) that ψ is an infinitely differentiable function. Therefore, one can safely take the nth derivative of $K(t)$ for $n \in \mathbb{N}$:

$$K^{(n)}(t) = \psi^{(n)}(t + \theta). \tag{3.3}$$

From the standard properties of cumulant generating function (CGF) (for example, section 5.6.2 in Khuri [Khu03]) that the first, second and third derivative of CGF evaluated at 0 are the first, second and third centered moments of Y, respectively:

$$\psi'(\theta) = \mathbb{E}_\theta[Y]\,, \tag{3.4}$$

$$\psi''(\theta) = \mathbb{E}_\theta[(Y - \mathbb{E}[Y])^2]\,, \tag{3.5}$$

$$\psi'''(\theta) = \mathbb{E}_\theta[(Y - \mathbb{E}[Y])^3]\,, \tag{3.6}$$

where $\psi'(\theta), \psi''(\theta), \psi'''(\theta)$ are the first order derivative, second order derivate and third order derivative of ψ evaluated at point θ. Recalling that we introduced $\mu(\theta) = \int yP_\theta(dy)$, from Eq. (3.4) we have

$$\mu(\theta) = \psi'(\theta)\,, \qquad \theta \in D\,.$$

3.8 Revisiting Generalized Linear Bandits

In generalized linear bandit (GLB), the conditional distribution of the observed reward $y_t \in \mathbb{R}$ associated to the pulled arm x_t with parameter $\theta \in \mathbb{R}^d$ in round t, is from the natural exponential family with parameter $x_t^\top \theta$. For every round t, as is stated in Eqs. (3.4) to (3.6), it follows that

$$\psi'(x_t^\top \theta) = \mathbb{E}_\theta[y_t|x_t] = \mu(x_t^\top \theta), \tag{3.7}$$

$$\psi''(x_t^\top \theta) = \mathbb{E}_\theta[(y - \mu(x_t^\top \theta_*))^2|x_t] = Var_\theta[y_t|x_t], \tag{3.8}$$

$$\psi'''(x_t^\top \theta) = \mathbb{E}_\theta[(y - \mu(x_t^\top \theta_*))^3|x_t]\,, \tag{3.9}$$

where $Var_\theta[y_t|x_t]$ denotes the conditional variance of the reward distribution of y_t given x_t and the parameter of the d.f. of y_t is θ. The mean function $\mu(u)$ is also known as *link function*, because it links the inner product $x_t^\top \theta$ to the mean of the reward associated to arm x_t.

For consistency, in what follows we assume the following, which is required to make the problem well-defined. As such, this assumption will not be repeated in the various results.

> **The link function μ is defined on $\mathbb{R}$, that is, the domain D of the natural exponential family considered (cf. Definition 1) is $\mathbb{R}$.**

3.8.1 Log-likelihood function

Let $\mathcal{D} = \{(x_s, y_s)\}_{s=1}^{n}$ be a subset of $\mathbb{R}^d \times \mathbb{R}$. We define the (unregularized) log-likelihood of a parameter θ as

$$\begin{aligned} \mathcal{L}\left(\theta; \mathcal{D}\right) &= \sum_{s=1}^{n} \log(p(y_s; x_s^\top \theta)) \\ &= \sum_{s=1}^{n} y_s x_s^\top \theta + \psi(x_s^\top \theta). \end{aligned}$$

We also define the regularized log-likelihood of θ with a regularizer $\lambda > 0$ to be:

$$\begin{aligned} \mathcal{L}_\lambda\left(\theta; \mathcal{D}\right) &= \sum_{s=1}^{n} \log(p(y_s; x_s^\top \theta)) - \frac{\lambda}{2}\|\theta\|_2^2, \\ &= \sum_{s=1}^{n} y_s x_s^\top \theta + \psi(x_s^\top \theta) - \frac{\lambda}{2}\|\theta\|_2^2. \end{aligned}$$

3.9 Subgaussian Distributions

Our assumption on the subgaussianity (Assumption 3.6.1) of the noise plays an important role in both the algorithm design and the analysis. The subgaussian assumption is commonly used in bandit literature Lattimore & Szepesvári [LS18]. It is reasonable to assume that rewards are bounded and, as we shall see later, subgaussianity is a generalization of boundedness. Subgaussianity essentially assumes that the tail of the noise is no heavier than a Gaussian distribution, or equivalently, the sample mean of a group of i.i.d. observations concentrates at least as fast as a Gaussian distribution. Informally, a random variable X is subgaussian if there is a Gaussian distribution whose tail of density function can fully "dominate" that of X, as is shown in Fig. 3.1. Formally, subgaussian random variables are defined to be:

Definition 2 (σ-subgaussian random variable)**.** *A random variable X with mean μ is σ-subgaussian if for all $\lambda \in \mathbb{R}$, it holds that* $\mathbb{E}[\exp(\lambda(X - \mu))] \leq \exp(\lambda^2 \sigma^2 / 2)$.

Example 1 (Exercise 2.4 of Wainwright [Wai19])**.** *If X is a zero-mean random variable that is supported on an interval $[a, b]$ then X is $\frac{b-a}{2}$-subgaussian.*

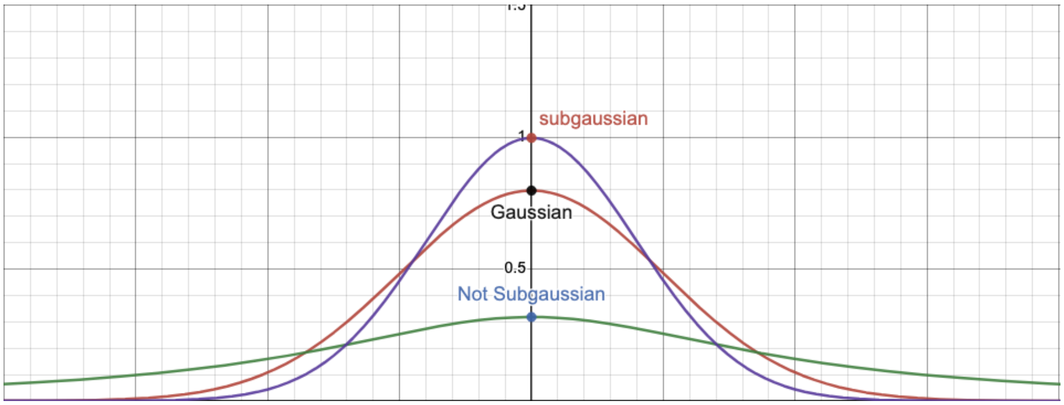

Figure 3.1: A plot visualizing subgaussianity in 1-d case. The red curve is Gaussian distribution with mean 0 and variance 0.5: $\mathcal{N}(0, 0.5)$. The green curve is the "notorious" Cauchy distribution whose tail cannot be dominated by a Gaussian distribution. The purple curve is Gaussian distribution with mean 0 and variance 0.4: $\mathcal{N}(0, 0.4)$, which itself is a subgaussian distribution. Note that one only tells if a distribution is subgaussian by its tail behavior so purple curve is still subgaussian even if it is higher than the red curve in the area around the mean.

We review a few important properties that will be used in the presented thesis. For the cited results and more the reader can refer to Vershynin [Ver18], Wainwright [Wai19] and Lattimore & Szepesvári [LS18].

One of the nice properties of subgaussianity is that it plays well with linear transformations:

Lemma 3.1. *Suppose X is a σ-subgaussian. Then $X - c$ is σ-subgaussian for all $c \in \mathbb{R}$.*

Proof. Let $\mu = \mathbb{E}[X]$. By definition of σ-subgaussian random variable, for all $\lambda \in \mathbb{R}$:

$$\mathbb{E}[\exp(\lambda(X - \mu))] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right).$$

The left hand side can also be written as

$$\mathbb{E}[\exp(\lambda(X - \mu))] = \mathbb{E}[\exp(\lambda(X - c - (\mu - c)))],$$

where

$$\mu - c = \mathbb{E}[X] - c = \mathbb{E}[X - c].$$

The proof finishes by noting that the definition of σ-subgaussian is satisfied for the random variable $X - c$. □

Lemma 3.2 (cf. Lemma 5.4 in [LS18])**.** *Suppose X, X_1 and X_2 are σ, σ_1 and σ_2 subgaussian respectively. Then*

1. *cX is $|c|\sigma$-subgaussian for all $c \in \mathbb{R}$;*
2. *$X_1 + X_2$ is $\sqrt{c_1^2 + c_2^2}$-subgaussian.*

We use Hoeffding's inequality to characterize the tail behavior of both a single subgaussian random variable and the sample mean of i.i.d random variable.

Lemma 3.3 (Hoeffding's inequality for subgaussian random variables)**.** *Let $X_1, ...X_n$ be independent random variables such that X_i has mean $\mu_i \in \mathbb{R}$ and subgaussian parameter $\sigma_i \geq 0$. Then for all $t \geq 0$, we have that*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}(X_i - \mu_i)\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right).$$

The tail of a subgaussian random variable decays exponentially, which can be seen by setting $n = 1$, that is, for a subgaussian random X with mean μ,

$$\mathbb{P}(|X - \mu| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{3.10}$$

Allowing for an error probability of at most $\delta \in (0, 1]$, the sample means of i.i.d. subgaussian random variables deviate from their average mean by at most $\sigma\sqrt{\log(1/\delta)/(2n)}$:

Corollary 1. *Let $X_1, ..., X_n$ be $n \geq 1$ independent σ-subgaussian random variables, $\mu_i = \mathbb{E}[X_i]$ for $i \in [n]$, and $\delta \in (0, 1]$. Then,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i)\right| \geq \sigma\sqrt{\frac{\log(1/\delta)}{2n}}\right) \leq \delta.$$

Since the tail decays exponentially, one would expect that its moments are also bounded because the integral of $x^k\exp(-x)$ is bounded.

Lemma 3.4. *Let X be a zero-mean σ-subgaussian random variable, then for any positive integer $k \geq 1$,*

$$\mathbb{E}[|X|^k] \leq (2\sigma^2)^{k/2}k\Gamma(k/2),$$

where $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx, t > 0$ *is the gamma function. Furthermore,*

$$\Gamma(t) \leq 3t^t\,.$$

Proof. Since for all integers $k \geq 1$, the random variable $|X|^k$ is non-negative, the expectation can be expressed in the following form:

$$\begin{aligned}
\mathbb{E}[|X|^k] &= \int_0^\infty \mathbb{P}(|X|^k \geq u)du \\
&= \int_0^\infty \mathbb{P}(|X| \geq t)kt^{k-1}dt && \text{(change of variable } u = t^k\text{)} \\
&\leq \int_0^\infty 2\exp\left(-\frac{t^2}{2\sigma^2}\right)kt^{k-1}dt && \text{(Eq. (3.10))} \\
&\leq (2\sigma^2)^{k/2}k\Gamma(k/2). && \text{(Definition of gamma function)}
\end{aligned}$$

□

3.10 Generalized Self-concordant Functions

Definition 3 (Definition 1 of Sun et al. [ST17]). *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a three times continuously differentiable function on the open domain dom(ϕ). Let $\nu > 0$ and $M_\phi \geq 0$ be two constants. We say that ϕ is (M_ϕ, ν)-generalized self-concordant if $|\phi'''(t)| \leq M_\phi \phi''(t)^{\nu/2}$ for all $t \in$ dom(ϕ).*

Examples of generalized self concordant functions include the logistic function, exponential function, and log-barrier functions.

Generalized self-concordant functions have several important properties that will prove essential in removing the trivial assumptions of Kveton et al. [Kve+19c] while simultaneously improving upon the results of Fillipi et al. [Fil+10], Li et al. [LLZ17]. The use of self-concordant analysis for stochastic bandits is due to Faury et al. [Fau+20]. However their work investigates UCB-style algorithms with a logistic link function. We notice that the negative (unregularized) log-likelihood function in sGLM is automatically self-concordant. To prove this we need the following result. For a function f, let dom(f) denote the domain of f.

Proposition 3.2 (Proposition 1 of Sun et al. [ST17]). *Let f_i be (M_{f_i}, ν)-generalized self-concordant functions satisfying Definition 3, where $M_{f_i} \geq 0$ and $\nu \geq 2$ for $i = 1, 2, ..., m$. Then for $\beta_i > 0$, $i = 1, 2, ..., m$, the function $f(x) = \sum_{i=1}^{m} \beta_i f_i(x)$ is well-defined on $dom(f) = \cap_{i=1}^{m} dom(f_i)$, and is (M_f, ν)-generalized self-concordant with the same order $\nu \geq 2$ and the constant*

$$M_f := \max\{\beta_i^{1-\frac{\nu}{2}} M_{f_i} | 1 \leq i \leq m\} \geq 0.$$

Lemma 3.5. *Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{m}$ be m observations. Under assumptions 3.6.1 and 3.6.4, for $\theta \in \Theta$ it holds that the negative unregularized log-likelihood of the reward distribution $L(\mathcal{D}; \theta)$ is generalized self-concordant with $M_{\mathcal{L}} = O(\sigma^3 \kappa)$ and $\nu = 2$.*

Proof. We first show that for a single (x, y)-pair, $-\mathcal{L}(y; x^\top \theta)$ is $(M_{\mathcal{L}}, 2)$-generalized self-concordant. Let $u = x^\top \theta$.

Take the derivatives of the negative unregularized log-likelihood function $-\mathcal{L}((x, y); \theta) = -\log p((x, y); \theta) = yu - \psi(u)$. Doing so gives $\frac{\partial^2}{\partial u^2}(-\mathcal{L}(y; u)) = \psi''(u)$ and $\frac{\partial^3}{\partial u^3}(-\mathcal{L}(y; u)) = \psi'''(u)$. By Eq. (3.8) and Eq. (3.9), we have that $\psi''(u) = Var_\theta[y|x]$ and $\psi'''(u) = \mathbb{E}_\theta[(y - \mu(x^\top \theta))^3 | x]$. This establishes that $M_{\mathcal{L}}$ depends on the second and third central moments of y given x. Thus in order to show that $-\mathcal{L}((x, y); \theta)$ is $(M_{\mathcal{L}}, 2)$-generalized self-concordant, it suffices to show that

$$\frac{\left|\mathbb{E}_\theta\left[\left(y - \mu\left(x^\top \theta\right)\right)^3 | x\right]\right|}{Var_\theta[y|x]} \leq M_{\mathcal{L}}$$

for some finite $M_{\mathcal{L}}$. By Assumption 3.6.4, it holds that $Var_\theta[y|x] \geq \frac{1}{\kappa}$. Furthermore since y is σ-subgaussian, from Lemma 3.4 it holds that

$$\left|\mathbb{E}_\theta\left[\left(y - \mu\left(x^\top \theta\right)\right)^3 | x\right]\right| \leq \mathcal{C} \sigma^3 3^{3/2},$$

where $\mathcal{C}$ is a universal constant. By setting $M_{\mathcal{L}} \leq \mathcal{C} \sigma^3 3^{3/2} \kappa$ we have shown $M_{\mathcal{L}}$ is finite. Now that it was shown for a single (x, y)-pair the negative log-likelihood is $(M_{\mathcal{L}}, 2)$-generalized self-concordant, it will be shown that the log-likelihood of a sequence of $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{m}$ remains $(M_{\mathcal{L}}, 2)$-generalized

Distribution of Y	Variance	Third Central Moment	$M_{\mathcal{L}}$
Normal(μ, σ^2)	σ^2	0	0
Exponential(λ)	$1/\lambda^2$	$2/\lambda^3$	$2\mathbb{E}[Y]$
Poisson(λ)	λ	λ	1
Binomial(n, p)	$np(1-p)$	$np(1-p)(1-2p)$	1
Geometric(k, p)	$(1-p)/p^2$	$(p-2)(p-1)/p^3$	$2\mathbb{E}[Y] - 1$ for $p \neq 1$
Gamma(k, θ)	$k\theta^2$	$2k\theta^3$	2θ
Beta(α, β)	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$\frac{-2\alpha\beta(\alpha-\beta)}{(\alpha+\beta)^3(\alpha+\beta+1)(\alpha+\beta+2)}$	1
Inverse Normal(μ, λ)	μ^3/λ	$3\mu^5/\lambda$	$3\mathbb{E}[Y]^2/\lambda$

Table 3.1: The self-concordance parameter for some common exponential family distributions. One can derive $M_{\mathcal{L}}$ by dividing the third central moment by the variance. Note for the distributions above, $M_{\mathcal{L}}$ either behaves like a constant or scales with the expected value of the distribution.

self-concordant. Recall that

$$-\mathcal{L}(\mathcal{D}; \theta) = \sum_{i=1}^{m} \psi(x_i^\top \theta) - y_i(x_i^\top \theta) = \sum_{i=1}^{m} -\mathcal{L}((x_i, y_i); \theta).$$

Since it was shown that for an arbitrary (x_i, y_i)-pair that $-\mathcal{L}((x_i, y_i); \theta)$ is $(M_{\mathcal{L}}, 2)$-generalized self-concordant, by Proposition 3.2, it follows that $-\mathcal{L}(\mathcal{D}; \theta)$ is also $(M_{\mathcal{L}}, 2)$-generalized self-concordant. The proof is complete. $\square$

Note that Lemma 3.5 is an extremely loose bound. Lots of the distributions that are frequently used in the exponential family do not have a $M_{\mathcal{L}}$ that scales with κ. For example, in logistic link function case, $M_{\mathcal{L}} = 1$. For other distributions, Table 3.1 summarizes some of the common exponential family distributions. We therefore believe there is a bound on $M_{\mathcal{L}}$ that is independent of κ and it is left for future work. In addition, with some further reasonable assumptions, for example, the density function $p_\theta(y)$ is log-concave with respect to Lebesgue measure or counting measure on their support, the dependence on κ can be removed.

We first present the formal definition of log-concave function:

Definition 4. *A non-negative function $f : \mathbb{R}^d \to \mathbb{R}_+$ is log-concave if its domain is a convex set and if it satisfies the inequality*

$$f(\theta x + (1-\theta)y) \geq f(x)^\theta f(y)^{1-\theta},$$

for all $\theta \in (0, 1)$ and $x, y \in dom(f)$.

In this thesis, we only consider all log-concave reward distributions that are either discrete or continuous.

Definition 5 (Log-concave distributions)**.** *We call a distribution ρ over the reals log-concave, if either one of the following two conditions is satisfied:*

1. *ρ has an uncountable domain: $\rho(dy) = f(y)\lambda(dy)$, where $f(y)$ is a log-concave function and λ is Lebesgue measure.*

2. *ρ has a domain subset the integers $\mathbb{N}$: $\rho(dy) = g(y)\nu(dy)$, where g satisfies that $g(i+1)^2 \geq g(i)g(i+2)$ for all $i \in \mathbb{N}$ and ν is the counting measure over the integers.*

In the first case, we say that ρ is a continuous, while in the second case we say that it is a discrete log-concave distribution. We will make the following assumption:

Assumption 3.10.1. *The base measure $\rho(\cdot)$ of reward distribution associated to each arm is a continuous log-concave distribution.*

We removed discrete distributions from the above assumption for technical reasons.

Example 2. *Normal distribution $p(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$ and exponential distribution $p(x; \lambda) = \mathbb{I}(x > 0)\lambda e^{-\lambda x}$ are both log-concave densities.*

A random variable with log-concave density function is called a log-concave random variable. We now present another definition and an auxiliary lemma that is necessary for us to prove the self-concordant property of a log-concave random variable.

Definition 6 (Definition 1.1 of Schudy et al. [SS12])**.** *A random variable Z is called moment bounded with parameter $L > 0$ if for any integer $i \geq 1$,*

$$\mathbb{E}[|Z|^i] \leq i \cdot L \cdot \mathbb{E}[|Z|^{i-1}].$$

Proposition 3.3 (Lemma 7.3 of Schudy et al. [SS12])**.** *Any log-concave random variable X satisfying condition 1 of Assumption 3.10.1 is moment-bounded with parameter $L = \frac{1}{\ln 2}\mathbb{E}[|X|] \approx 1.44\mathbb{E}[|X|]$.*

Proposition 3.4 (Lemma 7.7 of Schudy et al. [SS12])**.** *For any nonnegative-valued, log-concave random variable X satisfying condition 2 of Assumption 3.10.1 is moment-bounded with parameter $L = 1 + \mathbb{E}[|X|]$.*

The technical condition we did not allow discrete log-concave distributions in Assumption 3.10.1 is because it is not know whether the last proposition continuous to hold if the condition that X is nonnegative valued is removed.

We are now ready to prove that the unregularized log-likelihood of sGLMs are self-concordant with parameter independent of κ under the assumption that reward distribution is log-concave in y.

Lemma 3.6. *Under assumptions 3.6.1 and 3.10.1, for $\theta \in \mathbb{R}^d$, it holds that the negative unregularized log-likelihood of the reward distribution $-L(\mathcal{D};\theta)$ is generalized self-concordant with $M_{\mathcal{L}} = O(\sigma)$ and $\nu = 2$.*

Proof. Translation does not affect the concavity of a function, that is, if a function $x \mapsto \psi(x)$ is a concave function on its domain $\mathrm{dom}(\psi)$, then $x \mapsto \psi(x+c)$ for $c \in \mathbb{R}$ is still a concave function on $\mathrm{dom}(\psi) \oplus c$ where $\mathrm{dom}(\psi) \oplus c = \{x + c | x \in dom(\psi)\}$. Then if a random variable is log-concave, it is still log-concave after the random variable is translated. We also know that subgaussianity is translation-invariant from Lemmas 3.1 and 3.2. We therefore safely center the random variable y. We first show that for a single (x, y)-pair, $-\mathcal{L}(y; x^\top\theta)$ is $(M_{\mathcal{L}}, 2)$-generalized self-concordant. Let $u = x^\top\theta$.

Take the derivatives of the unregularized log-likelihood function $-\mathcal{L}((x,y);\theta) = -\log p((x,y);\theta) = -yu + \psi(u)$. Doing so gives $\frac{\partial^2}{\partial u^2}(-\mathcal{L}(y;u)) = \psi''(u)$ and $\frac{\partial^3}{\partial u^3}(-\mathcal{L}(y;u)) = \psi'''(u)$. By Eq. (3.8) and Eq. (3.9), we have that $\psi''(u) = Var_\theta[y|x]$ and $\psi'''(u) = \mathbb{E}_\theta[(y - \mu(x^\top\theta))^3|x]$. This establishes that $M_{\mathcal{L}}$ depends on the second and third central moments of y given x. Thus in order to show that $-\mathcal{L}((x,y);\theta)$ is $(M_{\mathcal{L}}, 2)$-generalized self-concordant, it suffices to show that

$$\frac{|\mathbb{E}_\theta\left[y^3|x\right]|}{Var_\theta[y|x]} \leq M_{\mathcal{L}},$$

for some $M_{\mathcal{L}}$. Note that moment can be upper bounded by the absolute moment:

$$\begin{aligned}\frac{|\mathbb{E}_u[y^3]|}{Var_u[y]} &\leq \frac{\mathbb{E}_u[|y|^3]}{\mathbb{E}_u[y^2]} \\ &\leq 3 \cdot \max\{1.44\mathbb{E}[|y|], 1 + \mathbb{E}[|y|]\} && \text{(Propositions 3.3 and 3.4)} \\ &\leq 3 \cdot 1.44 \cdot (1 + \sqrt{2} \cdot \Gamma(1/2)\sigma). && \text{(Lemma 3.4, } \max\{a, b\} \leq a + b\text{)}\end{aligned}$$

By setting $M_\mathcal{L} \leq \mathcal{C}\sigma$ we have shown $M_\mathcal{L}$ is finite. Now that it was shown for a single (x, y)-pair the negative log-likelihood is $(M_\mathcal{L}, 2)$-generalized self-concordant, it will be shown that the negative log-likelihood of a sequence of $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$ remains $(M_\mathcal{L}, 2)$-generalized self-concordant. Recall that

$$-\mathcal{L}(\mathcal{D}; \theta) = \sum_{i=1}^{m} \psi(x_i^\top \theta) - y_i(x_i^\top \theta) = \sum_{i=1}^{m} -\mathcal{L}((x_i, y_i); \theta).$$

Since it was shown that for an arbitrary (x_i, y_i)-pair that $-\mathcal{L}((x_i, y_i); \theta)$ is $(M_\mathcal{L}, 2)$-generalized self-concordant by Proposition 3.2, it follows that $-\mathcal{L}(\mathcal{D}; \theta)$ is also $(M_\mathcal{L}, 2)$-generalized self-concordant. The proof is complete. □

Note that for generalized linear models, $\dot{\mu}(z) = \psi''(z) = \frac{\partial^2}{\partial u^2}(-\mathcal{L}(\cdot; z))$ and $\ddot{\mu}(z) = \psi'''(z) = \frac{\partial^3}{\partial u^3}(-\mathcal{L}(\cdot; z))$. Thus, lemma 3.6 implies the following corollary.

Corollary 2. *Under assumptions 3.6.1 and 3.10.1, for all $z \in \mathbb{R}$, the link function $\mu(\cdot)$ satisfies*

$$|\ddot{\mu}(z)| \leq M_\mu \dot{\mu}(z)$$

where $M_\mu = 3 \cdot 1.44 \cdot (1 + \sqrt{2} \cdot \Gamma(1/2)\sigma)$.

3.10.1 Properties of the Link Function μ

A key idea when analyzing generalized linear bandits is to link reward deviations $\mu(x^\top \theta_1) - \mu(x^\top \theta_2)$ to parameter deviations $\theta_1 - \theta_2$. Linearizing μ using a first-order Taylor expansion is the standard technique to accomplish this. Specifically, for any $y, z \in \mathbb{R}$ define

$$\alpha(y, z) = \int_0^1 \dot{\mu}((1 - v)y + z))dv = \int_0^1 \dot{\mu}(vy + (1 - v)z)dv = \alpha(z, y)\,.$$

From the definition of α it follows that $\alpha(x, y) = \alpha(y, x)$, i.e., it is a symmetric function of its arguments. Now, the mean-value theorem gives the following identity: For any u_1, u_2 in the domain D of μ,

$$\mu(u_1) - \mu(u_2) = \alpha(u_1, u_2)(u_1 - u_2). \tag{3.11}$$

Therefore controlling α is an essential step in deriving tight bounds for generalized linear bandits. In order to accomplish this task, we will make abundant use of the properties of sGLB (Corollary 2) and the properties self-concordant functions. Our main tool is the following result, which is cited without a proof.

Lemma 3.7 (Corollary 2 of Sun et.al. [ST17]). *Under assumptions 3.6.1 and 3.10.1, for $x, y \in D$ it holds that:*

$$\dot{\mu}(x)\frac{1 - \exp(-M_\mu|x-y|)}{M_\mu|x-y|} \leq \alpha(x, y) \leq \dot{\mu}(x)\frac{\exp(M_\mu|x-y|) - 1}{M_\mu|x-y|}.$$

Furthermore,

$$\alpha(x, y) \geq \dot{\mu}(x)(1 + M_\mu|x-y|)^{-1}.$$

An immediate corollary of this result will help us bound the growth of the derivative of the link function:

Lemma 3.8. *Let $\Delta = |x - y|$ for $x, y \in D$. Then it holds that*

$$\dot{\mu}(x) \leq \exp(M_\mu|\Delta|)\dot{\mu}(y).$$

Proof. Applying Lemma 3.7 twice, one gets that

$$\begin{aligned}
\dot{\mu}(x) &= \left(\frac{1 - \exp(-M_\mu|\Delta|)}{M_\mu|\Delta|}\right)^{-1}\left(\frac{1 - \exp(-M_\mu|\Delta|)}{M_\mu|\Delta|}\dot{\mu}(x)\right) \\
&\leq \left(\frac{1 - \exp(-M_\mu|\Delta|)}{M_\mu|\Delta|}\right)^{-1}\alpha(x, y) && \text{(Lemma 3.7)} \\
&\leq \left(\frac{1 - \exp(-M_\mu|\Delta|)}{M_\mu|\Delta|}\right)^{-1}\frac{\exp(M_\mu|\Delta|) - 1}{M_\mu|\Delta|}\dot{\mu}(y) && \text{(Lemma 3.7)} \\
&= \frac{\exp(M_\mu|\Delta|) - 1}{1 - \exp(-M_\mu|\Delta|)}\dot{\mu}(y) = \exp(M_\mu|\Delta|)\dot{\mu}(y).
\end{aligned}$$

□

3.10.2 The Curvature of the Log-likelihood Function

Maximum likelihood estimation, where the Fisher information matrix plays an important role, is primarily used in GLMs (page 38 [McC19]). For $\mathcal{D} = \{x_i, y_i\}_{i=1}^{t-1}$ and $\theta \in \mathbb{R}^d$, we introduce the observed Fisher information matrix, that is, a sample-based version of Fisher information matrix that coincides with the Hessian of the negative log-likelihood:

$$\bar{H}_t(\theta) = \nabla_\theta^2 \left(-\mathcal{L}(\mathcal{D}; \theta)\right) = \sum_{s=1}^{t-1} \dot{\mu}(x_s^\top \theta) x_s x_s^\top .$$

We also introduce the Hessian of the negative regularized log-likelihood:

$$H_t(\theta) = \nabla_\theta^2 \left(-\mathcal{L}_\lambda(\mathcal{D}; \theta)\right) = \sum_{s=1}^{t-1} \dot{\mu}(x_s^\top \theta) x_s x_s^\top + \lambda I. \tag{3.12}$$

From Lemma 3.7, one is able to approximate $H_t(\theta)$ by replacing $\dot{\mu}(x_s^\top \theta)$ in each summand by $\alpha(x_s^\top \theta, x_s^\top \theta')$ for $\theta' \in \mathbb{R}^d$, the result of which we define as:

$$G_t(\theta, \theta') = \sum_{s=1}^{t-1} \alpha(x_s^\top \theta, x_s^\top \theta') x_s x_s^\top + \lambda I.$$

Lemma 3.9. *For any $\theta_1, \theta_2 \in \mathbb{R}^d$, we have*

$$G_t(\theta_1, \theta_2) \succeq (1 + M_\mu \mathfrak{D})^{-1} H_t(\theta_1),$$
$$G_t(\theta_1, \theta_2) \succeq (1 + M_\mu \mathfrak{D})^{-1} H_t(\theta_2),$$

where $\mathfrak{D} = \max_{s \in [t-1]} |x_s^\top (\theta_1 - \theta_2)|$.

Proof. By Lemma 3.7, we have that for all $s \in [t-1], \alpha(x_s^\top \theta_1, x_s^\top \theta_2) \geq \dot{\mu}(x_s^\top \theta_1)(1 + M_\mu |x_s^\top (\theta_1 - \theta_2)|)^{-1} \geq \dot{\mu}(x_s^\top \theta_1)(1 + M_\mu \mathfrak{D})^{-1}$. Thus, it follows that

$$\begin{aligned} G_t(\theta_1, \theta_2) &= \sum_{s=1}^{t-1} \alpha(x_s^\top \theta_1, x_s^\top \theta_2) x_s x_s^\top + \lambda I \\ &\succeq \frac{1}{1 + M_\mu \mathfrak{D}} \sum_{s=1}^{t-1} \dot{\mu}(x_s^\top \theta_1) x_s x_s^\top + \lambda I \succeq \frac{1}{1 + M_\mu \mathfrak{D}} H_t(\theta_1). \end{aligned}$$

Since the same bounds hold for $\alpha(x_s^\top \theta_2, x_s^\top \theta_1)$, which one can get from the symmetric property of α, repeating the same steps but with θ_2 completes the proof. □

3.11 Posterior Sampling and Perturbed History Exploration

There are two popular approaches to deal with the dilemma of exploration and exploitation: *optimism in face of uncertainty* (OFU) and *posterior sampling.*

3.11.1 Optimism in Face of Uncertainty

For OFU, the learner remains optimistic about each arm. A value, also known as the upper confidence bound (UCB), is assigned to each arm based on the data observed so far such that with high probability this value is an overestimate of the unknown mean. The chosen arm in each round is the one having the highest UCB. For each arm, its UCB is computed by adding an exploration bonus to the current estimation of the mean of the reward distribution associated to the arm. If an arm is only pulled very few times, its exploration bonus needs to be large in order to ensure UCB is an overestimate of the mean of its reward distribution with high probability. As more data are collected about an arm, its exploration bonus shrinks because the confidence level is higher that the empirical estimation of its reward mean is close to the true mean. Therefore, one can expect the UCB of a suboptimal arm will eventually fall under the UCB of an optimal arm.

3.11.2 Posterior Sampling

Posterior sampling, also known as Thompson Sampling, is a heuristic algorithm that was first introduced by Thompson [Tho33] to solve a two-armed bandit problem. The learner first assumes a prior distribution on all the possible bandit environments. In each round, it samples an environment from a posterior distribution, which is updated using the data collected so far with Bayes rule, and pulls the optimal arm in that environment. Even if the heuristic is based on assuming a prior, it still works when the true parameter is a fixed parameter instead of a random variable in some cases, for example, the linear bandit case.

3.11.3 Posterior Sampling in Linear Bandits

We briefly introduce posterior sampling in LBs here. The algorithm is presented as Algorithm 3 (Abeille et al. [AL+17]). In each round t, the algorithm solves the regularized least squares problem

$$\hat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t-1} (y_s - x_s^\top \theta)^2 + \lambda \|\theta\|_2^2.$$

There exists a closed form solution to this problem which is detailed in line 5. The algorithm then samples from the a suitable multivariate distribution that perturbs $\hat{\theta}_t$, that is, $\max_{x \in \mathcal{X}} x^\top \tilde{\theta}_t \geq x_*^\top \theta_*$ with probability at least p. Finally the algorithm pulls the arm that maximizes the reward mean if the model parameter were $\tilde{\theta}_t$: $x_t = \arg\max_{x \in \mathcal{X}} x^\top \tilde{\theta}_t$. The multivariate distribution $\mathcal{D}^{\mathrm{TS}}$ is defined to satisfy the following properties [AL+17]:

Definition 7 (Definition of the multivariate distribution $\mathcal{D}^{\mathrm{TS}}$)**.** *The multivariate distribution $\mathcal{D}^{TS}$ is defined such that the following properties are satisfied:*

1. *It is a multivariate distribution on $\mathbb{R}^d$ that is absolutely continuous w.r.t. the Lebesgue measure.*
2. *There exists a strictly positive probability p such that for any $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$, the probability $\mathbb{P}_{\eta \sim \mathcal{D}^{TS}}(u^\top \eta \geq 1) \geq p$*
3. *There exists c, c', positive constants such that for all $\delta \in (0,1)$, the probability*

$$\mathbb{P}_{\eta \sim \mathcal{D}^{TS}}(\|\eta\|_2 \leq \sqrt{cd \log \frac{c'd}{\delta}}) \geq 1 - \delta$$

Section 5 of Abeille et al. [AL+17] proves that any distribution satisfying Definition 7 incurs right amount of randomness to ensure low regret for Algorithm 3. Theorem 1 of Abeille et al. [AL+17] ensures that Algorithm 3 achieves a near-optimal regret of $\tilde{\mathcal{O}}(d^{3/2}\sqrt{T})$.

3.11.4 Perturbed History Exploration

One of the downsides of posterior sampling is that exact sampling from the posterior distribution may be intractable. One way to tackle it is approximate sampling, which can be costly as well as ruining the performance (Phan

Algorithm 3 Linear Thompson Sampling

```
**Input**: $\delta, n, \lambda > 0, \mathcal{D}^{\mathrm{TS}}$
$V_1 = \lambda I$
**for** $t = 1, ..., n$ **do**
    $V_t = \sum_{s=1}^{t-1} x_s x_s^\top + \lambda I$
    $\hat{\theta}_t \leftarrow V_t^{-1} \sum_{s=1}^{t-1} x_s y_s$
    Sample $\eta_t \sim \mathcal{D}^{\mathrm{TS}}$.
    Compute parameter $\tilde{\theta}_t = \hat{\theta}_t + \beta_t(\delta) V_t^{-1/2} \eta_t$
    Choose arm $x_t \in \arg\max_{x \in \mathcal{X}} x^\top \tilde{\theta}_t$
    Observe reward $y_t$
**end for**
```

et al. [PAD19]), that is, even with small approximation error, the performance degenerates to linear regret. Another option is to use *perturbed history exploration* or *follow the perturbed leader* ([Kve+19a; Kve+19b; Kve+19c; Kve+20]), which is another type of randomized algorithm that randomly perturbs the observed reward. To be more specific, in each round, after receiving the reward given by the environment, the learner injects random noise to the observed reward and learn about the mean of the reward distribution associated to the pulled arm from the noisy reward as if it was the true reward.

3.11.5 Perturbed History Exploration in Generalized Linear Bandit

We introduce GLM-FPL (Kveton et al. [Kve+19c]), perturbed history exploration in GLBs here. The algorithm is presented as Algorithm 4. After the initial exploration round, the algorithm samples fresh i.i.d. noises z_s for each data point $(x_s, y_s)_{s=1}^{t-1}$ from a Gaussian distribution with mean 0 and variance a^2. Note that:

1. The noises are resampled in each round, that is, the noises are not reused across rounds.
2. The variance of the noises are the same across different arms, that is, even if two arms in the dataset are different, the noises injected to them are the same within the certain round. We will see in Chapter 4 that injecting arm-dependent noises significantly improves the performance.

Algorithm 4 GLM-FPL

```
Input: $\delta, n, \lambda > 0$, exploration round $\tau$, $a$
for $t = \tau + 1, ..., n$ do
    Sample $z_s \sim \mathcal{N}(0,\, a^2)$ for each $s < t$ .
    Compute parameter $\theta_t = \arg\max_{\theta \in \mathbb{R}^d} \mathcal{L}_\lambda \left(\theta; \{(x_s, y_s + z_s)\}_{s=1}^{t-1}\right)$
    Choose arm $x_t \in \arg\max_{x \in \mathcal{X}} x^\top \theta_t$
    Observe reward $y_t$
end for
```

After injecting the noises, the algorithm solves a *randomly perturbed maximum likelihood estimation* problem with the solution being θ_t and pulls the arm that maximizes the reward mean if the model parameter were θ_t: $x_t = \arg\max_{x \in \mathcal{X}} x^\top \theta_t$. Algorithm 4 is proved in Theorem 5 of Kveton et al. [Kve+19c] to enjoy a regret bound of $\tilde{O}(\kappa^2 d \sqrt{n})$ with the assumption that the number of nonzero elements of the feature vector is **at most** one.

3.12 Optimal Design Problem

Experimental design problem is a subfield of statistics. Let $X \in \mathbb{R}^{n \times d}$ and $Y \in \mathbb{R}^n$. Each row of X can be viewed a design point which is also known as an input feature vector and Y is known as labels in machine learning. In linear experimental design, we deal with a linear model $Y = X\beta + \epsilon$ where the relationship between the labels and the input feature vectors is characterized by a linear map and $\epsilon \in \mathbb{R}^d$ is a zero-mean noise vector. The optimal design problem aims to select a small subset $S \subset \{1, 2, ..., n\}$ with r rows such that the precision of estimating β is maximized on the selected design points. We introduce two problems, G-optimal design problem and D-optimal design problem.

The G-optimal design problem is defined as follows:

Definition 8 (G-optimal Design). *Let $\mathcal{X} = \{[X_{i1}, ..., X_{id}]^\top | i \in [n]\} \subset \mathbb{R}^d$ be the set of design points and $\pi : \mathcal{X} \to [0, 1]$ be a distribution on $\mathcal{X}$ where X_{ij} refers to the entry corresponding to the ith row and the jth column. Let $V(\pi) = \sum_{x \in \mathcal{X}} \pi(x) x x^\top$. The G-optimal design problem is defined as*

$$\min_{\pi \in \Delta_{\mathcal{X}}} \max_{x \in \mathcal{X}} \|x\|^2_{V(\pi)^{-1}},$$

where $\Delta_{\mathcal{X}}$ is the set of all probability measures on $\mathcal{X}$.

The D-optimal design problem is defined as follows:

Definition 9 (D-optimal Design)**.** *Let $\mathcal{X} = \{[X_{i1}, ..., X_{id}]^\top | i \in [n]\} \subset \mathbb{R}^d$ be the set of design points and $\pi : \mathcal{X} \to [0,1]$ be a distribution on $\mathcal{X}$ where X_{ij} refers to the entry corresponding to the ith row and the jth column. Let $V(\pi) = \sum_{x \in \mathcal{X}} \pi(x) x x^\top$. The D-optimal design problem is defined as*

$$\max_{\pi \in \Delta_{\mathcal{X}}} \log(\det(V(\pi))),$$

where $\Delta_{\mathcal{X}}$ is the set of all probability measures on $\mathcal{X}$.

The following proposition characterizes the equivalence between G-optimal and D-optimal design.

Proposition 3.5 (Kiefer-Wolfowitz Theorem)**.** *Recall $\mathcal{X} \subset \mathbb{R}^d$ and assume that span($\mathcal{A}$) $= \mathbb{R}^d$. For $\pi \in \Delta_{\mathcal{X}}$, let $g(\pi) = \max_{x \in \mathcal{X}} \|x\|^2_{V(\pi)^{-1}}$ be the objective function of G-optimal design and $f(\pi) = \log(\det(V(\pi)))$ be the objective function of D-optimal design. The following are equivalent:*

1. *π^* is a minimizer of g.*
2. *π^* is a maximizer of f.*
3. *$g(\pi^*) = d$*

Furthermore, there exists a minimizer π^* of g such that $|supp(\pi^*)| \leq d(d+1)/2$ where $supp(\pi) = \{x \in \mathcal{X} : \pi(x) > 0\}$ denotes the support of π for $\pi \in \Delta_{\mathcal{X}}$.

The D-optimal design problem is a convex optimization problem [BV04] and can be solved by many algorithms, for example, Frank-Wolfe algorithm [Bub+15]. For G-optimal design, the way of solving it is given by Soare et al. [SLM14]. Given a π^* defined in Proposition 3.5, we are able to sample points from π^* approximately using standard rounding procedure: **round**$(\tau, \zeta, \varepsilon)$, detailed in Algorithm 5. Extracted from Fiez et.al [Fie+19], the definition of a rounding procedure is detailed in Definition 10.

Algorithm 5 The rounding procedure given by Chpater 12 of Pukelsheim et.al[Puk06] and is detailed in Fiez et al.[Fie+19].

```
**Input**: $\tau, \zeta \in \Delta_{\mathcal{X}}$ and $\varepsilon$
$r(\varepsilon) \leftarrow (d(d+1)/2+1)/\varepsilon$
$p \leftarrow |\mathrm{supp}(\zeta)|$
$\{x_i\}_{i=1}^{p} \leftarrow \mathrm{supp}(\zeta)$
$N_i \leftarrow \lceil (\tau - \frac{1}{2}p)\zeta_i \rceil$, for all $i \leq p$
**while** $\sum_{i=1}^{p} N_i \neq \tau$ **do**
    **if** $\sum_{i=1}^{p} N_i < \tau$ **then**
        $j \leftarrow \arg\min_{i \leq p} (N_i - 1)/\zeta_i$
        $N_j \leftarrow N_j + 1$
    **end if**
    **if** $\sum_{i=1}^{p} N_i > \tau$ **then**
        $j \leftarrow \arg\max_{i \leq p} (N_i - 1)/\zeta_i$
        $N_j \leftarrow N_j - 1$
    **end if**
**end while**
$N_i \leftarrow \max(N_i, r(\varepsilon)/p)$
**return** $N_1, \ldots, N_p$, for all $i \leq p$
```

Definition 10. *A rounding procedure is an algorithm that takes as input a real $\varepsilon \in (0,1)$, a set of vectors $\mathcal{X} \subset \mathbb{R}^d$ where the dimension of the vector space spanned by $\mathcal{X}$ is d, a probability measure over $\mathcal{X}$ with finite support: $\lambda \in \Delta_{\mathcal{X}}$, and a number of samples N. Let p be the cardinality of the support of λ and $\{x_i\}_{i=1}^{p} \subset \mathcal{X}$ be the support of λ, that is, $\lambda(\{x_i\}) > 0$. Further, denote $\lambda_i = \lambda(\{x_i\})$. The rounding procedure returns a finite allocation $s = (s_1, \ldots, s_p) \in \mathbb{N}^p$, where s_i is a number assigned to x_i, satisfying the following properties:*

1. *$\sum_{i=1}^{p} s_i = N$;*

2. *There exists a function $r(\varepsilon)$ such that if $N > r(\varepsilon)$, then*

$$\max_{y \in \mathbb{R}^d} \|y\|^2_{(\sum_{i=1}^{p} s_i x_i x_i^\top)^{-1}} \leq (1+\varepsilon) \max_{y \in \mathbb{R}^d} \|y\|^2_{(\sum_{i=1}^{p} \lambda_i x_i x_i^\top)^{-1}} / N.$$

Chapter 4

Perturbed History Exploration for Subgaussian Generalized Linear Model Bandits

In this chapter, the main algorithm, Subgaussian Generalized Linear Model Bandits - Perturbed History Exploration (sGLM-PHE), is introduced.

4.1 sGLM-PHE

Our algorithm, presented as Algorithm 6, consists of two stages:

Warm-up sGLM-PHE is initialized by calling a warm-up procedure that approximates a G-optimal design [Che72]. This is necessary in order to guarantee certain good events happen with high probability. The warm-up procedure detailed in Algorithm 7, which consists of solving a G-optimal design, follows that of Jun et al. [Jun+21]. A way of solving the G-optimal design is given by Soare et al. [SLM14]. Here, ζ is a probability measure over $\mathcal{X}$, denoted by $\zeta \in \Delta_{\mathcal{X}}$, set to minimize

$$\max_{x \in \mathcal{X}} \|x\|^2_{V(\zeta)^{-1}} \quad \text{where} \quad V(\zeta) = \sum_{x \in \mathcal{X}} \zeta(x) x x^\top,$$

that is the maximum linear model confidence width over the actions. We then sample points $x_1, \ldots, x_\tau$ from ζ approximately using Algorithm 5. In our context, the rounding procedure transforms a design $\zeta \in \Delta_{\mathcal{X}}$ into a discrete allocation for any fixed number of samples τ. It is clear that we only need to

Algorithm 6 sGLM-PHE: Perturbed History Exploration for Generalized Linear bandits

```
Input: $\mathcal{X}, M_\mu, L, \sigma, n, \delta$ and $S$
Calculate $\lambda, \tau, a$ according to Section 4.2.5
$\{(x_s, y_s)\}_{s=1}^{\tau} \leftarrow$ warm-up$(\mathcal{X}, \tau, \epsilon = 0.5)$
for $t = \tau + 1, ..., n$ do
    $\hat{\theta}_t \leftarrow \arg\max_{\theta \in \mathbb{R}^d} \mathcal{L}_\lambda \left(\theta; \{x_s, y_s\}_{s=1}^{t-1}\right)$
    Sample $z_{s,t} \sim \mathcal{N}\big(0,\, a^2 \, \dot{\mu}(x_s^\top \hat{\theta}_t)\big)$ for each $s < t$
    $\theta_t \leftarrow \arg\max_{\theta \in \mathbb{R}^d} \mathcal{L}_\lambda \left(\theta; \{(x_s, y_s + z_{s,t})\}_{s=1}^{t-1}\right)$
    Choose arm $x_t \in \arg\max_{x \in \mathcal{X}} x^\top \theta_t$
    Observe reward $y_t$
end for
```

Algorithm 7 warm-up

```
Input: $\mathcal{X} \subset \mathbb{R}^d, \tau, \epsilon \in (0, 1)$
Set $\zeta = \arg\min_{\zeta \in \Delta_{\mathcal{X}}} \max_{x \in \mathcal{X}} \|x\|^2_{V(\zeta)^{-1}}$
$x_1, ..., x_\tau = \mathbf{round}(\tau, \zeta, \epsilon)$
Observe the associated rewards $y_1, ..., y_\tau$
Return: $\{(x_s, y_s)\}_{s=1}^{\tau}$
```

take care of the allocation of the arms in the support of ζ: $\{x_i\}_{i=1}^{p}$ where p is the cardinality of the support of ζ and N_i is the number of pulls of arm x_i.

Main algorithm After the warm-up phase is completed, in each of the subsequent rounds, sGLM-PHE first fits a generalized linear model to its history, $\hat{\theta}_t$ then uses this model in order to fit a second generalized linear model to its perturbed history up to round t,

$$\theta_t = \arg\max_{\theta \in \mathbb{R}^d} \mathcal{L}_\lambda \left(\theta; \{(x_s, y_s + z_{s,t})\}_{s=1}^{t-1}\right) \text{ where } z_{s,t} \sim \mathcal{N}\left(0, a\dot{\mu}(x_s^\top \hat{\theta}_t)\right).$$

Unlike Kveton et al. [Kve+19c], our algorithm uses *data dependent* pseudo-rewards, $z_{s,t}$ which are freshly sampled in each round. Allowing the pseudo-rewards to depend on the solution to the unperturbed history, $\hat{\theta}_t$ enables the algorithm to be both instance-adaptive and achieve optimal κ dependencies. sGLM-PHE then pulls the arm with the highest estimated value under the linear model θ_t. Note that the derivative of μ is the second derivative of ψ, which by Eq. (3.8), is the variance of the associated reward hence non-negative. It is therefore equivalent to pull the arm that maximizes the esimated mean and the arm with the highest estimated value under the linear model θ_t.

sGLM-PHE has two tunable parameters. The perturbation scale a controls the variance of the pseudo-rewards in the perturbed history for each of the observed rewards. It controls the amount of exploration and exploitation the algorithm performs: with higher values of a, the probability of injecting a relatively large noise is higher, yielding more exploration. The second tunable parameter is the number of warm-up rounds τ. The number of exploration rounds τ is correlated with the non-linearity of the problem instance therefore needs to be scaled with this non-linearity.

4.2 The analysis of sGLM-PHE

In this section we report the mains steps of the regret analysis. We prove the following result.

Theorem 1 (Informal version)**.** *Under assumptions 3.6.1-3.6.3 and assumption 3.10.1, with appropriately chosen parameters, the n-round regret of sGLM-PHE, $R(n)$, is bounded as*

$$R(n) = \tilde{\mathcal{O}}\left(\sigma^2 d^{3/2}\sqrt{S\dot{\mu}(x_*^\top \theta_*)n} + \mathfrak{C}\right).$$

where $\mathfrak{C} = \tilde{\mathcal{O}}(\mathrm{poly}(d, L, \kappa))$ and the exact dependence is shown in Section 4.2.5.

This is the first regret bound for an algorithm that perturbs its history that holds for feature vectors with more than one nonzero component, i.e. non-tabular features, and non-linear function approximation. We enjoy this improved regret bound by both using properties of self concordant functions and choosing the perturbations to be adaptive to the current data, by setting the exploration term to a^2 times $\dot{\mu}(x_s^\top \hat{\theta}_t)$, a data-dependent perturbation, as is mentioend in Section 3.11.

Our algorithm's regret nearly recovers the minimax-optimal regret for logistic bandits regret presented in Abeille et al. [AFC21], $\tilde{\mathcal{O}}(d\sqrt{n\dot{\mu}(x_*^\top \theta_*)})$ up to a factor of $\sqrt{d}$, which is not reducible for the class of randomized exploration algorithms [HB20]. The regret matches that of the posterior sampling algorithm for logistic bandits introduced in Faury et al. [Fau+22] without the need to solve a constrained optimization problem and *significantly* improves the regret of the posterior sampling algorithm introduced in both Abeille et al. [AL17] and Kveton et al. [Kve+19c], without the need to sample parameters from a posterior. It is also simpler than the UCB algorithm of Faury et al. [Fau+22] that requires maintaining a set of plausible parameter vectors. However Faury et al. [Fau+22] does not require solving an optimal design for their warm-up procedure, whereas our algorithm does rely on solving one. Investigating the possibility of removing the warm-up phase for our algorithm is an interesting future research direction.

4.2.1 A Tail Inequality for Self-Normalized Martingales and the Confidence Set

At the core of our analysis, we use a confidence set that is constructed with a tail inequality result for self-normalized martingales, which extends the results of Faury et al. [Fau+20], designed for logistic bandits, to sGLBs. The next lemma generalizes Lemma 7 of Faury et al.[Fau+20] that states that if ϵ has zero mean, takes values in $[-1, 1]$ and the variance of ϵ is σ^2 and its subgaussian parameter is b, then for any $\lambda \in [-1, 1]$,

$$\mathbb{E}\exp(\lambda\epsilon) \leq 1 + \lambda^2\sigma^2 \,. \tag{4.1}$$

In the generalization we allow unbounded ϵ, though we require that ϵ is the noise underlying a one-parameter exponential family distribution.

Lemma 4.1. *Fix u in the interior of its domain and let $y \sim p(\cdot; u)$, $\epsilon = y - \mu(u)$, and $\sigma^2 = \mathbb{E}\epsilon^2$. Then for any $|\lambda| \leq \log(2)/M_\mu$*

$$\mathbb{E}[\exp(\lambda\varepsilon)] \leq \exp\left(\lambda^2\sigma^2\right) . \tag{4.2}$$

Proof. From Chebyshev's inequality it follows $\epsilon = 0$ with probability one when $\sigma^2 = 0$. Hence, the statement trivially holds when $\sigma^2 = 0$ and so for the rest of the proof we will assume that $\sigma^2 > 0$.

As it is well known (see, e.g., Equation 6.1 of Rigollet et al. [Rig12]) and also easy to check, the moment generating function of $y \sim p(\cdot; u)$ is given by

$$\mathbb{E}[e^{\lambda y}] = \exp\left(\psi(u+\lambda) - \psi(u)\right) .$$

Recalling that $\varepsilon = y - \mu(u)$ gives

$$\mathbb{E}[e^{\lambda\varepsilon}] = \mathbb{E}[e^{\lambda(y-\mu(u))}] = \exp\left(\psi(u+\lambda) - \psi(u) - \lambda\mu(u)\right) .$$

Hence, by taking logarithms, Eq. (4.2) is equivalent to that

$$\psi(u+\lambda) \leq \psi(u) + \lambda\mu(u) + \lambda^2\sigma^2. \tag{4.3}$$

for $|\lambda| \leq \log(2)/M_\mu$. By Taylor's theorem, which can be applied, since, as it is well known, ψ is analytic (c.f. Section 3 of Kakade et al. [Kak+10]), for some ξ which is in the interval with endpoints u and $u + \lambda$, we have

$$\psi(u+\lambda) = \psi(u) + \lambda\dot{\psi}(u) + \frac{\lambda^2\ddot{\psi}(\xi)}{2} = \psi(u) + \lambda\mu(u) + \frac{\lambda^2\ddot{\psi}(\xi)}{2} \tag{4.4}$$

where the last equality used that $\mu(u) = \dot{\psi}(u)$. This combined with Eq. (4.3) gives that it suffices to show that

$$\ddot{\psi}(\xi) \leq 2\sigma^2 .$$

Using Lemma 3.8, and recalling that $\sigma^2 = \ddot{\psi}(u)$, we have

$$\ddot{\psi}(\xi) \leq \exp(M_\mu|\xi - u|)\ddot{\psi}(u) = \exp(M_\mu|\xi - u|)\sigma^2 \leq \exp(M_u|\lambda|)\sigma^2 \leq 2\sigma^2 ,$$

where the last inequality follows from the assumption that $|\lambda| \leq \log(2)/M_\mu$. □

Theorem 2 (Theorem 3 of Russac et al. [Rus+20] with sub-gaussian noise assumption)**.** *selfNormalizedMartingales Fix* $\sigma > 0$, $\lambda > 0$. *Let* $\{(h_t, x_t, \epsilon_t)\}_{t=0}^{\infty}$ *be a stochastic process such that for each* $t \geq 0$,

1. x_t *is* $\sigma(h_t)$*-measurable and* x_t *takes values in* $B_2(d)$*;*

2. $\mathbb{E}[\exp(\eta\epsilon_t)|h_t] \leq \exp(\sigma_t^2\eta^2)$ *holds for all* $|\eta| \leq 1/m$*, where* $\sigma_t^2 = \mathbb{E}[\epsilon_t^2|h_t]$*;*

3. $\mathbb{E}[\exp(\eta\epsilon_t)|h_t] \leq \exp(\sigma^2\eta^2/2)$ *holds for all* $\eta \in \mathbb{R}$*.*

For any $t \geq 1$ define

$$H_t \coloneqq \sum_{s=1}^{t-1} \sigma_s^2 x_s x_s^\top + \lambda I_d, \qquad S_t \coloneqq \sum_{s=1}^{t-1} \varepsilon_s x_s.$$

Then for any $\delta \in (0, 1]$:

$$\mathbb{P}\left(\exists t \geq 1, \|S_t\|_{H_t^{-1}} \geq \sigma \left(\frac{\sqrt{\lambda}}{2m} + \frac{2m}{\sqrt{\lambda}} \log\left(\frac{\det(H_t)^{1/2}\lambda^{-d/2}}{\delta}\right) + \frac{2m}{\sqrt{\lambda}} d \log(2)\right)\right) \leq \delta.$$

Proof. This proof follows almost identically to the original proof. Firstly without loss of generality let that the subgaussian parameter $\sigma = 1$, since one can always rescale S_t such that this is holds as in done in Lemma 9 of Abbasi et al.[APS11]. As was done in Russac et al.[Rus+20], all one has to do is make the following modification to the proof of Lemma 5 Faury et al.[Fau+20] we which make below.

Lemma 4.2 (Lemma 5 of Faury et al.[Fau+20])**.** *Assume the conditions of Theorem 2 hold. For $t \geq 1$ define $\bar{H}_t = \sum_{s=1}^{t-1} \sigma_s^2 x_s x_s^\top$ and for $\xi \in B_2(d)$, let $M_0(\xi) = 1$ and for $t > 0$ let*

$$M_t(\xi) \doteq \exp\left(\frac{1}{m}\xi^\top S_t - \frac{1}{m^2}\|\xi\|_{\bar{H}_t}\right).$$

Then, $\{M_t(\xi)\}_{t=0}^\infty$ is a non-negative super-martingale adapted to the filtration $\{\sigma(h_{t-1})\}_{t=0}^\infty$, where $\sigma(h_{-1})$ is the trivial σ-algebra.

Proof. For all $t \geq 1$ we have that:

$$\mathbb{E}\left[\exp\left(\frac{1}{m}\xi^\top S_t\right) | h_{t-1}\right] = \exp\left(\frac{1}{m}\xi^\top S_{t-1}\right) \mathbb{E}\left[\exp\left(\frac{\xi^\top x_{t-1}}{m}\varepsilon_{t-1}\right) | h_{t-1}\right].$$

Since $|\xi^\top x_{t-1}/m| \leq \|\xi\|_2 \|x_{t-1}\|_2/m \leq 1/m$, we can apply Lemma 4.1 to bound the second term. With this, we get

$$\mathbb{E}\left[\exp\left(\frac{1}{m}\xi^\top S_t\right) | h_{t-1}\right] \leq \exp\left(\frac{1}{m}\xi^\top S_{t-1} + \frac{\left(\xi^\top x_{t-1}\right)^2}{m^2}\sigma_{t-1}^2\right). \tag{4.5}$$

Then, writing

$$M_t(\xi) = \exp\left(\frac{1}{m}\xi^\top S_t - \frac{1}{m^2}\xi^\top \bar{H}_{t-1}\xi - \frac{\left(\xi^\top x_{t-1}\right)^2}{m^2}\sigma_{t-1}^2\right),$$

and using the inequality in Eq. (4.5) gives

$$\mathbb{E}[M_t(\xi)|h_{t-1}] \leq \exp\left(\frac{1}{m}\xi^\top S_{t-1} - \frac{1}{m^2}\xi^\top \bar{H}_{t-1}\xi\right) = M_{t-1}(\xi) \quad \text{a.s.}$$

By definition, $M_t(\xi)$ is nonnegative, hence it is a nonnegative super-martingale with respect to the filtration stated. □

Since $\{M_t(\xi)\}_{t=1}^{\infty}$ is a non-negative super martingale, the rest of the proof follows identically to the proof of Theorem 1 in Faury et al.[Fau+20]. □

We now use this result to construct a confidence set that is key to bound the prediction error:

$$R_t^{\text{PRED}} = \mu(x_t^\top \theta_t) - \mu(x_t^\top \theta_*).$$

For $t \in [n]$ and $\delta \in (0,1)$, the confidence set $C_t(\delta, \lambda)$ is defined as follows:

$$C_t(\delta,\lambda) = \left\{\theta : \left\|g_t(\hat{\theta}_t) - g_t(\theta)\right\|_{H_t^{-1}(\theta_*)} \leq \gamma_t(\delta,\lambda)\right\}. \tag{4.6}$$

where for $\theta \in \mathbb{R}^d$, $g_t(\theta)$ is defined to be the gradient of the negative regularized log-likelihood of θ on $\mathcal{D}_t = \{x_s, y_s\}_{s=1}^{t-1}$:

$$g_t(\theta) = \nabla_\theta(-\mathcal{L}_\lambda(\theta, \mathcal{D})) = \sum_{s=1}^{t-1} \mu(x_s^\top \theta)x_s + \lambda\theta, \tag{4.7}$$

and

$$\gamma_t(\delta,\lambda) = \sqrt{\lambda}\left(S + \frac{1}{2m}\sigma\right) + \frac{2m\sigma}{\sqrt{\lambda}}\log\left(\frac{2^d}{\delta}\left(1 + \frac{Lt}{d\lambda}\right)^{d/2}\right). \tag{4.8}$$

If we choose $\lambda = \lambda^*$ as follows:

$$\lambda^* = \frac{2m\sigma}{S + \frac{1}{2m}\sigma} d\log\left(\frac{2}{\delta}\sqrt{1 + Ln/d}\right), \tag{4.9}$$

then $\gamma_n(\delta, \lambda)$ becomes

$$\gamma_n(\delta) := \sqrt{8m\sigma d\left(S + \frac{1}{2m}\sigma\right)\log\left(\frac{2}{\delta}\sqrt{1 + Ln/d}\right)}. \tag{4.10}$$

By the mean value theorem, for all $x, \theta_1, \theta_2 \in \mathbb{R}^d$, it immediately follows from Eq. (3.11) that

$$\mu(x^\top\theta_1) - \mu(x^\top\theta_2) = \alpha(x^\top\theta_1, x^\top\theta_2)x^\top(\theta_1 - \theta_2). \tag{4.11}$$

The above equality allows us to link $\theta_1 - \theta_2$ to $g_t(\theta_1) - g_t(\theta_2)$ which is stated in the lemma below.

Lemma 4.3. *For all $\theta_1, \theta_2 \in \mathbb{R}^d$, the following identity holds:*

$$g_t(\theta_1) - g_t(\theta_2) = G_t(\theta_1, \theta_2)(\theta_1 - \theta_2).$$

Proof. Note that for all $x \in \mathbb{R}^d$ and $\theta_1, \theta_2 \in \mathbb{R}^d$, the following equality holds by Eq. (4.11):

$$\mu(x^\top \theta_1 - x^\top \theta_2) = \alpha(x^\top \theta_1, x^\top \theta_2) x^\top (\theta_1 - \theta_2).$$

This result applied to $g_t(\theta_1) - g_t(\theta_2)$ yields,

$$\begin{aligned} g_t(\theta_1) - g_t(\theta_2) &= \sum_{s=1}^{t-1} \alpha(x_s^\top \theta_1, x_s^\top \theta_2) x_s x_s^\top (\theta_1 - \theta_2) + \lambda(\theta_1 - \theta_2) \\ &= G_t(\theta_1, \theta_2)(\theta_1 - \theta_2). \end{aligned}$$

□

We now show that for all $t \in [n]$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$, the true model parameter lies in the confidence set $C_t(\delta, \lambda)$. We need the following auxiliary result:

Lemma 4.4. *(Determinant-Trace Inequality, Lemma 10 of Abbasi et al. [APS11]) Let $\{x_s\}_{s=1}^{\infty}$ be a sequence in $\mathbb{R}^d$ such that $\|x_s\|_2 \leq X$ for all $s \in \mathbb{N}$ and $X \geq 0$. Let λ be a non-negative scalar. For $t \geq 1$ define $V_t := \sum_{s=1}^{t-1} x_s x_s^\top + \lambda I$. The following holds*

$$\det(V_{t+1}) \leq (\lambda t X^2/d)^d.$$

Lemma 4.5. *Recall $\hat{\theta}_t$ in line 5 of Algorithm 6, g_t in Eq. (4.7), $H_t(\theta)$ in Eq. (3.12) and $\gamma_t(\delta, \lambda)$ in Eq. (4.8). For $\delta \in (0, 1)$ and for all $t \geq 1$, with probability at least $1 - \delta$,*

$$\left\| g_t(\hat{\theta}_t) - g_t(\theta_*) \right\|_{H_t^{-1}(\theta_*)} \leq \gamma_t(\delta, \lambda).$$

Proof. For now fix $t \geq 1$. Since $\hat{\theta}_t$ is the maximizer of the original regularized log-likelihood and thus the unique minimizer of the negative log-likelihood because $H_t(\theta) \succ 0$ for all $\theta \in \mathbb{R}^d$:

$$-\mathcal{L}_\lambda(\theta; \mathcal{D}_t) = \sum_{s=1}^{t-1} \psi(x_s^\top \theta) - y_s x_s^\top \theta + \frac{\lambda}{2} \|\theta\|_2^2,$$

and therefore $\nabla_\theta \mathcal{L}_\lambda(\hat{\theta}_t; \mathcal{D}_t) = 0$ where the existence of the gradient is guaranteed by the standard properties of exponential family that $\psi(\cdot)$ is infinitely differentiable (page 38, Brown [Bro86]). Plugging in the definition of $\mathcal{L}_\lambda(\theta; \mathcal{D}_t)$ into the equation $\nabla_{\theta_t} \mathcal{L}_\lambda(\hat{\theta}; \mathcal{D}_t) = 0$ and taking the gradient, we get

$$\nabla_\theta \mathcal{L}_\lambda(\hat{\theta}_t; \mathcal{D}_t) = 0 \tag{4.12}$$

$$\sum_{s=1}^{t-1} \dot{\psi}(x_s^\top \theta) - y_s x_s + \lambda\theta = 0. \tag{4.13}$$

Rearranging the equation, and noting that $\dot{\psi} = \mu$, we obtain

$$g_t(\hat{\theta}_t) = \sum_{s=1}^{t-1} \mu(x_s^\top \hat{\theta}_t) x_s + \lambda \hat{\theta}_t = \sum_{s=1}^{t-1} y_s x_s. \tag{4.14}$$

Now, by the definition of ε_s, $y_s = \mu(x_s^\top \theta_*) + \varepsilon_s$. Hence,

$$\sum_{s=1}^{t-1} y_s x_s = \sum_{s=1}^{t-1} (\mu(x_s^\top \theta_*) + \varepsilon_s) x_s = g_t(\theta_*) - \lambda\theta_* + \sum_{s=1}^{t-1} \varepsilon_s x_s.$$

This expression, combined with Eq. (4.14) gives,

$$g_t(\hat{\theta}_t) - g_t(\theta_*) = \sum_{s=1}^{t-1} \varepsilon_{s+1} x_s - \lambda\theta_* = S_t - \lambda\theta_*,$$

where we let

$$S_t := \sum_{s=1}^{t-1} \varepsilon_s x_s.$$

Take the $\|\cdot\|_{H_t^{-1}(\theta_*)}$ norm of both sides and note that $H_t^{-1}(\theta_*) \preceq \lambda^{-1} I$,

$$\left\| g_t(\hat{\theta}_t) - g_t(\theta_*) \right\|_{H_t^{-1}(\theta_*)} \leq \|S_t\|_{H_t^{-1}(\theta_*)} + \sqrt{\lambda} S, \tag{4.15}$$

where we used that by Assumption 3.6.2 $\|\theta_*\| \leq S$. Note that by definition, $\{\varepsilon_t\}_{t=1}^\infty$ is a subgaussian martingale difference sequence adapted to $\{\sigma(h_t)\}_{t=1}^\infty$. Also, we have that for all $s \geq 1$,

$$\dot{\mu}(x_s^\top \theta_*) = \mathbb{E}[\varepsilon_s^2 | h_{s-1}] := \sigma_s^2.$$

This gives us $H_t(\theta_*) = \sum_{s=1}^{t-1} \sigma_s^2 x_s x_s^\top + \lambda I$. Therefore all the conditions of Theorem 2 have been satisfied and we have that with probability at least

$1-\delta$, for all $t \geq 1$,

$$\|S_t\|_{H_t^{-1}(\theta_*)} \leq \sigma \left(\frac{\sqrt{\lambda}}{2m} + \frac{2m}{\sqrt{\lambda}} \log \left(\frac{\det(H_t(\theta_*))^{1/2} \lambda^{-d/2}}{\delta} \right) + \frac{2m}{\sqrt{\lambda}} d \log(2) \right). \tag{4.16}$$

Now it remains to bound the determinant $\det(H_t(\theta_*))$. Using Lemma 4.4, it holds that

$$\det(H_t(\theta_*)) \leq L^d \det \left(\sum_{s=1}^{t-1} x_s x_s^\top + \frac{\lambda}{L} I \right) \leq L^d \left(\frac{\lambda}{L} + \frac{T}{d} \right)^d \leq \left(\lambda + \frac{Lt}{d} \right)^d.$$

Using the bound we just derived on the determinant of $H_t(\theta_*)$, we can bound Eq. (4.16) by

$$\begin{aligned} \|S_t\|_{H_t^{-1}(\theta_*)} &\leq \sigma \left(\frac{\sqrt{\lambda}}{2m} + \frac{2m}{\sqrt{\lambda}} \log \left(\frac{\left(\lambda + \frac{Lt}{d}\right)^{d/2} \lambda^{-d/2}}{\delta} \right) + \frac{2m}{\sqrt{\lambda}} d \log(2) \right) \\ &= \gamma_t(\delta, \lambda) - \sqrt{\lambda} S. \end{aligned}$$

Finally chaining Eq. (4.15) and Eq. (4.16) with the latter inequality, we get with probability at least $1-\delta$, for all $t \geq 1$,

$$\left\| g_t(\hat{\theta}_t) - g_t(\theta_*) \right\|_{H_t^{-1}(\theta_*)} \leq \gamma_t(\delta, \lambda).$$

□

We denote $\mathcal{E}(\delta)$ to be the event that θ_* lies in the confidence set $C_t(\delta, \lambda)$ with the choice of λ being λ^* (Eq. (4.9)).

Corollary 3. *For $\delta \in (0, 1]$ and for all $t \in [n]$, the event*

$$\mathcal{E}(\delta) = \left\{ \forall t \geq 1, \|g_t(\hat{\theta}_t) - g_t(\theta_*)\|_{H_t^{-1}(\theta_*)} \leq \gamma_n(\delta) \right\}$$

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

On event $\mathcal{E}(\delta)$ we can derive a few useful results where we need the following auxiliary lemma:

Lemma 4.6. *Let $b, c > 0$, and $x \in \mathbb{R}$. Then the following implication holds:*

$$x^2 \leq bx + c \implies x \leq b + \sqrt{c}.$$

Proof. Let $f(x) = x^2 - bx - c$. Then f is a strong convex function with roots $\lambda_1 = 1/2(b + \sqrt{b^2 + 4c})$ and $\lambda_2 = 1/2(b - \sqrt{b^2 + 4c})$. If $x^2 \leq -b - c$ then by convexity

$$x \leq \max(\lambda_1, \lambda_2) \leq 1/2(b + \sqrt{b^2 + 4c}) \leq b + \sqrt{c}.$$

□

Lemma 4.7. *On the good event $\mathcal{E}(\delta)$, it follows that*

$$\|g_t(\theta_*) - g_t(\hat{\theta}_t)\|_{G_t^{-1}(\theta_*, \hat{\theta}_t)} \leq \gamma_n(\delta) + M_\mu \frac{\gamma_n^2(\delta)}{\sqrt{\lambda}}.$$

Proof. From the definition of $G_t(\theta_*, \hat{\theta}_t)$, it follows that

$$\begin{aligned}
G_t(\theta_*, \hat{\theta}_t) &= \sum_{s=1}^{t-1} \alpha(x_s, \theta_*, \hat{\theta}_t) x_s x_s^\top + \lambda I \\
&\geq \sum_{s=1}^{t-1} \Big(1 + M_\mu |x_s^\top (\theta_* - \hat{\theta}_t)|\Big)^{-1} \dot{\mu}(x_s^\top \theta_*) x_s x_s^\top + \lambda I \qquad \text{(Lemma 3.7)} \\
&\geq \sum_{s=1}^{t-1} \Big(1 + M_\mu \|x_s\|_{G_t^{-1}(\theta_*, \hat{\theta}_t)} \|\theta_* - \hat{\theta}_t\|_{G_t(\theta_*, \hat{\theta}_t)}\Big)^{-1} \dot{\mu}(x_s^\top \theta_*) x_s x_s^\top + \lambda I \\
&\qquad \text{(Cauchy-schwarz)} \\
&\geq \Big(1 + M_\mu \lambda^{-1/2} \|\theta_* - \hat{\theta}_t\|_{G_t(\theta_*, \hat{\theta}_t)}\Big)^{-1} \sum_{s=1}^{t-1} \dot{\mu}(x_s^\top \theta_*) x_s x_s^\top + \lambda I_d \\
&\qquad (\|x\|_2 \leq 1 \text{ and } G_t^{-1}(\theta_*, \hat{\theta}_t) \preceq \lambda I) \\
&= \Big(1 + M_\mu \lambda^{-1/2} \|\theta_* - \hat{\theta}_t\|_{G_t(\theta_*, \hat{\theta}_t)}\Big)^{-1} H_t(\theta_*) \qquad \text{(Defn. of } H_t(\theta_*)) \\
&= \Big(1 + M_\mu \lambda^{-1/2} \|g_t(\theta_*) - g_t(\hat{\theta}_t)\|_{G_t^{-1}(\theta_*, \hat{\theta}_t)}\Big)^{-1} H_t(\theta_*). \quad \text{(Lemma 4.3)}
\end{aligned}$$

Using the above inequality, we obtain:

$$\begin{aligned}
\|g_t(\theta_*) - g_t(\hat{\theta}_t)\|_{G_t^{-1}(\theta_*, \hat{\theta}_t)}^2 &\leq \Big(1 + M_\mu \lambda^{-1/2} \|g_t(\theta_*) - g_t(\hat{\theta}_t)\|_{G_t^{-1}(\theta_*, \hat{\theta}_t)}\Big) \|g_t(\theta_*) - g_t(\hat{\theta}_t)\|_{H_t^{-1}(\theta_*)}^2 \\
&\leq M_\mu \lambda^{-1/2} \gamma_n^2(\delta) \|g_t(\theta_*) - g_t(\hat{\theta}_t)\|_{G_t^{-1}(\theta_*, \hat{\theta}_t)} + \gamma_n^2(\delta),
\end{aligned}$$

where the last inequality follows from the definition on $\mathcal{E}(\delta)$. By Lemma 4.6, solving the above quadratic inequality gives:

$$\|g_t(\theta_*) - g_t(\hat{\theta}_t)\|_{G_t^{-1}(\theta_*, \hat{\theta}_t)} \leq \gamma_n(\delta) + M_\mu \frac{\gamma_n^2(\delta)}{\sqrt{\lambda}}.$$

□

Lemma 4.8. *On event $\mathcal{E}(\delta)$, it holds that:*

$$\|\hat{\theta}_t - \theta_*\|_2 \leq 2(S + \frac{1}{2m}\sigma) + 4M_\mu(S + \frac{1}{2m}\sigma)^2.$$

Proof. We link $\hat{\theta}_t - \theta_*$ to $g_t(\hat{\theta}_t) - g_t(\theta_*)$ with Lemma 4.3 and upper bound $G_t^{-1}(\hat{\theta}_t, \theta_*)$:

$$\begin{aligned}
\|\hat{\theta}_t - \theta_*\|_2 &= \|g_t(\hat{\theta}_t) - g_t(\theta_*)\|_{G_t(\hat{\theta}_t,\theta_*)^{-2}} && \text{(Lemma 4.3)}\\
&\leq \frac{1}{\sqrt{\lambda}}\|g_t(\hat{\theta}_t) - g_t(\theta_*)\|_{G_t(\hat{\theta}_t,\theta_*)^{-1}} && (G_t(\hat{\theta}_t, \theta_*)^{-1} \preceq \tfrac{1}{\lambda}I)\\
&\leq \frac{1}{\sqrt{\lambda}}\left(\gamma_n(\delta) + M_\mu\frac{\gamma_n^2(\delta)}{\sqrt{\lambda}}\right).
\end{aligned}$$

Plugging in the value of $\gamma_n(\delta)$ and λ finishes the proof. □

4.2.2 The Good Events

In this section, we display "good" events that will allow us to upper bound the regret. In order to guarantee low regret, we first need to show that the arms our algorithm is selecting, x_t, quickly converges to x_* as $t \to \infty$. Since x_t is defined as the maximizer of $x^\top\theta_t$, it suffices to show that $\theta_t \to \theta_*$ as $t \to \infty$. For this, defining for $\beta > 0$,

$$\mathcal{E}_t(\beta) = \left\{\theta \in \mathbb{R}^d \,:\, \forall x \in \mathcal{X},\, |x^\top(\theta - \hat{\theta}_t)| \leq \beta\|x\|_{H_t(\hat{\theta}_t)^{-1}}\right\},$$

We will want two *good* events to hold, $E_{1,t}$ and $E_{2,t}$, given by

$$E_{1,t} = \{\theta_* \in \mathcal{E}_t(\beta'_{1,t}(\delta))\} \quad \text{and} \quad E_{2,t} = \{\theta_t \in \mathcal{E}_t(\beta'_{2,t}(\delta))\}, \tag{4.17}$$

That is, $\hat{\theta}_t$ is close to both θ_* and θ_t at a given time-step. The choices of the parameters $\beta'_{1,t}(\delta)$ and $\beta'_{2,t}(\delta)$ are detailed here:

$$\beta'_{1,t}(\delta) = (1 + D_t^* M_\mu)\gamma_n(\delta), \tag{4.18}$$

$$\beta'_{2,t}(\delta) = 8(1 + D_t M_\mu)C_d(\delta)\gamma_n(\delta). \tag{4.19}$$

where $\gamma_n(\delta)$ is defined in Eq. (4.10) and $C_d(\delta) = \sqrt{d + \log(n/\delta)}$. D_t^* is defined to be the maximum gap between the original MLE $\hat{\theta}_t$ and the true model parameter θ_* in directions of x_s for $s < t$, that is,

$$D_t^* = \max_{s<t} |x_s^\top(\hat{\theta}_t - \theta_*)|. \tag{4.20}$$

Similarly, D_t is defined to be the maximum gap between the original MLE $\hat{\theta}_t$ and the perturbed MLE θ_t in directions of x_s for $s < t$, that is,

$$D_t = \max_{s<t} |x_s^\top(\hat{\theta}_t - \theta_t)|. \tag{4.21}$$

For randomized algorithm to work, an additional event is needed: that the reward predicted using the optimal arm, x_* and the parameter θ_t is frequently large in comparison to using the true parameter θ_* for the same. That is

$$E_{3,t} = \{x_*^\top \theta_t > x_*^\top \theta_*\}.$$

holds with positive probability. We refer to $E_{3,t}$ as the anti-concentration event.

4.2.3 Analysis of the Warm-Up Procedure

To establish our guarantee on the warm-up procedure, we need $\dot{\mu}(x^\top \hat{\theta}_t) \geq 1/\kappa$, which is showed to hold on event $\mathcal{E}(\delta)$ with appropriately chosen Θ.

Lemma 4.9. *On the event $\mathcal{E}(\delta)$, for all $t \geq 1$ it holds that if*

$$\operatorname{diam}(\Theta) \leq 2(S + \frac{1}{2m}\sigma) + 4M_\mu(S + \frac{1}{2m}\sigma)^2 + S,$$

then $\theta_*, \hat{\theta}_t \in \Theta$.

Proof.

$$\begin{aligned}\|\hat{\theta}_t\|_2 = \|\hat{\theta}_t - \theta_* + \theta_*\|_2 &\leq \|\hat{\theta}_t - \theta_*\|_2 + \|\theta_*\|_2 \\ &\leq 2(S + \frac{1}{2m}\sigma) + 4M_\mu(S + \frac{1}{2m}\sigma)^2 + S,\end{aligned}$$

where the last inequality follows from Lemma 4.8 and the definition of S, see Assumption 3.6.2. Therefore if diam(Θ) satisfies the bound presented in the statement of the lemma, then on event $\mathcal{E}(\delta)$ it holds that $\hat{\theta}_t, \theta_* \in \Theta$. □

Since on event $\mathcal{E}(\delta)$, $\hat{\theta}_t \in \Theta$, it holds that $1/\kappa = \inf_{x\in\mathcal{X},\theta\in\Theta} \dot{\mu}(x^\top\theta) \leq \dot{\mu}(x^\top\hat{\theta}_t)$ for all $x \in \mathcal{X}$. Ensuring that with Θ being large enough, one has θ_* and $\hat{\theta}_t \in \Theta$ for all $t \in [n]$, we are able to establish a guarantee for the warm-up procedure:

Lemma 4.10. *Let ζ be the solution to the G-optimal design problem over a compact feature set $\mathcal{X}$ whose span is $\mathbb{R}^d$. Let $\iota > 0$ be a constant and $\varepsilon = 0.5$. By setting the number of warm-up rounds as $\tau = \lceil 1.5\iota^2 d\kappa \rceil$, Algorithm 7 returns a dataset $\{(x_s, y_s)\}_{s=1}^{\tau}$ such that for all $t > \tau$, on event $\mathcal{E}(\delta)$ the following holds:*

$$\max_x \|x\|_{H_t^{-1}(\hat{\theta}_t)} \leq 1/\iota. \tag{4.22}$$

Proof. For all $t > \tau$, one gets $H_t^{-1}(\hat{\theta}_t) \preceq H_\tau^{-1}(\hat{\theta}_t)$ and it holds that

$$\max_x \|x\|^2_{H_t^{-1}(\hat{\theta}_t)} \leq \max_x \|x\|^2_{H_\tau^{-1}(\hat{\theta}_t)}.$$

Furthermore define $H(\zeta, \theta) := \sum_{x \in \mathcal{X}} \zeta_x \dot{\mu}(x^\top \theta) x x^\top$ where ζ is a probability measure over $\mathcal{X}$ and with slight abuse of notation, we let $\zeta_x := \zeta(x)$. To show the result holds, it is sufficient to prove that for all $t > \tau$, $\max_x \|x\|_{H_\tau^{-1}(\hat{\theta}_t)} \leq 1/\iota$ holds. We have,

$$\begin{aligned}
\max_x \|x\|^2_{H_\tau^{-1}(\hat{\theta}_t)} &\leq \frac{1+\epsilon}{\tau} \max_x \|x\|^2_{H_\tau^{-1}(\zeta, \hat{\theta}_t)} && \text{(Lemma 13 of Jun et al. [Jun+21])} \\
&\leq \frac{\kappa(1+\epsilon)}{\tau} \max_x \|x\|^2_{V^{-1}(\zeta)} && (H(\zeta, \hat{\theta}_t) \succeq \kappa^{-1} V(\zeta) \text{ on } \mathcal{E}(\delta)) \\
&= \frac{d\kappa(1+\epsilon)}{\tau}. && \text{(Kiefer-Wolfowitz, Ch 21 [LS18])}
\end{aligned}$$

The proof is completed by setting $\tau = \iota^2 d\kappa(1+\varepsilon)$. □

The choice of ι in the main algorithm is:

$$\iota > \max\left(\beta'_{1,n}(\delta) M_\mu,\ \beta'_{2,n}(\delta) M_\mu,\ \beta'_{2,n}(\delta)^2 L M_\mu,\ 2\sqrt{\lambda},\ L\right). \tag{4.23}$$

4.2.4 Analysis of the Good Events

To reduce clutter, we define for all $t \in [n]$, $H_t := H_t(\hat{\theta}_t)$ and $G_t := G_t(\hat{\theta}_t, \theta_t) = G_t(\theta_t, \hat{\theta}_t)$ where the last equality is due to the definition of $G_t(\theta_1, \theta_2)$ and the fact that $\alpha(u_1, u_2) = \alpha(u_2, u_1)$(Section 3.10).

Event $E_{1,t}$ We now show that the good event $E_{1,t}$ holds with high probability, whose formal statement is detailed below.

Lemma 4.11. *For $\delta \in (0,1]$, recall D_t^* in Eq. (4.20) and $\beta'_{1,t}(\delta) = (1 + M_\mu D_t^*)\gamma_n(\delta)$ in Eq. (4.18), on the event $\mathcal{E}(\delta)$, we have $E_{1,t}$ holds for all $t \leq n$.*

Proof. Combining Lemma 4.3 with Cauchy-Schwarz gives

$$\begin{aligned}|x^\top(\theta_* - \hat{\theta}_t)| &= |x^\top G_t^{-1/2}(\theta_*, \hat{\theta}_t) G_t^{1/2}(\theta_*, \hat{\theta}_t)(\theta_* - \hat{\theta}_t)| \\ &\leq \|x\|_{G_t^{-1}(\theta_*,\hat{\theta}_t)} \|\theta_* - \hat{\theta}_t\|_{G_t(\theta_*\hat{\theta}_t)} \\ &= \|x\|_{G_t^{-1}(\theta_*,\hat{\theta}_t)} \|g_t(\theta_*) - g_t(\hat{\theta}_t)\|_{G_t^{-1}(\theta_*,\hat{\theta}_t)},\end{aligned}$$

where for the last inequality the mean value theorem was applied to derive the following,

$$\|g_t(\theta_*) - g_t(\hat{\theta}_t)\|_{G_t^{-1}(\theta_*,\hat{\theta}_t)} = \|G_t(\theta_*, \hat{\theta}_t)(\theta_* - \hat{\theta}_t)\|_{G_t^{-1}(\theta_*,\hat{\theta}_t)} = \|\theta_* - \hat{\theta}_t\|_{G_t(\theta_*\hat{\theta}_t)}.$$

Then, by lemma 3.9,

$$\|x\|_{G_t^{-1}(\theta_*,\hat{\theta}_t)} \|g_t(\theta_*) - g_t(\hat{\theta}_t)\|_{G_t^{-1}(\theta_*,\hat{\theta}_t)} \leq (1 + M_\mu D_t^*)\|x\|_{H_t^{-1}(\hat{\theta}_t)} \|g_t(\theta_*) - g_t(\hat{\theta}_t)\|_{H_t^{-1}(\theta_*)}.$$

Since we are on the good event $\mathcal{E}(\delta)$, it holds that $\|g_t(\theta_*) - g_t(\hat{\theta}_t)\|_{H_t^{-1}(\theta_*)} \leq \gamma_n(\delta)$, the statement of the lemma follows. □

On event $\mathcal{E}(\delta)$, D_t^* can be upper bounded by $\text{poly}(S, \sigma)$ for all $t \geq 1$.

Lemma 4.12. *On event $\mathcal{E}(\delta)$, for all $t \geq 1$, it holds that*

$$D_t^* \leq \frac{\gamma_n^2(\delta)}{\lambda} + \frac{\gamma_n(\delta)}{\sqrt{\lambda}} = 2(S + \frac{1}{2m}\sigma) + 4M_\mu(S + \frac{1}{2m}\sigma)^2.$$

Proof. We first use Cauchy schwarz to upper bound D_t^*:

$$\begin{aligned}D_t^* &= \max_{s<t} |x_s^\top(\theta_* - \hat{\theta}_t)| \\ &\leq \max_{s<t} \|x_s\|_{G_t^{-1}(\theta_*,\hat{\theta}_t)} \|\hat{\theta}_t - \theta_*\|_{G_t(\theta_*,\hat{\theta}_t)} \\ &= \max_{s<t} \|x_s\|_{G_t^{-1}(\theta_*,\hat{\theta}_t)} \|g_t(\hat{\theta}_t) - g_t(\theta_*)\|_{G_t^{-1}(\theta_*,\hat{\theta}_t)} && \text{(Lemma 4.3)} \\ &\leq \frac{1}{\sqrt{\lambda}} \left(\gamma_n(\delta) + M_\mu \frac{\gamma_n^2(\delta)}{\sqrt{\lambda}} \right). && \text{(Lemma 4.7)}\end{aligned}$$

The proof is completed by plugging in the values of $\gamma_n(\delta)$ and λ. □

Remark 4.12.1. *This upper bound on D_t^* only depends on $\mathcal{E}(\delta)$. Therefore it can be used in upper bounding ι (Eq. (4.23)), to be more specific, upper bounding $M_\mu \beta'_{1,t}(\delta)$ (Eq. (4.18)), that is, for all $t \in [n]$, on event $\mathcal{E}(\delta)$,*

$$\begin{aligned} \beta'_{1,t}(\delta) M_\mu &\leq \left(1 + 2M_\mu(S + \frac{1}{2m}\sigma) + 4M_\mu^2(S + \frac{1}{2m}\sigma)^2\right) \\ &\leq \left(1 + 2M_\mu(S + \frac{1}{2m}\sigma)\right)^2. \end{aligned} \tag{4.24}$$

Event $E_{2,t}$ Now that we have shown event $E_{1,t}$ holds with high probability, we move on to proving the same result for event $E_{2,t}$. Before we do so, we first need the following convenient way of expressing $\theta_t - \hat{\theta}_t$.

Lemma 4.13 (Lemma 1 of Kveton et al. [Kve+19c])**.** , *We can write $\theta_t - \hat{\theta}_t = G_t^{-1} W_t$ for $W_t := \sum_{s=1}^{t-1} \sqrt{\dot{\mu}(x_s^\top \hat{\theta}_t)} a z_{s,t} x_s$ where $(z_{1,t}, \ldots, z_{t-1,t}) \sim \mathcal{N}(0,1)^{t-1}$.*

Define $D_t := \max_{s<t} |x_s^\top(\theta_t - \hat{\theta}_t)|$ to be the maximum gap between the original MLE and the perturbed MLE in directions of x_s for $s < t$. We prove a similar result to Lemma 4.12 that D_t can be upper bounded for all $t \geq 1$. We will need the following auxiliary lemmas.

Lemma 4.14. *Let $A \in \mathbb{R}^{d\times d}$ be a positive semi-definite matrix and $x \in \mathbb{R}^d$ be a vector then*

$$\|x\|_{A^2}^2 \leq \lambda_{\max}(A)\|x\|_A^2.$$

Proof. For any positive semi-definite matrix

$$x^\top A^2 x = \lambda_{\max}^2(A) x^\top(\lambda_{\max}^{-2}(A)A^2)x \leq \lambda_{\max}^2(A) x^\top(\lambda_{\max}^{-1}(A)A)x = \lambda_{\max}(A)\|x\|_A^2.$$

They key observation is that all the eigenvalues of $\lambda_{\max}^{-2}(A)A^2$ are in $[0,1]$. □

Lemma 4.15 (Example 2.28 of Wainwright [Wai19])**.** *For an integer $k \geq 1$ and a given sequence $\{Z_j\}_{j=1}^k$ of i.i.d standard normal random variables, the random variable $Y := \sum_{j=1}^k Z_j^2$ follows a χ^2-distribution with k degrees of freedom. It holds that for all $\eta \geq 0$,*

$$\mathbb{P}(Y/k \geq (1+\eta)^2) \leq e^{-k\eta^2/2}.$$

Lemma 4.16. *Let $A \in \mathbb{R}^{d\times d}$ be positive semi-definite matrices and let $B = A + cI$ for some non-negative $c \in \mathbb{R}$. Let $x \in \mathbb{R}^d$ be a vector. Then it follows that*

$$\|Bx\|_A \geq \|Ax\|_A.$$

Proof. Since $B = A + cI$, we have that $B \succeq A$. By the condition that A is a p.s.d. matrix and $B = A + cI$, by spectral decomposition, we can write

$$A = V\Lambda_A V^{-1}, B = V\Lambda_B V^{-1}$$

where V is a orthonormal matrix, Λ_A and Λ_B are diagonal matrices because A, B are positive semi-definite matrices. Let $\Lambda_A = \mathrm{diag}(a_1, ...a_d)$ and $\Lambda_B = \mathrm{diag}(b_1, ...b_d)$ where $a_1 \geq a_2 \geq ... \geq a_d \geq 0$ and $b_i = a_i + c$ for all $i \in [d]$. Hence, $b_i - a_i \geq 0$. Take the difference $\|Bx\|_A - \|Ax\|_A$ and this yields

$$\begin{aligned} x^\top BABx - x^\top A^3 x &= x^\top V(\Lambda_B\Lambda_A\Lambda_B - \Lambda_A^3)V^\top x \\ &= x^\top V(\mathrm{diag}(a_1^2 b_1, ..., a_d^2 b_d) - \mathrm{diag}(a_1^3, ..., a_d^3))V^\top x. \end{aligned}$$

Since $a_i \leq b_i$, we have that $a_i^3 \leq a_i^2 b_i$ for all $i \in [d]$. Therefore, the matrix

$$V(\mathrm{diag}(a_1^2 b_1, ..., a_d^2 b_d) - \mathrm{diag}(a_1^3, ..., a_d^3))V^\top$$

is a positive semi-definite matrix and $\|Bx\|_A - \|Ax\|_A \geq 0$ follows by definition of a positive semi-definite matrix. □

We bound D_t by upper bounding $\|\theta_t - \hat{\theta}_t\|_2$.

Lemma 4.17. *Define the event*

$$E_{4,t} = \left\{ \|\theta_t - \hat{\theta}_t\|_2 \leq \frac{4a^2 M_\mu d(1 + \log(n/\delta))}{\lambda} \right\}.$$

Conditioned on the history h_t, for all $t \in [n]$ with probability $1 - \delta$, it holds that $\mathbb{P}_t(E_{4,t}) \geq 1 - \delta$ $\mathbb{P}$-almost surely.

Proof. Fix $t \geq 1$. By Lemma 4.13 it holds that $\|\theta_t - \hat{\theta}_t\|_2 = \|G_t^{-1} W_t\|_2$. To control this norm, notice that

$$\begin{aligned}
\|G_t^{-1} W_t\|_2 &\leq \frac{\|W_t\|_{G_t^{-1}}}{\sqrt{\lambda}} \\
&\leq \frac{\sqrt{1 + M_\mu D_t}\|W_t\|_{H_t^{-1}}}{\sqrt{\lambda}} && \text{(Lemma 3.9)} \\
&= \frac{\sqrt{1 + M_\mu \max_{s<t} x_s^\top(\theta_t - \hat{\theta}_t)}\|W_t\|_{H_t^{-1}}}{\sqrt{\lambda}} \\
&\leq \frac{\sqrt{1 + M_\mu \|\theta_t - \hat{\theta}_t\|}\|W_t\|_{H_t^{-1}}}{\sqrt{\lambda}} \\
& && \text{(Cauchy-Schwarz and Assumption 3.6.3)} \\
&= \frac{\sqrt{1 + M_\mu \|G_t^{-1} W_t\|}\|W_t\|_{H_t^{-1}}}{\sqrt{\lambda}} \\
& && \text{(Cauchy-Schwarz and Lemma 4.13)} \\
&\leq \frac{\|W_t\|_{H_t^{-1}}}{\sqrt{\lambda}} + \frac{\sqrt{M_\mu \|G_t^{-1} W_t\|}\|W_t\|_{H_t^{-1}}}{\sqrt{\lambda}}. && (\sqrt{a+b} \leq \sqrt{a} + \sqrt{b})
\end{aligned}$$

By Lemma 4.6, $x^2 \leq bx + c$ implies $x \leq b + \sqrt{c}$. Letting $x = \sqrt{\|G_t^{-1} W_t\|_2}$ gives

$$\|G_t^{-1} W_t\|_2 \leq \left(\sqrt{\frac{\|W_t\|_{H_t^{-1}}}{\sqrt{\lambda}}} + \frac{\sqrt{M_\mu}\|W_t\|_{H_t^{-1}}}{\sqrt{\lambda}} \right)^2.$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\|G_t^{-1} W_t\|_2 \leq \frac{2\|W_t\|_{H_t^{-1}}}{\sqrt{\lambda}} + \frac{2M_\mu \|W_t\|_{H_t^{-1}}^2}{\lambda}.$$

All that remains is to bound $\|W_t\|_{H_t^{-1}}$. Given the history, by definition W_t is a zero-mean normal random vector. The covariance of W_t is

$$\mathbb{E}_t\left[W_t W_t^\top\right] = \mathbb{E}_t\left(\sum_{s=1}^{t-1} a^2 \dot{\mu}(x_s^\top \hat{\theta}_t) x_s x_s^\top Z_{s,t}^2 \right) = a^2(H_t - \lambda I).$$

where $(z_{1,t}, \ldots, z_{t-1,t}) \sim \mathcal{N}(0,1)$ are independent. Define $Y = a^{-1}(H_t - \lambda I)^{-1/2} W_t$. This gives

$$\begin{aligned}
\|W_t\|_{H_t^{-1}} &= \|a^2(H_t - \lambda I)^{1/2} Y\|_{H_t^{-1}} \\
&= \sqrt{a^2 Y^\top (H_t - \lambda I)^{1/2} H_t^{-1} (H_t - \lambda I)^{1/2} Y} \\
&\leq a\|Y\|_2,
\end{aligned}$$

where the inequality holds since $H_t \leq LV_t$ and Lemma 4.16. Since $Y \sim \mathcal{N}(0, I)$, given the history, we have that $\|Y\|_2^2$ is χ^2 with d-degrees of freedom. Adapting Lemma 4.15 gives that with probability at least $1 - \delta$, given the history, it holds that $\|Y\|_2 \leq \sqrt{d(1 + \log(n/\delta))}$. Combining these bounds

$$\begin{aligned}\|G_t^{-1} W_t\|_2 &\leq \frac{2a\sqrt{d(1+\log(n/\delta))}}{\sqrt{\lambda}} + \frac{2a^2 M_\mu d(1+\log(n/\delta))}{\lambda} \\ &\leq \frac{4a^2 M_\mu d(1+\log(n/\delta))}{\lambda},\end{aligned}$$

which completes the proof. □

Remark 4.17.1. *On event $E_{4,t}$, by making use of Cauchy-schwarz inequality and Assumption 3.6.3, we can upper bound D_t in the following way:*

$$D_t \leq \max_{s<t} \|x_s\|_2 \|\theta_t - \hat{\theta}_t\|_2 \leq \frac{4a^2 M_\mu d(1+\log(n/\delta))}{.}\lambda \tag{4.25}$$

This bound only depends on event $E_{4,t}$. Therefore it can be used to upper bound ι (Eq. (4.23)), to be more specific, $\beta'_{2,t}(\delta)$ (Eq. (4.19)), that is, for all $t \in [n]$, on event $E_{4,t}$,

$$\beta'_{2,t}(\delta) \leq 8\left(1 + \frac{4a^2 M_\mu^2 d(1+\log(n/\delta))}{\lambda}\right) C_d(\delta)\gamma_n(\delta). \tag{4.26}$$

With Remarks 4.12.1 and 4.17.1, we can now give the exact value of ι.

Lemma 4.18. *On event $\mathcal{E}(\delta) \cap E_{1,t}$, by setting ι to be:*

$$\begin{array}{l}\max\{ \\ \quad \left(1 + 2M_\mu(S + \frac{1}{2m}\sigma)\right)^2, \\ \quad 8M_\mu\left(1 + \frac{4a^2 M_\mu^2 d(1+\log(n/\delta))}{\lambda}\right) C_d(\delta)\gamma_n(\delta), \\ \quad 64LM_\mu^2\left(1 + \frac{4a^2 M_\mu^2 d(1+\log(n/\delta))}{\lambda}\right)^2 C_d(\delta)^2\gamma_n(\delta)^2, \\ \quad 2\sqrt{\lambda}, \\ \quad L \\ \quad \}, \end{array}$$

It follows that for all $t > \tau$,

$$\sup_{x\in\mathcal{X}} \|x\|_{H_t^{-1}(\hat{\theta}_t)} \leq \min\left\{\frac{1}{\beta'_{2,t}(\delta)^2 M_\mu L}, \frac{1}{M_\mu \beta'_{1,t}(\delta)}, \frac{1}{M_\mu \beta'_{2,t}(\delta)}, \frac{1}{L}, \frac{1}{2\sqrt{\lambda}}\right\}.$$

Proof. By plugging the upper bounds on D_t and D_t^* into the choice of ι in Eq. (4.23) and applying Lemma 4.10, the proof is complete. □

We now begin the analysis of event $E_{2,t}$. By Lemma 4.13, we have that $x^\top(\theta_t - \hat{\theta}_t) = x^\top G_t^{-1} W_t$. One of the challenges in analyzing this quantity is that G_t, whose definition includes operations on θ_t, correlates with W_t because θ_t is the MLE of data perturbed by $z_{s,t}$. To deal with this issue, we start by characterizing the random variable $x^\top H_t^{-1} W_t$. Relating this random variable to the random variable $x^\top(\theta_t - \hat{\theta}_t)$ will be at the crux of our subsequent arguments in this section.

Lemma 4.19. *Let $a = \sqrt{2}\gamma_n(\delta)$. Fix $x \in \mathbb{R}^d$. Then conditioned on the history h_t, $x^\top H_t^{-1} W_t$ is zero-mean Gaussian with variance $\sigma_t^2(x)$. For all $t \geq 1$,*

$$\sigma_t^2(x) \leq 4\gamma_n^2(\delta)\|x\|^2_{H_t^{-1}},$$

and for all $t > \tau$, on the events $\mathcal{E}(\delta)$ and $E_{4,t}$,

$$\sigma_t^2(x) \geq \gamma_n^2(\delta)\|x\|^2_{H_t^{-1}}.$$

Proof. $x^\top H_t^{-1} W_t$ is a linear combination of the sum of zero-mean Gaussian random variables, and thus Gaussian with zero-mean. For the variance, $\sigma_t^2(x)$,

$$\begin{aligned}\mathbb{E}_t\left(x^\top H_t^{-1}\sum_{s=1}^{t-1} x_s\sqrt{\dot{\mu}(x_s^\top\hat{\theta}_t)}a z_{s,t}\right)^2 &= \sum_{s=1}^{t-1}\left(\sqrt{\dot{\mu}(x_s^\top\hat{\theta}_t)}a x^\top H_t^{-1} x_s\right)^2 \\ &= x^\top H_t^{-1}\left(\sum_{s=1}^{t-1}\dot{\mu}(x_s^\top\hat{\theta}_t)a^2 x_s x_s^\top\right)H_t^{-1}x,\end{aligned}$$

where $z_{1,t}, \ldots, z_{t-1,t} \sim \mathcal{N}(0,1)^{t-1}$. Consider the sum in the middle. Expanding the definition of $\dot{\mu}(x_s^\top\hat{\theta}_t)a^2$, this is equal to

$$M := \sum_{s=1}^{t-1}\dot{\mu}(x_s^\top\hat{\theta}_t)a^2 x_s x_s^\top = \sum_{s=1}^{t-1}\left(\sqrt{2\dot{\mu}(x_s^\top\hat{\theta}_t)}\gamma_n(\delta)\right)^2 x_s x_s^\top.$$

We have $M = 2\gamma_n^2(\delta)(H_t - \lambda I)$. Now all that is left is to bound $\sigma_t^2(x)$ from above and below. First

$$\sigma_t^2(x) = 2\gamma_n^2(\delta)x^\top H_t^{-1}(H_t - \lambda I)H_t^{-1}x \leq 4\gamma_n^2(\delta)\|x\|_{H_t^{-1}}.$$

This proves the first half of the lemma statement. Next we have that

$$\sigma_t^2(x) = 2\gamma_n^2(\delta)x^\top H_t^{-1}(H_t - \lambda I)H_t^{-1}x \geq 2\gamma_n^2(\delta)\|x\|_{H_t^{-1}}\left(1 - \sqrt{\lambda}\|x\|_{H_t^{-1}}\right).$$

where Lemma 4.14 was used for the last inequality. Since we are in rounds $t > \tau$ and on the events $\mathcal{E}(\delta) \cap E_{4,t}$ that ensures a bound on ι (Remarks 4.12.1 and 4.17.1), it holds that $\|x\|_{H_t^{-1}} \leq 1/(2\sqrt{\lambda})$, by choice of ι and Lemma 4.10. Substituting these bounds back into the definition of $\sigma_t^2(x)$ yields the claimed result. □

The proof of the next result relies on a standard covering argument (see, e.g. Lattimore & Szepesvári [LS18],Vershynin [Ver18]), where we will use Lemma 4.19 to control maxima on the cover elements.

We define an event $E'_{2,t}$ such that $E'_{2,t} \subset E_{2,t}$, which is as follows:

Lemma 4.20. *Let $C_d(\delta) = \sqrt{d + \log(n/\delta)}$. Then, for $s \in [t]$ define*

$$E'_{2,s} = \left\{\|G_s^{-1}W_s\|_{H_s} \leq 8(1 + M_\mu D_s)C_d(\delta)\gamma_n(\delta)\right\}.$$

Then $\mathbb{P}_t(\cap_{s=1}^t E'_{2,s}) \geq 1 - \delta$ holds $\mathbb{P}$-almost surely.

Proof. Fix s as is stated in the claim. By Lemma 3.9, $\|G_s^{-1}W_s\|_{H_s} \leq (1 + M_\mu D_s)\|W_s\|_{H_s^{-1}}$. We now bound $\|W_s\|_{H_s^{-1}}$ using a covering argument.

Let $z_1, \ldots, z_m$ be an ϵ-cover of $B_2(d)$; for any $z \in B_2(d)$, write z' for any cover element with $\|z' - z\|_2 \leq \epsilon$. Then,

$$\begin{aligned}\|H_s^{-\frac{1}{2}}W_s\|_2 &= \max_{z\in B_2(d)} z^\top H_s^{-\frac{1}{2}}W_s \\ &= \max_{z\in B_2(d)} (z - z')^\top H_s^{-\frac{1}{2}}W_s + z'H_s^{-\frac{1}{2}}W_s \\ &\leq \epsilon\|H_s^{-\frac{1}{2}}W_s\|_2 + \max_{i\leq m} z_i^\top H_s^{-\frac{1}{2}}W_s.\end{aligned}$$

Choosing $\epsilon = \frac{1}{2}$, rearranging, and letting $x_i = z_i^\top H_s^{1/2}$ for each $i \leq m$, we have

$$\|W_s\|_{H_s^{-1}} \leq 2\max_{i\leq m} x_i^\top H_s^{-1}W_s. \tag{4.27}$$

By Lemma 4.19, given the past, $x_iH_s^{-1}W_s$ is zero-mean Gaussian with variance $\sigma_s^2(x_i)$ satisfying

$$\sigma^2(x_i) \leq 4\gamma_n^2(\delta)\|x_i\|_{H_s^{-1}} \leq 4\gamma_s^2(\delta)\|z_i\| \leq 4\gamma_n(\delta). \tag{4.28}$$

Thus, for any $c_{s,i}(\delta) > 0$,

$$\mathbb{P}_s\left(\max_{i\le m} x_i^\top H_s^{-1} W_s > c_{s,i}(\delta)\right) \le \sum_{i=1}^m \mathbb{P}_s\left(x_i^\top H_s^{-1} W_s > c_{s,i}(\delta)\right)$$
$$\le \sum_{i=1}^m \exp\left(\frac{-c_{s,i}^2(\delta)}{2\sigma_s^2(x_i)}\right),$$

where in the last inequality, note that a zero-mean Gaussian random variable with variance $c_{s,i}(\delta)$ is also a $c_{s,i}(\delta)$-subgaussian random variable, therefore we apply Eq. (3.10). Choosing $c_{s,i}(\delta) = 8\gamma_n^2(\delta)\log(nm/\delta)$ yields

$$\max_{i\le m} x_i^\top H_s^{-1} W_s \le \gamma_n(\delta)\sqrt{8\log(nm/\delta)},$$

with probability at least $1-\delta$. The result follows by noting that $m = 6^d$ suffices (by, for example, Corollary 4.2.13 in Vershynin [Ver18]) and taking the union bound over t. □

Given the above result, analysing $E_{2,t}$ becomes trivial:

Lemma 4.21. *Fix any $t \in [n]$ and $\delta \in (0,1]$. Recall in Eq.* (4.19),

$$\beta'_{2,s}(\delta) = 8(1 + M_\mu D_s)C_d(\delta)\gamma_n(\delta),$$

Then $\mathbb{P}_t(\cap_{s=1}^t E_{2,s}) \ge 1-\delta$ holds $\mathbb{P}$-almost surely.

Proof. Using Lemma 4.13 and Cauchy-Schwarz, for all $\tau < s \le t$

$$|x^\top(\theta_s - \hat{\theta}_s)| = |x^\top G_s^{-1} W_s| \le \|x\|_{H_s^{-1}} \|G_s^{-1} W_s\|_{H_s}.$$

Lemma 4.20 then gives the result. □

Event $E_{3,t}$ Proving anti-concentration, the original difficulty in Kveton et al. [Kve+19c], turns out to be simple using the self-concordance properties of sGLBs. We find that after the initial exploration, under certain events that hold with high probability, the anti-concentration event $E_{3,t}$ holds with at least constant probability.

Lemma 4.22. *Fix $t > \tau$. On event $E_{1,t} \cap E_{2,t} \cap E_{4,t} \cap \mathcal{E}(\delta)$, it holds $\mathbb{P}_t(E_{3,t}) \ge 1 - \Phi(3)$ $\mathbb{P}$-almost surely where Φ is the distribution function of standard Gaussian distribution.*

Proof. Recalling that $x_t = \arg\max_{x \in \mathcal{X}} x^\top \theta_t$ it follows

$$\begin{aligned} \mathbb{P}_t\left(E_{3,t}\right) &= \mathbb{P}_t\left(x_t^\top \theta_t > x_t^\top \theta_*\right) \geq \mathbb{P}_t\left(x_*^\top \theta_t > x_*^\top \theta_*\right) \\ &= \mathbb{P}_t\left(x_*^\top(\theta_t - \hat{\theta}_t + \hat{\theta}_t - \theta_*) > 0\right). \end{aligned}$$

Since we are on event $E_{1,t}$ it holds that $|x_*^\top(\theta_* - \hat{\theta}_t)| \leq \beta'_{1,t}(\delta)\|x_*\|_{H_t^{-1}}$. This yields

$$\mathbb{P}_t\left(E_{3,t}\right) \geq \mathbb{P}_t\left(x_*^\top(\theta_t - \hat{\theta}_t) \geq \beta'_{1,t}(\delta)\|x_*\|_{H_t^{-1}}\right).$$

Using Lemma 4.13, $x_*^\top(\theta_t - \hat{\theta}_t) = x_*^\top G_t^{-1} W_t$. Using a standard identity for $(A+B)^{-1}$ and defining $U_t = G_t - H_t$,

$$G_t^{-1} = (H_t + U_t)^{-1} = H_t^{-1} - H_t^{-1} U_t (H_t + U_t)^{-1}.$$

This gives

$$\begin{aligned} &\mathbb{P}_t\left(x_*^\top(\theta_t - \hat{\theta}_t) \geq \beta'_{1,t}(\delta)\|x_*\|_{H_t^{-1}}\right) \\ =&\mathbb{P}_t\left(x_*^\top H_t^{-1} W_t - x_*^\top H_t^{-1} U_t (H_t + U_t)^{-1} W_t \geq \beta'_{1,t}(\delta)\|x_*\|_{H_t^{-1}}\right) \\ =&\mathbb{P}_t\left(x_*^\top H_t^{-1} W_t \geq x_*^\top H_t^{-1} U_t (H_t + U_t)^{-1} W_t + \beta'_{1,t}(\delta)\|x_*\|_{H_t^{-1}}\right). \end{aligned}$$

For the first term on the right hand side,

$$\begin{aligned} |x_*^\top H_t^{-1} U_t (H_t + U_t)^{-1} W_t| &= |x_* H_t^{-1} U_t G_t^{-1} W_t| \\ &\leq \|x_*\|_{H_t^{-1}} \|H_t^{-1/2} U_t H_t^{-1/2}\|_2 \|G_t^{-1} W_t\|_{H_t}. \end{aligned}$$

and thus we need to bound the two right hand side norms. To bound $\|H_t^{-1/2} U_t H_t^{-1/2}\|_2$, by Lemma 3.7 we have,

$$\begin{aligned} \max_{s \in [t]} |\alpha(x_s^\top \theta_t, x_s^\top \hat{\theta}_t) - \dot{\mu}(x_s^\top \hat{\theta}_t)| &\leq \frac{e^{M_\mu D_t} - 1}{M_\mu D_t} \dot{\mu}(x_s^\top \hat{\theta}_t) - \dot{\mu}(x_s^\top \hat{\theta}_t) \\ &= \underbrace{\left(\frac{e^{M_\mu D_t} - 1}{M_\mu D_t} - 1\right)}_{:= Q_t} \dot{\mu}(x_s^\top \hat{\theta}_t). \end{aligned}$$

Now observe that

$$\|A\|_2 \coloneqq \max_{x : \|x\| \leq 1} \max\{x^\top A x, x^\top(-A)x\},$$

and therefore for some x with $\|x\|_2 \leq 1$,$\|H_t^{-1/2} U_t H_t^{-1/2}\|_2$ is bounded by

$$\begin{aligned}
&x^\top H_t^{-1/2} \left(\sum_{s=1}^{t-1} |\alpha(x_s^\top \theta_t, x_s^\top \hat{\theta}_t) - \dot{\mu}(x_s^\top \hat{\theta}_t)| x_s x_s^\top \right) H_t^{-1/2} x \\
&\leq Q_t x^\top H_t^{-1/2} (H_t - \lambda I) H_t^{-1/2} x \leq Q_t.
\end{aligned}$$

By Lemma 4.20, $\|G_t^{-1} W_t\|_{H_t} \leq \beta'_{2,t}(\delta)$. Combining the bounds

$$|x_*^\top H_t^{-1} U_t (H_t + U_t)^{-1} W_t| \leq Q_t \beta'_{2,t}(\delta) \|x_*\|_{H_t^{-1}}.$$

Defining $b_t(\delta) = \beta'_{1,t}(\delta) + Q_t \beta'_{2,t}(\delta)$ gives

$$\mathbb{P}_t \left(x_*^\top (\theta_t - \hat{\theta}_t) \geq \beta'_{1,t}(\delta) \|x_*\|_{H_t^{-1}} \right) \geq \mathbb{P}_t \left(x_*^\top H_t^{-1} W_t \geq b_t(\delta) \|x_*\|_{H_t^{-1}} \right).$$

Now since we are on events $E_{1,t}$ and $E_{2,t}$ and in round $t > \tau$, we have that

$$\begin{aligned}
D_t^* &= \max_{s<t} |x_s^\top (\hat{\theta}_t - \theta_*)| \\
&\leq \sup_{x \in \mathcal{X}} |x^\top (\hat{\theta}_t - \theta_*)| \\
&\leq \beta'_{1,t}(\delta) \sup_{x \in \mathcal{X}} \|x\|_{H_t^{-1}} && (E_{1,t}) \\
&\leq \frac{\beta'_{1,t}(\delta)}{\iota}, && \text{(Warm-up, Lemma 4.10)}
\end{aligned}$$

and similarly,

$$D_t \leq \frac{\beta'_{2,t}(\delta)}{\iota}.$$

We select ι as in Lemma 4.18:

1. $D_t^* \leq \frac{\beta'_{1,t}(\delta)}{\iota} \leq M_\mu^{-1}$ then as a consequence $\beta'_{1,t}(\delta) = (1 + M_\mu D_t^*) \gamma_n(\delta) \leq 2\gamma_t(\delta)$.

2. $D_t \leq \frac{\beta'_{2,t}(\delta)}{\iota} \leq M_\mu^{-1}$ then as a consequence $Q_t \leq M_\mu D_t \leq \frac{\beta'_{2,t}(\delta) M_\mu}{\beta'_{2,t}(\delta)^2 M_\mu L}$.

At first sight it seems that we bound D_t^* with $\beta'_{1,t}(\delta)$ which in turn contains a factor of D_t^*. Note that there is no looping argument here because Lemma 4.12 and Eq. (4.25) gives an upper bound on D_t^* and D_t that holds regardlessly as long as the event $\mathcal{E}(\delta) \cap E_{4,t}$, holds. Upper bound $\beta'_{1,t}(\delta)$ and $\beta'_{2,t}(\delta)$ allows us to have a tighter bound on D_t and D_t^* by carefully setting ι. These statements yield

$$b_t(\delta) \leq 2\gamma_t(\delta) + \frac{M_\mu \beta'_{2,t}(\delta)^2}{\beta'_{2,t}(\delta)^2 M_\mu L} \leq 2\gamma_t(\delta) + 1,$$

where the last inequality holds by choice of ι. By Lemma 4.19, given the past, $x_*^\top H_t^{-1} W_t$ is a zero-mean Normal random variable whose variance is lower bounded by $\gamma_t^2(\delta)\|x_*\|^2_{H_t^{-1}}$. Dividing by the square root of the variance and using the bound on $b_t(\delta)$

$$\mathbb{P}_t\left(x_*^\top H_t^{-1} W_t \geq b_t(\delta)\|x_*\|_{H_t^{-1}}\right) \geq \mathbb{P}_t\left(Z > 2 + 1/\gamma_t(\delta)\right) \geq \mathbb{P}_t\left(Z > 2+1\right) > 0.0015,$$

where $Z \sim \mathcal{N}(0,1)$ and the second to last inequality used that $\gamma_t(\delta) \geq 1$. □

This lemma combined with earlier results that show that the events $E_{1,t}$ and $E_{2,t}$ occur with high probability imply that sGLM-PHE enjoys sublinear regret.

Remark 4.22.1. *While $E_{1,t}$ and $E_{2,t}$ are defined with respect to* linear *gaps $|x^\top(\theta_1 - \theta_2)|$, the upper bound on these, $\|x\|_{H_t^{-1}(\hat{\theta}_t)}$, contains curvature information about the link function $\mu(\cdot)$ through the norm induced by $H_t^{-1}(\hat{\theta}_t)$. Constructing our events with respect to this curvature information is what allows us to guarantee anti-concentration and obtain better dependencies on κ and $\dot{\mu}(x_*^\top\theta_*)$ in our regret bounds. In our analysis, we use self-concordance (via Lemma 4.23) to convert bounds on the linear gaps to those on $|\mu(x^\top\theta_1) - \mu(x^\top\theta_2)|$, the key quantity for controlling regret.*

4.2.5 Analysis of the regret bound of Algorithm 6

In previous sections, we only gave an informal version of the regret bound of Algorithm 6. In this section, we detail the formal version as well as the analysis.

Theorem 3. *Under assumptions 3.6.1-3.6.3 and assumption 3.10.1. Let $m = M_\mu/\log(2)$, $\delta \in (0,1)$ and $\beta'_n(\delta) = \beta'_{1,n}(\delta) + \beta'_{2,n}(\delta)$. With $\gamma_n(\delta)$ chosen in Eq.* (4.10)*, λ chosen in Eq.* (4.9)*, $\beta'_{1,t}(\delta)$ and $\beta'_{2,t}(\delta)$ chosen in Eq.* (4.18) *and Eq.* (4.19)*, a chosen in Lemma 4.19, ι chosen in Lemma 4.18 and τ chosen* *in Lemma 4.10, which we detail all of them below:*

$$\begin{aligned}
m &= M_\mu/\log(2),\\
\gamma_n(\delta) &= \sqrt{8m\sigma d\left(S+\frac{1}{2m}\sigma\right)\log\left(\frac{2}{\delta}\sqrt{1+Ln/d}\right)},\\
\lambda &= \frac{2m\sigma}{S+\frac{1}{2m}\sigma}d\log\left(\frac{2}{\delta}\sqrt{1+Ln/d}\right),\\
a^2 &= 2\gamma_n(\delta),\\
C_d(\delta) &= \sqrt{d+\log(n/\delta)},\\
\beta'_{1,n}(\delta) &= (1+M_\mu)\gamma_n(\delta),\\
\beta'_{2,n}(\delta) &= 8(1+M_\mu)C_d(\delta)\gamma_n(\delta),\\
\beta'_n(\delta) &= \beta'_{1,n}(\delta)+\beta'_{2,n}(\delta),\\
\iota &= 64LM_\mu^2\left(1+\frac{4a^2M_\mu^2 d(1+\log(n/\delta))}{\lambda}\right)^2 C_d(\delta)^2\gamma_n(\delta)^2\\
&\quad+\left(1+2M_\mu(S+\frac{1}{2m}\sigma)\right)^2+L,\\
\tau &= \iota^2 d\kappa = \tilde{\mathcal{O}}(d^9L^2\kappa),
\end{aligned}$$

the regret of Algorithm 6 can be upper bounded by

$$\begin{aligned}
R(n) \le{}& \frac{32e^3\beta'_n(\delta)^2}{p_3^2}\left(2M_\mu d\log(1+nL/(d\lambda))\right)\\
&+2\tau\Delta_{\max}+\frac{8e^{3/2}\beta'_n(\delta)}{p_3}\left(\sqrt{2d\log(1+nL/(d\lambda))}\sqrt{\dot\mu(x_*^\top\theta_*)n}+\sqrt{8n\log\frac{2}{\delta'}}\right)\\
&+2\Delta_{\max},
\end{aligned}$$

with probability at least $1-5\delta$.

In big-O notation, it can be written as

$$R(n)\le\tilde{\mathcal{O}}\left(\sigma d^{3/2}\sqrt{S\dot\mu(x_*^\top\theta_*)n}+\tau\right).$$

We now present the proof of the regret bound. In the definitions of $E_{1,t}$, $E_{2,t}$ and $E_{3,t}$, we only characterize the constraints on the linear gaps: $|x^\top(\hat\theta_t-\theta_*)|$ and $|x^\top(\hat\theta_t-\theta_t)|$ for $x\in\mathcal{X}$. The regret, however, is defined with respect to the non-linear gap: $\mu(x_*^\top\theta_*)-\mu(x_t^\top\theta_*)$. We therefore present a lemma that allows us to link the non-linear gap to the linear gap:

Lemma 4.23. *If for all $x, \theta_1, \theta_2 \in \mathbb{R}^d$ and $\delta \in (0,1]$, $|x^\top(\theta_1 - \theta_2)| \leq \beta'(\delta)\|x\|_{H_t^{-1}}$ holds and $\beta'(\delta)\|x\|_{H_t^{-1}} \leq M_\mu^{-1}$ for some $\beta'(\delta)$, then it follows that*

$$|(\mu(x^\top\theta_1) - \mu(x^\top\theta_2)| \leq \beta_t(\delta)\|x\|_{H_t^{-1}},$$

where

$$\beta_t(\delta) = \dot{\mu}(x^\top\theta_1)\beta'(\delta)\left(1 + eM_\mu\beta'(\delta)\|x\|_{H_t^{-1}}\right).$$

Proof. By Taylor's theorem and the property that $|\ddot{\mu}(\cdot)| \leq M_\mu\dot{\mu}(\cdot)$, we have that

$$\begin{aligned}|(\mu(x^\top\theta_1) - \mu(x^\top\theta_2)| &= \dot{\mu}(x^\top\theta_1)|x^\top(\theta_1 - \theta_2)| + \ddot{\mu}(\xi)(x^\top\theta_1 - x^\top\theta_2)^2 \\ &\leq \dot{\mu}(x^\top\theta_1)|x^\top(\theta_1 - \theta_2)| + M_\mu\dot{\mu}(\xi)(x^\top\theta_1 - x^\top\theta_2)^2,\end{aligned}$$

where $\xi \in \mathbb{R}$ is between $x^\top\theta_1$ and $x^\top\theta_2$. Now assume that $|x^\top(\theta_1 - \theta_2)| \leq \beta'(\delta)\|x\|_{H_t^{-1}}$. Then we have the following

$$\begin{aligned}&\dot{\mu}(x^\top\theta_1)|x^\top(\theta_1 - \theta_2)| + M_\mu\dot{\mu}(\xi)(x^\top\theta_1 - x^\top\theta_2)^2 \\ \leq&\dot{\mu}(x^\top\theta_1)\beta'(\delta)\|x\|_{H_t^{-1}} + M_\mu\dot{\mu}(\xi)\beta'(\delta)^2\|x\|^2_{H_t^{-1}}.\end{aligned}$$

Further by hypothesis we have $|x^\top(\theta_1 - \theta_2)| \leq \beta'(\delta)\|x\|_{H_t^{-1}} \leq 1/M_\mu$. Then from Lemma 3.8, we can upper bound $\dot{\mu}(\xi)$ by

$$\dot{\mu}(\xi) \leq \exp(M_\mu|\xi - x^\top\theta_1|)\dot{\mu}(x^\top\theta_1) \leq \exp(M_\mu|x^\top(\theta_1 - \theta_2)|)\dot{\mu}(x^\top\theta_1) \leq e\dot{\mu}(x^\top\theta_1).$$

Putting these results together it holds that

$$\begin{aligned}|(\mu(x^\top\theta_1) - \mu(x^\top\theta_2)| &\leq \dot{\mu}(x^\top\theta_1)\beta'(\delta)\|x\|_{H_t^{-1}} + e\dot{\mu}(x^\top\theta_1)M_\mu\beta'(\delta)^2\|x\|^2_{H_t^{-1}}, \\ &= \dot{\mu}(x^\top\theta_1)\beta'(\delta)\left(1 + eM_\mu\beta'(\delta)\|x\|_{H_t^{-1}}\right)\|x\|_{H_t^{-1}}.\end{aligned}$$

□

Note we chose $\iota > \beta'_{2,t}(\delta)^2 M_\mu L$ (Lemma 4.18). We also have that

$$D_t^* \leq \beta'_{1,t}(\delta)\|x\|_{H_t^{-1}(\hat{\theta}_t)} \leq \frac{\beta'_{1,t}(\delta)}{\beta'_{2,t}(\delta)^2 M_\mu L} \quad \text{on event } E_{1,t} \cap E_{4,t} \tag{4.29}$$

$$D_t \leq \beta'_{2,t}(\delta)\|x\|_{H_t^{-1}(\hat{\theta}_t)} \leq \frac{\beta'_{2,t}(\delta)}{\beta'_{2,t}(\delta)^2 M_\mu L} \quad \text{on event } E_{2,t} \cap E_{4,t}. \tag{4.30}$$

Denote $E_t = \mathcal{E}(\delta) \cap E_{1,t} \cap E_{2,t} \cap E_{4,t}$ and $\beta'_t(\delta) = \beta'_{1,t}(\delta) + \beta'_{2,t}(\delta)$. We then decompose the regret as follows:

$$R(n) = \sum_{t=1}^{n} \mu(x_*^\top \theta_*) - \mu(x_t^\top \theta_*) \le \tau\Delta_{\max} + \sum_{t=\tau+1}^{n} \mu(x_*^\top \theta_*) - \mu(x_t^\top \theta_*) \tag{4.31}$$

$$\le (\tau+1)\Delta_{\max} + \sum_{t=\tau+1}^{n} \underbrace{\mu(x_*^\top \theta_*) - \mu(x_t^\top \theta_t)}_{R_t^{\text{PHE}}} + \underbrace{\mu(x_t^\top \theta_t) - \mu(x_t^\top \theta_*)}_{R_t^{\text{PRED}}}. \tag{4.32}$$

We now bound R_t^{PHE} and R_t^{PRED} separately.

Bounding R_t^{PHE} We will need the following auxiliary lemmas from Abeille et.al [AL17].

Lemma 4.24 (Proposition 3 of Abeille et al. [AL17])**.** *For any set of arm $\mathcal{X}$ satisfying Assumption 3.3.2, $J(\theta) = \sup_x x^\top \theta$ has the following properties:*

1. *J is a real-valued as the supremum is attained in $\mathcal{X}$,*

2. *J is convex on $\mathbb{R}^d$,*

3. *J is continuous with continuous first derivative except for a zero-measure set w.r.t the Lebesgue measure*

Lemma 4.25 (Lemma 2 of Abeille et al. [AL17])**.** *Under Assumption 3.3.2, for any $\theta \in \mathbb{R}^d$, we have $\nabla J(\theta) = \arg\max_{x\in\mathcal{X}} \theta$ except for a zero-measure set w.r.t the Lebesgue measure.*

We now start to bound R_t^{PHE}.

Lemma 4.26. *On event E_t, for rounds $t > \tau$ it holds that*

$$R_t^{PHE} \le 4\dot{\mu}(x_t^\top \theta_t)\beta'_{2,t}(\delta)/p_3 \left(\|x_t\|_{H_t^{-1}} + \mathbb{E}[\|x_t\|_{H_t^{-1}} | h_t] - \|x_t\|_{H_t^{-1}} \right).$$

Proof. From the definition of R_t^{PHE} it follows that

$$R_t^{\text{PHE}} = \mu(x_*^\top \theta_*) - \mu(x_t^\top \theta_t) = \alpha(x_*^\top \theta_*, x_t^\top \theta_t)(x_*^\top \theta_* - x_t^\top \theta_t),$$

where the second equality follows from an exact first order Taylor's expansion. Now define $J(\theta) = \max_{x\in\mathcal{X}} x^\top \theta$. Therefore it holds that $R_t^{\text{PHE}} =$

$\alpha(J(\theta_*), J(\theta_t))(J(\theta_*) - J(\theta_t))$. By Lemma 4.24 we have that J is a convex function and from Lemma 4.25 we have that $\nabla J(\theta) = \arg\max_{x \in \mathcal{X}} x^\top \theta$. By the convexity of J

$$\begin{aligned}(J(\theta_*) - J(\theta_t)) &\leq \max\{\nabla J(\theta_*)^\top(\theta_* - \theta_t), \nabla J(\theta_t)^\top(\theta_* - \theta_t)\} \\ &\leq \max\{x_*^\top(\theta_* - \theta_t), x_t^\top(\theta_* - \theta_t)\},\end{aligned}$$

where the second inequality follows from the definition of $\nabla J(\theta)$. Finally since $t > \tau$, by choice of ι, and since we are on the event $E_t(\delta)$, it holds that $x^\top(\theta_* - \theta_t) = x^\top(\theta_* - \hat{\theta}_t + \hat{\theta}_t - \theta_t) \leq D_t^* + D_t \leq 2M_\mu^{-1}$ for all $x \in \mathcal{X}$. It immediately follows that

$$\begin{aligned}\alpha(x_*^\top \theta_*, x_t^\top \theta_t) &= \int_{v=0}^1 \dot{\mu}\left(J(\theta_*) + v(J(\theta_t) - J(\theta_*))\right) dv \\ &\leq \dot{\mu}(J(\theta_*)) \int_{v=0}^1 \exp(v M_\mu |J(\theta_*) - J(\theta_t)|) dv,\end{aligned}$$

where the inequality is due to Lemma 3.8. Since it was just shown that $(J(\theta_*) - J(\theta_t)) \leq 2M_\mu^{-1}$, the above integral can be further bounded by

$$\dot{\mu}(J(\theta_t)) \int_{v=0}^1 \exp(2v) dv \leq 4\dot{\mu}(J(\theta)) = 4\dot{\mu}(x_t^\top \theta_t).$$

This gives that

$$R_t^{\text{PHE}} = \alpha(J(\theta_*), J(\theta_t))(J(\theta_*) - J(\theta_t)) \leq 4\dot{\mu}(x_t^\top \theta_t)(J(\theta_*) - J(\theta_t)).$$

Now all that is left is to bound $J(\theta_*) - J(\theta_t)$. On event E_t, θ_t belongs to $\mathcal{E}_{2,t} := \mathcal{E}_t(\beta'_{2,t}(\delta))$ and therefore it follows that

$$J(\theta_*) - J(\theta_t) \leq J(\theta_*) - \inf_{\theta \in \mathcal{E}_{2,t}} J(\theta).$$

Now define the set of optimistic parameters $\Theta^{\text{opt}} = \{\theta : J(\theta) > J(\theta_*)\}$. As in Abeille et al. [AL17], we can bound the above expression by the expectation of any random choice of $\tilde{\theta}$ in $\Theta_t^{\text{opt}} = \Theta^{\text{opt}} \cap \mathcal{E}_{2,t}$. This gives

$$J(\theta_*) - J(\theta_t) \leq \mathbb{E}\left[J(\tilde{\theta}) - \inf_{\theta \in \mathcal{E}_{2,t}} J(\theta) \big| h_t, \tilde{\theta} \in \Theta_t^{\text{opt}}\right].$$

where $\tilde{\theta} = \arg\max_{\theta \in \mathbb{R}^d} \mathcal{L}_\lambda\left(\theta; \{(x_s, y_s + z_{s,t})\}_{s=1}^{t-1}\right)$ and $z_{s,t} \sim \mathcal{N}\left(0, a_t^2 \dot{\mu}(x_s^\top \hat{\theta}_t)\right)$ independently. Since J is convex, using Lemma 4.24, we can directly relate

R_t^{PHE} with the gradient of J as

$$\begin{aligned} J(\theta_*) - J(\theta_t) &\leq \mathbb{E}\left[\sup_{\theta\in\mathcal{E}_{2,t}} \nabla J(\tilde{\theta})^\top(\tilde{\theta}-\theta)\big|h_t, \tilde{\theta}\in\Theta_t^{\text{opt}}\right] \\ &\leq \mathbb{E}\left[\|\nabla J(\tilde{\theta})\|_{H_t^{-1}} \sup_{\theta\in\mathcal{E}_{2,t}} \|\tilde{\theta}-\theta\|_{H_t}\big|h_t, \tilde{\theta}\in\Theta_t^{\text{opt}}\right] \\ &\leq 2\beta'_{2,t}(\delta)\mathbb{E}\left[\|\nabla J(\tilde{\theta})\|_{H_t^{-1}}|h_t, \tilde{\theta}\in\Theta_t^{\text{opt}}\right] \end{aligned}$$

where the second inequality follows from Cauchy-Schwarz and the third inequality follows from the fact that we are on the event $E_t(\delta)$. Again following Abeille et al. [AL17], let $f(\theta_t)$ be an arbitrary non-negative function of θ_t, then we can write the full expectation as

$$\begin{aligned} \mathbb{E}[f(\theta_t)|h_t] &\geq \mathbb{E}[f(\theta_t)|\theta_t\in\Theta_t^{\text{OPT}}, h_t]\mathbb{P}_t(\theta_t\in\Theta_t^{\text{OPT}}) \\ &\geq \mathbb{E}[f(\theta_t)|\theta_t\in\Theta_t^{\text{OPT}}, h_t]\cdot p_3, \end{aligned}$$

where the last inequality holds by Section 4.2.4. Setting $f(\theta_t) = 2\beta'_{2,t}(\delta)\|x_t\|_{H_t^{-1}}$ one obtains

$$J(\theta_*) - J(\theta_t) \leq \frac{2\beta'_{2,t}(\delta)}{p_3}\mathbb{E}[\|x_t\|_{H_t^{-1}}|h_t] = \frac{2\beta'_{2,t}(\delta)}{p_3}\left(\|x_t\|_{H_t^{-1}} + \mathbb{E}[\|x_t\|_{H_t^{-1}}|h_t] - \|x_t\|_{H_t^{-1}}\right),$$

which completes the proof. □

To bound the right hand side of Lemma 4.26, we will need the following frequently used Azuma-Hoeffding lemma as well as its corollary:

Lemma 4.27 (Azuma-Hoeffding)**.** *If a super-martingale $(Y_t)_{t\geq 0}$ corresponding to a filtration $\mathcal{F}_t$ satisfies $|Y_t - Y_{t-1}| < c_t$ for some constant c_t for all $t = 1, 2, ..., T$, then for any $\alpha > 0$,*

$$\mathbb{P}(Y_T - Y_0 \geq \alpha) \leq 2\exp\left(-\frac{\alpha^2}{2\sum_{t=1}^T c_t^2}\right).$$

Corollary 4. *If a super-martingale $(Y_t)_{t\geq 0}$ corresponding to a filtration $\mathcal{F}_t$ satisfies $|Y_t - Y_{t-1}| < c$ for all $t = 1, 2, ..., T$, then for $\delta \in (0, 1]$, $\alpha = \sqrt{2c^2T\log\frac{2}{\delta}}$*

$$\mathbb{P}(Y_T - Y_0 \geq \alpha) \leq \delta.$$

Proof. Plug in $\alpha = \sqrt{2c^2 T \log \frac{2}{\delta}}$, the probability can be bounded as follows:

$$\mathbb{P}(Y_T - Y_0 \geq \alpha) \leq 2\exp(-\frac{2c^2 T \log \frac{2}{\delta}}{2c^2 T}) = \delta.$$

□

Lemma 4.28. *There exist some event B with $\mathbb{P}(B) \leq \delta$ such that on*

$$E^{A-H} := (\cap_{t=\tau+1}^{n} E_t) \cap B^c$$

we have

$$\frac{4}{p_3} \sum_{t=\tau+1}^{n} \dot{\mu}(x_t^\top \theta_t) \beta'_{2,t}(\delta) \left(\mathbb{E}[\|x_t\|_{H_t^{-1}} | h_t] - \|x_t\|_{H_t^{-1}} \right) \leq \frac{4e\beta'_{2,n}(\delta)}{p_3} \sqrt{8n \log \frac{2}{\delta}}. \tag{4.33}$$

Proof. For all $x \in \mathcal{X}$, by Lemma 3.8, in rounds $t > \tau$,

$$\dot{\mu}(x^\top \hat{\theta}_t) \leq \exp(|x^\top(\hat{\theta}_t - \theta_*)|)\dot{\mu}(x^\top \theta_*) \leq e\dot{\mu}(x^\top \theta_*),$$

where the last step is because we use the event $E_{1,t} \supset E_t$ to ensure that $x^\top(\hat{\theta}_t - \theta_*) \leq \beta'_{1,t}(\delta)\|x\|_{H_t^{-1}}$ and by choice of ι (Eq. (4.23) and Lemma 4.18), $\beta'_{1,t}(\delta) < 1$. Use Lemma 3.8, the sum is less than

$$S := \frac{4e}{p_3} \sum_{t=\tau+1}^{n} \sqrt{\dot{\mu}(x_t^\top \hat{\theta}_t)} \sqrt{\dot{\mu}(x_t^\top \theta_*)} \beta'_{2,t}(\delta) \left(\mathbb{E}[\|x_t\|_{H_t^{-1}} | h_t] - \|x_t\|_{H_t^{-1}} \right).$$

Since the warm-up phase guarantees that $\|x_t\|_{H_t^{-1}} \leq \frac{1}{\iota} \leq \frac{1}{L}$ (Lemma 4.18) while obviously $\|x_t\|_{H_t^{-1}} \geq 0$, one gets for every $t > \tau$,

$$\left| \mathbb{E}[\|x_t\|_{H_t^{-1}} | h_t] - \|x_t\|_{H_t^{-1}} \right| \leq \frac{2}{L}.$$

Define

$$K_t := \sqrt{\dot{\mu}(x_t^\top \theta_*)} \sqrt{\dot{\mu}(x_t^\top \hat{\theta}_t)} \left(\mathbb{E}[\|x_t\|_{H_t^{-1}} | h_t] - \|x_t\|_{H_t^{-1}} \right).$$

Then, thanks to $\dot{\mu}(\cdot) \leq L$,

$$|K_t| \leq \sqrt{L} \cdot \sqrt{L} \cdot \frac{2}{L} \leq 2.$$

Note that $\{K_t\}_{t>\tau}$ is a martingale difference sequence itself by construction. Applying Corollary 4 gives that with probability at least $1-\delta$,

$$\frac{4e\beta'_{2,n}(\delta)}{p_3}\sum_{t=\tau+1}^{n} K_t \le \frac{4e\beta'_{2,n}(\delta)}{p_3}\sqrt{8n\log\frac{2}{\delta}}.$$

The result follows by letting B^c be the event when this last inequality holds and combining this last inequality with the previous bound on S which holds on $\cap_{t=\tau+1}^{n} E_t$. □

Bounding $R_t^{\mathbf{PRED}}$

Lemma 4.29. *On event E_t, for $t>\tau$, it holds that*

$$R_t^{PRED} \le \sqrt{e}(1+e)\beta'_t(\delta)\sqrt{\dot\mu(x_t^\top\theta_*)}\sqrt{\dot\mu(x_t^\top\hat\theta_t)\|x_t\|^2_{H_t^{-1}}}.$$

Proof. By adding then subtracting $\mu(x_t^\top\hat\theta_t)$ and triangle inequality, we have that

$$R_t^{\text{PRED}} = \mu(x_t^\top\theta_t) - \mu(x_t^\top\theta_*) \le |\mu(x_t^\top\theta_t) - \mu(x_t^\top\hat\theta_t)| + |\mu(x_t^\top\hat\theta_t) - \mu(x_t^\top\theta_*)|.$$

On event E_t, it follows from Lemma 4.18 that for $t>\tau$

$$|x^\top(\theta_* - \hat\theta_t)| \le \beta'_{1,t}(\delta)\|x_t\|_{H_t^{-1}} \le \frac{\beta'_{1,t}(\delta)}{\iota} \le M_\mu^{-1}, \forall x\in\mathcal{X}, \tag{4.34}$$

$$|x^\top(\theta_t - \hat\theta_t)| \le \beta'_{2,t}(\delta)\|x_t\|_{H_t^{-1}} \le \frac{\beta'_{2,t}(\delta)}{\iota} \le M_\mu^{-1}, \forall x\in\mathcal{X}. \tag{4.35}$$

The conditions of Lemma 4.23 are satisfied. Therefore it holds that

$$\begin{aligned}|\mu(x_t^\top\theta_t) - \mu(x_t^\top\hat\theta_t)| &\le \dot\mu(x_t^\top\hat\theta_t)\beta'_{2,t}(\delta)(1+eM_{\mathcal{L}}\beta'_{2,t}(\delta)\|x_t\|_{H_t^{-1}})\|x_t\|_{H_t^{-1}}\\ &\le (1+e)\dot\mu(x_t^\top\hat\theta_t)\beta'_{2,t}(\delta)\|x_t\|_{H_t^{-1}},\end{aligned}$$

where the second inequality holds by choice of ι. Similarly by applying Lemma 4.23, it follows that

$$|\mu(x_t^\top\hat\theta_t) - \mu(x_t^\top\theta_*)| \le (1+e)\dot\mu(x_t^\top\hat\theta_t)\beta'_{1,t}(\delta)\|x_t\|_{H_t^{-1}}.$$

Putting these bounds together gives

$$\begin{aligned}R_t^{\text{PRED}} &\le (1+e)\dot\mu(x_t^\top\hat\theta_t)\beta'_{2,t}(\delta)\|x_t\|_{H_t^{-1}} + (1+e)\dot\mu(x_t^\top\hat\theta_t)\beta'_{1,t}(\delta)\|x_t\|_{H_t^{-1}}\\ &= (1+e)\dot\mu(x_t^\top\hat\theta_t)\beta'_t(\delta)\|x_t\|_{H_t^{-1}}\\ &\le \sqrt{e}(1+e)\beta'_t(\delta)\sqrt{\dot\mu(x_t^\top\theta_*)}\sqrt{\dot\mu(x_t^\top\hat\theta_t)\|x_t\|^2_{H_t^{-1}}},\end{aligned}$$

where in the last step we use Lemma 3.8 and the fact that $D_t^* \leq M_\mu^{-1}$ for $t > \tau$ to bound $\dot{\mu}(x_t^\top \theta_*) \leq e\dot{\mu}(x_t^\top \hat{\theta}_t)$. □

Bounding the regret We will need the following auxiliary lemmas to bound the regret:

Lemma 4.30 (Lemma 19.4 of Lattimore & Szepesvári [LS18])**.** *Let* $V_0 \in \mathbb{R}^{d \times d}$ *be a positive definite and* $a_1, ...a_n \in \mathbb{R}^d$ *be a sequence of vectors with* $\|a_t\|_2 \leq L < \infty$ *for all* $t \in [n]$, $V_t = V_0 + \sum_{s \leq t} a_s a_s^\top$. *Then*

$$\sum_{t=1}^{n} \min\{1, \|a_t\|^2_{V_{t-1}}\} \leq 2\log(\frac{\det V_n}{\det V_0}) \leq 2d\log(\frac{traceV_0 + nL^2}{d\det(V_0)^{1/d}}).$$

Lemma 4.31. *For* $\lambda > 1$ *and sequence* $\{y_t\}_{t=1}^n \in \mathcal{X}$ *such that* $\|y_t\|_2 \leq X$ *it holds that*

$$\sum_{t=1}^{n} \|y_t\|^2_{V_t^{-1}} \leq 2d\log\frac{d\lambda + X^2 n}{d\lambda}.$$

where $V_t = \sum_{t=1}^n y_t y_t^\top + \lambda I_d$

Proof. Following the proof of Lemma 11 of Abbasi et al.[APS11], we have

$$\sum_{t=0}^{n} \|y_t\|^2_{V_t^{-1}} \leq 2\log\frac{\det(V_n)}{\det(V_0)}.$$

Using Lemma 19.4 of Lattimore et al.[LS18] it follows that

$$2\log\frac{\det(V_n)}{\det(V_0)} \leq 2d\log\frac{\mathrm{trace}(V_0) + n\max_{y\in\mathcal{X}}\|y\|_2^2}{d\det(V_0)^{1/d}} \leq 2d\log\frac{d\lambda + X^2 n}{d\lambda},$$

where the final inequality holds by the assumption that $\|y\|_2 \leq X$ for all $y \in \mathcal{X}$. Now we have that

$$\sum_{t=\tau}^{n} \|y_t\|^2_{V_t^{-1}} \leq 2d\log\frac{d\lambda + X^2 n}{d\lambda}$$

□

Lemma 4.32. *Let* $\{x_t\}_{t=1}^n \subset \mathbb{R}^d$ *and* $\theta_* \in \mathbb{R}^d$. *For* μ *it holds*

$$\sum_{t=1}^{n} \dot{\mu}(x_t^\top \theta_*) \leq n\dot{\mu}(x_*^\top \theta_*) + M_\mu \sum_{t=1}^{n} \mu(x_*^\top \theta_*) - \mu(x_t^\top \theta_*).$$

Proof. By an exact first order Taylor expansion of $\dot{\mu}$, the summation $\sum_{t=1}^{n} \dot{\mu}(x_t^\top \theta_*)$ can be written as:

$$\begin{aligned}
&\sum_{t=1}^{n} \dot{\mu}(x_*^\top \theta_*) + \int_{v=0}^{1} \ddot{\mu}\left(x_*^\top \theta_* + v(x_t - x_*)^\top \theta_*\right) dv (x_t - x_*)^\top \theta_* \\
=& n\dot{\mu}(x_*^\top \theta_*) + \sum_{t=1}^{n} \int_{v=0}^{1} \ddot{\mu}\left(x_*^\top \theta_* + v(x_t - x_*)^\top \theta_*\right) dv (x_t - x_*)^\top \theta_* \\
\leq& n\dot{\mu}(x_*^\top \theta_*) + \sum_{t=1}^{n} \left| \dot{\mu}(x_*^\top \theta_*) + \int_{v=0}^{1} \ddot{\mu}\left(x_*^\top \theta_* + v(x_t - x_*)^\top \theta_*\right) dv (x_t - x_*)^\top \theta_* \right| \\
\leq& n\dot{\mu}(x_*^\top \theta_*) + \sum_{t=1}^{n} \left| \int_{v=0}^{1} \ddot{\mu}\left(x_*^\top \theta_* + v(x_t - x_*)^\top \theta_*\right) dv \right| (x_* - x_t)^\top \theta_* \\
& \hfill \text{(Since } x_t^\top \theta_* \leq x_*^\top \theta_*) \\
\leq& n\dot{\mu}(x_*^\top \theta_*) + M_\mu \sum_{t=1}^{n} \left| \int_{v=0}^{1} \dot{\mu}\left(x_*^\top \theta_* + v(x_t - x_*)^\top \theta_*\right) dv \right| (x_* - x_t)^\top \theta_* \\
& \hfill \text{(Since } |\ddot{\mu}| \leq M_\mu \dot{\mu}) \\
\leq& n\dot{\mu}(x_*^\top \theta_*) + M_\mu \sum_{t=1}^{n} \alpha(x_*^\top \theta_*, x_t^\top \theta_*)(x_* - x_t)^\top \theta_* \qquad \text{(def. } \alpha) \\
\leq& n\dot{\mu}(x_*^\top \theta_*) + M_\mu \sum_{t=1}^{n} \mu(x_*^\top \theta_*) - \mu(x_t^\top \theta_*). \qquad \text{(Mean Value Theorem)}
\end{aligned}$$

This completes the proof. □

We now bound the regret. For R_t^{PHE}, on event $\cap_{t=\tau+1}^{n} E_t \cap E^{A-H}$, using Lemma 4.26 and Lemma 4.28,

$$\begin{aligned}
\sum_{t=\tau+1}^{n} R_t^{\text{PHE}} &\leq \frac{4}{p_3} \sum_{t=\tau+1}^{n} \dot{\mu}(x_t^\top \theta_t) \beta'_{2,t}(\delta) \left(\|x_t\|_{H_t^{-1}} + \mathbb{E}[\|x_t\|_{H_t^{-1}} | h_t] - \|x_t\|_{H_t^{-1}} \right) \\
& \hfill \text{(Lemma 4.26)} \\
&\leq \frac{4e\beta'_{2,n}(\delta)}{p_3} \sqrt{8n \log \frac{2}{\delta}} + \frac{4}{p_3} \sum_{t=\tau+1}^{n} \dot{\mu}(x_t^\top \theta_t) \beta'_{2,t}(\delta) \|x_t\|_{H_t^{-1}}. \\
& \hfill \text{(Lemma 4.28 and } \delta \leq \delta')
\end{aligned}$$

Now, we focus our attention on the sum

$$\begin{aligned}
&\frac{4}{p_3}\sum_{t=\tau+1}^{n}\dot{\mu}(x_t^\top\theta_t)\beta'_{2,t}(\delta)\|x_t\|_{H_t^{-1}} && (4.36)\\
\le\;& \frac{4\beta'_{2,n}(\delta)}{p_3}\sum_{t=\tau+1}^{n}\dot{\mu}(x_t^\top\theta_t)\|x_t\|_{H_t^{-1}} && (4.37)\\
\le\;& \frac{4e\beta'_{2,n}(\delta)}{p_3}\sum_{t=\tau+1}^{n}\dot{\mu}(x_t^\top\hat{\theta}_t)\|x_t\|_{H_t^{-1}} && (4.38)\\
\le\;& \frac{4e^{3/2}\beta'_{2,n}(\delta)}{p_3}\sum_{t=\tau+1}^{n}\sqrt{\dot{\mu}(x_t^\top\theta_*)}\sqrt{\dot{\mu}(x_t^\top\hat{\theta}_t)}\|x_t\|_{H_t^{-1}} && (4.39)\\
=\;& \frac{4e^{3/2}\beta'_{2,n}(\delta)}{p_3}\sum_{t=\tau+1}^{n}\sqrt{\dot{\mu}(x_t^\top\theta_*)}\|\tilde{x}_t\|_{H_t^{-1}} && (4.40)\\
\le\;& \frac{4e^{3/2}\beta'_{2,n}(\delta)}{p_3}\sqrt{\sum_{t=\tau+1}^{n}\dot{\mu}(x_t^\top\theta_*)}\sqrt{\sum_{t=\tau+1}^{n}\|\tilde{x}_t\|^2_{H_t^{-1}}} && (4.41)\\
\le\;& \frac{4e^{3/2}\beta'_{2,n}(\delta)}{p_3}\sqrt{2d\log(1+nL/(d\lambda))}\sqrt{\sum_{t=\tau+1}^{n}\dot{\mu}(x_t^\top\theta_*)}, && (4.42)
\end{aligned}$$

where the first inequality is because of $\beta'_{2,t}(\delta)\le\beta'_{2,n}(\delta)$; the second inequality is because of $D_t\le 1$ and Lemma 3.8; the thrid inequality is because of $D_t^*\le 1$ and Lemma 3.8; in the fifth inequality we denote $\tilde{x}_t := \sqrt{\dot{\mu}(x_t^\top\hat{\theta}_t)}x_t$ and H_t can be written as $H_t=\sum_{s=1}^{t-1}\tilde{x}_s\tilde{x}_s^\top+\lambda I$; the sixth inequality is because of Cauchy-schwarz inequality and the final inequality is because of Lemma 4.31. Thus on the event E_t,

$$\sum_{t=\tau+1}^{n} R_t^{\mathrm{PHE}} \le \frac{4e\beta'_{2,t}(\delta)}{p_3}\left(\sqrt{2ed\log(1+nL/(d\lambda))}\sqrt{\sum_{t=\tau+1}^{n}\dot{\mu}(x_t^\top\theta_*)}+\sqrt{8n\log\frac{2}{\delta}}\right).$$

For bounding the sum of R_t^{PRED} on the event E_t

$$\begin{aligned}\sum_{t=\tau+1}^{n} R_t^{\text{PRED}} &\leq \sum_{t=\tau+1}^{n} \sqrt{e}(1+e)\beta_t'(\delta)\sqrt{\dot{\mu}(x_t^\top\theta_*)}\sqrt{\dot{\mu}(x_t^\top\hat{\theta}_t)\|x_t\|_{H_t^{-1}}^2} && (4.43)\\ &\leq 2e^{3/2}\beta_n'(\delta)\sum_{t=\tau+1}^{n}\sqrt{\dot{\mu}(x_t^\top\theta_*)}\sqrt{\dot{\mu}(x_t^\top\hat{\theta}_t)\|x_t\|_{H_t^{-1}}^2} && (4.44)\\ &\leq 2e^{3/2}\beta_n'(\delta)\sqrt{\sum_{t=\tau+1}^{n}\dot{\mu}(x_t^\top\theta_*)}\sqrt{\sum_{t=\tau+1}^{n}\dot{\mu}(x_t^\top\hat{\theta}_t)\|x_t\|_{H_t^{-1}}^2} && (4.45)\\ &\leq 2e^{3/2}\beta_n'(\delta)\sqrt{\sum_{t=\tau+1}^{n}\dot{\mu}(x_t^\top\theta_*)}\sqrt{\sum_{t=\tau+1}^{n}\|\tilde{x}_t\|_{H_t^{-1}}^2} && (4.46)\\ &\leq 2e^{3/2}\beta_n'(\delta)\sqrt{2d\log(1+nL/(d\lambda))}\sqrt{\sum_{t=\tau+1}^{n}\dot{\mu}(x_t^\top\theta_*)}. && (4.47)\end{aligned}$$

where the first inequality is because of Lemma 4.29; the second inequality is because of $\beta_t'(\delta) \leq \beta_n'(\delta)$; the third inequality is because of Cauchy-schwarz inequality; in the fourth inequality we denote $\tilde{x}_t := \sqrt{\dot{\mu}(x_t^\top\hat{\theta}_t)}x_t$ and H_t can be written as $H_t = \sum_{s=1}^{t-1}\tilde{x}_s\tilde{x}_s^\top + \lambda I$; the last inequality is because of Lemma 4.31.

Now combining the bounds on R_t^{PHE} and R_t^{PRED} on the event E_t and E^{A-H}.

$$\begin{aligned}&\sum_{t=\tau+1}^{n} R_t^{\text{PHE}} + R_t^{\text{PRED}}\\ \leq&\frac{4e^{3/2}\beta_n'(\delta)}{p_3}\left(\sqrt{2d\log(1+\frac{nL}{d\lambda})}\sqrt{\sum_{t=\tau+1}^{n}\dot{\mu}(x_t^\top\theta_*)} + \sqrt{8n\log\frac{2}{\delta}}\right). \qquad (4.48)\end{aligned}$$

Dealing with the remaining sum, Lemma 4.32 gives

$$\sqrt{\sum_{t=\tau+1}^{n}\dot{\mu}(x_t^\top\theta_*)} \leq \sqrt{\dot{\mu}(x_*^\top\theta_*)n + M_\mu R(n)}.$$

Therefore

$$\begin{aligned}&\sum_{t=\tau+1}^{n} R_t^{\text{PHE}} + R_t^{\text{PRED}}\\ &\leq \frac{4e^{3/2}\beta_n'(\delta)}{p_3}\left(\sqrt{2d\log(1+\frac{nL}{d\lambda})}\sqrt{\dot{\mu}(x_*^\top\theta_*)n + M_\mu R(n)} + \sqrt{8n\log\frac{2}{\delta}}\right) + 2\Delta_{\max}\\ &\leq \frac{4e^{3/2}\beta_n'(\delta)}{p_3}\left(\sqrt{2d\log(1+\frac{nL}{d\lambda})}\left(\sqrt{\dot{\mu}(x_*^\top\theta_*)n} + \sqrt{M_\mu R(n)}\right) + \sqrt{8n\log\frac{2}{\delta}}\right)\\ &\quad + 2\Delta_{\max}. \qquad (\sqrt{a+b} \leq \sqrt{a} + \sqrt{b})\end{aligned}$$

Recall $R(n) \leq \tau\Delta_{\max} + \sum_{t=\tau+1}^{n} R_t^{\text{PHE}} + R_t^{\text{PRED}}$. Now since $x^2 \leq bx + c$ implies $x \leq b + \sqrt{c}$, which is proved in Lemma 4.6, letting $x = \sqrt{R(n)}$ gives

$$\sqrt{R(n)} \leq \frac{4e^{3/2}\beta_n'(\delta)}{p_3}\sqrt{2M_\mu d\log(1+\frac{nL}{d\lambda})} \\ + \sqrt{\tau\Delta_{\max} + \frac{4e^{3/2}\beta_n'(\delta)}{p_3}\left(\sqrt{2d\log(1+\frac{nL}{d\lambda})}\sqrt{\dot{\mu}(x_*^\top\theta_*)n} + \sqrt{8n\log\frac{2}{\delta}}\right) + 2\Delta_{\max}}.$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$ finishes bounding the regret on the event E_δ,

$$R(n) \leq \frac{32e^3\beta_n'(\delta)^2}{p_3^2}\left(2M_\mu d\log(1+nL/(d\lambda))\right) \\ + 2\tau\Delta_{\max} + \frac{8e^{3/2}\beta_n'(\delta)}{p_3}\left(\sqrt{2d\log(1+nL/(d\lambda))}\sqrt{\dot{\mu}(x_*^\top\theta_*)n} + \sqrt{8n\log\frac{2}{\delta'}}\right) \\ + 2\Delta_{\max}.$$

Noting that $\beta_n'(\delta) = \tilde{\mathcal{O}}(d\sigma\sqrt{S})$ and $\tau = \tilde{\mathcal{O}}(d^9L^2\kappa)$ completes the proof

$$R(n) \leq \mathcal{O}\left(\sigma d^{3/2}\sqrt{S\dot{\mu}(x_*^\top\theta_*)n} + d^9L^2\kappa\right).$$

Chapter 5

The Rarely Switching Variant of sGLM-PHE: sGLM-PHE-RS

Besides the warm-up phase, an expensive step in the algorithm is that we need to solve two MLE optimization problems in every round. While both optimization problems are convex (they have the same form, with different data), and as such can be solved relatively efficiently (e.g., using Newton's method), they need to solve the optimization problem twice doubling the cost of running our algorithm.

Note that "pure" optimistic algorithms (i.e., algorithms that rely on optimization to achieve optimism) also need to solve an MLE problem, but sometimes their computational cost can be unchecked. Indeed, optimistic algorithms not only need to solve the MLE problem, but they also need to figure out an optimistic decision. When the action set is finite, such an optimistic decision can be computed by considering an upper bound on the reward that can be incurred for each action and then selecting the action that maximizes this upper bound. This incurs a cost that scales with the number of actions. Yet there are no known efficient ways of implementing optimistic algorithms when the action set is infinite, e.g., if the action set is the unit ball: The cost of the naive discretization approach that maintains $\tilde{O}(\sqrt{n})$ regret in this case becomes $n^{\Omega(d)}$, which scales poorly with the dimension. In contrast, the PHE method can be efficiently implemented as long as linear optimization over the action set can be efficiently implemented. Just to given an example, this holds when the action set is the unit ball [AL17].

Yet, the algorithm may still be relatively expensive if the MLE problem is costly to optimize. A simple idea to save computation is to only resolve the MLE problem when there is a chance that the solution will be substantially different than the previous solution, cf. Algorithm 8. This idea has been proposed in the context of linear bandits by Abbasi et al. [APS11], but goes back to econometrics [TC60], in which thresholds were designed in order to determine when to intervene in an ongoing treatment, or experiment, in order to determine whether or not the treatment was effective. The motivation being that switching treatments can be costly. This idea, however, has not been explored in the context of randomized algorithms, such as posterior sampling, or PHE. A priori, it may be indeed unclear whether *rarely switching* works for randomized algorithms that rely on randomness to produce optimistic choices: the concern is that the price of using a non-optimistic action for a longer time may be just too high. As it turns out, this is not the case, whcih we detail in Section 5.2.

5.1 sGLM-PHE-RS

The warm-up phase is remained the same as Algorithm 7 and the change is made after that. As is mentioned in the previous section, the MLEs are not computed until sufficient data are collected to make substantial changes. We introduce a map $\rho(t) : [n] \to [n]$ for each $t \in [n]$ to be the last time step prior to t that the MLEs are computed. With little abuse of notation, we denote $\rho_t := \rho(t)$. The map value ρ_t can be t itself, indicating the MLEs are computed in round t, otherwise $\rho_t = \rho_{t-1}$, implying the data accumulated since round ρ_t do not make substantial difference that is worth to compute MLEs. The map value $\rho_{\tau+1}$ is initialized to be τ in line 5 and $\hat{\theta}_{\rho_{\tau+1}}$ is computed in line 6 as an initialization. The switching criteria, which is the key to the algorithm, is detailed in line 8, where $\det(H_t(\hat{\theta}_{\rho_t}))$ is used to measure the "information" contained in dataset $\{(x_s, y_s)\}_{s=1}^{t-1}$. If $\det(H_t(\hat{\theta}_{\rho_t}))$ is sufficently different from $\det(H_{\rho_t}(\hat{\theta}_{\rho_t}))$, that is, $\det(H_t(\hat{\theta}_{\rho_t})) > (1+C)\det(H_{\rho_t}(\hat{\theta}_{\rho_t}))$ where C is a value specified by the agent, then it is believed that the data collected since round ρ_t makes enough difference. The two MLEs are therefore computed and the variables maintained by the algorithm: $\rho_t, x_t, \hat{\theta}_t, \theta_t$ are updated as well. Otherwise every variable maintained by the algorithm remains the same and $\rho_t := \rho_{t-1}$. Here for easy display, we initialize a new variable to represent ρ_t for each $t \geq \tau$ but in practice this can be stored in a single variable. Note that C is a user specified value where there is a tradeoff: larger C incurs less switches, hence computationally cheaper, while a price in the regret needs to be paid. See Section 5.2 for details.

5.2 Analysis of sGLM-PHE-RS

In this section we detail the analysis of *Algorithm* 8. The analysis structure highly resembles that of Section 4.2 and adaptating lemmas in Section 4.2 to the rarely switching structure is essentially what is done here. The choice of ι and τ hence remains the same as Eq. (4.23) and Lemma 4.18.

Theorem 4. *With appropriately chosen parameters, the n-round regret of Algorithm 8, $R(n)$, is bounded as*

$$R(n) = \tilde{\mathcal{O}}\left(\sigma d^{3/2}\sqrt{(1+C)S\dot{\mu}(x_*^\top \theta_*)n} + \mathfrak{C}\right).$$

where $\mathfrak{C} = \mathrm{poly}(d, L, \kappa)$ and the exact dependence is shown in Section 5.2.4.

5.2.1 An Auxillary Lemma

At the beginning of this section, we present Sherman-Morrison equality, a well-known result in linear algebra, without proof:

Lemma 5.1. *(Sherman-Morrison) Let G a be nonsingular matrices and be E rank one matrix. Define $g = tr(EG^{-1})$, then*

$$(E+G)^{-1} = G^{-1} - \frac{1}{1+g}G^{-1}EG^{-1}$$

We need the following lemma that is a modification of Lemma 12 of Abbasi et al. [APS11] to support our analysis of Algorithm 8. It characterizes the ratio $\|x\|_{A^{-1}}/\|x\|_{B^{-1}}$ where B is a p.d. matrix and A is resulted from one or more rank-1 updates on B.

Algorithm 8 Rarely Switching sGLM-PHE

```
**Input**: $\mathcal{X}, M_\mu, L, \sigma, n, \delta$ and $S$
Calculate $\lambda, \tau$ according to Section 4.2.5
$\{(x_s, y_s)\}_{s=1}^{\tau} \leftarrow$ warm-up$(\mathcal{X}, \tau, \epsilon = 0.5)$
$\rho_{\tau+1} \leftarrow \tau$        {$\rho_t \leq t$ is the last time step prior to $t$ that the MLEs were updated}
$\hat{\theta}_{\rho_{\tau+1}} \leftarrow \arg\max_{\theta \in \mathbb{R}^d} \mathcal{L}_\lambda(\theta; \mathcal{D}_\tau)$
**for** $t = \tau + 1, ..., n$ **do**
    $\mathcal{D}_t = \{(x_s, y_s)\}_{s=1}^{t-1}$
    **if** $\det(H_t(\hat{\theta}_{\rho_t})) > (1 + C)\det(H_{\rho_t}(\hat{\theta}_{\rho_t}))$ **then**
        $\hat{\theta}_t \leftarrow \arg\max_{\theta \in \mathbb{R}^d} \mathcal{L}_\lambda(\theta; \mathcal{D}_t)$
        Sample $z_{s,t} \sim \mathcal{N}(0, a^2 \dot{\mu}(x_t^\top \hat{\theta}_t))$ for $s \in [t-1]$
        Set $\tilde{\mathcal{D}}_t = \{(x_s, y_s + z_{s,t})\}_{s=1}^{t-1}$
        $\theta_t \leftarrow \arg\max_{\theta \in \mathbb{R}^d} \mathcal{L}_\lambda(\theta; \tilde{\mathcal{D}}_t)$
        $\rho_t \leftarrow t$
        $x_t = \arg\max_{x \in \mathcal{X}} x^\top \theta_t$
    **else**
        $\rho_t \leftarrow \rho_{t-1}$, $x_t \leftarrow x_{t-1}$
    **end if**
    Observe $y_t$
**end for**
```

Lemma 5.2 (Modified lemma 12 of Abbasi et al. [APS11])**.** *Let B be a positive definite matrix, C be a positive semi-definite matrix and $A = B + C$. Then we have that*

$$\sup_{x \neq 0} \frac{x^\top B^{-1} x}{x^\top A^{-1} x} \leq \frac{\det(B^{-1})}{\det(A^{-1})}$$

Proof. We consider first a simple case. Assume $C = mm^\top$ where $m \in \mathbb{R}^d$ is a nonzero vector. By lemma 5.1,

$$A^{-1} = B^{-1} - \frac{1}{1 + \mathrm{tr}(mm^\top B^{-1})} B^{-1} mm^\top B^{-1}$$

Note that $\mathrm{tr}(mm^\top B^{-1}) = \mathrm{tr}(m^\top B^{-1} m) = \|m\|_{B^{-1}}^2 > 0$. Hence, it follows that

$$\frac{1}{1 + \mathrm{tr}(mm^\top B^{-1})} B^{-1} mm^\top B^{-1} = \tilde{m}\tilde{m}^\top$$

where

$$\tilde{m} = \frac{1}{\sqrt{1 + \|m\|_{B^{-1}}^2}} B^{-1} m$$

Then we have

$$B^{-1} = A^{-1} + \tilde{m}\tilde{m}^\top$$

Let x be an arbitrary d-dimensional vector. Using Cauchy-schawrz inequality, we have

$$(x^\top \tilde{m})^2 \leq \|x\|_{A^{-1}}^2 \|\tilde{m}\|_A^2$$

Therefore,

$$x^\top (A^{-1} + \tilde{m}\tilde{m}^\top) x \leq x^\top A^{-1} x + \|x\|_{A^{-1}}^2 \|\tilde{m}\|_A^2 = (1 + \|\tilde{m}\|_A^2)\|x\|_{A^{-1}}^2.$$

Furthermore ,

$$\det(B^{-1}) = \det(A^{-1})(1 + \|\tilde{m}\|_A^2)$$

which yields

$$\frac{x^\top B^{-1} x}{x^\top A^{-1} x} \leq 1 + \|\tilde{m}\|_A^2 = \frac{\det(B^{-1})}{\det(A^{-1})}.$$

Finishing the proof of this case. If $A = B + m_1 m_1^\top + ... + m_{t-1} m_{t-1}^\top$, then by the above argument, $B^{-1} = A^{-1} + \tilde{m}_1 \tilde{m}_1^\top + ... + \tilde{m}_{t-1} \tilde{m}_{t-1}^\top$. Let $V_s = A^{-1} + \tilde{m}_1 \tilde{m}_1^\top + ... + \tilde{m}_{s-1} \tilde{m}_{s-1}^\top$. Then again by the above argument

$$\begin{aligned} \frac{x^\top B^{-1} x}{x^\top A^{-1} x} &= \frac{x^\top V_1 x}{x^\top A^{-1} x} \frac{x^\top V_2 x}{x^\top V_1 x} \cdots \frac{x^\top B^{-1} x}{x^\top V_{t-1} x} \\ &\leq \frac{\det(V_1)}{\det(A^{-1})} \frac{\det(V_2)}{\det(V_1)} \cdots \frac{\det(B^{-1})}{\det(V_{t-1})} \\ &= \frac{\det(B^{-1})}{\det(A^{-1})}. \end{aligned}$$

This completes the proof. □

5.2.2 The Good Events

Similar to Section 4.2.2, in this section we present the events that allows us to upper bound the regret. The analysis still goes through by analyzing the three events whose definition need to be modified according to $\hat{\theta}_{\rho_t}$ as follows.

1. The event $E_{1,t}^{\text{RS}}$ is that the absolute value of the projection of $\theta_* - \hat{\theta}_t$ on any $x \in \mathcal{X}$ is small. Throughout this section, we define an ellipsoid-like confidence set centered at $\hat{\theta}_t$ similar to the analysis of the main algorithm. Specifically, for $\beta > 0$, letting

$$\mathcal{E}_{t,\rho_t}(\beta) = \{\theta \in \mathbb{R}^d : \forall x \in \mathcal{X}, |x^\top(\theta - \hat{\theta}_{\rho_t})| \leq \beta \|x\|_{H_t^{-1}(\hat{\theta}_{\rho_t})}\}$$

we then let

$$E_{1,t}^{\mathrm{RS}} = \left\{\theta_* \in \mathcal{E}_{t,\rho_t}(\sqrt{1+C}\bar{\beta}_1(\delta))\right\}$$

for $\bar{\beta}_1(\delta) = 2\gamma_n(\delta)$ where $\gamma_n(\delta)$ is defined in Eq. (4.10).

2. The event $E_{2,t}^{\mathrm{RS}}$ is that the absolute value of the projection of $\theta_{\rho_t} - \hat{\theta}_{\rho_t}$ onto any $x \in \mathcal{X}$ is small. Specifically,

$$E_{2,t}^{\mathrm{RS}} = \left\{\theta_{\rho_t} \in \mathcal{E}_{t,\rho_t}(\sqrt{1+C}\bar{\beta}_2(\delta))\right\},$$

for $\bar{\beta}_2(\delta) = 16C_d(\delta)\gamma_n(\delta)$ where $\gamma_n(\delta)$ is defined in Eq. (4.10) and $C_d(\delta)$ is defined in Section 4.2.2.

3. The event $E_{3,t}^{\mathrm{RS}}$ is that the perturbed MLE frequently encourages optimism, i.e., $x_t^\top \theta_{\rho_t}$ is large in comparison to $x_*\theta_*$ holds with constant probability. Specifically,

$$E_{3,t}^{\mathrm{RS}} = \left\{x_t^\top \theta_{\rho_t} > x_*^\top \theta_*\right\}$$

5.2.3 Analysis of the Good Events

Event E_{1,ρ_t} We now show that event E_{1,ρ_t} holds with high probability by defining E''_{1,ρ_t} and showing that $E_{1,\rho_t} \subset E''_{1,\rho_t} \subset E_{1,t}^{\mathrm{RS}}$. We now define E''_{1,ρ_t}:

$$E''_{1,\rho_t} := \{\theta_* \in \mathcal{E}_{\rho_t}(\bar{\beta}_1(\delta))\}$$

Lemma 5.3. *For $\bar{\beta}_1(\delta) = 2\gamma_n(\delta)$. For $\delta \in (0,1]$, on event $\mathcal{E}(\delta)$, it follows that $E_{1,\rho_t} \subset E''_{1,\rho_t}$, hence E''_{1,ρ_t} holds for all $t \leq n$.*

Proof. From Lemma 4.17 and the choice of ι in Section 4.2.5, one gets $D^*_{\rho_t} \leq 1/M_\mu$ thus $\bar{\beta}_1(\delta) \geq \beta'_{1,\rho_t}(\delta)$ for all $t \leq n$ after the warm-up phase, whose guarantee on D^*_t holds on event $\mathcal{E}(\delta)$. Then the ellipsoid-like set $\mathcal{E}_{\rho_t}(\beta'_{1,\rho_t}) \subset \mathcal{E}_{\rho_t}(\bar{\beta}_{1,\rho_t})$. It thus holds that $E_{1,\rho_t} \subset E''_{1,\rho_t}$. Recall that Lemma 4.11 gives $\mathbb{P}(\cap_{t\in[n]} E_{1,t}) \geq 1-\delta$. Noting that $\cap_{t\in[n]} E_{1,t} \subset \cap_{t\in[n]} E_{1,\rho_t}$ finishes the proof. □

The next step is to show that $E''_{1,\rho_t} \subset E_{1,t}^{\mathrm{RS}}$ and hence $\mathbb{P}(E_{1,t}^{\mathrm{RS}}) \geq 1-\delta$ for any t, which is detailed in the next lemma:

Lemma 5.4. *For $\delta \in (0,1]$, on event $\mathcal{E}(\delta)$, $E_{1,t}^{RS}$ holds for all $t \le n$.*

Proof. By design of Algorithm 8, the determinants satisfy

$$\det(H_t(\hat{\theta}_{\rho_t})) \le (1+C)\det(H_{\rho_t}(\hat{\theta}_{\rho_t})).$$

Then by Lemma 5.2, it follows that for all $x \in \mathcal{X}$,

$$\|x\|_{H_{\rho_t}^{-1}(\hat{\theta}_{\rho_t})} \le \sqrt{1+C}\|x\|_{H_t^{-1}(\hat{\theta}_{\rho_t})}.$$

because

$$x^\top H_{\rho_t}^{-1}(\hat{\theta}_{\rho_t})x \le \frac{\det(H_t(\hat{\theta}_{\rho_t}))}{\det(H_{\rho_t}(\hat{\theta}_{\rho_t}))}\left(x^\top H_t^{-1}(\hat{\theta}_{\rho_t})x\right) \le (1+C)x^\top H_t^{-1}(\hat{\theta}_{\rho_t})x$$

and note that $\|x\|_M = \sqrt{x^\top M x}$ for a positive-semidefinite matrix by definition. Therefore we have that $E''_{1,\rho_t} \subset E^{\mathrm{RS}}_{1,t}$. □

Event $E_{2,t}^{\mathbf{RS}}$ We show that $\cap_{t=\tau+1}^n E^{\mathrm{RS}}_{2,t}$ happens with probability at least $1-\delta$ given the past by defining E''_{2,ρ_t} and showing that $E_{2,\rho_t} \subset E''_{2,\rho_t} \subset E^{\mathrm{RS}}_{2,t}$. We now define E''_{2,ρ_t}:

$$E''_{2,\rho_t} := \{\theta_{\rho_t} \in \mathcal{E}_{\rho_t}(\bar{\beta}_2(\delta))\}$$

The next lemma shows that the events $E''_{2,t}$ holds with high probability given the past.

Lemma 5.5. *Let $\mathcal{I}_t := \{\tau < s \le t : \rho_s = s\}$ be the set of indices that the MLEs are recomputed prior to round t. For $\bar{\beta}_2(\delta) = 16C_d(\delta)\gamma_n(\delta)$, on event $\cap_{s\in\mathcal{I}_t} E_{4,s}$, it follows that $\mathbb{P}_t(\cap_{s\in\mathcal{I}_t} E''_{2,s}) \ge 1-\delta$ holds $\mathbb{P}$-almost surely.*

Proof. From Lemma 4.17 and the choice of ι Lemma 4.18 and Eq. (4.23), one gets for all $s \in \mathcal{I}_t$, it follows that $D_s \le 1/M_\mu$ on event $E_{4,s}$. Replacing D_s in $\beta'_{2,s}(\delta)$ with this upper bound results in $\bar{\beta}_2(\delta)$. Thus it holds that

$$E_{2,s} \subset E''_{2,s}, \forall s \in \mathcal{I}_t. \tag{5.1}$$

Note that

$$\bigcap_{s=1}^{t} E_{2,s} \subset \bigcap_{s\in\mathcal{I}_t} E_{2,s} \subset \bigcap_{s\in\mathcal{I}_t} E''_{2,s}$$

Finally Lemma 4.21 gives $\mathbb{P}_t(\cap_{s=\tau+1}^t E_{2,s}) \ge 1-\delta$ holds $\mathbb{P}$-almost surely for all $t > \tau$. Then the statement follows by using the monotone property of probability measure. □

The next step is to show that $E''_{2,\rho_t} \subset E^{\mathrm{RS}}_{2,t}$ and hence $\mathbb{P}_t(\cap_{s=\tau+1}^{t} E^{\mathrm{RS}}_{2,s}) \geq 1-\delta$ for all $t \in [n]$, which is detailed in the proof Lemma 5.6:

Lemma 5.6. *On event E_{4,ρ_t}, for $\bar{\beta}_2(\delta) = 16C_d(\delta)\gamma_n(\delta)$, it follows that $\mathbb{P}_t(\cap_{s=\tau+1}^{t} E^{RS}_{2,s}) \geq 1-\delta$ holds $\mathbb{P}$-almost surely for $t > \tau$ and $\delta \in (0,1]$.*

Proof. By the design of Algorithm 8, the determinants satisfy

$$\det(H_t(\hat{\theta}_{\rho_t})) \leq (1+C)\det(H_{\rho_t}(\hat{\theta}_{\rho_t})).$$

Then by Lemma 5.2, it follows that for all $x \in \mathcal{X}$,

$$\|x\|_{H_{\rho_t}^{-1}(\hat{\theta}_{\rho_t})} \leq \sqrt{1+C}\|x\|_{H_t^{-1}(\hat{\theta}_{\rho_t})}.$$

Therefore we have that

$$E''_{2,\rho_t} \subset E^{\mathrm{RS}}_{2,t} \tag{5.2}$$

for all t and note that

$$\bigcap_{s \in \mathcal{I}_t} E''_{2,s} \subset \bigcap_{s=\tau+1}^{t} E^{\mathrm{RS}}_{2,s},$$

therefore it follows that $\mathbb{P}_t(\cap_{s=\tau+1}^{t} E^{\mathrm{RS}}_{2,s}) \geq 1-\delta$. □

Event $E^{\mathbf{RS}}_{3,t}$ We show that $E^{\mathrm{RS}}_{3,t}$ happens with probability at least p_3, which is detailed in the next lemma:

Lemma 5.7. *On event $E^{RS}_{1,\rho_t}, E^{RS}_{2,\rho_t}$ and E_{4,ρ_t}, it follows that for any $t > \tau$ $\mathbb{P}_t(E^{RS}_{3,t}) \geq p_3$.*

Proof. Note that by design of the algorithm $x_t = x_{\rho_t}, \theta_t = \theta_{\rho_t}$, therefore it follows that:

$$\mathbb{P}_t(E^{\mathrm{RS}}_{3,t}) = \mathbb{P}_t(x_t^\top \theta_t > x_* \theta_*) = \mathbb{P}_{\rho_t}(x_{\rho_t}^\top \theta_{\rho_t} > x_* \theta_*) = \mathbb{P}_{\rho_t}(E_{3,\rho_t})$$

Note that E_{1,ρ_t} and E_{2,ρ_t} holds because $E^{\mathrm{RS}}_{1,\rho_t}$ and $E^{\mathrm{RS}}_{2,\rho_t}$ holds (Eqs. (5.1) and (5.2)). Therefore, all of the conditions of Lemma 4.22 are satisfied:

$$\mathbb{P}_t(E^{\mathrm{RS}}_{3,t}) = \mathbb{P}_{\rho_t}(E_{3,\rho_t}) \geq p_3$$

where the last inequality is an immediate result of Lemma 4.22. □

5.2.4 Analysis of the regret bound of Algorithm 8

In previous sections, we only give an informal version of the regret bound of Algorithm 6. In this section, we detail the formal version as well as the analysis.

Theorem 5. *Under assumptions 3.6.1-3.6.3 and assumption 3.10.1. Let $m = M_\mu/\log(2)$, $\delta \in (0,1)$ and $\beta'_n(\delta) = \beta'_{1,n}(\delta) + \beta'_{2,n}(\delta)$. With $\gamma_n(\delta)$ chosen in Eq. (4.10), λ chosen in Eq. (4.9), $\bar\beta_{1,t}(\delta)$ and $\bar\beta_{2,t}(\delta)$ chosen in Section 5.2.2, a chosen in Lemma 4.19, ι chosen in Lemma 4.18 and τ chosen in Lemma 4.10, which we detail all of them below:*

$$\begin{aligned}
m &= M_\mu/\log(2),\\
\gamma_n(\delta) &= \sqrt{8m\sigma d\left(S + \frac{1}{2m}\sigma\right)\log\left(\frac{2}{\delta}\sqrt{1+Ln/d}\right)},\\
\lambda &= \frac{2m\sigma}{S+\frac{1}{2m}\sigma}d\log\left(\frac{2}{\delta}\sqrt{1+Ln/d}\right),\\
a^2 &= 2\gamma_n(\delta),\\
C_d(\delta) &= \sqrt{d+\log(n/\delta)},\\
\bar\beta_{1,n}(\delta) &= 2\gamma_n(\delta),\\
\bar\beta_{2,n}(\delta) &= 16C_d(\delta)\gamma_n(\delta),\\
\bar\beta_n(\delta) &= \bar\beta_{1,n}(\delta) + \bar\beta_{2,n}(\delta),\\
\iota &= 64LM_\mu^2\left(1+\frac{4a^2M_\mu^2 d(1+\log(n/\delta))}{\lambda}\right)^2 C_d(\delta)^2\gamma_n(\delta)^2\\
&\quad + \left(1+2M_\mu(S+\frac{1}{2m}\sigma)\right)^2 + L,\\
\tau &= \iota^2 d\kappa = \tilde{\mathcal{O}}(d^9L^2\kappa),
\end{aligned}$$

the regret of Algorithm 6 can be upper bounded by

$$\begin{aligned}
R(n) \le{}& \frac{32e^3(1+C)\bar\beta_n(\delta)^2}{p_3^2}\left(2M_\mu d\log(1+nL/(d\lambda))\right) + 2\tau\Delta_{\max}\\
&+ \frac{8e^{3/2}\sqrt{1+C}\bar\beta_n(\delta)}{p_3}\left(\sqrt{2d\log\left(1+\frac{nL}{d\lambda}\right)}\sqrt{\dot\mu(x_*^\top\theta_*)n} + \sqrt{8n\log\frac{2}{\delta'}}\right)\\
&+ 2\Delta_{\max},
\end{aligned}$$

with probability at least $1-5\delta$*.*

In big-O notation, it can be written as

$$R(n) \leq \tilde{\mathcal{O}}\left(\sigma d^{3/2}\sqrt{(1+C)S\dot{\mu}(x_*^\top\theta_*)n} + d^9L^2\kappa\right).$$

Denote $E_t^{\text{RS}} = \mathcal{E}(\delta) \cap E_{1,t}^{\text{RS}} \cap E_{2,t}^{\text{RS}} \cap E_{4,\rho t}$ and $\bar{\beta}(\delta) = \bar{\beta}_1(\delta) + \bar{\beta}_2(\delta)$. We then decompose the regret into:

$$\begin{aligned} R(n) = \sum_{t=1}^{n} \mu(x_*^\top\theta_*) - \mu(x_t^\top\theta_*) &\leq \tau\Delta_{\max} + \sum_{t=\tau+1}^{n} \mu(x_*^\top\theta_*) - \mu(x_t^\top\theta_*) \\ &\leq (\tau+1)\Delta_{\max} + \sum_{t=\tau+1}^{n} \underbrace{\mu(x_*^\top\theta_*) - \mu(x_t^\top\theta_{\rho t})}_{R_t^{\text{PHE,RS}}} + \underbrace{\mu(x_t^\top\theta_{\rho t}) - \mu(x_t^\top\theta_*)}_{R_t^{\text{PRED,RS}}}. \end{aligned}$$

Following exactly the same steps in the proof of Theorem 1, we bound $R_t^{\text{PHE,RS}}$ and $R_t^{\text{PRED,RS}}$ separately on event E_t^{RS}.

Bound $R_t^{\textbf{PHE,RS}}$

Lemma 5.8. *On event* E_t^{RS}*, for rounds* $t > \tau$ *it holds that*

$$R_t^{PHE,RS} \leq 4\dot{\mu}(x_t^\top\theta_{\rho t})\sqrt{1+C}\bar{\beta}_2(\delta)/p_3\left(\|x_t\|_{H_t^{-1}(\hat{\theta}_{\rho t})} + \mathbb{E}[\|x_t\|_{H_t^{-1}(\hat{\theta}_{\rho t})}|h_t] - \|x_t\|_{H_t^{-1}(\hat{\theta}_{\rho t})}\right).$$

Proof. By replacing R_t^{PHE}, θ_t $\beta'_{2,t}(\delta)$, $\mathcal{E}_t(\beta'_{2,t}(\delta))$, $\|x_t\|_{H_t^{-1}}$ with $R_t^{\text{PHE,RS}}$, $\theta_{\rho t}$, $\sqrt{1+C}\bar{\beta}_2$, $\mathcal{E}_{t,\rho t}(\sqrt{1+C}\bar{\beta}_2(\delta))$, $\|x_t\|_{H_t^{-1}(\hat{\rho}_t)}$, the proof finishes by following exactly the same steps in the proof of Lemma 4.26 □

Lemma 5.9. *On event* E_t^{RS}*, with probability at least* $1-\delta$*, it holds that*

$$\begin{aligned} &\frac{4}{p_3}\sum_{t=\tau+1}^{n} \dot{\mu}(x_t^\top\theta_{\rho t})\sqrt{1+C}\bar{\beta}_2(\delta)\left(\mathbb{E}[\|x_t\|_{H_t^{-1}(\hat{\theta}_{\rho t})}|h_t] - \|x_t\|_{H_t^{-1}(\hat{\theta}_{\rho t})}\right) \\ \leq &\frac{4e\sqrt{1+C}\bar{\beta}_2(\delta)}{p_3}\sqrt{8n\log\frac{2}{\delta}}. \end{aligned}$$

Proof. Note that for $t > \tau$, by the choice of ι in Section 5.2.4, it follows that,

$$\|x_t\|_{H_t^{-1}(\hat{\theta}_{\rho t})} \leq \|x_t\|_{H_{\rho t}^{-1}} \leq \frac{1}{\iota} \leq \frac{1}{L}$$

By replacing $\hat{\theta}_t$, θ_t, $\beta'_{2,t}$, $\|x_t\|_{H_t^{-1}}$ with $\hat{\theta}_{\rho t}$,$\theta_{\rho t}$,$\sqrt{1+C}\bar{\beta}_2$, $\|x_t\|_{H_t^{-1}(\hat{\rho}_t)}$ and following exactly the same steps in the proof of Lemma 4.28, the statement follows. □

Denote the event that Lemma 5.9 holds to be $E^{A-H,RS}$

Bound on $R_t^{\textbf{PRED,RS}}$

Lemma 5.10. *On Event E_t^{RS}, for any $t > \tau$, it holds that*

$$R_t^{PRED,RS} \leq \sqrt{e}(1+e)\sqrt{1+C}\bar{\beta}(\delta)\sqrt{\dot{\mu}(x_t^\top \theta_*)}\sqrt{\dot{\mu}(x_t^\top \hat{\theta}_{\rho t})\|x_t\|_{H_t^{-1}(\hat{\theta}_{\rho t})}}$$

Proof. Note that on event E_{4,ρ_t}, $D_t^* \leq 1/M_\mu$ and $D_t \leq 1/M_\mu$. Replacing R_t^{PRED},$\|x_t\|_{H_t^{-1}}$ and $\beta_t'(\delta)$ with $R_t^{\text{PRED,RS}}$, $\|x_t\|_{H_t^{-1}(\hat{\rho}_t)}$ and $\sqrt{1+C}\bar{\beta}(\delta)$ gives the proof by following exact the same steps in the proof of Lemma 4.29. □

Bound the regret For $R_t^{\text{PHE,RS}}$, by using Lemma 5.8 and Lemma 5.9, which are the adapted versions of Lemma 4.26 and Lemma 4.28, on event $E^{A-H,RS}$ and E_t^{RS}, it follows that

$$\begin{aligned} R_t^{\text{PHE,RS}} &\leq \frac{4}{p_3}\dot{\mu}(x_t^\top \theta_{\rho t})\sqrt{1+C}\bar{\beta}(\delta)\left(\|x_t\|_{H_t^{-1}(\hat{\theta}_{\rho t})} + \mathbb{E}[\|x_t\|_{H_t^{-1}(\hat{\theta}_{\rho t})}|h_t] - \|x_t\|_{H_t^{-1}(\hat{\theta}_{\rho t})}\right) \\ &\leq \frac{4e\sqrt{1+C}\bar{\beta}(\delta)}{p_3}\left(\dot{\mu}(x_t^\top \theta_{\rho t})\|x_t\|_{H_t^{-1}(\hat{\theta}_{\rho t})} + \sqrt{8n\log\frac{2}{\delta}}\right). \end{aligned}$$

Following Eqs. (4.36) to (4.42) and noting that $D_{\rho_t}^* \leq 1, D_{\rho_t} \leq 1, \sum_{t=\tau+1}^n R_t^{\text{PHE,RS}}$ can be bounded on event E_t^{RS} as

$$\sum_{t=\tau+1}^n R_t^{\text{PHE,RS}} \leq \frac{4e\sqrt{1+C}\bar{\beta}(\delta)}{p_3}\left(\sqrt{2ed\log\left(1+\frac{nL}{d\lambda}\right)}\sqrt{\sum_{t=\tau+1}^n \dot{\mu}(x_t^\top \theta_*)} + \sqrt{8n\log\frac{2}{\delta}}\right)$$

For bounding $\sum_{t=\tau+1}^n R_t^{\text{PRED,RS}}$ on event E_t^{RS}, we follow Eqs. (4.43) to (4.47), where Lemma 4.29 used in the first inequality is replaced with Lemma 5.10, and the sum can be bounded as:

$$\sum_{t=\tau+1}^n R_t^{\text{PRED,RS}} \leq 2e^{3/2}\sqrt{1+C}\bar{\beta}(\delta)\sqrt{2d\log(1+nL/(d\lambda))}\sqrt{\sum_{t=\tau+1}^n \dot{\mu}(x_t^\top \theta_*)}.$$

and therefore, on the event E_t^{RS}, it follows that

$$\begin{aligned} \sum_{t=\tau+1}^n R_t^{\text{PHE,RS}} + R_t^{\text{PRED,RS}} \leq& \frac{4e^{\frac{3}{2}}\sqrt{1+C}\bar{\beta}(\delta)}{p_3}\times \\ &\left(\sqrt{2d\log\left(1+\frac{nL}{d\lambda}\right)}\sqrt{\sum_{t=\tau+1}^n \dot{\mu}(x_t^\top \theta_*)} + \sqrt{8n\log\frac{2}{\delta}}\right). \end{aligned}$$

Using Lemma 4.6 and following the calculations starting from Eq. (4.48) in the proof of Theorem 1 gives the final regret bound

$$\begin{aligned}R(n) \leq \frac{32e^3(1+C)\bar{\beta}(\delta)^2}{p_3^2}\left(2M_\mu d\log(1+nL/(d\lambda))\right) + 2\tau\Delta_{\max} \\ + \frac{8e^{3/2}\sqrt{1+C}\bar{\beta}(\delta)}{p_3}\left(\sqrt{2d\log(1+nL/(d\lambda))}\sqrt{\dot{\mu}(x_*^\top\theta_*)n} + \sqrt{8n\log\frac{2}{\delta'}}\right) \\ + 2\Delta_{\max}.\end{aligned}$$

Noting that $\bar{\beta}(\delta') = \tilde{\mathcal{O}}(d\sigma\sqrt{S})$ and $\tau = \tilde{\mathcal{O}}(d^5L^2\kappa)$ completes the proof

$$R(n) \leq \mathcal{O}\left(\sigma d^{3/2}\sqrt{S\dot{\mu}(x_*^\top\theta_*)n} + d^9L^2\kappa\right).$$

Chapter 6

Conclusion and Future Work

In this thesis we extended perturbed history exploration (PHE) to subgaussian generalized linear bandits (sGLBs), pushing both our understanding of sGLBs and the boundary of the problems that PHE applies to. For the former, we showed that the mean function in sGLBs automatically enjoys a self-concordance property, showing that, amongst other things, all sGLBs with log-concave density function are not too different from logistic bandits. This insight allowed us to extend the scope of previous worked that considered only logistic bandits to sGLBs with log-concave density function, partially answering the question proposed by Fillipi et al. [Fil+10]. On the front of PHE, we show that adding perturbations that reflect the uncertainty at each data point, through the curvature of the likelihood function, gives state-of-the-art guarantees, at least in logistic bandits. This leads us to conjecture that this is generally necessary: that in all settings, PHE should add noise to each observation in relation to the amount of uncertainty. We also showed that the rarely switchiing prcedure, which is able to significantly increase the computational efficiency of an algorithm, has theoretical guarantees on randomized algorithms such as Posterior Sampling and Perturbed History Exploration, which previously has only been shown to hold in UCB-type of pure optimistic algorithms.

An important direction for future work is removing the warm-up phase. While the warm-up phase is not a problem for parametric bandits—therein Algorithm 7 is efficient, that is, poly(d)—it does not scale to reinforcement learning, where we effectively must take sequences of actions, and not a single action, within each round. For such more general settings, adding fake data, as in Russo et al. [Rus19], seems like a promising direction. Another direction which may be worth exploring is in making the results of Ishfaq et al. [Ish+21] for model-free RL with bounded eluder dimension rigorous. Therein, the authors *assume* that the probability of optimism is constant to show their result. Removing this assumption and proving that, in their setting, anti-concentration occurs with constant probability using the analysis and techniques developed in this work would further establish the universality of perturbed history exploration.

Another crucial direction of future work is to investigate how to remove the dependency of M_μ on κ for the link function of all sGLMs. As is mentioned in Section 3.10 that the self-concordant parameter of many of frequently used distributions are independent very small.

References

[APS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. "Improved algorithms for linear stochastic bandits". In: *Advances in Neural Information Processing Systems*. 2011, pp. 2312–2320.

[AFC21] Marc Abeille, Louis Faury, and Clément Calauzènes. "Instance-wise minimax-optimal algorithms for logistic bandits". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 3691–3699.

[AL+17] Marc Abeille, Alessandro Lazaric, et al. "Linear thompson sampling revisited". In: *Electronic Journal of Statistics* 11.2 (2017), pp. 5165–5197.

[AL17] Marc Abeille and Alessandro Lazaric. "Thompson sampling for linear-quadratic control problems". In: *arXiv preprint arXiv:1703.08972* (2017).

[ACF02] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. "Finite-time analysis of the multiarmed bandit problem". In: *Machine learning* 47.2-3 (2002), pp. 235–256.

[BV04] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[Bro86] Lawrence D Brown. "Fundamentals of statistical exponential families: with applications in statistical decision theory". In: Ims. 1986.

[Bub+15] Sébastien Bubeck et al. "Convex optimization: Algorithms and complexity". In: *Foundations and Trends® in Machine Learning* 8.3-4 (2015), pp. 231–357.

[Che72] Herman Chernoff. *Sequential Analysis and Optimal Design*. Society for Industrial and Applied Mathematics, 1972.

[Fau+20] Louis Faury et al. *Improved Optimistic Algorithms for Logistic Bandits*. 2020. arXiv: `2002.07530 [cs.LG]`.

[Fau+22] Louis Faury et al. "Jointly Efficient and Optimal Algorithms for Logistic Bandits". In: *AISTATS*. 2022.

[Fie+19] Tanner Fiez et al. "Sequential Experimental Design for Transductive Linear Bandits". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/8ba6c657b03fc7c8dd4dff8e45defcd2-Paper.pdf.

[Fil+10] Sarah Filippi et al. "Parametric bandits: The generalized linear case". In: *Advances in Neural Information Processing Systems* 23 (2010).

[Fos+21] Dylan J Foster et al. "The statistical complexity of interactive decision making". In: *arXiv preprint arXiv:2112.13487* (2021).

[FGH23] Dylan J. Foster, Noah Golowich, and Yanjun Han. *Tight Guarantees for Interactive Decision Making with the Decision-Estimation Coefficient*. 2023. arXiv: 2301.08215 [cs.LG].

[HB20] Nima Hamidi and Mohsen Bayati. *On Worst-case Regret of Linear Thompson Sampling*. 2020. arXiv: 2006.06790 [cs.LG].

[Han57] James Hannan. "Approximation to Bayes risk in repeated play". In: *Contributions to the Theory of Games*. Vol. 3. Princeton, NJ: Princeton University Press, 1957, pp. 97–140.

[Ish+21] Haque Ishfaq et al. "Randomized Exploration in Reinforcement Learning with General Value Function Approximation". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 4607–4616. URL: https://proceedings.mlr.press/v139/ishfaq21a.html.

[JOA10] Thomas Jaksch, Ronald Ortner, and Peter Auer. "Near-optimal regret bounds for reinforcement learning". In: *Journal of Machine Learning Research* 11.Apr (2010), pp. 1563–1600.

[Jun+17] Kwang-Sung Jun et al. "Scalable Generalized Linear Bandits: Online Computation and Hashing". In: *NIPS*. 2017.

[Jun+21] Kwang-Sung Jun et al. "Improved Confidence Bounds for the Linear Logistic Model and Applications to Bandits". In: *ICML*. 2021.

[Kak+10] Sham Kakade et al. "Learning Exponential Families in High-Dimensions: Strong Convexity and Sparsity". In: *Artificial Intelligence and Statistics*. 2010.

[KV05] Adam Kalai and Santosh Vempala. "Efficient algorithms for online decision problems". In: *Journal of Computer and System Sciences* 71.3 (2005), pp. 291–307.

[Khu03] André I Khuri. *Advanced calculus with applications in statistics*. John Wiley & Sons, 2003.

[Kve+19a] Branislav Kveton et al. “Garbage In, Reward Out: Bootstrapping Exploration in Multi-Armed Bandits”. In: *Proceedings of the 36th International Conference on Machine Learning.* 2019, pp. 3601–3610.

[Kve+19b] Branislav Kveton et al. “Perturbed-History Exploration in Stochastic Multi-Armed Bandits”. In: *IJCAI.* 2019.

[Kve+19c] Branislav Kveton et al. *Randomized Exploration in Generalized Linear Bandits.* 2019. arXiv: `1906.08947 [cs.LG]`.

[Kve+20] Branislav Kveton et al. “Perturbed-History Exploration in Stochastic Linear Bandits”. In: *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference.* Vol. 115. Proceedings of Machine Learning Research. PMLR, 22–25 Jul 2020, pp. 530–540.

[LR85] T.L Lai and Herbert Robbins. “Asymptotically Efficient Adaptive Allocation Rules”. In: *Adv. Appl. Math.* 6.1 (Mar. 1985), pp. 4–22. ISSN: 0196-8858. DOI: `10.1016/0196-8858(85)90002-8`. URL: `https://doi.org/10.1016/0196-8858(85)90002-8`.

[LS18] Tor Lattimore and Csaba Szepesvári. “Bandit algorithms”. In: *preprint* (2018), p. 28.

[LS20] Tor Lattimore and Csaba Szepesvári. “Exploration by Optimisation in Partial Monitoring”. In: *COLT.* June 2020.

[LLZ17] 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.* Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 2071–2080. URL: `http://proceedings.mlr.press/v70/li17c.html`.

[LR17] Xiuyuan Lu and Benjamin Van Roy. “Ensemble sampling”. In: *Advances in Neural Information Processing Systems.* 2017.

[McC19] Peter McCullagh. *Generalized linear models.* Routledge, 2019.

[OAC18] I. Osband, J. Aslanides, and A. Cassirer. “Randomized prior functions for deep reinforcement learning”. In: *Advances in Neural Information Processing Systems.* 2018.

[OV15] Ian Osband and Benjamin Van Roy. “Bootstrapped Thompson sampling and deep exploration”. In: *arXiv preprint arXiv:1507.00300* (2015).

[OVW14] Ian Osband, Benjamin Van Roy, and Zheng Wen. *Generalization and Exploration via Randomized Value Functions.* 2014. DOI: `10.48550/ARXIV.1402.0635`. URL: `https://arxiv.org/abs/1402.0635`.

[Osb+16] Ian Osband et al. “Deep exploration via bootstrapped DQN”. In: *Advances in Neural Information Processing Systems.* 2016.

[Osb+17] Ian Osband et al. “Deep Exploration via Randomized Value Functions”. In: *CoRR* abs/1703.07608 (2017). arXiv: `1703.07608`. URL: `http://arxiv.org/abs/1703.07608`.

[PAD19] My Phan, Yasin Abbasi-Yadkori, and Justin Domke. “Thompson Sampling and Approximate Inference”. In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

[Puk06] Friedrich Pukelsheim. *Optimal Design of Experiments (Classics in Applied Mathematics) (Classics in Applied Mathematics, 50)*. USA: Society for Industrial and Applied Mathematics, 2006. ISBN: 0898716047.

[Qin+22] Chao Qin et al. “An Analysis of Ensemble Sampling”. In: *Advances in Neural Information Processing Systems*. 2022.

[Rig12] Philippe Rigollet. “Kullback–Leibler aggregation and misspecified generalized linear models”. In: (2012).

[Rus+20] Yoan Russac et al. “Self-Concordant Analysis of Generalized Linear Bandits with Forgetting”. In: *ArXiv* abs/2011.00819 (2020).

[Rus19] Daniel Russo. “Worst-case regret bounds for exploration via randomized value functions”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 14410–14420.

[RV13] Daniel Russo and Benjamin Van Roy. “Eluder Dimension and the Sample Complexity of Optimistic Exploration”. In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. Curran Associates, Inc., 2013. URL: `https://proceedings.neurips.cc/paper/2013/file/41bfd20a38bb1b0bec75acf0845530a7-Paper.pdf`.

[RV18] Daniel Russo and Benjamin Van Roy. “Learning to optimize via information-directed sampling”. In: *Operations Research* 66.1 (2018), pp. 230–252.

[SS12] Warren Schudy and Maxim Sviridenko. “Concentration and moment inequalities for polynomials of independent random variables”. In: *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*. SIAM. 2012, pp. 437–446.

[SLM14] Marta Soare, Alessandro Lazaric, and Rémi Munos. “Best-Arm Identification in Linear Bandits”. In: *ArXiv* abs/1409.6110 (2014).

[ST17] Tianxiao Sun and Quoc Tran-Dinh. “Generalized self-concordant functions: a recipe for Newton-type methods”. In: *Mathematical Programming* (2017), pp. 1–69.

[TC60] Donald L. Thistlewaite and Donald T. Campbell. “Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment”. In: *Observational Studies* 3 (1960), pp. 119–128.

[Tho33] William R. Thompson. "On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples". In: *Biometrika* 25.3/4 (1933), pp. 285–294. ISSN: 00063444. URL: `http://www.jstor.org/stable/2332286` (visited on 09/30/2022).

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.