Exploration via linearly perturbed loss minimisation

David Janz[⋆] University of Alberta
Shuai Liu[⋆] University of Alberta
Alex Ayoub[⋆] University of Alberta
Csaba Szepesvári[⋆] University of Alberta

Abstract

We introduce *exploration via linear loss perturbations* (EVILL), a randomised exploration method for structured stochastic bandit problems that works by solving for the minimiser of a linearly perturbed regularised negative log-likelihood function. We show that, for the case of generalised linear bandits, EVILL reduces to perturbed history exploration (PHE), a method where exploration is done by training on randomly perturbed rewards. In doing so, we provide a simple and clean explanation of when and why random reward perturbations give rise to good bandit algorithms. We propose data-dependent perturbations not present in previous PHE-type methods that allow EVILL to match the performance of Thompson-sampling-style parameter-perturbation methods, both in theory and in practice. Moreover, we show an example outside generalised linear bandits where PHE leads to inconsistent estimates, and thus linear regret, while EVILL remains performant. Like PHE, EVILL can be implemented in just a few lines of code.

1 INTRODUCTION

Effective exploration is key to the success of algorithms that learn to optimise long term reward while interacting with their environments (Lattimore and Szepesvári, 2020). Algorithms based on *perturbed history exploration* (PHE) follow the optimal policy given a model fitted to randomly perturbed data. The appeal of PHE is that, alike to the widely used ϵ-greedy strategy, it is *minimally invasive*: PHE can be implemented by adding a few extra lines of code on top of that used to fit models and optimise policies against fitted models.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

The challenge in using PHE is to design perturbations that induce sufficient exploration while controlling the computational cost. Two strategies have been proposed for this purpose. First, Kveton et al. (2019, 2020a) proposed to add a fixed number of copies of each observation, but with the rewards replaced by some randomly chosen ones. An alternative, which avoids increasing the size of the data, is to *additively perturb the observed rewards* (Kveton et al., 2020b). Kveton et al. motive this latter approach—a special case of randomised least-squares value iteration (Osband et al., 2016)—by 'the equivalence of posterior sampling and perturbations by Gaussian noise in linear models (Lu and Van Roy, 2017), when the prior of $\theta_\star$ and rewards are Gaussian'. In this paper, we follow up on the work of Kveton et al. (2020b) and ask when and why will additive reward perturbations induce the right amount of exploration in structured, nonlinear bandits.

The first difficulty that one encounters when attempting to answer this question is best illustrated by considering *logistic bandit problems*, where the rewards associated with the individual arms are binary-valued. Then, when real-valued rewards are used in place of binary rewards—which is what happens when the observed binary rewards are additively perturbed by Gaussian noise—the log-likelihood associated with the data becomes *undefined*. Kveton et al. (2020b) note this, but also observe that the term in the loss that causes the problem is constant (as a function of the unknown parameter to be estimated), and hence can be dropped without changing the estimate. While convenient, this critically relies on the properties of logistic bandits, namely that the reward distribution is a member of the natural exponential family of distributions. For more general problems, additive reward perturbations may lead to algorithms that are not well-defined.

Moreover, even when additive reward perturbations lead to a well-defined method, it is unclear why these are reasonable. While in the case of linear bandits, additive reward perturbations relate closely to posterior sampling, which has a firm theoretical basis (Agrawal

⋆`{djanz,shuai14,aayoub,szepesva}@ualberta.ca`

and Goyal, 2013; Abeille and Lazaric, 2017), no such justification exists for, say, generalised linear bandits (GLBs). Indeed, for GLBs, the only theoretical result for PHE, that of Kveton et al. (2020b), relies on the troublesome assumption that the arms of the bandit are orthogonal—an assumption under which the problem becomes equivalent to that of learning and acting in an unstructured multi-armed bandit. Thus, Kveton et al. (2020b) pose as an open problem the construction and analysis of an appropriate PHE-style data-perturbation method for structured, nonlinear bandit problems.

Our work both resolves the above two challenges, and provides a new view on additive reward perturbations. We take a step back, and design a new algorithm, *exploration via linear loss perturbations* (EVILL), which minimises loss functions that are perturbed by adding random linear components. We prove that this perturbed loss is *equivalent* to the loss used by Kveton et al. (2020b) in *generalised linear bandits*, but is applicable even when PHE with additive perturbations is not well-defined. We justify our linear perturbations by considering a quadratic approximation to the negative log-likelihood, and show how to set these in a way such that the minimiser of the perturbed loss gives rise to an optimistic model with constant probability, a property that is widely recognised as key to the design of effective random exploration methods for stochastic environments (Agrawal and Goyal, 2013; Abeille and Lazaric, 2017; Lattimore and Szepesvári, 2020).

In contrast with the PHE method for GLBs of Kveton et al. (2020b), the perturbations induced by EVILL have a data-dependent variance that scales with an estimate of the Fisher information of the reward distribution. This scaling arises naturally by considering a quadratic approximation to the loss, and similar scaling was used in previous works that approximate Thompson sampling (e.g., Kveton et al., 2020b). Our experiments show that these data-dependent perturbations lead to a significant boost in performance, and demonstrate a specific bandit problem outside the class of GLBs where PHE with additive data perturbations leads to *inconsistent* estimates and *linear regret*, while the estimates of EVILL are consistent and the algorithm is performant.

To summarise, we extend and strengthen the literature on randomised exploration in stochastic bandits as follows:

- We propose a new way of inducing exploration, EVILL, which adds a random linear term to the model-fitting loss, chosen to induce optimism.
- We show that EVILL is equivalent to PHE with additive perturbations in the generalised linear bandit setting, but is also applicable in settings where PHE is not well-defined or is inconsistent.
- We propose that the equivalence of EVILL and PHE is the best way to understand when and why additive reward perturbations can be successful.
- We establish that, in self-concordant generalised linear bandits, the regret of EVILL enjoys guarantees that parallel those available for Thompson sampling. Due to the equivalence of EVILL and PHE, this resolves the open problem of giving a rigorous theoretical justification for PHE with additive reward perturbations.
- We establish experimentally that our data-dependent perturbations are effective and that EVILL is competitive with alternatives.

In addition, we establish that the concentration results of Abeille et al. (2021) and Russac et al. (2021) hold in all self-concordant generalised linear bandits, removing a previous assumption of bounded responses.

2 PROBLEM SETTING

In this section we introduce notation, and describe the bandit problems we consider. These will, for now, be quite general, which facilitates the description of our algorithm and its relation to PHE. Our upcoming theoretical results will consider the narrower setting, that of generalised linear bandits equipped with some extra technical assumptions, described in Section 4.

Notation Throughout, $\|\cdot\|$ will denote the Euclidean 2-norm. For a positive semidefinite matrix M and a vector v of compatible dimensions, $\|v\|_M$ will denote the M-weighted Euclidean 2-norm given by $\|v\|_M^2 = v^\top M v$. For any positive integer $n \in \mathbb{N}^+$, $[n]$ denotes the integers $\{1, \dots, n\}$. We use $\mathbb{R}$ to denote the set of reals. We let $\operatorname{diag}(a_1, \dots, a_d)$ stand for the $d \times d$ diagonal matrix with diagonal entries $a_1, \dots, a_d$. We use I_d to denote the $d \times d$ identity matrix. We use upper-case letters to denote random variables, which are defined over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with expectation operator $\mathbb{E}$. We write $\operatorname{supp}(Q)$ to denote the support of a distribution Q. For a real-valued differentiable map f on a suitable subset of $\mathbb{R}^d$, we use $\dot{f}$ ($\ddot{f}$, $\dddot{f}$, ...) to denote the first (respectively, second, third, etc.) derivative of f, and view $\dot{f}$ as a column vector. We denote the interior of a set $\mathcal{U} \subset \mathbb{R}$ by $\mathcal{U}^\circ$.

Bandits & regret We consider bandit problems described by a non-empty (possibly infinite) arm set $\mathcal{X}$, a parameter vector $\theta_\star \in \Theta \subset \mathbb{R}^d$, a family of smooth *localisation maps* $(\phi_x)_{x\in\mathcal{X}}$ with $\phi_x : \Theta \to \mathbb{R}$, and a family of distributions $(P(\cdot;u))_{u\in\mathcal{U}}$ over the reals, where $\mathcal{U} \subset \mathbb{R}$ is a (possibly infinite) interval of real numbers with non-empty interior such that for all $x \in \mathcal{X}$, $\phi_x(\theta_\star) \in \mathcal{U}$. We will use $\mu(u)$ to denote the mean of $P(\cdot\,;u)$:

$$\mu(u) = \int y\ P(dy;u)\,.$$

We assume that the learner is given the family $(P(\cdot\,;u))_{u\in\mathcal{U}}$, the arm set $\mathcal{X}$, the family of maps $(\phi_x)_{x\in\mathcal{X}}$, but not the parameter $\theta_\star$. In each round $t \in \{1,2,\dots\}$ of interaction, the learner selects an action $X_t \in \mathcal{X}$ based on the available past information. In response, the learner receives a random reward $Y_t \in \mathbb{R}$, which they observe. The distribution of Y_t, given the past, and in particular X_t, is $P(\cdot\,;\phi_{X_t}(\theta_\star))$. The goal of the learner is to maximise its mean reward. The performance of the learner is measured by their (pseudo)regret, which, for an interaction that lasts for n rounds, is given

$$R(n) = \sum_{t=1}^{n} \mu(\phi_{x_\star}(\theta_\star)) - \mu(\phi_{X_t}(\theta_\star))\,,$$

where $x_\star$ is any arm in $\arg\max_{x\in\mathcal{X}} \mu(\phi_x(\theta_\star))$. For simplicity, we assume the existence of the maximising arm $x_\star$.

To make the above concrete, consider how the generalised linear bandits fit into our framework:

Example 1 (Generalised linear bandits (Filippi et al., 2010))**.** *Generalised linear bandits (GLBs) are obtained by choosing* $(P(\cdot\,;u))_{u\in\mathcal{U}}$ *to be a natural exponential family of distributions (for more on these, see Section 4.1),* $\mathcal{X} \subset \mathbb{R}^d$, $\Theta = \mathbb{R}^d$, *and* $\phi_x(\theta) = x^\top\theta$.

Regularised negative log-likelihood The algorithms we design for the learner will rely on regularised maximum likelihood estimation, or, equivalently, on minimising a regularised negative log-likelihood loss.

To define the likelihood function, assume that each distribution in the family $(P(\cdot;u))_{u\in\mathcal{U}}$ is dominated by a common σ-finite measure ν over the reals. This allows us to consider the density of $P(\cdot;u)$ with respect to ν, which we will denote by $p(\cdot;u)$; that is, $p(y;u) = \frac{dP(\cdot;u)}{d\nu}$. Then, given a list of pairs $\mathcal{D} = ((x_1,y_1),\dots,(x_s,y_s))$ of length s where $x_i \in \mathcal{X}$ and $y_i \in \mathbb{R}$, and a regularisation parameter $\lambda > 0$, the λ-regularised *negative* log-likelihood function is defined to be

$$\mathcal{L}(\theta;\mathcal{D}) = \frac{\lambda}{2}\|\theta\|^2 - \sum_{i=1}^{s} \log p(y_i;\phi_{x_i}(\theta))\,. \qquad (1)$$

We assume throughout that the minimisation of $\theta \mapsto \mathcal{L}(\theta;\mathcal{D}) + w^\top\theta$ is efficient for data that can be generated by the model and any arbitrary weight vector $w \in \mathbb{R}^d$. This holds, for example, for generalised linear bandits.

Examples To wrap up our setting section, we give two more examples of suitable bandit problems.

Example 2 (Logistic linear bandits)**.** *This is a special case of generalised linear bandits when* $\mathcal{U} = \mathbb{R}$ *and* $P(\cdot\,;u)$ *is the Bernoulli distribution with mean*

$$\mu(u) = \exp(u)/(1+\exp(u))\,.$$

Then, picking ν *to be the counting measure over* $\{0,1\}$, *we have the density*

$$p(y;u) = (\mu(u))^y(1-\mu(u))^{1-y}$$

for $y \in \{0,1\}$.

Example 3 (Linear bandits with Gaussian rewards)**.** *Another special case of generalised linear bandits is when* $\mathcal{U} = \mathbb{R}$ *and* $P(\cdot\,;u)$ *is a Gaussian distribution with (say) variance one and mean* $\mu(u) = u$. *Here, taking* ν *to be the Lebesgue measure over the reals, we have a density*

$$p(y;u) = \exp(-(y-u)^2/2)/\sqrt{2\pi}\,.$$

We will give one further example in Section 4.2, that of a Rayleigh linear bandit, on which the previous PHE approaches lead to inconsistent estimation and linear regret. Rayleigh linear bandits fall outside GLBs.

3 THE EVILL ALGORITHM

Consider the EVILL algorithm, listed in Algorithm 1. EVILL takes as arguments a horizon length $n \in \mathbb{N}^+$, a regularisation parameter $\lambda > 0$, a perturbation scale parameter $a > 0$ and prior observations $\mathcal{D}_\tau = ((X_i,Y_i))_{i=1}^{\tau}$ of length $\tau \le n$, where τ can be a stopping time (i.e., chosen in a data-dependent way). In addition, the algorithm needs to have access to the maps $(\phi_x)_{x\in\mathcal{X}}$, the arm-set $\mathcal{X}$, a feasible region for the parameters $\Theta' \subset \mathbb{R}^d$, the Fisher-information map $I : \mathcal{U}^\circ \to [0,\infty)$, and the mean maps $\mu : \mathcal{U} \to [0,\infty)$ underlying the single-parameter reward distribution family $(P(\cdot;u))_{u\in\mathcal{U}}$. We define the Fisher information map shortly, and discuss its importance in Section 3.1.

The role of prior observations $\mathcal{D}_\tau$ is to ensure that, right from the start, for all arms $x \in \mathcal{X}$, $\phi_x(\theta_\star)$ is estimated up to a constant accuracy by $\phi_x(\hat\theta_\tau)$, where $\hat\theta_\tau \in \arg\min_{\theta\in\Theta'} \mathcal{L}(\theta;\mathcal{D}_\tau)$. The prior observations may be offline data, already available before any interaction occurs, or may be collected using a standard warm-up procedure, as discussed in Appendix B.

With the prior observations in-place, at each round $t \in \{\tau+1,\dots,n\}$, EVILL first finds $\hat\theta_{t-1}$, the minimiser

Algorithm 1 Exploration via linear loss perturbations (EVILL)

Require: horizon $n \in \mathbb{N}^+$, arm-set $\mathcal{X}$, a feasible parameter set $\Theta' \subset \mathbb{R}^d$, localisation maps $(\phi_x)_{x\in\mathcal{X}}$, Fisher information map $I \colon \mathcal{U}^\circ \to [0,\infty)$ and mean map $\mu \colon \mathcal{U} \to \mathbb{R}$, regularisation parameter $\lambda > 0$, perturbation scale $a > 0$, prior observations $((X_i, Y_i))_{i=1}^{\tau}$ of length $\tau \in [n]$

for all $t \in \{\tau+1, \dots, n\}$ **do**
- Compute $\hat{\theta}_{t-1} \in \arg\min_{\theta\in\Theta'} \mathcal{L}(\theta; ((X_i, Y_i))_{i=1}^{t-1})$ ▷ *MLE of $\theta_\star$*
- Sample perturbations $Z_t \sim \mathcal{N}(0, I_d)$, $Z_t' \sim \mathcal{N}(0, I_{t-1})$
- Compute vector $W_t = a\lambda^{1/2} Z_t + a \sum_{i=1}^{t-1} I(\phi_{X_i}(\hat{\theta}_{t-1}))^{1/2} Z'_{t,i} \dot{\phi}_{X_i}(\hat{\theta}_{t-1})$
- Compute $\theta_t \in \arg\min_{\theta\in\Theta'} \mathcal{L}(\theta; ((X_i, Y_i))_{i=1}^{t-1}) + W_t^\top \theta$ ▷ *Optimise the linearly perturbed loss*
- Select arm $X_t \in \arg\max_{x\in\mathcal{X}} \mu(\phi_x(\theta_t))$ and receive reward Y_t

of $\mathcal{L}(\cdot; \mathcal{D}_{t-1})$, where $\mathcal{D}_{t-1} = ((X_i, Y_i))_{i=1}^{t-1}$ collects all past data. That is, we find

$$\hat{\theta}_{t-1} \in \arg\min_{\theta\in\Theta'} \mathcal{L}(\cdot; \mathcal{D}_{t-1}) \,.$$

This preliminary estimate $\hat{\theta}_{t-1}$ then is used to construct a random perturbation vector, W_t, given by

$$W_t = a\lambda^{1/2} Z_t + a \sum_{i=1}^{t-1} I(\phi_{X_i}(\hat{\theta}_{t-1}))^{1/2} Z'_{t,i} \dot{\phi}_{X_i}(\hat{\theta}_{t-1}) \,,$$

where $Z_t \sim \mathcal{N}(0, I_d)$, $Z_t' \sim \mathcal{N}(0, I_{t-1})$, $\dot{\phi}_x$ is the derivative of ϕ_x and $I : \mathcal{U}^\circ \to [0,\infty)$ is the Fisher-information map underlying $(P(\cdot; u))_{u\in\mathcal{U}}$, given by

$$I(u) = \int \frac{\partial^2}{\partial u^2} \log \frac{1}{p(y;u)} P(dy; u) \,. \tag{2}$$

Next, the loss minimiser is invoked on the loss perturbed by $W_t^\top \theta$, to compute

$$\theta_t \in \arg\min_{\theta\in\Theta'} \mathcal{L}(\theta; \mathcal{D}_{t-1}) + W_t^\top \theta \,,$$

and θ_t is used to select the action X_t as

$$X_t \in \arg\max_{x\in\mathcal{X}} \mu(\phi_x(\theta_t)) \,. \tag{3}$$

Again, merely to simplify the exposition, we assume that the minimisers and this latter maximiser exist.

If μ is increasing, which holds in natural exponential families, the choice in Equation (3) is the same as $X_t \in \arg\max_{x\in\mathcal{X}} \phi_x(\theta_t)$. When ϕ_x is the bilinear map as in generalised linear models, this simplifies to $X_t \in \arg\max_{x\in\mathcal{X}} x^\top \theta_t$, while $\dot{\phi}_{X_i}(\hat{\theta}_{t-1})$ (used in the construction of W_t) simplifies to X_i.

In generalised linear bandits, under mild regularity conditions and when $x^\top \theta_*$ belongs to the interior of $\mathcal{U}$ for every $x \in \mathcal{X}$, $\hat{\theta}_t$ and θ_t are guaranteed to be such that $x^\top \hat{\theta}_t$ and $x^\top \theta_t$ both belong to the interior of $\mathcal{U}$.[1] Otherwise, this can always be achieved by modifying $-\log p(y; u)$ to return infinity for u not in $\mathcal{U}$.

3.1 Why do linear perturbations work?

We now motivate our choice of linear perturbations. Our starting point is the observation that to guarantee the success of algorithms that choose actions greedily with respect to a model with a randomised parameter, it suffices that the random parameter is *optimistic* in the sense that the mean reward for the best arm under the random parameter, say θ_t, exceeds that under the true model parameter $\theta_\star$, with some probability bounded away from zero in a uniform manner (for intuition, see Chapters 7 and 36 in Lattimore and Szepesvári, 2020).

A general approach to guarantee to achieve optimism is to sample the random parameters from a distribution that is nearly uniform over a set of the form

$$C(\Delta) = \{\theta \in \Theta \,:\, \ell_{t-1}(\theta) \le \ell_{t-1}(\hat{\theta}_{t-1}) + \Delta\} \,,$$

where $\Delta > 0$ is a tuning parameter (think 'suitably large constant') and ℓ_{t-1} is the unregularised negative log-likelihood function corresponding to the observations available at the beginning of round t. For the purposes of this section, for simplicity, we set $\lambda = 0$ and consider the case when $\Theta = \mathbb{R}^d$.

Considering using a quadratic approximation to the negative log-likelihood function, we obtain the set

$$\widetilde{C}(\Delta) = \{\theta \in \Theta \,:\, \frac{1}{2} \|\theta - \hat{\theta}_{t-1}\|^2_{F_{t-1}(\hat{\theta}_{t-1})} \le \Delta\} \,,$$

where

$$F_{t-1}(\theta) = \sum_{i=1}^{t-1} I(\phi_{X_i}(\theta)) \dot{\phi}_{X_i}(\theta) \dot{\phi}_{X_i}^\top(\theta) \tag{4}$$

[1] To illustrate why regularity conditions are needed, we borrow an example from page 153 of the book by Barndorff-Nielsen (2014). Letting Q be a Pareto distribution with shape parameter $\alpha > 1$ and scale parameter of one and $\mathcal{X} = \{1\}$, we have $U_Q = (-\infty, 0]$, ψ is left-differentiable at 0 with a finite derivative of $1 + 1/(\alpha - 1)$, while $\psi(0) = 0$. One can then show that if the average of observed values exceeds $1 + 1/(\alpha - 1)$, the maximum likelihood estimate of $u = \theta$ is $0 \in \partial U_Q$.

is an approximation to the curvature (second derivative) of $\ell_{t-1}(\theta)$, and corresponds to the Fisher-information underlying the parametric model $(P(dy;\phi_x(\theta)))_{\theta\in\Theta}$ with the design $X_1,\dots,X_{t-1}$. This approximation is based on the following observation.

Proposition 1. *Let $\ell(x,y;\theta) = -\log p(y;\phi_x(\theta))$ and $\ell(x;\theta) = \int \ell(x,y;\theta)P(dy;\phi_x(\theta_\star))$. Then, under suitable regularity assumptions,*

$$\frac{d^2}{d\theta^2}\ell(x,\theta_\star) = I(\phi_x(\theta_\star))\,\dot\phi_x(\theta_\star)\dot\phi_x^\top(\theta_\star)\,.$$

We now argue that the said sampling goal can be accomplished by minimising a linearly perturbed negative log-likelihood function. For this let $g_i = \dot\phi_{x_i}(\hat\theta_{t-1})$ and

$$W = a\sum_{i=1}^{t-1} Z'_{t,i-1} I(\phi_{X_i}(\hat\theta_{t-1}))^{1/2} g_i\,,$$

where $Z'_{t,1},\dots,Z'_{t,t-1}$ are zero mean random variables that are independent of each other and of the past. Let $\theta_t = \arg\min_\theta \ell_{t-1}(\theta) - W^\top\theta$ and $F = F_{t-1}(\hat\theta_{t-1})$. Using the first-order optimality condition and if we also replace the gradient of ℓ_{t-1} with its local quadratic approximation, $\tilde\ell_{t-1}(\theta) = \ell_{t-1}(\hat\theta_{t-1}) + \frac{1}{2}\|\theta-\hat\theta_{t-1}\|_F^2$, we get that

$$F^{1/2}(\theta_t - \hat\theta_{t-1}) \approx F^{-1/2}W\,.$$

Now, note that $W\sim\mathcal{N}(0,a^2F)$, hence $F^{-1/2}W\sim\mathcal{N}(0,a^2I_d)$. It follows that, as long as the approximations used are precise enough, given the past, the distribution of $F^{1/2}(\theta_t-\hat\theta_{t-1})$ is close to $\mathcal{N}(0,a^2I_d)$, hence, for a sufficiently large, θ_t is approximately uniform over $C(\Delta)$. Examining the perturbations in Algorithm 1, we see these feature an additional term $a\lambda^{1/2}Z_t$ where $Z_t\sim\mathcal{N}(0,I_d)$—this accounts for the effect of regularisation, which we omitted here.

The linear perturbation of the loss used by EVILL is appealing from an implementation point of view: to implement EVILL, one only needs to be able to construct a random vector W; since we expect that existing libraries that fit models can also deal with losses with the extra linear term, the implementation of EVILL should take only a few extra lines of code.

We provide an in-depth discussion of the computational complexity of EVILL in Appendix A.

3.2 Relation to Thompson sampling

From the description of the previous section, it is clear that EVILL is a close relative of direct parameter randomisation methods, such as Thompson (posterior) sampling. To make this concrete, fix some (prior) distribution π_0 over Θ and let p_{t-1} denote the posterior corresponding to π_0 and the likelihood functions $\theta\mapsto p(Y_i;\psi_{X_i}(\theta))$ with $i\in\{1,\dots,t-1\}$. Thus,

$$p_{t-1}(d\theta)\propto\pi_0(d\theta)\exp(-\mathcal{L}(\theta;\mathcal{D}_{t-1}))\,.$$

Now note that for as long as θ is 'close' to $\hat\theta_{t-1}$, $\mathcal{L}(\theta;\mathcal{D}_{t-1})\approx\mathcal{L}(\hat\theta_{t-1};\mathcal{D}_{t-1})+\frac{1}{2}\|\theta-\hat\theta_{t-1}\|_F^2 =: \tilde\ell_{t-1}(\theta)$, where F is as above. For reasons that will become clear in a moment, let $\pi_0=\mathcal{N}(0,\lambda^{-1}I)$. Now, if most of the probability mass of p_{t-1} concentrates in a small neighbourhood of $\hat\theta_{t-1}$, we get that the Gaussian distribution

$$\tilde p_{t-1}(d\theta)\propto\exp\left(-\frac{\lambda}{2}\|\theta\|^2-\tilde\ell_{t-1}(\theta)\right)d\theta$$

is a good approximation to p_{t-1}. The distribution $\tilde p_{t-1}$, which is based on the best local quadratic fit to the log-density of p_{t-1} around the mode $\hat\theta_{t-1}$ of p_{t-1}, is known as the *Laplace approximation* to p_{t-1}.

Observe that both $\tilde p_{t-1}$ and the sampling distribution induced by EVILL are based on a quadratic approximation to the negative log-likelihood. In fact, sampling from the Laplace approximation $\tilde p_{t-1}$ can also be implemented using a 'sample-then-optimise' method (Papandreou and Yuille, 2010; Antoran et al., 2022), which can be shown to coincide with choosing the minimiser of $\tilde\ell_{t-1}(\theta)-W^\top\theta$; meanwhile, EVILL chooses the minimiser of $\ell_{t-1}(\theta)-W^\top\theta$.

Sampling from the Laplace approximation to the posterior has been used in many prior works, including Chapelle and Li (2011), Russo et al. (2018), Abeille and Lazaric (2017), and Kveton et al. (2020b), and is generally a computationally efficient approach. Yet, because the matrix F changes with the parameter estimates, naïve approaches to sampling from the Laplace approximation scale poorly when either d or t is large—efficient implementations use the 'doubling trick' or the aforementioned sample-then-optimise approach. In contrast, EVILL is a simple to implement alternative.

4 RESULTS FOR GENERALISED LINEAR BANDITS

In this section we consider EVILL in generalised linear bandits (GLBs) as defined in Example 1. We start with recalling the definition natural exponential families (NEFs). Next we show that when the reward distributions are given by a natural exponential family of distributions, EVILL and PHE with additive data perturbations are equivalent (Proposition 2), but that the equivalence may break outside this setting; indeed, we show that on a slightly enlarged class of bandits, PHE introduces a 'bias' that may lead to linear regret, which is avoided by EVILL. We then introduce our main theoretical result, Theorem 1, which gives a guarantee on the regret of EVILL that is comparable to the existing results for Thompson sampling based approaches. This puts EVILL—and thus, in the special case of generalised linear bandits, PHE—on equal footing with Thompson sampling. We end by highlighting a technical result, Lemma 1, that states that members of a self-concordant NEF distribution are sub-exponential—this allows us to remove the boundedness condition from the prior works of Abeille et al. (2021) and Russac et al. (2021).

4.1 Natural exponential families

For any probability distribution Q over the reals and $u \in \mathbb{R}$, we let $\psi_Q(u) = \log \int e^{uy} Q(dy)$ be the *cumulant-generating function* of Q, and let $\mathcal{U}_Q \subset \mathbb{R}$ be the largest set on which $\psi_Q(u)$ is finite. The natural exponential family generated by a distribution Q is the family $(Q_u(dy))_{u \in \mathcal{U}_Q}$ of probability measures on $\mathbb{R}$ given by

$$Q_u(dy) = \exp(yu - \psi(u))Q(dy) \quad \text{for} \quad u \in \mathcal{U}_Q\,.$$

As is well known, $\mathcal{U}_Q$ is always an interval. We only consider *regular* families, that is, families where $\mathcal{U}_Q^\circ$, the interior of $\mathcal{U}_Q$, is not empty.

Our earlier examples, that is, the Bernoulli and Gaussian distributions are examples of regular families with $\mathcal{U} = \mathcal{U}_Q = \mathbb{R}$. The Bernoulli distribution corresponds to the base distribution that is the uniform distribution on $\{0, 1\}$, while the Gaussian distribution corresponds to Q chosen as the standard normal distribution. The latter might be considered a somewhat unusual choice (as opposed to choosing Q to be Lebesgue measure), though this makes no difference.

4.2 The equivalence of EVILL and PHE

In this section we show that in generalised linear bandits, EVILL reduces to PHE with additive, data-dependent perturbations, and demonstrate how this reduction can fail outside that setting. For the first part, we work under the following assumption.

Assumption 1 (Generalised linear bandit)**.** *The family $(P(\cdot; u))_u$ that determines the rewards in our bandit instances is a natural exponential family generated by some base Q. That is,*

$$P(\cdot; u) = Q_u(\cdot) \quad \textit{for} \quad u \in \mathcal{U}\,,$$

where $\mathcal{U} \subset \mathcal{U}_Q^\circ$ is an interval with non-empty interior. Furthermore, we assume that the localisation maps are linear; that is, $\phi_x(\theta) = x^\top\theta$ for each $x \in \mathcal{X}$ and $\theta \in \Theta$.

We now show how the PHE algorithm of Kveton et al. (2020b) may be understood in terms of linear extensions of the likelihood function. Under Assumption 1, and choosing base measure $\nu = Q$, we get the density $p(y; u) = \exp(yu - \psi(u))$. Recalling the notation $\ell(y; u) = -\log p(y; u)$, this gives

$$\ell(y; u) = \begin{cases} -yu + \psi(u)\,, & \text{if } y \in \operatorname{supp}(Q)\,; \\ +\infty\,, & \text{otherwise}\,. \end{cases}$$

Now consider the linear extension of ℓ to all of $\mathbb{R}$:

$$\tilde{\ell}(y; u) = -yu + \psi(u)\,, \qquad y \in \mathbb{R}\,.$$

Using this function, for any $y \in \operatorname{supp}(Q)$ and any $z \in \mathbb{R}$, we have

$$\tilde{\ell}(y + z; u) = \ell(y; u) - zu\,. \tag{5}$$

For $\mathcal{D} = ((x_1, y_1), \dots, (x_s, y_s))$, $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$, let

$$\tilde{\mathcal{L}}(\theta; \mathcal{D}) = \sum_{i=1}^{s} \tilde{\ell}(y_i; \phi_{x_i}(\theta)) + \frac{\lambda}{2}\|\theta\|^2\,.$$

Assume also that $y_1, \dots, y_s \in \operatorname{supp}(Q)$. For $z \in \mathbb{R}^s$ let

$$\mathcal{D}^z = ((x_1, y_1 + z_1), \dots, (x_s, y_s + z_s))\,.$$

Then, from Equation (5) and $\phi_{x_i}(\theta) = x_i^\top\theta$ we get

$$\tilde{\mathcal{L}}(\theta; \mathcal{D}^z) = \mathcal{L}(\theta; \mathcal{D}) - \sum_{i=1}^{t-1} z_i x_i^\top \theta\,.$$

The PHE algorithm of Kveton et al. (2019) for GLBs computes at round $t \in \mathbb{N}^+$ a random parameter $\theta_t \in \arg\min_\theta \tilde{\mathcal{L}}(\theta; \mathcal{D}_{t-1}^{Z_t})$, where $\mathcal{D}_{t-1}^{Z_t}$ are the thus far collected observations and where $Z_t \sim \mathcal{N}(0, aI_{t-1})$ is chosen independently of the past. Then PHE chooses the action that is optimal under this random parameter θ_t for the action to be played in round t.

Comparing the above with EVILL, we can see that, in this GLB setting, EVILL is a variant of PHE provided that we replace the scaled identity covariance of Z_t with a diagonal covariance that uses the Fisher information of the respective data-entries:

Proposition 2. *In generalised linear bandits, EVILL reduces to a variant of GLM-PHE where the perturbation vector used by PHE in round t is*

$$Z_t \sim \mathcal{N}(0, a^2 \operatorname{diag}(I(X_1^\top \hat{\theta}_{t-1}), \dots, I(X_{t-1}^\top \hat{\theta}_{t-1})))\,.$$

However, the equivalence of EVILL and PHE *breaks in a strong sense* beyond generalised linear bandits. In particular, when the natural exponential family is replaced with a slightly bigger class, such as the exponential family, additive reward perturbations can lead to linear regret. Consider the following example.

Example 4 (Rayleigh linear bandit). *Let $\mathcal{U} = [0, \infty)$ and for any $u \in \mathcal{U}$, let $P(\cdot; u)$ be the Rayleigh distribution with parameter u, that is one with density with respect to $\nu(dy) = 2y\mathbf{1}_{[0,\infty)}(y)m(dy)$ given by*

$$p(y; u) = u \exp(-uy^2)\,,$$

where $u \in \mathcal{U} = [0, \infty)$, $\mathbf{1}_{[0,\infty)}$ is the characteristic function of $[0, \infty)$ and m is the Lebesgue measure. Let $\phi_x(\theta) = x^\top \theta$, for all $x \in \mathcal{X}$, $\theta \in \Theta$, with both $\mathcal{X}$ and Θ subsets of the positive orthant of $\mathbb{R}^d$.

In the above Rayleigh linear bandit example, $\tilde{\ell}$, the linear extension of the negative log-likelihood induced by p to the whole real line satisfies

$$\tilde{\ell}(y; u) = -\log(u) + uy^2, \quad y \in \mathbb{R}\,.$$

Now, for $Z \sim \mathcal{N}(0, 1)$,

$$\mathbb{E}[\tilde{\ell}(y + Z; u)] = -\log(u) + u(y^2 + 1) \neq \tilde{\ell}(y; u)$$

and so, unlike in generalised linear bandits, the additive-perturbation-based estimate is *biased*. As we shall see in Section 5, this bias may lead PHE to converge to choosing a suboptimal arm indefinitely. In contrast, EVILL does not incur such bias.

4.3 Regret guarantees

In this section we state our theoretical results in which we give an upper bound on the regret of EVILL. The bounds are given for a broad subclass of generalised linear bandits, captured by the following assumptions.

Assumption 2. *The following hold:*

(i) *The family $(P(\cdot; u))_{u \in \mathcal{U}}$ corresponds to a regular natural exponential family $\mathcal{Q}$ with some base Q, and $\mathcal{U} = \mathcal{U}_Q = \mathbb{R}$.*

(ii) *For all $u \in \mathbb{R}$ and a known $L > 0$, $\dot{\mu}(u) \le L$.*

(iii) *$\mathcal{Q}$ is M-self-concordant with a known constant $M > 0$:*

$$|\ddot{\mu}(u)| \le M\dot{\mu}(u) \quad \text{for all} \quad u \in \mathbb{R}\,.$$

(iv) *$\mathcal{X} \subset B_2^d = \{x \in \mathbb{R}^d\,;\, \|x\|_2 \le 1\}$ is closed*

(v) *$\Theta \subset S \cdot B_2^d = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le S\}$ where $S > 0$ is known.*

(vi) *The localisation maps are linear: $\phi_x(\theta) = x^\top \theta$ for all $x \in \mathcal{X}$ and $\theta \in \Theta$.*

Remark 1. *Observe that in Examples 2 and 3, these assumptions are met. Indeed, for the logistic linear bandit, $M = 1/4$ and $L = 1/2$, and for a Gaussian linear bandit with unit variance, $M = 0$ and $L = 1$.*

Assumptions (i)–(ii) and (iv)–(vi) essentially appeared in the work of Filippi et al. (2010), who calls bandit models where these conditions hold generalised linear bandits (GLBs). The *self-concordance property*, assumption (iii), was introduced to the GLB literature by Russac et al. (2021), and ensures that I, the Fisher information map, does not change too rapidly. Given that the algorithm uses a preliminary estimate of $\theta_\star$ together with function I, it is natural that the results require the control of the smoothness of I. In particular, in natural exponential families, $I(u) = \dot{\mu}(u)$ for any $u \in \mathcal{U}_Q^\circ$, and so this condition guarantees that $I(u)/I(v) \le e^{M|u-v|}$ for any $u, v \in \mathcal{U}_Q$, which is what we will need in our proofs. An additional property of natural exponential families is that, there, $\dot{\mu}(u)$ is also the same as the variance of Q_u, and thus L is simply an upper bound on the variance of Q_u for any $u \in \mathbb{R}$.

With the assumptions in place, the main result of this section is the following bound on the regret of EVILL:

Theorem 1. *Let Assumption 2 hold and let $n \in \mathbb{N}^+$, the horizon, be given. Assume that, for some sufficiently small $b > 0$, the prior observations satisfy*

$$\max_{x \in \mathcal{X}} \|x\|_{V_\tau^{-1}} \le b\,,$$

where

$$V_\tau = \sum_{i=1}^{\tau} X_i X_i^\top + \lambda I_d\,.$$

Then, with appropriate choices of parameters $a, \lambda > 0$, with high probability, the regret of EVILL, Algorithm 1, satisfies

$$R(n) \le \tilde{O}(d^{3/2}\sqrt{n\dot{\mu}(x_\star^\top \theta_\star)}) + O(\tau)\,,$$

where $\tilde{O}$ hides problem dependent constants and the last term accounts for the regret incurred during the τ steps during which the prior observations are obtained.

Remark 2. *When the prior observations are chosen by using a standard warm-up routine, discussed in Appendix B,*

$$\tau = O((\log n)^2)\,,$$

with constants as detailed in Remark 5 of Appendix E.

When specialised to logistic bandits, our result is comparable to that proven for the TS-ECOLog algorithm of Faury et al. (2022), which is the state-of-the-art for randomised methods. In particular, EVILL, just like TS-ECOLog, adapts to $\dot{\mu}(x_\star^\top \theta_\star)$, the variance of rewards associated with the optimal arm. See Appendix E for a formal statement of this regret bound, its proof—which builds on the works of Abeille and Lazaric (2017), Faury et al. (2020), Russac et al. (2021), and Faury et al. (2022)—and more discussion.

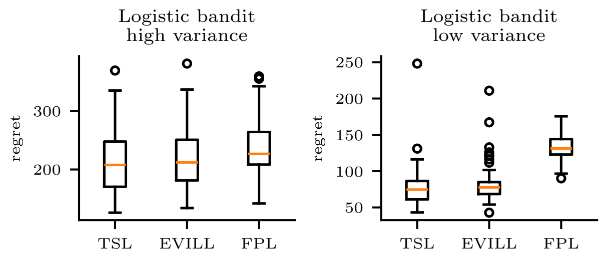

Figure 1: Regret of TSL, EVILL and FPL on two logistic linear bandit tasks: high variance, where $\dot{\mu}(x_\star^\top \theta_\star)$ is high, and low variance, where $\dot{\mu}(x_\star^\top \theta_\star)$ is small. Box plots are based on the regret from 100 instances, and show the median and the interquartile (IQ) range, with whiskers restricted to 1.5× the IQ range.

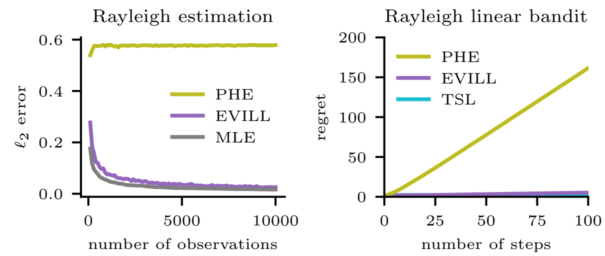

Figure 2: Rayleigh parameter estimation and bandit experiments. The left panel shows the ℓ_2 error in estimating the parameter of a Rayleigh bandit when data is i.i.d., uncontrolled by the bandit algorithm. In addition to PHE and EVILL's estimates, the MLE estimate using the same data is also shown for reference. The right panel shows regret for TSL, PHE and EVILL for the same Rayleigh bandit problem, over 100 steps of interaction. Both panels plot means over 100 iterations and no confidence intervals—the latter were too tight to be visible. EVILL and TSL lines in right panel overlap.

4.4 Confidence sets for NEFs

Our proofs also extend the confidence sets developed for natural exponential family distributions by Faury et al. (2020) and Russac et al. (2021), removing the assumption that rewards are uniformly bounded with probability one. This is based on the following moment generating function bound:

Lemma 1. *Consider an M-self-concordant NEF $\mathcal{Q}$. Then, for any $u \in \mathcal{U}_Q^\circ$ and all $|s| \leq \log(2)/M$,*

$$\psi_{Q_u}(s) \leq s\mu(u) + s^2\dot{\mu}(u)\,.$$

Thus, Q_u is what Wainwright (2019) would describe as a *sub-exponential distribution* with parameters

$$(\nu, \alpha) = (\sqrt{2\dot{\mu}(u)},\, M/\log(2))\,.$$

The above bound and the resulting confidence sets are proven in Appendix C.1 and Appendix D respectively.

5 EXPERIMENTS

We present two experiments.[2] The first, on a synthetic logistic bandit problem, shows the importance of our data-dependent perturbations. The second, on a Rayleigh parameter estimation problem and a Rayleigh linear bandit problem, shows that PHE can suffer catastrophic bias outside the exponential family.

Logistic bandit Consider the classic logistic linear bandit setting, Example 2, with arm set $\mathcal{X}$ consisting of all $x \in \{0,1\}^{10}$ such that exactly three components of x are nonzero—real world problems often exhibit this kind of combinatorial arm structure, features can either be present, or not, with the choice of 10 and 3 being arbitrary. W test on two $\theta_\star$, one giving a high variance problem, with $\dot{\mu}(x_\star^\top \theta_\star) \approx 0.15$, and a low variance problem, with $\dot{\mu}(x_\star^\top \theta_\star) \approx 0.02$. Specifically, we consider $\theta^\star$ of the form

$$[\theta_\star]_i = \frac{c_1}{(1+i+c_2)^2}$$

for each $i \in [d]$, a simple inverse polynomial form, and set $c_1, c_2 \in \mathbb{R}$ using numerical optimisation to give the right variances, and satisfy $\min_{x \in \mathcal{X}} x^\top \theta_\star = 0.1$.

As baselines, we use the Thompson sampling algorithm with Laplace approximation (TSL) and the follow-the-perturbed-leader (FPL) algorithms, both of Kveton et al. (2020b). FPL is the previous additive noise PHE method where the noise magnitude is not data-dependent. We set perturbation scale parameter a to $a = 1$ for TSL and EVILL (an arbitrary choice), and $a = 1/2 = \sqrt{L}$ for FPL, the latter such that the behaviour of FPL matches that of the other two algorithms in the high variance setting.

We provide all algorithms with a warm-up of 120 (number of arms) observations, with the actions selected uniformly at random, and run for a total of 10,000 iterations (long enough so that the regret of all algorithms is clearly sublinear), and run 100 independent copies of the experiment, which suffices to make confidence intervals negligible. The results, shown in Figure 1, confirm that the regret of EVILL, alike that of TSL, scales favourably with $\dot{\mu}(x_\star^\top \theta_\star)$. In contrast, FPL does not perform as well on the lower variance instance.

Rayleigh experiment Our second experiment is based on the linear Rayleigh bandit of Example 4, and

[2]Code: `https://github.com/DavidJanz/EVILL-code`

consists of two parts: estimation, and the full bandit problem. We will use $\theta_\star = (0.9, 0.85)$ for both.

For the estimation part, we take $\mathcal{X} = \{(1,0),(0,1)\}$ and construct a data set composed of $X_1, \dots, X_n$ sampled uniformly at random from $\mathcal{X}$, and corresponding observations $Y_1, \dots, Y_n$ sampled from a Rayleigh distributions with parameters $X_i^\top \theta^\star$, $i \in [n]$. The bandit algorithms PHE and EVILL are then used to produce parameter estimates with the modification that their action choice for round t is overwritten to be X_t. This way, we can test in isolation whether they are able to produce good parameter estimates. In Figure 2, left panel, we plot the mean ℓ_2-error of the estimates of $\theta_\star$ as a function of data subset size $n \in \{100, 200, \dots, 10,000\}$ for the two methods, as well as that of a maximum likelihood estimate, for reference. The results show that PHE is hopelessly inconsistent, while the EVILL estimates are only a little worse than those of the unperturbed MLE.

For the bandit experiment, we have arms $\mathcal{X} = \{(1, 0.99), (0.1, 0.05)\}$, chosen specifically such that the bias we expect PHE to exhibit leads to it choosing the suboptimal arm. We run TSL, PHE and EVILL for 100 steps of interaction with no warm-up. All methods use the same noise scale, which was somewhat arbitrarily chosen to be $a = 1.0$. As shown in the right panel of Figure 2, both EVILL and TSL suffer almost no regret, while PHE suffers linear regret—as expected.

6 CONCLUSION

The main contribution of this work suggests replacing perturbed-history exploration with EVILL. The main appeal of EVILL when compared to its alternatives is that it can be implemented by adding a few extra lines to any code that implements model fitting and policy optimisation, while we also expect to be competitive with its alternatives. We have showed that this holds in generalised linear bandits. Multiple intriguing avenues for further research, including the interplay between computation, memory, and performance, remain to be explored. Another potential research direction involves evaluating EVILL's interaction with nonlinear models, such as neural networks. Drawing from experience gained with Thompson sampling, we also anticipate EVILL to be useful in reinforcement learning.

Acknowledgements

Csaba Szepesvári would like to express his gratitude to Branislav Kveton and Benjamin van Roy for their invaluable feedback during earlier stages of this research. Csaba Szepesvári also gratefully acknowledges funding from the Canada CIFAR AI Chairs Program, Amii and NSERC.

References

M. Abeille, L. Faury, and C. Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, 2021. Cited on pages 2, 6.

M. Abeille and A. Lazaric. Linear Thompson sampling revisited. *Electronic Journal of Statistics*, 11(2), 2017. Cited on pages 2, 5, 7.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, 2013. Cited on pages 1, 2.

J. Antoran, S. Padhy, R. Barbano, E. Nalisnick, D. Janz, and J. M. Hernández-Lobato. Sampling-based inference for large linear models with application to linearised Laplace. In *International Conference on Learning Representations*, 2022. Cited on page 5.

O. E. Barndorff-Nielsen. *Information and Exponential Families In Statistical Theory.* Wiley Series in Probability and Statistics. Wiley, 2014. Cited on page 4.

O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. *Advances in Neural Information Processing Systems*, 2011. Cited on page 5.

K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. *Handbook of the geometry of Banach spaces*, 1(317-366):131, 2001. Cited on page 23.

L. Faury, M. Abeille, C. Calauzènes, and O. Fercoq. Improved Optimistic Algorithms for Logistic Bandits. In *International Conference on Machine Learning*, 2020. Cited on pages 7, 8, 18, 19, 26.

L. Faury, M. Abeille, K.-S. Jun, and C. Calauzènes. Jointly efficient and optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, 2022. Cited on pages 7, 20, 24, 26.

S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, 2010. Cited on pages 3, 7.

M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization.* Springer, 2011. Cited on pages 12, 14.

E. Hazan and Z. Karnin. Volumetric spanners: an efficient exploration basis for learning. *Journal of Machine Learning Research*, 2016. Cited on page 14.

B. Kveton, C. Szepesvari, M. Ghavamzadeh, and C. Boutilier. Perturbed-History Exploration in Stochastic Multi-Armed Bandits. In *International Joint Conference on Artificial Intelligence*, 2019. Cited on pages 1, 6.

B. Kveton, C. Szepesvári, M. Ghavamzadeh, and C. Boutilier. Perturbed-History Exploration in Stochastic Linear Bandits. In *Conference on Uncertainty in Artificial Intelligence*, 2020. Cited on page 1.

B. Kveton, M. Zaheer, C. Szepesvari, L. Li, M. Ghavamzadeh, and C. Boutilier. Randomized Exploration in Generalized Linear Bandits. In *International Conference on Artificial Intelligence and Statistics*, 2020. Cited on pages 1, 2, 5, 6, 8, 24.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020. Cited on pages 1, 2, 4, 13–15, 19, 26.

G. Letac and M. Mora. Natural real exponential families with cubic variance functions. *The Annals of Statistics*, 18(1):1–37, 1990. Cited on page 15.

X. Lu and B. Van Roy. Ensemble sampling. *Advances in Neural Information Processing Systems*, 2017. Cited on page 1.

C. N. Morris. Natural exponential families with quadratic variance functions: statistical theory. *The Annals of Statistics*:515–529, 1983. Cited on page 15.

C. N. Morris and K. F. Lock. Unifying the named natural exponential families and their relatives. *The American Statistician*, 63(3):247–253, 2009. Cited on page 15.

I. Osband, B. Van Roy, and Z. Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, 2016. Cited on page 1.

G. Papandreou and A. L. Yuille. Gaussian sampling by local perturbations. In *Advances in Neural Information Processing Systems*, 2010. Cited on page 5.

Y. Russac, L. Faury, O. Cappé, and A. Garivier. Self-concordant analysis of generalized linear bandits with forgetting. In *International Conference on Artificial Intelligence and Statistics*, 2021. Cited on pages 2, 6–8.

D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018. Cited on page 5.

T. Sun and Q. Tran-Dinh. Generalized self-concordant functions: a recipe for Newton-type methods. *Mathematical Programming*, 178(1-2):145–213, 2019. Cited on pages 14, 16, 24.

M. J. Todd. *Minimum Volume Ellipsoids*. SIAM, 2016. Cited on page 14.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019. Cited on pages 8, 27.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A ON COMPUTATIONAL COMPLEXITY

A naive implementation of the algorithm stores all the data. This uses $O(nd)$ memory.

Regarding the compute cost, first we note that the algorithm is tolerant to approximation errors in solving the optimisation problems. In particular, it follows with a simple argument that it is sufficient to optimise the losses in Steps 2 and 5 up to a constant accuracy, while it suffices to find an optimal arm up to an accuracy of $1/\sqrt{n}$.

In regard to the computational complexity of the various steps of the algorithm, the following holds: In Steps 2 and 5, the algorithm needs to solve a convex optimisation problem and the starting point of the design of our algorithm was that this can be done efficiently. Indeed, the ellipsoidal algorithm can be used for this purpose with cost that scales polynomially in n, d and $\log(1/\epsilon)$ (Grötschel et al., 2011), though we expect numerical algorithms tailored to the properties of the specific natural exponential family underlying a specific bandit model to be more efficient.

Steps 3 and 4 take $O(n(c + d))$ time, where c is the cost of evaluating $\dot{\phi}_{X_i}$ at some input of its domain and the cost of evaluating the Fisher-information. For generalised linear models, $\dot{\phi}_{X_i}$ is the identity map, which means that c represents just the cost of evaluating the Fisher-information. In "named" natural exponential families $c = O(1)$, assuming that the standard transcendental functions can be evaluated in $O(1)$ time.

Another starting point for the design of our algorithm was that arm selection (Step 6) can also be implemented in an efficient way. When $\mathcal{X}$ is finite, this can be done by looping through all the arms. With this, the cost of selecting an arm is at most $O(|\mathcal{X}|c)$ where c represents the cost of evaluating μ and ϕ_x. For generalised linear models, μ can be dropped (as it is an increasing function), and thus c becomes the cost of evaluating ϕ_x, which is $O(d)$. When $\mathcal{X}$ is a convex set (which one may assume in any case without loss of generality) and μ is an increasing function, the assumption that arm-selection can be done efficiently is equivalent to that linear optimisation over $\mathcal{X}$ can be efficiently implemented. This is known to be possible when $\mathcal{X}$ is represented with either a membership or a separation oracle (Grötschel et al., 2011).

Putting everything together, we find that in generalised linear models the total cost is polynomial in n and d. For a naive implementation, the cost, however, is at least quadratic in n. It remains to be seen whether this quadratic cost can be avoided with *any* randomised method that is not relying on maximising upper confidence bounds.

When compute time is a bottleneck, one approach is to parallelise the computation. Here, with $O(n)$ computers, accessing the same central storage, one can parallelise the computation of the perturbed estimates to reduce the per iteration compute time to $O(1)$ at the price of adding $O(n)$ memory per computer (compute t would store $(Z_{t,i})_{i \le t}$). The speed-up is possible because the losses change by a little only between rounds, hence, computer t, which is used to get the weight vector to used in round t, can start the optimisation from the result that it obtained by optimising the objective just with the data available in rounds $1, \dots, t-2$.

B PRIOR OBSERVATIONS, AND THE BASICS OF SELF-CONCORDANCE

Recall that Algorithm 1 asks for prior observations $(X_1, Y_1), \dots, (X_\tau, Y_\tau)$. Then, Theorem 1 asks that, for $V_\tau = \sum_{i=1}^{\tau} X_i X_i^\top + \lambda I$ and some $b > 0$, the *precondition*

$$\max_{x \in \mathcal{X}} \|x\|_{V_\tau^{-1}} \le b$$

holds. In this appendix, we look at what these prior observations buy us—which gives us an excuse to introduce some basic results on self-concordance—and how to gather these prior observations efficiently. We start with the latter.

B.1 Warm-up procedures

Let $X_1, X_2, \dots \in \mathbb{R}^d$ and for $t \ge 1$ let $V_t = \lambda I + \sum_{s \le t} X_s X_s^\top$ with $V_0 = \lambda I$. Suppose we want to produce prior observations $X_1, X_2, \dots, X_\tau$ so that $\max_{x \in \mathcal{X}} \|x\|_{V_\tau^{-1}} \le b$. We are interested in stopping as early as possible. Hence, in round $t+1$ we stop if $\max_{x \in \mathcal{X}} \|x\|_{V_t^{-1}}$, while if this is not satisfied we choose X_{t+1} in some way. Define

$$\tau = \min\{t \in [n] \colon \max_{x \in \mathcal{X}} \|x\|_{V_t^{-1}} \le b\}$$

as the index of the round when we stop. How small can we make τ (or $\mathbb{E}[\tau]$ when the (X_t) are chosen by using a randomised method).

A slight modification of Exercise 19.3 of Lattimore and Szepesvári (2020) gives the following:

Lemma 2 (Elliptical potentials: You cannot have more than $O(d)$ big intervals)**.** *Let $V_0 = \lambda I$ and $a_1, \dots, a_n \in \mathbb{R}^d$ be a sequence of vectors with $\|a_t\|_2 \le L$ for all $t \in [n]$. Further, let $V_t = V_0 + \sum_{s=1}^t a_s a_s^\top$. Then, the number of times $\|a_t\|_{V_{t-1}^{-1}} \ge b$ is at most*

$$\frac{3d}{\log(1+b^2)} \log\left(1 + \frac{L^2}{\lambda \log(1+b^2)}\right) .$$

For completeness, we include the proof:

Proof. Let $\mathcal{T}$ be the set of rounds t when $\|a_t\|_{V_{t-1}^{-1}} \ge b$ and $G_t = V_0 + \sum_{s=1}^t 1\{s \in \mathcal{T}\} a_s a_s^\top$. Then

$$\begin{aligned}
\left(\frac{d\lambda + |\mathcal{T}|L^2}{d}\right)^d &\ge \left(\frac{\text{trace}(G_n)}{d}\right)^d \\
&\ge \det(G_n) && \text{(determinant-trace inequality)} \\
&= \det(V_0) \prod_{t \in \mathcal{T}} (1 + \|a_t\|^2_{G_{t-1}^{-1}}) && \text{(expanding determinant)} \\
&\ge \det(V_0) \prod_{t \in \mathcal{T}} (1 + \|a_t\|^2_{V_{t-1}^{-1}}) && \text{(dropping indicators)} \\
&\ge \lambda^d (1+b^2)^{|\mathcal{T}|} . && \text{(definition of } \mathcal{T})
\end{aligned}$$

Rearranging and taking the logarithm shows that

$$|\mathcal{T}| \le \frac{d}{\log(1+b^2)} \log\left(1 + \frac{|\mathcal{T}|L^2}{d\lambda}\right) .$$

Abbreviate $x = d/\log(1+b^2)$ and $y = L^2/d\lambda$, which are both positive. Then

$$x \log\left(1 + y(3x\log(1+xy))\right) \le x \log\left(1 + 3x^2y^2\right) \le x \log((1+xy)^3) = 3x \log(1+xy) .$$

Since $z - x\log(1+yz)$ is decreasing for $z \ge 3x\log(1+xy)$ it follows that

$$|\mathcal{T}| \le 3x \log(1+xy) = \frac{3d}{\log(1+b^2)} \log\left(1 + \frac{L^2}{\lambda \log(1+b^2)}\right) .$$

□

Now, by Assumption 2(iv), we can set $L = 1$. It follows that regardless the choice of X_{t+1}, the total number of times $\|X_{t+1}\|_{V_t^{-1}} \ge b$ holds for $t = 0, 1, \dots$ is bounded by

$$\tau_{\text{naive}} = \frac{3d}{\log(1+b^2)} \log\left(1 + \frac{1}{\lambda \log(1+b^2)}\right) .$$

Thus, we can simply choose $X_{t+1} = \arg\max_{x \in \mathcal{X}} \|x\|_{V_t^{-1}}$, which guarantees that $\tau \le \tau_{\text{naive}}$.

The dependence on b can be slightly improved by a refined approach. To see how this can be done we need to recall the Kiefer-Wolfowitz theorem and some definitions. For any distribution π over $\mathcal{X}$, let

$$\bar{V}(\pi) = \int x x^\top \pi(dx) \quad \text{and} \quad g(\pi) = \max_{x \in \mathcal{X}} \|x\|^2_{\bar{V}(\pi)^{-1}} .$$

The Kiefer-Wolfowitz theorem (e.g., Theorem 21.1 in Lattimore and Szepesvári (2020)) states that when $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$, there exists a probability distribution $\pi_\star$ over $\mathcal{X}$ such that $g(\pi_\star) = d$, and the support of $\pi_\star$ is a finite set with at most $d(d+1)/2$-many points in it. With that, sampling $X_1, X_2, \dots$ independently of $\pi_\star$ gives us a stopping time τ that is, with high probability, on the order of d/b^2. Alternatively, one can use $\lceil n\pi_\star(y)\rceil$ observations from y in the support of $\pi_\star$ with $n = \lceil d/b^2 \rceil \geq d/b^2$ for a total of at most $d(d+1)/2 + n \leq d(d+1)/2 + 1 + d/b^2$ observations. The distribution $\pi_\star$ is called the G-optimal design, for 'globally optimal'. This improves the dependency of τ on b from slightly worse than $3d/\log(1+b^2)$ to d/b^2.

It remains to be seen how a G-optimal design can be found. Finding the G-optimal design, precisely, or up to a tiny approximation error, can be expensive. Fortunately, there is no need for doing this.

In particular, finding a distribution π, such that, say, $g(\pi) \leq 2g(\pi_\star)$, can be done in time almost linear in d, and using such a distribution in place of $\pi_\star$ increases the resulting stopping time τ by at most a factor of 2. Specifically, when $\arg\max_{x\in\mathcal{X}} c^\top x$ and $\arg\max_{x\in\mathcal{X}} \|x\|^2_{V^{-1}}$ can both be efficiently computed for any $c \in \mathbb{R}^d$ and positive definite V, an appropriate distribution π can be found in $O(d \log\log d)$ time using the classical Frank-Wolfe algorithm (see Lattimore and Szepesvári (2020), Note 3 of section 21.2, together with Algorithm 3.3 and Theorem 3.9 of Todd (2016)). Furthermore, the size of the support of the resulting distribution π will match the number of steps that algorithm, $O(d\log\log d)$. When only a linear optimisation oracle over $\mathcal{X}$ is available, Hazan and Karnin (2016) gives an alternate poly-time approach. As mentioned earlier, for $\mathcal{X}$ convex, the ellipsoidal algorithm can be used at a polynomial cost for implementing both $\arg\max_{x\in\mathcal{X}} c^\top x$ and $\arg\max_{x\in\mathcal{X}} \|x\|^2_{V^{-1}}$ (Grötschel et al., 2011).

B.2 Prior observations and self-concordance: what do our prior observations guarantee?

Per Proposition 1, for an observation at location $x \in \mathcal{X}$, we might like to introduce perturbations with variance that scales with $I(x^\top\theta_\star) = \dot\mu(x^\top\theta_\star)$, but we do not know $\theta_\star$. We have only some estimate $\hat\theta_{t-1}$, and use $\dot\mu(x^\top\hat\theta_{t-1})$ instead. Self-concordance lets us bound how close $\dot\mu(x^\top\theta_\star)$ and $\dot\mu(x^\top\hat\theta_t)$ are. For this, the following result, shown in Sun and Tran-Dinh (2019), will be useful:

Lemma 3. *Consider an M-self-concordant NEF domain $\mathcal{U}_Q$. Then, for any $u, u' \in \mathcal{U}_Q^\circ$,*

$$\dot\mu(u) \leq \dot\mu(u')e^{M|u-u'|}\,.$$

With that, if we know that, say,

$$\max_{x\in\mathcal{X}} |x^\top(\theta_\star - \hat\theta_{t-1})| \leq \frac{1}{M}\,, \tag{6}$$

then it follows that $\dot\mu(x^\top\hat\theta_{t-1})$ is an e-multiplicative approximation to $\dot\mu(x^\top\theta_\star)$. Our requirements on the prior observations will be so as to make this happen—and slightly more, in that we actually end up needing that not just $\dot\mu(x^\top\hat\theta_{t-1})$ but $\dot\mu(x^\top\theta_t)$ is also close to $\dot\mu(x^\top\theta_\star)$. For this, it will suffice to slightly tighten Equation (6), but for now, to keep things simple, we keep considering only Equation (6).

To connect Equation (6) to the prior observations, we first upper bound the right-hand side of Equation (6). For this, for $t \in [n]$, let

$$H_t = \sum_{i=1}^t \dot\mu(X_i^\top\theta_\star)X_iX_i^\top + \lambda I \tag{7}$$

(and $H_0 = \lambda I$) and observe that since these matrices are positive definite, by Cauchy-Schwarz,

$$|x^\top(\theta_\star - \hat\theta_{t-1})| \leq \|x\|_{H_{t-1}^{-1}}\|\theta_\star - \hat\theta_{t-1}\|_{H_{t-1}}.$$

Thus, if we have a high probability upper bound on $\|\theta_\star - \hat\theta_{t-1}\|_{H_{t-1}}$ that holds for all $t \in [n]$, then it suffices that we ensure that, for all $t > \tau$, $\|x\|_{H_t^{-1}}$ times that upper bound is less than $1/M$. Since our warm-up is in terms of V_t and not H_t, we need a way to convert between these. To do so, we define the constant $\kappa \geq 1$ as follows.

Definition 1 (Constant κ)**.** *Let κ be the smallest upper bound on $\{1\} \cup \{1/\dot\mu(u) : |u| \leq S\}$:*

$$\kappa = 1 \vee \max_{u:|u|\leq S} \frac{1}{\dot\mu(u)}\,.$$

With that, we have the bound $H_{t-1}^{-1} \preceq \kappa V_{t-1}^{-1}$, from which the following claim follows immediately. Here, and in what follows, we use $\preceq$ to denote the Loewner partial ordering of positive semidefinite matrices (i.e., $A \preceq B$ if $B - A$ is positive semidefinite).

Claim 1 (Precondition)**.** *For all* $t \in \{\tau + 1, \dots, n\}$*,*

$$\max_{x\in\mathcal{X}} \|x\|_{H_{t-1}^{-1}} \leq b\sqrt{\kappa}\,.$$

It remains to establish just how small we need to set b, which will depend on how well we can upper bound $\|\theta_\star - \hat\theta_{t-1}\|_{H_{t-1}}$ (and $\|\theta_\star - \theta_t\|_{H_{t-1}}$, but that will not be hard given the previous bound). We return to this after establishing some additional definitions.

C PRELIMINARIES FOR THE CONSTRUCTION OF CONFIDENCE SETS AND THE REGRET BOUND

In this appendix, we list some standard properties of NEFs, define some quantities related to the likelihood function and use self-concordance results to provide bounds for these. Quantities defined here will be used throughout the remainder of the appendices.

C.1 Natural exponential families and their cumulant generating functions

Consider a natural exponential family $\mathcal{Q}$ with base Q, cumulant generating function $\psi(s) = \log \int e^{sx} Q(dx)$ and domain $\mathcal{U}_Q$. The cumulant generating function ψ is analytic on $\mathcal{U}_Q^\circ$ (see Exercise 5.9 in Lattimore and Szepesvári (2020)), and, for any $u \in \mathcal{U}_Q^\circ$, its derivatives are given by

$$\begin{aligned}
\dot\psi(u) &= \int x P(dx; u) = \mu(u) \\
\ddot\psi(u) &= \int (x - \mu(u))^2 P(dx; u) = \dot\mu(u) \\
\dddot\psi(u) &= \int (x - \mu(u))^3 P(dx; u) = \ddot\mu(u).
\end{aligned}$$

That is, the derivatives of ψ at u give the first three central moments of the distribution $P(\cdot; u)$. For these and more properties of NEFs, see Morris and Lock (2009). For even more, see Morris (1983) and Letac and Mora (1990).

With these properties fresh in mind, we now restate and quickly prove Lemma 1.

Lemma 1. *Consider an* M*-self-concordant NEF* $\mathcal{Q}$*. Then, for any* $u \in \mathcal{U}_Q^\circ$ *and all* $|s| \leq \log(2)/M$*,*

$$\psi_{Q_u}(s) \leq s\mu(u) + s^2\dot\mu(u)\,.$$

Proof. Writing out the expressions involved, it is immediate that $\psi_{Q_u}(s) = \psi(u+s) - \psi(u)$. Taking a second order Taylor expansion of $\psi(u+s)$ about u, assuming without loss of generality that $s > 0$, we thus see that there exists a $\xi \in [u, u+s]$ such that

$$\psi_{Q_u}(s) = s\dot\psi(u) + \frac{s^2\ddot\psi(\xi)}{2} = s\mu(u) + \frac{s^2\dot\mu(\xi)}{2},$$

where we used that $\dot\psi = \mu$. By self-concordance, that is, using Lemma 3, and considering only s that satisfy $|s| < \log(2)/M$, we have that

$$\dot\mu(\xi) \leq \exp(M|\xi - u|)\dot\mu(u) \leq \exp(M|s|)\dot\mu(u) \leq 2\dot\mu(u),$$

completing the proof. ☐

C.2 The negative log-likelihood function of natural exponential families

Recall from Equation (1) that the λ-regularised negative log-likelihood function given $\mathcal{D}_t = ((X_i, Y_i))_{i=1}^t$ is

$$\mathcal{L}(\theta; \mathcal{D}_t) = \frac{\lambda}{2}\|\theta\|^2 - \sum_{i=1}^t \log p(Y_i; X_i^\top \theta)\,.$$

For an NEF with base Q, defining the density p in the above with respect to Q, we have that $p(y; u) = \exp(yu - \psi(u))$, and so

$$\mathcal{L}(\theta; \mathcal{D}_t) = \frac{\lambda}{2}\|\theta\|^2 - \sum_{i=1}^t (Y_i X_i^\top \theta - \psi(X_i^\top \theta))\,,$$

where recall that ψ is the cumulant-generating function of Q. Since we will be interested in the global minimisers of $\mathcal{L}(\theta; \mathcal{D}_t)$, and ψ is known to be strictly convex and smooth, $\mathcal{L}$ has a unique minimiser, which is the unique stationary point of $\mathcal{L}$. Taking the derivative of both sides of the previous display with respect to θ, and using that $\dot{\psi} = \mu$, we get

$$\left(\frac{\partial}{\partial \theta}\mathcal{L}(\theta; \mathcal{D}_t)\right)^\top = \underbrace{\sum_{i=1}^t \mu(X_i^\top \theta) X_i + \lambda \theta}_{=: g_t(\theta)} - \sum_{i=1}^t X_i Y_i\,, \tag{8}$$

where the previous display defines the 'gradient function' g_t. It follows that $\hat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta; \mathcal{D}_t)$ satisfies

$$g_t(\hat{\theta}_t) = \sum_{i=1}^t X_i Y_i\,. \tag{9}$$

We will also need to reason about the Hessian $H(\theta; \mathcal{D}_t) = \frac{\partial^2}{\partial \theta^2}\mathcal{L}(\theta; \mathcal{D}_t)$ of $\mathcal{L}(\theta; \mathcal{D}_t)$. Further differentiation gives

$$H(\theta; \mathcal{D}_t) = \sum_{i=1}^t \dot{\mu}(X_i^\top \theta) X_i X_i^\top + \lambda I.$$

Now, notice that

$$H_t = H(\theta_\star; \mathcal{D}_t)\,. \tag{10}$$

We further let

$$\widehat{H}_t = H(\hat{\theta}_t, \mathcal{D}_t) \tag{11}$$

The reader will be reminded of these definitions frequently.

C.3 The secant gradient approximation, the Hessian, and self-concordance

We let $\alpha : \mathcal{U}_Q^\circ \times \mathcal{U}_Q^\circ \to \mathbb{R}$ denote a 'secant approximation' to $\dot{\mu}$. In particular, for any $u, u' \in \mathcal{U}_Q^\circ$,

$$\alpha(u, u') = \begin{cases} \dot{\mu}(u)\,, & \text{if } u = u'\,; \\ \frac{\mu(u) - \mu(u')}{u - u'}\,, & \text{otherwise}\,. \end{cases}$$

For $u \neq u'$, the value of $\alpha(u, u')$ is the gradient of the secant line from u to u' and $\lim_{u' \to u} \alpha(u, u') = \dot{\mu}(u)$. Furthermore, clearly, α is a symmetric function of its arguments:

$$\alpha(u, u') = \alpha(u', u) \text{ for all } u, u' \in \mathcal{U}_Q^\circ\,.$$

With self-concordance, α can be a relatively good approximation to $\dot{\mu}$, as shown by the following result, a consequence of Lemma 3 combined with the fundamental theorem of calculus. This result is a special case of Corollary 2 of Sun and Tran-Dinh (2019), hence the proof is omitted.

Claim 2. *For all $u, u' \in \mathcal{U}_Q^\circ$,*

$$\dot{\mu}(u)h(-M|u-u'|) \leq \alpha(u,u') \leq \dot{\mu}(u)h(M|u-u'|)\,,$$

where for $x \neq 0$, $h(x) = \frac{e^x-1}{x}$ and $h(0) = 1$.

Based on the secant approximation to the gradient, we construct the following approximation to the Hessian $H(\theta; \mathcal{D}_t)$:

$$G(\theta, \theta'; \mathcal{D}_t) = \sum_{i=1}^{t} \alpha(X_i^\top \theta, X_i^\top \theta') X_i X_i^\top + \lambda I.$$

Thanks to α being symmetric, G is also a symmetric function of θ and θ'.

By Claim 2, it is clear that when θ' is close, in a suitable sense, to θ, $G(\theta, \theta'; \mathcal{D}_t) \approx H(\theta; \mathcal{D}_t)$. The suitable sense here is captured by the pseudo-norm

$$D(v) = \max_{x \in \mathcal{X}} |x^\top v|, \qquad v \in \mathbb{R}^d\,.$$

Indeed, Claim 2 with a crude bound gives the following upper bound on $H(\theta; \mathcal{D}_t)$ in terms of $G(\theta, \theta'; \mathcal{D}_t)$:

Claim 3. *For any $\theta, \theta' \in \mathbb{R}^d$, $H(\theta; \mathcal{D}_t) \preceq (1 + MD(\theta - \theta'))G(\theta, \theta'; \mathcal{D}_t) = (1 + MD(\theta - \theta'))G(\theta', \theta; \mathcal{D}_t)$.*

Proof. Just notice that for h from Claim 2 it holds that for $x \geq 0$, $h(-x) \geq 1/(1+x)$. This together with the definition of D and Claim 2 gives the result. □

We can, of course, get a two-sided bound from Claim 2, but we happen not to need it. Regarding the pseudonorm D, clearly, for any $v \in \mathbb{R}^d$,

$$D(v) \leq \|v\| \tag{12}$$

holds thanks to $\mathcal{X} \subset B_2^d$. We will often rely on this inequality.

Finally, directly from the definitions of $G(\theta, \theta'; \mathcal{D}_t)$ and g_t, we see the following.

Claim 4. *For any $\theta, \theta' \in \mathbb{R}^d$, $g_t(\theta) - g_t(\theta') = G(\theta, \theta'; \mathcal{D}_t)(\theta - \theta') = G(\theta', \theta; \mathcal{D}_t)(\theta - \theta')$.*

We will use these when working with first order optimality conditions for the negative log-likelihood function.

D CONFIDENCE SETS FOR THE REGULARISED MLE IN SELF-CONCORDANT NATURAL EXPONENTIAL FAMILIES

In this appendix, we show the following concentration result.

Lemma 4. *For any $\delta, \lambda > 0$, let $\gamma_t(\delta, \lambda)$ be given by*

$$\gamma_t(\delta, \lambda) = \sqrt{\lambda}\left(\frac{1}{2M} + S\right) + \frac{2Md}{\sqrt{\lambda}}\left(1 + \frac{1}{2}\log\left(1 + \frac{tL}{\lambda d}\right)\right) + \frac{2M}{\sqrt{\lambda}}\log(1/\delta),$$

and let $\mathcal{E}(\delta, \lambda)$ be the event

$$\mathcal{E}(\delta, \lambda) = \left\{\forall t \geq 0,\ \|g_t(\hat{\theta}_t) - g_t(\theta_\star)\|_{H_t^{-1}} \leq \gamma_t(\delta, \lambda)\right\}.$$

Then, under Assumption 2, $\mathbb{P}(\mathcal{E}(\delta, \lambda)) \geq 1 - \delta$.

The above lemma is a direct consequence of Lemma 1, the bound on the cumulant generating function of a self-concordant natural exponential family distribution, which we just proved in Appendix C.1. We start the proof with an intermediate result in the next section.

D.1 An intermediate concentration result

The following concentration inequality is a generalisation of Theorem 1 of Faury et al. (2020), who prove a similar statement under the assumption that each response Y_t is absolutely bounded by 1, that is $|Y_t| \leq 1$ almost surely for all $t \in \mathbb{N}^+$. We relax this to a condition that bounds the conditional cumulant generating function of Y_t by a quadratic term in a neighbourhood of zero. Comparing the condition in the statement of the theorem with our Lemma 1, it is clear that, under Assumption 2, we will be able to satisfy this condition.

Theorem 2. *Fix $\lambda, M > 0$. Let $(X_t)_{t\in\mathbb{N}^+}$ be a B_2^d-valued random sequence, $(Y_t)_{t\in\mathbb{N}^+}$ a real valued random sequence, $(\nu_t)_{t\in\mathbb{N}}$ be a nonnegative valued random sequence. Let $\mathbb{F}' = (\mathbb{F}'_t)_{t\in\mathbb{N}}$ be a filtration such that* (i) *$(X_t)_{t\in\mathbb{N}^+}$ is $\mathbb{F}'$-predictable and* (ii) *$(Y_t)_{t\in\mathbb{N}^+}$ and $(\nu_t)_{t\in\mathbb{N}}$ are $\mathbb{F}'$-adapted.*[3] *Let $\epsilon_t = Y_t - \mathbb{E}[Y_t \mid \mathbb{F}'_{t-1}]$ and assume that the following condition holds:*

$$\mathbb{E}[\exp(s\epsilon_t) \mid \mathbb{F}'_{t-1}] \leq \exp(s^2\nu_{t-1}) \quad \text{for all} \quad |s| \leq 1/M \text{ and } t \in \mathbb{N}^+ \,. \tag{13}$$

Then, for $\widetilde{H}_t = \sum_{i=1}^t \nu_{i-1}X_iX_i^\top + \lambda I$ and $S_t = \sum_{i=1}^t \epsilon_i X_i$ and any $\delta > 0$,

$$\mathbb{P}\left(\exists t \in \mathbb{N}^+ \colon \|S_t\|_{\widetilde{H}_t^{-1}} \geq \frac{\sqrt{\lambda}}{2M} + \frac{2M}{\sqrt{\lambda}}\log\left(\frac{\det(\widetilde{H}_t)^{1/2}\lambda^{-d/2}}{\delta}\right) + \frac{2M}{\sqrt{\lambda}}d\log(2)\right) \leq \delta.$$

The proof of Theorem 2 follows identically to that of Theorem 1 in Faury et al. (2020), once their Lemma 5 is replaced with the following lemma. As the proof of their Theorem 1 is rather tedious, we do not reproduce it.

Lemma 5. *Assume the conditions of Theorem 2 hold. For each $t \in \mathbb{N}^+$, let $\bar{H}_t = \sum_{i=1}^t \nu_{i-1}X_iX_i^\top$ and $S_t = \sum_{i=1}^t \epsilon_i X_i$. For any $\xi \in B_2^d$, define the real-valued process starting with $M_0(\xi) = 1$ and given by*

$$M_t(\xi) = \exp\left(\frac{1}{M}\xi^\top S_t - \frac{1}{M^2}\|\xi\|^2_{\bar{H}_t}\right) \quad \text{for} \quad t \in \mathbb{N}^+.$$

Then, $(M_t(\xi))_{t=0}^\infty$ is a nonnegative supermartingale with respect to $\mathbb{F}'$.

Remark 3. *We use the notation $M_0(\xi), M_1(\xi), \dots$ for the process to match that of Theorem 1 and Lemma 5 of Faury et al. (2020). This process is not to be confused with the constant $M > 0$, which while only an abstract constant here, will of course become the self-concordance constant when we use the result.*

In the next and subsequent proofs the statements concerning conditional expectations hold almost surely, even when this is not explicitly stated.

Proof. Fix $t \in \mathbb{N}^+$, $\xi \in B_2^d$. From Equation (13), we get that for any $\mathbb{F}'_{t-1}$-measurable random variable R_t such that $|R_t| \leq 1/M$, the inequality

$$\mathbb{E}\left[\exp\left(R_t\epsilon_t - R_t^2\nu_{t-1}\right) \mid \mathbb{F}'_{t-1}\right] \leq 1$$

holds. Define now $R_t = \xi^\top X_t/M$. Since, by assumption, $\xi, X_t \in B_2^d$, it holds that $|\xi^\top X_t| \leq 1$ and hence $|R_t| \leq 1/M$. Further, since $(X_t)_{t\in\mathbb{N}^+}$ is $\mathbb{F}'$-predictable, X_t and hence also R_t is $\mathbb{F}'_{t-1}$-measurable. Thus, by the previous inequality,

$$\mathbb{E}\left[\exp\left(\frac{\epsilon_t}{M}\xi^\top X_t - \frac{\nu_{t-1}}{M^2}\xi^\top X_tX_t^\top\xi\right) \mid \mathbb{F}'_{t-1}\right] \leq 1\,. \tag{14}$$

It then follows that

$$\begin{aligned}
\mathbb{E}[M_t(\xi) \mid \mathbb{F}'_{t-1}] &= \mathbb{E}\left[\exp\left(\frac{1}{M}\xi^\top S_t - \frac{1}{M^2}\|\xi\|^2_{\bar{H}_t}\right) \mid \mathbb{F}'_{t-1}\right] \\
&= \exp\left(\frac{1}{M}\xi^\top S_{t-1} - \frac{1}{M^2}\|\xi\|^2_{\bar{H}_{t-1}}\right)\mathbb{E}\left[\exp\left(\frac{\epsilon_t}{M}\xi^\top X_t - \frac{\nu_{t-1}}{M^2}\xi^\top X_tX_t^\top\xi\right) \mid \mathbb{F}'_{t-1}\right] \\
&\leq \exp\left(\frac{1}{M}\xi^\top S_{t-1} - \frac{1}{M^2}\|\xi\|^2_{\bar{H}_{t-1}}\right) && \text{(by Equation (14))} \\
&= M_{t-1}(\xi)\,,
\end{aligned}$$

finishing the proof. □

[3]Thus, for $t \geq 1$, X_t is $\mathbb{F}'_{t-1}$-measurable, Y_t, ν_t are $\mathbb{F}'_t$-measurable, while ν_0 is also $\mathbb{F}'_0$-measurable.

Remark 4. *As noted beforehand, the lemma just proved corresponds to that of Lemma 5 of Faury et al. (2020). Faury et al. (2020) prove this lemma based on their Lemma 7, which states that for a centred random variable* ϵ *almost surely absolutely bounded by* 1*, for* $|s| \le 1$, $\log \mathbb{E}[\exp(s\epsilon)] \le \log(1 + s^2 \operatorname{Var}[\epsilon])$*. Noting that in their proof of Lemma 5 they use* $\log(1 + s^2 \operatorname{Var}[\epsilon]) \le s^2 \operatorname{Var}[\epsilon]$*, we see that our proof simply replaces Lemma 7 with requiring* $\log \mathbb{E}[\exp(s\epsilon)] \le s^2 \operatorname{Var}[\epsilon]$ *directly (cf. Equation* (13)*), which, ultimately will follow from Lemma 1.*

D.2 Proof of the confidence set result, Lemma 4

We can now prove Lemma 4. This is effectively just combining Assumption 2, Lemma 1 and Theorem 2.

Proof of Lemma 4. Consider an arbitrary $t \in \mathbb{N}$. By first order optimality and the definition of g_t, the following inequality, which we copy from Equation (9) for the convenience of the reader, holds:

$$g_t(\hat{\theta}_t) = \sum_{i=1}^{t} Y_i X_i \,.$$

Also, by the definition of g_t,

$$g_t(\theta_\star) = \sum_{i=1}^{t} \mu(X_i^\top \theta_\star) X_i + \lambda \theta_\star.$$

Writing $\epsilon_i = Y_i - \mu(X_i^\top \theta_\star)$ and $S_t = \sum_{i=1}^t \epsilon_i X_i$, we thus have that

$$g_t(\hat{\theta}_t) - g_t(\theta_\star) = S_t - \lambda \theta_\star.$$

Taking the H_t^{-1}-weighted 2-norm of the above and using the triangle inequality and recalling that $H_t \succeq \lambda I$ and $\|\theta_\star\| \le S$, we get

$$\|g_t(\hat{\theta}_t) - g_t(\theta_\star)\|_{H_t^{-1}} \le \|S_t\|_{H_t^{-1}} + \lambda \|\theta_\star\|_{H_t^{-1}} \le \|S_t\|_{H_t^{-1}} + \sqrt{\lambda} S.$$

We complete the proof by bounding $\|S_t\|_{H_t^{-1}}$ using Theorem 2 applied to $(X_t, Y_t)_{t \in \mathbb{N}^+}$ and $(\nu_t)_{t \in \mathbb{N}}$, where the latter is defined by

$$\nu_{t-1} = \dot{\mu}(X_t^\top \theta_\star)\,, \qquad t \in \mathbb{N}^+ \,. \tag{15}$$

First, we verify the conditions of this theorem. Choose any filtration $\mathbb{F}'$ that makes $(X_t)_{t \in \mathbb{N}^+}$ $\mathbb{F}'$-predictable and $(Y_t)_{t \ge 1}$ $\mathbb{F}'$-adapted. Then, $(\nu_t)_{t \in \mathbb{N}}$ is also $\mathbb{F}'$-adapted.

By Assumption 2(iv), $X_t \in B_2^d$ for $t \ge 1$. Our definition of ϵ_i and S_t matches the definition in Theorem 2. In particular, $\epsilon_t = Y_t - \mathbb{E}[Y_t | \mathbb{F}'_{t-1}] = Y_t - \mu(X_t^\top \theta_\star)$. Then, Equation (13) holds because $\log(2) \le 1$, Lemma 1 and the choice of $(\nu_t)_{t \in \mathbb{N}}$ in Equation (15). Note that Lemma 1 is applicable because Assumption 2(i) and (iii) hold.

Now, note that Equation (15), together with the definition of H_t (see Equation (7)) implies that

$$\widetilde{H}_t = H_t \,.$$

Applying Theorem 2, we conclude that, with probability at least $1 - \delta$,

$$\|S_t\|_{H_t^{-1}} \le \frac{\sqrt{\lambda}}{2M} + \frac{2M}{\sqrt{\lambda}} \left(d + \frac{1}{2} \log\left(\det H_t / \lambda^d\right) + \log(1/\delta) \right),$$

where we upper-bounded the $\log(2)$ featuring in Theorem 2 by 1 for convenience.

We prepare for upper bounding the determinant $\det H_t$ that appeared on the right-hand side of the last display. Applying the usual trace-determinant inequality, say, Note 1, Section 20.2 of Lattimore and Szepesvári (2020), with the bound $\dot{\mu}(X_i^\top \theta_\star) \le L$ and $\|X_i\| \le 1$ which hold due to Assumption 2(ii) and Assumption 2(iv), respectively, we have that

$$\det H_t / \lambda^d \le \left(1 + \frac{tL}{\lambda d} \right)^d,$$

which we substitute into the above bound on $\|S_t\|_{H_t^{-1}}$ to complete the proof. ∎

E FORMAL STATEMENT OF REGRET BOUND FOR EVILL

In the reminder of these appendices, we will show that under the above assumptions, the following formal version of the regret bound given in Theorem 1 holds. In the theorem below, and in what follows, we use $\vee$ to denote the binary operator on reals that return the maximum of its arguments, and $\wedge$ the minimum.

Theorem 3. *Fix $\delta > 0$ and $n \in \mathbb{N}^+$ and let $\delta' = (\delta/n) \wedge (1/200)$. Consider Algorithm 1 under Assumption 2, with parameters $\lambda = \lambda_n$ and $a = \gamma_n$ given by*

$$\lambda_n = 1 \vee \frac{2dM}{S} \log\left(e\sqrt{1+nL/d} \vee 1/\delta\right)$$

and

$$\gamma_n = \sqrt{\lambda_n}\left(\frac{1}{2M} + S\right) + \frac{2dM}{\sqrt{\lambda_n}} \log\left(e\sqrt{1+nL/d} \vee 1/\delta\right).$$

Suppose further that the prior observations $(X_1, Y_1), \dots, (X_\tau, Y_\tau)$ satisfy

$$\max_{x\in\mathcal{X}} \|x\|_{V_\tau^{-1}} \le b \quad \textit{for} \quad b = \left[22M(1 + MC_d^2(\delta')\widehat{D}_\star)C_d^2(\delta')\gamma_n \cdot \sqrt{\kappa}\right]^{-1},$$

where $V_\tau = \sum_{i=1}^{\tau} X_i X_i^\top + \lambda I$ and

$$\kappa = 1 \vee \max_{|u|\le S} 1/\dot\mu(u), \quad C_d(\delta') = \sqrt{d} + \sqrt{2\log 1/\delta'}, \quad \widehat{D}_\star = \Xi + M\Xi^2 \quad \textit{where} \quad \Xi = \sqrt{2}\left(\frac{1}{2M} + 2S\right),$$

where τ is a stopping time with respect to the filtration induced by $(X_t, Y_t)_{t\in[n]}$. Then, the regret of Algorithm 1 incurred over the n rounds of interaction, denoted $R(n)$, is upper bounded with probability $1 - 3\delta$ as

$$R(n) \preccurlyeq C_d(\delta')\gamma_n\sqrt{n\dot\mu_\star dL\log(1+nL/d \vee 1/\delta)} + C_d^2(\delta')\gamma_n^2 MdL\log(1+nL/d) + \tau\Delta,$$

where $\preccurlyeq$ hides absolute constants and Δ is the maximum per-step regret.

Remark 5. *By our discussion in Appendix B, the τ featuring in the regret bound of Theorem 3 needs only be as large as order d/b^2, which itself is $O(\kappa \cdot (d^{7/2} + d^{3/2}(\log n)^2) \cdot M(1 \vee S^2M^2))$. This is the only term of our regret bound that features κ. Observe also that, since NEF distributions are non-degenerate by definition, and κ depends on the minimum variance of such a distribution over a closed subset of the natural parameter space, κ is finite.*

Layout of proof Throughout our proofs, we work in the setting of Theorem 3, but we may highlight specific parameter choices when these matter. Our proof is laid out as follows:

- In Appendix F we introduce the good event $\mathcal{E}$, a specialisation of the event where the confidence sets of Appendix D hold, and establish properties that hold on it.
- In Appendix G we examine the distribution of θ_t, introducing along the way a second set of good events, $\mathcal{G}_t$. We establish useful properties that hold on $\mathcal{G}_t$.
- In Appendix H we use the properties that hold on $\mathcal{E}$ and $\mathcal{G}_t$ to establish that the parameters θ_t are, in an appropriate sense, optimistic for the true parameters $\theta_\star$ with constant probability.
- In Appendix I, we use the constant probability optimism together with techniques of Faury et al. (2022), extended to the GLB setting, to prove Theorem 3.

We will rely on the notations and definitions of Appendix C. We would also encourage the reader to consult Appendix B for some intuition on self-concordance and the effect of the prior observations.

Filtration and probability In the remainder, we will need the filtration $\mathbb{F} = (\mathbb{F}_0, \mathbb{F}_1, \dots)$, where

$$\mathbb{F}_t = \sigma(X_1, Y_1, Z_1, Z_1', \dots, X_t, Y_t, Z_t, Z_t'),$$

the smallest σ-algebra that makes $X_1, Y_1, Z_1, Z_1', \dots, X_t, Y_t, Z_t, Z_t'$ measurable. Here, and later in the proof, we introduce Z_t, Z_t' for all $t \in [n]$, including $t \leq \tau$. This makes writing the proof easier. In fact, we will think of running the algorithm for all $t \in [n]$, including $t \leq \tau$, except that when $t \leq \tau$, the algorithm 'chooses' X_t as given to it in the prior data. This allows us to define $\hat{\theta}_0$, $\hat{\theta}_t$, W_t and θ_t for all $t \in [n]$. Note also that since $\{\tau \leq t\} \in \sigma(X_1, Y_1, \dots, X_t, Y_t)$ by assumption, $\{\tau \leq t\} \in \mathbb{F}_t$—in other words, τ is an $\mathbb{F}$-stopping time.

We will write $\mathbb{E}_t$ for the $\mathbb{F}_t$-conditional expectation, and $\mathbb{P}_t$ for the $\mathbb{F}_t$-conditional probability. Inequalities of random variables will hold almost surely. Throughout, we index $\mathbb{F}_t$-measurable quantities by t.

A couple useful claims The following claims will be used throughout the proofs. Both follow by rote algebra.

Claim 5. *We have* $\gamma_n/\sqrt{\lambda_n} \leq \Xi$.

Claim 6. *For* $x \in \mathbb{R}$ *and* $b, c \geq 0$, $x^2 \leq bx + c \implies x \leq b + \sqrt{c}$.

F THE GOOD EVENT $\mathcal{E}$

Define the good event

$$\mathcal{E} = \cap_{t=0}^{n} \mathcal{E}_t\,, \tag{16}$$

where

$$\mathcal{E}_t = \left\{ \|g_t(\hat{\theta}_t) - g_t(\theta_\star)\|_{H_t^{-1}} \leq \gamma_n \right\}\,.$$

Note that $\mathcal{E}_t \in \mathbb{F}_t$ holds for all $t \in [n]$. From Lemma 4 and observing that for all $\delta > 0$ and $t \in [n]$, $\gamma_n \geq \gamma_t(\delta, \lambda_n)$, the following is immediate.

Claim 7. *We have* $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$.

The next claim will require a little more work.

Claim 8. *For any* $t \in [n]$*, on* $\mathcal{E}_t$*,* $D(\hat{\theta}_t - \theta_\star) \leq \|\hat{\theta}_t - \theta_\star\| \leq \widehat{D}_\star$ *and consequently the same holds also on* $\mathcal{E}$.

Proof. The first inequality follows from that the pseudonorm D is bounded by the 2-norm (see Equation (12)). Turning to the second equality, for any $t \in \mathbb{N}$, let $G_t = G(\hat{\theta}_t, \theta_\star; \mathcal{D}_t)$. Then,

$$\begin{aligned}
\lambda_n \|\hat{\theta}_t - \theta_\star\|^2 &= \lambda_n \|G_t^{-1}(g_t(\hat{\theta}_t) - g_t(\theta_\star))\|^2 && \text{(Claim 4 and } G_t \succeq \lambda_n I \text{ invertible)} \\
&\leq \|g_t(\hat{\theta}_t) - g_t(\theta_\star)\|^2_{G_t^{-1}} && (G_t \succeq \lambda_n I) \\
&\leq (1 + MD(\hat{\theta}_t - \theta_\star)) \|g_t(\hat{\theta}_t) - g_t(\theta_\star)\|^2_{H_t^{-1}} && \text{(Claim 3 and } H_t = H(\theta_\star; \mathcal{D}_t)) \\
&\leq (1 + M\|\hat{\theta}_t - \theta_\star\|) \|g_t(\hat{\theta}_t) - g_t(\theta_\star)\|^2_{H_t^{-1}}\,. && \text{(by } D(v) \leq \|v\|) \\
&\leq (1 + M\|\hat{\theta}_t - \theta_\star\|)\gamma_n^2\,. && \text{(holds on } \mathcal{E}_t \text{ by the definition of } \mathcal{E}_t) \\
&= \gamma_n^2 + M\gamma_n^2 \|\hat{\theta}_t - \theta_\star\|\,.
\end{aligned}$$

Using Claim 6 with $x = \|\hat{\theta}_t - \theta_\star\|$, we obtain that for all $t \in [n]$,

$$\|\hat{\theta}_t - \theta_\star\| \leq \sqrt{\frac{\gamma_n^2}{\lambda_n}} + M\frac{\gamma_n^2}{\lambda_n},$$

which is bounded by $\widehat{D}_\star = \Xi + M\Xi^2$ by Claim 5. ☐

The above together with our choice of b, Claim 1 and Lemma 3 yield the following claim, which we use repeatedly.

Claim 9. *Fix any $t \in [n]$. With the choice of b prescribed by Theorem 3, on $\mathcal{E}_t \cap \{t \geq \tau\}$,*

$$MD(\hat{\theta}_t - \theta_\star) \leq \frac{1}{20C_d^2(\delta')} \leq \frac{1}{20} \quad \textit{and} \quad \|\theta_\star - \hat{\theta}_t\|_{H_t} \leq \frac{21}{20}\gamma_n.$$

Consequently, for any $x \in \mathcal{X}$, on $\mathcal{E}_t \cap \{t \geq \tau\}$, $\dot{\mu}(x^\top \hat{\theta}_t)$ and $\dot{\mu}(x^\top \theta_\star)$ are within a factor of $e^{1/20} \leq 11/10$ of one another.

Note, we will sometimes use e in place of $e^{1/20}$ or $11/10$ in the above, and 2 in place of $21/20$, so as not to introduce too many unsightly fractions.

Proof. For the first inequality, note that by Cauchy-Schwarz and Claim 1, which can be applied on $\mathcal{E}_t \cap \{t \geq \tau\} \subset \{t \geq \tau\}$, we have

$$MD(\hat{\theta}_t - \theta_\star) \leq M\|x\|_{H_t^{-1}}\|\hat{\theta}_t - \theta_\star\|_{H_t} \leq b\sqrt{\kappa}M\|\hat{\theta}_t - \theta_\star\|_{H_t}$$

and, letting $G_t = G(\hat{\theta}_t, \theta_\star; \mathcal{D}_t)$, on $\mathcal{E}_t$,

$$\begin{aligned}
\|\theta_\star - \hat{\theta}_t\|_{H_t} &= \|G_t^{-1}(g_t(\theta_\star) - g_t(\hat{\theta}_t))\|_{H_t} && \text{(Claim 4)}\\
&\leq (1 + MD(\hat{\theta}_t - \theta_\star))\|g_t(\theta_\star) - g_t(\hat{\theta}_t)\|_{H_t^{-1}} && \text{(by Claim 9, } H_t \preceq (1 + MD(\hat{\theta}_t - \theta_\star))G_t)\\
&\leq (1 + MD(\hat{\theta}_t - \theta_\star))\gamma_n && \text{(Definition on } \mathcal{E}_t)\\
&\leq (1 + M\widehat{D}_\star)\gamma_n. && \text{(Claim 8)}
\end{aligned}$$

Chaining the inequalities obtained, together with the definition of b, and $C_d(\delta') \geq 1$ gives the first inequalities. For the second inequality, apply the first inequality at the penultimate line of the above display.

For the final conclusion, note that by Lemma 3 and the first part of the result,

$$\max\left(\frac{\dot{\mu}(x^\top \hat{\theta}_t)}{\dot{\mu}(x^\top \theta_\star)}, \frac{\dot{\mu}(x^\top \theta_\star)}{\dot{\mu}(x^\top \hat{\theta}_t)},\right) \leq \exp(MD(\theta_\star - \hat{\theta}_t)) \leq \exp(1/20)\,. \qquad \square$$

G THE OTHER GOOD EVENTS, $\mathcal{G}_t$

The next lemma follows from considering the first order optimality conditions at θ_t and $\hat{\theta}_{t-1}$ and using Claim 4.

Lemma 6. *For any $t \in [n]$,*

$$-W_t = g_{t-1}(\theta_t) - g_{t-1}(\hat{\theta}_{t-1}) = G(\theta_t, \hat{\theta}_{t-1}; \mathcal{D}_{t-1})(\theta_t - \hat{\theta}_{t-1})\,. \tag{17}$$

Furthermore, with $a = \gamma_n$, EVILL induces parameter perturbations of the form

$$G(\theta_t, \hat{\theta}_{t-1}; \mathcal{D}_{t-1})(\theta_t - \hat{\theta}_{t-1}) = \gamma_n H(\hat{\theta}_{t-1}; \mathcal{D}_{t-1})^{1/2} A_t \tag{18}$$

where $A_t \in \mathbb{R}^d$ is such that its distribution, given $\mathbb{F}_{t-1}$, is $\mathcal{N}(0, I)$.

Proof. We start by establishing Equation (17). From the first order optimality conditions and the definition of g_{t-1} (copying partly from Equation (9)),

$$g_{t-1}(\hat{\theta}_{t-1}) = \sum_{i=1}^{t} X_i Y_i = g_{t-1}(\theta_t) + W_t\,,$$

where

$$W_t = a\lambda^{1/2} Z_t + a \sum_{i=1}^{t-1} \dot{\mu}(X_i^\top \hat{\theta}_{t-1})^{1/2} Z'_{t,i} X_i$$

and $Z_t \sim \mathcal{N}(0, I_d)$, $Z'_t \sim \mathcal{N}(0, I_{t-1})$. Hence,

$$-W_t = g_{t-1}(\theta_t) - g_{t-1}(\hat{\theta}_{t-1}) = G(\theta_t, \hat{\theta}_{t-1}; \mathcal{D}_{t-1})(\theta_t - \hat{\theta}_{t-1})\,,$$

where the last equality is from Claim 4. This proves Equation (17).

Now, notice that, given $\mathbb{F}_{t-1}$, W_t is zero mean Gaussian with covariance $a^2 H(\hat\theta_{t-1}; \mathcal{D}_{t-1})$. It follows that, given $\mathbb{F}_{t-1}$, $A_t = \frac{1}{a} H(\hat\theta_{t-1}; \mathcal{D}_{t-1})^{-1/2}(-W_t)$ is zero mean Gaussian with identity covariance. $\square$

We define the event $\mathcal{G}_t$ to be that on which the norm of A_t is not too extreme. In particular, recalling that $C_d(\delta') = \sqrt{d} + \sqrt{2\log 1/\delta'}$, we set

$$\mathcal{G}_t = \{\|A_t\| \le C_d(\delta')\}.$$

From the standard Gaussian concentration bound (e.g., Theorem II.6 of Davidson and Szarek (2001)), we have the following bound for the probability of $\mathcal{G}_t$:

Claim 10. *We have* $\mathbb{P}_{t-1}(\mathcal{G}_t) \ge 1 - \delta'$.

Mirroring Appendix F, we now upper bound $D(\theta_t - \hat\theta_{t-1})$ on $\mathcal{G}_t$.

Claim 11. *With the choice* $a = \gamma_n$, $\lambda = \lambda_n$, *on the event* $\mathcal{G}_t \cap \{t > \tau\}$,

$$D(\theta_t - \hat\theta_{t-1}) \le \|\theta_t - \hat\theta_{t-1}\| \le C_d^2(\delta')\widehat{D}_\star.$$

Proof. Again, the first inequality follows from $D(v) \le \|v\|$ with $v = \theta_t - \hat\theta_{t-1}$ (see Equation (12)).

For the second inequality, introduce the shorthand $\widehat{H}_{t-1} = H(\hat\theta_{t-1}; \mathcal{D}_{t-1})$ and $G_t = G(\theta_t, \hat\theta_{t-1}; \mathcal{D}_{t-1})$. Let also S_2^{d-1} be the unit sphere in $\mathbb{R}^d$: $S_2^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$. Then, we have the following:

$$\begin{aligned}
\lambda_n\|\theta_t - \hat\theta_{t-1}\|^2 &= \lambda_n\gamma_n^2\|G_t^{-1}\widehat{H}_{t-1}^{1/2}A_t\|^2 && \text{(Equation (18) of Lemma 6)}\\
&= \gamma_n^2\|G_t^{-1/2}\widehat{H}_{t-1}^{1/2}A_t\|^2 && (G_t = G(\theta_t, \hat\theta_{t-1}; \mathcal{D}_{t-1}) \succeq \lambda_n I)\\
&\le \gamma_n^2 C_d^2(\delta')\|G_t^{-1/2}\widehat{H}_{t-1}^{1/2}\|^2 && \text{(definition of } \mathcal{G}_t)\\
&= \gamma_n^2 C_d^2(\delta') \sup_{x\in S_2^{d-1}} x^\top \widehat{H}_{t-1}^{1/2} G_t^{-1}\widehat{H}_{t-1}^{1/2} x\\
&\le \gamma_n^2 C_d^2(\delta')\,(1 + MD(\theta_t - \hat\theta_{t-1})) && \text{(Claim 3)}\\
&\le \gamma_n^2 C_d^2(\delta')\,(1 + M\|\theta_t - \hat\theta_{t-1}\|). && (D(v) \le \|v\|)\\
&= \gamma_n^2 C_d^2(\delta') + M\gamma_n^2 C_d^2(\delta')\|\theta_t - \hat\theta_{t-1}\|.
\end{aligned}$$

Now, by Claim 6,

$$\|\theta_t - \hat\theta_{t-1}\| \le C_d(\delta')\sqrt{\frac{\gamma_n^2}{\lambda_n}} + MC_d^2(\delta')\frac{\gamma_n^2}{\lambda_n}.$$

Since $C_d(\delta') \ge \sqrt{d} \ge 1$, by Claim 5, this is upper bounded by $C_d^2(\delta')\widehat{D}_\star$. $\square$

Claim 12. *With the choice of b prescribed by Theorem 3, for all* $t \in [n]$, *on* $\mathcal{E}_{t-1} \cap \mathcal{G}_t \cap \{t > \tau\}$,

$$MD(\theta_t - \hat\theta_{t-1}) \le \frac{1}{20} \quad \textit{and} \quad \|\theta_t - \hat\theta_{t-1}\|_{H(\hat\theta_{t-1};\mathcal{D}_{t-1})} \le \frac{21}{20}C_d(\delta')\gamma_n.$$

Consequently, for any $x \in \mathcal{X}$, $\dot\mu(x^\top\theta_t)$ *and* $\dot\mu(x^\top\hat\theta_{t-1})$ *are within a factor of* $e^{1/20} \le 11/10$ *of one another.*

Again, we will sometimes use e in place of $e^{1/20}$ or $11/10$ in the above, and 2 in place of $21/20$.

Proof. Again write $\widehat{H}_{t-1} = H(\hat\theta_{t-1}; \mathcal{D}_{t-1})$ and $G_t = G(\theta_t, \hat\theta_{t-1}; \mathcal{D}_{t-1})$. For the first inequality, letting $X = \arg\max_{y\in\mathcal{X}} |y^\top(\theta_t - \hat\theta_{t-1})|$,

$$\begin{aligned}
MD(\theta_t - \hat\theta_{t-1}) &\le M\|X\|_{\widehat{H}_{t-1}^{-1}}\|\theta_t - \hat\theta_{t-1}\|_{\widehat{H}_{t-1}} && \text{(Cauchy-Schwarz)}\\
&\le \frac{11}{10}M\|X\|_{H_{t-1}^{-1}}\|\theta_t - \hat\theta_{t-1}\|_{\widehat{H}_{t-1}} && \text{(by Claim 9, } H_{t-1} \preceq \tfrac{11}{10}\widehat{H}_{t-1} \text{ on } \mathcal{E}_{t-1} \cap \{t > \tau\})\\
&\le \frac{11}{10}b\sqrt{\kappa}M\|\theta_t - \hat\theta_{t-1}\|_{\widehat{H}_{t-1}} && \text{(on } \{t > \tau\}\text{, by Claim 1)}
\end{aligned}$$

and

$$\begin{aligned}
\|\theta_t - \hat{\theta}_{t-1}\|_{\widehat{H}_{t-1}} &= \gamma_n \|\widehat{H}_{t-1}^{1/2} G_t^{-1} \widehat{H}_{t-1}^{1/2} A_t\| \\
&\leq \gamma_n C_d(\delta') \|\widehat{H}_{t-1}^{1/2} G_t^{-1} \widehat{H}_{t-1}^{1/2}\| && \text{(on } \mathcal{G}_t\text{)} \\
&\leq \gamma_n C_d(\delta')(1 + MD(\theta_t - \hat{\theta}_{t-1})) && \text{(Claim 3)} \\
&\leq \gamma_n C_d(\delta')(1 + C_d^2(\delta') M \widehat{D}_\star). && \text{(Claim 11)}
\end{aligned}$$

The first inequality then follows from chaining the last two displays and using the definition of b. For the second inequality, apply the first inequality at the penultimate line of the above display. The final part follows with the same proof as the analogue statement in Claim 9. □

H LOWER BOUNDING THE PROBABILITY OF OPTIMISM

The aim of this appendix is to show that each θ_t is optimistic with at least a constant probability, where optimistic means that it will belong to the set of optimistic parameters, defined as follows.

Definition 2 (Optimistic parameters)**.** *We let Θ^{opt} denote the set of parameters optimistic for $\theta_\star$, defined as*

$$\Theta^{opt} = \{\theta \in \mathbb{R}^d \colon \max_{x \in \mathcal{X}} x^\top \theta \geq x_\star^\top \theta_\star\}.$$

This proof of constant optimism is where Kveton et al. (2020b) needed to introduce the orthogonality assumption, that we do away with by employing tools based on self-concordance from Sun and Tran-Dinh (2019) and Faury et al. (2022). This appendix is the main technical contribution in the context of proving Theorem 3.

In this section, we will use the following shorthands:

$$G_t = G(\theta_t, \hat{\theta}_{t-1}; \mathcal{D}_t), \qquad \widehat{H}_{t-1} = H(\hat{\theta}_{t-1}; \mathcal{D}_{t-1}), \qquad H_{t-1} = H(\theta_\star, \mathcal{D}_{t-1})$$

(the shorthands $\widehat{H}_{t-1}$ and H_{t-1} were introduced beforehand.) We will also work with the quantity

$$E_t = G_t - \widehat{H}_{t-1}\,.$$

We start with a result concerning E_t:

Claim 13. *If $MD(\theta_t - \hat{\theta}_{t-1}) \leq 1$, then $\|\widehat{H}_{t-1}^{-1/2} E_t \widehat{H}_{t-1}^{-1/2}\| \leq MD(\theta_t - \hat{\theta}_{t-1})$.*

Proof. Choose $x \in \mathbb{R}^d$ with $\|x\| = 1$ such that $\|\widehat{H}_{t-1}^{-1/2} E_t \widehat{H}_{t-1}^{-1/2}\| = x^\top \widehat{H}_{t-1}^{-1/2} E_t \widehat{H}_{t-1}^{-1/2} x$, and write $y = \widehat{H}_{t-1}^{-1/2} x$. By Claim 2,

$$\begin{aligned}
y^\top E_t y &= y^\top \sum_{i=1}^{t} \Big(\alpha(X_i^\top \theta_t, X_i^\top \hat{\theta}_{t-1}) - \dot{\mu}(X_i^\top \hat{\theta}_{t-1})\Big) X_i X_i^\top y \\
&\leq \Big(h(MD(\theta_t - \hat{\theta}_{t-1})) - 1\Big) \cdot y^\top \sum_{i=1}^{t} \dot{\mu}(X_i^\top \hat{\theta}_{t-1}) X_i X_i^\top y.
\end{aligned}$$

Now, we note that for $z \leq 1$, $h(z) - 1 = \frac{e^z - 1}{z} - 1 \leq z$, and recall that we assumed $MD(\theta_t - \hat{\theta}_{t-1}) \leq 1$. Thus,

$$\|\widehat{H}_{t-1}^{-1/2} E_t \widehat{H}_{t-1}^{-1/2}\| \leq MD(\theta_t - \hat{\theta}_{t-1}) \cdot x^\top \widehat{H}_{t-1}^{-1/2} (\widehat{H}_{t-1} - \lambda I) \widehat{H}_{t-1}^{-1/2} x \leq MD(\theta_t - \hat{\theta}_{t-1}).$$ □

Lemma 7. *Fix $t \in [n]$. On the event $\mathcal{E} \cap \{t > \tau\}$,*

$$\mathbb{P}_{t-1}\left(\{\theta_t \in \Theta^{opt}\} \cap \mathcal{G}_t\right) \geq 1/200.$$

Proof. Consider an arbitrary $t \in [n]$. Write $\widetilde{\mathbb{P}}_{t-1}(\,\cdot\,) = \mathbb{P}_{t-1}(\,\cdot\, \cap \mathcal{G}_t)$. Let $\tilde{\mathcal{E}}_{t-1} = \{\|\theta_\star - \hat\theta_{t-1}\|_{H_{t-1}} \le 2\gamma_n\}$. By Claim 9, $\mathcal{E}_{t-1} \cap \{t-1 \ge \tau\} = \mathcal{E}_{t-1} \cap \{t > \tau\} \subset \tilde{\mathcal{E}}_{t-1}$. Hence, on $\mathcal{E} \cap \{t > \tau\}$,

$$\begin{aligned}
p_{t-1} &\doteq \widetilde{\mathbb{P}}_{t-1}\left(\theta_t \in \Theta^{\text{opt}}\right) \\
&= \widetilde{\mathbb{P}}_{t-1}\left(X_t^\top \theta_t \ge x_\star^\top \theta_\star\right) \\
&\ge \widetilde{\mathbb{P}}_{t-1}\left(x_\star^\top(\theta_t - \hat\theta_{t-1}) \ge x_\star^\top(\theta_\star - \hat\theta_{t-1})\right) && \text{(definition of } X_t) \\
&\ge \widetilde{\mathbb{P}}_{t-1}\left(x_\star^\top(\theta_t - \hat\theta_{t-1}) \ge x_\star^\top(\theta_\star - \hat\theta_{t-1})\,, \tilde{\mathcal{E}}_{t-1}\right) \\
&\ge \widetilde{\mathbb{P}}_{t-1}\left(x_\star^\top(\theta_t - \hat\theta_{t-1}) \ge 2\gamma_n \|x_\star\|_{H_{t-1}^{-1}}\,, \tilde{\mathcal{E}}_{t-1}\right), && \text{(Cauchy-Schwarz and the definition of } \tilde{\mathcal{E}}_{t-1}) \\
&= \mathbf{1}_{\tilde{\mathcal{E}}_{t-1}} \widetilde{\mathbb{P}}_{t-1}\left(x_\star^\top(\theta_t - \hat\theta_{t-1}) \ge 2\gamma_n \|x_\star\|_{H_{t-1}^{-1}}\right), && (\tilde{\mathcal{E}}_{t-1} \in \mathbb{F}_{t-1}) \\
&= \widetilde{\mathbb{P}}_{t-1}\left(x_\star^\top(\theta_t - \hat\theta_{t-1}) \ge 2\gamma_n \|x_\star\|_{H_{t-1}^{-1}}\right). && \text{(we are on } \mathcal{E} \cap \{t > \tau\} \text{ and } \mathcal{E} \cap \{t > \tau\} \subset \tilde{\mathcal{E}}_{t-1})
\end{aligned}$$

By Lemma 6, $x_\star^\top(\theta_t - \hat\theta_{t-1}) = x_\star^\top G_t^{-1} \widehat{H}_{t-1}^{1/2} A_t$, where, as before, $G_t = G(\theta_t, \hat\theta_{t-1}; \mathcal{D}_{t-1})$. By the Woodbury matrix identity, letting $E_t = G_t - \widehat{H}_{t-1}$,

$$G_t^{-1} = (\widehat{H}_{t-1} + E_t)^{-1} = \widehat{H}_{t-1}^{-1} - \widehat{H}_{t-1}^{-1} E_t (\widehat{H}_{t-1} + E_t)^{-1} = \widehat{H}_{t-1}^{-1} - \widehat{H}_{t-1}^{-1} E_t G_t^{-1},$$

and so we have that

$$\begin{aligned}
p_{t-1} &\ge \widetilde{\mathbb{P}}_{t-1}\left(x_\star^\top \widehat{H}_{t-1}^{-1/2} A_t \ge x_\star^\top \widehat{H}_{t-1}^{-1} E_t G_t^{-1} \widehat{H}_{t-1}^{1/2} A_t + 2\gamma_n \|x_\star\|_{H_{t-1}}\right) \\
&\ge \widetilde{\mathbb{P}}_{t-1}\left(x_\star^\top \widehat{H}_{t-1}^{-1/2} A_t \ge x_\star^\top \widehat{H}_{t-1}^{-1} E_t G_t^{-1} \widehat{H}_{t-1}^{1/2} A_t + 2\gamma_n \|x_\star\|_{H_{t-1}}, \mathcal{E}_{t-1} \cap \{t > \tau\}\right). \qquad (19)
\end{aligned}$$

Using Cauchy-Schwarz, and then Lemma 6 and Claim 12, on $\mathcal{G}_t$,

$$\begin{aligned}
x_\star^\top \widehat{H}_{t-1}^{-1} E_t G_t^{-1} \widehat{H}_{t-1}^{1/2} A_t &\le \|x_\star\|_{\widehat{H}_{t-1}^{-1}} \|\widehat{H}_{t-1}^{-1/2} E_t \widehat{H}_{t-1}^{-1/2}\| \|G_t^{-1} \widehat{H}_{t-1}^{1/2} A_t\|_{\widehat{H}_{t-1}} \\
&= \|x_\star\|_{\widehat{H}_{t-1}^{-1}} \|\widehat{H}_{t-1}^{-1/2} E_t \widehat{H}_{t-1}^{-1/2}\| \|\theta_t - \hat\theta_{t-1}\|_{\widehat{H}_{t-1}} \\
&\le 2\gamma_n C_d(\delta') \|x_\star\|_{\widehat{H}_{t-1}^{-1}} \|\widehat{H}_{t-1}^{-1/2} E_t \widehat{H}_{t-1}^{-1/2}\|\,.
\end{aligned}$$

Now, observe that on $\mathcal{G}_t \cap \mathcal{E}_{t-1} \cap \{t > \tau\}$, by Claim 9 and Claim 12, $MD(\theta_t - \hat\theta_{t-1}) \le \frac{1}{20 C_d(\delta')} \le \frac{1}{20}$ and $MD(\hat\theta_{t-1} - \theta_\star) \le \frac{1}{20}$. The first of these lets us apply Claim 13 to obtain that

$$\|\widehat{H}_{t-1}^{-1/2} E_t \widehat{H}_{t-1}^{-1/2}\| \le \frac{1}{20 C_d(\delta')}, \quad \text{which gives the bound} \quad x_\star^\top \widehat{H}_{t-1}^{-1} E_t G_t^{-1} \widehat{H}_{t-1}^{1/2} A_t \le \frac{1}{10} \gamma_n \|x_\star\|_{\widehat{H}_{t-1}^{-1}}.$$

Combining the first with the second, $MD(\theta_t - \theta_\star) \le MD(\theta_t - \hat\theta_{t-1}) + MD(\hat\theta_{t-1} - \theta_\star) \le \frac{1}{10}$, and so by Lemma 3, $\|x_\star\|_{H_{t-1}^{-1}} \le \sqrt{e^{\frac{1}{10}}} \|x_\star\|_{\widehat{H}_{t-1}^{-1}} \le \frac{11}{10} \|x_\star\|_{\widehat{H}_{t-1}^{-1}}$ (note the norm changed from H_{t-1}^{-1}-weighted to $\widehat{H}_{t-1}^{-1}$-weighted). Using these two bounds in Equation (19), on $\mathcal{E} \cap \{t > \tau\}$, we have that

$$\begin{aligned}
p_{t-1} &\ge \widetilde{\mathbb{P}}_{t-1}\left(x_\star^\top \widehat{H}_{t-1}^{-1/2} A_t \ge \frac{23}{10} \gamma_n \|x_\star\|_{\widehat{H}_t^{-1}}, \mathcal{E}_{t-1} \cap \{t > \tau\}\right) \\
&= \mathbf{1}_{\mathcal{E}_{t-1} \cap \{t > \tau\}} \widetilde{\mathbb{P}}_{t-1}\left(x_\star^\top \widehat{H}_{t-1}^{-1/2} A_t \ge \frac{23}{10} \gamma_n \|x_\star\|_{\widehat{H}_t^{-1}}\right) && (\mathcal{E}_{t-1}, \{t > \tau\} \in \mathbb{F}_{t-1}, \text{ as } \tau \text{ is an } \mathbb{F}\text{-stopping time}) \\
&= \widetilde{\mathbb{P}}_{t-1}\left(x_\star^\top \widehat{H}_{t-1}^{-1/2} A_t \ge \frac{23}{10} \gamma_n \|x_\star\|_{\widehat{H}_t^{-1}}\right) && (\mathcal{E} \cap \{t > \tau\} \subset \mathcal{E}_{t-1} \cap \{t > \tau\}) \\
&\ge \mathbb{P}_{t-1}\left(x_\star^\top \widehat{H}_{t-1}^{-1/2} A_t \ge \frac{23}{10} \gamma_n \|x_\star\|_{\widehat{H}_t^{-1}}\right) - (1 - \mathbb{P}_{t-1}(\mathcal{G}_t)) \\
&\ge \mathbb{P}_{t-1}\left(x_\star^\top \widehat{H}_{t-1}^{-1/2} A_t \ge \frac{23}{10} \gamma_n \|x_\star\|_{\widehat{H}_t^{-1}}\right) - \delta',
\end{aligned}$$

where the middle inequality is a standard result for lower bounding the probability of an intersection of two events, and the third follows from Claim 10. Now, observe that conditioned on $\mathbb{F}_{t-1}$, $x_\star^\top \widehat{H}_{t-1}^{-1/2} A_t$ is a centred Gaussian random variable with standard deviation $\gamma_n \|x_\star\|_{\widehat{H}_t^{-1}}$. Thus, together with our assumption on δ', looking up the relevant Gaussian tail probability, we conclude that, on $\mathcal{E} \cap \{t > \tau\}$,

$$p_{t-1} \geq \mathbb{P}_{t-1}\left(x_\star^\top \widehat{H}_{t-1}^{-1/2} A_t \geq \frac{23}{10}\gamma_n \|x_\star\|_{\widehat{H}_t^{-1}} \right) - \delta' = \mathcal{N}(0,1)(\{z \in \mathbb{R} \colon z > 23/10\}) - \delta' \geq 1/200. \qquad \square$$

I PROOF OF REGRET BOUND FOR EVILL

Much of the proofs in this final appendix largely follows the regret bound given by Faury et al. (2022) for their Thompson-sampling-type algorithm for the logistic bandit setting, but fixes an error present therein.[4]

To start our bound, we write $\widetilde{R}(n)$ for the regret over steps $\tau+1, \ldots, n$, and break this up as

$$\widetilde{R}(n) = \sum_{t=\tau+1}^{n} \underbrace{\mu(x_\star^\top \theta_\star) - \mu(X_t^\top \theta_t)}_{r_t^{\mathrm{OPT}}} + \underbrace{\mu(X_t^\top \theta_t) - \mu(X_t^\top \theta_\star)}_{r_t^{\mathrm{PRED}}}. \tag{20}$$

The first term is controlled if the algorithm plays optimistically, while the second is controlled if the reward parameter of the arm chosen is well predicted by the parameter vector that governs the action choice. Hence, the naming of these two terms.

The following two lemmas, proven in Appendix I.1 and Appendix I.2 respectively, bound these two terms.

Lemma 8. *Let $p = 1/200$ and fix $t \in [n]$. On the event $\mathcal{E} \cap \mathcal{G}_t \cap \{t > \tau\}$, for δ' and b as in Theorem 3,*

$$r_t^{\mathrm{OPT}} \leq 7 C_d(\delta') \gamma_n \sqrt{\dot\mu(x_\star^\top \theta_\star)} \mathbb{E}_{t-1}[\sqrt{\dot\mu(X_t^\top \hat\theta_{t-1})} \|X_t\|_{\widehat{H}_{t-1}^{-1}}]/p.$$

Lemma 9. *Fix $t \in [n]$. On event $\mathcal{E} \cap \mathcal{G}_t \cap \{t > \tau\}$, for b as in Theorem 3,*

$$r_t^{\mathrm{PRED}} \leq 5(1 + C_d(\delta'))\gamma_n \dot\mu(X_t^\top \hat\theta_{t-1}) \|X_t\|_{\widehat{H}_{t-1}^{-1}}.$$

Our proof of Theorem 3 will also use the following simple claim, proven in Appendix I.3, and the version of the elliptical potential lemma stated thereafter, which follows from Lemma 15 of Faury et al. (2020) combined with the usual trace-determinant inequality, say, Note 1, Section 20.2 of Lattimore and Szepesvári (2020).

Claim 14. $\sum_{t=\tau+1}^n \dot\mu(X_t^\top \theta_\star) \leq n \dot\mu(x_\star^\top \theta_\star) + M \widetilde{R}(n)$.

Lemma 10 (Elliptical potential lemma)**.** *Fix $\lambda, A > 0$. Let $\{a_t\}_{t=1}^\infty$ be a sequence in AB_2^d and let $V_0 = \lambda I$. Define $V_{t+1} = V_t + a_{t+1}a_{t+1}^\top$ for each $t \in \mathbb{N}$. Then, for all $n \in \mathbb{N}^+$,*

$$\sum_{t=1}^n \|a_t\|_{V_{t-1}^{-1}}^2 \leq 2d \max\left\{1, \frac{A^2}{\lambda}\right\} \log\left(1 + \frac{nA^2}{d\lambda}\right).$$

Proof of Theorem 3. Assume that $\mathcal{E}$, defined via Equation (16), holds. We start with the regret decomposition of Equation (20), use Lemmas 8 and 9 to bound the two terms respectively, and write the result in terms of

$$M_t = \mathbb{E}_{t-1}\left[\sqrt{\dot\mu(X_t^\top \hat\theta_{t-1})} \|X_t\|_{\widehat{H}_{t-1}^{-1}}\right] - \sqrt{\dot\mu(X_t^\top \hat\theta_{t-1})} \|X_t\|_{\widehat{H}_{t-1}^{-1}}.$$

That is, using $\preccurlyeq$ to indicate inequalities up to absolute constants (including $p = 1/200$),

$$\widetilde{R}(n) \preccurlyeq \gamma_n C_d(\delta') \sum_{t=1}^n 1\{t > \tau\} \left(\underbrace{\dot\mu(X_t^\top \hat\theta_{t-1}) \|X_t\|_{\widehat{H}_{t-1}^{-1}}}_{A_t} + \sqrt{\dot\mu(x_\star^\top \theta_\star)} \underbrace{\sqrt{\dot\mu(X_t^\top \hat\theta_{t-1})} \|X_t\|_{\widehat{H}_{t-1}^{-1}}}_{B_t} + \sqrt{\dot\mu(x_\star^\top \theta_\star)} M_t \right).$$

[4] We thank Marc Abeille for helping us fix the error in the Faury et al. (2022) manuscript.

Consider first bounding $\sum_{t=1}^n 1\{t > \tau\} M_t$. For this bound we drop the assumption that $\mathcal{E}$ holds: we give an upper bound that holds with probability $1-\delta$ (so for the final bound we will need to use an extra union bound with the error event introduced in this step). The sum is upper bounded with the help of the Azuma-Hoeffding's inequality (say, Corollary 2.20 in Wainwright (2019)). For this, notice that $(1\{t > \tau\} M_t)_{t\in[n]}$ is a martingale difference sequence with respect to $\mathbb{F}$ (this follows because $1\{t>\tau\} = 1\{t-1 \geq \tau\} \in \mathbb{F}_{t-1}$ as τ is an $\mathbb{F}$-stopping time) and each M_t satisfies $|M_t| \leq 2\sqrt{L}$. This latter follows because $\|X_t\|_{\widehat{H}_{t-1}^{-1}} \leq \|X_t\|/\lambda \leq \|X_t\| \leq 1$ since by choice $\lambda \geq 1$. Then, from Azuma-Hoeffding, we conclude that with probability $1-\delta$,

$$\sum_{t=1}^n 1\{t > \tau\} M_t \leq \sqrt{8Ln \log 2/\delta}\,.$$

Now consider bounding $\sum_{t=1}^n 1\{t > \tau\} A_t$. We upper-bound this sum on $\mathcal{E}$. By Claim 9, which can be used since $\mathcal{E}_t \subset \mathcal{E}$,

$$\dot\mu(X_t^\top \hat\theta_{t-1}) \leq \sqrt{e\dot\mu(X_t^\top \theta_\star)\dot\mu(X_t^\top \hat\theta_{t-1})}.$$

Using this and Cauchy-Schwarz,

$$\sum_{t=1}^n 1\{t > \tau\} A_t \leq \sqrt{e\sum_{t=\tau+1}^n \dot\mu(X_t^\top \theta_\star)}\sqrt{\sum_{t=\tau+1}^n \dot\mu(X_t^\top \hat\theta_{t-1})\|X_t\|^2_{\widehat{H}_{t-1}^{-1}}}$$

Now, by Claim 14, $\sum_{t=\tau+1}^n \dot\mu(X_t^\top\theta_\star) \leq n\dot\mu(x_\star^\top\theta_\star) + M\widetilde{R}(n)$. Furthermore, by Lemma 10 applied to the sequence $a_1, a_2, \dots$ given by $a_t = \sqrt{\dot\mu(X_t^\top \hat\theta_{t-1})}X_t$, using $\|a_t\| \leq \sqrt{L}$, we get

$$\sum_{t=\tau+1}^n \dot\mu(X_t^\top \hat\theta_{t-1})\|X_t\|^2_{\widehat{H}_{t-1}^{-1}} \leq 2dL \log\left(1 + \frac{nL}{d\lambda_n}\right) =: Q_n.$$

Now, with an entirely analogous argument that is omitted as it would just repeat the previous argument, we get than on $\mathcal{E}$, $\sum_{t=1}^n 1\{t > \tau\} B_t \leq Q_n$ also holds.

Putting together these three bounds and using the concavity of the square root function, we have that on $(\cap_{t=1}^n \mathcal{G}_t) \cap \mathcal{E} \cap \{t > \tau\}$, with probability $1-\delta$,

$$\widetilde{R}(n) \preccurlyeq \gamma_n C_d(\delta')\left(\sqrt{Q_n M \widetilde{R}(n)} + \sqrt{n\dot\mu_\star(Q_n + L\log 2/\delta)}\right).$$

Now, note the above is of the form $x^2 \leq bx + c$ for $x^2 = \widetilde{R}(n)$ and $b, c \geq 0$. Applying Claim 6 and squaring, we conclude that on $(\cap_{t=1}^n \mathcal{G}_t) \cap \mathcal{E} \cap \{t > \tau\}$, with probability $1-\delta$,

$$\widetilde{R}(n) \preccurlyeq C_d(\delta')\gamma_n\sqrt{n\dot\mu_\star(Q_n + L\log 2/\delta)} + C_d^2(\delta')\gamma_n^2 M Q_n\,.$$

We then take $R(n) \leq \widetilde{R}(n) + \tau\Delta$, where Δ is the maximal per-step regret, and observe that, by Claim 7 and Claim 10, with our choice of δ', $\mathbb{P}((\cap_{t=1}^n \mathcal{G}_t) \cap \mathcal{E}) \geq 1 - 2\delta$. We then apply a union bound for a $1-3\delta$ probability bound, and simplify somewhat. □

I.1 Bounding r_t^{OPT} (proof of Lemma 8)

For this proof, we will work with the function $J(\theta) = \max_{x\in\mathcal{X}} x^\top\theta$. In the convex analysis literature, the function J is called the *support function* of $\mathcal{X}$. It is well known that the support function J is convex, and for $\mathcal{X} \subset \mathbb{R}^d$ compact, any $\theta \in \mathbb{R}^d$, any $x \in \arg\max_{y\in\mathcal{X}} y^\top\theta$ is a subgradient of J at θ, but we include a proof for completeness.

Claim 15. *Let J be the support function of a compact set $\mathcal{X} \subset \mathbb{R}^d$. Then for any $\theta \in \mathbb{R}^d$, any $x \in \arg\max_{y\in\mathcal{X}} y^\top\theta$ is a subgradient of J at θ.*

Proof. Note that for any $\theta' \in \mathbb{R}^d$,

$$J(\theta) + x^\top(\theta' - \theta) = x^\top \theta' \leq \max_{y \in \mathcal{X}} y^\top \theta' = J(\theta'),$$

which is the definition of x being a subgradient of J at θ. □

Proof of Lemma 8. From definition of α,

$$r_t^{\mathrm{OPT}} = \mu(x_\star^\top \theta_\star) - \mu(X_t^\top \theta_t) = \alpha(x_\star^\top \theta_\star, X_t^\top \theta_t)(x_\star^\top \theta_\star - X_t^\top \theta_t) = \alpha(J(\theta_\star), J(\theta_t))(J(\theta_\star) - J(\theta_t)).$$

Assume from now on that $\mathcal{E} \cap \mathcal{G}_t \cap \{t > \tau\}$ holds.

Using the convexity of J, that X_t is a subgradient of J at θ_t, and that $x_\star$ is a subgradient of J at $\theta_\star$,

$$M|J(\theta_\star) - J(\theta_t)| \leq M \max\{|x_\star^\top(\theta_\star - \theta_t)|, |X_t^\top(\theta_\star - \theta_t)|\} \leq MD(\theta_t - \theta_\star) \leq MD(\theta_t - \hat\theta_{t-1}) + MD(\hat\theta_{t-1} - \theta_\star) \leq 1,$$

where the last inequality follows by Claim 9 and Claim 12. Now, for $\alpha(J(\theta_\star), J(\theta_t))$, using self-concordance, that is Lemma 3 together with the last display, bounding crudely, we have that $\alpha(J(\theta_\star), J(\theta_t)) \leq 2\dot\mu(J(\theta_\star))$, and also that $\alpha(J(\theta_\star), J(\theta_t)) \leq 2\dot\mu(J(\theta_t))$, and so

$$r_t^{\mathrm{OPT}} \leq 2\sqrt{\dot\mu(x_\star^\top \theta_\star)} \cdot \sqrt{\dot\mu(X_t^\top \theta_t)}(J(\theta_\star) - J(\theta_t)). \tag{21}$$

We turn to bounding $\sqrt{\dot\mu(X_t^\top \theta_t)}(J(\theta_\star) - J(\theta_t))$. Let

$$\Theta^{\mathrm{opt}} = \{\theta : J(\theta) \geq J(\theta_\star)\} \quad \text{and} \quad \Theta_{t-1} = \left\{\theta \in \mathbb{R}^d \colon \|\theta - \hat\theta_{t-1}\|_{\widehat{H}_{t-1}} \leq \tfrac{21}{20} C_d(\delta')\gamma_n\right\},$$

and observe that Θ^{opt} matches that defined in Definition 2. Now let $p_{t-1} = \mathbb{P}_{t-1}(\theta_t \in \Theta^{\mathrm{opt}} \cap \Theta_{t-1})$ and

$$Q(\,\cdot\,) = \begin{cases} \mathbb{P}_{t-1}(\theta_t \in \cdot \cap \Theta^{\mathrm{opt}} \cap \Theta_{t-1})/p_{t-1}\,, & p_{t-1} > 0\,; \\ \text{an arbitrary probability measure}\,, & \text{otherwise.} \end{cases}$$

On $\mathcal{E} \cap \{t > \tau\}$, we claim that

$$p_{t-1} \geq p = 1/200 > 0 \tag{22}$$

hence we are in the first case of the above definition. Indeed,

$$\begin{aligned}
p_{t-1} &= \mathbb{P}_{t-1}(\theta_t \in \Theta^{\mathrm{opt}} \cap \Theta_{t-1}) \\
&= \mathbb{P}_{t-1}(\{\theta_t \in \Theta^{\mathrm{opt}}\} \cap \{\theta_t \in \Theta_{t-1}\}) \\
&\geq \mathbb{P}_{t-1}(\{\theta_t \in \Theta^{\mathrm{opt}}\} \cap \mathcal{G}_t \cap \{t > \tau\}) && \text{(because } \{\theta_t \in \Theta_{t-1}\} \supseteq \mathcal{G}_t \cap \{t > \tau\} \text{ by Claim 12)} \\
&= 1\{t > \tau\}\mathbb{P}_{t-1}(\{\theta_t \in \Theta^{\mathrm{opt}}\} \cap \mathcal{G}_t) && \text{(because } \tau \text{ is } \mathbb{F}\text{-adapted)} \\
&= \mathbb{P}_{t-1}(\{\theta_t \in \Theta^{\mathrm{opt}}\} \cap \mathcal{G}_t) && \text{(because we are on } \mathcal{E} \cap \{t > \tau\}) \\
&\geq p = 1/200\,. && \text{(by Lemma 7)}
\end{aligned}$$

Since we are on $\mathcal{G}_t \cap \{t > \tau\}$, by Claim 12, $\theta_t \in \Theta_{t-1}$, and so

$$\sqrt{\dot\mu(J(\theta_t))}(J(\theta_\star) - J(\theta_t)) \leq \sqrt{\dot\mu(X_t^\top \theta_-)}(J(\theta_\star) - J(\theta_-))\,,$$

where we let θ_- be a maximiser of $\sqrt{\dot\mu(X_t^\top \theta)}(J(\theta_\star) - J(\theta))$ over $\theta \in \Theta_{t-1}$. Now, observe that

$$\begin{aligned}
\sqrt{\dot\mu(X_t^\top \theta_-)}(J(\theta_\star) - J(\theta_-)) &\leq \int \sqrt{\dot\mu(X_t^\top \theta_-)}\,(J(\theta_+) - J(\theta_-))Q(d\theta_+) \\
&\quad \text{(on } \mathcal{E} \cap \{t > \tau\}\text{, by Equation (22), } \operatorname{supp} Q \subset \Theta^{\mathrm{opt}} \text{ and for } \theta_+ \in \Theta^{\mathrm{opt}},\ J(\theta_+) \geq J(\theta_\star),\ Q \text{ is a prob. m.)} \\
&= \frac{1}{p_{t-1}} \mathbb{E}_{t-1}[\sqrt{\dot\mu(X_t^\top \theta_-)}\,(J(\theta_t) - J(\theta_-))1\{\theta_t \in \Theta^{\mathrm{opt}} \cap \Theta_{t-1}\}] && \text{(definition of } Q) \\
&\leq \frac{1}{p} \mathbb{E}_{t-1}[\sqrt{\dot\mu(X_t^\top \theta_-)}\,|J(\theta_t) - J(\theta_-)|1\{\theta_t \in \Theta_{t-1}\}]\,. && (p_{t-1} \geq p\text{, introducing } |\cdot| \text{ and dropping } \Theta^{\mathrm{opt}})
\end{aligned}$$

Now, observe that using that $\theta_-, \hat\theta_{t-1} \in \Theta_{t-1}$, the proof of Claim 12 gives us that $\sqrt{\dot\mu(X_t^\top \theta_-)} \le \sqrt{e\dot\mu(X_t^\top \hat\theta_{t-1})}$. Also,

$$
\begin{aligned}
|J(\theta_t) - J(\theta_-)|\, 1\{\theta_t \in \Theta_{t-1}\} &\le |X_t^\top(\theta_t - \theta_-)|\, 1\{\theta_t \in \Theta_{t-1}\} && \text{(Claim 15)}\\
&\le \|X_t\|_{\widehat H_{t-1}^{-1}} \|\theta_t - \theta_-\|_{\widehat H_{t-1}} 1\{\theta_t \in \Theta_{t-1}\} && \text{(Cauchy-Schwarz)}\\
&\le \frac{21}{10} C_d(\delta')\gamma_n \|X_t\|_{\widehat H_{t-1}^{-1}}\,. && \text{(definition of } \Theta_{t-1} \text{ and Claim 12)}
\end{aligned}
$$

When combined with the inequalities developed thus far, this yields the statement of the claim. □

I.2 Bounding r_t^{PRED} (proof of Lemma 9)

Proof of Lemma 9. Observe that

$$
r_t^{\mathrm{PRED}} = \mu(X_t^\top \theta_t) - \mu(X_t^\top \theta_\star) \le |\mu(X_t^\top \theta_t) - \mu(X_t^\top \hat\theta_{t-1})| + |\mu(X_t^\top \hat\theta_{t-1}) - \mu(X_t^\top \theta_\star)|.
$$

Now, by Taylor's theorem and self-concordance,

$$
\begin{aligned}
|\mu(X_t^\top \hat\theta_{t-1}) - \mu(X_t^\top \theta_\star)| &\le (\dot\mu(X_t^\top \hat\theta_{t-1}) + |\ddot\mu(\xi)| D(\hat\theta_{t-1} - \theta_\star)) |X_t^\top(\hat\theta_{t-1} - \theta_\star)|\\
&\le (\dot\mu(X_t^\top \hat\theta_{t-1}) + \dot\mu(\xi) M D(\hat\theta_{t-1} - \theta_\star)) |X_t^\top(\hat\theta_{t-1} - \theta_\star)| \qquad (23)
\end{aligned}
$$

for some ξ between $X_t^\top \hat\theta_{t-1}$ and $X_t^\top \theta_\star$. Recall that on $\mathcal{E} \cap \{t > \tau\}$ by Claim 9,

$$
MD(\hat\theta_{t-1} - \theta_\star) \le 1\,. \qquad (24)
$$

Hence,

$$
\begin{aligned}
&\dot\mu(\xi) M D(\hat\theta_{t-1} - \theta_\star) \le \dot\mu(\xi)\\
&\quad\le \exp(M|\xi - X_t^\top \hat\theta_{t-1}|)\dot\mu(X_t^\top \hat\theta_{t-1}) && \text{(Lemma 3)}\\
&\quad\le \exp(MD(\hat\theta_{t-1} - \theta_\star))\dot\mu(X_t^\top \hat\theta_{t-1}) && (\xi \text{ is between } X_t^\top\theta_\star \text{ and } X_t^\top \hat\theta_{t-1}, \text{ definition of } D(\cdot))\\
&\quad\le e\dot\mu(X_t^\top \hat\theta_{t-1})\,. && \text{(by Equation (24))}
\end{aligned}
$$

which, together with Equation (23) gives

$$
\begin{aligned}
|\mu(X_t^\top \hat\theta_{t-1}) - \mu(X_t^\top \theta_\star)| &\le (e+1)\dot\mu(X_t^\top \hat\theta_{t-1}) |X_t^\top(\hat\theta_{t-1} - \theta_\star)|\\
&\le (e+1)\dot\mu(X_t^\top \hat\theta_{t-1}) \|\hat\theta_{t-1} - \theta_\star\|_{\widehat H_{t-1}} \|X_t\|_{\widehat H_{t-1}} && \text{(Cauchy-Schwarz)}\\
&\le \sqrt{\frac{11}{10}}(e+1)\dot\mu(X_t^\top \hat\theta_{t-1}) \|\hat\theta_{t-1} - \theta_\star\|_{H_{t-1}} \|X_t\|_{\widehat H_{t-1}} && \left(\text{by Claim 9, } \widehat H_{t-1} \preceq \sqrt{\tfrac{11}{10}} H_{t-1}\right)\\
&\le 5\gamma_n \dot\mu(X_t^\top \hat\theta_{t-1}) \|X_t\|_{\widehat H_{t-1}^{-1}}\,, && \text{(other part of Claim 9)}
\end{aligned}
$$

where in the last inequality we also bounded $\sqrt{\frac{11}{10}} \times \frac{21}{20} \times (e+1) \le 5$.

By the same sequence of arguments, this time using Claim 12 in place of Claim 9,

$$
|\mu(X_t^\top \theta_t) - \mu(X_t^\top \hat\theta_{t-1})| \le 5C_d(\delta')\gamma_n \dot\mu(X_t^\top \hat\theta_{t-1}) \|X_t\|_{\widehat H_{t-1}^{-1}}.
$$

Summing the two bounds per the initial decomposition of r_t^{PRED} of this proof gives the stated bound. □

I.3 Bounding sum of derivatives (proof of Claim 14)

Proof. By a first order Taylor expansion of $\dot{\mu}$,

$$\begin{aligned}
\sum_{t=\tau+1}^{n} \dot{\mu}(X_t^\top \theta_\star) &= \sum_{t=\tau+1}^{n} \dot{\mu}(x_\star^\top \theta_\star) + \sum_{t=\tau+1}^{n} (X_t - x_\star)^\top \theta_\star \int_{v=0}^{1} \ddot{\mu}\left(x_\star^\top \theta_\star + v(X_t - x_\star)^\top \theta_\star\right) dv \\
&\leq n\dot{\mu}(x_\star^\top \theta_\star) + \sum_{t=\tau+1}^{n} \left| (X_t - x_\star)^\top \theta_\star \int_{v=0}^{1} \ddot{\mu}\left(x_\star^\top \theta_\star + v(X_t - x_\star)^\top \theta_\star\right) dv \right| \\
&\leq n\dot{\mu}(x_\star^\top \theta_\star) + \sum_{t=\tau+1}^{n} (x_\star - X_t)^\top \theta_\star \int_{v=0}^{1} \left|\ddot{\mu}\left(x_\star^\top \theta_\star + v(X_t - x_\star)^\top \theta_\star\right)\right| dv && (X_t^\top \theta_\star \leq x_\star^\top \theta_\star) \\
&\leq n\dot{\mu}(x_\star^\top \theta_\star) + M \sum_{t=\tau+1}^{n} (x_\star - X_t)^\top \theta_\star \int_{v=0}^{1} \dot{\mu}\left(x_\star^\top \theta_\star + v(X_t - x_\star)^\top \theta_\star\right) dv && (|\ddot{\mu}| \leq M\dot{\mu}) \\
&= n\dot{\mu}(x_\star^\top \theta_\star) + M \sum_{t=\tau+1}^{n} \left(\mu(x_\star^\top \theta_\star) - \mu(X_t^\top \theta_\star)\right) && \text{(fundemental theorem of calculus)} \\
&= n\dot{\mu}(x_\star^\top \theta_\star) + M\widetilde{R}(n).
\end{aligned}$$

□