VITS : Variational Inference Thompson Sampling for contextual bandits

Pierre Clavier [*1 2 3] **Tom Huix** [*1] **Alain Durmus** [1]

Abstract

In this paper, we introduce and analyze a variant of the Thompson sampling (TS) algorithm for contextual bandits. At each round, traditional TS requires samples from the current posterior distribution, which is usually intractable. To circumvent this issue, approximate inference techniques can be used and provide samples with distribution close to the posteriors. However, current approximate techniques yield to either poor estimation (Laplace approximation) or can be computationally expensive (MCMC methods, Ensemble sampling...). In this paper, we propose a new algorithm, Varational Inference TS (VITS), based on Gaussian Variational Inference. This scheme provides powerful posterior approximations which are easy to sample from, and is computationally efficient, making it an ideal choice for TS. In addition, we show that VITS achieves a sub-linear regret bound of the same order in the dimension and number of round as traditional TS for linear contextual bandit. Finally, we demonstrate experimentally the effectiveness of VITS on both synthetic and real world datasets.

1. Introduction

In traditional Multi-Armed Bandit (MAB) problems, an agent, has to sequentially choose between several actions (referred to as "arms"), from which he receives a reward from the environment. The arm selection process is induced by a sequence of policies, which is inferred and refined at each round from past observations. These policies are designed to optimize the cumulative rewards over the entire process. The main challenge in this task is to effectively manage a suitable exploitation and exploration trade-off (Robbins, 1952; Katehakis & Veinott, 1987; Berry & Fristedt, 1985; Auer et al., 2002; Lattimore & Szepesvári, 2020; Kveton et al., 2020). Here, exploitation refers to selecting an arm that is currently believed to be the best based on past observations, while exploration refers to selecting arms that have not been selected frequently in the past in order to gather more information.

Contextual bandit problems is a particular instance of MAB problem, which supposes, at each round, that the set of arms and the corresponding reward depend on a d-dimensional feature vector called a contextual vector or context. This scenario has been extensively studied over the past decades and learning algorithms have been developed to address this problem (Langford & Zhang, 2007; Abbasi-Yadkori et al., 2011a; Agrawal & Goyal, 2013; Kveton et al., 2020), and they have been successfully applied in several real-world problem such as recommender systems, mobile health and finance (Li et al., 2010; Agarwal et al., 2016; Tewari & Murphy, 2017; Bouneffouf et al., 2020). The existing algorithms for addressing contextual bandit problems can be broadly categorized into two groups. The first category is based on maximum likelihood and the principle of optimism in the face of uncertainty (OFU) and has been studied in (Auer et al., 2002; Chu et al., 2011; Abbasi-Yadkori et al., 2011b; Li et al., 2017; Ménard & Garivier, 2017; Zhou et al., 2020; Foster & Rakhlin, 2020; Zenati et al., 2022).

The second category consists in randomized probability matching algorithms, which is based on Bayesian belief and posterior sampling. Thompson Sampling (TS) is one of the most famous algorithms that fall into this latter category. Since its introduction by Thompson (1933), it has been widely studied, both theoretically and empirically (Agrawal & Goyal, 2012; Kaufmann et al., 2012; Agrawal & Goyal, 2013; Russo & Van Roy, 2014; 2016; Lu & Van Roy, 2017; Riquelme et al., 2018; Jin et al., 2021). Despite the fact that OFU algorithms offer better theoretical guarantees compared to classic TS-based algorithms, traditional TS methodologies still appeal to us due to their straightforward implementation and empirical advantages. In (Agrawal & Goyal, 2012), the authors claimed that: "In applications like display advertising and news article recommendation, TS is competitive with or better than popular methods such as

[*]Equal contribution [1]CMAP, CNRS, Ecole Polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France. [2]Inria Paris, 75015 Paris, France [3]Centre de Recherche des Cordeliers, INSERM, Universite de Paris, Sorbonne Universite, 75006 Paris, France . Correspondence to: Tom Huix <tom.huix@polytechnique.edu>, Pierre Clavier <pierre.clavier@polytechnique.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

UCB". Similarly, (Chapelle & Li, 2011) has examined the empirical performances of TS on both simulated and real data. Their experiments demonstrate that TS outperforms OFU methods, leading them to conclude: "In any case, TS is very easy to implement and should thus be considered as a standard baseline". Taking all these factors into account, we have decided to focus on TS-based algorithms for addressing contextual bandit problems.

Despite its relative simplicity, effectiveness and convergence guarantees, TS comes with a computational burden which is to sample, at each iteration $t \in \mathbb{N}^\star$, from an appropriate Bayesian posterior distribution $\hat{\mathrm{p}}_t$ defined from the previous observations. Indeed, these posteriors are usually intractable and approximate inference methods have to be used to obtain samples with distributions "close" to the posterior. The family of TS methods using approximate inference methods will be referred to as approximate inference TS in the sequel. Among the simplest approximate inference methods, Laplace approximation has been proposed for TS in Chapelle & Li (2011). This method consists of approximating the posterior distribution $\hat{\mathrm{p}}_t$ by a Gaussian distribution with a carefully chosen mean and covariance matrix. More precisely, the mean is a mode of the target distribution which is typically found using an optimization algorithm, while the covariance matrix is taken to be the negative Hessian matrix of the log posterior at the considered mode. Despite this method is easy to implement, it may lead to poor posterior representations. Indeed, while Laplace method achieves minimal optimality in terms of regret (Faury et al., 2022), it doesn't dictate the posterior convergence rate. More precisely, in (Katsevich & Rigollet, 2023) it has been demonstrated that VI outperforms Laplace in terms of mean convergence by a factor of $1/n$. It is worth noting that the covariance rates remain the same for both methods. This discrepancy can lead to inadequate approximations, especially in high-dimensional settings, as highlighted in section I.4 of (Katsevich & Rigollet, 2023).

Another class of popular approximate inference methods are Markov Chain Monte Carlo (MCMC) methods, such as Metropolis or Langevin Monte Carlo (LMC) algorithms. In the bandit literature, LMC has been proposed to get approximate samples from TS posteriors for solving traditional bandit problem in (Mazumdar et al., 2020) and for contextual bandit problems in Xu et al. (2022); Huix et al. (2023). Also, Lu & Van Roy (2017) have proposed to adapt Ensemble Methods to the bandit setting. Roughly, the idea here is to maintain and incrementally update an ensemble of statistically plausible models and to draw a uniform sample from this family at each iteration.

Finally, Variational Inference (VI) (Blei et al., 2017) is another class of approximate method that could be used to get samples from the posterior distribution. The core concept behind VI is to find a distribution $\tilde{\mathrm{q}}$, referred to as the variational posterior, to closely match the true posterior $\hat{\mathrm{p}}$ in terms of Kullback-Leibler divergence (KL) within a predefined family of distributions known as the variational family $\mathcal{G}$. In general, the variational family is chosen to make the optimization of the KL tractable and to be easy to sample from. In their work Urteaga & Wiggins (2018) propose the mean-field mixture of Gaussian variational family for TS. This family of distributions is quite extensive and provides an accurate approximation for a wide range of posterior distributions. However, in our perspective, it might not be the most suitable choice for TS. Firstly, the optimization algorithm at each time step can be computationally expensive. Secondly, the mean-field assumption assumes that the parameters are independent, a premise that holds true in the regime of large, overparameterized models. In our perspective, this regime may not align with the Bandit problem, which often operates in a setting where the number of data points tends towards infinity in comparison to the model size. Finally, Yu et al. (2020) also employs VI in more general graphical models but focuses on structured arms and rewards, where the rewards are correlated through latent variables.

In this paper, we develop an efficient VI method that makes use of the whole family of non-degenerate Gaussian distributions. This choice of VI family is supported by the Bernstein-Von Mises theorem (Van der Vaart, 2000) . This theorem, subject to specific regularity conditions, asserts that a properly scaled version of the posterior converges to a Gaussian as the sample size grows. When applied to contextual bandits, the data points progressively accumulate over time, leading to the gradual concentration of the posterior around a dominant mode. As a consequence, the Gaussian approximation becomes increasingly suitable for representing the posterior in this particular setting. Furthermore, the covariance of the rescaled posterior distribution tends to converge towards the inverse Fisher information matrix, which may not necessarily be diagonal, thus justifying the need for a non-mean-field hypothesis. Our main contributions can be summarized as follows:

Our **first contribution** is methodological. We develop a novel variant of the TS algorithm, referred to as Variational Inference TS (`VITS`). Our method addresses the main challenges encountered by the existing approximate TS algorithms and can be applied to a very large class of TS posteriors. Moreover, it enjoys a low computational cost both theoretically and empirically, since it boils down to adding a few optimization steps per round. We also propose two approximate versions of `VITS`, called `VITS − II` and `VITS − II Hessian-free`, that scale with the problem dimension.

Our **second contribution** is theoretical. We establish that

our proposed methodology achieves a sub-linear regret of order $\tilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ (up to logarithmic term) in the linear contextual bandit framework, where T is the number of rounds and d is the dimension of the policy parameter. To the best of our knowledge, this is the first regret bound derived for VI in the context of sequential learning.

Finally, our **last contribution** is to illustrate the empirical performances of our method on a synthetic and on the real world dataset MovieLens (Lam & Herlocker). It has been shown that in many cases, VITS outperforms existing approximate TS algorithms such as LMC algorithm.

Related work. The theoretical foundations of TS for linear contextual bandits were initially explored by Agrawal & Goyal (2013). In this paper, the authors establish a sub-linear cumulative regret bound $\tilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ for Linear TS (Lin-TS). Compared to this study, our method achieves a similar regret bound in the linear framework. However, it should be noted that Lin-TS is a specialized algorithm that can be only used when the posterior is known and can be efficiently sampled from.

As mentioned previously, VI has been suggested for TS in Urteaga & Wiggins (2018). This paper introduces a TS algorithm called VTS that utilizes a mixture of mean-field Gaussian distributions to approximate the sequence of posteriors. In comparison to this work, the setting and the variational family we consider are richer than (Urteaga & Wiggins, 2018). A more detailed comparison is postponed in Appendix D. Moreover, the methodology developed in (Urteaga & Wiggins, 2018) does not come with any convergence guarantees. An empirical and theoretical study of using LMC as approximate inference method for TS for contextual bandit problems was carried out in Xu et al. (2022). This paper establishes that the resulting algorithm, called LMC-TS, achieves a state-of-the-art sub-linear cumulative regret for linear contextual bandits. Compared to this method, our approach yields a similar sub-linear regret in the same setting. Finally, (Zhang et al., 2020) suggests a TS method based on Neural Tangent Kernel. While this performs well on real datasets, their method is much more expensive than previously mentioned approaches, as it requires training a neural network.

Notation. For $n \geq 1$, $[n]$ represents the set of integers between 1 and n. $\mathrm{N}(\mu, \Sigma)$ denotes the d-multidimensional Gaussian probability distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. The transpose of a matrix M is denoted by $M^\top$. For any symmetric-real matrix A, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ represent the maximum and minimum eigenvalues of A respectively. The norm $\|\cdot\|_2$ will refer to the 2-norm for vectors, and the operator norm for matrices. For any semi-definite positive matrix A, the norm $\|x\|_A$ denotes the Mahalanobis norm, i.e., $\|x\|_A = \sqrt{xAx^\top}$. For any event E on a probability space, $\bar{\mathsf{E}}$ refers to the complementary of E. Finally, $\mathbb{1}$ is the indicator function and tr is the trace of a matrix.

2. Thompson sampling for contextual bandits

Contextual bandit: We now present in more details the contextual bandit framework. Let X be a contextual space and consider $\mathcal{A} : \mathsf{X} \to 2^{\mathsf{A}}$ a set-valued action map, where 2^{A} stands for the power set of the action space A. For simplicity, we assume here that $\sup_{x\in\mathsf{X}} \mathrm{Card}(\mathcal{A}(x)) < +\infty$. A (deterministic or random) function $\pi : \mathsf{X} \to \mathsf{A}$ is said to be a policy if for any $x \in \mathsf{X}$, $\pi(x) \in \mathcal{A}(x)$. Then, for a fixed horizon $T \in \mathbb{N}^\star$, a contextual bandit process can be defined as follows: at each iteration $t \in [T]$ and given the past observations $\mathrm{D}_{t-1} = \{(x_s, a_s, r_s)\}_{s<t}$:

- The agent receives a contextual feature $x_t \in \mathsf{X}$;
- The agent chooses an action $a_t = \pi_t(x_t)$ where π_t is a policy sampled from $\mathbb{Q}_t(\cdot|\mathrm{D}_{t-1})$;
- Finally, the agent receives a reward r_t sampled from $\mathcal{R}(\cdot|x_t, a_t)$ given D_{t-1}. Here, $\mathcal{R}$ is a Markov kernel on $(\mathsf{A} \times \mathsf{X}) \times \mathsf{R}$, where $\mathsf{R} \subset \mathbb{R}$

For a fixed family of conditional distributions $\mathbb{Q}_{1:T} = \{\mathbb{Q}_t\}_{t\leq T}$, this process defines a random sequence of policies, $\pi_{1:T} = \{\pi_t\}_{t\leq T}$ with distribution still denoted by $\mathbb{Q}_{1:T}$ by abuse of notation. Let's defined the optimal expected reward for a contextual vector $x \in \mathsf{X}$ and the expected reward given x and any action $a \in \mathcal{A}(x)$ as follow

$$f_\star(x) = \max_{a\in\mathcal{A}(x)} f(x,a)\ , f(x,a) = \int r\mathcal{R}(\mathrm{d}r|x,a)\ . \quad (1)$$

The main challenge of a contextual bandit problem is to find the distribution $\mathbb{Q}_{1:T}$ that minimizes the cumulative regret defined as

$$\begin{aligned} &\mathrm{CRegret}(\mathbb{Q}_{1:T}) = \textstyle\sum_{s\leq T} \mathrm{Regret}_s^{\pi_s} \\ &\text{with} \quad \mathrm{Regret}_s^{\pi_s} = f_\star(x_s) - f(x_s, \pi_s(x_s)). \end{aligned} \quad (2)$$

The main difficulty in the contextual bandit problem, comes from the fact that the reward distribution $\mathcal{R}$ is intractable and must be inferred to find the best policy to minimize the instantaneous regret $\pi \mapsto f_\star(x) - f(x, \pi(x))$ for a context $x \in \mathsf{X}$. However, the estimation of $\mathcal{R}$ may be in contradiction with the primary objective to minimize the cumulative regret (2), since potential non-effective arms has to be chosen to obtain a complete description of $\mathcal{R}$. Therefore, bandit learning algorithms have to achieve an appropriate trade-off between exploitation of arms which have been confidently learned and exploration of misestimated arms.

Thompson sampling: To achieve such a trade-off, we consider the popular Thompson Sampling (TS) algorithm. Consider a parametric model $\{\mathcal{R}_\theta : \theta \in \mathbb{R}^d\}$ for the reward distribution, where for any θ, $\mathcal{R}_\theta$ is a Markov kernel on $(\mathsf{A} \times \mathsf{X}) \times \mathsf{R}$ parameterized by $\theta \in \mathbb{R}^d$. We assume in this paper that $\mathcal{R}_\theta$ admits a density with respect to some dominating measure λ_{ref}. An important example are generalized linear bandits (Filippi et al., 2010; Kveton et al., 2020). In particular, it assumes that $\{\mathcal{R}_\theta(\cdot|x,a) : \theta \in \Theta\}$ is an exponential family with respect to λ_{ref}, i.e., for $x \in \mathsf{X}$ and $a \in \mathsf{A}$,

$$\frac{\mathrm{d}\mathcal{R}_\theta}{\mathrm{d}\lambda_{\mathrm{ref}}}(r|x,a) = \mathrm{h}(r)\exp(g(\theta,x,a)\mathrm{T}(r) - \mathrm{C}(\theta,x,a)), \tag{3}$$

for $h : \mathsf{R} \to \mathbb{R}_+$, natural parameter and log-partition function $g, \mathrm{C} : \mathbb{R}^d \times \mathsf{X} \times \mathsf{A} \to \mathbb{R}$ and sufficient statistics $\mathrm{T} : \mathsf{R} \to \mathbb{R}$. The family is said to be in canonical form if $g(\theta,x,a) = \langle \phi(x,a), \theta \rangle$ for some feature map $\phi : \mathsf{X} \times \mathsf{A} \to \mathbb{R}$ and $\mathrm{C}(\theta,x,a) = \sigma(\langle \phi(x,a), \theta \rangle)$ for some link function σ. Linear contextual bandits (Chu et al., 2011; Abbasi-Yadkori et al., 2011b) fall into this model taking $\lambda_{\mathrm{ref}} = \mathrm{Leb}$, T equals to the identity function,

$$\mathrm{h}(r) = \exp(-\eta r^2/2) \text{ and } g(\theta,x,a) = \eta \langle \phi(x,a), \theta \rangle, \tag{4}$$

for some $\eta > 0$. As a result, $\mathcal{R}_\theta(\cdot|x,a)$ is simply the Gaussian distribution with mean $\langle \phi(x,a), \theta \rangle$ and variance $1/\eta$. Finally (Riquelme et al., 2018; Zhou et al., 2020; Xu et al., 2020) introduced an extension of linear contextual bandits, referred to as linear neural contextual bandits where g is a neural network with weights θ and taking as input a pair (x,a). With the introduced notations, the likelihood function associated to the observations D_t at step $t > 1$ is given by

$$\mathrm{L}_t(\theta) \propto \exp\left\{\sum_{s=1}^{t-1} \ell(\theta|x_s,a_s,r_s)\right\}, \tag{5}$$

where the log-likelihood is given by $\ell(\theta|x_s,a_s,r_s) = \log(\mathrm{d}\mathcal{R}_\theta/\mathrm{d}\lambda_{\mathrm{ref}})(r_s|x_s,a_s)$. Choosing a prior on θ with density p_0 with respect to Leb, and applying Bayes formula, the posterior distribution at round $t \in [T]$ is given by

$$\hat{\mathrm{p}}_t = \mathrm{L}_t(\theta)\mathrm{p}_0(\theta)/\mathfrak{Z}_t \tag{6}$$

where $\mathfrak{Z}_t = \int \mathrm{L}_t(\theta)\mathrm{p}_0(\theta)\mathrm{d}\theta$ denotes the normalizing constant and we used the convention that $\hat{\mathrm{p}}_1 = \mathrm{p}_0$. Moreover we define the potential function $U(\theta) \propto -\log \hat{\mathrm{p}}_t(\theta)$. Then, at each iteration $t \in [T]$, TS consists in sampling a sample θ_t from the posterior $\hat{\mathrm{p}}_t$ and from it, use as a policy, $\pi_t^{(\mathrm{TS})}(x)$ defined for any x by

$$\pi_t^{(\mathrm{TS})}(x) = a^{\theta_t}(x)\ , a^\theta(x) = \arg\max_a \int r\mathcal{R}_\theta(\mathrm{d}r|x,a) \tag{7}$$

Since $\mathfrak{Z}_t$ is generally intractable, sampling from the posterior distribution is not in general an option.

Variational inference TS: To address this challenge, practitioners often employ approximate inference methods to generate samples from a distribution that is expected to be "close" to the actual posterior distribution. In this context, we specifically concentrate on the application of VI. In this scenario, we consider a variational family $\mathcal{G}$ which is a set of probability densities with respect to the Lebesgue measure, from which it is typically easy to sample from. Then ideally, at each round $t \in [T]$, the posterior distribution $\hat{\mathrm{p}}_t$ is approximated by the variational posterior distribution $\tilde{\mathrm{q}}_t$ which is defined as:

$$\tilde{\mathrm{q}}_t = \arg\min_{\mathrm{p} \in \mathcal{G}} \mathrm{KL}(\mathrm{p}|\hat{\mathrm{p}}_t)\ , \tag{8}$$

where KL is the Kullback-Leibler divergence. However, we have to determine at each round a solution to the problem specified in (8). In this paper, we consider as variational family the set of non-degenerate Gaussian distribution $\mathcal{G} = \{\mathrm{N}(\mu,\Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathcal{S}_+^*\}$ where $\mathrm{N}(\mu,\Sigma)$ is the Gaussian distribution with mean μ and covariance matrix Σ and $\mathcal{S}_+^*$ is the set of symmetric positive definite matrices. As explained in the introduction, this Gaussian variational family is particularly relevant in bandit framework according to Bernstein-Von Mises theorem.

Presentation of `VITS - I`: As we will see, this choice of variational family will allow to derive an efficient method for solving (8) using the Riemannian structure of $\mathcal{G}$. As noted in (Lambert et al., 2022b), $\mathcal{G}$ equipped with the Wasserstein distance of order 2 is a complete metric space as a closed subset of $\mathcal{P}_2(\mathbb{R}^d)$, the set of probability distributions with finite second moment. Recall that for two Gaussian distributions $p_0 = \mathrm{N}(\mu_0,\Sigma_0)$ and $p_1 = \mathrm{N}(\mu_1,\Sigma_1)$, their Wasserstein distance has a closed form:

$$W_2^2(p_0,p_1) = \|\mu_0 - \mu_1\|^2 + \mathrm{tr}(\Sigma_0 + \Sigma_1 - 2(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2})\ .$$

This Wasserstein distance on $\mathcal{G}$ allows to derive a Riemannian metric denoted $\mathfrak{g}$. The corresponding geodesic is given through the exponential map. More precisely, for a Gaussian distribution $p = \mathrm{N}(\mu_p,\Sigma_p)$, this map is defined as

$$\begin{aligned}\exp_p(\mu_v,\Sigma_v) =& (\mu_p + \mu_v + (\Sigma_v + \mathrm{I})(\cdot - \mu_p))_\# p \\ =& \mathrm{N}(\mu_p + \mu_v,\ (\Sigma_v + \mathrm{I})\ \Sigma_p\ (\Sigma_v + \mathrm{I}))\ .\end{aligned} \tag{9}$$

With all these preliminaries, we can now present and motivate the algorithm developed in (Lambert et al., 2022b) to efficiently solve (8). This method can be formalized as a Riemannian gradient descent scheme on $\mathcal{G}$. Firstly, we define the loss function $\mathcal{F}_t : p \to \mathrm{KL}(p|\hat{\mathrm{p}}_t)$. Then, following (Lambert et al., 2022b), we derive the gradient operator of $\mathcal{F}_t$ on $\mathcal{G}$ equipped with $\mathfrak{g}$ as

$$\nabla_{\mathfrak{g}}\mathcal{F}_t(p) = (\int \nabla U_t(\theta)\mathrm{d}p(\theta),\ \int \nabla^2 U_t(\theta)\mathrm{d}p(\theta) - \Sigma_p^{-1}) \tag{10}$$

where Σ_p is the covariance matrix of p. From this expression, the corresponding Riemannian gradient descent (Bonnabel, 2013) using a step size $h_t > 0$ defines the sequence of iterates $\{\mathrm{q}_{t,k}\}_{k=1}^{K_t}$ recursively as:

$$\mathrm{q}_{t,k+1} = \exp_{\mathrm{q}_{t,k}}(-h_t \nabla_{\mathfrak{g}} \mathcal{F}_t(\mathrm{q}_{t,k})) \ .$$

At each time step t, this sequence is initialized with variational posterior at the previous step, ie, $\mathrm{q}_{t,0} = \mathrm{q}_{t-1,K_{t-1}}$. Please note that this warm initialization of the posterior results in an efficient algorithm and has been directly used in our main theoretical result (see (31)). Combining (9) and (10), this recursion amounts defining a sequence of means $\{\mu_{t,k}\}_{k=1}^{K_t}$ and covariance matrices $\{\Sigma_{t,k}\}_{k=1}^{K_t}$ by the recursions

$$\begin{aligned}
&\mu_{t,k+1} = \mu_{t,k} - h_t \int \nabla U_t(\theta) \mathrm{d}\mathrm{q}_{t,k}(\theta), \\
&\Sigma_{t,k+1} = A_{t,k} \Sigma_{t,k} A_{t,k}, \\
&\mathrm{q}_{t,k+1} = \mathrm{N}(\mu_{t,k+1}, \Sigma_{t,k+1}) \\
&\text{where} \quad A_{t,k} = \mathrm{I} - h_t(\int \nabla^2 U_t(\theta) \mathrm{d}\mathrm{q}_{t,k}(\theta) - \Sigma_{t,k}^{-1})
\end{aligned}$$

The main computational challenge in this recursion stems is that the integrals involved are typically intractable. To overcome this issue, we employ a Monte Carlo procedure to approximate these integrals. Subsequently, we consider a sequence of mean values denoted as $\{\tilde{\mu}_{t,k}\}_{k=1}^{K_t}$ and covariance matrices $\{\tilde{\Sigma}_{t,k}\}_{k=1}^{K_1}$ such that:

$$\begin{aligned}
&\tilde{\mu}_{t,k+1} = \tilde{\mu}_{t,k} - h_t \nabla U_t(\tilde{\theta}_{t,k}), \quad \tilde{\Sigma}_{t,k+1} = \tilde{A}_{t,k} \tilde{\Sigma}_{t,k} \ \tilde{A}_{t,k} \\
&\text{with} \quad \tilde{A}_{t,k} = \mathrm{I} - h_t(\nabla^2 U_t(\tilde{\theta}_{t,k}) - \tilde{\Sigma}_{t,k}^{-1}) \ ,
\end{aligned}$$

where $\tilde{\theta}_{t,k} \sim \mathrm{N}(\tilde{\mu}_{t,k}, \tilde{\Sigma}_{t,k})$. Consequently, following (Lambert et al., 2022b) we obtain an algorithm capable of addressing the problem defined in (8). However, this algorithm exhibits computational inefficiency, particularly in high-dimensional scenarios. This inefficiency arises from the necessity to sample from a Gaussian distribution with a non-diagonal covariance matrix during each updating step $k \in [K_t]$. As a result, it becomes impractical for use in a contextual bandit problem, where, at each time step t, we must solve the problem described in (8). This paper introduces an improved version of the earlier algorithm, designed to efficiently address the problem presented in (8). To achieve this, we begin by examining a sequence of matrices denoted as $B_{t,k}$, defined by the following

$$B_{t,k+1} = \{\mathrm{I} - h_t \nabla^2 U_t(\tilde{\theta}_{t,k})\} B_{t,k} + h_t (B_{t,k}^{-1})^\top \ . \quad (11)$$

It is important to note that $B_{t,k}$ is a square-root matrix of the covariance of the variational distribution $\tilde{\Sigma}_{t,k}$, ie, $B_{t,k} B_{t,k}^\top = \tilde{\Sigma}_{t,k}$. Then we can sample efficiently from the variational distribution using $B_{t,k}$ with $\tilde{\theta}_{t,k} = \tilde{\mu}_{t,k} + B_{t,k} \epsilon_{t,k}$, $\epsilon \sim \mathrm{N}(0, \mathrm{I})$. As a result, note that our method does not require any Cholesky decomposition, which has a complexity of $\mathcal{O}(d^3)$, contrary to the algorithm derived in (Lambert et al., 2022a) and also in LinTS. The updating strategies for the sequence of $\tilde{\mu}_{t,k}$ and $B_{t,k}$ are given by

$$\begin{aligned}
&\tilde{\mu}_{t,k+1} = \tilde{\mu}_{t,k} - h_t \nabla U_t(\tilde{\theta}_{t,k}) \\
&B_{t,k+1} = \{\mathrm{I} - h_t \nabla^2 U_t(\tilde{\theta}_{t,k})\} B_{t,k} + h_t (B_{t,k}^{-1})^\top \\
&\tilde{\theta}_{t,k} \sim \mathrm{N}(\tilde{\mu}_{t,k}, B_{t,k}^\top B_{t,k}) \ .
\end{aligned}$$

From this methodology, we can now complete the description of our first algorithm, referred to as VITS-I. At each step t, we consider the variational distribution $\tilde{\mathrm{q}}_t = \tilde{\mathrm{q}}_{t,K_t} = \mathrm{N}(\tilde{\mu}_{t,K_t}, B_{t,k}^\top B_{t,k})$ which approximates the solution of (8). Then, at round $t+1$, VITS-I consists in sampling $\tilde{\theta}_{t+1}$ according to $\tilde{\mathrm{q}}_t$ and choosing

$$\pi_{t+1}^{\texttt{VITS-I}}(x) = \arg\max_{a \in \mathcal{A}(x)} a^{\tilde{\theta}_{t+1}}(x) \ . \quad (12)$$

As in TS, the likelihood function and the posterior distribution $\hat{\mathrm{p}}_{t+1}$ are updated following equations (5) and (6) using the new observed reward r_{t+1} distributed according to $\mathcal{R}(\cdot|x_{t+1}, a_{t+1})$ with $a_{t+1} = \pi_{t+1}^{\texttt{VITS-I}}(x)$. The round $t+1$ is then concluded by solving $\tilde{\mathrm{q}}_{t+1} = \tilde{\mathrm{q}}_{t+1,K_{t+1}}$. The pseudo-code associated with this algorithm is given in Algorithm 1 and Algorithm 2.

Algorithm 1 VITS algorithm

$B_{1,1} = \mathrm{I}/\sqrt{\lambda\eta}$, $\tilde{W}_{1,1} = \mathrm{I}/(\eta\lambda)$,$\tilde{\mu}_{1,1} \sim \mathrm{N}(0, \tilde{W}_{1,1})$
for $t = 1, \ldots, T$ **do**
  receive $x_t \in \mathsf{X}$
  sample $\tilde{\theta}_t$ from $\tilde{\mathrm{q}}_{t,K_t} = \mathrm{N}(\tilde{\mu}_{t,K_t}, B_{t,K_t}^\top B_{t,k})$
  choose $a_t = \pi^{(\mathrm{VITS})}(x_t)$ presented in (12)
  receive $r_t \sim \mathcal{R}(\cdot|x_t, a_t)$
  update $\tilde{\mathrm{q}}_{t+1,K_{t+1}}$ using Alg. 2 or 3.
end for

Algorithm 2 VITS-I

Parameters: step-size h_t, number of iterations K_t
$\tilde{\mu}_{t,1} \leftarrow \tilde{\mu}_{t-1,K_{t-1}}$,$B_{t,1} \leftarrow B_{t-1,K_{t-1}}$
for $k = 1, \ldots, K_t$ **do**
  draw $\tilde{\theta}_{t,k} \sim \tilde{\mathrm{q}}_{t,k} = \mathrm{N}(\tilde{\mu}_{t,k}, B_{t,k}^\top B_{t,k})$
  $\tilde{\mu}_{t,k+1} \leftarrow \tilde{\mu}_{t,k} - h_t \nabla U_t(\tilde{\theta}_{t,k})$
  $B_{t,k+1} \leftarrow \{\mathrm{I} - h_t \nabla^2 (U_t(\tilde{\theta}_{t,k}))\} B_{t,k} + h_t (B_{t,k}^{-1})^\top$
end for

Presentation of VITS-II: In high dimension, the computational cost of the recursion of mean values and covariance matrices may be prohibitive since at each iteration $k \in [K_t]$, it requires inverting the matrix $B_{t,k}$. To tackle this computational issue, we propose a new version of VITS. More

precisely, the inverse of the square root covariance matrix $B_{t,k}^{-1}$ can be approximated using a first order Taylor expansion in h_t; see Appendix C for more details. We denote by $C_{t,k}$ the approximation of $B_{t,k}^{-1}$, and we obtain recursions for the sequence of $\{C_{t,k}\}_{k\leq K_t}$ and $\{B_{t,k}\}_{k\leq K_t}$ such that:

$$C_{t,k+1} = C_{t,k}\{\mathrm{I} - h_t(C_{t,k}^\top C_{t,k} - \nabla^2 U_t(\tilde{\theta}_{t,k}))\} \ ,$$
$$B_{t,k+1} = (\mathrm{I} - h_t \nabla^2 U_t(\tilde{\theta}_{t,k}))B_{t,k} + h_t C_{t,k}^\top \ .$$

This trick reduces the complexity from $\mathcal{O}(d^3)$ to $\mathcal{O}(d^2)$ for the computation of the inverse. This version of VITS is referred to as VITS − II and is given in Algorithm 1 and 3.

Presentation of VITS − II Hessian-free: The most computationally intensive step in VITS − II remains the computation of the Hessian of U_t. In scenarios with a large number of data points and high dimensions, this step can become highly demanding. To avoid computing the Hessian of U_t, we suggest to use the following property of Gaussian distribution which is the result of a simple integration by part:

$$\int \nabla^2 U_t \mathrm{dN}(\mu, \Sigma) = \int \Sigma^{-1}(\mathrm{I} - \mu)\nabla U_t^\top \ \mathrm{dN}(\mu, \Sigma) \ . \quad (13)$$

After approximating this right side integral using Monte Carlo, we derive a new sequence of square-root covariance matrix $\{B_{t,k}\}_{k\leq K_t}$ and inverse square-root covariance matrix $\{C_{t,k}\}_{k\leq K_t}$, defined recursively by:

$$C_{t,k+1} = C_{t,k}\{\mathrm{I} - h_t(C_{t,k}^\top C_{t,k} - A_{t,k})\} \ ,$$
$$B_{t,k+1} = (\mathrm{I} - h_t A_{t,k})B_{t,k} + h_t C_{t,k}^\top \ ,$$

where $A_{t,k} = C_{t,k}^\top C_{t,k}(\tilde{\theta}_{t,k} - \tilde{\mu}_{t,k})\nabla U_t^\top(\tilde{\theta}_{t,k})$ and $\tilde{\theta}_{t,k} \sim \mathrm{N}(\tilde{\mu}_{t,k}, B_{t,k}^\top B_{t,k})$. This last version of VITS is referred to as VITS − II Hessian-free and its pseudo-code is given in Algorithm 1 and Algorithm 3.

Algorithm 3 VITS − II / VITS − II Hessian-free

Parameters: step-size h_t, number of iterations K_t
$\tilde{\mu}_{t,1} \leftarrow \tilde{\mu}_{t-1,K_{t-1}}$, $B_{t,1} \leftarrow B_{t-1,K_{t-1}}$
for $k = 1, \ldots, K_t$ **do**
  draw $\tilde{\theta}_{t,k} \sim \tilde{\mathrm{q}}_{t,k} = \mathrm{N}(\tilde{\mu}_{t,k}, B_{t,k}^\top B_{t,k})$
  $\tilde{\mu}_{t,k+1} \leftarrow \tilde{\mu}_{t,k} - h_t \nabla U_t(\tilde{\theta}_{t,k})$
  $A_{t,k} = \begin{cases} \nabla^2(U_t(\tilde{\theta}_{t,k})) & \text{(H)} \\ C_{t,k}^2(\tilde{\theta}_{t,k} - \tilde{\mu}_{t,k})(\nabla U_t(\tilde{\theta}_{t,k}))^\top & \text{(H free)} \end{cases}$
  $B_{t,k+1} \leftarrow \{\mathrm{I} - h_t A_{t,k}\}B_{t,k} + h_t C_{t,k}^\top$
  $C_{t,k+1} \leftarrow C_{t,k}(\mathrm{I} - h_t(C_{t,k}^\top C_{t,k} - A_{t,k}))$
end for

where (H) and (H free) are for respectively Hessian and Hessian Free version. The computational complexity of all methods has been experimentally studied in a simple case, as discussed in Section I.

3. Main results

3.1. Linear Bandit

In this section, we are interested in convergence guarantees for VITS − I and VITS − II applied to the linear contextual bandit framework. This framework consists in assuming that $\mathcal{R}_\theta$ has form (3) with $\lambda_{\mathrm{ref}} = \mathrm{Leb}$, T is the identity function and h and g are specified by (4):

$$\frac{\mathrm{d}R_\theta}{\mathrm{dLeb}}(r|x,a) \propto \exp[\eta(r - \langle\phi(x,a),\theta\rangle)^2/2] \ . \quad (14)$$

Assumption on the reward kernel $\mathcal{R}$ is the following:

Assumption 3.1. (Sub-Gaussian Reward Distribution) There exists $R > 1$ such that for any $x \in \mathsf{X}$, $a \in \mathcal{A}(x)$, $\rho > 0$, $\log \int \exp\{\rho(r - f(x,a))\}\mathcal{R}(\mathrm{d}r|x,a) \leq R\rho^2$, where f is defined in 1

We could only assume that $R > 0$ in Assumption 3.1 since if a distribution is R-sub-Gaussian, it is also R'-sub-Gaussian for any $R' \geq R$, however, we choose to set $R \geq 1$ to ease the presentation of our main results. We also assume that the model is well-specified.

Assumption 3.2. There exists $\theta^\star$ such that $\mathcal{R} = \mathcal{R}_{\theta^\star}$ and satisfying $\|\theta^\star\|_2 \leq 1$. Feature map ϕ satisfies the boundedness condition.

Assumption 3.3. For any contextual vector $x \in \mathbb{R}^d$ and action $a \in \mathcal{A}(x)$, it holds that $\|\phi(x,a)\|_2 \leq 1$.

Uniform boundedness condition on the feature map is relatively common for obtaining regret bounds for linear bandit problems (Agrawal & Goyal, 2013; Xu et al., 2022; Kveton et al., 2020; Abbasi-Yadkori et al., 2011b). Note that Assumption (3.3) is equivalent to $\sup_{x\in\mathsf{X},\, a\in\mathcal{A}(a)}\|\phi(x,a)\|_2 \leq M_\phi$ for some arbitrary but fixed constant $M_\phi > 0$, changing the feature map ϕ by ϕ/M_ϕ. Finally, we specify the prior distribution.

Assumption 3.4. The prior distribution is assumed to be zero-mean Gaussian distribution with variance $1/(\lambda\eta)$, where η also appears in the definition $\mathcal{R}_\theta$ in (14),

While our theoretical results can readily be extended to accommodate a non-zero mean Gaussian prior, for the sake of simplicity, we have chosen to center the prior. Under Assumption 3.4, combining (6) and (14), the negative log posterior $-\log \hat{\mathrm{p}}_t$ denoted by U_t is given by

$$U_t(\theta) = \frac{\eta}{2}\left(\sum_{s=1}^{t-1}(\phi(a_s,x_s)^\top\theta - r_s)^2 + \lambda\|\theta\|_2^2\right) \quad (15)$$

$$= \frac{\eta}{2}(\theta^\top V_t\theta - 2\theta^\top b_t + \sum_{s=1}^{t-1} r_s^2) \ , \quad (16)$$

$$V_t = \lambda \mathrm{I}_d + \sum_{s=1}^{t-1}\phi_s\phi_s^\top \in \mathbb{R}^{d\times d}, \ b_t = \sum_{s=1}^{t-1} r_s\phi_s \in \mathbb{R}^{d\times 1}.$$

Therefore, it follows that the gradient of U_t is given by $\nabla U_t(\theta) = \eta(V_t\theta - b_t)$ and its hessian matrix is equal to $\nabla^2 U_t(\theta) = \eta V_t$. Consequently, we recover the well-known fact that the posterior is a Gaussian distribution with mean $\hat{\mu}_t = V_t^{-1} b_t$ and covariance matrix $\hat{\Sigma}_t = (\eta V_t)^{-1}$. Denote by $\tilde{\mathbb{Q}}_{1:T}$ the distribution on the sequence of policies induced by the sequence of variational posterior $\{\tilde{\mathrm{q}}_t = \mathrm{N}(\tilde{\mu}_{t,K_t}, B_{t,K_t}^\top B_{t,K_t})\}_{t\in[T]}$ obtained with VITS − I or VITS − II. We now state our main result on the cumulative regret associated to VITS − I and VITS − II for linear contextual bandit, where a the proof is provided in Appendix A.

Theorem 3.5. *Assume Assumptions 3.1 to 3.4 hold. For the choice of hyperparameters $\{K_t, h_t\}_{t\in[T]}$ and η specified in Section A.2, for any $\delta \in (0,1)$, with probability at least $1-\delta$, the cumulative regret is bounded by*

$$\mathrm{CRegret}(\tilde{\mathbb{Q}}_{1:T}) \leq \frac{CR^2 d\sqrt{dT}\log(3T^3)}{\lambda^2} \log\left(\frac{(1+T/\lambda d)}{\delta}\right)$$

where $C \geq 0$ is a constant independent of the problem. Our main result shows that the distribution of the sequence of policies generated by VITS − I or VITS − II results in a cumulative regret of order $\tilde{\mathcal{O}}(d\sqrt{dT})$. It is in the same order as the state-of-the-art cumulative regret obtained in (Agrawal & Goyal, 2013) for LinTS. The number of optimization steps K_t we found are of order $\kappa_t^2 \log(dT\log(T))$ where $\kappa_t = \lambda_{\max}(V_t)/\lambda_{\min}(V_t)$. Following (Hamidi & Bayati, 2020; Wu et al., 2020), if the diverse context assumption holds, the condition number is $\kappa_t = \mathcal{O}(1)$. Therefore, under this previous assumption, VITS − I and VITS − II require a number of optimization steps that scale as $\log(dT\log(T))$. Finally, (Xu et al., 2022) derived similar bounds for TS using LMC for linear contextual bandit problems. Although our proof is based on the linear case, it could be extended to more general cases insofar as our updates remain Gaussian by definition of the variational family. This allows the use of Gaussian (anti) concentration bound in the theoretical analysis. This is in contrast to other approximation methods, which do not possess this advantage.

Comparison table. In this paragraph we have added a comparison table between Linear TS (LinTS), Linear UCB (LinUCB), Feel-Good TS (Huix et al., 2023; Zhang, 2022), VITS − I, VITS − II (VITS-I/II), VITS − II Hessian-free (VITS-II HF), Langevin Monte Carlo TS (LMCTS) and Variational TS (VTS). The column "Regret" corresponds to the theoretical regret bound obtained by the algorithm. "Complexity" is the computational complexity, more precisely the symbol $(++)$ corresponds to a regret $O(\sqrt{dT})$, $(+)$ to $O(d^{3/2}\sqrt{T})$ and $(-)$ to no existing regret bound. " Linear" is set to Yes when the algorithm is designed only for the Linear Bandit setting and No for general setting including Linear. The "Conditioning" column describes the algorithm's robustness against the conditioning of the problem.

	Regret	Complexity	Linear	Conditioning
LinTS	+	++	Yes	++
LinUCB	++	++	Yes	++
FG-TS	++		No	
VITS-I/II	+	+	No	+
VITS-II HF	-	+ +	No	+
LMC-TS	+	++	No	-
VTS	-	-	No	

4. Numerical experiments

4.1. Linear and quadratic bandit

Our initial investigation focused on a toy setting where contextual vectors are sampled from a Gaussian distribution. However, in this specific setting, the contextual vectors exhibit high diversity, resulting in a posterior covariance matrix with a condition number of $O(1)$. This condition makes the optimization problem overly simplistic, as a result, all approximation methods seem to perform identically in this simple well-conditioned problem. So we introduce a novel setting in which the diversity of arms is controlled by a parameter, denoted as ζ. Firstly, we consider a fixed pool of arms denoted as $P = [\tilde{x}_1, \ldots, \tilde{x}_n]$ with $n = 50$, where each arm $\tilde{x}_i$ follows a normal distribution $\mathrm{N}(0_d, \mathrm{I}_d)$. This fixed pool is relevant in real-world scenarios, such as in a Recommender system, where this pool corresponds to the concept of a meta-user. Then, at each step $t \in [T]$, for every arm, we randomly sample a vector $\tilde{x}_i$ from the pool P, and the contextual vector associated with this arm is defined as $x = \tilde{x}_i + \zeta\epsilon$, where $\epsilon \sim \mathcal{N}(0_d, \mathrm{I}_d)$. When ζ has a high value, the corresponding user is far from the meta-user. Consequently, the diversity among arms is high, resulting in a well-conditioned problem. However, in cases where ζ is low, the problem is ill-conditioned and the optimization becomes challenging.

We consider the linear bandit and the quadratic bandit problems. In both settings, the bandit environment is simulated using a random vector $\theta^\star$ sampled from a normal distribution $\mathrm{N}(0_d, \sigma^\star \mathrm{I}_d)$. We opted for $\sigma^\star = 1/d$ to ensure that the variance of the scalar product $x^\top\theta^\star$ remains independent of the dimension d. The parameter dimension d is set to 20 and we consider a number of arms $K = 50$. In the linear bandit setting, the reward associated with the contextual vector x, is $r = x^\top\theta^\star + \alpha\epsilon$ where $\epsilon \sim \mathrm{N}(0_d, \mathrm{I}_d)$. However, to maintain problem complexity independent of ζ, we have set the signal-to-noise ratio to a fixed value of 1, meaning $\mathbb{E}[(x^\top\theta^\star)^2]/\mathbb{E}[(\alpha\epsilon)^2] = 1$. This implies that $\sqrt{1+\zeta^2} = \alpha$. See Appendix H for more details about the setting. In these experiments, we have chosen to com-

pare VITS − II, VITS − II Hessian-free, Linear TS (LinTS), and LMC-TS, with 10 and 50 iterations of Langevin diffusion at each step. For VITS based algorithm, we have only used 10 updating steps. We have omitted the performance of VITS − I since it experimentally performs identically to VITS − II. For the algorithm VITS − II Hessian-free, we approximate the integral presented in (13) using 20 Monte Carlo samples. This choice is made due to the observed instability caused by the Monte Carlo error when considering high values of η. However, in our setting, even with 20 Monte Carlo samples, VITS − II Hessian-free remains a faster method compared to VITS − II. We also attempted to assess the performance of VTS, but, in the ill-conditioned setting, it exhibited a linear and notably high cumulative regret. Consequently, we have opted to exclude it from the figure for the sake of clarity and visibility. The mean and standard error are reported for all experiments over 50 runs. The hyperparameter is provided in Appendix E.

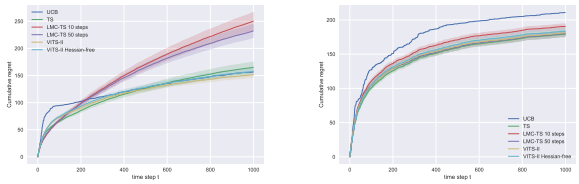


Figure 1. Linear bandits, $\zeta = 0.1$ (left), $\zeta = 1$ (right).

Figure 1 illustrates the cumulative regret with respect to the time step t for a well-conditioned problem ($\zeta = 1$) and a ill-conditioned problem ($\zeta = 0.1$). Firstly, for $\zeta = 1$, it appears that all methods exhibit similar performance, with the exception of LMC-TS with 10 steps, which slightly underperforms. However, for $\zeta = 0.1$, the optimization problem becomes harder and LMC-TS underperforms even with 50 Langevin steps. This behaviour was expected in our setting, because LMC requires a lot of iterations to converge to the posterior compared to VI. A more complete explanation of this phenomenon can be found in Appendix F. Finally, we can conclude that VITS − II performs similarly to LinTS and that its Hessian-free version slightly underperforms but is computationally more efficient.

For Quadratic bandit in Fig 2, the reward is $r = (x^\top \theta^\star)^2 + \alpha\epsilon$. This setting is similar to the Linear setting, but we ensure the condition $\mathbb{E}[(x^\top \theta^\star)^4]/\mathbb{E}[(\alpha\epsilon)^2] = 1$ to still get the signal-to-noise ratio equals to 1. This implies a slight different condition $\alpha = (\zeta^2 + 1)\sqrt{3 + 6/d}$, see Appendix H. Moreover, a simple MLP with two hidden layers of 20 neurons is used for LMC, VITS − II, and its Hessian-free version as neural network architecture. Performance in Fig 2 are similar to linear bandits where VITS − II slightly performs better than its Hessian-free version but outperforms both LMC and LinTS algorithms as LinTS is not adapted for this setting. The gap between LMC and our algorithm is smaller in the well-conditioned setting than in the ill-conditioned, which was also expected. Finally, additional experience on non-contextual bandits can also be found in Appendix G.

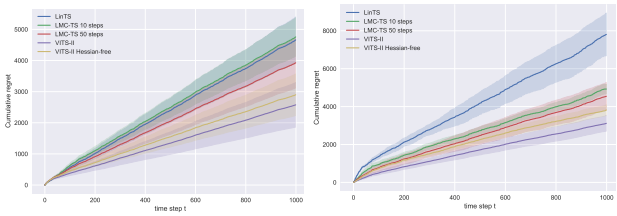


Figure 2. Quadratic bandit, $\zeta = 0.1$(left), $\zeta = 1$(right).

5. MovieLens Dataset

In this section, we evaluate VITS on the MovieLens dataset, consisting of one million ratings by 6040 users for 3952 movies. We adopt the setup proposed in (Aouali et al., 2022), involving a low-rank factorization of the rating matrix to yield 5-dimensional representations for users ($x_j \in \mathbb{R}^5$) and movies ($\theta_i \in \mathbb{R}^5$). Movies are treated as potential actions, and context x_t is uniformly sampled from the pool of user vectors. We consider logistic rewards, sampled from $\text{Ber}(\mu(x_j^\top \theta_i))$, where μ is the sigmoid function. We conduct 50 simulations, each involving 100 randomly selected movies.Our prior distribution employs a Gaussian distribution with mean μ_0 and covariance $\Sigma_0 = \text{diag}(\sigma_0)$. Here, μ_0 and σ_0 represent the mean and variance of movie vectors across all dimensions. This setting deviates somewhat from our theoretical framework, where we consider a unified posterior distribution for all arms using a feature map function ϕ representing context-action pairs. In the MovieLens context, each arm possesses an individual posterior distribution. These two settings closely align when the feature map is the vector concatenation function. In practice, we can apply VITS or LMC at each arm to obtain posterior samples. In this experiment, we compare LinTS against LMC-TS, VITS − II, and the VITS − II Hessian-free variant. LMC-TS uses 10 Langevin updating steps. It's crucial to note that for each time step t and each arm a, LMC-TS requires running Langevin diffusion to obtain a new parameter with low correlation to the previous one. This leads to a high computational complexity for LMC-TS. In contrast, VITS for each arm only involves sampling from a low-dimensional Gaussian distribution and updating the variational posterior corresponding to the chosen arm. This approach offers significant computational efficiency.

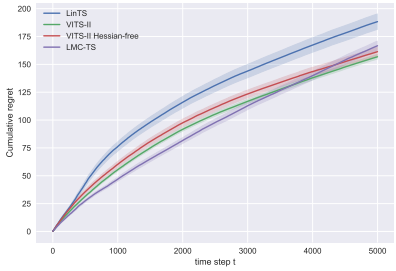


Figure 3. Cumulative regret for MovieLens dataset.

Figure 3 reveals that LinTS is ill-suited for this particular task, as it assumes rewards to be linear while the approximated algorithms outperform LinTS, as they specifically target the logistic posterior. Remarkably, VITS appears to slightly outperform LMC-TS, despite its computational efficiency advantages.

6. Conclusion and perspectives

This paper presents two novel TS algorithms called VITS − I, VITS − II that use VI as an approximation method. These algorithms provide robust theoretical guarantees, in particular a cumulative regret bound of $\tilde{\mathcal{O}}(d\sqrt{dT})$ in the linear setting.

One limitation of our analysis is that the regret bound derived is limited to the linear setting while the interest of our algorithm relies on nonlinear tasks. Additionally, we introduce a third algorithm named VITS − II Hessian-free, which offers enhanced computational efficiency. This algorithm removes the computations of Hessian, resulting in faster execution. Finally, all algorithms have been extensively evaluated in both simulated and real problems.

Acknowledgements

Pierre Clavier has been supported by a grant from Région Île-de-France. Tom Huix acknowledge the support of ANR-CHIA-002, "Statistics, computation and Artificial Intelligence". Part of the work has been developed under the auspice of the Lagrange Center for Mathematics and Calculus. This work was granted access to the HPC resources of IDRIS under the allocation AD011014442 and AD011013313R2 made by GENCI (Grand Equipement National de Calcul Intensif) Alain Durmus is Funded by the European Union (ERC, Ocean, 101071601). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Impact Statement

This paper presents a work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011a.

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011b. URL https://proceedings.neurips.cc/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Abramowitz, M. and Stegun, I. A. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1964.

Agarwal, A., Bird, S., Cozowicz, M., Hoang, L., Langford, J., Lee, S., Li, J., Melamed, D., Oshri, G., Ribas, O., Sen, S., and Slivkins, A. Making contextual decisions with low technical debt, 2016. URL https://arxiv.org/abs/1606.03966.

Agrawal, S. and Goyal, N. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pp. 39–1. JMLR Workshop and Conference Proceedings, 2012.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pp. 127–135. PMLR, 2013.

Aouali, I., Kveton, B., and Katariya, S. Generalizing hierarchical bayesian bandits. *arXiv preprint arXiv:2205.15124*, 2022.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Berry, D. A. and Fristedt, B. Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall*, 5 (71-87):7–7, 1985.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Bonnabel, S. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58 (9):2217–2229, 2013.

Bouneffouf, D., Rish, I., and Aggarwal, C. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8. IEEE, 2020.

Chapelle, O. and Li, L. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In Gordon, G., Dunson, D., and Dudík, M. (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/chu11a.html.

Faury, L., Abeille, M., Jun, K.-S., and Calauzènes, C. Jointly efficient and optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 546–580. PMLR, 2022.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric bandits: The generalized linear case. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/c2626d850c80ea07e7511bbae4c76f4b-Paper.pdf.

Foster, D. and Rakhlin, A. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 3199–3210. PMLR, 2020.

Hamidi, N. and Bayati, M. On worst-case regret of linear thompson sampling. *arXiv preprint arXiv:2006.06790*, 2020.

Huix, T., Zhang, M., and Durmus, A. Tight regret and complexity bounds for thompson sampling via langevin monte carlo. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8749–8770. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/huix23a.html.

Jin, T., Xu, P., Shi, J., Xiao, X., and Gu, Q. Mots: Minimax optimal thompson sampling. In *International Conference on Machine Learning*, pp. 5074–5083. PMLR, 2021.

Katehakis, M. N. and Veinott, A. F. The multi-armed bandit problem: Decomposition and computation. *Math. Oper. Res.*, 12:262–268, 1987.

Katsevich, A. and Rigollet, P. On the approximation accuracy of gaussian variational inference. *arXiv preprint arXiv:2301.02168*, 2023.

Kaufmann, E., Korda, N., and Munos, R. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pp. 199–213. Springer, 2012.

Kveton, B., Zaheer, M., Szepesvari, C., Li, L., Ghavamzadeh, M., and Boutilier, C. Randomized exploration in generalized linear bandits. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2066–2076. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/kveton20a.html.

Lam, S. and Herlocker, J. Movielens dataset. URL http://grouplens.org/datasets/movielens/,2016.

Lambert, M., Bonnabel, S., and Bach, F. The recursive variational gaussian approximation (r-vga). *Statistics and Computing*, 32(1):10, 2022a.

Lambert, M., Chewi, S., Bach, F., Bonnabel, S., and Rigollet, P. Variational inference via wasserstein gradient flows. *arXiv preprint arXiv:2205.15902*, 2022b.

Langford, J. and Zhang, T. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

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

Li, L., Lu, Y., and Zhou, D. Provably optimal algorithms for generalized linear contextual bandits. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2071–2080. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/li17c.html.

Lu, X. and Van Roy, B. Ensemble sampling. *Advances in neural information processing systems*, 30, 2017.

Mazumdar, E., Pacchiano, A., Ma, Y.-a., Bartlett, P. L., and Jordan, M. I. On thompson sampling with langevin algorithms. *arXiv preprint arXiv:2002.10002*, 2020.

Ménard, P. and Garivier, A. A minimax and asymptotically optimal algorithm for stochastic bandits. In *International Conference on Algorithmic Learning Theory*, pp. 223–237. PMLR, 2017.

Riquelme, C., Tucker, G., and Snoek, J. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SyYe6k-CW.

Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Russo, D. and Van Roy, B. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39 (4):1221–1243, 2014.

Russo, D. and Van Roy, B. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

Tewari, A. and Murphy, S. A. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health - Sensors, Analytic Methods, and Applications*, 2017.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Urteaga, I. and Wiggins, C. Variational inference for the multi-armed contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pp. 698–706. PMLR, 2018.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Wu, W., Yang, J., and Shen, C. Stochastic linear contextual bandits with diverse contexts. In *International Conference on Artificial Intelligence and Statistics*, pp. 2392–2401. PMLR, 2020.

Xu, P., Wen, Z., Zhao, H., and Gu, Q. Neural contextual bandits with deep representation and shallow exploration. *arXiv preprint arXiv:2012.01780*, 2020.

Xu, P., Zheng, H., Mazumdar, E. V., Azizzadenesheli, K., and Anandkumar, A. Langevin monte carlo for contextual bandits. In *International Conference on Machine Learning*, pp. 24830–24850. PMLR, 2022.

Yu, T., Kveton, B., Wen, Z., Zhang, R., and Mengshoel, O. J. Graphical models meet bandits: A variational thompson sampling approach. In *International Conference on Machine Learning*, pp. 10902–10912. PMLR, 2020.

Zenati, H., Bietti, A., Diemert, E., Mairal, J., Martin, M., and Gaillard, P. Efficient kernelized ucb for contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 5689–5720. PMLR, 2022.

Zhang, T. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022.

Zhang, W., Zhou, D., Li, L., and Gu, Q. Neural thompson sampling. *arXiv preprint arXiv:2010.00827*, 2020.

Zhou, D., Li, L., and Gu, Q. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, pp. 11492–11502. PMLR, 2020.

A. Proof of the regret bound

A.1. Proof of Theorem 3.5

While (Lambert et al., 2022b) establishes quantitative bounds on the bias introduced by Algorithm 2 for the VI of the posterior. Combining this result with the one derived in (Agrawal & Goyal, 2013) for TS leads to sub-optimal regret bounds. It is similar to LMC-TS (Xu et al., 2022) which had to make a clever adaptation of (Agrawal & Goyal, 2013). Similar to this work, we need here to revise the proof of (Agrawal & Goyal, 2013) to VITS. We give in this section the main steps of our proofs. Each step is based on Lemmas which are stated and proved in the next sections. First, we define the filtration $(\mathcal{F}_t)_{t\in\{0,\dots,T-1\}}$ such that for any $t \in [T]$, $\mathcal{F}_{t-1}$ is the σ-field generated by $\mathcal{H}_{t-1}$ and x_t where $\mathcal{H}_{t-1} = \{(x_s, a_s, r_s)\}_{s\leq t-1}$ is the observations up to $t-1$ and x_t is the contextual vector at step t. For some feature map $\phi : \mathsf{X} \times \mathsf{A} \to \mathbb{R}$ and for any $t \in [T]$, we denote by

$$\overset{*}{\phi}_t = \phi(x_t, a_t^\star) \ , \ \phi_t = \phi(x_t, a_t) \ ,$$

the features vector of the best arm $a_t^\star$ and the features vector of the arm a_t chosen by VITS at time t respectively. the difference between the best expected reward and the expected reward obtained by VITS is denoted by

$$\Delta_t = \overset{*}{\phi}_t \theta^\star - \phi_t^\top \theta^\star \ .$$

At each round $t \in [T]$, we consider the set of saturated arms $\mathcal{S}_t$ and unsaturated arms $\mathcal{U}_t$ defined by

$$\mathcal{S}_t = \bigcap_{a\in\mathcal{A}(x_t)} \{\Delta_t(a) > g(t) \ \|\phi(x_t, a)\|_{V_t^{-1}}\} \ , \tag{17}$$

and $\mathcal{U}_t = \mathcal{A}(x_t) \setminus \mathcal{S}_t$ where V_t^{-1} is defined in (15) and

$$g(t) = CR^2 d\sqrt{\log(t)\log(T)}/\lambda^{3/2} \ ,$$

for some constant $C \geq 0$ independent of d, t and T. In addition, consider the events $\mathrm{E}_t^{\text{true}}$ and $\mathrm{E}_t^{\text{var}}$ such that

$$\bigcap_{a\in\mathcal{A}(x_t)} \{|\phi(x_t, a)^\top \hat{\mu}_t - \phi(x_t, a)^\top \theta^\star| \leq \ g_1(t) \ \ \|\phi(x_t, a)\|_{V_t^{-1}}\} \subset \mathrm{E}_t^{\text{true}}$$

$$\mathrm{E}_t^{\text{var}} = \bigcap_{a\subset\mathcal{A}(x_t)} \{|\phi(x_t, a)^\top \tilde{\theta}_t - \phi(x_t, a)^\top \hat{\mu}_t| \leq \ g_2(t) \ \ \|\phi(x_t, a)\|_{V_t^{-1}}\} \ ,$$

where $\hat{\mu}_t$ is given by $\hat{\mu}_t = V^{-1} b_t$ and b_t is given in (15). The specific definitions of $\mathrm{E}_t^{\text{true}}$, g, g_1 and g_2 are given in Section A.3 of the supplementary. Nevertheless, by definition, it holds that $g_1(t) + g_2(t) \leq g(t)$.

1. For ease of notation, the conditional expectation $\mathbb{E}_{\pi_{1:T}\sim\mathbb{Q}_{1:T}}[\cdot]$ and probabilities $\mathbb{P}_{\pi_{1:T}\sim\mathbb{Q}_{1:T}}(\cdot)$ with respect to the σ-field $\mathcal{F}_{t-1}$ are denoted by $\mathbb{E}_t[\cdot]$ and $\mathbb{P}_t(\cdot)$ respectively. Therefore, with these notations, we have by definition of the cumulative regret:

$$\mathrm{CRegret}(\tilde{\mathbb{Q}}_{1:T}) = \sum_{t=1}^{T} \Delta_t \ .$$

We now bound for any $t \in [T]$, with high probability, $\Delta_t(a_t)$. To this end, in the next step of the proof, we show that the stochastic process $(X_t)_{t\in[T]}$ defined below is a $(\mathcal{F}_t)_{t\in[T]}$ super-martingale.

$$X_t = \sum_{s=1}^{t} Y_s$$

with

$$Y_s = \Delta_s - cg(s)\|\phi_s\|_{V_s^{-1}}/p - 2/s^2 \ ,$$

where $p \in (0, 1)$ and c is a sufficiently large real number, independent of d, T and s.

2. **Showing that $(X_t)_{t\in[T]}$ is a super-martingale.** We consider the following decomposition

$$\begin{aligned}\mathbb{E}_t[\Delta_t(a_t)] &= \mathbb{E}_t[\Delta_t(a_t)\mathbb{1}_{\mathrm{E}_t^{\text{true}}}] + \mathbb{E}_t[\Delta_t(a_t)|\bar{\mathrm{E}}_t^{\text{true}}]\mathbb{P}_t(\bar{\mathrm{E}}_t^{\text{true}}) \\ &\leq \mathbb{E}_t[\Delta_t(a_t)\mathbb{1}_{\mathrm{E}_t^{\text{true}}}] + \mathbb{P}_t(\bar{\mathrm{E}}_t^{\text{true}}) \ , \end{aligned} \tag{18}$$

where we used for the last inequality that $\|\theta^\star\|_2 \leq 1$ and Assumption 3.3. Then, since $\mathrm{E}_t^{\text{true}} \in \mathcal{F}_{t-1}$, we have,

$$\begin{aligned}\mathbb{E}_t[\Delta_t(a_t)\mathbb{1}_{\mathrm{E}_t^{\text{true}}}] &= \mathbb{1}_{\mathrm{E}_t^{\text{true}}}\mathbb{E}_t[\Delta_t(a_t)|\mathrm{E}_t^{\text{var}}]\mathbb{P}_t(\mathrm{E}_t^{\text{var}}) + \mathbb{1}_{\mathrm{E}_t^{\text{true}}}\mathbb{E}_t[\Delta_t(a_t)|\bar{\mathrm{E}}_t^{\text{var}}]\mathbb{P}_t(\bar{\mathrm{E}}_t^{\text{var}}) \\ &\leq \mathbb{1}_{\mathrm{E}_t^{\text{true}}}[\mathbb{E}_t[\Delta_t(a_t)|\mathrm{E}_t^{\text{var}}] + \mathbb{P}_t(\bar{\mathrm{E}}_t^{\text{var}})] \end{aligned} \tag{19}$$

where in the last line we have used that $\Delta_t(a_t) \leq 1$ again. Denote by $\bar{a}_t = \arg\min_{a\in\mathcal{U}_t} \|\phi(x_t,a)\|_{V_t^{-1}}$ and $\overline{\phi}_t = \phi(x_t,\bar{a}_t)$. Then, given $\mathrm{E}_t^{\text{true}}$ and $\mathrm{E}_t^{\text{var}}$ we have

$$\begin{aligned}\Delta_t(a_t) &= \overset{*}{\phi}_t^\top\theta^\star - \phi_t^\top\theta^\star \\ &= \overset{*}{\phi}_t^\top\theta^\star - \overline{\phi}_t^\top\theta^\star + \overline{\phi}_t^\top\theta^\star - \phi_t^\top\theta^\star \\ &\overset{(a)}{\leq} g(t)\|\overline{\phi}_t\|_{V_t^{-1}} + \overline{\phi}_t^\top\theta^\star - \phi_t^\top\theta^\star \\ &\overset{(b)}{\leq} g(t)\|\overline{\phi}_t\|_{V_t^{-1}} + (\overline{\phi}_t^\top\tilde{\theta}_t + g(t)\|\overline{\phi}_t\|_{V_t^{-1}}) - (\phi_t^\top\tilde{\theta}_t - g(t)\|\phi_t\|_{V_t^{-1}}) \\ &\overset{(c)}{\leq} (2\|\overline{\phi}_t\|_{V_t^{-1}} + \|\phi_t\|_{V_t^{-1}})g(t) \end{aligned} \tag{20}$$

where inequality (a) is due to $\bar{a}_t \in \mathcal{U}_t$, and therefore $\Delta_t(\bar{a}_t) \leq g(t)\|\overline{\phi}_t\|_{V_t^{-1}}$, inequality (b) uses that given $\mathrm{E}_t^{\text{true}}$ and $\mathrm{E}_t^{\text{var}}$, for any $\phi \in \mathbb{R}^d$, $|\phi^\top\tilde{\theta}_t - \phi^\top\theta^\star| \leq g(t)\|\phi\|_{V_t^{-1}}$ since by definition $g_1(t) + g_2(t) \leq g(t)$; finally, the arm a_t maximizes the quantity $\phi(x_t,a_t)^\top\tilde{\theta}_t$, $\overline{\phi}_t^\top\tilde{\theta}_t - \phi_t^\top\tilde{\theta}_t$ is obviously negative, which implies inequality (c).
Moreover, given $\mathrm{E}_t^{\text{true}}$ and $\mathrm{E}_t^{\text{var}}$,

$$\begin{aligned}\mathbb{E}_t[\|\phi_t\|_{V_t^{-1}}] &= \mathbb{E}_t[\|\phi_t\|_{V_t^{-1}}|a_t \in \mathcal{U}_t]\mathbb{P}_t(a_t \in \mathcal{U}_t) + \mathbb{E}_t[\|\phi_t\|_{V_t^{-1}}|a_t \in \mathcal{S}_t]\mathbb{P}_t(a_t \in \mathcal{S}_t) \\ &\overset{(a)}{\geq} \|\overline{\phi}_t\|_{V_t^{-1}}\mathbb{P}_t(a_t \in \mathcal{U}_t) \\ &\overset{(b)}{\geq} (p - 1/t^2)\|\overline{\phi}_t\|_{V_t^{-1}} \end{aligned}$$

where (a) is due to the definition of $\overline{\phi}_t$, i.e. for any $a \in \mathcal{U}_t$, $\|\overline{\phi}_t\|_{V_t^{-1}} \leq \|\phi(x_t,a)\|_{V_t^{-1}}$, and (b) uses Lemma A.4 with $p \in (0,1)$. Here is one of the main differences with the proof conducted by (Agrawal & Goyal, 2013). Indeed, to obtain such a bound, we need to carefully dig into the convergence of the the sequence of means $\{\tilde{\mu}_{t,K_t}\}_{k\in[1,K_t]}$ and covariance matrices $\{\tilde{\Sigma}_{t,K_t}\}_{k\in[1,K_t]}$ to obtain a fine-grained analysis of the distribution of $\tilde{\mathrm{q}}_t$. Therefore, using equations (19) and (20)

$$\mathbb{1}_{\mathrm{E}_t^{\text{true}}}\mathbb{E}_t[\Delta_t(a_t)] \leq \Big(\frac{2}{p - 1/t^2} + 1\Big)g(t)\mathbb{E}_t[\|\phi_t\|_{V_t^{-1}}] + \frac{1}{t^2} \leq \frac{cg(t)}{p}\mathbb{E}_t[\|\phi_t\|_{V_t^{-1}}] + \frac{1}{t^2} \ ,$$

where c is a sufficiently large real number independent of the problem. Plugging this bounds in 18, we obtain

$$\mathbb{E}_t[\Delta_t(a_t)] \leq \frac{cg(t)}{p}\mathbb{E}_t[\|\phi\|_{V_t^{-1}}] + \frac{1}{t^2} + \mathbb{P}_t(\bar{\mathrm{E}}_t^{\text{true}})$$

Applying Lemma A.3 yields

$$\mathbb{E}_t[\Delta_t(a_t)] \leq \frac{cg(t)}{p}\mathbb{E}_t[\|\phi\|_{V_t^{-1}}] + \frac{2}{t^2} \ .$$

This is another important difference with the original proof of (Agrawal & Goyal, 2012) which uses our precise convergence study for $\{\tilde{\mu}_{t,K_t}\}_{k\in[1,K_t]}$. Then, it follows that $(X_t)_{t\in[T]}$ is a $(\mathcal{F}_t)_{t\in[T]}$-super martingale.

3. **Concentration for $(X_t)_{t\in[T]}$.** Note that $(X_t)_{t\in[T]}$ is a super-martingale with bounded increments: for any $t \in [T]$

$$\begin{aligned}|X_{t+1} - X_t| &= |Y_{t+1}| \\ &= |\Delta_t(a_t) - \frac{cg(t)}{p}\|\phi_t\|_{V_t^{-1}} - \frac{2}{t^2}| \\ &\overset{(a)}{\leq} |\Delta_t(a_t)(a_t) - \frac{cg(t)}{\sqrt{\lambda}p} - \frac{2}{t^2}| \\ &\leq \frac{3cg(t)}{\sqrt{\lambda}p} \ ,\end{aligned}$$

where in (a) we have used that

$$\|\phi_t\|_{V_t^{-1}} \leq \|\phi_t\|_{V_1^{-1}} \leq 1/\sqrt{\lambda}$$

and inequality (b) is due to $\Delta_t(a_t) \leq 1$, $2/t^2 \leq 2$ and $3cg(t)/(p\sqrt{\lambda}) > 2$ for an appropriate choice of the numerical constant c. Therefore, applying Azuma-Hoeffding inequality (Lemma (B.3)), with probability $1-\delta$ it holds that

$$X_T \leq \sqrt{2\log(1/\delta)\sum_{s=1}^{T}\frac{9c^2g(s)^2}{p^2\lambda}} \leq \sqrt{18\log(1/\delta)\frac{c^2}{p^2\lambda}g(T)^2T} \ ,$$

using that $g(T) \geq g(t)$.

4. **Conclusion.** The super-martingale $(X_t)_{t=1}^T$ is directly linked to the cumulative regret by

$$\begin{aligned}X_T &= \sum_{t=1}^{T} Y_t \\ &= \sum_{s=1}^{T} \Delta_t - cg(t)\|\phi_t\|_{V_t^{-1}}/p - 2/t^2 \\ &= \mathrm{CRegret}(\tilde{\mathbb{Q}}_{1:T}) - \sum_{t=1}^{T} cg(t)\|\phi_t\|_{V_t^{-1}}/p + 2/t^2\end{aligned}$$

then taking the expectation and using the super-martingale previous argument of the proof, we obtain the following upper bound for the cumulative regret :

$$\mathrm{CRegret}(\tilde{\mathbb{Q}}_{1:T}) \leq \sum_{t=1}^{T}\frac{cg(t)}{p}\|\phi_t\|_{V_t^{-1}} + \sqrt{18\log(1/\delta)\frac{c^2}{p^2\lambda}g(T)^2T} + \frac{\pi^2}{3} \ .$$

using that $\sum_{t=1}^{+\infty} 1/t^2 \leq \pi^2/6$. As a result, applying Lemma A.5 yields

$$\mathrm{CRegret}(\tilde{\mathbb{Q}}_{1:T}) \leq \frac{cg(T)}{p}\sqrt{2dT\log\left(1+T/(\lambda d)\right)} + \frac{cg(T)}{p\sqrt{\lambda}}\sqrt{18\log(1/\delta)T} + \frac{\pi^2}{3} \ .$$

Using the definition of $g(T)$ in (18), we get

$$\mathrm{CRegret}(\tilde{\mathbb{Q}}_{1:T}) \leq \frac{CR^2d}{\lambda^2}\log(3T^3)\sqrt{dT\log\left(1+T/(\lambda d)\right)\log\left(1/\delta\right)} \ ,$$

where $C \geq 0$ is a constant, independent of the problem, which completes the proof.

A.2. Hyperparameters choice and values

In this section, we define and discuss the values of the main hyperparameters.

Parameter η : is the inverse of the temperature. The lower is η, the better is the exploration. It is fixed to

$$\eta = 4\lambda^2/(81R^2 d\log(3T^3)) \leq 1 \tag{21}$$

Parameter λ : is the inverse of the standard deviation of the prior distribution. It controls the regularization. The lower is λ, the better is the exploitation. This parameter is fixed but lower than 1.

Parameter h_t : is the step size used in all Algorithms. It is fixed to

$$h_t = \lambda_{\min}(V_t)/(2\eta(\lambda_{\min}(V_t)^2 + 2\lambda_{\max}(V_t)^2)) \tag{22}$$

Parameter K_t : is the number of gradient descent steps performed. It is fixed to

$$K_t = 1 + 2(1+2\kappa_t^2)\log\left(2R\kappa_t d^2 T^2 \log^2(3T^3)\right) . \tag{23}$$

Therefore the number of gradient descent steps is $K_t \leq \mathcal{O}(\kappa_t^2 \log(dT\log(T)))$.

A.3. Useful definitions

Definition A.1. (Variational approximation) Recall that $\hat{\mathrm{p}}_t(\theta) \propto \exp(-U_t(\theta))$ is the posterior distribution. And $\tilde{\mathrm{q}}_t$ is the variational posterior distribution in the sense that

$$\tilde{\mathrm{q}}_t = \arg\min_{\mathrm{p}\in\mathcal{G}} \mathrm{KL}(\mathrm{p}|\hat{\mathrm{p}}_t) ,$$

where $\mathcal{G}$ is a variational family. In this paper we focus on the Gaussian variational family and we denote by $\tilde{\mu}_t$ and B_t respectively the mean and the square root covariance matrix of the variational distribution, ie,

$$\tilde{\mathrm{q}}_t = \mathrm{N}(\tilde{\mu}_t, B_t B_t^\top) .$$

The values of $\tilde{\mu}_t$ and B_t are obtained after running K_t steps of algorithm 2 or 3. Note that the sequence of means $\{\tilde{\mu}_{t,k}\}_{k=1}^{K_t}$ is defined recursively by

$$\begin{aligned}\tilde{\mu}_{t,k+1} &= \tilde{\mu}_{t,k} - h_t \nabla U_t(\tilde{\theta}_{t,k}) \\ &= \tilde{\mu}_{t,k} - h_t \eta V_t(\tilde{\theta}_{t,k} - \hat{\mu}_t)\end{aligned}$$

where $\tilde{\theta}_{t,k} \sim \mathrm{N}(\tilde{\mu}_{t,k}, B_{t,k}^\top B_{t,k})$ and we have used that $\nabla U_t(\theta) = \eta(V_t\theta - b_t)$ (see equation (15)). Consequently, $\tilde{\mu}_{t,k}$ is also Gaussian and we denote by $\tilde{m}_{t,k}$ and $\tilde{W}_{t,k}$ its mean and covariance matrix, ie, $\tilde{\mu}_{t,k} \sim \mathrm{N}(\tilde{m}_{t,k}, \tilde{W}_{t,k})$. Furthermore, the sequence of square root covariance matrix $\{B_{t,k}\}_{k=1}^{K_t}$ is defined recursively in Algorithm 2 by

$$\begin{aligned}B_{t,k+1} &= \left\{\mathrm{I} - h_t \nabla^2(U_t(\tilde{\theta}_{t,k}))\right\} B_{t,k} + (B_{t,k}^\top)^{-1} \\ &= \{\mathrm{I} - \eta h_t V_t\} B_{t,k} + h_t (B_{t,k}^\top)^{-1}\end{aligned}$$

where we have used that $\nabla^2(U_t(\theta)) = \eta V_t$ for the linear bandit case (see (15)). Let denote by $\tilde{\Sigma}_{t,k} = B_{t,k}B_{t,k}^\top$ the covariance of the variational posterior $\tilde{\mathrm{q}}_{t,k}$. For ease of notation we denote by $A_t = \mathrm{I} - \eta h_t V_t$, it follows that

$$\tilde{\Sigma}_{t,k+1} = A_t \tilde{\Sigma}_{t,k} A_t + 2h_t A_t + h_t^2 \tilde{\Sigma}_{t,k}^{-1}$$

If $\Lambda_{t,k} = \tilde{\Sigma}_{t,k} - 1/\eta V_t^{-1}$ denotes the difference between the covariance matrix of the varational posterior and the true posterior, therefore it holds that

$$\begin{aligned}\Lambda_{t,k+1} &= A_t \tilde{\Sigma}_{t,k} A_t + 2h_t A_t + h_t^2 \tilde{\Sigma}_{t,k}^{-1} - 1/\eta V_t^{-1} \\ &= A_t \Lambda_{t,k} A_t + 2h_t A_t - 2h_t \mathrm{I} + \eta h_t^2 V_t + h_t^2 \tilde{\Sigma}_{t,k}^{-1} \\ &= A_t \Lambda_{t,k} A_t - \eta h_t^2 V_t + h_t^2 \tilde{\Sigma}_{t,k}^{-1} \\ &= A_t \Lambda_{t,k} A_t - h_t^2 \eta V_t \Lambda_{t,k} \tilde{\Sigma}_{t,k}^{-1}\end{aligned}$$

In the case of VITS − II, the sequence of square root covariance matrix $\{B_{t,k}\}_{k\le K_t}$ and the sequence of inverse square root covariance matrix $\{C_{t,k}\}_{k\le K_t}$ are defined recursively in Algorithm 3 by

$$C_{t,k+1} = C_{t,k}\{\mathrm{I} - h_t(C_{t,k}^\top C_{t,k} - \nabla^2 U_t(\tilde\theta_{t,k}))\}$$
$$B_{t,k+1} = (\mathrm{I} - h_t\nabla^2 U_t(\tilde\theta_{t,k}))B_{t,k} + h_t C_{t,k}^\top \ .$$

Nevertheless, we will show that this approximation does not impact the cumulative regret bound. Note that all sequences are deterministic in the specific setting of linear bandit, because the Hessian of $\nabla^2 U_t(\tilde\theta)$ does not depend on $\tilde\theta$. In VITS-II, we obtain the following form for $\Lambda_{t,k}$, see Lemma A.12.

$$\Lambda_{t,k+1} = A_t B_{t,k}^2 A_t + 2h_t A_t(B_{t,k}C_{t,k} + C_{t,k}^T B_{t,k}^T) + h_t^2 C_{t,k}^2 - 1/\eta V_t^{-1}$$

Definition A.2. (Concentration events)

The main challenge for the proof of Theorem 3.5, is to control the probability of the following events: for any $t \in [T]$ we define

- $\widehat{\mathrm{E}_t^{\text{true}}} = \Big\{\text{for any } a \in \mathcal{A}(x_t):\ |\phi(x_t,a)^\top\hat\mu_t - \phi(x_t,a)^\top\theta^\star| \le\ g_1(t)\ \ \|\phi(x_t,a)\|_{V_t^{-1}}\Big\}$
- $\mathrm{E}_t^{\text{true}} = \widehat{\mathrm{E}_t^{\text{true}}} \bigcap \Big\{|\xi_t| < R\sqrt{1+\log 3t^2}\Big\} \bigcap \Big\{\|\tilde W_{t,K_t}^{-1/2}(\tilde\mu_{t,K_t} - \tilde m_{t,K_t})\| \le \sqrt{4d\log 3t^3}\Big\}$
- $\mathrm{E}_t^{\text{var}} = \Big\{\text{for any } a \in \mathcal{A}(x_t):\ |\phi(x_t,a)^\top\tilde\theta_t - \phi(x_t,a)^\top\hat\mu_t| \le\ g_2(t)\ \ \|\phi(x_t,a)\|_{V_t^{-1}}\Big\}\ ,$

where $g_1(t) = R\sqrt{d\log(3t^3)} + \sqrt{\lambda}$ and $g_2(t) = 10\sqrt{d\log(3t^3)/(\eta\lambda)}$ and ξ_t is the R-sub Gaussian noise of the reward definition defined by the relation

$$r_t = \phi_t^\top\theta^\star + \xi_t\ . \tag{24}$$

The first event $\widehat{\mathrm{E}_t^{\text{true}}}$ controls the concentration of $\phi(x_t,a)^\top\hat\mu_t$ around its mean. Similarly, event $\mathrm{E}_t^{\text{var}}$ controls the concentration of $\phi(x_t,a)^\top\tilde\theta_t$ around its mean. Note that compared to (Agrawal & Goyal, 2013), in our case, it is important to include within $\mathrm{E}_t^{\text{true}}$, the concentration of the distributions ξ_t and $\tilde W_{t,K_t}^{-1/2}(\tilde\mu_{t,K_t} - \tilde m_{t,K_t})$. Consequently, conditionally on $\mathrm{E}_t^{\text{true}}$ and $\mathrm{E}_t^{\text{var}}$ it holds that: for any $a \in \mathcal{A}(x_t)$

$$\begin{aligned}
|\phi(x_t,a)^\top\tilde\theta_t - \phi(x_t,a)^\top\theta^\star| &\le \left(R\sqrt{d\log(3t^3)} + \sqrt{\lambda} + 10\sqrt{d\log(3t^3)/(\eta\lambda)}\right)\|\phi(x_t,a)\|_{V_t^{-1}} \\
&\le 12R\sqrt{d\log(3t^3)/(\eta\lambda)}\ \|\phi(x_t,a)\|_{V_t^{-1}} \\
&\overset{(a)}{=} \frac{108dR^2}{\lambda^{3/2}}\sqrt{\log(3t^3)\log(3T^3)}\|\phi(x_t,a)\|_{V_t^{-1}} \\
&:= g(t)\|\phi(x_t,a)\|_{V_t^{-1}}\ ,
\end{aligned} \tag{25}$$

where in (a), we have used that $\eta = 4\lambda/\Big(81R^2 d\log(3T^3)\Big)$ and in the last inequality we have used that $g(t) = CR^2 d\sqrt{\log(t)\log(T)}/\lambda^{3/2}$.

A.4. Main lemmas

Lemma A.3. ***(Concentration lemma for $\hat\mu_t$)***

Recall the definition of the event E_t^{true} in (A.2). Therefore, for any $t \in [T]$, it holds that

$$\mathbb{P}(\mathrm{E}_t^{true}) \ge 1 - \frac{1}{t^2} \tag{26}$$

This lemma shows that the mean of the posterior distribution $\hat\mu_t$ is concentrated around the true parameter $\theta^\star$ with high probability.

Proof. Firstly, we apply Lemma B.4, with $m_t = \phi_t/\sqrt{\lambda} = \phi(x_t, a_t)/\sqrt{\lambda}$ and $\epsilon_t = \left(r_{a_t}(t) - \phi_t^\top \theta^\star\right)/\sqrt{\lambda}$, where $r_{a_t}(t)$ is sampled from the R-sub-Gaussian reward distribution of mean $\phi_t^\top \theta^\star$. Let's define the filtration $\mathcal{F}'_t = \{a_{\tau+1}, m_{\tau+1}, \epsilon_\tau\}_{\tau \leq t}$. By the definition of $\mathcal{F}'_t$, m_t is $\mathcal{F}'_{t-1}$-measurable. Moreover, ϵ_t is conditionally $R/\sqrt{\lambda}$-sub-Gaussian due to Assumption 3.1 and is a martingale difference process because $\mathbb{E}[\epsilon_t|\mathcal{F}'_{t-1}] = 0$. If we denote by

$$M_t = \mathrm{I}_d + 1/\lambda \sum_{\tau=1}^{t} m_\tau m_\tau^\tau = 1/\lambda V_{t+1} \ ,$$

and

$$\zeta_t = \sum_{\tau=1}^{t} m_\tau \epsilon_\tau \ ,$$

Then, Lemma B.4 shows that $\|\zeta_t\|_{M_t^{-1}} \leq R/\sqrt{\lambda}\sqrt{d\log(\frac{t+1}{\delta'})}$ with probability at least $1-\delta'$. Moreover, note that

$$\begin{aligned} M_{t-1}^{-1}(\zeta_{t-1} - \theta^\star) &= M_t^{-1}(1/\lambda b_t - 1/\lambda \sum_{\tau=1}^{t-1} \phi_\tau \phi_\tau^\top \theta^\star - \theta^\star) \\ &= M_{t-1}^{-1}(1/\lambda b_t - M_{t-1}\theta^\star) \\ &= \hat{\mu}_t - \theta^\star \ . \end{aligned}$$

Note that $||\theta^\star||_{M_{t-1}^{-1}} = \|\theta^\star M_{t-1}^{-1/2}\|_2 \leq \|\theta^\star\|_2 \|M_{t-1}^{-1/2}\|_2 \leq \|\theta^\star\|_2$, where the last inequality is due to Assumption 3.2. Then, for any arm $a \in \mathcal{A}(x_t)$ we have

$$\begin{aligned} |\phi(x_t,a)^\top \hat{\mu}_t - \phi(x_t,a)^\top \theta^\star| &= |\phi(x_t,a) M_{t-1}^{-1}(\xi_{t-1} - \theta^\star)| \\ &\leq \|\phi(x_t,a)\|_{M_{t-1}^{-1}} \|\xi_{t-1} - \theta^\star\|_{M_{t-1}^{-1}} \\ &\leq \|\phi(x_t,a)\|_{M_{t-1}^{-1}} (\|\xi_{t-1}\|_{M_{t-1}^{-1}} + \|\theta^\star\|_{M_{t-1}^{-1}}) \\ &\leq \sqrt{\lambda}\left(R/\sqrt{\lambda}\sqrt{d\log(\frac{t}{\delta'})} + 1\right) \|\phi(x_t,a)\|_{V_t^{-1}} \\ &= \sqrt{\lambda}\left(R/\sqrt{\lambda}\sqrt{d\log(3t^3)} + 1\right) \|\phi(x_t,a)\|_{V_t^{-1}} \\ &= \left(R\sqrt{d\log(3t^3)} + \sqrt{\lambda}\right) \|\phi(x_t,a)\|_{V_t^{-1}} \\ &:= g_1(t) \|\phi(x_t,a)\|_{V_t^{-1}} \ . \end{aligned}$$

This inequality holds with probability at least $\delta' = 1/(3t^2)$.

Moreover, recall the definition of the R-subGaussian noise of the reward definition in section A.2

$$r_t = \phi_t^\top \theta^\star + \xi_t$$

Then it holds that $\mathbb{P}(|\xi_t| > x) \leq \exp(1 - x^2/R^2)$. It follows that $\mathbb{P}(|\xi_t| \leq R\sqrt{1+\log 3t^2}) \geq 1 - 1/(3t^2)$, for any $t \leq 1$. Finally, recall the definition of $\tilde{W}_{t,k}$, $\tilde{\mu}_{t,k}$ and $\tilde{m}_{t,k}$ in section A.1. Consequently, the term $\tilde{W}_{t,K_t}^{-1/2}(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t})$ is gaussian with mean 0 and an identity covariance matrix. Therefore, it holds that

$$\mathbb{P}\Big(\|\tilde{W}_{t,K_t}^{-1/2}(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t})\| \leq \sqrt{4d\log 3t^3}\Big) \geq 1 - 1/(3t^2) \tag{27}$$

Consequently, we have

$$\mathbb{P}\Big(\widehat{\mathrm{E}_t^{\text{true}}} \bigcap \Big\{|\xi_i| < R\sqrt{1+\log 3t^2}\Big\} \bigcap \Big\{\|\tilde{W}_{t,K_t}^{-1/2}(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t})\| \leq \sqrt{4d\log 3t^3}\Big\}\Big) \geq 1 - \frac{1}{t^2} \ ,$$

where $\widehat{\mathrm{E}_t^{\text{true}}}$ is defined is $A.2$

□

Lemma A.4. ***Probability of playing an unsaturated arm***

Given E_t^{true} defined in section (A.2), the conditional probability of playing an unsaturated arm is strictly positive and is lower bounded as

$$\mathbb{1}_{\mathrm{E}_t^{true}}\mathbb{P}_t(a_t \in \mathcal{U}_t) := \mathbb{P}(a_t \in \mathcal{U}_t|\mathcal{F}_{t-1}) \geq \mathbb{1}_{\mathrm{E}_t^{true}}(p - 1/t^2) \ , \tag{28}$$

where $p = 1/\sqrt{2\pi \mathrm{e}}$ and $\mathcal{U}_t$ is defined in (17).

Proof. If we suppose that $\forall a \in \mathcal{S}_t,\ \phi(x_t, a_t^\star)^\top \tilde{\theta}_t \geq \phi(x_t, a)^\top \tilde{\theta}_t$, then $a_t \in \mathcal{U}_t$. Indeed, The optimal arm $a_t^\star$ is obviously in the unsaturated arm set ($\mathcal{U}_t$) and $\phi(x_t, a_t)^\top \tilde{\theta}_t \geq \phi(x_t, a_t^\star)^\top \tilde{\theta}_t$ by construction of the algorithm. Hence we have

$$\mathbb{P}(a_t \in \mathcal{U}_t|\mathcal{F}_{t-1}) \geq \mathbb{P}(\overset{*}{\phi}{}^\top \tilde{\theta}_t \geq \phi(x_t, a)^\top \tilde{\theta}_t, \forall a \in \mathcal{S}_t|\mathcal{F}_{t-1})$$

Subsequently, given events $\mathrm{E}_t^{\text{true}}$ and $\mathrm{E}_t^{\text{var}}$ we have

$$\left\{\overset{*}{\phi}{}^\top \tilde{\theta}_t \geq \phi(x_t, a)^\top \tilde{\theta}_t, \forall a \in \mathcal{S}_t\right\} \supset \left\{\overset{*}{\phi}{}^\top \tilde{\theta}_t \geq \overset{*}{\phi}{}^\top \theta^\star\right\} \ .$$

Indeed, for any $a \in \mathcal{S}_t$,

$$\begin{aligned}\phi(x_t, a)^\top \tilde{\theta}_t &\overset{(a)}{\leq} \phi(x_t, a)^\top \theta^\star + g(t)\|\phi(x_t, a)\|_{\hat{\Sigma}_t} \\ &\overset{(b)}{\leq} \overset{*}{\phi}{}^\top \theta^\star \ ,\end{aligned}$$

where (a) uses that $\mathrm{E}_t^{\text{true}}$ and $\mathrm{E}_t^{\text{var}}$ hold. And in inequality (b) we have used that $a \in \mathcal{S}_t$, ie, $\overset{*}{\phi}{}_t^\top \theta^\star - \phi(x_t, a)^\top \theta^\star := \Delta_t(a) > g(t)\|\phi(x_t, a)\|_{\hat{\Sigma}_t}$.

Consequently,

$$\begin{aligned}\mathbb{P}(\overset{*}{\phi}{}^\top \tilde{\theta}_t \geq \overset{*}{\phi}{}^\top \theta^\star|\mathcal{F}_{t-1}) &= \mathbb{P}(\overset{*}{\phi}{}^\top \tilde{\theta}_t \geq \overset{*}{\phi}{}^\top \theta^\star|\mathcal{F}_{t-1}, \mathrm{E}_t^{\text{var}})\mathbb{P}(\mathrm{E}_t^{\text{var}}) + \mathbb{P}(\overset{*}{\phi}{}^\top \tilde{\theta}_t \geq \overset{*}{\phi}{}^\top \theta^\star|\mathcal{F}_{t-1}, \overline{\mathrm{E}_t^{\text{var}}})\mathbb{P}(\overline{\mathrm{E}_t^{\text{var}}}) \\ &\leq \mathbb{P}(\overset{*}{\phi}{}^\top \tilde{\theta}_t \geq \phi(x_t, a)^\top \tilde{\theta}_t, \forall a \in \mathcal{S}_t|\mathcal{F}_{t-1}) + \mathbb{P}(\overline{\mathrm{E}_t^{\text{var}}})\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{P}(a_t \in \mathcal{U}_t|\mathcal{F}_{t-1}) &\geq \mathbb{P}(\overset{*}{\phi}{}_t^\top \tilde{\theta}_t \geq \overset{*}{\phi}{}_t^\star \theta^\star|\mathcal{F}_{t-1}) - \mathbb{P}(\overline{\mathrm{E}_t^{\text{var}}}) \\ &\geq p - \frac{1}{t^2} \ ,\end{aligned}$$

where the last inequality is due to Lemma B.2 and Lemma B.1 with $p = 1/(2\sqrt{2\pi e})$. □

Lemma A.5. ***(Upper bound of $\sum_{t=1}^T \|\phi_t\|_{\hat{\Sigma}_t}$)*** *The following lemma we will be useful in the derivation of the regret bound later in the proof.*

$$\sum_{t=1}^T \|\phi_t\|_{V_t^{-1}} \leq \sqrt{2dT \log\left(1 + \frac{T}{\lambda d}\right)}$$

Proof. Recall the relation between the 1-norm and 2-norm for a d-dimensional vector, ie, $\|\cdot\|_1 \leq \sqrt{d}\|\cdot\|_2$. Hence, it follows that

$$\sum_{t=1}^T \|\phi_t\|_{V_t^{-1}} \leq \sqrt{T\sum_{t=1}^T \|\phi_t\|^2_{V_t^{-1}}}$$

Firs, recall the definition of $V_t = \lambda \mathrm{I} \sum_{s=1}^{t-1} \phi_s \phi_s^\top$ in (15). Therefore, we apply Lemma 11 and Lemma 10 of (Abbasi-Yadkori et al., 2011b), then we have

$$\begin{aligned}\sum_{t=1}^{T} \|\phi_t\|^2_{V_t^{-1}} &\leq 2 \log \frac{\det V_t}{\det \lambda \mathrm{I}} \\ &\leq 2 \log \frac{(\lambda + T/d)^d}{\lambda^d} \\ &= 2d \log \left(1 + \frac{T}{\lambda d}\right) .\end{aligned}$$

Consequently,

$$\sum_{t=1}^{T} \|\phi_t\|_{V_t^{-1}} \leq \sqrt{2dT \log \left(1 + \frac{T}{\lambda d}\right)}$$

□

A.5. Technical Lemmas

A.5.1. Upper bound of variational mean concentration term

In this section the objective is to bound the mean variational concentration term, ie, $|\phi^\top (\tilde{m}_{t,k} - \hat{\mu})|$.

Lemma A.6. *Given E_t^{true} defined in section (A.2), the expected mean of the variational posterior at time step t after K_t steps of gradient descent $\tilde{m}_{t,K_t}$, defined in section (A.1), is equal to:*

$$\tilde{m}_{t,K_t} = \sum_{j=1}^{t-1} \prod_{i=j}^{t-1} A_i^{K_i - 1} (\hat{\mu}_j - \hat{\mu}_{j+1}) + \hat{\mu}_t \tag{29}$$

where $A_i = \mathrm{I} - \eta h_i V_i$.

Proof. Recall the definitions of $\tilde{\mu}_{t,k}$ and $\tilde{m}_{t,k}$ in section A.1. Moreover, this section also presents the sequence $\{\tilde{\mu}_{t,k}\}_{k=1}^{K_t}$ defined recursively in Algorithm 1 by:

$$\tilde{\mu}_{t,k+1} = \tilde{\mu}_{t,k} - h_t \eta V_t (\tilde{\theta}_{t,k} - \hat{\mu}_t) .$$

Note that $\{\tilde{\mu}_{t,k}\}_{k=1}^{K_t}$ is a sequence of Gaussian samples with mean and covariance matrix $\tilde{m}_{t,k}$ and $\tilde{W}_{t,k+1}$ respectively (see (A.1)). Then, we have,

$$\begin{aligned}\tilde{m}_{t,k+1} &= \mathbb{E}[\tilde{\mu}_{t,k+1}] \\ &= \tilde{m}_{t,k} - \eta h_t V_t (\tilde{m}_{t,k} - \hat{\mu}_t) \\ &= (\mathrm{I} - h_t \eta V_t) \tilde{m}_{t,k} + \eta h_t V_t \hat{\mu}_t\end{aligned}$$

Now, we recognise an arithmetico-geometric sequence, therefore the solution is:

$$\tilde{m}_{t,k} = (\mathrm{I} - h_t \eta V_t)^{k-1} (\tilde{m}_{t,1} - \hat{\mu}_t) + \hat{\mu}_t$$

Moreover, in the algorithm we use that $\tilde{\mu}_{t,1} = \tilde{\mu}_{t-1,K_{t-1}}$, which implies that $\tilde{m}_{t,1} = \tilde{m}_{t-1,k_{t-1}}$ and $W_{t,1} = W_{t-1,K_{t-1}}$. Hence, we have

$$\tilde{m}_{t,K_t} = \prod_{i=1}^{t} (\mathrm{I} - \eta h_i V_i)^{K_i - 1} (\tilde{m}_{1,1} - \hat{\mu}_1) + \sum_{j=1}^{t-1} \prod_{i=j+1}^{t} (\mathrm{I} - \eta h_i V_i)^{K_i - 1} (\hat{\mu}_j - \hat{\mu}_{j+1}) + \hat{\mu}_t \tag{30}$$

Moreover, the mean of the variational posterior is initialized at $\tilde{\mu}_{1,1} = 0_d$, then the expected mean of the variational posterior $\tilde{m}_1 = \hat{\mu}_1 = 0_d$. Therefore the first term of (30) is null. □

Lemma A.7. *Given E_t^{true}, for any $\phi \in \mathbb{R}^d$, it holds that*

$$|\phi(\tilde{m}_{t,K_t} - \hat{\mu}_t)| \leq \sum_{j=1}^{t-1} \prod_{i=j}^{t-1} (1 - \eta h \lambda_{min}(V_i))^{K_i - 1} \|\phi\|_{V_j^{-1}} \|\phi_j\|_{V_j^{-1}} \left(g_1(t)/\sqrt{\lambda} + R\sqrt{1 + \log(3t^2)} \right)$$

where $\tilde{m}_{t,K_t}$ is the expected mean of the variational posterior at time step t after K_t steps of gradient descent, ie, $\tilde{m}_{t,K_t} = \mathbb{E}[\tilde{\mu}_{t,K_t}]$, (see section A.1). Recall that $g_1(t) = R\sqrt{d \log(3t^3)} + \sqrt{\lambda}$ (see section: A.2).

Proof. Lemma A.6 gives us that $\tilde{m}_{t,K_t} = \sum_{j=1}^{t-1} \prod_{i=j}^{t-1} A_i^{K_i - 1}(\hat{\mu}_j - \hat{\mu}_{j+1}) + \hat{\mu}_t$ where $A_i = \mathrm{I} - \eta h_i V_i$. Then, for any $\phi \in \mathbb{R}^d$, the term we want to upper bound is:

$$|\phi^\top(\tilde{m}_{t,K_t} - \hat{\mu}_t)| \leq \sum_{j=1}^{t-1} |\phi^\top \prod_{i=j}^{t-1} A_i^{K_i - 1}(\hat{\mu}_j - \hat{\mu}_{j+1})| \,, \tag{31}$$

We can notice that the previous term only depends on the difference between the mean posterior at time j and the one at time j+1, which can be upper bounded. Recall the different relations between V_j, b_j, r_j, ϕ_j and $\hat{\Sigma}_j$ in the linear bandit setting (see equation (15)): $V_{j+1} = V_j + \phi_j \phi_j^\top$, $b_{j+1} = b_j + r_j \phi_j$ and $\hat{\mu}_j = V_j^{-1} b_j$, then by Sherman–Morrison formula we have:

$$V_{j+1}^{-1} = (V_j + \phi_j \phi_j^\top)^{-1} = V_j^{-1} - \frac{V_j^{-1} \phi_j \phi_j^\top V_j^{-1}}{1 + \phi_j^\top V_j^{-1} \phi_j} \tag{32}$$

The difference between the mean posterior at time $j+1$ and the one at time j becomes:

$$\begin{aligned}
\hat{\mu}_{j+1} - \hat{\mu}_j &= V_{j+1}^{-1} b_{j+1} - V_j^{-1} b_j \\
&= (V_j^{-1} - \frac{V_j^{-1} \phi_j \phi_j^\top V_j^{-1}}{1 + \phi_j^\top V_j^{-1} \phi_j})(b_j + r_j \phi_j) - V_j^{-1} b_j \\
&= r_j V_j^{-1} \phi_j - \frac{V_j^{-1} \phi_j \phi_j^\top V_j^{-1}}{1 + \phi_j^\top V_j^{-1} \phi_j}(b_j + r_j \phi_j) \\
&= \frac{V_j^{-1} \phi_j}{1 + \phi_j^\top V_j^{-1} \phi_j} \left\{ -\phi_j^\top \hat{\mu}_j - r_j \phi_j^\top V_j^{-1} \phi_j + r_j (1 + \phi_j^\top V_j^{-1} \phi_j^\top) \right\} \\
&= \frac{V_j^{-1} \phi_j (r_j - \phi_j^\top \hat{\mu}_j)}{1 + \phi_j^\top V_j^{-1} \phi_j} \\
&\overset{(a)}{=} \frac{V_j^{-1} \phi_j (\phi_j^\top \theta^\star + \xi_j - \phi_j^\top \hat{\mu}_j)}{1 + \phi_j^\top V_j^{-1} \phi_j} \\
&\overset{(b)}{\leq} V_j^{-1} \phi_j (\phi_j^\top \theta^\star + \xi_j - \phi_j^\top \hat{\mu}_j)
\end{aligned} \tag{33}$$

where in (a) we have used that $r_j = \phi_j^\top \theta^\star + \xi_j$ with ξ_j is sampled from a R-Subgaussian distribution. Inequality (b) is due to $\phi_j^\top V_j^{-1} \phi_i = \|\phi_j\|^2_{V_j^{-1}} > 0$.

Subsequently, combining equations (31) and (33), we obtain the following upper bound

$$
\begin{aligned}
|\phi(\tilde{m}_{t,K_t} - \hat{\mu}_t)| &\leq \sum_{j=1}^{t-1}\prod_{i=j}^{t-1} |\phi^\top A_i^{K_i-1}(\hat{\mu}_j - \hat{\mu}_{j+1})| \\
&\overset{(a)}{\leq} \sum_{j=1}^{t-1}\prod_{i=j}^{t-1} \left|\phi^\top A_i^{K_i-1} V_j^{-1}\phi_j(\phi_j^\top\theta^\star + \xi_j - \phi_j^\top\hat{\mu}_j)\right| \\
&\overset{(b)}{\leq} \sum_{j=1}^{t-1}\prod_{i=j}^{t-1} (1-\eta h\lambda_{\min}(V_i))^{K_i-1}\phi^\top V_j^{-1/2}V_j^{-1/2}\phi_j|(\phi_j^\top\theta^\star + \xi_j - \phi_j^\top\hat{\mu}_j)| \\
&\overset{(c)}{\leq} \sum_{j=1}^{t-1}\prod_{i=j}^{t-1} (1-\eta h\lambda_{\min}(V_i))^{K_i-1}\|\phi\|_{V_j^{-1}}\|\phi_j\|_{V_j^{-1}}|(\phi_j^\top\theta^\star + \xi_j - \phi_j^\top\hat{\mu}_j)| \\
&\overset{(d)}{\leq} \sum_{j=1}^{t-1}\prod_{i=j}^{t-1} (1-\eta h\lambda_{\min}(V_i))^{K_i-1}\|\phi\|_{V_j^{-1}}\|\phi_j\|_{V_j^{-1}}\left(g_1(t)/\sqrt{\lambda} + R\sqrt{1+\log(3t^2)}\right)
\end{aligned}
$$

In the inequality (a) we have used equation (33), in (b) the relation $A_i^{K_i-1} = (\mathrm{I}-\eta h_i V_i)^{K_i-1} \preceq (1-\eta h_i\lambda_{\min}(V_i))^{K_i-1}\,\mathrm{I}_d$, in (c) the definition of $\|\phi\|_{V_t^{-1}} = \sqrt{\phi^\top V_t^{-1}\phi} = \sqrt{\phi^\top V_t^{-1/2}}\sqrt{V_t^{-1/2}\phi} = \phi^\top V_t^{-1/2}$, and finally (d) is due to $|\xi_i| < R\sqrt{1+\log 3t^2}$ as $\mathrm{E}_t^{\text{true}}$ holds and $|\phi_i^\top(\theta^\star - \hat{\mu}_t)| \leq g_1(t)\|\phi_i^\top\|_{V_t^{-1}} \leq g_1(t)/\sqrt{\lambda}$ □

Lemma A.8. *Given E_t^{true}, for any $\phi \in \mathbb{R}^d$, if the number of gradient descent of Algorithm 2 is such that for $t \geq 2$*

$$
K_t \geq 1 + 2(1+2\kappa_t^2)\log\left(4R\sqrt{dT\log(3T^3)}\right),
$$

then it holds that

$$
|\phi(\tilde{m}_{t,K_t} - \hat{\mu}_t)| \leq \frac{2\|\phi\|_{V_t^{-1}}}{\lambda}
$$

This lemma provides the upper bound for variational mean concentration term.

Proof. Firstly, we can apply Lemma A.7, it gives us

$$
|\phi(\tilde{m}_{t,K_t} - \hat{\mu}_t)| \leq \sum_{j=1}^{t-1}\prod_{i=j}^{t-1} (1-\eta h_t\lambda_{\min}(V_t))^{K_i-1}\|\phi\|_{V_j^{-1}}\|\phi_j\|_{V_j^{-1}}\left(g_1(t)/\sqrt{\lambda} + R\sqrt{1+\log(3t^2)}\right)
$$

where $g_1(t) = R\sqrt{d\log(3t^3)} + \sqrt{\lambda}$. Moreover, for $t \geq 2$,

$$
\begin{aligned}
R\sqrt{1+\log 3t^2} + g_1(t)/\sqrt{\lambda} &\leq R\sqrt{\log 3t^2} + R\sqrt{d\log(3t^3)/\lambda} + R + 1 && (34)\\
&\leq 4R\sqrt{d\log 3t^2/\lambda} && (35)
\end{aligned}
$$

where we have used that $R \geq 1$ and $\lambda \leq 1$. Moreover, for any $j \in [1,t]$ we have

$$
\begin{aligned}
\|\phi\|_{V_j^{-1}} &\leq \|\phi\|_2/\sqrt{\lambda} && (36)\\
&\leq \lambda_{\max}(V_t)^{1/2}\|\phi\|_{V_t^{-1}}/\sqrt{\lambda} && (37)\\
&= \lambda_{\max}(V_t)^{1/2}\|\phi\|_{V_t^{-1}}/\lambda^{1/2} && (38)
\end{aligned}
$$

Let's define $\epsilon = \left(4R\sqrt{dt\log(3t^2)}\right)^{-1} \leq 1/2$ and let's take K_i such that $(1-h_t\lambda_{\min}(V_t))^{K_i-1} \leq \epsilon$, this condition will be explained later in the proof. It follows that

$$
\begin{aligned}
|\phi(\tilde{m}_{t,K_t} - \hat{\mu}_t)| &\leq \sum_{j=1}^{t-1} \prod_{i=j}^{t-1} (1 - \eta h_t \lambda_{\min}(V_t))^{K_i - 1} \|\phi\|_{V_j^{-1}} \|\phi_j\|_{V_j^{-1}} \epsilon^{-1} \\
&\overset{(a)}{\leq} \frac{\|\phi\|_{V_t^{-1}} \lambda_{\max}(V_t)^{1/2}}{\lambda} \sum_{j=1}^{t-1} \prod_{i=j}^{t-1} (1 - h_t \lambda_{\min}(V_T))^{K_i - 1} \epsilon^{-1} \\
&\overset{(b)}{\leq} \frac{\|\phi\|_{V_t^{-1}}}{\lambda} \sum_{j=1}^{t-1} \epsilon^{t-j-1} \\
&\overset{(c)}{\leq} \frac{\|\phi\|_{V_t^{-1}}}{\lambda} \times \frac{1}{1-\epsilon} \\
&\overset{(d)}{\leq} \frac{2\|\phi\|_{V_t^{-1}}}{\lambda} \ ,
\end{aligned}
$$

where (a) comes from equations (36) and (35). The point (b) comes that $\lambda_{\max}(V_t) \leq \sqrt{t}$ because $\lambda \leq 1$ and definition of ϵ, then (c) from the geometric series formula. Finally, in (d), we have used $\epsilon \leq 1/2$.

Now, let's focus on that condition on K_i presented previously. For any $i \in [t]$, recall the definition of the step size h_t in A.1.

$$
h_i = \frac{\lambda_{\min}(V_i)}{2\eta(\lambda_{\min}(V_i)^2 + 2\lambda_{\max}(V_i)^2)} \ ,
$$

and define $\kappa_i = \lambda_{\max}(V_i)/\lambda_{\min}(V_i)$. Therefore, it holds that

$$
(1 - \eta h_i \lambda_{\min}(V_i))^{K_i - 1} = \left(1 - \frac{1}{2(1 + 2\kappa_i^2)}\right)^{K_i - 1}
$$

For any $\epsilon > 0$, we want that $(1 - h_i \lambda_{\min}(V_i))^{K_i - 1} \leq \epsilon$. Hence we deduce that

$$
K_i \geq 1 + \frac{\log(1/\epsilon)}{\log(1 - 1/(2(1 + 2\kappa_i^2)))} \ .
$$

Moreover, if $0 < x < 1$ then we have $-x > \log(1 - x)$, if follows that

$$
K_i \geq 1 + 2(1 + 2\kappa_i^2) \log(1/\epsilon) \ .
$$

We note that,

$$
\begin{aligned}
\log(1/\epsilon) &= \log\left(4R\sqrt{dt \log 3t^3}\right) \\
&\leq \log\left(4R\sqrt{dT \log 3T^3}\right) \ .
\end{aligned}
$$

Finally, taking $K_i \geq 1 + 2(1 + 2\kappa_i^2) \log\left(4R\sqrt{dT \log(3T^3)}\right)$, we obtain the condition

$$
(1 - \eta h_i \lambda_{\min}(V_t))^{K_i - 1} \leq \epsilon \ ,
$$

which concludes the proof. □

A.5.2. Control the variational covariance matrix

The objective of this section is to control the following term: $|\boldsymbol{\phi}^\top \left(\tilde{\boldsymbol{\theta}}_{\boldsymbol{t,k}} - \tilde{\boldsymbol{\mu}}_{\boldsymbol{t,k}}\right)|$. As $\tilde{\theta}_{t,k}$ is a sample from a Gaussian distribution with mean $\tilde{\mu}_{t,k}$, the previous term will be controlled using Gaussian concentration and an upper bound of the norm of the variational covariance matrix $\tilde{\Sigma}_{t,k}$. Recall the definitions of parameters $\hat{\Sigma}_t$, $B_{t,k}$, $\tilde{\theta}_{t,k}$ and $\tilde{\mu}_{t,k}$ in section A.1.

Lemma A.9. *For any $t \in [T]$ and $k \in [K_t]$, the following relation holds:*

$$(\mathrm{H}): \ \tilde{\Sigma}_{t,k} \succeq V_t^{-1}/(2\eta) \tag{39}$$

Proof. The sequence $\{\tilde{\Sigma}_{t,n}\}_{t\in[T],\, n\in[k_t]}$ is initialized by $\tilde{\Sigma}_{1,1} = \mathrm{I}/(\lambda\eta) = V_1^{-1}/\eta \succeq V_t^{-1}/(2\eta)$. Hence, (H) holds for the pair $t = 1$ and $k = 1$. Therefore, to conclude the proof, we have to show that the following transitions are true:

- for any $t \in [T]$, if (H) holds at step (t, K_t) then it stays true at step $(t + 1, 1)$ (`recursion in t`),
- for any $k \in [K_t]$, if (H) holds at step (t, k) then it stays true at step $(t, k + 1)$ (`recursion in k`).

Firstly, let's focus on the first implication and suppose that (H) holds at step (t, K_t). Therefore we have

$$\tilde{\Sigma}_{t+1,1} \overset{(a)}{=} \tilde{\Sigma}_{t,K_t} \overset{(b)}{\succeq} V_t^{-1}/(2\eta) \overset{(c)}{\succeq} V_{t+1}^{-1}/(2\eta)$$

where (a) comes from the initialization of the sequence $\{\tilde{\Sigma}_{t,k}\}_{k\in[k_t]}$, (b) from the hypothesis (H) at step (t, K_t). And finally, (c) is due to $V_{t+1} = V_t + \phi_t\phi_t^\top \succeq V_t$. Then we can conclude that (H) holds at step $(t + 1, 1)$.

Now we focus on the second implication and we suppose that (H) holds at step (t, k). For ease of notation we denote by $Z_{t,k} := \tilde{\Sigma}_{t,k} - V_t^{-1}/(2\eta)$. Therefore using the recursive definition of $\tilde{\Sigma}_{t,k}$ given in section (A.1), we have

$$\begin{aligned}
Z_{t,k+1} &= A_t\tilde{\Sigma}_{t,k}A_t + 2h_tA_t + h_t^2\tilde{\Sigma}_{t,k}^{-1} - V_t^{-1}/(2\eta) \\
&= A_tZ_{t,k}A_t + 2h_tA_t + h_t^2\tilde{\Sigma}_{t,k}^{-1} - h_t\mathrm{I} + \eta h_t^2V_t/2 \\
&= A_tZ_{t,k}A_t + h_t\mathrm{I} - 3h_t^2\eta V_t/4 + h_t^2\tilde{\Sigma}_{t,k}^{-1}
\end{aligned}$$

where in the last inequalities we have used that $A_t = (\mathrm{I} - \eta h_t^2V_t)$. Moreover, all terms in the previous inequality are positive semi-definite. Indeed, as (H) holds at step (t, k), we know that $Z_{t,k} \succeq 0$ and then that $A_tZ_{t,k}A_t \succeq 0$. Moreover, $\tilde{\Sigma}_{t,k} \succeq V_t^{-1}/(2\eta) \succeq 0$, so $\tilde{\Sigma}_{t,k}^{-1} \succeq 0$. Finally, recall the definition of h_t in section (A.1)

$$\begin{aligned}
h_t &\leq \frac{\lambda_{\min}(V_t)}{2\eta(\lambda_{\min}(V_t)^2 + 2\lambda_{\max}(V_t)^2)} \\
&= \frac{1/\kappa_t}{2\eta\lambda_{\max}(V_t)((1/\kappa_t)^2 + 1)} \\
&= \frac{4}{3\eta\lambda_{\max}(V_t)} \times \frac{3/\kappa_t}{8(1 + (1/\kappa_t^2))} \\
&\leq \frac{4}{3\eta\lambda_{\max}(V_t)} \ ,
\end{aligned}$$

where $\kappa_t = \lambda_{\max}(V_t)/\lambda_{\min}(V_t) \geq 1$. Consequently, the matrix $\mathrm{I} - 3\eta h_tV_t/4$ is also positive semi-definite. Subsequently, we have

$$Z_{t,k+1} \succeq 0 \ .$$

□

Lemma A.10. *For any $\phi \in \mathbb{R}^d$, let $B_{t,k}$ the square root of the covariance matrix defined in Algorithm (2). It holds that*

$$\begin{aligned}
\|B_{t,K_t}\phi\|_2 &\leq 1/\sqrt{\eta}\Big(1 + \sqrt{\|V_t\|_2C_t}\Big)\|\phi\|_{V_t^{-1}} \\
\|B_{t,K_t}\phi\|_2 &\geq 1/\sqrt{\eta}\Big(1 - \sqrt{\|V_t\|_2C_t}\Big)\|\phi\|_{V_t^{-1}}
\end{aligned}$$

where $C_t = 1/\lambda \sum_{j=1}^{t-1}\prod_{i=j+1}^{t}\left(1 - \frac{3h_t\eta}{2}\lambda_{min}(V_i)\right)^{K_i-1}$.

Proof. Recall the recursive relation of $\Lambda_{t,k}$ defined in Section (A.1).

$$\Lambda_{t,k+1} = A_t \Lambda_{t,k} A_t - \eta h_t^2 V_t \Lambda_{t,k} \tilde{\Sigma}_{t,k}^{-1} ,$$

Hence, we have the following relation on the norm of $\Lambda_{t,k+1}$:

$$\begin{aligned}
\|\Lambda_{t,k+1}\|_2 &\leq \|A_t\|_2 \|\Lambda_{t,k}\|_2 \|A_t\|_2 + \eta h_t^2 \|V_t\|_2 \|\Lambda_{t,k}\|_2 \|\tilde{\Sigma}_{t,k}^{-1}\|_2 \\
&= \left(\lambda_{\max}(A_t)^2 + \eta h_t^2 \lambda_{\max}(V_t) \lambda_{\max}(\tilde{\Sigma}_{t,k}^{-1})\right) \|\Lambda_{t,k}\|_2 \\
&\overset{(a)}{=} \left(1 - 2\eta h_t \lambda_{\min}(V_t) + \eta^2 h_t^2 \lambda_{\min}(V_t)^2 + \eta h_t^2 \lambda_{\max}(V_t) \lambda_{\max}(\tilde{\Sigma}_{t,k}^{-1})\right) \|\Lambda_{t,k}\|_2 \\
&= \left(1 - \frac{3h_t\eta}{2} \lambda_{\min}(V_t) + \eta h_t \{h_t(\eta \lambda_{\min}(V_t)^2 + \lambda_{\max}(V_t)\lambda_{\max}(\tilde{\Sigma}_{t,k}^{-1})) - \lambda_{\min}(V_t)/2\}\right) \|\Lambda_{t,k}\|_2 \\
&\overset{(b)}{\leq} \left(1 - \frac{3h_t\eta}{2} \lambda_{\min}(V_t)\right) \|\Lambda_{t,k}\|_2 ,
\end{aligned}$$

where (a) uses that $\lambda_{\max}(A_t) = 1 - \eta h \lambda_{\min}(V_t)$. Finally, inequality (b) is due to: $\tilde{\Sigma}_{t,k} \succeq V_t^{-1}/(2\eta)$ (Lemma A.9). Indeed it implies that

$$\begin{aligned}
h_t(\eta \lambda_{\min}(V_t)^2 + \lambda_{\max}(V_t)\lambda_{\max}(\tilde{\Sigma}_{t,k}^{-1})) &\leq h_t(\eta \lambda_{\min}(V_t)^2 + 2\eta \lambda_{\max}(V_t)^2) \\
&\leq \lambda_{\min}(V_t)/2 ,
\end{aligned}$$

where the inequality comes from the definition of the step size: $h_t \leq \lambda_{\min}(V_t)/\Big(2\eta(\lambda_{\min}(V_t)^2 + 2\lambda_{\max}(V_t)^2)\Big)$. Subsequently,

$$\begin{aligned}
\|\Lambda_{t,K_t}\|_2 &\leq \left(1 - \frac{3h_t\eta}{2} \lambda_{\min}(V_t)\right) \|\Lambda_{t,K_t-1}\|_2 \\
&\leq \left(1 - \frac{3h_t\eta}{2} \lambda_{\min}(V_t)\right)^{K_t-1} \|\tilde{\Sigma}_{t,1} - 1/\eta V_t^{-1}\|_2 \\
&= \left(1 - \frac{3h_t\eta}{2} \lambda_{\min}(V_t)\right)^{K_t-1} \|\tilde{\Sigma}_{t-1,k_{t-1}} - 1/\eta V_t^{-1}\|_2 \\
&\leq \prod_{i=1}^{t-1} \left(1 - \frac{3h_t\eta}{2} \lambda_{\min}(V_i)\right)^{K_i-1} \|\tilde{\Sigma}_{1,1} - 1/\eta V_1^{-1}\|_2 \\
&+ \sum_{j=1}^{t-1} \prod_{i=j+1}^{t} \left(1 - \frac{3h_t\eta}{2} \lambda_{\min}(V_i)\right)^{K_i-1} \|1/\eta V_j^{-1} - 1/\eta V_{j+1}^{-1}\|_2 \\
&\overset{(a)}{\leq} \frac{1}{\eta} \sum_{j=1}^{t-1} \prod_{i=j+1}^{t} \left(1 - \frac{3h_t\eta}{2} \lambda_{\min}(V_i)\right)^{K_i-1} \|V_j^{-1} - V_{j+1}^{-1}\|_2 \\
&\overset{(b)}{\leq} \frac{1}{\lambda\eta} \sum_{j=1}^{t-1} \prod_{i=j+1}^{t} \left(1 - \frac{3h_t\eta}{2} \lambda_{\min}(V_i)\right)^{K_i-1} , \\
&:= C_t/\eta
\end{aligned} \tag{40}$$

where in (a) we have used that $\tilde{\Sigma}_{1,1} = \frac{1}{\lambda\eta}\mathrm{I} = 1/\eta V_1^{-1}$. Moreover $\|V_j^{-1} - V_{j+1}^{-1}\|_2 = \|\big(V_j^{-1}\phi_j\phi_j^\top V_j^{-1}\big)/\big(1 - \phi_j^\top V_j^{-1}\phi_j\big)\|_2$ see result (32). It implies that $\|V_j^{-1} - V_{j+1}^{-1}\|_2 \leq \|V_j^{-1}\|_2^2 \leq \|V_1^{-1}\|_2^2 = 1/\lambda$.

Finally, for any $\phi \in \mathbb{R}^d$,

$$\begin{aligned}
\|B_{t,K_t}\phi\|_2 &= \sqrt{\phi^\top B_{t,K_t}^\top B_{t,K_t}\phi} \\
&= \sqrt{\phi^\top \tilde{\Sigma}_{t,K_t}\phi} \\
&= \sqrt{\phi^\top(\tilde{\Sigma}_{t,K_t} - 1/\eta V_t^{-1})\phi + 1/\eta\phi^\top V_t^{-1}\phi} \\
&\leq \|\phi\|_2\sqrt{\|\tilde{\Sigma}_{t,K_t} - 1/\eta V_t^{-1}\|_2} + 1/\sqrt{\eta}\|\phi\|_{V_t^{-1}}
\end{aligned}$$

where the last inequality comes from the fact that for $a, b > 0$, $\sqrt{a+b} < \sqrt{a} + \sqrt{b}$, Moreover,

$$\begin{aligned}
\|\phi\|_2 &= \|\phi V_t^{-1/2} V_t^{1/2}\|_2 \\
&\leq \|\phi\|_{V_t^{-1}}\|V_t^{1/2}\|_2
\end{aligned}$$

Consequently, we have

$$\|B_{t,K_t}\phi\|_2 \leq 1/\sqrt{\eta}\Big(1 + \sqrt{\|V_t\|_2 C_t}\Big)\|\phi\|_{V_t^{-1}}$$

The lower bound of this lemma

$$\|B_{t,K_t}\phi\|_2 \geq 1/\sqrt{\eta}\Big(1 - \sqrt{\|V_t\|_2 C_t}\Big)\|\phi\|_{V_t^{-1}}$$

is obtained because

$$\begin{aligned}
\|B_{t,K_t}\phi\|_2 &= \sqrt{\phi^\top B_{t,K_t}^\top B_{t,K_t}\phi} \\
&= \sqrt{\phi^\top \tilde{\Sigma}_{t,K_t}\phi} \\
&= \sqrt{\phi^\top(\tilde{\Sigma}_{t,K_t} - 1/\eta V_t^{-1})\phi + 1/\eta\phi^\top V_t^{-1}\phi} \\
&\geq -\|\phi\|_2\sqrt{\|\tilde{\Sigma}_{t,K_t} - 1/\eta V_t^{-1}\|_2} + 1/\sqrt{\eta}\|\phi\|_{V_t^{-1}} \quad \leq \|B_{t,K_t}\phi\|_2 \geq 1/\sqrt{\eta}\Big(1 - \sqrt{\|V_t\|_2 C_t}\Big)\|\phi\|_{V_t^{-1}},
\end{aligned}$$

where the first inequality comes from remarkable identity $\sqrt{a} - \sqrt{b} < \sqrt{a+b}$ for $a, b > 0$.

□

A.6. Control of covariance matrix in VITS − II

The analysis of VITS − II is similar to VITS − I in the only part that changes is the equivalent of the previous lemma. The other lemmas are still valid for the rest of the analysis.

First, lemma A.11 is the equivalent of A.9 for VITS − I.

Lemma A.11. *For any $t \in [T]$ and $k \in [K_t]$, the following relation holds:*

$$(\mathrm{H}): \; C_{t,k}^2 \succeq \eta V_t \tag{41}$$

Proof. By recurrence,

$$\begin{aligned}
C_{t,k+1}^2 - \eta V_t &= (\mathrm{I} - h_t(C_{t,k}^\top C_{t,k} - \eta V_t))C_{t,k}^2(\mathrm{I} - h_t(C_{t,k}^\top C_{t,k} - \eta V_t)) - \eta V_t \\
&= (C_{t,k}^2 - \eta V_t) - h_t C_{t,k}^2(C_{t,k}^\top C_{t,k} - \eta V_t)) - h_t(C_{t,k}^\top C_{t,k} - \eta V_t))C_{t,k}^2 + h_t^2 C_{t,k}^2(C_{t,k}^\top C_{t,k} - \eta V_t)) \\
&= (C_{t,k}^2 - \eta V_t)(Id - h_t C_{t,k}^2)^2.
\end{aligned}$$

By hypothesis , $(C_{t,k}^2 - \eta V_t)$ is positive definite and the last term $(\mathrm{I} - h_t C_{t,k}^2)^2$ is also positive definite so by recurrence the property is true.

□

Finally lemma A.12 is the equivalent of lemma A.10 for $\texttt{VITS}-\texttt{I}$.

Lemma A.12. *For any $\phi \in \mathbb{R}^d$, let $B_{t,k}$ the square root of the covariance matrix defined in Algorithm $\texttt{VITS}-\texttt{II}$. It holds that*

$$\|B_{t,K_t}\phi\|_2 \leq 1/\sqrt{\eta}\Big(1 + \sqrt{\|V_t\|_2 C_t}\Big)\|\phi\|_{V_t^{-1}}$$
$$\|B_{t,K_t}\phi\|_2 \geq 1/\sqrt{\eta}\Big(1 - \sqrt{\|V_t\|_2 C_t}\Big)\|\phi\|_{V_t^{-1}}$$

where $C_t = 1/\lambda \sum_{j=1}^{t-1} \prod_{i=j+1}^{t} \left(1 - \frac{3h_t\eta}{2}\lambda_{min}(V_i)\right)^{K_i - 1}$.

Once this lemma is proved, $\|B_{t,K_t}\phi\|_2$ is upper and lower bounded by the same quantity for $\texttt{VITS}-\texttt{I}$ and $\texttt{VITS}-\texttt{II}$ and the rest of the proof is then similar for both algorithms.

Proof. Let's define $\Lambda_{t,k} = B_{t,k}^2 - 1/\eta V_t^{-1}$. Recall that where $B_{t,k}$ is the square root inverse with approximation and is defined by:

$$B_{t,k+1} = (\mathrm{I} - h_t A_{t,k})B_{t,k} + h_t C_{t,k}^\top$$

Let's use a Taylor expansion to obtain the relation between $C_{t,k}$ and $B_{t,k}^{-1}$:

$$C_{t,k} = B_{t,k}^{-1} + o(h_t D_{t,k})$$

where

$$\begin{aligned} D_{t,k} &= ((C_{t,k}^\top)^{-1} C_{t,k}^{-1} - \eta V_t) \\ &= -\eta V_t \Lambda_{t,k} C_{t,k}^2. \end{aligned}$$

Using the recurrent definition of sequences :

$$\begin{aligned}
\Lambda_{t,k+1} &= A_t B_{t,k}^2 A_t + 2h_t A_t(B_{t,k}C_{t,k} + C_{t,k}^T B_{t,k}^T) + h_t^2 C_{t,k}^2 - 1/\eta V_t^{-1} \\
&= A_t \Lambda_{t,k} A_t + 2h_t A_t(B_{t,k}C_{t,k} + C_{t,k}^T B_{t,k}^T) - 2h_t \mathrm{I} + \eta h_t^2 V_t + h_t^2 C_{t,k}^2 \\
&= A_t \Lambda_{t,k} A_t - h_t(2\mathrm{I} - (B_{t,k}C_{t,k} + C_{t,k}^T B_{t,k}^T)) + h_t^2(C_{t,k}^2 + V_t\eta(\mathrm{I} - (B_{t,k}C_{t,k} + C_{t,k}^T B_{t,k}^T)) \\
&= A_t \Lambda_{t,k} A_t - h_t(2\mathrm{I} - (B_{t,k}C_{t,k} + C_{t,k}^T B_{t,k}^T)) + h_t^2(C_{t,k}^2 + V_t\eta(\mathrm{I} - (B_{t,k}C_{t,k} + C_{t,k}^T B_{t,k}^T)) \\
&\overset{(a)}{=} A_t \Lambda_{t,k} A_t - h_t o(h_t D_{t,k}) + h_t^2(C_{t,k}^2 - V_t\eta(1 + o(h_t^2 D_{t,k}))) \\
&= A_t \Lambda_{t,k} A_t - h_t^2 \eta V_t \Lambda_{t,k} C_{t,k}^2 + o(h_t^2 D_{t,k}).
\end{aligned}$$

Where in (a) we have used that $B_{t,k}C_{t,k} + C_{t,k}^T B_{t,k}^T = 2\mathrm{I} + o(h_t D_{t,k})$. Finally, we obtain the following relation on the norm of $\Lambda_{t,k+1}$:

$$\begin{aligned}
\|\Lambda_{t,k+1}\|_2 &\leq \|A_t\|_2 \|\Lambda_{t,k}\|_2 \|A_t\|_2 + \eta h_t^2 \|V_t\|_2 \|\Lambda_{t,k}\|_2 \|C_{t,k}^2\|_2 + o(h_t^2 \eta \|\Lambda_{t,k}\|_2 \|V_t\|_2 \|C_{t,k}^2\|_2) \\
&= \left(\lambda_{\max}(A_t)^2 + \eta h_t^2 \lambda_{\max}(V_t)\lambda_{\max}(C_{t,k}^2) + o(h_t^2 \eta \|V_t\|_2 \|C_{t,k}^2\|_2)\right)\|\Lambda_{t,k}\|_2 \\
&\overset{(a)}{=} \left(1 - 2\eta h_t \lambda_{\min}(V_t) + \eta^2 h_t^2 \lambda_{\min}(V_t)^2 + \eta h_t^2 \lambda_{\max}(V_t)\lambda_{\max}(C_{t,k}^2) + o(h_t^2 \eta \|V_t\|_2 \|C_{t,k}^2\|_2)\right)\|\Lambda_{t,k}\|_2 \\
&= \left((1 - \frac{3h_t\eta}{2}\lambda_{\min}(V_t) + \eta h_t\{h_t(\eta\lambda_{\min}(V_t)^2 + \lambda_{\max}(V_t)\lambda_{\max}(C_{t,k}^2))(1 + o(1)) - \lambda_{\min}(V_t)/2\}\right)\|\Lambda_{t,k}\|_2 \\
&\overset{(b)}{\leq} \left(1 - \frac{3h_t\eta}{2}\lambda_{\min}(V_t) + \eta h_t\{h_t(\eta\lambda_{\min}(V_t)^2 + 2\lambda_{\max}(V_t)\lambda_{\max}(C_{t,k}^2)) - \lambda_{\min}(V_t)/2\}\right)\|\Lambda_{t,k}\|_2 \\
&\overset{(c)}{\leq} \left(1 - \frac{3h_t\eta}{2}\lambda_{\min}(V_t)\right)\|\Lambda_{t,k}\|_2 \ ,
\end{aligned}$$

where (a) uses that $\lambda_{\max}(A_t) = 1 - \eta h \lambda_{\min}(V_t)$, (b) that $1 + o(1) < 2$ for a sufficiently small h_t, and finally, inequality (c) is due to: $C_{t,k}^2 \succeq \eta V_t$ (Lemma A.11). It follows that

$$\begin{aligned} h_t(\eta\lambda_{\min}(V_t)^2 + 2\lambda_{\max}(V_t)\lambda_{\max}(C_{t,k})) &\leq h_t(\eta\lambda_{\min}(V_t)^2 + 2\eta\lambda_{\max}(V_t)^2) \\ &\leq \lambda_{\min}(V_t)/2 \ , \end{aligned}$$

where the inequality comes from the definition of the step size: $h_t \leq \lambda_{\min}(V_t)/\Big(2\eta(\lambda_{\min}(V_t)^2 + 2\lambda_{\max}(V_t)^2)\Big)$. Subsequently, we have

$$\begin{aligned} \|\Lambda_{t,K_t}\|_2 &\leq \left(1 - \frac{3h_t\eta}{2}\lambda_{\min}(V_t)\right)\|\Lambda_{t,K_t-1}\|_2 \\ &\leq \left(1 - \frac{3h_t\eta}{2}\lambda_{\min}(V_t)\right)^{K_t-1}\|\tilde{\Sigma}_{t,1} - 1/\eta V_t^{-1}\|_2 \\ &= \left(1 - \frac{3h_t\eta}{2}\lambda_{\min}(V_t)\right)^{K_t-1}\|\tilde{\Sigma}_{t-1,k_{t-1}} - 1/\eta V_t^{-1}\|_2 \\ &\leq \prod_{i=1}^{t-1}\left(1 - \frac{3h_t\eta}{2}\lambda_{\min}(V_i)\right)^{K_i-1}\|\tilde{\Sigma}_{1,1} - 1/\eta V_1^{-1}\|_2 \\ &+ \sum_{j=1}^{t-1}\prod_{i=j+1}^{t}\left(1 - \frac{3h_t\eta}{2}\lambda_{\min}(V_i)\right)^{K_i-1}\|1/\eta V_j^{-1} - 1/\eta V_{j+1}^{-1}\|_2 \\ &\overset{(a)}{\leq} \frac{1}{\eta}\sum_{j=1}^{t-1}\prod_{i=j+1}^{t}\left(1 - \frac{3h_t\eta}{2}\lambda_{\min}(V_i)\right)^{K_i-1}\|V_j^{-1} - V_{j+1}^{-1}\|_2 \\ &\overset{(b)}{\leq} \frac{1}{\lambda\eta}\sum_{j=1}^{t-1}\prod_{i=j+1}^{t}\left(1 - \frac{3h_t\eta}{2}\lambda_{\min}(V_i)\right)^{K_i-1} \ , \\ &:= C_t/\eta \end{aligned} \tag{42}$$

where in (a) we have used that $\tilde{\Sigma}_{1,1} = \frac{1}{\lambda\eta}\mathrm{I} = 1/\eta V_1^{-1}$. Moreover $\|V_j^{-1} - V_{j+1}^{-1}\|_2 = \|\big(V_j^{-1}\phi_j\phi_j^\top V_j^{-1}\big)/\big(1 - \phi_j^\top V_j^{-1}\phi_j\big)\|_2$ see result (32). It implies that $\|V_j^{-1} - V_{j+1}^{-1}\|_2 \leq \|V_j^{-1}\|_2^2 \leq \|V_1^{-1}\|_2^2 = 1/\lambda$.

Finally, for any $\phi \in \mathbb{R}^d$,

$$\begin{aligned} \|B_{t,K_t}\phi\|_2 &= \sqrt{\phi^\top B_{t,K_t}^\top B_{t,K_t}\phi} \\ &= \sqrt{\phi^\top \tilde{\Sigma}_{t,K_t}\phi} \\ &= \sqrt{\phi^\top(\tilde{\Sigma}_{t,K_t} - 1/\eta V_t^{-1})\phi + 1/\eta\phi^\top V_t^{-1}\phi} \\ &\leq \|\phi\|_2\sqrt{\|\tilde{\Sigma}_{t,K_t} - 1/\eta V_t^{-1}\|_2} + 1/\sqrt{\eta}\|\phi\|_{V_t^{-1}} \end{aligned}$$

where the last inequality comes from the fact that for $a, b > 0$, $\sqrt{a+b} < \sqrt{a} + \sqrt{b}$, Moreover,

$$\begin{aligned} \|\phi\|_2 &= \|\phi V_t^{-1/2}V_t^{1/2}\|_2 \\ &\leq \|\phi\|_{V_t^{-1}}\|V_t^{1/2}\|_2. \end{aligned}$$

Consequently, we have

$$\|B_{t,K_t}\phi\|_2 \leq 1/\sqrt{\eta}\Big(1 + \sqrt{\|V_t\|_2 C_t}\Big)\|\phi\|_{V_t^{-1}}$$

The lower bound of this lemma

$$\|B_{t,K_t}\phi\|_2 \geq 1/\sqrt{\eta}\Big(1-\sqrt{\|V_t\|_2 C_t}\Big)\|\phi\|_{V_t^{-1}}$$

is obtained because

$$\begin{aligned}
\|B_{t,K_t}\phi\|_2 &= \sqrt{\phi^\top B_{t,K_t}^\top B_{t,K_t}\phi} \\
&= \sqrt{\phi^\top \tilde{\Sigma}_{t,K_t}\phi} \\
&= \sqrt{\phi^\top(\tilde{\Sigma}_{t,K_t} - 1/\eta V_t^{-1})\phi + 1/\eta\phi^\top V_t^{-1}\phi} \\
&\geq -\|\phi\|_2\sqrt{\|\tilde{\Sigma}_{t,K_t} - 1/\eta V_t^{-1}\|_2} + 1/\sqrt{\eta}\|\phi\|_{V_t^{-1}} \\
&\geq 1/\sqrt{\eta}\Big(1-\sqrt{\|V_t\|_2 C_t}\Big)\|\phi\|_{V_t^{-1}},
\end{aligned}$$

where the first inequality comes from identity $\sqrt{a}-\sqrt{b} < \sqrt{a+b}$ for $a, b > 0$.

□

As $\|B_{t,K_t}\phi\|_2$ is upper and lower bounded by the same quantity for VITS − I and VITS − II, the rest of the proof is exactly similar for both algorithms.

Lemma A.13. *For any $t \in [T]$ and $a \in \mathcal{A}(x_t)$, if the number of gradient descent steps of Algorithm 2 is $K_t \geq 1 + 4(1+2\kappa_t^2)\log(2T)/(3\eta)$, therefore it holds that*

$$\begin{aligned}
\|\phi(x_t,a)^\top B_{t,K_t}\|_2 &\leq 1/\sqrt{\eta}\Big(1+1/\sqrt{\lambda}\Big)\|\phi\|_{V_t^{-1}} \\
\|\phi(x_t,a)^\top B_{t,K_t}\|_2 &\geq 1/\sqrt{\eta}\Big(1-1/\sqrt{\lambda}\Big)\|\phi\|_{V_t^{-1}} .
\end{aligned}$$

Proof. Firstly, Lemma A.10 or A.10 for VITS-II, gives us

$$\begin{aligned}
\|B_{t,K_t}\phi\|_2 &\leq 1/\sqrt{\eta}\Big(1+\sqrt{\|V_t\|_2 C_t}\Big)\|\phi\|_{V_t^{-1}} \\
\|B_{t,K_t}\phi\|_2 &\geq 1/\sqrt{\eta}\Big(1-\sqrt{\|V_t\|_2 C_t}\Big)\|\phi\|_{V_t^{-1}}
\end{aligned}$$

with $C_t = 1/\lambda\sum_{j=1}^{t-1}\prod_{i=j+1}^{t}(1-\frac{3h_t\eta}{2}\lambda_{\min}(V_t))^{K_i-1}$. Furthermore, for any $t \in [T]$, recall that

$$h_t = \frac{\lambda_{\min}(V_t)}{2\eta(\lambda_{\min}(V_t)^2 + 2\lambda_{\max}(V_t)^2)} ,$$

and define $\kappa_t = \lambda_{\max}(V_t)/\lambda_{\min}(V_t)$. Therefore, it holds that

$$\left(1-\frac{3\eta h_t}{2}\lambda_{\min}(V_t)\right)^{K_t-1} = \left(1-\frac{3}{4(1+2\kappa_t^2)}\right)^{K_t-1}$$

For any $\epsilon > 0$, we want that $\left(1-\frac{3h_t\eta}{2}\lambda_{\min}(V_t)\right)^{K_t-1} \leq \epsilon$. Hence we deduce the following relation for K_t:

$$K_t \geq 1 + \frac{\log(\epsilon)}{\log(1-3\eta/(4(1+2\kappa_t^2)))} .$$

Moreover, if $0 < x < 1$ then we have $-x > \log(1-x)$, then we have

$$K_t \geq 1 + 4(1+2\kappa_t^2)\log(1/\epsilon)/3 \ .$$

Subsequently, let's apply the last result to $\epsilon = 1/(2t)$. Then for $K_t \geq 1 + 4(1+2\kappa_t^2)\log(2T)/3$, we have

$$\begin{aligned}
\|V_t\|_2/\lambda \sum_{j=1}^{t-1} \prod_{i=j+1}^{t} \left(1 - \frac{3\eta h_t}{2}\lambda_{\min}(V_t)\right)^{K_i - 1} &\leq \|V_t\|_2 \epsilon/\lambda \sum_{j=1}^{t-1} \epsilon^{t-j-1} \\
&\leq \|V_t\|_2 \epsilon/\lambda \sum_{j=0}^{+\infty} \epsilon^{j} \\
&\overset{(a)}{\leq} \frac{\|V_t\|_2}{2\lambda t(1-\epsilon)} \\
&\overset{(b)}{\leq} 1/\lambda \ ,
\end{aligned}$$

where (a) and (b) come from the geometric serie because $\epsilon \leq 1/2$ and we have used that $\|V_t\|_2 = \|\lambda \mathrm{I} + \sum_{s=1}^{t-1} \phi\phi^\top\|_2 \leq \lambda + t - 1 \leq t$, as $\lambda \leq 1$. Consequently, we have

$$\begin{aligned}
\|B_{t,K_t}\phi\|_2 &\leq 1/\sqrt{\eta}\Big(1 + 1/\sqrt{\lambda}\Big)\|\phi\|_{V_t^{-1}} \\
\|B_{t,K_t}\phi\|_2 &\geq 1/\sqrt{\eta}\Big(1 - 1/\sqrt{\lambda}\Big)\|\phi\|_{V_t^{-1}}
\end{aligned}$$

□

Lemma A.14. *For any $t \in [T]$ and $a \in \mathcal{A}(x_t)$, if the number of gradient descent steps of Algorithm 2 is $K_t \geq 1 + 4(1+2\kappa_t^2)\log(2T)/3$, then with probability at least $1 - 1/t^2$, we have*

$$|\phi(x_t,a)^\top \left(\tilde{\theta}_{t,K_t} - \tilde{\mu}_{t,K_t}\right)| \leq \sqrt{4d\log(t^3)/\eta}\Big(1 + 1/\sqrt{\lambda}\Big)\|\phi(x_t,a)\|_{V_t^{-1}}$$

Proof. For any $a \in \mathcal{A}(x_t)$, if $K_t \geq 1 + 4(1+2\kappa_t^2)\log(2T)/(3\eta)$, Lemma A.13 gives us that

$$\begin{aligned}
|\phi(x_t,a)^\top \left(\tilde{\theta}_{t,K_t} - \tilde{\mu}_{t,K_t}\right)| &\leq \|B_{t,K_t}^{-1}(\tilde{\theta}_{t,K_t} - \tilde{\mu}_{t,K_t})\|_2 \|\phi(x_t,a)^\top B_{t,K_t}\|_2 \\
&\leq \|B_{t,K_t}^{-1}(\tilde{\theta}_{t,K_t} - \tilde{\mu}_{t,K_t})\|_2 (1/\sqrt{\eta})\Big(1 + 1/\sqrt{\lambda}\Big)\|\phi\|_{V_t^{-1}} \ .
\end{aligned}$$

where first inequality comes from classical matrix norm inequality and the second one is previous lemma A.13, recall that $\tilde{\theta}_{t,K_t} \sim \mathrm{N}(\tilde{\mu}_{t,K_t}, B_{t,K_t}B_{t,K_t}^\top)$, hence $B_{t,K_t}^{-1}(\tilde{\theta}_{t,K_t} - \tilde{\mu}_{t,K_t}) \sim \mathrm{N}(0, \mathrm{I}_d)$. Therefore, with probability $1 - 1/t^2$ we have

$$B_{t,K_t}^{-1}(\tilde{\theta}_{t,K_t} - \tilde{\mu}_{t,K_t}) \leq \sqrt{4d\log(t^3)} \ .$$

Finally, we conclude that with probability $1 - 1/t^2$, it holds that

$$|\phi(x_t,a)^\top \left(\tilde{\theta}_{t,K_t} - \tilde{\mu}_{t,K_t}\right)| \leq \sqrt{4d\log(t^3)/\eta}\left(1 + 1/\sqrt{\lambda}\right)\|\phi(x_t,a)\|_{V_t^{-1}} \ .$$

□

A.6.1. Concentration of the mean of the Variational posterior around its mean

In this section, the objective is the show to concentration of $\tilde{\mu}_{t,k}$ around its mean $\tilde{m}_{t,k}$ More precisely, we want an upper bound of $|\boldsymbol{\phi}^\top \left(\boldsymbol{\tilde{\mu}_{t,k} - \tilde{m}_{t,k}}\right)|$.

Lemma A.15. *For any $t \in [T]$ and $k \in [K_t]$, we have the following relation*

$$\tilde{W}_{t,k+1} = (\mathrm{I} - \eta h_t V_t)\tilde{W}_{t,k}(\mathrm{I} - \eta h_t V_t)^T + \eta^2 h_t^2 V_t \mathbb{E}[\tilde{\Sigma}_{t,k}]V_t$$

where the sequence $\{\tilde{W}_{t,k}\}_{k=1}^{K_t}$ is introduced in section A.1. (Recall : $\tilde{\mu}_{t,k} \sim \mathrm{N}(\tilde{m}_{t,k}, \tilde{W}_{t,k})$)

Proof. In this section, we focus on the covariance matrix $\tilde{W}_{t,k}$ (see definition A.1), by definition we have

$$\begin{aligned}\tilde{W}_{t,k+1} &= \mathbb{E}[(\tilde{\mu}_{t,k+1} - \tilde{m}_{t,k+1})(\tilde{\mu}_{t,k+1} - \tilde{m}_{t,k+1})^\top] \\ &= \mathbb{E}[a_{t,k+1} a_{t,k+1}^\top] \ ,\end{aligned}$$

where $a_{t,k}$ is the difference between $\tilde{\mu}_{t,k}$ and its mean. For ease of notation, let's define $\Omega_{t,k} := \tilde{\theta}_{t,k} - \tilde{m}_{t,k}$, then we have

$$a_{t,k+1} = \tilde{\mu}_{t,k} - \tilde{m}_{t,k} - \eta h_t V_t(\tilde{\theta}_{t,k} - \tilde{m}_{t,k}) = \tilde{\mu}_k - \tilde{m}_{t,k} - \eta h_t V_t \Omega_{t,k}.$$

Consequently,

$$\begin{aligned}a_{t,k+1} a_{t,k+1}^\top = &(\tilde{\mu}_{t,k} - \tilde{m}_{t,k})(\tilde{\mu}_{t,k} - \tilde{m}_{t,k})^\top - \eta h_t V_t \Omega_{t,k}(\tilde{\mu}_{t,k} - \tilde{m}_{t,k})^\top \\ &- \eta h_t(\tilde{\mu}_{t,k} - \tilde{m}_{t,k})\Omega_{t,k}^\top V_t + \eta^2 h_t^2 V_t \Omega_{t,k}\Omega_{t,k}^\top V_t\end{aligned}$$

Moreover, note that $\tilde{\theta}_{t,k} = \tilde{\mu}_{t,k} + B_{t,k}\epsilon_{t,k}$ where $\epsilon_{t,k} \sim \mathrm{N}(0, \mathrm{I})$. Subsequently we have $\Omega_{t,k} = \tilde{\mu}_{t,k} - \tilde{m}_{t,k} + \tilde{\Sigma}_{t,k}^{1/2}\epsilon_{t,k}$. Then we have $\mathbb{E}[\Omega_{t,k}\Omega_{t,k}^\top] = \tilde{W}_{t,k} + \mathbb{E}[B_{t,k}B_{t,k}^\top]$, $\mathbb{E}[\Omega_{t,k}(\tilde{\mu}_{t,k} - \tilde{m}_{t,k})^\top] = W_{t,k}$, and $\mathbb{E}[(\tilde{\mu}_{t,k} - \tilde{m}_{t,k})\Lambda_{t,k}^\top] = \tilde{W}_{t,k}$. Finally we obtain that

$$\begin{aligned}\tilde{W}_{t,k+1} &= \mathbb{E}[a_{t,k+1} a_{t,k+1}^\top] \\ &= \tilde{W}_{t,k} - \eta h_t V_t \tilde{W}_{t,k} - \eta h_t \tilde{W}_{t,k} V_t + \eta^2 h_t^2 V_t \tilde{W}_{t,k} V_t + \eta^2 h_t^2 V_t \mathbb{E}[B_{t,k}B_{t,k}^\top] V_t \\ &= (\mathrm{I} - \eta h_t V_t)\tilde{W}_{t,k}(\mathrm{I} - \eta h_t V_t)^\top + \eta^2 h_t^2 V_t \mathbb{E}[B_{t,k}B_{t,k}^\top] V_t \ .\end{aligned}$$

□

Lemma A.16. *Recall that $\tilde{\mu}_{t,K_t}$, the mean of the variational posterior after K_t steps of gradient descent, is a sample from the Gaussian with mean $\tilde{m}_{t,K_t}$ and covariance matrix $\tilde{W}_{t,K_t}$, ie, $\tilde{\mu}_{t,K_t} \sim \mathrm{N}(\tilde{m}_{t,K_t}, \tilde{W}_{t,K_t})$. Recall the definition of $\Lambda_{t,k} = \tilde{\Sigma}_{t,k} - 1/\eta V_t^{-1}$, and let denote by $\Gamma_{t,k} = \tilde{W}_{t,k} - J_t V_t^{-1}$, where $J_t = h_t(2\mathrm{I} - \eta h_t V_t)^{-1} V_t$.*

This Lemma shows that the 2-norm of Γ_{t,K_t} is controlled by

$$\|\Gamma_{t,K_t}\|_2 \leq \sum_{j=1}^{t} \frac{6\kappa_j h_j \|V_j\|_2}{\lambda} \sum_{r=1}^{j} \prod_{i=r}^{t} D_i^{K_i - 1} \ ,$$

where $\kappa_j = \lambda_{max}(V_j)/\lambda_{min}(V_j)$ and $D_i = \mathrm{I} - 3\eta h_i \lambda_{min}(V_i)/2$.

Proof. Lemma (A.15) gives us that

$$\tilde{W}_{t,k+1} = A_t \tilde{W}_{t,k} A_t + \eta^2 h_t^2 V_t \tilde{\Sigma}_{t,k} V_t \ ,$$

where $A_t = \mathrm{I} - \eta h_t V_t$.

Note that J_t and V_t commute, therefore we have

$$A_t J_t V_t^{-1} A_t = J_t V_t^{-1} - 2h_t \eta J_t + \eta^2 h_t^2 J_t V_t \ .$$

Consequently, by combining the two previous equations we obtain

$$\begin{aligned}\Gamma_{t,k+1} &= A_t \Gamma_{t,k} A_t - 2h_t \eta J_t + \eta^2 h_t^2 J_t V_t + \eta h_t^2 V_t + \eta^2 h_t^2 V_t \Lambda_{t,k} V_t \\ &= A_t \Gamma_{t,k} A_t + \eta^2 h_t^2 V_t \Lambda_{t,k} V_t - h_t \eta J_t(2\mathrm{I} - \eta h_t V_t) + \eta h_t^2 V_t \\ &= A_t \Gamma_{t,k} A_t + \eta^2 h_t^2 V_t \Lambda_{t,k} V_t \ .\end{aligned}$$

It follows that

$$\|\Gamma_{t,k+1}\|_2 \leq \|A_t\|_2^2 \|\Gamma_{t,k}\|_2 + \eta^2 h_t^2 \|V_t\|_2^2 \|\Lambda_{t,k}\|_2$$

Therefore, iterating over k gives us

$$\|\Gamma_{t,k+1}\|_2 \leq \|A_t\|_2^{2k}\|\Gamma_{t,k}\|_2 + \eta^2 h_t^2 \sum_{j=0}^{k-1}\|A_t\|_2^{2j}\|V_t\|_2^2\|\Lambda_{t,k-j}\|_2 \ .$$

Moreover, Equation (40) is used to controls the following quantity

$$\|\Lambda_{t,k}\|_2 \leq \left(1 - \frac{3\eta h_t}{2}\lambda_{\min}(V_t)\right)^{k-1}\|\Lambda_{t,1}\|_2 \ .$$

Let's denote by $D_t = 1 - 3\eta h_t \lambda_{\min}(V_t)/2$, Subsequently

$$\|\Gamma_{t,k+1}\|_2 \leq \|A_t\|_2^{2k}\|\Gamma_{t,k}\|_2 + \eta^2 h_t^2 \sum_{j=0}^{k-1}\|A_t\|_2^{2j}\|V_t\|_2^2 D_t^{k-j-1}\|\Lambda_{t,1}\|_2 \ .$$

However, $\|A_t\|_2^2 = (1-\eta h_t\lambda_{\min}(V_t))^2 < (1 - \frac{3\eta h_t}{2}\lambda_{\min}(V_t))$, because $\eta h_t \leq 1/(4\lambda_{\min}(V_t))$. Consequently, the geometric sum has a common ratio strictly lower than 1, then it is upper bounded by:

$$\begin{aligned}
\sum_{j=0}^{k-1}\left(\frac{\|A_t\|_2^2}{(1-\frac{3\eta h_t}{2}\lambda_{\min}(V_t))}\right)^j &\leq \sum_{j=0}^{+\infty}\left(\frac{\|A_t\|_2^2}{(1-\frac{3\eta h_t}{2}\lambda_{\min}(V_t))}\right)^j \\
&= \frac{1-\frac{3\eta h_t}{2}\lambda_{\min}(V_t)}{1-\frac{3\eta h_t}{2}\lambda_{\min}(V_t) - \|A_t\|_2^2} \\
&\leq \frac{1-\frac{3\eta h_t}{2}\lambda_{\min}(V_t)}{1/2\ \eta h_t\lambda_{\min}(V_t) - \eta^2 h_t^2\lambda_{\min}(V_t)^2} \\
&\leq \frac{6}{\eta h_t \lambda_{\min}(V_t)} \ , \qquad (43)
\end{aligned}$$

where in the first inequality we have used that the ratio of the previous sum is positive. In the last inequality we have used that $\eta h_t \leq 1/(6\lambda_{\min}(V_t))$ in the denominator and we can remove the negative part of the numerator. Therefore, it holds that

$$\|\Gamma_{t,k+1}\|_2 \leq \|A_t\|_2^{2k}\|\Gamma_{t,k}\|_2 + 6\eta\kappa_t h_t D_t^{k-1}\|V_t\|_2\|\Lambda_{t,1}\|_2 \ ,$$

where the last inequality comes from (43) and the definition of $\kappa_t = \lambda_{\max}(V_t)/\lambda_{\min}(V_t)$. Finally, iterating over t yields to:

$$\begin{aligned}
\|\Gamma_{t,k+1}\|_2 &\leq \|A_t\|_2^{2k}\prod_{j=1}^{t-1}\|A_j\|_2^{2(K_j-1)}\|\Gamma_{1,1}\|_2 + \sum_{j=1}^{t-1}\|A_t\|_2^{2k}\prod_{i=j+1}^{t-1}\|A_i\|_2^{2(K_i-1)}\Big(6\eta\kappa_j h_j D_j^{K_j-1}\|V_j\|_2\|\Lambda_{t,1}\|_2\Big) \\
&\quad + 6\eta\kappa_t h_t D_t^{k-1}\|V_t\|_2\|\Lambda_{t,1}\|_2 \\
&\leq \sum_{j=1}^{t-1}\|A_t\|_2^{2k}\prod_{i=j+1}^{t-1}\|A_i\|_2^{2(K_i-1)}\Big(6\eta\kappa_j h_j D_j^{K_j-1}\|V_j\|_2\|\Lambda_{t,1}\|_2\Big) + 6\eta\kappa_t h_t D_t^{k-1}\|V_t\|_2\|\Lambda_{t,1}\|_2 \ ,
\end{aligned}$$

where in the last inequalities we have used that $W_{1,1}$ is initialized such that $W_{1,1} = 1/(11\eta\lambda)\mathrm{I}$ and that $J_1V_1 = h_1(2\mathrm{I} - \eta h_1\lambda\mathrm{I})^{-1} = 1/(11\eta\lambda)\mathrm{I}$ because $h_1 = 1/(6\eta\lambda)$. Finally, we can conclude

$$\begin{aligned}
\|\Gamma_{t,K_t}\|_2 &\leq \sum_{j=1}^{t-1}\|A_t\|_2^{2(K_t-1)}\prod_{i=j+1}^{t-1}\|A_i\|_2^{2(K_i-1)}\Big(6\eta\kappa_j h_j D_j^{K_j-1}\|V_j\|_2\|\Lambda_{t,1}\|_2\Big) + 6\eta\kappa_t h_t D_t^{K_t-1}\|V_t\|_2\|\Lambda_{t,1}\|_2 \\
&= \sum_{j=1}^{t}\prod_{i=j+1}^{t}\|A_i\|_2^{2(K_i-1)}\Big(6\eta\kappa_j h_j D_j^{K_j-1}\|V_j\|_2\|\Lambda_{t,1}\|_2\Big) \\
&\leq \sum_{j=1}^{t}\prod_{i=j}^{t} D_i^{K_i-1}\Big(6\eta\kappa_j h_j\|V_j\|_2\|\Lambda_{t,1}\|_2\Big) \ ,
\end{aligned}$$

where in the last inequality we have used that $\|A_t\|_2^2 \leq D_t$. Moreover, equation (40) gives us that $\|\Lambda_{j,1}\|_2 \leq 1/(\eta\lambda)\sum_{r=1}^{j}\prod_{l=r}^{j-1} D_l^{K_l-1}$. Consequently, it holds that

$$\begin{aligned}\|\Gamma_{t,K_t}\|_2 &\leq \sum_{j=1}^{t}\frac{6\kappa_j h_j\|V_j\|_2}{\lambda}\sum_{r=1}^{j}\prod_{l=r}^{j-1} D_l^{K_l-1}\prod_{i=j}^{t} D_i^{K_i-1}\\ &= \sum_{j=1}^{t}\frac{6\kappa_j h_j\|V_j\|_2}{\lambda}\sum_{r=1}^{j}\prod_{i=r}^{t} D_i^{K_i-1}\ .\end{aligned}$$

□

Lemma A.17. *For any $t \geq 2$, given E_t^{true}, if the number of gradient descent steps is $K_t \geq 1 + 4(1 + 2\kappa_t^2)\log\left(2\kappa_t d^2 T\log^2(3T^3)\right)/3$, therefore it holds that*

$$|\phi^\top(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t})| \leq \left(\sqrt{\frac{3}{\eta\lambda d\log(3t^3)}} + \sqrt{4d\log(3t^3)/(11\eta)}\right)\|\phi\|_{V_t^{-1}}\ .$$

Proof. For any $\phi \in \mathbb{R}^d$,

$$|\phi^\top(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t})| \leq \|\phi^\top\tilde{W}_{t,K_t}^{1/2}\|_2\|\tilde{W}_{t,K_t}^{-1/2}(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t}))\|_2\ , \tag{44}$$

where $\tilde{W}_{t,K_t}^{1/2}$ is the unique symmetric square root of $\tilde{W}_{t,K_t}$. Firstly, given $\mathrm{E}_t^{\text{true}}$, the term $\|\tilde{W}_{t,K_t}^{-1/2}(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t})\|_2 < \sqrt{4d\log(3t^3)}$.

Then, we observe that

$$\begin{aligned}\sqrt{4d\log(3t^3)}\|\tilde{W}_{t,K_t}^{1/2}\phi\|_2 &= \sqrt{4d\log(3t^3)\phi\tilde{W}_{t,K_t}\phi^\top}\\ &\leq \sqrt{4d\log(3t^3)\phi\Gamma_{t,K_t}\phi^\top} + \sqrt{4d\log(3t^3)\phi J_t V_t^{-1}\phi^\top}\ ,\end{aligned} \tag{45}$$

where $J_t = h_t(2\mathrm{I} - \eta h_t V_t)^{-1}V_t = (2V_t^{-1}/h_t - \eta\mathrm{I})^{-1}$ and $\Gamma_{t,k} = \tilde{W}_{t,k} - J_t V_t^{-1}$.

Moreover,

$$\begin{aligned}\sqrt{\phi J_t V_t^{-1}\phi^\top} &= \|(J_t V_t^{-1})^{1/2}\phi\|_2\\ &\overset{(a)}{=} \|J_t^{1/2}V_t^{-1/2}\phi\|_2\\ &\leq \|J_t^{1/2}\|_2\|\phi\|_{V_t^{-1}}\ ,\end{aligned}$$

where in inequality (a) we have used that J_t and V_t^{-1} commute.

Recall that V_t is a symmetric matrix, therefore we have $\lambda_{\min}(V_t)\mathrm{I} \preceq V_t \preceq \lambda_{\max}(V_t)\mathrm{I}$. It follows that

$$\frac{2}{h_t\lambda_{\max}(V_t)}\mathrm{I} \preceq \frac{2}{h_t}V_t^{-1} \preceq \frac{2}{h_t\lambda_{\min}(V_t)}\mathrm{I}.$$

Recall the definition of $h_t = \lambda_{\min}(V_t)/\big(2\eta(\lambda_{\min}(V_t)^2 + \lambda_{\max}(V_t)^2)\big)$. Consequently, the previous relation becomes

$$\Big(\frac{4\eta(1+2\kappa_t^2)}{\kappa_t} - \eta\Big)\mathrm{I} \preceq \frac{2}{h_t}V_t^{-1} - \eta\mathrm{I} \preceq \Big(3\eta + 8\eta\kappa_t^2\Big)\mathrm{I}\ . \tag{46}$$

The left hand term is obviously positive, therefore it holds that

$$\begin{aligned}\|J_t^{1/2}\|_2 &= \|(\frac{2}{h_t}V_t^{-1} - \eta \mathrm{I})^{-1/2}\|_2 \\ &\leq \sqrt{\frac{\kappa_t}{\eta(4+8\kappa_t^2-\kappa_t)}} \\ &\leq \frac{1}{\sqrt{\eta(4+7\kappa_t^2)}} \\ &\leq \frac{1}{\sqrt{11\eta}} \ .\end{aligned}$$

Finally,

$$\sqrt{4d\log(3t^3)\phi J_t V_t^{-1}\phi^\top} \leq \sqrt{4d\log(3t^3)/(11\eta)}\|\phi\|_{V_t^{-1}} \ .$$

Now, we focus on the first term of equation 45. Lemma A.16 gives us that

$$\begin{aligned}\|\Gamma_{t,K_t}\|_2 &\leq \sum_{j=1}^{t} \frac{6\kappa_j h_j \|V_j\|_2}{\lambda} \sum_{r=1}^{j}\prod_{i=r}^{t} D_i^{K_i-1} \\ &\leq \sum_{j=1}^{t} \frac{\kappa_j}{\eta\lambda} \sum_{r=1}^{j}\prod_{i=r}^{t} D_i^{K_i-1} \ ,\end{aligned}$$

where in the last inequality we have used that $h_t\|V_t\|_2 = \kappa_t/\big(2\eta(1+2\kappa_t)\big) \leq 1/(6\eta)$. For any $j \in [2,t]$, let's define $\epsilon_j = 1/(2(\kappa_j d^2 t^2 \log^2(3t^3)))$. Additionally, let's fix K_i such that $D_i^{K_i-1} \leq \epsilon_j$ (this condition will be explained later in the Lemma). Subsequently, we have

$$\begin{aligned}4d\log(3t^3)\|V_t\|_2/(\eta\lambda)\sum_{j=1}^{t}\kappa_j\sum_{r=1}^{j}\prod_{i=r}^{t} D_i^{K_i-1} &\leq 4d\log(3t^3)\|V_t\|_2/(\eta\lambda)\sum_{j=1}^{t}\kappa_j\sum_{r=1}^{j}\epsilon_j^{t-r+1} \\ &\leq \frac{2\|V_t\|_2}{t^2\eta\lambda d\log(3t^3)}\sum_{r=1}^{t}\sum_{j=r}^{t}\epsilon_j^{t-r} \\ &\overset{(a)}{\leq} \frac{2\|V_t\|_2}{t^2\eta\lambda d\log(3t^3)}\sum_{r=1}^{t}\sum_{j=r}^{t}\Big(\frac{1}{2d^2t^2\log^2(3t^3)}\Big)^{t-r} \\ &\leq \frac{2\|V_t\|_2}{t\eta\lambda d\log(3t^3)}\sum_{r=1}^{t}\frac{t-r+1}{t}\Big(\frac{1}{2d^2t^2\log^2(3t^3)}\Big)^{t-r} \\ &\overset{(b)}{\leq} \frac{2\|V_t\|_2}{t\eta\lambda d\log(3t^3)}\sum_{r=1}^{t}\Big(\frac{1}{2d^2t^2\log^2(3t^3)}\Big)^{t-r} \\ &\overset{(c)}{\leq} \frac{2\|V_t\|_2}{\eta t\lambda d\log(3t^3)}\sum_{u=0}^{t-1}\Big(\frac{1}{25}\Big)^{u} \\ &\leq \frac{3}{\eta\lambda d\log(3t^3)} \ ,\end{aligned}$$

where in (a) we have used that $\epsilon_j \leq 1/(2(d^2t^2\log^2(3t^3))$. Inequality (b) is due to $t-r+1 \leq t$. The inequality (c) is obtained because $1/(2d^2t^2\log^2 3t^3) \leq 1/(4\times\log^2(8)) \leq 1/25$ and $u = t-r$. For the last inequality we have used the geometric series formula and $\|V_t\|_2 = \|\lambda\mathrm{I} + \sum_{s=1}^{t-1}\phi\phi^\top\|_2 \leq \lambda + t - 1 \leq t$, because $\lambda \leq 1$.

Consequently, as $\|\phi\|_2 \leq \|V_t^{1/2}\|_2 \|\phi\|_{V_t^{-1}}$, we obtain

$$\sqrt{4d\log(3t^3)\phi\Gamma_{t,K_t}\phi^\top} \leq \sqrt{\frac{3}{\eta\lambda d\log(3t^3)}}\|\phi\|_{V_t^{-1}} \ . \tag{47}$$

Moreover, the previous inequalities hold if $(1-(3/2)\eta h_i \lambda_{\min}(V_i))^{K_i-1} \leq \epsilon$, following a similar reasoning than in section A.5.2, it follows that we need

$$K_t \geq 1 + 4(1+2\kappa_t^2)\log\left(2\kappa_t d^2 T^2 \log^2(3T^3)\right)/3$$

□

B. Concentration and anti-concentration

Lemma B.1. ***(Concentration lemma for $\tilde{\theta}_t$)***

For any $t \in [T]$, given E_t^{true}, the following event is controlled

$$\mathbb{P}(\mathrm{E}_t^{var}|\mathcal{F}_{t-1}) \geq 1 - \frac{1}{t^2}$$

Proof. Firstly, if $t=1$, the condition is obvious because $\mathbb{P}(\mathrm{E}_t^{\text{var}}|\mathcal{F}_{t-1}) \geq 0$. For the rest of the proof, we assume that $t \geq 2$. Recall the definition of the event $\mathrm{E}_t^{\text{var}}$:

$$\mathrm{E}_t^{\text{var}} = \left\{\text{for any } a \in \mathcal{A}(x_t), |\phi(x_t,a)^\top\tilde{\theta}_t - \phi(x_t,a)^\top\hat{\mu}_t| \leq g_2(t)\|\phi(x_t,a)\|_{V_t^{-1}}\right\} \ .$$

with $g_2(t) = 10\sqrt{d\log(3t^3)/(\eta\lambda)}$.

Let $a \in \mathcal{A}(x_t)$, it holds that

$$|\phi(x_t,a)^\top(\tilde{\theta}_t - \hat{\mu}_t)| \leq |\phi(x_t,a)^\top(\tilde{\theta}_{t,K_t} - \tilde{\mu}_{t,K_t})| + |\phi(x_t,a)^\top(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t})| + |\phi(x_t,a)^\top(\tilde{m}_{t,K_t} - \hat{\mu})| \ ,$$

where $\tilde{\theta}_t = \tilde{\theta}_{t,K_t}$ is a sample from the variational posterior distribution trained after K_t steps of Algorithm 2. $\tilde{\mu}_{t,K_t}$ and $\tilde{\Sigma}_{t,K_t}$ are, respectively, the mean and covariance matrix of the variational posterior. Moreover, $\tilde{\mu}_{t,K_t}$ is gaussian with mean $\tilde{m}_{t,K_t}$ and covariance matrix $\tilde{W}_{t,K_t}$ (see section A.1). If the number of gradient descent steps is $K_t^{(1)} \geq 1+4(1+2\kappa_t^2)\log(2T)/3$, then Lemma A.14 shows that with probability at least $1-1/t^2$, we have

$$\begin{aligned}|\phi(x_t,a)^\top\left(\tilde{\theta}_{t,K_t} - \tilde{\mu}_{t,K_t}\right)| &\leq \sqrt{4d\log(t^3)/\eta}\left(1+1/\sqrt{\lambda}\right)\|\phi(x_t,a)\|_{V_t^{-1}} \\ &\leq 4\sqrt{d\log(t^3)/(\eta\lambda)}\|\phi(x_t,a)\|_{V_t^{-1}},\end{aligned}$$

where the last inequality is due to $\lambda \leq 1$.

Similarly, Lemma A.17 shows that for any $t \geq 2$, given $\mathrm{E}_t^{\text{true}}$, if $K_t^2 \geq 1+4(1+2\kappa_t^2)\log\left(2\kappa_t d^2T^2\log^2(3T^3)\right)/3$, therefore we have

$$|\phi(x_t,a)^\top(\tilde{\mu}_{t,K_t} - \tilde{m}_{t,K_t})| \leq \left(\sqrt{\frac{3}{\eta\lambda d\log(3t^3)}} + \sqrt{4d\log(3t^3)/(11\eta)}\right)\|\phi(x_t,a)\|_{V_t^{-1}}$$

where in the last simplification we have used $\lambda \leq 1$.

Finally, Given $\mathrm{E}_t^{\text{true}}$, let's apply Lemma A.8 with a number of gradient descent steps such $K_t^{(3)} \geq 1+2(1+2\kappa_t^2)\log\left(4R\sqrt{dT\log(3T^3)}\right)$, we obtain that

$$|\phi(\tilde{m}_{t,K_t} - \hat{\mu}_t)| \leq 2/\lambda\|\phi(x_t,a)\|_{V_t^{-1}}.$$

Note that $K_t = 1 + 2(1+2\kappa_t^2)\log\left(2R\kappa_t d^2T^2\log^2(3T^3)\right) \geq \max\{K_t^{(1)}, K_t^{(2)}, K_t^{(3)}\}$ (see Equation (23)), then with probability at least $1 - 1/t^2$ we have

$$\begin{aligned}
|\phi(x_t,a)^\top\left(\tilde{\theta}_{t,k} - \hat{\mu}_t\right)| &\leq |\phi(x_t,a)^\top\left(\tilde{\theta}_{t,k} - \tilde{\mu}_{t,k}\right)| + |\phi(x_t,a)^\top\left(\tilde{\mu}_{t,k} - \tilde{m}_{t,k}\right)| + |\phi(x_t,a)^\top\left(\tilde{m}_{t,k} - \hat{\mu}\right)| \\
&\leq \left(4\sqrt{d\log(t^3)/(\eta\lambda)} + \sqrt{\frac{3}{\eta\lambda d\log(3t^3)}} + \sqrt{4d\log(3t^3)/(11\eta)} + 2/\lambda\right)\|\phi(x_t,a)\|_{V_t^{-1}} \\
&\leq 10\sqrt{d\log(3t^3)/(\eta\lambda)}\|\phi(x_t,a)\|_{V_t^{-1}} \\
&\leq g_2(t)\|\phi(x_t,a)\|_{V_t^{-1}} \ .
\end{aligned}$$

where the last inequality holds because $t \geq 2$, $\lambda \leq 1$ and $\eta \leq 1$. □

Lemma B.2. *(Anti-concentration lemma) Given E_t^{true}, if the number of gradient steps is $K_t = 1 + 2(1+\kappa_t^2)\log\left(2R\kappa_t d^2T^2\log^2(3T^3)\right)$ Therefore, it holds that*

$$\mathbb{P}\left(\overset{*}{\phi}_t^\top\tilde{\theta}_{t,k} > \overset{*}{\phi}_t^\top\theta^\star\right) \leq p \ ,$$

where $p = 1/(2\sqrt{2\pi e})$

Proof. Firstly, note that

$$\mathbb{P}\left(\overset{*}{\phi}_t^\top\tilde{\theta}_{t,K_t} > \overset{*}{\phi}_t^\top\theta^\star\right) = \mathbb{P}\left(\frac{\overset{*}{\phi}_t^\top\tilde{\theta}_{t,K_t} - \overset{*}{\phi}_t^\top\tilde{m}_{t,K_t}}{\sqrt{\overset{*}{\phi}_t^\top\tilde{\Sigma}_{t,K_t}\overset{*}{\phi}_t + \overset{*}{\phi}_t^\top\tilde{W}_{t,K_t}\overset{*}{\phi}_t}} > \frac{\overset{*}{\phi}_t^\top\theta^\star - \overset{*}{\phi}_t^\top\tilde{m}_{t,K_t}}{\sqrt{\overset{*}{\phi}_t^\top\tilde{\Sigma}_t\overset{*}{\phi}_t + \overset{*}{\phi}_t^\top\tilde{W}_{t,K_t}\overset{*}{\phi}_t}}\right) \ .$$

Recall that

$$\overset{*}{\phi}_t^\top\tilde{\mu}_t \sim \mathrm{N}\big(\overset{*}{\phi}_t^\top\tilde{m}_t,\ \overset{*}{\phi}_t\tilde{W}_{t,k}\overset{*}{\phi}_t^\top\big) \text{ and } \quad \overset{*}{\phi}_t^\top\tilde{\theta}_{t,K_t} \sim \mathrm{N}\big(\overset{*}{\phi}_t^\top\tilde{\mu}_{t,k},\ \overset{*}{\phi}_t^\top\tilde{\Sigma}_t\overset{*}{\phi}_t\big) \ .$$

Therefore, using the conditional property of Gaussian vectors, we have

$$\overset{*}{\phi}_t^\top\tilde{\theta}_t \sim \mathrm{N}\big(\overset{*}{\phi}_t^\top\tilde{m}_t,\ \overset{*}{\phi}_t\tilde{\Sigma}_t\overset{*}{\phi}_t^\top + \overset{*}{\phi}_t\tilde{W}_t\overset{*}{\phi}_t^\top\big) \ .$$

Consequently, we have to control the term

$$Y_t := \big(\overset{*}{\phi}_t^\top\theta^\star - \overset{*}{\phi}_t^\top\tilde{m}_{t,K_t}\big)/\big(\sqrt{\overset{*}{\phi}_t^\top\tilde{\Sigma}_{t,K_t}\overset{*}{\phi}_{t,K_t} + \overset{*}{\phi}_t^\top\tilde{W}_{t,K_t}\overset{*}{\phi}_{t,K_t}}\big)$$

and use the Gaussian anti-concentration lemma (Lemma B.5). First, in this lemma, we suppose that $\mathrm{E}_t^{\text{true}}$ holds, therefore we have

$$\begin{aligned}
|\overset{*}{\phi}_t^\top(\hat{\mu}_t - \theta^\star)| &\leq g_1(t)\|\overset{*}{\phi}_t\|_{V_t^{-1}} \\
&= \left(R\sqrt{d\log(3t^3)} + \sqrt{\lambda}\right)\|\overset{*}{\phi}_t\|_{V_t^{-1}} \ .
\end{aligned}$$

Moreover, as the number of gradient descent, defined in section A.2 is upper than $K_t^{(1)} = 1 + 2(1+2\kappa_i^2)\log\left(4R\sqrt{dT\log(3T^3)}\right)$, then Lemma A.8 gives us that

$$|\overset{*}{\phi}_t^\top(\tilde{m}_{t,K_t} - \hat{\mu}_t)| \leq \frac{2\|\overset{*}{\phi}_t\|_{V_t^{-1}}}{\lambda}.$$

Consequently, the numerator of Y_t is upper bounded by

$$
\begin{aligned}
|\phi_t^{*\top}(\theta^\star - \tilde{m}_{t,K_t})| &\overset{a}{\leq} |\overset{*}{\phi}_t^\top(\theta^\star - \hat{\mu}_{t,K_t})| + |\overset{*}{\phi}_t^\top(\hat{\mu}_{t,K_t} - \tilde{m}_{t,K_t})| \\
&\overset{(b)}{\leq} \left(R\sqrt{d\log(3t^3)} + \sqrt{\lambda} + \frac{2}{\lambda} \right) \|\overset{*}{\phi}_t^\top\|_{V_t^{-1}}
\end{aligned}
$$

Regarding the denominator of Y_t, we need a lower bound for $\|B_{t,k}\overset{*}{\phi}_t\|_2$. Lemma A.10 for VITS-Ior A.12 for VITS-IIgives us that

$$
\|B_{t,K_t}\phi\|_2 \geq 1/\sqrt{\eta}\Big(1 - \sqrt{\|V_t\|_2 C_t}\Big)\|\overset{*}{\phi}_t\|_{V_t^{-1}}
$$

with

$$
-C_t^{1/2} = -\Big(\frac{1}{\lambda}\sum_{j=1}^{t-1}\prod_{i=j+1}^{t}(1 - \frac{3h_t\eta}{2}\lambda_{\min}(V_t))^{K_i-1}\Big)^{1/2}. \tag{48}
$$

Finally, we find a lower bound to this quantity.

$$
\begin{aligned}
\|B_{t,K_t}\phi\|_2 &\geq 1/\sqrt{\eta}\Big(1 - \sqrt{\|V_t\|_2 C_t}\Big)\|\overset{*}{\phi}_t\|_{V_t^{-1}} \\
&\overset{(a)}{=} \frac{\|\overset{*}{\phi}_t\|_{V_t^{-1}}}{\sqrt{\eta}}\Big(1 - \sqrt{\|V_t\|_2}\Big(\frac{1}{\lambda}\sum_{j=1}^{t-1}\prod_{i=j+1}^{t}(1 - \frac{3h_t\eta}{2}\lambda_{\min}(V_t))^{K_i-1}\Big)^{1/2}\Big) \\
&\overset{(b)}{=} \frac{\|\overset{*}{\phi}_t\|_{V_t^{-1}}}{\sqrt{\eta}}\left(1 - \Big(\sum_{j=1}^{t-1}\epsilon^{t-j-1}\Big)^{1/2}\right) \\
&\overset{(c)}{=} \frac{\|\overset{*}{\phi}_t\|_{V_t^{-1}}}{\sqrt{\eta}}\left(1 - \Big(\sum_{j=0}^{t-2}\epsilon^{j}\Big)^{1/2}\right) \\
&\overset{(d)}{\geq} \frac{\|\overset{*}{\phi}_t\|_{V_t^{-1}}}{\sqrt{\eta}}\Big(1 - \frac{1}{9t^{1/4}}\Big) \\
&\overset{(e)}{\geq} \frac{\|\overset{*}{\phi}_t\|_{V_t^{-1}}}{\sqrt{\eta}}\Big(1 - \frac{1}{9}\Big) \\
&= \frac{8\|\overset{*}{\phi}_t\|_{V_t^{-1}}}{9\sqrt{\eta}}
\end{aligned}
$$

with (a) is 48. Where (b) we use $\|V_t\|_2 \leq t$ and setting $\epsilon = (4t)^{-1}$, point (c) comes from a change of variable, (d) comes from the fact that for any $t \geq 1$, $\sum_{j=0}^{t-2}\epsilon^j < 1/(81\sqrt{t})$. Finally, (e) comes from that $1/t$ by can be upper bounded by 1 for any t.

Finally, regrouping the nominator and the denominator, we have the following expression for Y_t:

$$\begin{aligned} Y_t &\leq \frac{\overset{*}{\phi}_t^\top \theta^\star - \overset{*}{\phi}_t^\top \tilde{m}_{t,K_t}}{\sqrt{\overset{*}{\phi}_t^\top \tilde{\Sigma}_{t,K_t} \overset{*}{\phi}_{t,K_t} + \overset{*}{\phi}_t^\top \tilde{W}_{t,K_t} \overset{*}{\phi}_{t,K_t}}} \\ &\leq \frac{\overset{*}{\phi}_t^\top \theta^\star - \overset{*}{\phi}_t^\top \tilde{m}_{t,K_t}}{\|\overset{*}{\phi}_t\|_{\tilde{\Sigma}_{t,K_t}}} \\ &\leq \frac{R\sqrt{d\log(3t^3)} + \sqrt{\lambda} + \frac{2}{\lambda}}{8/(9\sqrt{\eta})} \\ &\leq \frac{9R\sqrt{d\log(3t^3)}\sqrt{\eta}}{2\lambda} \end{aligned}$$

Recall the definition of η in Section A.2

$$\eta = \frac{4\lambda^2}{81R^2 d\log(3T^3)}$$

Consequently, it yields that $|Y_t| \leq 1$.

Finally, Lemma B.5 gives us that

$$\mathbb{P}\Big(\overset{*}{\phi}_t^\top \tilde{\theta}_{t,K_t} > \overset{*}{\phi}_t^\top \theta^\star\Big) \geq \frac{1}{2\sqrt{2\pi e}}$$

□

B.1. Auxiliary Lemmas

Lemma B.3. ***(Azuma-Hoeffding inequality)*** *We define $\{X_s\}_{s\in[T]}$ a super-martingale associated to the filtration $\mathcal{F}_t$. If it holds that for any $s \geq 1$, $|X_{s+1} - X_s| \leq c_{s+1}$. Then for any $\epsilon > 0$, we have*

$$\mathbb{P}(X_T - X_0 \geq \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2\sum_{s=1}^T c_s^2}\right) .$$

Lemma B.4. ***(Martingale Lemma (Abbasi-Yadkori et al., 2011b))*** *Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration, $(m_t)_{t\geq 1}$ be an $\mathbb{R}^d$-valued stochastic process such that m_t is $(\mathcal{F}'_{t-1})$-measurable, $(\epsilon_t)_{t\geq 1}$ be a real-valued martingale difference process such that ϵ_t is $(\mathcal{F}'_t)$-measurable. For $t \geq 0$, define $\zeta_t = \sum_{\tau=1}^t m_\tau \epsilon_\tau$ and $M_t = \mathrm{I}_d + \sum_{\tau=1}^t m_\tau m_\tau^\top$, where I_d is the d-dimensional identity matrix. Assume ϵ_t is conditionally R-sub-Gaussian. Then, for any $\delta' > 0$, $t \geq 0$, with probability at least $1 - \delta'$,*

$$\|\zeta_t\|_{M_t^{-1}} \leq R\sqrt{d\log(\frac{t+1}{\delta'})}$$

where $\|\zeta_t\|_{M_t^{-1}} = \sqrt{\zeta_t^\top M_t^{-1} \zeta_t}$

Lemma B.5. ***(Gaussian concentration (Abramowitz & Stegun, 1964))*** *Suppose Z is a Gaussian random variable $Z \sim \mathcal{N}\left(\mu, \sigma^2\right)$, where $\sigma > 0$. For $0 \leq z \leq 1$, we have*

$$\mathbb{P}(Z > \mu + z\sigma) \geq \frac{1}{\sqrt{8\pi}} e^{-\frac{z^2}{2}}, \quad \mathbb{P}(Z < \mu - z\sigma) \geq \frac{1}{\sqrt{8\pi}} e^{-\frac{z^2}{2}} \tag{49}$$

And for $z \geq 1$, we have

$$\frac{e^{-z^2/2}}{2z\sqrt{\pi}} \leq \mathbb{P}(|Z - \mu| > z\sigma) \leq \frac{e^{-\frac{z^2}{2}}}{z\sqrt{\pi}}$$

C. Approximation of our algorithm and complexity

In this section, the objective is to approximate the inversion of the matrix $B_{t,k}$ of Algorithm 2. Indeed, Algorithm 2, requires to compute the inversion of a $d \times d$ matrix at each step t and k, which represents a complexity of $\mathcal{O}(d^3)$. In the approximated version of Algorithm 2, we consider both the sequence of square root covariance matrix $\{B_{t,k}\}_{k=1}^{K_t}$ and the sequence of their approximations $\{C_{t,k}\}_{k=1}^{K_t}$ such that: for any $t \in [T]$ and $k \in [K_t]$

$$C_{t,k} \approx B_{t,k}^{-1} \ .$$

Recall the recursive definition of $B_{t,k}$,

$$\begin{aligned} B_{t,k+1} &= \{\mathrm{I} - h_t A_{t,k})\} \, B_{t,k} + h_t (B_{t,k}^\top)^{-1} \\ &\approx \{\mathrm{I} - h_t A_{t,k})\} \, B_{t,k} + h_t C_{t,k}^\top \ , \end{aligned} \tag{50}$$

where $A_{t,k} = B_{t,k}^2(\tilde{\theta}_{t,k} - \tilde{\mu}_{t,k})(\nabla U_t(\tilde{\theta}_{t,k})^\top$ if the hessian free algorithm is used or $A_{t,k} = \nabla^2 U(\tilde{\theta}_{t,k})$ otherwise. Recall that $\tilde{\theta}_{t,k} \sim \mathrm{N}(\tilde{\mu}_{t,k}, B_{t,k}B_{t,k}^\top)$. Furthermore, we can now focus on the definition of the sequence $\{C_{t,k}\}_{k=1}^{K_t}$. Firstly, we recall that

$$\begin{aligned} B_{t,k+1} &= \{\mathrm{I} - h_t A_{t,k}\} \, B_{t,k} + h_t (B_{t,k}^\top)^{-1} \\ &= \{\mathrm{I} + h_t((B_{t,k}^\top)^{-1}(B_{t,k})^{-1} - A_{t,k}\} B_{t,k}. \end{aligned}$$

Then, let's use a first order Taylor expansion of the previous equation in h_t, we obtain the approximated inverse square root covariance matrix:

$$C_{t,k+1} = C_{t,k}^{-1}\{(\mathrm{I} - h_t(C_{t,k}^\top C_{t,k} - A_{t,k})\}. \tag{51}$$

Note that the lower is h_t, the better is the approximation and in our case the step size h_t is decreasing with t. The approximated recursive definition of the square root covariance matrix defined in equation (50) and its approximated inverse defined in equation (51) are used to defined our the approximated version of VITS called VITS − II and is presented in Algorithm 3. Moreover, note that the updating step of Algorithm 3 uses only matrix multiplication and sampling from independent Gaussian distribution $\mathrm{N}(0, \mathrm{I})$. Therefore the global complexity of the overall algorithm is $\mathcal{O}(d^2)$.

D. Discussion on the difference between the algorithm of (Urteaga & Wiggins, 2018) and our algorithm VITS.

The main difference between our setting and the one of (Urteaga & Wiggins, 2018) is the bandit modelisation. Indeed, given a context x and an action a, in our setting, the agent receives a reward $r \sim \mathrm{R}(\cdot|x, a)$. Consequently, a parametric model R_θ is used to approximate the reward distribution and it yields to a posterior distribution $\hat{\mathrm{p}}$. In the setting of (Urteaga & Wiggins, 2018), the agent receives a reward $r \sim \mathrm{R}_a(\cdot|x)$. Then, it considers a set of parametric models $\{\mathrm{R}_{\theta_a}\}_{a=1}^K$ and a set of posterior distributions: $\{\hat{\mathrm{p}}_a\}_{a=1}^K$. The setting we have used in this paper is richer as it consider the correlation between the arms distributions compared to (Urteaga & Wiggins, 2018) which consider that the arm distributions are independents. For example, if we consider the case of the Linear bandit. In this setting, the posterior distribution is Gaussian. With the modelisation of (Urteaga & Wiggins, 2018), we have for any $a \in [K]$, $\hat{\mathrm{p}}_a := \mathrm{N}(\mu_a, \Sigma_a)$, where $\mu_a \in \mathbb{R}^d$ and $\Sigma_a \in \mathcal{S}_+^d$. However, with our modelisation, $\hat{\mathrm{p}} := \mathrm{N}(\mu, \Sigma)$, where $\mu \in \mathbb{R}^{d \times K}$ and $\Sigma \in \mathcal{S}_+^{d \times K}$. We can see that the covariance matrix Σ encodes the correlations between the different arms, which is not the case of $\{\Sigma_a\}_{a=1}^K$. In addition, in our setting, we can consider any model for the mean of the reward distribution. For example we can choose $g(\theta, x, a)$ as a Neural Networks. This kind of model is unusable in the formulation of (Urteaga & Wiggins, 2018).

Moreover, the approximate families used in both papers are different. Indeed, we consider the family of non-degenerate Gaussian distributions, and (Urteaga & Wiggins, 2018) is focused on the family of mixture of mean-field Gaussian distribution. The mixture of Gaussian distribution is richer than the classic Gaussian distribution. However, the non mean-field hypothesis allow to keep the correlation between arms distributions.

Furthermore, VTS from (Urteaga & Wiggins, 2018) scales very poorly with the size of the problem. The variational parameters are very large: $\alpha \in \mathbb{R}^{K \times M}$, $\beta \in \mathbb{R}^{K \times M}$, $\gamma \in \mathbb{R}^{K \times M}$, $u \in \mathbb{R}^{K \times M \times d}$, $V \in \mathbb{R}^{K \times M \times d \times d}$ where K is the number

of arms, M is the number of mixtures and d the parameter dimension. In addition, the parameter updating step is also very costly in term of memory and speed. We have re-implemented an efficient version of their algorithm in JAX in order to scale as much as possible but many memory problems occur.

Finally, our algorithm comes with theoretical guarantees in the Linear Bandit case and outperforms empirically the others approximate TS methods. VTS performs poorly in practice and has no theoretical guarantee, even in the Linear case.

E. Hyper-parameters tuning

This section summarizes the different grid-search used to compute all the plots in this paper for the algorithms: LinTS, LMC-TS,, VITS − I, VITS − II and VITS − II Hessian-free.

Parameter	Value
inverse temperature η	$10, 100, 500, 1000$
regularization λ	$0.1, 1, 10$

Table 1. LinTS hyperparameter grid-search

Parameter	Value
inverse temperature η	$10, 100, 500, 1000$
regularization λ	$0.1, 1, 10$
Nb gradient steps K_t	10, 50
learning rate h	$0.001, 0.01, 0.1$

Table 2. LMC-TS hyperparameter grid-search

Parameter	Value
inverse temperature η	$10, 100, 500, 1000$
regularization λ	$0.1, 1, 10$
Nb gradient steps K_t	10
learning rate h	$0.001/\eta, 0.01/\eta, 0.1/\eta$
Monte Carlo samples	1 (Hessian) and 20 (Hessian-free)

Table 3. VITS hyperparameter grid-search

F. Experimental comparison between Langevin Monte Carlo and VI

In this section, we conduct an experimental comparison between Langevin Monte Carlo (LMC) and two variants of Variational Inference (VI), denoted as VI-I and VI-II, in approximating a specific target distribution. Our target distribution is a straightforward Gaussian distribution, represented as $\mathrm{p}_\star = \mathrm{N}(\mu_\star, \Sigma_\star)$. We perform LMC, VI-I, and VI-II for a designated number of iterations. In each iteration, we calculate the Kullback-Leibler distance between the approximated distribution and the target distribution. In this context, all distributions generated by LMC, VI-I, and VI-II take the form of Gaussians. To compute the mean and covariance matrix for LMC, we perform parameter averaging over the results obtained after 1000 burn-in steps (which are excluded from the plotted data). Then, the training is stopped when

$$\mathrm{KL}(q_k, \mathrm{p}_\star) \leq \epsilon \ , \tag{52}$$

or if the number of steps exceeds 50000 steps.

Figure 4 illustrates the relationship between the condition number of $\Sigma_\star$ and the number of steps needed to achieve (52). We conducted these experiments with three different step sizes and repeated them across 100 different seeds. The red dashed line in the figure represents the maximum allowable number of iterations.

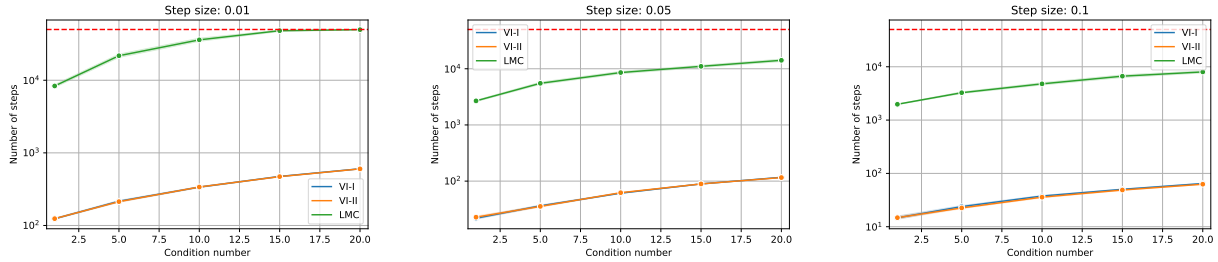


Figure 4. Comparison Langevin Monte Carlo and Variational inference

The first observation drawn from these figures is that VI-I and VI-II exhibit identical behavior, even when using a relatively large step size of 0.1. The second finding suggests that both LMC and VI exhibit a linear dependency on the condition number. However, we cannot definitively conclude that one algorithm is more robust in the face of varying condition numbers. Lastly, the third conclusion highlights that VI consistently requires fewer iterations to achieve (52).

G. Additional Results on non-contextual bandits

G.1. Linear and logistic bandit on synthetic data (non contectual)

In this subsection, we consider a contextual bandit setting with a parameter dimension $d = 10$ and a number of arms $K = 10$. The bandit environment is simulated by a random vector $\theta^\star \in \mathbb{R}^d$ sampled from a normal distribution $\mathrm{N}(0, \mathrm{I}_d)$ and subsequently scaled to unit norm. To create a complex environment that necessitates exploration, we define the set of contextual vectors as $\mathsf{X} := \{\theta^\star, \theta^\star_\epsilon, x_2, \dots, x_K\}$. Here, $\theta^\star_\epsilon$ is defined as $(\theta^\star + \epsilon)/\|(\theta^\star + \epsilon)\|_2$, where ϵ is sampled from a normal distribution with mean 0 and standard deviation 0.1. This contextual vector corresponds to a small modification of $\theta^\star$. The other contextual vectors are sampled from a normal distribution $\mathrm{N}(0, 1)$ and then scaled to unit norm.

Linear bandit scenario. Here, the true reward $\mathcal{R}(\cdot|x_a, a)$ associated to an action $a \in \{1, \dots, K\}$ and an arm $x_a \in \mathbb{R}^d$ corresponds to the distribution of $r_a = x_a^\top \theta^\star + \xi$, where the noise ξ is sampled from $\mathrm{N}(0, \mathrm{I})$. In this complex setting, we can calculate the expected reward for each arm as follows: $\mu_0 = \mathbb{E}[r_0] = 1$, $\mu_1 \approx 1 < \mu_0$, and for any $i > 1$, $\mu_i < \mu_1$. Intuitively, the first and second arms offered high rewards, while the remaining arms offered low rewards. On the other hand, finding the optimal arm is challenging and needs a significant amount of exploration.

Logistic bandit framework. We consider the same contextual set X, but the true reward $\mathcal{R}(\cdot|x_a, a)$ associated to an action $a \in \{1, \dots, K\}$ and an arm x_a now corresponds to $r_a \sim \mathrm{Ber}(\sigma(\langle x_a, \theta^\star \rangle))$, where Ber is the Bernoulli distribution, and $\sigma(x) = 1/(1 + e^{-x})$ is the logistic function. Similarly to the linear bandit, the logistic framework introduces a complex environment where a significant amount of exploration is required to accurately distinguish between the first and second arm.

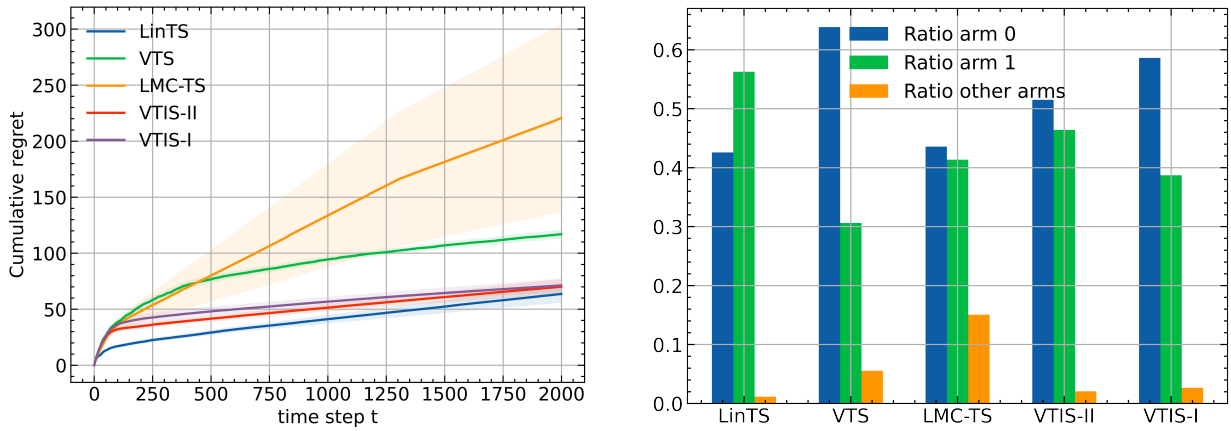


Figure 5. Linear Bandits

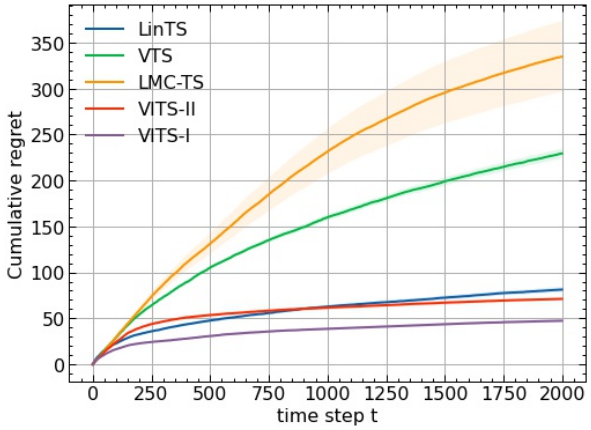


Figure 6. Logistic Bandits

Figures 5 and 6 display the cumulative regret (2) obtained by various TS algorithms, namely Linear TS (LinTS), Langevin Monte Carlo TS (LMC-TS), Variational TS (VTS), VITS-I and VITS-II in the linear and logistic bandit settings. The figure shows the mean and standard error of the cumulative regret over 20 samples. As depicted in Figure 5, VITS-I outperforms the other approximate TS algorithms in the linear bandit scenario. Note that the cumulative regret of VITS-I and VITS-II is comparable to that of Lin-TS, which uses the true posterior distribution. This observation highlights the efficiency of the variational TS algorithms in approximating the true posterior distribution and achieving similar performance to the Lin-TS algorithm. Figure 6 shows that VITS outperforms all other TS algorithms in the logistic setting too. This highlights the importance of employing approximation techniques in scenarios where the true posterior distribution cannot be sampled exactly. Moreover, both figures illustrate that VITS-II achieves a comparable regret to VITS-I while significantly reducing the computational complexity of the algorithm. Finally, as emphasized earlier, the settings we have chosen require a good tradeoff between exploration and exploitation that LMC-TS cannot achieve, as illustrated by the histogram in Figure 5.

H. Details about experiences in synthetic contextual bandits with synthetic data

In this subsection, we provide more details about the toy example derive in this paper. Firstly, we consider a fixed pool of arms denoted as $P = [\tilde{x}_1, \ldots, \tilde{x}_n]$ with $n = 50$, where each arm $\tilde{x}_i$ follows a normal distribution $\mathrm{N}(0_d, \mathrm{I}_d)$. Then, at each step $t \in [T]$, for every arm, we randomly sample a vector $\tilde{x}_i$ from the pool P, and the contextual vector associated with this arm is defined as $x = \tilde{x}_i + \zeta\epsilon$, where $\epsilon \sim \mathcal{N}(0_d, \mathrm{I}_d)$. The bandit environment is simulated using a random vector $\theta^\star$ sampled from a normal distribution $\mathrm{N}(0_d, \sigma^\star \mathrm{I}_d)$. We opted for $\sigma^\star = 1/d$ to ensure that the variance of the scalar product $x^\top \theta^\star$ remains independent of the dimension d. Indeed, both linear and quadratic settings, the reward only depends on the scalar product between the context and the true parameter. If we denote by $x[i]$ and $\theta^\star[i]$ the ith coordinate of the vector x and $\theta^\star$ respectively, then the scalar product is defined by

$$x^\top \theta^\star = \sum_{i=1}^{d} x[i]\theta^\star[i],$$

and its variance is defined by

$$\begin{aligned}\mathbb{V}[x^\top \theta^\star] &= \mathbb{V}[\sum_{i=1}^{d} x[i]\theta^\star[i]] \\ &= \sum_{i=1}^{d} \mathbb{V}[x[i]]\mathbb{V}[\theta^\star[i]] \\ &= d\sigma^\star\ \mathbb{V}[x[i]].\end{aligned}$$

In the previous equations we have used that all coordinates are independents identically distributed and centered. Therefore, taking $\sigma^\star = 1/d$ ensure that the variance of the scalar product remains independent of d. In the linear bandit setting, the

reward depends linearly on the contextual vector x, more precisely,

$$r = x^\top \theta^\star + \alpha\epsilon \ ,$$

where $\epsilon \sim \mathrm{N}(0_d, \mathrm{I}_d)$. However, to maintain problem complexity independent of ζ, we have set the signal-to-noise ratio to a fixed value of 1. This signal-to-noise ratio is the ratio between $\mathbb{E}[(x^\top\theta^\star)^2]$ and $\mathbb{E}[(\alpha\epsilon)^2]$. Firstly,

$$\begin{aligned}\mathbb{E}[(x^\top\theta^\star)^2] &= \mathbb{V}[x^\top\theta^\star] \\ &= \mathbb{V}[x[i]] \\ &= 1 + \zeta^2 \ ,\end{aligned}$$

where in the last equation we have used that $x = \tilde{x}_i + \zeta\epsilon$ and $\mathbb{V}[x[i]] = 1 + \zeta^2$. Moreover, the denominator of the signal-to-noise ratio is $\mathbb{E}[(\alpha\epsilon)^2] = \alpha^2$. Consequently, a signal-to-noise ratio equals to 1 implies that $\sqrt{1+\zeta^2} = \alpha$.

In the quadratic bandit setting, the reward depends quadratically on the contextual vector x, more precisely,

$$r = (x^\top\theta^\star)^2 + \alpha\epsilon \ ,$$

where $\epsilon \sim \mathrm{N}(0, \mathrm{I})$. In this setting, the reward also depends only on the scalar product between x and $\theta^\star$, thus, we also choose $\sigma^\star = 1/d$. We also ensure a signal-to-noise equal to 1, it implies a more sophisticated condition on the noise: $\alpha = (\zeta^2+1)\sqrt{3+6/d}$. More precisely, in the quadratic setting, the signal-to-noise ratio is defined as follow

$$\frac{\mathbb{E}[(x^\top\theta^\star)^4]}{\mathbb{E}[(\alpha\epsilon)^2]} = 1.$$

Firstly,

$$\begin{aligned}\mathbb{E}(x^\top\theta^\star)^4] &= \mathbb{E}[(\sum_{i=1}^d x[i]\theta^\star[i])^4] \\ &= \mathbb{E}\Bigg[\sum_{i=1}^d (x[i]\theta^\star[i])^4 + 4\sum_{i=1}^d\sum_{j\neq i}(x[i]\theta^\star[i])^3 x[j]\theta^\star[j] + 6\sum_{i=1}^d\sum_{j<i}(x[i]\theta^\star[i])^2(x[j]\theta^\star[j])^2 \\ &+ 12\sum_{i=1}^d\sum_{j\neq i}\sum_{k\neq i,k<j}(x[i]\theta^\star[i])^2 x[j]\theta^\star[j]x[k]\theta^\star[k] + 24\sum_{i=1}^d\sum_{j<i}\sum_{k<j}\sum_{l<k} x[i]\theta^\star[i]x[j]\theta^\star[j]x[k]\theta^\star[k]x[l]\theta^\star[l]\Bigg] \\ &= \sum_{i=1}^d \mathbb{E}[x[i]^4]\mathbb{E}[\theta^\star[i]^4] + 6\sum_{i=1}^d\sum_{j<i}\mathbb{E}[x_i^2]\mathbb{E}[x_j^2]\mathbb{E}[\theta^\star[i]^2]\mathbb{E}[\theta^\star[j]^2] \\ &= \frac{9(\zeta^2+1)^2}{d} + 6\binom{d}{2}\frac{(\zeta^2+1)^2}{d^2} \\ &= (\zeta^2+1)^2(\frac{9}{d} + \frac{3(d-1)}{d}) \\ &= (\zeta^2+1)^2(\frac{6}{d}+3)\end{aligned}$$

which gives that $\alpha = (\zeta^2+1)\sqrt{3+6/d}$

I. Computation complexity

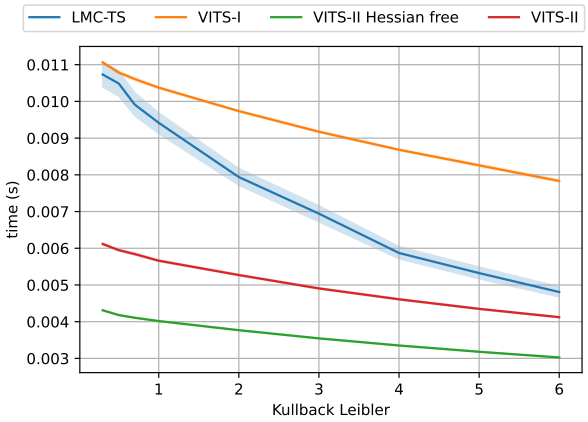
We conduct an experimental comparison between Langevin Monte Carlo (LMC) and three variants of Variational Inference, denoted as `VITS - I`, `VITS - II` and `VITS - II Hessian-free`, in approximating a specific target distribution. Our target distribution is a straightforward Gaussian distribution, represented as $p^\star = \mathrm{N}(\mu^\star, \Sigma^\star)$. At each iteration, we calculate the Kullback-Leibler distance between the approximated distribution and the target distribution. In this context, all

distributions generated by LMCTS, VITS − I, VITS − II and VITS − II Hessian-free take the form of Gaussians. To compute the mean and covariance matrix for LMC, we perform parameter averaging. As both the posterior and its approximation are Gaussians, the Kullback-Leibler divergence is easily tractable. Then, the training is stopped when

$$\mathrm{KL}(q_k, p^\star) \leq \epsilon$$

or if the number of steps exceeds 10000 steps.

The following Figure illustrates the relationship between the obtained Kullback- Leibler divergence and the computational time needed to achieve I. The computational time is the total time (in second) required to run all updating steps of the algorithm. This experiment is repeated across 1000 different seeds to compute the confidence interval. We decide not to compare with LinTS or LinUCB algorithms as they do not allow to approximate complex posteriors compared to LMCTS and VITS algorithms.



This figure shows that VITS − II and VITS − II Hessian-free are faster (in term of computational time) than LMCTS to obtain a certain Kullback-Leibler divergence. Note that VITS − I is the slowest algorithm, this is due to the costly inverse matrix calculation.

J. Computational Power

In this work, we use GPUs v100-16g or v100-32g for running our code with GPU Nvidia Tesla V100 SXM2 16 Go and CPUs with 192 Go per node.