Metalearning Linear Bandits by Prior Update

Amit Peleg
Viterbi Faculty of ECE
Technion, Israel
samit22@campus.technion.ac.il

Naama Pearl
University of Haifa, Israel
npearl@campus.haifa.ac.il

Ron Meir
Viterbi Faculty of ECE
Technion, Israel
rmeir@ee.technion.ac.il

Abstract

Fully Bayesian approaches to sequential decision-making assume that problem parameters are generated from a *known* prior. In practice, such information is often lacking. This problem is exacerbated in setups with partial information, where a misspecified prior may lead to poor exploration and performance. In this work we prove, in the context of stochastic linear bandits and Gaussian priors, that as long as the prior is sufficiently close to the true prior, the performance of the applied algorithm is close to that of the algorithm that uses the true prior. Furthermore, we address the task of learning the prior through metalearning, where a learner updates her estimate of the prior across multiple task instances in order to improve performance on future tasks. We provide an algorithm and regret bounds, demonstrate its effectiveness in comparison to an algorithm that knows the correct prior, and support our theoretical results empirically. Our theoretical results hold for a broad class of algorithms, including Thompson Sampling and Information Directed Sampling.

1 INTRODUCTION

Stochastic bandit problems involve sequential decision-making in the face of partial feedback, aiming to maximize cumulative reward or minimize regret gained over a series of interactions with the environment (for a comprehensive overview see Lattimore and Szepesvári (2020)). Bandit algorithms often differ in their prior knowledge about the nature of the rewards. In a frequentist setting, one assumes a reward distribution with fixed, but unknown parameters, while in a Bayesian setting, these parameters are generated from a known prior. While much effort has been devoted to devising effective algorithms with provably low regret in both settings, the situation is far less clear in a mixed setup, where the reward parameters are drawn from some *unknown* or partially known prior distribution. A particular challenge in this case is that exploration based on an incorrect prior assumption may lead an algorithm to waste resources by exploring irrelevant actions or, on the other hand, to disregard good ones (for earlier discussions of the influence of prior choice see Chapelle and Li (2011); Bubeck and Liu (2013); Honda and Takemura (2014); Liu and Li (2016)).

In the frequentist settings, for algorithmic reasons, some algorithms treat the parameters as if they arise from a prior distribution even though it does not reflect nature (Agrawal and Goyal, 2013; Abeille and Lazaric, 2017). Although these algorithms can be applied in the mixed setup, it is natural to expect improved performance when an adequate estimate of the prior exists. In this work we demonstrate in Theorem 1, for Gaussian prior distributions, that as long as the prior estimate is sufficiently accurate, the performance of an algorithm that uses the approximate prior is close to that of the same algorithm that uses the true prior. This analysis is challenging, since it compares two learning algorithms, both evolving throughout their interaction with the environment.

One natural approach to acquire a good prior estimation is based on metalearning. We study d-dimensional linear bandits in a metalearning setup where, at the beginning of each one of the N instances, each of duration T, a linear bandit task is sampled from an unknown prior distribution. The meta-learner maintains a continually updated meta-prior estimator across instances, and uses it as a prior for each instance. Then, within an instance, she selects actions in pursuance

Proceedings of the 25$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

of maximizing accumulated rewards, based on an updated within-instance posterior. We provide an explicit algorithm and establish regret bounds with respect to (WRT) the algorithm that knows the prior.

The main contributions of this work are the following:

- In the single instance setting, we prove that when the prior deviation is small, an algorithm's regret is close, up to a multiplicative constant, to the regret of the same algorithm that uses the correct prior. This result holds even when the prior deviation is not restricted by a function of the instance duration as implied in previous works, e.g., (Bastani et al., 2021).

- We present a class of algorithms that can use *any* single-instance prior-based approach in a metalearning setup to derive regret bounds with $\tilde{\mathcal{O}}(\sqrt{NT})$ regret, as opposed to previous results with $\tilde{\mathcal{O}}(\sqrt{NT^{\alpha}})$ regret, $\alpha \geq 3/2$. As far as we are aware, our results provide the first regret bounds of order $\tilde{\mathcal{O}}(\sqrt{NT})$ when *both* the prior mean and covariance are unknown.

- Technically, we develop a two-stage approach to compare algorithms using different priors, and hence *different* actions along the run. This significantly reduces the time-dependence of the regret bounds, and allows us to deal with the uncertainty in both the mean and covariance of the prior (See Table 1).

- We demonstrate empirically the importance of meta-prior learning in general, and covariance estimation in particular.

2 PRELIMINARIES AND SETTING

We use the following convention: variables appear with small letters x, vectors with capital letters X and matrices with bold capital letters $\mathbf{X}$. For $X \in \mathbb{R}^d, \mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$, $\|X\|_p$ is the l_p norm, $\|X\|$ is the l_2 norm and $\|\mathbf{A}\|_{\text{op}}$ is the l_2 operator norm. The smallest and largest eigenvalues of a matrix $\mathbf{A}$ are $\lambda_{\min}(\mathbf{A}), \lambda_{\max}(\mathbf{A})$ and $\mathbf{A} \succeq \mathbf{B}$ represents that $\mathbf{A}-\mathbf{B}$ is PSD. The unique square root of a PSD matrix $\mathbf{A}$ is $\mathbf{A}^{1/2}$. We introduce the notation $\mathbf{A}[\mathbf{B}]$, when we wish to emphasize that $\mathbf{A}$ is a function of a matrix $\mathbf{B}$. This notation is used for vectors and variables as well. The set $\{1,\ldots,n\}$ is denoted by $[n]$ for $n \in \mathbb{N}$ and the indicator function is denoted by $\mathbb{1}\{\cdot\}$. Finally $\tilde{\mathcal{O}}$ represents the $\mathcal{O}$ notation up to polylogarithmic factors and so does $\tilde{\Omega}$ and Ω.

2.1 Setting and Assumptions

We consider a metalearning problem where a learner interacts with N instances sequentially. At the start of each instance $n \in [N]$, a random vector $\theta_n \in \mathbb{R}^d$ is sampled from a multivariate Gaussian distribution $\mathcal{N}(\mu_*, \mathbf{\Sigma}_*)$ with unknown parameters. At each time $t \in [T]$, the learner chooses an action $A_{n,t} \in \mathbb{R}^d$ from a subset of available actions $\mathcal{A}_{n,t}$ presented to her and receives a reward $x_{n,t}[A_{n,t}] = A_{n,t}^{\top}\theta_n + \xi_{n,t}$, where $\xi_{n,t}$ is a noise term sampled independently from a known distribution $\mathcal{N}(0,\sigma^2)$. We also define the vector $X_{n,t} = [x_{n,1},\ldots,x_{n,t}]$ and the matrix $\mathbf{A}_{n,t}$, which is formed by concatenating the vectors $\left\{A_{n,s}^{\top}\right\}_{s=1}^{t}$ in its rows.

The following technical assumptions are required for the proofs. We first define $B_a(0)$ as a d-dimensional ball of radius a centered at 0 and the density function $f_A(A) \triangleq \tilde{f}_A(A)\,\mathbb{1}\{\|A\| \leq a\}/Z_a$ for some function $\tilde{f}_A$ and an appropriate normalization constant Z_a.

Assumption 1. *The set of actions can be either deterministic, $\mathcal{A}_{n,t} = B_a(0)$, or a set of actions of any size, $\mathcal{A}_{n,t} \subset B_a(0)$, each of which is sampled i.i.d. from a distribution f_A with a covariance matrix whose minimal eigenvalue is lower bounded by a known constant $\lambda_{\min}(\mathbf{\Sigma}_{\mathcal{A}}) \geq \underline{\lambda}_{\mathbf{\Sigma}_{\mathcal{A}}} > 0$. The function $\tilde{f}_A(A)$, can be either a zero mean Gaussian distribution or one which satisfies monotonicity, i.e., for every $\|A_1\| \leq \|A_2\|$ in the support of f_A, $\tilde{f}_A(A_1) \geq \tilde{f}_A(A_2)$.*

Assumption 2. *The minimal and maximal eigenvalues of the prior covariance matrix are lower and upper bounded by known constants, $\lambda_{\min}(\mathbf{\Sigma}_*) \geq \underline{\lambda}_{\mathbf{\Sigma}_*} > 0$, $\lambda_{\max}(\mathbf{\Sigma}_*) \leq \bar{\lambda}_{\mathbf{\Sigma}_*}$.*

Assumption 3. *The norm of the prior mean is upper bounded by a known constant, $\|\mu_*\| \leq m$.*

Regarding Assumption 1, only the boundedness of the actions is necessary during all time-steps, while the monotonicity and the eigenvalues bound are used just during the exploration steps of the algorithm.

2.2 QB$_\tau$ Algorithms and Regret Definition

While optimal Bayesian approaches operate by an exact computation of predictive distributions, we consider algorithms that work with posterior estimates, and which are not committed to Bayesian optimality. We refer to such algorithms as *Quasi-Bayesian* (QB), including, for example, Thompson Sampling (TS) (Thompson, 1933; Russo and Van Roy, 2014), and Information Directed Sampling (IDS) (Russo and Van Roy, 2018). The Bayesian regret of a QB algorithm that uses a prior $\mathcal{N}(\mu_n, \mathbf{\Sigma}_n)$ in the n_{th} instance is defined WRT an oracle that chooses at each step the action that yields the highest expected reward, i.e.,

$A^*_{n,t} = \text{argmax}_{A_{n,t} \in \mathcal{A}_{n,t}} A_{n,t}^\top \theta_n$,

$$\mathbb{E}\left[\mathrm{R}^*_{\mathrm{QB}}\left(\mu_n, \mathbf{\Sigma}_n, T\right)\right] \triangleq \sum_{t=1}^{T} \mathbb{E}\left[x_{n,t}\left[A^*_{n,t}\right] - x_{n,t}\left[A_{n,t}\right]\right]. \tag{1}$$

The expectation is taken over the prior used by the learner, which may be random, due to previous observations, the realization of θ_n, the actions that were presented during the instance, the randomness of the algorithm and the received noise terms.

We analyze algorithms that use the first τ steps of each instance to explore the actions uniformly at random in order to gain information. We refer to such algorithms as QB_τ (for example TS_τ, IDS_τ) and select τ so as to minimally affect the regret.

For clarification, there are three degrees of knowledge in this problem setup. The highest one is direct knowledge of the realization of each instance $\{\theta_n\}_{n=1}^N$. The oracle which knows these realizations always chooses the best actions $A^*_{n,t}$ and does not need to learn anything in the environment. The second level is knowledge of the prior. We use the term KQB_τ for the special version of each QB_τ algorithm that *knows* the true prior and denote its actions by $A^{\mathrm{K}}_{n,t}$. Such algorithms attempt to learn the realization of θ_n within the instance, but do not need to learn the meta environment between the instances. Hence their regret scales linearly in the number of instances N. This type of algorithm is the one usually analyzed under the Bayesian setting, e.g., (Russo and Van Roy, 2014, 2018). The last level of knowledge includes general QB_τ algorithms that are unaware of the prior and the realizations and may learn both within and between instances.

The regret of a QB_τ algorithm incurred by the incorrect prior is defined WRT KQB_τ and essentially measures the cost of 'not knowing' the true prior. We refer to it as the *relative regret*,

$$\begin{aligned}\mathbb{E}\left[\mathrm{R}^{\mathrm{K}}_{\mathrm{QB}_\tau}\left(\mu_n, \mathbf{\Sigma}_n, T\right)\right] &\triangleq \sum_{t=1}^{T} \mathbb{E}\left[x_{n,t}\left[A^{\mathrm{K}}_{n,t}\right] - x_{n,t}\left[A_{n,t}\right]\right] \\ &= \mathbb{E}\left[\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\mu_n, \mathbf{\Sigma}_n, T\right) - \mathrm{R}^*_{\mathrm{QB}_\tau}\left(\mu_*, \mathbf{\Sigma}_*, T\right)\right].\end{aligned} \tag{2}$$

Note that a naive approach that uses the same initial prior, without transferring knowledge between instances, yields a relative regret linear in N.

By rewriting (2), we can view the Bayesian regret of a QB_τ algorithm as a sum of the 'cost of not knowing the realization of θ_n when the prior is known' and the 'cost of not knowing the prior',

$$\begin{aligned}&\mathbb{E}\left[\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\mu_n, \mathbf{\Sigma}_n, T\right)\right] \\ &\quad = \mathbb{E}\left[\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\mu_*, \mathbf{\Sigma}_*, T\right) + \mathrm{R}^{\mathrm{K}}_{\mathrm{QB}_\tau}\left(\mu_n, \mathbf{\Sigma}_n, T\right)\right].\end{aligned} \tag{3}$$

For brevity, we now omit the index n until presenting the meta setting in Section 4. Given an assumed prior $\mathcal{N}(\mu, \mathbf{\Sigma})$ at the beginning of an instance, a QB_τ algorithm updates its posterior at time t based on the actions taken and the rewards received,

$$\begin{aligned}\mathbf{\Sigma}_t &= \left(\mathbf{\Sigma}^{-1} + \frac{1}{\sigma^2}\mathbf{V}_{t-1}\right)^{-1}, \\ \mu_t &= \mathbf{\Sigma}_t\left(\mathbf{\Sigma}^{-1}\mu + \frac{1}{\sigma^2}\mathbf{A}_{t-1}^\top X_{t-1}\right),\end{aligned} \tag{4}$$

for the Gram matrix $\mathbf{V}_t \triangleq \mathbf{A}_t^\top \mathbf{A}_t$. We remind the reader that $\mathbf{A}_t, X_t$ contain the actions and rewards up to time t respectively. The full derivation of the posterior calculation can be found in Appendix B.

Algorithm 1 presents a general scheme of a QB_τ algorithm. The specific mechanism of each algorithm is reflected in Line 6. For example, at time $t > \tau$, TS_τ samples from the posterior $\tilde{\theta} \sim \mathbb{P}(\mu_t, \mathbf{\Sigma}_t)$ and then plays the best action given that sample, $\text{argmax}_{A_t \in \mathcal{A}_t} A_t^\top \tilde{\theta}$.

Algorithm 1: $\mathrm{QB}_\tau(\mu, \mathbf{\Sigma}, \tau, \sigma)$

Inputs : μ, $\mathbf{\Sigma}$, τ, σ
Outputs: $\mathbf{A}_\tau, X_\tau$ // for meta estimation
1 **Initialization:** empty matrix $\mathbf{A}_0$ and vector X_0
2 **for** $t = 1, \ldots, T$ **do**
3 **if** $t \leq \tau$ **then** // within instance exploration
4 Sample A_t uniformly from $\mathcal{A}_t$, observe a reward x_t
5 **else**
6 Play A_t according to the specific algorithm scheme, observe a reward x_t
7 Concatenate the actions and rewards $\mathbf{A}_t \leftarrow \mathbf{A}_{t-1} \circ A_t$, $X_t \leftarrow X_{t-1} \circ x_t$
8 Update the posterior $\mathcal{N}(\mu_{t+1}, \mathbf{\Sigma}_{t+1})$ by (4)

3 SINGLE INSTANCE REGRET

Our main result, Theorem 1, bounds the relative regret (2) of *any* QB_τ algorithm in a Gaussian prior setting. In order to establish the result, we follow common practice in the bandit literature of dividing random events into the set of 'good events' and their complement, e.g., (Lattimore and Szepesvári, 2020). The former refers to situations where the various estimates are 'reasonably' close to their true or expected

values, and the latter is the complementary event that is shown to occur with low probability. The bulk of the proof consists of bounding the regret for the good event. In our setting, the good event $\mathcal{E}$ is defined as the intersection of four basic events for $\delta > 0$,

$$\begin{aligned}
\mathcal{E}_\theta &\triangleq \left\{ \left\| \boldsymbol{\Sigma}_*^{-1/2} (\theta - \mu_*) \right\|_\infty^2 \le 2 \ln \left(\frac{d^2 T}{\delta} \right) \right\}, \\
\mathcal{E}_v &\triangleq \left\{ \lambda_{\min} (\mathbf{V}_\tau) \ge \frac{\lambda_{\boldsymbol{\Sigma}_\mathcal{A}} d}{2} \right\}, \\
\mathcal{E}_m &\triangleq \left\{ \|\hat{\mu} - \mu_*\| \le \sqrt{f_m \delta} \right\}, \\
\mathcal{E}_s &\triangleq \left\{ \left\| \hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_* \right\|_{\text{op}} \le \sqrt{f_s \delta}, \quad \hat{\boldsymbol{\Sigma}} \succeq \boldsymbol{\Sigma}_* \right\}, \\
\mathcal{E} &\triangleq \{ \mathcal{E}_\theta \cap \mathcal{E}_v \cap \mathcal{E}_m \cap \mathcal{E}_s \}.
\end{aligned} \tag{5}$$

The event $\mathcal{E}_\theta$ is an instance-based event, unrelated to the performed algorithm, and represents the event that the realization of θ is not too far from its mean. The event $\mathcal{E}_v$ indicates that the QB_τ algorithm explores sufficiently in all directions during the exploration steps. The events $\mathcal{E}_m, \mathcal{E}_s$ represent the distance between the prior of the QB_τ algorithm $\mathcal{N}(\hat{\mu}, \hat{\boldsymbol{\Sigma}})$ and the true unknown prior $\mathcal{N}(\mu_*, \boldsymbol{\Sigma}_*)$. The arguments f_m and f_s, introduced in (5), quantify these distances. Moreover, the event $\mathcal{E}_s$ specifies that the estimated covariance is wider than the true covariance, reflecting the learner's lower level of certainty compared to an oracle that *knows* the true prior and thus prevents under-exploration. This issue can be also realized from a Bayesian point of view, where in the case that both the mean and covariance are unknown, the posterior mean distribution is broader compared to the case that only the mean is unknown (see section 4.6 in Murphy (2012)). The arguments f_m, f_s, as well as τ, may depend on the dimension and the horizon, and may also depend logarithmically on $1/\delta$.

Theorem 1. *Let $\theta \sim \mathcal{N}(\mu_*, \boldsymbol{\Sigma}_*)$ and let $\mathcal{N}(\hat{\mu}, \hat{\boldsymbol{\Sigma}})$ be the prior of a QB_τ algorithm. For $\tau < T$, if for some $0 < \delta \le 1/M$ the event $\mathcal{E}$ holds with probability larger than $1 - \frac{9\delta}{dT}$, then the relative regret is bounded by,*

$$\underbrace{\mathbb{E}\left[\text{R}^{\text{K}}_{\text{QB}_\tau} \left(\hat{\mu}, \hat{\boldsymbol{\Sigma}}, T \right) \right]}_{\substack{\text{cost of not knowing} \\ \text{the prior}}} \le k_1 \cdot \underbrace{\mathbb{E}\left[\text{R}^*_{\text{QB}_\tau} \left(\mu_{*,\tau+1}, \boldsymbol{\Sigma}_{*,\tau+1}, T - \tau \right) \right]}_{\substack{\text{cost of not knowing the realization of } \theta \\ \text{when the prior is known}}} + \underbrace{\frac{c_{\text{bad}} \delta}{\sqrt{d}}}_{\substack{\text{bad} \\ \text{event}}},$$

where $M \in \tilde{\mathcal{O}}\left(f_m + \tau^2 f_s \right)$, $k_1 \in \tilde{\mathcal{O}}\left(\sqrt{f_m \delta} + \tau \sqrt{f_s \delta} \right)$.

The definitions of M, k_1 and c_{bad} are in (21) in Appendix C as well as further details. The relationship between the performance of the algorithm and the initial prior deviation in the events $\mathcal{E}_m$ and $\mathcal{E}_s$ is represented by k_1. The term M ties δ to the arguments f_m and f_s, thus forcing the prior deviations to be small, and $c_{\text{bad}} \in \tilde{\mathcal{O}}(1)$ stems from the bad event.

An immediate consequence of Theorem 1 and (3) is a bound on the Bayesian regret (1) of any QB_τ algorithm,

$$\begin{aligned}
&\underbrace{\mathbb{E}\left[\text{R}^*_{\text{QB}_\tau} \left(\hat{\mu}, \hat{\boldsymbol{\Sigma}}, T \right) \right]}_{\substack{\text{cost of not knowing both the} \\ \text{realization of } \theta \text{ and the prior}}} \\
&\quad \le (1 + k_1)\, \mathbb{E}\left[\text{R}^*_{\text{QB}_\tau} \left(\mu_{*,\tau+1}, \boldsymbol{\Sigma}_{*,\tau+1}, T - \tau \right) \right] + \tilde{\mathcal{O}}(\tau).
\end{aligned} \tag{6}$$

Note that for $\tau \in \tilde{\mathcal{O}}(d)$, $\|\hat{\mu} - \mu_*\| \in \tilde{\mathcal{O}}(1)$ and $\left\| \hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_* \right\|_{\text{op}} \in \tilde{\mathcal{O}}(1/d)$, QB_τ is a $(1+\alpha)$-*approximation* of KQB_τ for some constant $\alpha > 0$ which is determined by the constants in τ, $\|\hat{\mu} - \mu_*\|$ and $\left\| \hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_* \right\|_{\text{op}}$. See Appendix C.6 for a concrete example.

Having bounded the regret of a QB_τ algorithm by the standard Bayesian regret of KQB_τ, we can leverage previous results for Bayesian algorithms. For example, proposition 6 and Lemma 7 in Lu and Van Roy (2019), adjusted to the Gaussian prior in Basu et al. (2021), bound the prior-dependent Bayesian regret for TS and a Bayesian version of UCB (Upper Confidence Bound) in the case of finite action spaces. Plugging this bound with $\delta = 1/T^2$ into (6) we get,

$$\begin{aligned}
&\mathbb{E}\left[\text{R}^*_{\text{QB}_\tau} \left(\hat{\mu}, \hat{\boldsymbol{\Sigma}}, T \right) \right] \\
&\quad \le (1 + k_1) \left[4 \sqrt{\frac{\bar{\lambda}_{\boldsymbol{\Sigma}_*} a^2}{\ln\left(1 + \frac{\bar{\lambda}_{\boldsymbol{\Sigma}_*} a^2}{\sigma^2} \right)} \ln\left(4 |\mathcal{A}| T^2 \right)} \right. \\
&\qquad \left. \times \sqrt{\frac{1}{2} dT \ln\left(1 + \frac{\bar{\lambda}_{\boldsymbol{\Sigma}_*} T}{\sigma^2} \right)} + \sqrt{2 \bar{\lambda}_{\boldsymbol{\Sigma}_*} a^2} \right] + \tilde{\mathcal{O}}(\tau).
\end{aligned} \tag{7}$$

An interesting implication of Theorem 1 is for the offline learning setup. With the increasing amount of data available, the opportunity arises to form more informative priors, which are guaranteed by the theorem to have the same regret (up to constants) as any KQB_τ algorithm in a single instance. Another implication is for sequential settings, where N instances are sampled from the same distribution one by one. We elaborate on the latter in Section 4 and show that the suggested meta-algorithm produces the conditions for the good event to hold with high probability.

Proof sketch The difficulty in bounding the regret based on the comparison between an algorithm that knows the prior (KQB_τ) and another that estimates it (QB_τ), is twofold. First, since the posteriors of both algorithms depend on the actions and the rewards throughout the instance, it is hard to track the distance between the posteriors as the instance progresses. Second, although regret bounds on TS with

a known prior are proved to be tighter as the prior is more informative (Russo and Van Roy, 2016; Dong and Van Roy, 2018), an improved bound does not ensure an actual improvement in the regret of the algorithm. Therefore, establishing low estimation error at the start or during the instance, does not suffice.

To establish a within-instance regret bound between QB_τ and KQB_τ, we adapt the idea of mean alignment from Bastani et al. (2021) and adjust it to cover covariance alignment as well. This analytic tool is used to cause the two algorithms to mathematically posses an identical posterior at a specific time and thus to behave identical (on average) until the end of the instance. Specifically, with a two stage technique, we use the randomness of the first τ exploration steps to align both the means and the covariance matrices at time $\tau+1$. Since the two compared algorithms start with different covariance matrices, they can only align if the learners would take *different* actions (see (4)). Practically, for every set of actions chosen by QB_τ with a certain probability, there is a nonzero probability for KQB_τ to choose the set of actions that would give rise to covariance alignment. This occurs due to the randomness in the actions selection and due to the action space properties in Assumption 1. The cost of analytically switching between these two probabilities in order to align the covariance matrices is termed *covariance alignment cost* and it is reduced as a function of the distance between the covariance matrices at the beginning of the instance. The covariance alignment is analyzed in Appendix C.2.1.

Yet, aligning the covariance matrices does not imply mean alignment as well, but rather leads to an alignment with a surrogate algorithm, that has started the interaction with the true covariance but with an incorrect mean. Since the mean is a function of the actions taken and the reward noise received, after using the randomness in the actions to align the covariance matrices, we can use the noise terms to align the different means (Appendix C.2.2) and thus to align the surrogate algorithm to KQB_τ at time $\tau+1$. See the full proof in Appendix C.

4 MQB_τ ALGORITHM

QB algorithms are designed to minimize the regret within a single instance, by refining their estimation of θ, while exploiting their knowledge, as the interaction with the instance proceeds. Using the same line of thought, MQB_τ aims to minimize the regret along multiple instances by learning the meta-prior, while using the improved prior to reduce the per-instance regret. Since the prior distribution is Gaussian, one may think to form MLE estimators for the mean and covariance prior to the n_{th} instance. However, it is inapplicable in the linear bandits environment, since the learner has no access to the true realizations of $\{\theta_j\}_{j=1}^{n-1}$. A simple approach would be to utilize the inner-instance estimation of the QB algorithm. This straightforward approach has two problems that the MQB_τ algorithm solves, using two levels of exploration.

The first problem rises from the adaptive nature of bandit algorithms, which leads to biased instance-estimators, as discussed in Shin et al. (2019). This in turn, would lead to an inconsistent meta-estimation of the prior. The solution is within-instance exploration. At the beginning of each instance the learner performs τ *exploration steps*, in which she chooses actions uniformly at random to ensure sufficient estimation of θ in all directions. The information from steps $\tau+1,\ldots,T$ is ignored during the meta-estimation to keep it unbiased. The inner mechanism of the QB_τ algorithm remains the same, i.e., all the actions taken during the instance participate in the inner estimation of θ. The number of exploration steps, defined so as to balance the regret incurred and the quality of the estimators, is set to

$$\tau = \max\left\{d, \frac{8a^2}{\lambda_{\mathbf{\Sigma}_\mathcal{A}}}\ln(d^2N^2T)\right\}. \tag{8}$$

The second problem is the limited amount of knowledge on the meta-prior during the first instances. Using the inaccurate meta-prior in these instances may result in poor performance as compared to an algorithm with frequentist guarantees which by its nature explores sufficiently in arbitrary environments. Hence, the MQB_τ algorithm uses a second level of exploration and during the first N_0 *exploration instances* gathers information on the environment without exploiting it yet. The number of exploration instances, derived from several requirements along the regret proof, is discussed in Section 5.

The full scheme of the MQB_τ algorithm is presented in Algorithm 2. Next, we elaborate on the main idea behind the prior estimation. We use the Ordinary Least Squares (OLS) estimator to obtain a meta estimation for θ_j in every instance,

$$\hat{\theta}_j = \left(\sum_{t=1}^{\tau} A_{j,t}A_{j,t}^\top\right)^{-1}\sum_{t=1}^{\tau} A_{j,t}x_{j,t} = \mathbf{V}_{j,\tau}^{-1}\mathbf{A}_{j,\tau}^\top X_{j,\tau}. \tag{9}$$

The estimator for the mean before the n_{th} instance, $\hat{\mu}_n$, uses these estimations from all previous instances,

$$\hat{\mu}_n = \frac{1}{n-1}\sum_{j=1}^{n-1}\hat{\theta}_j. \tag{10}$$

The bias-corrected MLE for the covariance before the beginning of the n_{th} instance would be $\frac{1}{n-2}\sum_{j=1}^{n-1}(\hat{\theta}_j -$

$\hat{\mu}_n)(\hat{\theta}_j - \hat{\mu}_n)^\top$. However, as we show in Appendix D.3, the estimation errors of $\{\theta_j\}_{j=1}^{n-1}$ cause it to converge to $\mathbf{\Sigma}_* + \frac{\sigma^2}{n-1}\sum_{j=1}^{n-1}\mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\right]$. In order to cancel out the added variance, we add a further term,

$$\hat{\mathbf{\Sigma}}_n = \frac{1}{n-2}\sum_{j=1}^{n-1}\left(\hat{\theta}_j - \hat{\mu}_n\right)\left(\hat{\theta}_j - \hat{\mu}_n\right)^\top - \mathbf{G}_\Sigma, \quad (11)$$

where $\mathbf{G}_\Sigma = \frac{\sigma^2}{n-1}\sum_{j=1}^{n-1}\mathbf{V}_{j,\tau}^{-1}$. Although this estimator is unbiased, in practice it can be wider or narrower than the true covariance, and, as explained in Section 3, we aim for the former. Hence we use a widened version of the covariance as suggested by Bastani et al. (2021) and proved in Lemma 25. Given the initial estimation $\hat{\mathbf{\Sigma}}_n$ and a confidence level $\left\|\hat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}_*\right\|_{\text{op}} \leq s$, the widened version is given by $\hat{\mathbf{\Sigma}}_n^{\text{w}} \triangleq \hat{\mathbf{\Sigma}}_n + s \cdot \mathbf{I}$, which ensures that $\hat{\mathbf{\Sigma}}_n^{\text{w}} \succeq \mathbf{\Sigma}_*$ with high probability. In Lemma 1 in Section 5 we show that the confidence level prior to the n_{th} instance dictates,

$$\hat{\mathbf{\Sigma}}_n^{\text{w}} = \hat{\mathbf{\Sigma}}_n + c_{\text{w}} \cdot \sqrt{\frac{5d + 2\ln(dnT)}{n-1}}\,\mathbf{I}, \quad (12)$$

where $c_{\text{w}} = 50\left(\frac{2\sigma^2}{\lambda_{\mathbf{\Sigma}_\mathcal{A}} d} + \bar{\lambda}_{\mathbf{\Sigma}_*}\right)$. To align with the MQB$_\tau$ scheme, we adjust the QB$_\tau$ algorithm to output the actions taken and the rewards received during the first τ steps. Any algorithm can be used in Line 3, as long as it is adapted to perform τ exploration steps and to return $\mathbf{A}_\tau, X_\tau$.

Algorithm 2: MQB$_\tau$

Inputs: N, T, $a, \underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}}, \underline{\lambda}_{\mathbf{\Sigma}_*}, \bar{\lambda}_{\mathbf{\Sigma}_*}$, m, σ

1 **Initialization:** set τ by (8)
2 **for** $n = 1, \ldots, N_0$ **do** // meta exp. instances
3 $(\mathbf{A}_{n,\tau}, X_{n,\tau}) \leftarrow$ Run any QB$_\tau$ algorithm with frequentist guarantees
4 Compute $\hat{\theta}_n$ by (9)
5 **for** $n = N_0 + 1, \ldots, N$ **do**
6 Update $\hat{\mu}_n$ by (10)
7 Update $\hat{\mathbf{\Sigma}}_n$ by (11) and $\hat{\mathbf{\Sigma}}_n^{\text{w}}$ by (12)
8 $(\mathbf{A}_{n,\tau}, X_{n,\tau}) \leftarrow \text{QB}_\tau\left(\hat{\mu}_n, \hat{\mathbf{\Sigma}}_n^{\text{w}}, \tau, \sigma\right)$
9 Compute $\hat{\theta}_n$ by (9)

5 MQB$_\tau$ REGRET

The meta algorithm consists of two key phases, as depicted in Figure 1:

1. *Within-instance* phase, where actions are taken based on the estimated meta-prior and on the within-instance updated posterior.
2. *Between-instance* phase, where the estimated meta-prior is updated based on information from previous instances.

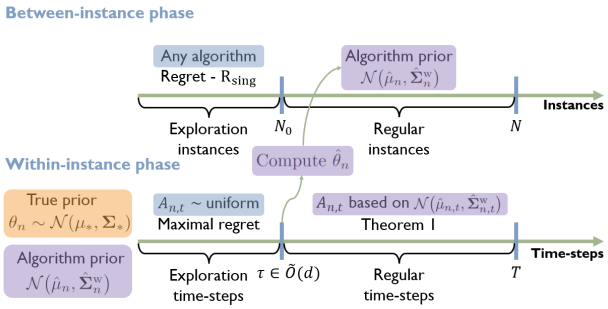

Figure 1: **MQB$_\tau$** algorithm and regret scheme. See Section 3 for the regret analysis in a single instance, and Section 4 for the complete algorithm scheme. R_{sing} is defined later in the section.

We address the first in Theorem 1, which bounds the per-instance relative regret given a bound on the deviations between the estimated and the true prior. The second, addressed in Lemma 1, explained below, demonstrates that as the number of instances increases, these prior deviations approach zero. Finally, Theorem 2 combines these two basic components in order to establish a regret bound for MQB$_\tau$ over N instances. In order to emphasize the instance dependence, we denote in this section several of the arguments with a subscript n.

Next, we present Lemma 1, which provides bounds on the distance between the prior constructed by MQB$_\tau$ and the true prior. In particular, we show that before the n_{th} instance, MQB$_\tau$ meets the events $\mathcal{E}_m, \mathcal{E}_s$ defined in (5) with $\delta_n \triangleq 1/(n-1)$ and closed-form expressions of $f_{m,n}, f_{s,n} \in \tilde{\mathcal{O}}(d)$ (see Appendix D.1). We denote this adjusted per instance *good event* by $\mathcal{E}_{n(\text{MQB}_\tau)}$. The adjusted formalization and the proof of Lemma 1 can be found in Appendix D.1, based on the mean estimation error (Appendix D.2) and the covariance estimation error (Appendix D.3).

Lemma 1. *(MQB$_\tau$ conditions) For every instance* $n > 10d + 4\ln(16dT)$, $\mathbb{P}\left(\mathcal{E}_{n(\text{MQB}_\tau)}\right) \geq 1 - 8/dnT$.

The expressions of $f_{m,n}, f_{s,n}, \delta_n$ and the expression of τ in (8) define M and k_1 in Theorem 1 as a function of n, i.e. $M_n, k_{1,n}$. We define the number of exploration instances for MQB$_\tau$ as $N_0 \triangleq \lceil M_{N+1} \rceil$, which ensures for every $n > N_0$ that $\delta_n < 1/M_n$. Having established the two components described at the beginning of this section, we can bound the regret incurred by the incorrect prior of MQB$_\tau$ using the following theorem. The single instance regret of the algorithm used during the first N_0 instances is denoted by R_{sing}.

Theorem 2. *For $N_0 \leq N$, the* MQB_τ *N-instance relative regret is bounded by,*

$$\sum_{n=1}^{N} \mathbb{E}\left[\mathrm{R}^{\mathrm{K}}_{\mathrm{QB}_\tau}\left(\hat{\mu}_n, \hat{\mathbf{\Sigma}}^{\mathrm{w}}_n, T\right)\right] \leq \underbrace{k_2\sqrt{N}\mathbb{E}\left[\mathrm{R}^{*}_{\mathrm{QB}_\tau}\left(\mu_{*,\tau+1}, \mathbf{\Sigma}_{*,\tau+1}, T-\tau\right)\right]}_{\substack{\text{regular}\\ \text{instances}}} + \underbrace{N_0 \mathrm{R}_{\mathrm{sing}}}_{\substack{\text{exploration}\\ \text{instances}}},$$

where $k_2 \in \tilde{\mathcal{O}}\left(d^{3/2}\right)$, $N_0 \in \tilde{\mathcal{O}}\left(d^3\right)$. The definitions of k_2 and N_0 are in (82) and the proof is in (83) in Appendix E.

During the first N_0 instances, MQB_τ suffers as much regret as the frequentist algorithm it chooses, for example when using the IDS algorithm of Kirschner and Krause (2018), $\mathrm{R}_{\mathrm{sing}} \in \tilde{\mathcal{O}}(d\sqrt{T})$.

An immediate consequence of Theorem 2 is a bound on the Bayesian regret (1) of the MQB_τ algorithm,

$$\sum_{n=1}^{N} \mathbb{E}\left[\mathrm{R}^{*}_{\mathrm{QB}}\left(\hat{\mu}_n, \hat{\mathbf{\Sigma}}^{\mathrm{w}}_n, T\right)\right] \leq \left(N + k_2\sqrt{N}\right)\mathbb{E}\left[\mathrm{R}^{*}_{\mathrm{QB}_\tau}\left(\mu_{*,\tau+1}, \mathbf{\Sigma}_{*,\tau+1}, T-\tau\right)\right] + \tilde{\mathcal{O}}\left(N_0\mathrm{R}_{\mathrm{sing}} + Nd\right). \tag{13}$$

The multiplicative factor $\left(N + k_2\sqrt{N}\right)$ is a sum of two parts; N, that stems from the inherent regret of the KQB_τ algorithm and cannot be avoided, and $k_2\sqrt{N}$ which represents the price of the 'prior alignment'. The second term on the right-hand-side of the inequality is the 'cost' of the two exploration levels. Both the 'prior alignment' and the exploration costs become vanishingly small WRT the inherent regret as N and T increase, implying negligible cost for 'not knowing the prior'.

Using the same prior-dependent bound as in (7), now with $\delta = 1/(NT)^2$, (13) can be further extended to a prior-dependent bound,

$$\sum_{n=1}^{N} \mathbb{E}\left[\mathrm{R}^{*}_{\mathrm{QB}}\left(\hat{\mu}_n, \hat{\mathbf{\Sigma}}^{\mathrm{w}}_n, T\right)\right] \leq \left(N + k_2\sqrt{N}\right)\left[\sqrt{\frac{8\bar{\lambda}_{\mathbf{\Sigma}_*}a^2\ln\left(4\left|\mathcal{A}\right|(NT)^2\right)}{\ln\left(1+\frac{\bar{\lambda}_{\mathbf{\Sigma}_*}a^2}{\sigma^2}\right)}}\times\sqrt{\ln\left(1+\frac{\bar{\lambda}_{\mathbf{\Sigma}_*}T}{\sigma^2}\right)dT}\right] + \tilde{\mathcal{O}}\left(N_0\mathrm{R}_{\mathrm{exp}} + Nd\right). \tag{14}$$

Note that a simple policy that runs all the tasks separately will incur regret of $\tilde{O}(Nd\sqrt{T})$ regardless of the "informativeness" of the correct prior. On the other hand, for the MQB_τ algorithm with TS as a subroutine, as the prior is more informative, the regret is lower, and in the extreme case ($\lambda_{\max}(\mathbf{\Sigma}_*) \to 0$ and N becomes large), only the inner exploration cost remains, i.e., $\tilde{O}(Nd)$.

6 RELATED WORK

While a significant amount of empirical and theoretical work has been devoted to metalearning in the domain of supervised learning (see recent review in Hospedales et al. (2021)), including methods based on prior update (Pentina and Lampert, 2014; Amit and Meir, 2018), there has been far less theoretical work on this topic in sequential decision-making problems (for a recent survey of algorithmic issues, see Ortega et al. (2019)).

We mention several works that deal with metalearning of stochastic bandits. Cella et al. (2020) consider linear bandits tasks drawn from a more general prior distribution, but assume a *known* variance. They establish *prior-dependent* regret bound for their proposed regularized optimism-based algorithm, similar to (7). However, our result is a consequence of the tighter bound in Theorem 1 that holds for every QB_τ algorithm relatively to its best scenario when the prior is known. Two recent papers that answer a question similar to Theorem 1 are Kveton et al. (2021) and Bastani et al. (2021), both suggest TS based meta-algorithms. The main difference between the approaches is the analysis technique, leading to the gap in the regret bounds, summarized in Table 1.

Kveton et al. (2021) focus on a fully Bayesian multi-armed bandits (MAB) setting, where tasks are drawn from a Gaussian prior. The prior is parameterized by a *known* scalar covariance and an unknown mean, that is itself drawn from a *known* hyper-prior. The authors derive a regret bound which depends on T as $\tilde{\mathcal{O}}(T^2)$. Our result preserves their linear dependence in the initial mean deviation, while keeping the same time dependence as the algorithm that knows the prior. When using TS, this leads to a worst-case regret whose T dependence is $\tilde{\mathcal{O}}(\sqrt{T})$. Note that in the setting of known covariance, it is possible to use our proof scheme and still achieve the same regret guarantee of $\tilde{\mathcal{O}}(\sqrt{T})$, even if we drop the somewhat restrictive action assumption (Assumption 1). Bastani et al. (2021) consider contextual linear bandits in a dynamic pricing setting. Their $\tilde{\mathcal{O}}(d^4\sqrt{N}T^{3/2})$ regret bound is effective after $N_0 \in \tilde{\mathcal{O}}(d^4T^2)$ instances, while we obtain $\tilde{\mathcal{O}}(d^{5/2}\sqrt{NT})$ regret, effective after $N_0 \in \tilde{\mathcal{O}}(d^3)$ instances in which the learner suffers regret of $\mathrm{R}_{\mathrm{sing}} \in \tilde{\mathcal{O}}(d\sqrt{T})$.

Finally, three very recent papers warrant mention. Basu et al. (2021) assume a fully Bayesian framework

Table 1: Comparison of worst-case relative regret bounds for state-of-the-art meta TS algorithms

	SETTING	ACTIONS ASSUMPTIONS	COVARIANCE ASSUMPTIONS	MEAN ASSUMPTIONS	RELATIVE REGRET
Kveton et al. (2021)	K-arms MAB	—irrelevant—	Known $c \cdot I$	Bayesian, known hyper-prior	$\tilde{\mathcal{O}}(K\sqrt{NT^2})$
Bastani et al. (2021)	Dynamic pricing[i]	known bounds on actions and eigenvalues (Assumption 1[ii])	Unknown, known bounds on eigenvalues (Assumption 2)	Unknown, known bound	$\tilde{\mathcal{O}}(d^4\sqrt{NT^{3/2}})$, $N_0 \in \tilde{\mathcal{O}}(d^4T^2)$
Ours, MQB$_\tau$[iii]	Linear bandits			(Assumption 3)	$\tilde{\mathcal{O}}(d^{5/2}\sqrt{NT})$, $N_0 \in \tilde{\mathcal{O}}(d^3)$

[i] Adapting the analysis to a linear bandits setting results in a reduction of $\tilde{\mathcal{O}}(\sqrt{d})$ from the regret.
[ii] In our work there exists an extra requirement on the distribution 'monotonicity'.
[iii] With an exploration algorithm for which $\mathrm{R}_{\mathrm{sing}} \in \tilde{\mathcal{O}}(d\sqrt{T})$, for example IDS (Kirschner and Krause, 2018).

where the covariance is *known* and the mean is sampled from a *known* Gaussian distribution. These assumptions allow the authors to elegantly expand the information theory analysis previously used in the single instance setup (Lu and Van Roy, 2019) to the new framework of multiple instances. However, relaxing the assumption of a known covariance within their Bayesian setting complicates their analysis significantly and was not pursued in their paper. They establish a *prior-dependent* regret bound whose worst-case dependence on T is $\tilde{\mathcal{O}}(\sqrt{T})$. Simchowitz et al. (2021) bound the single instance misspecification error for a wide class of priors and settings and achieve an upper-bound of $\tilde{\mathcal{O}}(\varepsilon T^2)$, where ε is the initial total-variation prior estimation error, while our bound from Theorem 1 is $\tilde{\mathcal{O}}(\sqrt{T})$. In addition, they derive a lower bound of $\tilde{\Omega}(\varepsilon T^2)$ for MAB with $T \ll |\mathcal{A}|$ ($|\mathcal{A}|$ is the number of actions). To the best of our knowledge, this is the only lower bound in the literature, and it is not applicable for most settings, including ours. For multiple instances, they derive a bound only for the MAB setting. Wan et al. (2021) studies a generalized version of a meta MAB environment, in which they allow the distribution to depend on task-specific features. Their algorithm uses TS in a Bayesian hierarchical model.

We briefly highlight differences in the proof techniques. Kveton et al. (2021) performs *history alignment*, focusing on the probability that the two algorithms have the same history. The alignment process separates each time-step into two events. *(i)* Both algorithms perform the same action and receive the same reward, hence have zero regret WRT each other. *(ii)* The algorithms perform different actions that violate the alignment, and therefore suffer a worst case regret of $\tilde{\mathcal{O}}(T)$ over the rest of the instance. Summation over the time-steps leads to regret of $\tilde{\mathcal{O}}(T^2)$. Bastani et al. (2021) first performs τ exploration steps, in which the two algorithms choose the *same* actions but receive different rewards due to noise, thus enabling the mean alignment. From this point, the proof continues using tools from importance sampling (Precup et al., 2000). We believe this technique has a shortcoming. While aligning the means in the first τ steps facilitates the analysis at time $\tau + 1$, the resulting posterior updates of the means do not render them equal in subsequent steps, even if the two algorithms choose the same actions, since the covariance matrices differ (see (4)). Our work aligns both the means and covariance matrices. This line of proof establishes at a specific time a full prior alignment at a single cost that scales with the distances between the priors, while the two other techniques are applied separately for each step, thus their per-step cost is multiplied by the horizon. These differences lead to a significant gap in the upper bounds.

7 EXPERIMENTS

We demonstrate the effectiveness of MQB$_\tau$ with TS as a subroutine (MTS$_\tau$) in a synthetic environment as in Kveton et al. (2021); Simchowitz et al. (2021), comparing it to several baselines. *(i)* TS algorithm that does not know the prior and uses a zero vector as μ and a diagonal covariance matrix $\mathbf{\Sigma}$ with $\lambda_{\max}(\mathbf{\Sigma}_*)$ in its diagonal (UKTS); *(ii)* TS algorithm that *knows* the correct mean and uses the above covariance $\mathbf{\Sigma}$ (KMTS); *(iii)* TS algorithm that *knows* the correct prior (KTS). None of the above perform any forced exploration. Other metalearning algorithms in the literature, which do not assume known covariance, mostly differ from our work in their settings and analysis. Adapting the algorithms to our setting with empirical adjustments results in an algorithm similar to ours. Therefore we do not use them as baselines.

We compare three versions of the algorithm. The first, Th-MTS$_\tau$, uses only the first τ steps in each instance to form the meta estimator as suggested by theory; the second, All-MTS$_\tau$, still performs the τ exploration steps which ensures an invertible Gram matrix $\mathbf{V}_{j,\tau}$,

however it uses the information gathered from all time-steps for the meta estimation; the third, All-MTS, is similar to All-MTS$_\tau$, but does not perform inner-instance exploration (exploration is only used towards the end of an instance in case that the Gram matrix $\mathbf{V}_{j,t}$ remains singular). Since in realistic environments the learner is often unaware of T and N, τ was adjusted to be the first time in each instance in which $\lambda_{\min}\left(\frac{1}{\sigma^2}\mathbf{V}_{j,t}\right) \geq 0.03$. We also set N_0 to be d^3 instead of the exact definition of $N_0 \in \tilde{\mathcal{O}}(d^3)$ and the covariance widening constant c_{w} to be 10 for Th-MTS$_\tau$ and 1 for the versions that use all samples.

We use a linear bandits framework with $d = 5$ and $N = 10{,}000$ instances all drawn from a Gaussian distribution $\mathcal{N}(\mu_*, \mathbf{\Sigma}_*)$, where $\mu_* = [2, 2, 2, 2, 2]$ and $\mathbf{\Sigma}_*$ is a non-diagonal covariance matrix, with ones along the diagonal and 0.8 elsewhere. The horizon is $T = 200$ and in each time-step, 20 actions are available to the learners, all sampled from a uniform distribution over an $a = 0.25$ radius ball. The reward observed by the learners is corrupted by a standard Gaussian noise $\mathcal{N}(0, 1)$.

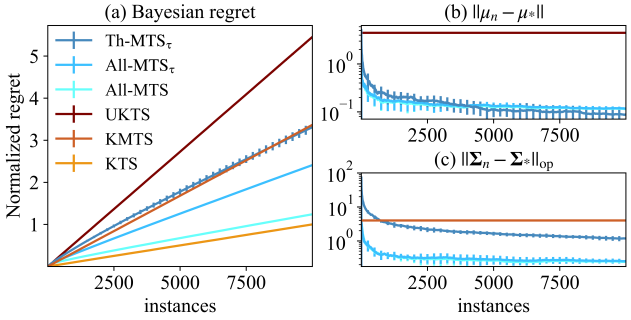

Figure 2: Comparing MQB$_\tau$ using TS as a sub-routine to several baselines, averaged over 10 runs with error bars equivalent to one std (hardly noticeable in the scale of the plots). The Bayesian regret is normalized by the KTS regret, such that its maximal regret is equal to 1. The prior convergence graphs ((b) and (c)) are in logarithmic scale.

As can be seen in Figure 2(a), all versions of the MTS$_\tau$ algorithm achieve better results than UKTS, indicating the importance of prior learning. Meta algorithms that assume a known covariance (Kveton et al., 2021; Basu et al., 2021) may achieve good results WRT KTS when their assumed covariance is correct. However in realistic environments when the covariance is unknown, KMTS represents their best scenario of estimating the mean alone. The results of the MTS$_\tau$ versions that reach and even outperform this oracle that knows the correct mean, demonstrate the significance of covariance learning. As can be seen in Figure 2(b) the mean of Th-MTS$_\tau$ approaches the true prior mean, in contrast to the other two versions that used all the samples, and incur the known bias of adaptive algorithms (Shin et al., 2019). However, due to scarcity of samples, the covariance convergence is slower (Figure 2(c)) and results in higher regret compared to All-MTS$_\tau$. This, and the additional 'cost' of exploration, as demonstrated by the gap between All-MTS$_\tau$ and All-MTS, suggest that empirical adjustments are needed for purely theoretically justified algorithms, perhaps using advances in bias reducing techniques, e.g., (Deshpande et al., 2018).

8 CONCLUSIONS

We presented algorithms and expected regret bounds for stochastic linear bandits where the expected rewards originate from a vector θ, sampled from a Gaussian distribution with unknown mean and covariance. For QB$_\tau$ algorithms with a good estimation of the prior, we derived single instance regret bounds, which are a multiplicative constant away from the regret of the algorithm that uses the true prior. For TS and IDS in the metalearning setup we established a relative regret of $\tilde{\mathcal{O}}(d^{5/2}\sqrt{NT})$ when using frequentist IDS during the exploration instances, a multiplicative improvement of $\tilde{\mathcal{O}}(d^{3/2}T)$ from previous results of $\tilde{\mathcal{O}}(d^4\sqrt{N}T^{3/2})$. Two limitations of our approach are the somewhat restrictive Assumption 1 and the need to compare with algorithms whose first τ within-instance steps are purely exploratory. We believe that Assumption 1 can be made more flexible without harming performance. Removing the initial τ exploratory steps, in a theoretically justifiable way, is left as an open problem.

Acknowledgements

We are grateful to Nadav Merlis for helpful discussions and insights. The work is partially supported by the Ollendorff Center of the Viterbi Faculty of Electrical and Computer Engineering at the Technion, and by the Skillman chair in biomedical sciences.

References

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, 2017.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, 2013.

Ron Amit and Ron Meir. Meta-learning by adjusting priors based on extended pac-bayes theory. In *International Conference on Machine Learning*, 2018.

Hamsa Bastani, David Simchi-Levi, and Ruihao Zhu. Meta dynamic pricing: Transfer learning across experiments. *Management Science*, 2021.

Soumya Basu, Branislav Kveton, Manzil Zaheer, and Csaba Szepesvári. No regrets for learning the prior in bandits. *Advances in Neural Information Processing Systems*, 2021.

Rajendra Bhatia. *Matrix analysis*. Springer, 1997.

Sébastien Bubeck and Che-Yu Liu. Prior-free and prior-dependent regret bounds for thompson sampling. *Advances in Neural Information Processing Systems*, 2013.

Leonardo Cella, Alessandro Lazaric, and Massimiliano Pontil. Meta-learning with stochastic linear bandits. In *International Conference on Machine Learning*, 2020.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in Neural Information Processing Systems*, 2011.

Yash Deshpande, Lester Mackey, Vasilis Syrgkanis, and Matt Taddy. Accurate inference for adaptive linear models. In *International Conference on Machine Learning*, 2018.

Shi Dong and Benjamin Van Roy. An information-theoretic analysis for thompson sampling with many actions. *Advances in Neural Information Processing Systems*, 2018.

William Feller. *An Introduction to Probability Theory and its Applications*. Wiley, New York, 1968.

Junya Honda and Akimichi Takemura. Optimality of thompson sampling for gaussian bandits depends on priors. In *Artificial Intelligence and Statistics*, 2014.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Timothy M Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J Storkey. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Daniel Hsu, Sham Kakade, Tong Zhang, et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, 2018.

Branislav Kveton, Mikhail Konobeev, Manzil Zaheer, Chih-wei Hsu, Martin Mladenov, Craig Boutilier, and Csaba Szepesvari. Meta-thompson sampling. In *International Conference on Machine Learning*, 2021.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Che-Yu Liu and Lihong Li. On the prior sensitivity of thompson sampling. In *International Conference on Algorithmic Learning Theory (ALT)*, 2016.

Xiuyuan Lu and Benjamin Van Roy. Information-theoretic confidence bounds for reinforcement learning. *Advances in Neural Information Processing Systems*, 2019.

Jan R Magnus and Heinz Neudecker. The commutation matrix: some properties and applications. *The Annals of Statistics*, pages 381–394, 1979.

Jan R Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 2019.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Pedro A Ortega, Jane X Wang, Mark Rowland, Tim Genewein, Zeb Kurth-Nelson, Razvan Pascanu, Nicolas Heess, Joel Veness, Alex Pritzel, Pablo Sprechmann, et al. Meta-learning of sequential strategies. *arXiv preprint arXiv:1905.03030*, 2019.

Anastasia Pentina and Christoph Lampert. A pac-bayesian bound for lifelong learning. In *International Conference on Machine Learning*, 2014.

Doina Precup, Richard S Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *International Conference on Machine Learning*, 2000.

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

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

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. Are sample means in multi-armed bandits positively or negatively biased? *Advances in Neural Information Processing Systems*, 2019.

Max Simchowitz, Christopher Tosh, Akshay Krishnamurthy, Daniel Hsu, Thodoris Lykouris, Miroslav Dudík, and Robert E Schapire. Bayesian decision-making under misspecified priors with applications to meta-learning. *Advances in Neural Information Processing Systems*, 2021.

W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 1933.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

Runzhe Wan, Lin Ge, and Rui Song. Metadata-based multi-task bandits with bayesian hierarchical models. *Advances in Neural Information Processing Systems*, 2021.

Shilin Zhan. On the determinantal inequalities. *Journal of Inequalities in Pure and Applied Mathematics*, 6(4):105, 2005.

Supplementary Material: Metalearning Linear Bandits by Prior Update

A SUMMARY OF NOTATION

Some additional notations we use in the appendix are the trace and the determinant of a matrix $\mathbf{A}$, $\mathrm{Tr}(\mathbf{A})$ and $\mathrm{Det}(\mathbf{A})$, respectively. We use $\|X\|_{\mathbf{A}} = \sqrt{X^\top \mathbf{A} X}$ for the weighted norm matrix of a PD matrix $\mathbf{A}$. We define the vector $\Xi_{n,t} = [\xi_{n,1}, \ldots, \xi_{n,t}]$ and often use $\int_E$ with an abuse of notation to indicate that we are integrating over all the terms that maintain the event E. Next, we summarize the notation used along the paper.

$\mathcal{A}_{n,t}$	The actions available to the learner at the n_{th} instance at time-step t.
a	Bound on the actions available to the learner.
$\mathbf{\Sigma}_{\mathcal{A}}$	The action covariance matrix.
$A^*_{n,t}$	The optimal action at the n_{th} instance at time-step t.
$A^{\mathrm{K}}_{n,t}$	The action taken by KQB_τ at the n_{th} instance at time-step t.
$A_{n,t}$	The action taken by the QB_τ learner at the n_{th} instance at time-step t.
$\mathbf{A}_{n,t}$	A matrix formed by concatenating the vectors $\left\{A^\top_{n,s}\right\}_{s=1}^{t}$ in it's rows.
$\mathbf{V}_{n,t}$	The Gram matrix $\sum_{t=1}^{T} A_{n,s} A^\top_{n,s} = \mathbf{A}^\top_{n,t} \mathbf{A}_{n,t}$.
$x_{n,t}$	The reward at the n_{th} instance at time-step t.
$X_{n,t}$	A vector containing all the rewards at the n_{th} instance up to time-step t.
$\xi_{n,t}$	The reward noise at the n_{th} instance at time-step t, sampled from $\mathcal{N}(0, \sigma^2)$.
$\Xi_{n,t}$	Vector containing all the reward noises at the n_{th} instance up to time-step t.
$S_{n,t}$	Summation of the action-noise terms $\sum_{t=1}^{T} A_{n,s}\xi_{n,s} = \mathbf{A}^\top_{n,s}\Xi_{n,s}$.
τ	The number of exploration time-steps taken in each instance.
θ_n	The realization of θ at the n_{th} instance.
$\hat{\theta}_n$	Meta approximation of θ_n using the rewards from the first τ time-steps.
ρ_n	The inner n_{th} instance error, $\hat{\theta}_n - \theta_n = \mathbf{V}^{-1}_{n,\tau} \sum_{s=1}^{\tau} A_{n,s}\xi_{n,s}$.
μ_*	The true unknown prior mean.
m	Bound on the prior mean, $\|\mu_*\| \leq m$.
$\hat{\mu}_n$	The prior mean of the learner for the n_{th} instance.
Δ_n	The difference between the realization of the instance and its mean. Can be viewed as it were sampled from $\mathcal{N}(0, \mathbf{\Sigma}_*)$.
$\mathbf{\Sigma}_*$	The true unknown prior covariance.
$\hat{\mathbf{\Sigma}}_n$	The estimated prior covariance for the n_{th} instance

$\hat{\mathbf{\Sigma}}_n^{\mathrm{w}}$	The widend prior covariance for the n_{th} instance.
$\mathbf{B}$	Represents the error between the covariance matrices that should be canceled during the covariance alignment phase.
$\mathbf{A}^{1/2}$	The unique square root of a PSD matrix $\mathbf{A}$.
$\mathbf{A}[\mathbf{B}]$	Represents the matrix $\mathbf{A}$ as a function of the matrix $\mathbf{B}$, used for vectors and scalars as well.
$\mathbf{A} \succeq \mathbf{B}$	Represents that $\mathbf{A} - \mathbf{B}$ is PSD.
$\|X\|_p$	The l_p norm.
$\|X\|$	The l_2 norm.
$\|X\|_{\mathbf{A}}$	The weighted norm matrix for a PD matrix $\mathbf{A}$, $\sqrt{X^\top \mathbf{A} X}$.
$\|\mathbf{A}\|_{\mathrm{op}}$	The l_2-operator norm.
$\lambda_j(\mathbf{A}), \sigma_j(\mathbf{A})$	The j_{th} eigenvalue and singular value of a matrix $\mathbf{A}$ respectively, arranged in a decreasing manner.
$\lambda_{\min}(\mathbf{A}), \lambda_{\max}(\mathbf{A})$	Smallest and largest eigenvalues of a matrix $\mathbf{A}$ respectively.
$\underline{\lambda}_{\mathbf{A}}, \bar{\lambda}_{\mathbf{A}}$	Bounds on the smallest and largest eigenvalues of a matrix $\mathbf{A}$ respectively.
$\mathrm{R}^*_{\mathrm{QB}_\tau}(\mu, \mathbf{\Sigma}, t)$	The regret of QB_τ with mean μ, covariance $\mathbf{\Sigma}$ and remaining horizon t WRT the optimal algorithm (an oracle that knows the true realization of θ_n).
$\mathrm{R}^{\mathrm{K}}_{\mathrm{QB}_\tau}(\mu, \mathbf{\Sigma}, t)$	The relative regret of QB_τ with mean μ, covariance $\mathbf{\Sigma}$ and remaining horizon t WRT the KQB_τ algorithm.

B WITHIN INSTANCE POSTERIOR CALCULATIONS

We recall a basic result from Bayesian statistics.

Lemma 2. *(Bayes rule for linear Gaussian systems - Theorem 4.4.1 in Murphy (2012))*

Suppose we have two variables, X and Y. Let $X \in \mathbb{R}^{D_x}$ be a hidden variable and $Y \in \mathbb{R}^{D_y}$ be a noisy observation of X. Let us assume we the the following prior and likelihood:

$$\mathbb{P}(X) = \mathcal{N}(\mu_X, \mathbf{\Sigma}_{\mathbf{X}}), \quad \mathbb{P}(Y \mid X) = \mathcal{N}(\mathbf{A}X + B, \mathbf{\Sigma}_{\mathbf{Y}}).$$

The posterior $\mathbb{P}(X \mid Y)$ is given by the following:

$$\begin{aligned}
&\mathbb{P}(X \mid Y) = \mathcal{N}\left(\mu_{X|Y}, \mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}}\right), \\
&\mathbf{\Sigma}^{-1}_{\mathbf{X}|\mathbf{Y}} = \mathbf{\Sigma}^{-1}_{\mathbf{X}} + \mathbf{A}^\top \mathbf{\Sigma}^{-1}_{\mathbf{Y}} \mathbf{A}, \\
&\mu_{X|Y} = \mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}} \left(\mathbf{\Sigma}^{-1}_{\mathbf{X}} \mu_X + \mathbf{A}^\top \mathbf{\Sigma}^{-1}_{\mathbf{Y}} (Y - B)\right).
\end{aligned}$$

Given a prior $\mathcal{N}(\mu_n, \mathbf{\Sigma}_n)$ and using Lemma 2, the prior before choosing an action at time-step t is,

$$\begin{aligned}
\mathbf{\Sigma}_{n,t} &= \left(\mathbf{\Sigma}^{-1}_{n,t-1} + \frac{1}{\sigma^2} A_{n,t-1} A^\top_{n,t-1}\right)^{-1} \\
&= \left(\mathbf{\Sigma}^{-1}_{n} + \frac{1}{\sigma^2} \sum_{s=1}^{t-1} A_{n,s} A^\top_{n,s}\right)^{-1} \\
&= \left(\mathbf{\Sigma}^{-1}_{n} + \frac{1}{\sigma^2} \mathbf{A}^\top_{n,t-1} \mathbf{A}_{n,t-1}\right)^{-1} \\
&= \left(\mathbf{\Sigma}^{-1}_{n} + \frac{1}{\sigma^2} \mathbf{V}_{n,t-1}\right)^{-1},
\end{aligned} \tag{15}$$

$$\begin{aligned}\mu_{n,t} &= \mathbf{\Sigma}_{n,t}\left(\mathbf{\Sigma}_{n,t-1}^{-1}\mu_{n,t-1} + \frac{1}{\sigma^2}A_{n,t-1}x_{n,t-1}\right)\\ &= \mathbf{\Sigma}_{n,t}\left(\mathbf{\Sigma}_{n}^{-1}\mu_{n} + \frac{1}{\sigma^2}\mathbf{A}_{n,t-1}^{\top}X_{n,t-1}\right)\\ &= \mathbf{\Sigma}_{n,t}\left(\mathbf{\Sigma}_{n}^{-1}\mu_{n} + \frac{1}{\sigma^2}\mathbf{V}_{n,t-1}\theta_n + \frac{1}{\sigma^2}\sum_{s=1}^{t-1}A_{n,s}\xi_{n,s}\right)\\ &= \mathbf{\Sigma}_{n,t}\left(\mathbf{\Sigma}_{n}^{-1}\mu_{n} + \frac{1}{\sigma^2}\mathbf{V}_{n,t-1}\theta_n + \frac{1}{\sigma^2}\mathbf{A}_{n,t-1}^{\top}\Xi_{n,t-1}\right).\end{aligned} \tag{16}$$

Specifically, for MQB_τ, the meta-prior is $\mathcal{N}\left(\hat{\mu}_n, \hat{\mathbf{\Sigma}}_n^{\mathrm{w}}\right)$, hence the inner-instance posterior is,

$$\hat{\mathbf{\Sigma}}_{n,t}^{\mathrm{w}} = \left(\left(\hat{\mathbf{\Sigma}}_n^{\mathrm{w}}\right)^{-1} + \frac{1}{\sigma^2}\mathbf{V}_{n,t-1}\right)^{-1}, \tag{17}$$

$$\hat{\mu}_{n,t} = \hat{\mathbf{\Sigma}}_{n,t}^{\mathrm{w}}\left(\left(\hat{\mathbf{\Sigma}}_n^{\mathrm{w}}\right)^{-1}\hat{\mu}_n + \frac{1}{\sigma^2}\mathbf{V}_{n,t-1}\theta_n + \frac{1}{\sigma^2}\mathbf{A}_{n,t-1}^{\top}\Xi_{n,t-1}\right). \tag{18}$$

For KQB_τ, the prior is $\mathcal{N}(\mu_*, \mathbf{\Sigma}_*)$ and the inner-instance posterior is,

$$\mathbf{\Sigma}_{*,t} = \left(\mathbf{\Sigma}_*^{-1} + \frac{1}{\sigma^2}\mathbf{V}_{n,t-1}^{\mathrm{K}}\right)^{-1}, \tag{19}$$

$$\mu_{*,t} = \mathbf{\Sigma}_{*,t}\left(\mathbf{\Sigma}_*^{-1}\mu_* + \frac{1}{\sigma^2}\mathbf{V}_{n,t-1}^{\mathrm{K}}\theta_n + \frac{1}{\sigma^2}\left(\mathbf{A}_{n,t-1}^{\mathrm{K}}\right)^{\top}\Xi_{n,t-1}^{\mathrm{K}}\right). \tag{20}$$

C SINGLE INSTANCE REGRET PROOF

In this section we prove Theorem 1, which bounds the regret incurred by the incorrect prior within a single instance. We do so by decomposing the regret (Appendix C.1) into the good and bad events defined in (5). By (85) in Lemma 29 and a union bound argument, we have that $\mathbb{P}(\mathcal{E}_\theta) > 1 - \frac{\delta}{dT}$. Thus, if the QB_τ algorithm maintains $\mathbb{P}(\mathcal{E}_v \cap \mathcal{E}_m \cap \mathcal{E}_s) \geq 1 - \frac{8\delta}{dT}$ the conditions for Theorem 1 hold.

Theorem 1. *Let $\theta \sim \mathcal{N}(\mu_*, \mathbf{\Sigma}_*)$ and let $\mathcal{N}(\hat{\mu}, \hat{\mathbf{\Sigma}})$ be the prior of a QB_τ algorithm. For $\tau < T$, if for some $0 < \delta \leq 1/M$ the event $\mathcal{E}$ holds with probability larger than $1 - \frac{9\delta}{dT}$, then the relative regret is bounded by,*

$$\underbrace{\mathbb{E}\left[\mathrm{R}_{\mathrm{QB}_\tau}^{\mathrm{K}}\left(\hat{\mu}, \hat{\mathbf{\Sigma}}, T\right)\right]}_{\substack{\text{cost of not knowing}\\ \text{the prior}}} \leq k_1 \cdot \underbrace{\mathbb{E}\left[\mathrm{R}_{\mathrm{QB}_\tau}^{*}\left(\mu_{*,\tau+1}, \mathbf{\Sigma}_{*,\tau+1}, T-\tau\right)\right]}_{\substack{\text{cost of not knowing the realization of } \theta\\ \text{when the prior is known}}} + \underbrace{\frac{c_{\mathrm{bad}}\delta}{\sqrt{d}}}_{\substack{\text{bad}\\ \text{event}}},$$

$$\begin{aligned}M &= \max\left\{3, c_s^2\tau^2 f_s, 18c_\xi^2 c_s\left(f_m + \left(c_1 d + c_\xi^2 c_s/36\right) f_s\right)\right\},\\ k_1 &= 12\sqrt{c_\xi^2 c_s}\sqrt{f_m\delta} + \left(c_s\tau + 12\sqrt{c_\xi^2 c_s c_1 d} + 2c_\xi^2 c_s\right)\sqrt{f_s\delta},\\ c_s &= \frac{2\sigma^2}{\lambda_{\mathbf{\Sigma}_*}^2 \lambda_{\mathbf{\Sigma}_\mathcal{A}}}, \quad c_\xi = \sigma\sqrt{5\ln\left(\frac{dT}{\delta}\right)}, \quad c_1 = \frac{2}{\lambda_{\mathbf{\Sigma}_*}}\ln\left(\frac{d^2T}{\delta}\right), \quad c_{\mathrm{bad}} = 22a\left(m + \sqrt{4\bar{\lambda}_{\mathbf{\Sigma}_*}\ln\left(\frac{d^2T}{\delta}\right)}\right).\end{aligned} \tag{21}$$

Note The expression for M depends polylogarithmicly on $1/\delta$, which in turn has to satisfy $0 < \delta \leq 1/M$, leading to an implicit inequality for δ. We show in Appendix C.3 that there exist $\tilde{M} \geq M$, independent of δ, such that $0 < \delta \leq 1/\tilde{M} \leq 1/M$ is well defined, while maintaining the same asymptotic behavior.

The expectation in the regret analysis includes all sources of randomness in the problem: the prior of the QB_τ algorithm at the start of the instance, the realization of θ, the actions that were presented to the learners during the instance, the randomness of the algorithms and the received noises. It is worth mentioning that in this work both the actions and the noises can differ between the algorithms compared. In most sections we abbreviate some or all of the notations to improve readability.

$$\mathbb{E} = \underset{\hat{\mu}}{\mathbb{E}}\underset{\hat{\mathbf{\Sigma}}}{\mathbb{E}}\underset{\theta}{\mathbb{E}}\underset{\mathcal{A}_\tau}{\mathbb{E}}\underset{\mathcal{A}_\tau^{\text{K}}}{\mathbb{E}}\underset{\mathbf{A}_\tau}{\mathbb{E}}\underset{\mathbf{A}_\tau^{\text{K}}}{\mathbb{E}}\underset{\Xi_\tau}{\mathbb{E}}\underset{\Xi_\tau^{\text{K}}}{\mathbb{E}}\mathbb{E}.$$

Intuition on Theorem 1

In order to provide intuition on Theorem 1 we analyze the demand $0 < \delta \leq 1/M$. To do so, we start from the equivalent demand, $0 < \delta M \leq 1$ and derive a stricter version of it. Plugging the definition of M and rearranging,

$$0 < \max\left\{3\delta, c_s^2\tau^2 f_s\delta, 18c_\xi^2 c_s\left(f_m\delta + \left(c_1 d + c_\xi^2 c_s/36\right) f_s\delta\right)\right\} \leq 1.$$

Splitting into two demands,

$$0 < \max\left\{c_s^2\tau^2 f_s\delta, 18c_\xi^2 c_s\left(f_m\delta + \left(c_1 d + c_\xi^2 c_s/36\right) f_s\delta\right)\right\} \leq 1; \quad 0 < \delta \leq 1/3.$$

Focusing on the first demand, since all the terms are positive, using $\max\{a,b\} \leq a+b$ yields the stricter demand,

$$18c_\xi^2 c_s f_m\delta + \left(c_s^2\tau^2 + 18c_\xi^2 c_s\left(c_1 d + c_\xi^2 c_s/36\right)\right) f_s\delta \leq 1.$$

Taking the square root, using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, and demanding a stricter condition,

$$\sqrt{18c_\xi^2 c_s}\sqrt{f_m\delta} + \left(c_s\tau + \sqrt{18c_\xi^2 c_s c_1 d} + \frac{1}{\sqrt{2}}c_\xi^2 c_s\right)\sqrt{f_s\delta} \leq 1.$$

From the similarity between the above expression and that for k_1 we can conclude the following. First, every δ that meets the demand, dictates k_1 to be bounded by a constant. Second, the initial prior deviations in (5) are bounded by $\sqrt{f_m\delta}$ and $\sqrt{f_s\delta}$, therefore k_1 and hence the relative regret have linear dependence in the initial prior deviations. Third, by Lemma 24, $\tau \in \tilde{\mathcal{O}}(d)$ meets the event $\mathcal{E}_v$, thus the demand holds for $\|\hat{\mu} - \mu_*\| \in \tilde{\mathcal{O}}(1)$ and $\left\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_*\right\|_{\text{op}} \in \tilde{\mathcal{O}}(1/d)$.

C.1 Single Instance Regret Decomposition

$$\begin{aligned}\mathbb{E}\left[\text{R}_{\text{QB}_\tau}^{\text{K}}\left(\hat{\mu}, \hat{\mathbf{\Sigma}}, T\right)\right] &= \mathbb{E}\left[\text{R}_{\text{QB}_\tau}^{\text{K}}\left(\hat{\mu}_{\tau+1}, \hat{\mathbf{\Sigma}}_{\tau+1}, T-\tau\right)\right] \\ &= \underbrace{\mathbb{E}\left[\text{R}_{\text{QB}_\tau}^{*}\left(\hat{\mu}_{\tau+1}, \hat{\mathbf{\Sigma}}_{\tau+1}, T-\tau\right)\right]}_{\text{QB}_\tau} - \underbrace{\mathbb{E}\left[\text{R}_{\text{QB}_\tau}^{*}\left(\mu_{*,\tau+1}, \mathbf{\Sigma}_{*,\tau+1}, T-\tau\right)\right]}_{\text{KQB}_\tau}.\end{aligned} \tag{22}$$

The first equality uses that the QB_τ algorithm does not incur regret during the exploration time-steps WRT KQB_τ since both algorithms choose actions randomly with the same distribution and for the same period of time. Decomposing the regret of QB_τ in (22) based on the event $\mathcal{E}$,

$$\mathbb{E}\left[\text{R}_{\text{QB}_\tau}^{*}\left(\hat{\mu}_{\tau+1}, \hat{\mathbf{\Sigma}}_{\tau+1}, T-\tau\right)\right] = \underbrace{\mathbb{E}\left[\text{R}_{\text{QB}_\tau}^{*}\left(\hat{\mu}_{\tau+1}, \hat{\mathbf{\Sigma}}_{\tau+1}, T-\tau\right)\mathbb{1}\left\{\bar{\mathcal{E}}\right\}\right]}_{\text{“Bad event”}} + \underbrace{\mathbb{E}\left[\text{R}_{\text{QB}_\tau}^{*}\left(\hat{\mu}_{\tau+1}, \hat{\mathbf{\Sigma}}_{\tau+1}, T-\tau\right)\mathbb{1}\left\{\mathcal{E}\right\}\right]}_{\text{“Good event”}}. \tag{23}$$

In (76) in Section C.5 we bound the regret incurred under the bad event. We state here the final result,

$$\mathbb{E}\left[\text{R}_{\text{QB}_\tau}^{*}\left(\hat{\mu}_{\tau+1}, \hat{\mathbf{\Sigma}}_{\tau+1}, T-\tau\right)\mathbb{1}\left\{\bar{\mathcal{E}}\right\}\right] \leq \frac{9c_{\text{bad}}\delta}{11\sqrt{d}}, \; c_{\text{bad}} \triangleq 22a\left(m + \sqrt{4\bar{\lambda}_{\mathbf{\Sigma}_*}\ln\left(\frac{d^2T}{\delta}\right)}\right). \tag{24}$$

Next, we bound the single instance regret under the good event.

C.2 Regret Incurred Under the "Good event"

Before the beginning of the instance, the learner possesses a prior mean μ and a prior covariance $\mathbf{\Sigma}$. Within the instance, this prior is being used by the learner, regardless its origin, whether it is an estimation, heuristics or comes from previous knowledge. At time-step t, given θ and these priors, the covariance matrix of the learner is only influenced by the actions taken (15), while the mean is influenced by both the actions and the noise terms received (16). Thus, we denote,

$$\mathbf{\Sigma}_t\left[\mathbf{A}_{t-1}\right]=\left(\mathbf{\Sigma}^{-1}+\frac{1}{\sigma^2}\mathbf{A}_{t-1}^\top\mathbf{A}_{t-1}\right)^{-1}=\left(\mathbf{\Sigma}^{-1}+\frac{1}{\sigma^2}\mathbf{V}_{t-1}\right)^{-1},$$
$$\mu_t\left[\mathbf{A}_{t-1},\Xi_{t-1}\right]=\mathbf{\Sigma}_t\left[\mathbf{A}_{t-1}\right]\left(\mathbf{\Sigma}^{-1}\mu+\frac{1}{\sigma^2}\mathbf{A}_{t-1}^\top\mathbf{A}_{t-1}\theta+\frac{1}{\sigma^2}\mathbf{A}_{t-1}^\top\Xi_{t-1}\right).$$

We adapt the notations for the specific case of QB_τ and KQB_τ summarized in (17), (18), (19), (20).

C.2.1 Covariance Alignment

The first step towards bounding the regret incurred under the "Good event" is the alignment of the covariance matrices of QB_τ and KQB_τ. For each set of actions $\mathbf{A}_\tau$, we define $\mathbf{A}_\tau^{\mathrm{K}}\left[\mathbf{A}_\tau\right]$ as a specific set of actions that may be taken by KQB_τ and brings (17) and (19) into equality. This requires,

$$\mathbf{V}_\tau^{\mathrm{K}}\left[\mathbf{A}_\tau\right]=\mathbf{V}_\tau-\mathbf{B}\quad;\quad\mathbf{B}\triangleq\sigma^2\left(\mathbf{\Sigma}_*^{-1}-\hat{\mathbf{\Sigma}}^{-1}\right).\tag{25}$$

Even though this requirement is not unique, we may choose a specific mapping between the two set of actions,

$$\mathbf{A}_\tau^{\mathrm{K}}\left[\mathbf{A}_\tau\right]\triangleq\mathbf{A}_\tau\left(\mathbf{A}_\tau^\top\mathbf{A}_\tau\right)^{-1/2}\left(\mathbf{A}_\tau^\top\mathbf{A}_\tau-\mathbf{B}\right)^{1/2}=\mathbf{A}_\tau\mathbf{V}_\tau^{-1/2}\left(\mathbf{V}_\tau-\mathbf{B}\right)^{1/2}.\tag{26}$$

We first prove that $\mathbf{V}_\tau\succ\mathbf{B}$ under the event $\mathcal{E}_v$ and for $\delta\leq 1/M$, thus the square root and the inverse exist and $\mathbf{A}_\tau^{\mathrm{K}}\left[\mathbf{A}_\tau\right]$ is well defined,

$$\lambda_{\min}\left(\mathbf{V}_\tau-\mathbf{B}\right)\underset{(a)}{\geq}\lambda_{\min}\left(\mathbf{V}_\tau\right)+\lambda_{\min}\left(-\mathbf{B}\right)\underset{(b)}{=}\lambda_{\min}\left(\mathbf{V}_\tau\right)-\left\|\mathbf{B}\right\|_{\mathrm{op}}\underset{(c)}{\geq}\frac{\lambda_{\mathbf{\Sigma}_{\mathcal{A}}}d}{2}-\frac{\sigma^2\sqrt{f_s\delta}}{\lambda_{\mathbf{\Sigma}_*}^2}\underset{(d)}{>}0,\tag{27}$$

where (a) uses Weyl's inequality, (b) uses that $\mathbf{B}$ is PSD by Lemma 33 and that for PSD matrices under the l_2-operator norm $\lambda_{\max}\left(\mathbf{A}\right)=\sigma_{\max}\left(\mathbf{A}\right)=\left\|\mathbf{A}\right\|_{\mathrm{op}}$, (c) uses $\mathcal{E}_v$ for the first term and Lemma 36, $\mathcal{E}_s$ for the latter and (d) uses that $\frac{\sigma^2\sqrt{f_s\delta}}{\lambda_{\mathbf{\Sigma}_*}^2}<\frac{\lambda_{\mathbf{\Sigma}_{\mathcal{A}}}d}{2}$ for $\delta\leq 1/M$.

For $\mathbf{A}_\tau^{\mathrm{K}}$ in the image of (26), $\mathbf{V}_\tau^{\mathrm{K}}\succ 0$, so we may define the inverse function,

$$\mathbf{A}_\tau\left[\mathbf{A}_\tau^{\mathrm{K}}\right]=\mathbf{A}_\tau^{\mathrm{K}}\left(\mathbf{V}_\tau^{\mathrm{K}}\right)^{-1/2}\left(\mathbf{V}_\tau^{\mathrm{K}}+\mathbf{B}\right)^{1/2}.\tag{28}$$

Since every action in the first τ time-steps is chosen independently from the previous actions, we may view the actions as they are drawn from the following distribution, $f_{\mathbf{A}}\left(\mathbf{A}\right)=\prod_{t=1}^{\tau}f_A(A_t)$. We continue from (23) by evaluating the integral of the good event over the action space. Since the actions in the first τ time-steps are independent of $\mathcal{E}_\theta,\mathcal{E}_m,\mathcal{E}_s$ we often omit them to improve readability.

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\hat{\mu}_{\tau+1},\hat{\mathbf{\Sigma}}_{\tau+1},T-\tau\right)\mathbb{1}\left\{\mathcal{E}\right\}\right]\\
&\quad=\mathbb{E}\int_{\mathcal{E}_v}f_{\mathbf{A}}\left(\mathbf{A}_\tau\right)\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\hat{\mu}_{\tau+1}\left[\mathbf{A}_\tau,\Xi_\tau\right],\hat{\mathbf{\Sigma}}_{\tau+1}\left[\mathbf{A}_\tau\right],T-\tau\right)d\mathbf{A}_\tau\\
&\quad\leq\mathbb{E}\Bigg[\underbrace{\max_{\mathcal{E}_v}\left\{\frac{f_{\mathbf{A}}\left(\mathbf{A}_\tau\right)}{f_{\mathbf{A}}\left(\mathbf{A}_\tau^{\mathrm{K}}\left[\mathbf{A}_\tau\right]\right)}\right\}}_{\text{Term A}}\cdot\underbrace{\int_{\mathcal{E}_v}f_{\mathbf{A}}\left(\mathbf{A}_\tau^{\mathrm{K}}\left[\mathbf{A}_\tau\right]\right)\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\hat{\mu}_{\tau+1}\left[\mathbf{A}_\tau,\Xi_\tau\right],\hat{\mathbf{\Sigma}}_{\tau+1}\left[\mathbf{A}_\tau\right],T-\tau\right)d\mathbf{A}_\tau}_{\text{Term B}}\Bigg],
\end{aligned}\tag{29}$$

where the inequality uses that the regret is non-negative.

Analyzing "Covariance alignment" - Term A

In order to bound term A, we first prove two auxiliary lemmas. The first lemma proves that for A_t in the support of f_A, A_t^{K} is in the support as well, where $A_t^\top, \left(A_t^{\mathrm{K}}\right)^\top$ are the t_{th} rows of the matrices $\mathbf{A}_\tau, \mathbf{A}_\tau^{\mathrm{K}}[\mathbf{A}_\tau]$, respectively.

Lemma 3. *For every $t \leq \tau$,*

$$\frac{\mathbb{1}\left\{\|A_t\| \leq a\right\}}{\mathbb{1}\left\{\left\|A_t^{\mathrm{K}}\right\| \leq a\right\}} \leq 1.$$

Proof.

$$\begin{aligned}\left\|A_t^{\mathrm{K}}\right\|^2 &= A_t^\top \mathbf{V}_\tau^{-1/2}\left(\mathbf{V}_\tau - \mathbf{B}\right)\mathbf{V}_\tau^{-1/2} A_t \\ &\underset{(a)}{=} \|A_t\|^2 - A_t^\top \mathbf{V}_\tau^{-1/2}\mathbf{B}\mathbf{V}_\tau^{-1/2} A_t \\ &\leq \|A_t\|^2\end{aligned}$$

Where (a) uses that $\mathbf{V}_\tau$ is PD and $\mathbf{B}$ is PSD by Lemma 33, thus $\mathbf{V}_\tau^{-1/2}\mathbf{B}\mathbf{V}_\tau^{-1/2}$ is PSD. □

The second lemma implies that the probability to sample the set of actions $\mathbf{A}_\tau$ is lower than the probability to sample $\mathbf{A}_\tau^{\mathrm{K}}[\mathbf{A}_\tau]$,

Lemma 4.

$$\prod_{t=1}^{\tau}\left(\frac{\tilde{f}_A(A_t)}{\tilde{f}_A\left(A_t^{\mathrm{K}}\right)}\right) \leq 1 \tag{30}$$

Proof. For $\tilde{f}_A(A)$, s.t. for every $\|A_1\| \leq \|A_2\|$ in the support of f_A, $\tilde{f}_A(A_1) \geq \tilde{f}_A(A_2)$ the result can be obtained directly from the proof of Lemma 3. For the case where $\tilde{f}_A = \mathcal{N}(0, \mathbf{\Sigma})$, for some general $\mathbf{\Sigma}$,

$$\begin{aligned}\prod_{t=1}^{\tau}\left(\frac{\tilde{f}_A(A_t)}{\tilde{f}_A\left(A_t^{\mathrm{K}}\right)}\right) &= \prod_{t=1}^{\tau}\left(\frac{\exp\left(-\frac{1}{2}A_t^\top \mathbf{\Sigma}^{-1}A_t\right)}{\exp\left(-\frac{1}{2}\left(A_t^{\mathrm{K}}\right)^\top \mathbf{\Sigma}^{-1}A_t^{\mathrm{K}}\right)}\right) \\ &= \frac{\exp\left(-\frac{1}{2}\sum_{t=1}^{\tau}A_t^\top \mathbf{\Sigma}^{-1}A_t\right)}{\exp\left(-\frac{1}{2}\sum_{t=1}^{\tau}\left(A_t^{\mathrm{K}}\right)^\top \mathbf{\Sigma}^{-1}A_t^{\mathrm{K}}\right)} \\ &\underset{(a)}{=} \frac{\exp\left(-\frac{1}{2}\sum_{t=1}^{\tau}\mathrm{Tr}\left(A_t^\top \mathbf{\Sigma}^{-1}A_t\right)\right)}{\exp\left(-\frac{1}{2}\sum_{t=1}^{\tau}\mathrm{Tr}\left(\left(A_t^{\mathrm{K}}\right)^\top \mathbf{\Sigma}^{-1}A_t^{\mathrm{K}}\right)\right)} \\ &\underset{(b)}{=} \frac{\exp\left(-\frac{1}{2}\mathrm{Tr}\left(\sum_{t=1}^{\tau}\mathbf{\Sigma}^{-1}A_tA_t^\top\right)\right)}{\exp\left(-\frac{1}{2}\mathrm{Tr}\left(\sum_{t=1}^{\tau}\mathbf{\Sigma}^{-1}A_t^{\mathrm{K}}\left(A_t^{\mathrm{K}}\right)^\top\right)\right)} \\ &= \frac{\exp\left(-\frac{1}{2}\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{V}_\tau\right)\right)}{\exp\left(-\frac{1}{2}\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{V}_\tau^{\mathrm{K}}\right)\right)} \\ &\underset{(c)}{=} \exp\left(-\frac{1}{2}\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{B}\right)\right) \\ &\underset{(d)}{\leq} 1,\end{aligned} \tag{31}$$

where (a) applies trace on a scalar, (b) uses $\mathrm{Tr}\left(A^\top B\right) = \mathrm{Tr}\left(BA^\top\right)$ and the linearity of the trace, (c) uses the definition of $\mathbf{V}_\tau^{\mathrm{K}}$ in (25) and (d) uses Lemma 32. □

Term A is finally bounded by Lemma 3 and Lemma 4,

$$\max_{\mathcal{E}_v}\left\{\frac{f_{\mathbf{A}}(\mathbf{A}_\tau)}{f_{\mathbf{A}}\left(\mathbf{A}_\tau^{\mathrm{K}}[\mathbf{A}_\tau]\right)}\right\} = \max_{\mathcal{E}_v}\left\{\prod_{t=1}^{\tau}\left(\frac{\tilde{f}_A(A_t)}{\tilde{f}_A\left(A_t^{\mathrm{K}}\right)}\frac{\mathbb{1}\left\{\|A_t\| \leq a\right\}}{\mathbb{1}\left\{\left\|A_t^{\mathrm{K}}\right\| \leq a\right\}}\right)\right\} \leq 1, \tag{32}$$

Analyzing "Covariance alignment" - Term B

Term B in (29) is an integration over $\mathbf{A}_\tau$. Denote $\mathbf{J}_\mathbf{A} = \frac{\partial \mathbf{A}_\tau^{\mathrm{K}}}{\partial \mathbf{A}_\tau}$ as the Jacobian matrix that transforms the integral to $\mathbf{A}_\tau^{\mathrm{K}}$. We further refer to $\mathbf{J}_\mathbf{A}$ as the *actions Jacobian*.

$$\begin{aligned}
&\int_{\mathcal{E}_v} f_\mathbf{A}\left(\mathbf{A}_\tau^{\mathrm{K}}\left[\mathbf{A}_\tau\right]\right) \mathrm{R}^*_{\mathrm{QB}_\tau}\left(\hat{\mu}_{\tau+1}\left[\mathbf{A}_\tau, \Xi_\tau\right], \hat{\mathbf{\Sigma}}_{\tau+1}\left[\mathbf{A}_\tau\right], T-\tau\right) d\mathbf{A}_\tau \\
&\quad \leq \underbrace{\max_{\mathcal{E}_v}\left\{\frac{1}{\left|\mathrm{Det}\left(\mathbf{J}_\mathbf{A}\right)\right|}\right\}}_{\text{The Jacobian}} \cdot \underbrace{\int_{\mathcal{E}_v} f_\mathbf{A}\left(\mathbf{A}_\tau^{\mathrm{K}}\left[\mathbf{A}_\tau\right]\right) \mathrm{R}^*_{\mathrm{QB}_\tau}\left(\hat{\mu}_{\tau+1}\left[\mathbf{A}_\tau, \Xi_\tau\right], \hat{\mathbf{\Sigma}}_{\tau+1}\left[\mathbf{A}_\tau\right], T-\tau\right) \left|\mathrm{Det}\left(\mathbf{J}_\mathbf{A}\right)\right| d\mathbf{A}_\tau}_{\text{The integral}},
\end{aligned} \tag{33}$$

where the inequality uses that the regret is non-negative.

Analyzing "Covariance alignment" - Term B - The Jacobian

The following lemma bounds the determinant of the actions Jacobian.

Lemma 5. *Let matrices* $\mathbf{X} \in \mathbb{R}^{n\times d}, \mathbf{B} \in \mathbb{R}^{d\times d}$, *such that* $\mathbf{B} \succeq 0, \mathbf{X}^\top\mathbf{X} \succ \mathbf{B}$.

Define the matrix $\mathbf{U} \triangleq \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}$ *and denote the Jacobian matrix* $\mathbf{J} = \frac{\partial \mathbf{U}}{\partial \mathbf{X}}$, *then*

$$\frac{1}{\left|\mathrm{Det}\left(\mathbf{J}\right)\right|} \leq \left(\frac{\mathrm{Det}\left(\mathbf{X}^\top\mathbf{X}\right)}{\mathrm{Det}\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)}\right)^{n/2}.$$

Lemma 5 is highly important in the proof, since it allows later to perform a change of measure over the actions space. The proof is mostly technical and uses properties of Kronecker product and PSD matrices. Due to its length, it can be found in Appendix C.4.

Define $c_s \triangleq \frac{2\sigma^2}{\lambda^2_{\mathbf{\Sigma}_*}\lambda_{\mathbf{\Sigma}_\mathcal{A}}}$, bounding the Jacobian term in (33),

$$\begin{aligned}
\max_{\mathcal{E}_v}\left\{\frac{1}{\left|\mathrm{Det}\left(\mathbf{J}_\mathbf{A}\right)\right|}\right\} &\underset{(a)}{\leq} \max_{\mathcal{E}_v}\left\{\left(\frac{\mathrm{Det}\left(\mathbf{V}_\tau\right)}{\mathrm{Det}\left(\mathbf{V}_\tau-\mathbf{B}\right)}\right)^{\tau/2}\right\} \\
&\underset{(b)}{\leq} \max_{\mathcal{E}_v}\left\{\left(\frac{\prod_{j=1}^d \lambda_j\left(\mathbf{V}_\tau\right)}{\prod_{j=1}^d\left(\lambda_j\left(\mathbf{V}_\tau\right)+\lambda_j\left(-\mathbf{B}\right)\right)}\right)^{\tau/2}\right\} \\
&= \max_{\mathcal{E}_v}\left\{\left(\prod_{j=1}^d\left(1+\frac{\lambda_j\left(-\mathbf{B}\right)}{\lambda_j\left(\mathbf{V}_\tau\right)}\right)\right)^{-\tau/2}\right\} \\
&\underset{(c)}{\leq} \max_{\mathcal{E}_v}\left\{\left(1-\frac{\lambda_{\max}\left(\mathbf{B}\right)}{\lambda_{\min}\left(\mathbf{V}_\tau\right)}\right)^{-d\tau/2}\right\} \\
&\underset{(d)}{=} \max_{\mathcal{E}_v}\left\{\left(1-\frac{\|\mathbf{B}\|_{\mathrm{op}}}{\lambda_{\min}\left(\mathbf{V}_\tau\right)}\right)^{-d\tau/2}\right\} \\
&\underset{(e)}{\leq} \left(1-\frac{c_s}{d}\left\|\hat{\mathbf{\Sigma}}-\mathbf{\Sigma}_*\right\|_{\mathrm{op}}\right)^{-d\tau/2} \\
&\underset{(f)}{\leq} \left(1-\frac{c_s}{d}\sqrt{f_s\delta}\right)^{-d\tau/2} \\
&\underset{(g)}{\leq} \left(1-\frac{c_s\tau}{2}\sqrt{f_s\delta}\right)^{-1} \\
&= 1+\frac{c_s\tau\sqrt{f_s\delta}}{2-c_s\tau\sqrt{f_s\delta}}
\end{aligned}$$

$$\underset{(h)}{\leq} 1 + c_s \tau \sqrt{f_s \delta}, \tag{34}$$

where (a) uses Lemma 5 and (27), (b) uses Lemma 35, that both matrices are symmetric and (27), (c) uses that both $\mathbf{B}$ and $\mathbf{V}_\tau$ are PSD matrices, (d) uses that for PSD matrices under the l_2-operator norm $\lambda_{\max}(\mathbf{A}) = \sigma_{\max}(\mathbf{A}) = \|\mathbf{A}\|_{\text{op}}$, in (e) the numerator uses Lemma 36 and the denominator uses the event $\mathcal{E}_v$, (f) uses event $\mathcal{E}_s$, (g) uses Bernoulli inequality and that $c_s \tau \sqrt{f_s \delta} < 2$ for $\delta \leq 1/M$ and (h) uses that $c_s \tau \sqrt{f_s \delta} \leq 1$ for $\delta \leq 1/M$.

Analyzing "Covariance alignment" - Term B - The integral

Imagine a QB_τ algorithm with the correct prior covariance $\mathbf{\Sigma}_*$ and a prior mean $\tilde{\mu}$ which is defined by the following scenario: If this algorithm would have taken the specific set of actions $\mathbf{A}_\tau^{\text{K}}[\mathbf{A}_\tau]$ and would have received the specific set of noises Ξ_τ, it would end up with the same mean as QB_τ at time-step $\tau+1$, i.e. the vector that brings the following two equations to equality,

$$\hat{\mu}_{\tau+1} = \hat{\mathbf{\Sigma}}_{\tau+1}[\mathbf{A}_\tau]\left(\hat{\mathbf{\Sigma}}^{-1}\hat{\mu} + \frac{1}{\sigma^2}\mathbf{V}_\tau\theta + \frac{1}{\sigma^2}\mathbf{A}_\tau^\top \Xi_\tau\right),$$
$$\tilde{\mu}_{\tau+1} = \mathbf{\Sigma}_{*,\tau+1}\left[\mathbf{A}_\tau^{\text{K}}[\mathbf{A}_\tau]\right]\left(\mathbf{\Sigma}_*^{-1}\tilde{\mu} + \frac{1}{\sigma^2}\mathbf{V}_\tau^{\text{K}}[\mathbf{A}_\tau]\theta + \frac{1}{\sigma^2}\left(\mathbf{A}_\tau^{\text{K}}[\mathbf{A}_\tau]\right)^\top \Xi_\tau\right).$$

Describing $\tilde{\mu}$ as a function of the terms determined during the instance,

$$\tilde{\mu}[\mathbf{A}_\tau, \Xi_\tau] = \mathbf{\Sigma}_*\left(\hat{\mathbf{\Sigma}}^{-1}\hat{\mu} + \frac{1}{\sigma^2}\mathbf{B}\theta + \frac{1}{\sigma^2}\left(\mathbf{A}_\tau - \mathbf{A}_\tau^{\text{K}}[\mathbf{A}_\tau]\right)^\top \Xi_\tau\right), \tag{35}$$
$$\tilde{\mu}\left[\mathbf{A}_\tau^{\text{K}}, \Xi_\tau\right] = \mathbf{\Sigma}_*\left(\hat{\mathbf{\Sigma}}^{-1}\hat{\mu} + \frac{1}{\sigma^2}\mathbf{B}\theta + \frac{1}{\sigma^2}\left(\mathbf{A}_\tau\left[\mathbf{A}_\tau^{\text{K}}\right] - \mathbf{A}_\tau^{\text{K}}\right)^\top \Xi_\tau\right). \tag{36}$$

Analyzing the integral in (33),

$$\begin{aligned}
&\int_{\mathcal{E}_v} f_{\mathbf{A}}\left(\mathbf{A}_\tau^{\text{K}}[\mathbf{A}_\tau]\right) \text{R}^*_{\text{QB}_\tau}\left(\hat{\mu}_{\tau+1}[\mathbf{A}_\tau, \Xi_\tau], \hat{\mathbf{\Sigma}}_{\tau+1}[\mathbf{A}_\tau], T-\tau\right) |\text{Det}(\mathbf{J}_A)| \, d\mathbf{A}_\tau \\
&\underset{(a)}{\leq} \int_{\mathbf{V}_\tau \succ \mathbf{B}} f_{\mathbf{A}}\left(\mathbf{A}_\tau^{\text{K}}[\mathbf{A}_\tau]\right) \text{R}^*_{\text{QB}_\tau}\left(\hat{\mu}_{\tau+1}[\mathbf{A}_\tau, \Xi_\tau], \hat{\mathbf{\Sigma}}_{\tau+1}[\mathbf{A}_\tau], T-\tau\right) |\text{Det}(\mathbf{J}_A)| \, d\mathbf{A}_\tau \\
&\underset{(b)}{=} \int_{\mathbf{V}_\tau \succ \mathbf{B}} f_{\mathbf{A}}\left(\mathbf{A}_\tau^{\text{K}}[\mathbf{A}_\tau]\right) \text{R}^*_{\text{QB}_\tau}\left(\tilde{\mu}_{\tau+1}\left[\mathbf{A}_\tau^{\text{K}}[\mathbf{A}_\tau], \Xi_\tau\right], \mathbf{\Sigma}_{*,\tau+1}\left[\mathbf{A}_\tau^{\text{K}}[\mathbf{A}_\tau]\right], T-\tau\right) |\text{Det}(\mathbf{J}_A)| \, d\mathbf{A}_\tau \\
&\underset{(c)}{=} \int_{\mathbf{V}_\tau^{\text{K}} \succ 0} f_{\mathbf{A}}\left(\mathbf{A}_\tau^{\text{K}}\right) \text{R}^*_{\text{QB}_\tau}\left(\tilde{\mu}_{\tau+1}\left[\mathbf{A}_\tau^{\text{K}}, \Xi_\tau\right], \mathbf{\Sigma}_{*,\tau+1}\left[\mathbf{A}_\tau^{\text{K}}\right], T-\tau\right) d\mathbf{A}_\tau^{\text{K}} \\
&= \mathop{\mathbb{E}}_{\mathbf{A}_\tau^{\text{K}}}\left[\text{R}^*_{\text{QB}_\tau}\left(\tilde{\mu}_{\tau+1}\left[\mathbf{A}_\tau^{\text{K}}, \Xi_\tau\right], \mathbf{\Sigma}_{*,\tau+1}, T-\tau\right) \mathbb{1}\left\{\mathbf{V}_\tau^{\text{K}} \succ 0\right\}\right],
\end{aligned} \tag{37}$$

where (a) uses (27) and that the regret is non-negative, (b) uses the definition of $\mathbf{A}_\tau^{\text{K}}[\mathbf{A}_\tau]$ and (c) uses a change of measure.

Plugging (32), (33), (34) and (37) back to (29), the regret incurred under the "Good event",

$$\mathbb{E}\left[\text{R}^*_{\text{QB}_\tau}\left(\hat{\mu}_{\tau+1}, \hat{\mathbf{\Sigma}}_{\tau+1}, T-\tau\right) \mathbb{1}\{\mathcal{E}\}\right] \leq \underbrace{\left(1 + c_s \tau \sqrt{f_s \delta}\right)}_{\substack{\text{Covariance} \\ \text{alignment cost}}} \mathbb{E} \mathop{\mathbb{E}}_{\mathbf{A}_\tau^{\text{K}}}\left[\text{R}^*_{\text{QB}_\tau}\left(\tilde{\mu}_{\tau+1}\left[\mathbf{A}_\tau^{\text{K}}, \Xi_\tau\right], \mathbf{\Sigma}_{*,\tau+1}, T-\tau\right) \mathbb{1}\left\{\mathbf{V}_\tau^{\text{K}} \succ 0\right\}\right]. \tag{38}$$

C.2.2 Mean Alignment

After the covariance alignment, in order to bound QB_τ by KQB_τ, we still need to align the mean. At this stage of the proof, the actions at the exploration time-steps have already been determined, yet we still have a degree of freedom in the randomness of the noise terms. In the Bayesian update rule for the mean (16) the noise terms

appear only in the expression ${\mathbf{A}_\tau}^\top \Xi_\tau$. Since in this stage of the proof, both algorithms use the set of actions $\mathbf{A}_\tau^{\mathrm{K}}$, we denote $S_\tau \triangleq \left(\mathbf{A}_\tau^{\mathrm{K}}\right)^\top \Xi_\tau$ and $S_\tau^{\mathrm{K}} \triangleq \left(\mathbf{A}_\tau^{\mathrm{K}}\right)^\top \Xi_\tau^{\mathrm{K}}$. To comply with this definition we further denote $\mu_{\tau+1}\left[S_\tau\right]$ instead of $\mu_{\tau+1}\left[\mathbf{A}_\tau^{\mathrm{K}}, \Xi_\tau\right]$. During the exploration time-steps, the actions are chosen independently of the rewards achieved, hence, independent of the noise terms and of θ. Given these actions, S_τ is a Gaussian vector, i.e.

$$S_\tau \mid \mathbf{A}_\tau^{\mathrm{K}} \sim \mathcal{N}\left(0, \left(\mathbf{A}_\tau^{\mathrm{K}}\right)^\top \mathbb{E}\left[\Xi_\tau \Xi_\tau^\top\right] \mathbf{A}_\tau^{\mathrm{K}}\right) = \mathcal{N}\left(0, \sigma^2 \mathbf{V}_\tau^{\mathrm{K}}\right).$$

We denote by $S_\tau^{\mathrm{K}}\left[S_\tau\right]$ the vector S_τ^{K} that brings the following equations to equality,

$$\begin{aligned}
\tilde{\mu}_{\tau+1}\left[S_\tau\right] &= \mathbf{\Sigma}_{*,\tau+1}\left(\mathbf{\Sigma}_*^{-1}\tilde{\mu} + \frac{1}{\sigma^2}\mathbf{V}_\tau^{\mathrm{K}}\theta + \frac{1}{\sigma^2}S_\tau\right), \\
\mu_{*,\tau+1}\left[S_\tau^{\mathrm{K}}\right] &= \mathbf{\Sigma}_{*,\tau+1}\left(\mathbf{\Sigma}_*^{-1}\mu_* + \frac{1}{\sigma^2}\mathbf{V}_\tau^{\mathrm{K}}\theta + \frac{1}{\sigma^2}S_\tau^{\mathrm{K}}\right).
\end{aligned}$$

Define,

$$G \triangleq \sigma^2 \hat{\mathbf{\Sigma}}^{-1}\left(\hat{\mu} - \mu_*\right) + \mathbf{B}\left(\theta - \mu_*\right). \tag{39}$$

We get,

$$\begin{aligned}
S_\tau^{\mathrm{K}}\left[S_\tau\right] &= S_\tau + \sigma^2 \mathbf{\Sigma}_*^{-1}\left(\tilde{\mu} - \mu_*\right) \\
&\underset{(a)}{=} S_\tau + \sigma^2\left(\hat{\mathbf{\Sigma}}^{-1}\hat{\mu} + \frac{1}{\sigma^2}\mathbf{B}\theta + \frac{1}{\sigma^2}\left(\mathbf{A}_\tau\left[\mathbf{A}_\tau^{\mathrm{K}}\right] - \mathbf{A}_\tau^{\mathrm{K}}\right)^\top \Xi_\tau - \mathbf{\Sigma}_*^{-1}\mu_*\right) \\
&= S_\tau + \sigma^2\hat{\mathbf{\Sigma}}^{-1}\hat{\mu} + \mathbf{B}\theta + \mathbf{A}_\tau\left[\mathbf{A}_\tau^{\mathrm{K}}\right]^\top \Xi_\tau - S_\tau - \sigma^2\mathbf{\Sigma}_*^{-1}\mu_* \\
&= \sigma^2\hat{\mathbf{\Sigma}}^{-1}\left(\hat{\mu} - \mu_*\right) + \mathbf{B}\theta + \mathbf{A}_\tau\left[\mathbf{A}_\tau^{\mathrm{K}}\right]^\top \Xi_\tau + \sigma^2\left(\hat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}_*^{-1}\right)\mu_* \\
&= G + \mathbf{A}_\tau\left[\mathbf{A}_\tau^{\mathrm{K}}\right]^\top \Xi_\tau \\
&\underset{(b)}{=} G + \left(\mathbf{V}_\tau^{\mathrm{K}} + \mathbf{B}\right)^{1/2}\left(\mathbf{V}_\tau^{\mathrm{K}}\right)^{-1/2} S_\tau,
\end{aligned} \tag{40}$$

where (a) uses the definition of $\tilde{\mu}$ from (36) and (b) uses the definition of $\mathbf{A}_\tau\left[\mathbf{A}_\tau^{\mathrm{K}}\right]$ from (28).

Mean Alignment - Regret Decomposition

Denote $c_\xi \triangleq \sigma\sqrt{5\ln\left(\frac{dT}{\delta}\right)}$ and the event,

$$\mathcal{E}_\xi \triangleq \left\{\|S_\tau\|_{\left(\mathbf{V}_\tau^{\mathrm{K}}\right)^{-1}} \le c_\xi\sqrt{d}\right\}. \tag{41}$$

By Lemma 28 with $X = \Xi_\tau$ and $\mathbf{A} = \left(\mathbf{V}_\tau^{\mathrm{K}}\right)^{-1/2}\mathbf{A}_\tau^{\mathrm{K}}$ we have that,

$$\mathbb{P}\left(\mathcal{E}_\xi \mid \mathbf{A}_\tau^{\mathrm{K}}\right) > 1 - \frac{\delta}{dT}. \tag{42}$$

Continue from equation (38),

$$\begin{aligned}
&\mathbb{E}\underset{\Xi_\tau}{\mathbb{E}}\underset{\mathbf{A}_\tau^{\mathrm{K}}}{\mathbb{E}}\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\tilde{\mu}_{\tau+1}\left[S_\tau\right], \mathbf{\Sigma}_{*,\tau+1}, T-\tau\right)\cdot\mathbb{1}\left\{\mathbf{V}_\tau^{\mathrm{K}} \succ 0\right\} \\
&\quad= \mathbb{E}\underset{\mathbf{A}_\tau^{\mathrm{K}}}{\mathbb{E}}\underbrace{\int_{\mathcal{E}_\xi}\frac{e^{-\frac{1}{2}\|S_\tau\|^2_{\left(\mathbf{V}_\tau^{\mathrm{K}}\right)^{-1}}}}{\sqrt{(2\pi)^\tau \mathrm{Det}\left(\mathbf{V}_\tau^{\mathrm{K}}\right)}}\cdot\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\tilde{\mu}_{\tau+1}\left[S_\tau\right], \mathbf{\Sigma}_{*,\tau+1}, T-\tau\right)dS_\tau}_{\text{small noise terms}}\cdot\mathbb{1}\left\{\mathbf{V}_\tau^{\mathrm{K}} \succ 0\right\} \\
&\qquad+\underbrace{\mathbb{E}\underset{\mathbf{A}_\tau^{\mathrm{K}}}{\mathbb{E}}\underset{\mathbf{S}_\tau}{\mathbb{E}}\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\tilde{\mu}_{\tau+1}\left[S_\tau\right], \mathbf{\Sigma}_{*,\tau+1}, T-\tau\right)\mathbb{1}\left\{\bar{\mathcal{E}}_\xi\right\}}_{\text{large noise terms}}\cdot\mathbb{1}\left\{\mathbf{V}_\tau^{\mathrm{K}} \succ 0\right\}.
\end{aligned} \tag{43}$$

The second term complements the first, but uses an expectation notation. It is bounded by,

$$\begin{aligned}\mathbb{E}\left[\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\tilde{\mu}_{\tau+1}\left[S_\tau\right],\boldsymbol{\Sigma}_{*,\tau+1},T-\tau\right)\mathbb{1}\left\{\bar{\mathcal{E}}_\xi\right\}\right] &\underset{(a)}{\leq}\mathbb{E}\left[\max_{\Xi_\tau}\left\{\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\tilde{\mu}_{\tau+1}\left[S_\tau\right],\boldsymbol{\Sigma}_{*,\tau+1},T-\tau\right)\right\}\,\middle|\,\mathcal{E}_\theta\right]\cdot\mathbb{P}\left(\bar{\mathcal{E}}_\xi\mid\mathbf{A}^{\mathrm{K}}_\tau\right)\\ &\underset{(b)}{\leq}\frac{\delta}{dT}\mathbb{E}\left[\max_{\Xi_\tau}\left\{\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\tilde{\mu}_{\tau+1}\left[S_\tau\right],\boldsymbol{\Sigma}_{*,\tau+1},T-\tau\right)\right\}\,\middle|\,\mathcal{E}_\theta\right]\\ &\underset{(c)}{\leq}\frac{c_{\mathrm{bad}}\delta}{11\sqrt{d}},\end{aligned}\tag{44}$$

where (a) uses that the regret is non-negative and (b) uses (42) and (c) uses the same derivations as Lemma 11.

Analyzing "Mean alignment" - small noise terms

We bound the small noise term by a change of measure of the integral. Given $\mathbf{A}^{\mathrm{K}}_\tau$,

$$\mathbf{J}_S=\frac{\partial S^{\mathrm{K}}_\tau}{\partial S_\tau}=\left(\mathbf{V}^{\mathrm{K}}_\tau+\mathbf{B}\right)^{1/2}\left(\mathbf{V}^{\mathrm{K}}_\tau\right)^{-1/2}.$$

Hence,

$$\left|\mathrm{Det}\left(\mathbf{J}_S\right)\right|=\left|\frac{\left(\mathrm{Det}\left(\mathbf{V}^{\mathrm{K}}_\tau+\mathbf{B}\right)\right)^{1/2}}{\left(\mathrm{Det}\left(\mathbf{V}^{\mathrm{K}}_\tau\right)\right)^{1/2}}\right|=\left(\frac{\mathrm{Det}\left(\mathbf{V}^{\mathrm{K}}_\tau+\mathbf{B}\right)}{\mathrm{Det}\left(\mathbf{V}^{\mathrm{K}}_\tau\right)}\right)^{1/2}\geq 1,\tag{45}$$

where the last equality uses determinant laws and that both matrices are PD and the inequality uses $\mathbf{B}\succeq 0$.

$$\begin{aligned}&\int_{\mathcal{E}_\xi}\frac{e^{-\frac{1}{2}\|S_\tau\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}}}}{\sqrt{(2\pi)^\tau\mathrm{Det}\left(\mathbf{V}^{\mathrm{K}}_\tau\right)}}\cdot\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\tilde{\mu}_{\tau+1}\left[S_\tau\right],\boldsymbol{\Sigma}_{*,\tau+1},T-\tau\right)dS_\tau\\ &\leq\underbrace{\max_{\mathcal{E}_\xi}\left\{\frac{e^{-\frac{1}{2}\|S_\tau\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}}}}{e^{-\frac{1}{2}\|S^{\mathrm{K}}_\tau[S_\tau]\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}}}}\right\}}_{\text{Term A}}\cdot\underbrace{\int_{\mathcal{E}_\xi}\frac{e^{-\frac{1}{2}\left\|S^{\mathrm{K}}_\tau[S_\tau]\right\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}}}}{\sqrt{(2\pi)^\tau\mathrm{Det}\left(\mathbf{V}^{\mathrm{K}}_\tau\right)}}\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\tilde{\mu}_{\tau+1}\left[S_\tau\right],\boldsymbol{\Sigma}_{*,\tau+1},T-\tau\right)\left|\mathrm{Det}\left(\mathbf{J}_S\right)\right|dS_\tau}_{\text{Term B}},\end{aligned}\tag{46}$$

where the inequality uses that the regret is non-negative and (45).

Analyzing "Mean alignment" - small noise terms - Term A

$$\begin{aligned}\max_{\mathcal{E}_\xi}\left\{\frac{e^{-\frac{1}{2}\|S_\tau\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}}}}{e^{-\frac{1}{2}\|S^{\mathrm{K}}_\tau[S_\tau]\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}}}}\right\}&=\max_{\mathcal{E}_\xi}\left\{\exp\left(\frac{1}{2}\left\|S^{\mathrm{K}}_\tau\left[S_\tau\right]\right\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}}-\frac{1}{2}\left\|S_\tau\right\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}}\right)\right\}\\ &\underset{(a)}{=}\max_{\mathcal{E}_\xi}\left\{\exp\left(\frac{1}{2}\left\|G+\left(\mathbf{V}^{\mathrm{K}}_\tau+\mathbf{B}\right)^{1/2}\left(\mathbf{V}^{\mathrm{K}}_\tau\right)^{-1/2}S_\tau\right\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}}-\frac{1}{2}\left\|S_\tau\right\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}}\right)\right\}\\ &\underset{(b)}{\leq}\max_{\mathcal{E}_\xi}\Bigg\{\exp\Bigg(\frac{1}{2}\|G\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}}+\|G\|_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}}\left\|\left(\mathbf{V}^{\mathrm{K}}_\tau+\mathbf{B}\right)^{1/2}\left(\mathbf{V}^{\mathrm{K}}_\tau\right)^{-1/2}S_\tau\right\|_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}}\\ &\qquad+\frac{1}{2}\left\|\left(\mathbf{V}^{\mathrm{K}}_\tau+\mathbf{B}\right)^{1/2}\left(\mathbf{V}^{\mathrm{K}}_\tau\right)^{-1/2}S_\tau\right\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}}-\frac{1}{2}\left\|S_\tau\right\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}}\Bigg)\Bigg\},\end{aligned}\tag{47}$$

where (a) uses the definition of $S^{\mathrm{K}}_\tau\left[S_\tau\right]$ in (40) and (b) uses Cauchy–Schwarz inequality.

Defining $c_1 \triangleq \frac{2}{\lambda \boldsymbol{\Sigma}_*} \ln\left(\frac{d^2 T}{\delta}\right)$ and analyzing the first term of the exponent in (47),

$$\begin{aligned}
\|G\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}} &\underset{(a)}{=} \left\|\sigma^2 \hat{\boldsymbol{\Sigma}}^{-1}\left((\hat{\mu}-\mu_*)+\left(\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_*\right)\boldsymbol{\Sigma}_*^{-1}(\theta-\mu_*)\right)\right\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}} \\
&\underset{(b)}{\leq} \frac{2\sigma^4}{\lambda_{\min}(\mathbf{V}^{\mathrm{K}}_\tau)}\left\|\hat{\boldsymbol{\Sigma}}^{-1}\right\|^2_{\mathrm{op}}\left(\|\hat{\mu}-\mu_*\|^2+\left\|\left(\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_*\right)\boldsymbol{\Sigma}_*^{-1}(\theta-\mu_*)\right\|^2\right) \\
&\underset{(c)}{\leq} \frac{2c_s}{d}\left(\|\hat{\mu}-\mu_*\|^2+\left\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_*\right\|^2_{\mathrm{op}}\left\|\boldsymbol{\Sigma}_*^{-1/2}\right\|^2_{\mathrm{op}}\left\|\boldsymbol{\Sigma}_*^{-1/2}(\theta-\mu_*)\right\|^2\right) \\
&\underset{(d)}{\leq} \frac{2c_s\delta\left(f_m+c_1 d f_s\right)}{d},
\end{aligned} \tag{48}$$

where (a) uses the definition of G in (39) and $\mathbf{B}$ in (25), (b) uses Lemma 26, $\|\mathbf{A}Y\| \leq \|\mathbf{A}\|_{\mathrm{op}}\|Y\|$, triangle inequality and $(a+b)^2 \leq 2(a^2+b^2)$, (c) uses the event $\mathcal{E}_v$, that for PSD matrices $\|\mathbf{A}\|_{\mathrm{op}} = \lambda_{\max}(\mathbf{A})$, $\hat{\boldsymbol{\Sigma}} \succeq \boldsymbol{\Sigma}_*$, Assumption 2, $\|\mathbf{A}Y\| \leq \|\mathbf{A}\|_{\mathrm{op}}\|Y\|$ and the sub-multiplicative norm property, (d) uses the triangle inequality, $(a+b)^2 \leq 2(a^2+b^2)$, the events $\mathcal{E}_\theta, \mathcal{E}_m, \mathcal{E}_s$, that for PSD matrices $\|\mathbf{A}\|_{\mathrm{op}} = \lambda_{\max}(\mathbf{A})$, $\hat{\boldsymbol{\Sigma}} \succeq \boldsymbol{\Sigma}_*$ and Assumption 2.

Analyzing the third term in the exponent in (47),

$$\begin{aligned}
\left\|\left(\mathbf{V}^{\mathrm{K}}_\tau+\mathbf{B}\right)^{1/2}\left(\mathbf{V}^{\mathrm{K}}_\tau\right)^{-1/2}S_\tau\right\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}} &= S_\tau^\top\left(\mathbf{V}^{\mathrm{K}}_\tau\right)^{-1/2}\left(\mathbf{V}^{\mathrm{K}}_\tau+\mathbf{B}\right)^{1/2}\left(\mathbf{V}^{\mathrm{K}}_\tau\right)^{-1}\left(\mathbf{V}^{\mathrm{K}}_\tau+\mathbf{B}\right)^{1/2}\left(\mathbf{V}^{\mathrm{K}}_\tau\right)^{-1/2}S_\tau \\
&\underset{(a)}{\leq} \lambda_{\max}\left(\left(\mathbf{V}^{\mathrm{K}}_\tau+\mathbf{B}\right)^{1/2}\left(\mathbf{V}^{\mathrm{K}}_\tau\right)^{-1}\left(\mathbf{V}^{\mathrm{K}}_\tau+\mathbf{B}\right)^{1/2}\right)\|S_\tau\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}} \\
&\underset{(b)}{=} \lambda_{\max}\left(\left(\mathbf{V}^{\mathrm{K}}_\tau\right)^{-1/2}\left(\mathbf{V}^{\mathrm{K}}_\tau+\mathbf{B}\right)\left(\mathbf{V}^{\mathrm{K}}_\tau\right)^{-1/2}\right)\|S_\tau\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}} \\
&= \lambda_{\max}\left(\mathbf{I}+\left(\mathbf{V}^{\mathrm{K}}_\tau\right)^{-1/2}\mathbf{B}\left(\mathbf{V}^{\mathrm{K}}_\tau\right)^{-1/2}\right)\|S_\tau\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}} \\
&= \left(1+\lambda_{\max}\left(\left(\mathbf{V}^{\mathrm{K}}_\tau\right)^{-1/2}\mathbf{B}\left(\mathbf{V}^{\mathrm{K}}_\tau\right)^{-1/2}\right)\right)\|S_\tau\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}} \\
&\underset{(c)}{\leq} \left(1+\frac{\lambda_{\max}(\mathbf{B})}{\lambda_{\min}(\mathbf{V}^{\mathrm{K}}_\tau)}\right)\|S_\tau\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}} \\
&\underset{(d)}{\leq} \left(1+\frac{c_s}{d}\left\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_*\right\|_{\mathrm{op}}\right)\|S_\tau\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}} \\
&\underset{(e)}{\leq} \left(1+\frac{c_s}{d}\sqrt{f_s\delta}\right)\|S_\tau\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}},
\end{aligned} \tag{49}$$

where (a) uses the same derivation as in Lemma 26, (b) uses Lemma 31, (c) uses Lemma 31 and sub-multiplicative norm properties, (d) uses Lemma 36 and event $\mathcal{E}_v$ and (e) uses event $\mathcal{E}_s$.

Plugging (48), (49) to (47)

$$\begin{aligned}
\max_{\mathcal{E}_\xi}\left\{\frac{e^{-\frac{1}{2}\|S_\tau\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}}}}{e^{-\frac{1}{2}\|S^{\mathrm{K}}_\tau[S_\tau]\|^2_{(\mathbf{V}^{\mathrm{K}}_\tau)^{-1}}}}\right\} &\underset{(a)}{\leq} \exp\left(\frac{c_s\delta\left(f_m+c_1 d f_s\right)}{d}+2c_\xi\sqrt{c_s\delta\left(f_m+c_1 d f_s\right)}+\frac{c_\xi^2 c_s\sqrt{f_s\delta}}{2}\right) \\
&\underset{(b)}{\leq} \exp\left(3\sqrt{c_\xi^2 c_s\delta\left(f_m+c_1 d f_s\right)}+\frac{c_\xi^2 c_s\sqrt{f_s\delta}}{2}\right) \\
&\underset{(c)}{\leq} 1+2\left(3\sqrt{c_\xi^2 c_s\delta\left(f_m+c_1 d f_s\right)}+\frac{c_\xi^2 c_s\sqrt{f_s\delta}}{2}\right) \\
&\underset{(d)}{\leq} 1+6\sqrt{c_\xi^2 c_s}\sqrt{f_m\delta}+6\left(\sqrt{c_\xi^2 c_s c_1 d}+c_\xi^2 c_s\right)\sqrt{f_s\delta},
\end{aligned} \tag{50}$$

where (a) uses that $1+\frac{c_s}{d}\sqrt{f_s\delta} \leq 2$ for $\delta \leq 1/M$ and (41), (b) uses that $\sqrt{c_s\delta\left(f_m+c_1 d f_s\right)} \leq c_\xi d$ for $\delta \leq 1/M$, (c) uses Lemma 37 and $\left(3\sqrt{c_\xi^2 c_s\left(f_m+c_1 d f_s\right)}+\frac{1}{2}c_\xi^2 c_s\sqrt{f_s}\right)\sqrt{\delta} \leq 1$ for $\delta \leq 1/M$ and (d) uses $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$.

Analyzing "Mean alignment" - small noise terms - Term B

$$
\begin{aligned}
&\int_{\mathcal{E}_\xi} \frac{e^{-\frac{1}{2}\left\|S_\tau^{\mathrm{K}}[S_\tau]\right\|^2_{(\mathbf{V}_\tau^{\mathrm{K}})^{-1}}}}{\sqrt{(2\pi)^\tau \mathrm{Det}\left(\mathbf{V}_\tau^{\mathrm{K}}\right)}} \mathrm{R}^*_{\mathrm{QB}_\tau}\left(\tilde{\mu}_{\tau+1}\left[S_\tau\right], \mathbf{\Sigma}_{*,\tau+1}, T-\tau\right) \left|\mathrm{Det}\left(\mathbf{J}_S\right)\right| dS_\tau \\
&\underset{(a)}{\leq} \int \frac{e^{-\frac{1}{2}\left\|S_\tau^{\mathrm{K}}[S_\tau]\right\|^2_{(\mathbf{V}_\tau^{\mathrm{K}})^{-1}}}}{\sqrt{(2\pi)^\tau \mathrm{Det}\left(\mathbf{V}_\tau^{\mathrm{K}}\right)}} \mathrm{R}^*_{\mathrm{QB}_\tau}\left(\tilde{\mu}_{\tau+1}\left[S_\tau\right], \mathbf{\Sigma}_{*,\tau+1}, T-\tau\right) \left|\mathrm{Det}\left(\mathbf{J}_S\right)\right| dS_\tau \\
&\underset{(b)}{=} \int \frac{e^{-\frac{1}{2}\left\|S_\tau^{\mathrm{K}}[S_\tau]\right\|^2_{(\mathbf{V}_\tau^{\mathrm{K}})^{-1}}}}{\sqrt{(2\pi)^\tau \mathrm{Det}\left(\mathbf{V}_\tau^{\mathrm{K}}\right)}} \mathrm{R}^*_{\mathrm{QB}_\tau}\left(\mu_{*,\tau+1}\left[S_\tau^{\mathrm{K}}\left[S_\tau\right]\right], \mathbf{\Sigma}_{*,\tau+1}, T-\tau\right) \left|\mathrm{Det}\left(\mathbf{J}_S\right)\right| dS_\tau \\
&\underset{(c)}{=} \mathbb{E}\left[\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\mu_{*,\tau+1}, \mathbf{\Sigma}_{*,\tau+1}, T-\tau\right)\right],
\end{aligned}
\tag{51}
$$

where (a) uses that the regret is non-negative, (b) uses the definition of $S_\tau^{\mathrm{K}}\left[S_\tau\right]$ and (c) uses change of measure. Denote $k_1 \triangleq 12\sqrt{c_\xi^2 c_s}\sqrt{f_m\delta} + \left(c_s\tau + 12\sqrt{c_\xi^2 c_s c_1 d} + 2c_\xi^2 c_s\right)\sqrt{f_s\delta}$ and plugging equations (50), (51), (46) and (44) back to (43) and (38) we get that the regret incurred under the "Good event",

$$
\begin{aligned}
&\mathbb{E}\left[\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\hat{\mu}_{\tau+1}, \hat{\mathbf{\Sigma}}_{\tau+1}, T-\tau\right)\mathbb{1}\left\{\mathcal{E}\right\}\right] \\
&\leq \left(1 + c_s\tau\sqrt{f_s\delta}\right)\left[\left(1 + 6\sqrt{c_\xi^2 c_s}\sqrt{f_m\delta} + \left(6\sqrt{c_\xi^2 c_s c_1 d} + c_\xi^2 c_s\right)\sqrt{f_s\delta}\right)\mathbb{E}\left[\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\mu_{*,\tau+1}, \mathbf{\Sigma}_{*,\tau+1}, T-\tau\right)\right] + \frac{c_{\mathrm{bad}}\delta}{11\sqrt{d}}\right] \\
&\leq \left(1 + k_1\right)\mathbb{E}\left[\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\mu_{*,\tau+1}, \mathbf{\Sigma}_{*,\tau+1}, T-\tau\right)\right] + \frac{2c_{\mathrm{bad}}\delta}{11\sqrt{d}},
\end{aligned}
\tag{52}
$$

where the last inequality uses that $c_s\tau\sqrt{f_s\delta} \leq 1$ for $\delta \leq 1/M$.

Plugging back the regret from the bad event (24) and the good event (52) to (23) and then to (22) the proof of Theorem 1 follows.

C.3 Derivation of M

Along the proof of Theorem 1 there were several upper bounds on δ that hold for $\delta \leq 1/M$. Next, we summarize them to derive the exact expression for M.

$$
\begin{aligned}
&\text{From (27), (50) (transition } (a)\text{):} \ \frac{\sigma^2\sqrt{f_s\delta}}{\lambda^2_{\mathbf{\Sigma}_*}} < \frac{\lambda_{\mathbf{\Sigma}_\mathcal{A}} d}{2} \iff \delta < \left(\frac{\lambda^2_{\mathbf{\Sigma}_*}\lambda_{\mathbf{\Sigma}_\mathcal{A}} d}{2\sigma^2\sqrt{f_s}}\right)^2 \iff \delta \underset{\text{(i)}}{<} \frac{d^2}{c_s^2 f_s}, \\
&\text{From (34), (52)} : \ c_s\tau\sqrt{f_s\delta} < 1 \iff \delta \underset{\text{(ii)}}{<} \frac{1}{c_s^2\tau^2 f_s}, \\
&\text{From (50) (transition } (b)\text{):} \ \sqrt{c_s\delta\left(f_m + c_1 d f_s\right)} \leq c_\xi d \iff \delta \underset{\text{(iii)}}{\leq} \frac{c_\xi^2 d^2}{c_s\left(f_m + c_1 d f_s\right)} \\
&\text{From (50) (transition } (c)\text{):} \ \left(3\sqrt{c_\xi^2 c_s\left(f_m + c_1 d f_s\right)} + \frac{1}{2}c_\xi^2 c_s\sqrt{f_s}\right)\sqrt{\delta} \leq 1 \\
&\iff \delta \underset{\text{(iv)}}{\leq} \frac{1}{\left(3\sqrt{c_\xi^2 c_s\left(f_m + c_1 d f_s\right)} + \frac{1}{2}c_\xi^2 c_s\sqrt{f_s}\right)^2}.
\end{aligned}
\tag{53}
$$

Note that inequalities (i), (iii) in (53) are directly implied by inequalities (ii), (iv) respectively. To ensure inequality (iv) we tighten its Right-Hand-Side (RHS),

$$\begin{aligned}\left(3\sqrt{c_\xi^2 c_s\left(f_m+c_1 d f_s\right)}+\frac{1}{2}c_\xi^2 c_s\sqrt{f_s}\right)^2 &\underset{(v)}{\leq} 2\left(9c_\xi^2 c_s\left(f_m+c_1 d f_s\right)+\frac{1}{4}c_\xi^4 c_s^2 f_s\right)\\ &= 2c_\xi^2 c_s\left(9\left(f_m+c_1 d f_s\right)+\frac{1}{4}c_\xi^2 c_s f_s\right)\\ &= 18c_\xi^2 c_s\left(f_m+c_1 d f_s+\frac{1}{36}c_\xi^2 c_s f_s\right)\\ &= 18c_\xi^2 c_s\left(f_m+\left(c_1 d+\frac{c_\xi^2 c_s}{36}\right)f_s\right)\end{aligned} \tag{54}$$

Using (ii), (v) and adding 3 to ensure that $M \geq e$,

$$M = \max\left\{3, c_s^2\tau^2 f_s, 18c_\xi^2 c_s\left(f_m+\left(c_1 d+c_\xi^2 c_s/36\right)f_s\right)\right\}.$$

In the remainder of the section, we emphasize the dependence of M on δ by $M(\delta)$, even though it is only poly-logarithmic. Specifically, since the constants and f_m, f_s, τ have at most logarithmic dependence, the expression for M is a p-degree polynomial of $\ln\left(1/\delta\right)$, for $p \leq 3$. By definition, $M\left(\delta\right) > e$, thus $\delta < 1/e$ and $\ln\left(1/\delta\right) > 1$, hence we can upper bound the expression by,

$$M(\delta) \leq a\ln^p\left(\frac{1}{\delta}\right)+b, \text{ for } p \leq 3,$$

for the minimal a, b that upper bound $M\left(\delta\right)$ and satisfy $a \geq 1/p^p$, $b > 0$. Next, we define $\tilde{M}$, which is not a function of δ, yet preserving the same asymptotic dependence, such that for $\delta \leq 1/\tilde{M}$, it follows that $\delta \leq \frac{1}{a\ln^p\left(\frac{1}{\delta}\right)+b} \leq \frac{1}{M(\delta)}$.

$$\delta^{-1} \geq \tilde{M} \triangleq \left(4pa^{1/p}\ln\left(2pa^{1/p}\right)+2b^{1/p}\right)^p \iff \delta^{-1/p} \geq 4pa^{1/p}\ln\left(2pa^{1/p}\right)+2b^{1/p}.$$

Thus using Lemma 38 with $x = \delta^{-1/p}$, $a_1 = pa^{1/p}$, $b_1 = b^{1/p}$,

$$\delta^{-1/p} \geq pa^{1/p}\ln\left(\delta^{-1/p}\right)+b^{1/p} = a^{1/p}\ln\left(\frac{1}{\delta}\right)+b^{1/p} \geq \left(a\ln^p\left(\frac{1}{\delta}\right)+b\right)^{1/p},$$

where the last inequality uses that $(x+y)^{1/p} \leq x^{1/p}+y^{1/p}$ for $p \in \mathbb{N}^+$ and $x, y > 0$. Since both sides are positive, we can raise both sides to the power of $-p$ and get,

$$\delta \leq \frac{1}{a\ln^p\left(\frac{1}{\delta}\right)+b} \leq \frac{1}{M\left(\delta\right)}.$$

C.4 Action Jacobian

In this section we prove Lemma 5. The proof uses some technical matrix relations appearing in Section C.4.3.

Lemma 5. *Let matrices $\mathbf{X} \in \mathbb{R}^{n\times d}, \mathbf{B} \in \mathbb{R}^{d\times d}$, such that $\mathbf{B} \succeq 0, \mathbf{X}^\top\mathbf{X} \succ \mathbf{B}$.*

Define the matrix $\mathbf{U} \triangleq \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}$ and denote the Jacobian matrix $\mathbf{J} = \frac{\partial\mathbf{U}}{\partial\mathbf{X}}$, then

$$\frac{1}{\left|\mathrm{Det}\left(\mathbf{J}\right)\right|} \leq \left(\frac{\mathrm{Det}\left(\mathbf{X}^\top\mathbf{X}\right)}{\mathrm{Det}\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)}\right)^{n/2}.$$

Proof.

C.4.1 Jacobian Derivation

Using differentials,

$$\begin{aligned} d\mathbf{U} &= d\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2} \\ &+ \mathbf{X}d\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2} \\ &+ \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2} d\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}. \end{aligned} \tag{55}$$

Applying vectorization (Lemma 6) on (55), where $\otimes$ is the *Kronecker product* and $\oplus$ is the *Kronecker sum*,

$$\begin{aligned} \operatorname{vec}(d\mathbf{U}) &= \underbrace{\left(\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\otimes\mathbf{I}_n\right)}_{\triangleq\mathbf{J}_1}\operatorname{vec}(d\mathbf{X}) \\ &+ \underbrace{\left(\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\otimes\mathbf{X}\right)}_{\triangleq\mathbf{J}_{2a}}\underbrace{\operatorname{vec}\left(d\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)}_{\text{term vec}_2} \\ &+ \underbrace{\left(\mathbf{I}_d\otimes\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)}_{\triangleq\mathbf{J}_{3a}}\underbrace{\operatorname{vec}\left(d\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\right)}_{\text{term vec}_3}. \end{aligned} \tag{56}$$

Define $\mathbf{N}_d \triangleq \frac{1}{2}\left(\mathbf{I}_{d^2}+\mathbf{K}_d\right)$ as in page 55 in Magnus and Neudecker (2019), for the commutation matrix $\mathbf{K}_d$ and analyzing term vec_2,

$$\begin{aligned} &\operatorname{vec}\left(d\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right) \\ &\underset{(a)}{=} -\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\otimes\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)\operatorname{vec}\left(d\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\right) \\ &\underset{(b)}{=} -\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\otimes\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\oplus\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\right)^{-1}\operatorname{vec}\left(d\left(\mathbf{X}^\top\mathbf{X}\right)\right) \\ &\underset{(c)}{=} -\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\otimes\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\oplus\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\right)^{-1}\underbrace{2\mathbf{N}_d\left(\mathbf{I}_d\otimes\mathbf{X}^\top\right)}_{\triangleq\mathbf{J}_{2c}}\operatorname{vec}\left(d\mathbf{X}\right) \\ &\underset{(d)}{=} -\left(\mathbf{I}_d\otimes\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\otimes\mathbf{I}_d\right)\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\oplus\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\right)^{-1}\mathbf{J}_{2c}\operatorname{vec}\left(d\mathbf{X}\right) \\ &\underset{(e)}{=} \underbrace{-\left(\mathbf{I}_d\otimes\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)\left(\left(\mathbf{X}^\top\mathbf{X}\right)\otimes\mathbf{I}_d+\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\otimes\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\right)^{-1}}_{\triangleq\mathbf{J}_{2b}}\mathbf{J}_{2c}\operatorname{vec}\left(d\mathbf{X}\right). \end{aligned} \tag{57}$$

Similarly for term vec_3,

$$\begin{aligned} \operatorname{vec}\left(d\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\right) &\underset{(b)}{=} \left(\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\oplus\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\right)^{-1}\operatorname{vec}\left(d\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)\right) \\ &\underset{(c)}{=} \underbrace{\left(\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\oplus\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\right)^{-1}2\mathbf{N}_d\left(\mathbf{I}_d\otimes\mathbf{X}^\top\right)}_{\triangleq\mathbf{J}_{3b}}\operatorname{vec}\left(d\mathbf{X}\right), \end{aligned} \tag{58}$$

where (a) uses table 9.7 in Magnus and Neudecker (2019), (b) uses Lemma 9, (c) uses table 9.6 in Magnus and Neudecker (2019), (d) uses (67) in Lemma 7 and (e) uses (69), (67), (66) in Lemma 7.

Plugging back (57), (58) to (56) and using the definition of the Jacobian (bottom of page 196 in Magnus and Neudecker (2019)),

$$\mathbf{J} = \mathbf{J}_1 + \underbrace{\mathbf{J}_{2a}\mathbf{J}_{2b}\mathbf{J}_{2c}}_{\triangleq\mathbf{J}_2} + \underbrace{\mathbf{J}_{3a}\mathbf{J}_{3b}}_{\triangleq\mathbf{J}_3}.$$

C.4.2 Non Negative Eigenvalues of C

Denote,

$$\mathbf{C} \triangleq \frac{1}{2}\mathbf{J}_1^{-1}(\mathbf{J}_2 + \mathbf{J}_3). \tag{59}$$

In order to use Corollary 2.2 in Zhan (2005) (Lemma 10) with $\mathbf{A}_1 = \mathbf{J}_2 + \mathbf{J}_3$ and $\mathbf{B}_1 = \mathbf{J}_1$, we need to prove that the matrix $\mathbf{C}$ has non negative eigenvalues. We start by deriving an expression for the matrix $\mathbf{C}$ and then show that its eigenvalues are the same as the eigenvalues of a multiplication of two PSD matrices $\mathbf{G}_1\mathbf{G}_2$, defined later in the proof.

Finding a Matrix With the Same Eigenvalues as C

Starting with the following auxiliary expressions,

$$\begin{aligned}
\mathbf{D}_1 &\triangleq \mathbf{J}_1^{-1}\left(\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\otimes\mathbf{X}\right)\left(\mathbf{I}_d\otimes\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)\\
&= \left(\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\otimes\mathbf{I}_n\right)^{-1}\left(\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\otimes\mathbf{X}\right)\left(\mathbf{I}_d\otimes\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)\\
&\underset{(a+b)}{=} \left(\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{-1/2}\otimes\mathbf{I}_n\right)\left(\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\otimes\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)\\
&\underset{(b)}{=} \left(\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\otimes\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right),
\end{aligned} \tag{60}$$

$$\begin{aligned}
\mathbf{D}_2 &\triangleq \mathbf{J_1}^{-1}\left(\mathbf{I}_d\otimes\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)\\
&= \left(\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\otimes\mathbf{I}_n\right)^{-1}\left(\mathbf{I}_d\otimes\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)\\
&\underset{(a)}{=} \left(\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{-1/2}\otimes\mathbf{I}_n\right)\left(\mathbf{I}_d\otimes\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)\\
&\underset{(b)}{=} \left(\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{-1/2}\otimes\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)\\
&\underset{(b)}{=} \left(\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\otimes\mathbf{X}\right)\left(\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{-1/2}\otimes\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)\\
&\underset{(b)}{=} \left(\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\otimes\mathbf{X}\right)\left(\mathbf{I}_d\otimes\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)\left(\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{-1/2}\otimes\mathbf{I}_d\right)\\
&\underset{(a+b)}{=} \left(\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\otimes\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)\left(\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\otimes\mathbf{I}_d\right)^{-1},
\end{aligned} \tag{61}$$

where (a) uses (69) and (b) uses (67), both from Lemma 7.

Using (60),

$$\begin{aligned}
\frac{1}{2}\mathbf{J}_1^{-1}\mathbf{J}_2 &= -\mathbf{D}_1\left(\left(\mathbf{X}^\top\mathbf{X}\right)\otimes\mathbf{I}_d + \left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\otimes\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\right)^{-1}\mathbf{N}_d\left(\mathbf{I}_d\otimes\mathbf{X}^\top\right)\\
&= -\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\otimes\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)\left(\left(\mathbf{X}^\top\mathbf{X}\right)\otimes\mathbf{I}_d + \left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\otimes\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\right)^{-1}\mathbf{N}_d\left(\mathbf{I}_d\otimes\mathbf{X}^\top\right).
\end{aligned} \tag{62}$$

Using (61),

$$\begin{aligned}
\frac{1}{2}\mathbf{J}_1^{-1}\mathbf{J}_3 &= \mathbf{D}_2\left(\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\oplus\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\right)^{-1}\mathbf{N}_d\left(\mathbf{I}_d\otimes\mathbf{X}^\top\right)\\
&= \left(\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\otimes\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)\left(\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\otimes\mathbf{I}_d\right)^{-1}\left(\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\oplus\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\right)^{-1}\\
&\quad\ \mathbf{N}_d\left(\mathbf{I}_d\otimes\mathbf{X}^\top\right)\\
&\underset{(a)}{=} \left(\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\otimes\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\right)\left(\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)\otimes\mathbf{I}_d + \left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\otimes\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\right)^{-1}\\
&\quad\ \mathbf{N}_d\left(\mathbf{I}_d\otimes\mathbf{X}^\top\right),
\end{aligned} \tag{63}$$

where (a) uses (67), (66) from Lemma 7.

Denote,

$$\mathbf{G}_1 \triangleq \left((\mathbf{X}^\top\mathbf{X}-\mathbf{B})\otimes\mathbf{I}_d+(\mathbf{X}^\top\mathbf{X}-\mathbf{B})^{1/2}\otimes(\mathbf{X}^\top\mathbf{X}-\mathbf{B})^{1/2}\right)^{-1}-\left((\mathbf{X}^\top\mathbf{X})\otimes\mathbf{I}_d+(\mathbf{X}^\top\mathbf{X})^{1/2}\otimes(\mathbf{X}^\top\mathbf{X})^{1/2}\right)^{-1}.$$

Plugging (62) and (63) to (59),

$$\mathbf{C}=\frac{1}{2}\mathbf{J}_1^{-1}(\mathbf{J}_2+\mathbf{J}_3)=\left((\mathbf{X}^\top\mathbf{X})^{1/2}\otimes\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1/2}\right)\mathbf{G}_1\mathbf{N}_d(\mathbf{I}_d\otimes\mathbf{X}^\top). \tag{64}$$

Using Lemma 30, $\mathbf{C}$ has the same non-zero eigenvalues as,

$$\mathbf{G}_1\mathbf{N}_d(\mathbf{I}_d\otimes\mathbf{X}^\top)\left((\mathbf{X}^\top\mathbf{X})^{1/2}\otimes\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1/2}\right)=\mathbf{G}_1\mathbf{G}_2, \tag{65}$$

where the equality uses (67) from Lemma 7 and $\mathbf{G}_2 \triangleq \mathbf{N}_d\left((\mathbf{X}^\top\mathbf{X})^{1/2}\otimes(\mathbf{X}^\top\mathbf{X})^{1/2}\right)$.

The Matrices $\mathbf{G}_1,\mathbf{G}_2$ are PSD

In order to prove that $\mathbf{G}_2$ is a PSD matrix, we prove that it is symmetric and that all its eigenvalues are non negative. Using Theorem 3.1 in Magnus and Neudecker (1979), $\mathbf{K}_d$ is a symmetric matrix with eigenvalues equal $\{-1,1\}$, so using observation 1.1.8 in Horn and Johnson (2012), $\mathbf{N}_d=\frac{1}{2}(\mathbf{I}_{d^2}+\mathbf{K}_d)$ is a PSD matrix. Using (68) from Lemma 7 and Theorem 2.1 in Magnus and Neudecker (2019) we get, that $\left((\mathbf{X}^\top\mathbf{X})^{1/2}\otimes(\mathbf{X}^\top\mathbf{X})^{1/2}\right)$ is PD. Using Lemma 31 we get that the eigenvalues of $\mathbf{G}_2$ are the same as the eigenvalues of,

$$\left((\mathbf{X}^\top\mathbf{X})^{1/2}\otimes(\mathbf{X}^\top\mathbf{X})^{1/2}\right)^{1/2}\mathbf{N}_d\left((\mathbf{X}^\top\mathbf{X})^{1/2}\otimes(\mathbf{X}^\top\mathbf{X})^{1/2}\right)^{1/2},$$

so the eigenvalues of $\mathbf{G}_2$ are non negative. Showing that $\mathbf{G}_2$ is symmetric,

$$\begin{aligned}\mathbf{G}_2^\top&=\left(\mathbf{N}_d\left((\mathbf{X}^\top\mathbf{X})^{1/2}\otimes(\mathbf{X}^\top\mathbf{X})^{1/2}\right)\right)^\top\\&=\left((\mathbf{X}^\top\mathbf{X})^{1/2}\otimes(\mathbf{X}^\top\mathbf{X})^{1/2}\right)^\top\mathbf{N}_d^\top\\&\underset{(a)}{=}\left((\mathbf{X}^\top\mathbf{X})^{1/2}\otimes(\mathbf{X}^\top\mathbf{X})^{1/2}\right)\mathbf{N}_d\\&\underset{(b)}{=}\mathbf{N}_d\left((\mathbf{X}^\top\mathbf{X})^{1/2}\otimes(\mathbf{X}^\top\mathbf{X})^{1/2}\right),\end{aligned}$$

where (a) uses that $\mathbf{N}_d$ is symmetric and (b) is from Theorem 3.9 in Magnus and Neudecker (2019).

Next, we prove that $\mathbf{G}_1$ is a PSD matrix. Using Lemma 33 it holds only if the following matrix is PSD,

$$\begin{aligned}\mathbf{H}_1&\triangleq\left((\mathbf{X}^\top\mathbf{X})\otimes\mathbf{I}_d+(\mathbf{X}^\top\mathbf{X})^{1/2}\otimes(\mathbf{X}^\top\mathbf{X})^{1/2}\right)-\left((\mathbf{X}^\top\mathbf{X}-\mathbf{B})\otimes\mathbf{I}_d+(\mathbf{X}^\top\mathbf{X}-\mathbf{B})^{1/2}\otimes(\mathbf{X}^\top\mathbf{X}-\mathbf{B})^{1/2}\right)\\&\underset{(a)}{=}\left((\mathbf{X}^\top\mathbf{X})-(\mathbf{X}^\top\mathbf{X}-\mathbf{B})\right)\otimes\mathbf{I}_d+(\mathbf{X}^\top\mathbf{X})^{1/2}\otimes(\mathbf{X}^\top\mathbf{X})^{1/2}-(\mathbf{X}^\top\mathbf{X}-\mathbf{B})^{1/2}\otimes(\mathbf{X}^\top\mathbf{X}-\mathbf{B})^{1/2}\\&\underset{(b)}{=}\underbrace{\mathbf{B}\otimes\mathbf{I}_d}_{\triangleq\mathbf{H}_{1a}}+\underbrace{\left((\mathbf{X}^\top\mathbf{X})\otimes(\mathbf{X}^\top\mathbf{X})\right)^{1/2}-\left((\mathbf{X}^\top\mathbf{X}-\mathbf{B})\otimes(\mathbf{X}^\top\mathbf{X}-\mathbf{B})\right)^{1/2}}_{\triangleq\mathbf{H}_{1b}},\end{aligned}$$

where (a) uses (66) from Lemma 7 and (b) uses Lemma 8.

Since $\mathbf{H}_{1a}$ is a PSD matrix, it is left to show that $\mathbf{H}_{1b}$ is a PSD matrix. Define $\mathbf{M}\triangleq\mathbf{X}^\top\mathbf{X}-\mathbf{B}$,

$$\begin{aligned}\mathbf{H}_{1b}&=\left((\mathbf{X}^\top\mathbf{X})\otimes(\mathbf{X}^\top\mathbf{X})\right)^{1/2}-(\mathbf{M}\otimes\mathbf{M})^{1/2}\\&=((\mathbf{M}+\mathbf{B})\otimes(\mathbf{M}+\mathbf{B}))^{1/2}-(\mathbf{M}\otimes\mathbf{M})^{1/2}\\&=\left(\underbrace{(\mathbf{M}\otimes\mathbf{M})}_{\triangleq\mathbf{K}}+\underbrace{(\mathbf{M}\otimes\mathbf{B}+\mathbf{B}\otimes(\mathbf{M}+\mathbf{B}))}_{\triangleq\mathbf{L}}\right)^{1/2}-\underbrace{(\mathbf{M}\otimes\mathbf{M})}_{\mathbf{K}}{}^{1/2},\end{aligned}$$

where the last equality uses (66) from Lemma 7.

The matrix $\mathbf{K}$ is PSD, the matrix $\mathbf{L}$ is PSD since it is an addition of two PSD matrices, thus using Lemma 34 $\mathbf{H}_{1b}$ is PSD and subsequently $\mathbf{G}_1$ is PSD.

C.4.3 Bounding the Jacobian Determinant

Using (65) and Lemma 31 we get that the eigenvalues of $\mathbf{C}$ are the same as the eigenvalues of $\mathbf{G}_1^{1/2}\mathbf{G}_2\mathbf{G}_1^{1/2}$, which are non negative, hence the demands of Lemma 10 are met. Finally,

$$\begin{aligned}
\frac{1}{|\mathrm{Det}\left(\mathbf{J}\right)|} &\underset{(a)}{\leq} \frac{1}{|\mathrm{Det}\left(\mathbf{J}_1\right)|} \\
&= \left|\mathrm{Det}\left(\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\otimes\mathbf{I}_n\right)^{-1}\right| \\
&\underset{(b)}{=} \mathrm{Det}\left(\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{1/2}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\otimes\mathbf{I}_n\right)^{-1} \\
&\underset{(c)}{=} \mathrm{Det}\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{-1/2}\otimes\mathbf{I}_n\right) \\
&\underset{(d)}{=} \left(\mathrm{Det}\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{1/2}\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)^{-1/2}\right)\right)^n\left(\mathrm{Det}\left(\mathbf{I}_n\right)\right)^d \\
&\underset{(e)}{=} \left(\frac{\mathrm{Det}\left(\mathbf{X}^\top\mathbf{X}\right)}{\mathrm{Det}\left(\mathbf{X}^\top\mathbf{X}-\mathbf{B}\right)}\right)^{n/2},
\end{aligned}$$

where (a) uses Lemma 10, (b) uses that all the eigenvalues are positive, (c) uses (69), (d) uses Corollary 2.2 in Magnus and Neudecker (2019) and (e) uses $\mathrm{Det}\left(\mathbf{I}\right)=1$. □

Lemma 6. *(Vectorization) Theorem 2.2 in Magnus and Neudecker (2019)*

For matrices $\mathbf{A}\in\mathbb{R}^{n\times d}$; $\mathbf{B},\mathbf{C}\in\mathbb{R}^{d\times d}$,

$$\begin{aligned}
\mathrm{vec}(\mathbf{ABC}) &= (\mathbf{C}^\top\mathbf{B}^\top\otimes\mathbf{I}_n)\mathrm{vec}(\mathbf{A}) \\
\mathrm{vec}(\mathbf{ABC}) &= (\mathbf{C}^\top\otimes\mathbf{A})\mathrm{vec}(\mathbf{B}) \\
\mathrm{vec}(\mathbf{ABC}) &= (\mathbf{I}_d\otimes\mathbf{AB})\mathrm{vec}(\mathbf{C}).
\end{aligned}$$

Lemma 7. *(Kronecker properties) From page 32 in Magnus and Neudecker (2019)*

If $\mathbf{A}+\mathbf{B}$ *and* $\mathbf{C}+\mathbf{D}$ *exist,*

$$(\mathbf{A}+\mathbf{B})\otimes(\mathbf{C}+\mathbf{D}) = \mathbf{A}\otimes\mathbf{C}+\mathbf{A}\otimes\mathbf{D}+\mathbf{B}\otimes\mathbf{C}+\mathbf{B}\otimes\mathbf{D}. \tag{66}$$

If $\mathbf{AC}$ *and* $\mathbf{BD}$ *exist,*

$$(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = (\mathbf{AC})\otimes(\mathbf{BD}). \tag{67}$$

$$(\mathbf{A}\otimes\mathbf{B})^\top = \mathbf{A}^\top\otimes\mathbf{B}^\top. \tag{68}$$

If $\mathbf{A}$ *and* $\mathbf{B}$ *are nonsingular,*

$$(\mathbf{A}\otimes\mathbf{B})^{-1} = \mathbf{A}^{-1}\otimes\mathbf{B}^{-1}. \tag{69}$$

Lemma 8. *(Square root of Kronecker product)*

Let $\mathbf{A},\mathbf{B}$ *be PSD matrices. Then,*

$$(\mathbf{A}\otimes\mathbf{B})^{1/2} = \mathbf{A}^{1/2}\otimes\mathbf{B}^{1/2}.$$

Proof. The square root of a matrix $\mathbf{X}$ is defined such that $\mathbf{X}=\mathbf{X}^{1/2}\mathbf{X}^{1/2}=\left(\mathbf{X}^{1/2}\right)^\top\mathbf{X}^{1/2}$ and equivalently, defined in terms of the eigenvalue decomposition $\mathbf{X}=\mathbf{U}\mathbf{\Sigma}_\mathbf{X}\mathbf{U}^\top$ as $\mathbf{X}^{1/2}=\mathbf{U}\mathbf{\Sigma}_\mathbf{X}^{1/2}\mathbf{U}^\top$. Using (67), $\left(\mathbf{A}^{1/2}\otimes\mathbf{B}^{1/2}\right)\left(\mathbf{A}^{1/2}\otimes\mathbf{B}^{1/2}\right)=\mathbf{A}\otimes\mathbf{B}$. Hence we can define $(\mathbf{A}\otimes\mathbf{B})^{1/2}=\left(\mathbf{A}^{1/2}\otimes\mathbf{B}^{1/2}\right)$. □

Lemma 9. *(Square root vectorization)*
Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ *be PD matrix. Then,*

$$\operatorname{vec}\left(d\mathbf{A}^{1/2}\right) = \left(\mathbf{A}^{1/2} \oplus \mathbf{A}^{1/2}\right)^{-1} \operatorname{vec}\left(d\mathbf{A}\right).$$

Proof.

$$\mathbf{A} = \mathbf{A}^{1/2}\mathbf{A}^{1/2}.$$

Taking the differential from both sides,

$$d\mathbf{A} = (d\mathbf{A}^{1/2})\mathbf{A}^{1/2} + \mathbf{A}^{1/2}(d\mathbf{A}^{1/2}),$$

$$\begin{aligned}
\operatorname{vec}(d\mathbf{A}) &\underset{(a)}{=} \left(\left(\left(\mathbf{A}^{1/2}\right)^{\top} \otimes \mathbf{I}_d\right) + \left(\mathbf{I}_d \otimes \mathbf{A}^{1/2}\right)\right) \operatorname{vec}(d\mathbf{A}^{1/2}) \\
&\underset{(b)}{=} \left(\left(\mathbf{A}^{1/2} \otimes \mathbf{I}_d\right) + \left(\mathbf{I}_d \otimes \mathbf{A}^{1/2}\right)\right) \operatorname{vec}(d\mathbf{A}^{1/2}) \\
&= \left(\mathbf{A}^{1/2} \oplus \mathbf{A}^{1/2}\right) \operatorname{vec}(d\mathbf{A}^{1/2}),
\end{aligned}$$

where (a) is from Lemma 6 and (b) uses that $\mathbf{A}$ is symmetric. Rearranging the equation, the proof follows. □

Lemma 10. *(Corollary 2.2 in Zhan (2005))*
Let $\mathbf{A}_1, \mathbf{B}_1 \in \mathbf{C}^{n\times n}$ $(n \geq 2)$. *If* $\mathbf{B}_1$ *is invertible and* $\operatorname{Re}\{\lambda_k\} \geq 0$ $(k = 1, 2, ..., n)$, *where* $\lambda(\mathbf{B}_1^{-1}\mathbf{A}_1) = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, *then,*

$$\left|\operatorname{Det}\left(\mathbf{A}_1 + \mathbf{B}_1\right)\right| \geq \left|\operatorname{Det}\left(\mathbf{A}_1\right)\right| + \left|\operatorname{Det}\left(\mathbf{B}_1\right)\right|.$$

C.5 Regret Incurred Under the "Bad event"

Lemma 11. *(Expected maximal regret incurred during a single instance)*

$$\mathop{\mathbb{E}}_{\theta}\left[\max_{\mu,\mathbf{\Sigma},A,\Xi}\left\{\mathrm{R}^{*}_{\mathrm{QB}_\tau}\left(\mu,\Sigma,T\right)\right\} \,\middle|\, \bar{\mathcal{E}}_\theta\right] \leq \frac{c_{bad}\sqrt{dT}}{11}, \quad c_{bad} \triangleq 22a\left(m + \sqrt{4\bar{\lambda}_{\mathbf{\Sigma}_*}\ln\left(\frac{d^2T}{\delta}\right)}\right).$$

Proof.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\theta}\left[\max_{\mu,\Sigma,A,\Xi}\left\{\mathrm{R}^{*}_{\mathrm{QB}_\tau}\left(\mu,\mathbf{\Sigma},T\right)\right\} \,\middle|\, \bar{\mathcal{E}}_\theta\right] &\underset{(a)}{\leq} \mathop{\mathbb{E}}_{\theta}\left[\sum_{t=1}^{T}\max_{A\in\mathcal{A}_t}\left\{A^{\top}\theta\right\} \,\middle|\, \bar{\mathcal{E}}_\theta\right] - \mathop{\mathbb{E}}_{\theta}\left[\sum_{t=1}^{T}\min_{A\in\mathcal{A}_t}\left\{A^{\top}\theta\right\} \,\middle|\, \bar{\mathcal{E}}_\theta\right] \\
&\leq 2\mathop{\mathbb{E}}_{\theta}\left[\sum_{t=1}^{T}\max_{A\in\mathcal{A}_t}\left\{\left|A^{\top}\theta\right|\right\} \,\middle|\, \bar{\mathcal{E}}_\theta\right] \\
&\underset{(b)}{\leq} 2\mathop{\mathbb{E}}_{\theta}\left[\sum_{t=1}^{T}\max_{A\in\mathcal{A}_t}\left\{\|A\|\,\|\theta\|\right\} \,\middle|\, \bar{\mathcal{E}}_\theta\right] \\
&\underset{(c)}{\leq} 2aT\mathop{\mathbb{E}}_{\theta}\left[\|\theta\| \mid \bar{\mathcal{E}}_\theta\right],
\end{aligned} \tag{70}$$

where (a) is the maximal regret of any algorithm, (b) uses Cauchy-schwarz inequality and (c) uses Assumption 1. Denote $Z \triangleq \mathbf{\Sigma}_*^{-1/2}(\theta - \mu_*)$ and analyzing the expectation,

$$\begin{aligned}
\mathbb{E}\left[\|\theta\| \mid \bar{\mathcal{E}}_\theta\right] &\underset{(a)}{\leq} \|\mu_*\| + \mathbb{E}\left[\|\theta - \mu_*\| \mid \bar{\mathcal{E}}_\theta\right] \\
&\underset{(b)}{\leq} m + \sqrt{\bar{\lambda}_{\mathbf{\Sigma}_*}}\,\mathbb{E}\left[\|Z\| \mid \bar{\mathcal{E}}_\theta\right] \\
&= m + \sqrt{\bar{\lambda}_{\mathbf{\Sigma}_*}}\,\mathbb{E}\left[\sqrt{\sum_{i=1}^{d} Z_i^2}\,\middle|\,\bar{\mathcal{E}}_\theta\right] \\
&\underset{(c)}{\leq} m + \sqrt{\bar{\lambda}_{\mathbf{\Sigma}_*}}\sqrt{\sum_{i=1}^{d}\mathbb{E}\left[Z_i^2 \mid \bar{\mathcal{E}}_\theta\right]},
\end{aligned} \tag{71}$$

where (a) uses the triangle inequality, (b) uses Lemma 26 and Assumptions 2 and 3, and (c) uses Jensen inequality.

The expectation in the last expression can be written as $\mathbb{E}\left[Z_i^2 \mid \bar{\mathcal{E}}_\theta\right] = \frac{\mathbb{E}\left[Z_i^2 \cdot \mathbb{1}\left\{Z_i^2 > z\right\}\right]}{\mathbb{P}\left(Z_i^2 > z\right)}$ for $z \triangleq \sqrt{2\ln\left(\frac{d^2T}{\delta}\right)}$. Taking note that Z_i is a standard normal variable and calculating the numerator,

$$\begin{aligned}
\mathbb{E}\left[Z_i^2 \cdot \mathbb{1}\left\{Z_i^2 > z\right\}\right] &\underset{(a)}{=} 2\mathbb{E}\left[Z_i^2 \cdot \mathbb{1}\left\{Z_i < -\sqrt{z}\right\}\right] \\
&= \frac{2}{\sqrt{2\pi}}\int_{-\infty}^{-\sqrt{z}} Z_i \cdot Z_i e^{-\frac{1}{2}Z_i^2} dZ_i \\
&\underset{(b)}{=} -\frac{2}{\sqrt{2\pi}} Z_i \cdot e^{-\frac{1}{2}Z_i^2}\Big|_{-\infty}^{-\sqrt{z}} + \frac{2}{\sqrt{2\pi}}\int_{-\infty}^{-\sqrt{z}} e^{-\frac{1}{2}Z_i^2} dZ_i \\
&\underset{(c)}{=} \sqrt{\frac{2z}{\pi}} \cdot e^{-\frac{1}{2}z} + 2\Phi\left(-\sqrt{z}\right),
\end{aligned} \tag{72}$$

where (a) uses the symmetry of a standard Gaussian distribution, (b) uses integration by parts and in (c) $\Phi(\cdot)$ stands for the standard Gaussian CDF.

Calculating the denominator,

$$\mathbb{P}\left(Z_i^2 > z\right) = \mathbb{P}\left(|Z_i| > \sqrt{z}\right) = 2\Phi\left(-\sqrt{z}\right). \tag{73}$$

Using the symmetry of a standard Gaussian distribution and (84) in Lemma 29,

$$\Phi\left(-\sqrt{z}\right) \geq \frac{1}{\sqrt{2\pi}}\frac{z-1}{z^{3/2}} e^{-\frac{1}{2}z}. \tag{74}$$

Bounding the second moment given $\bar{\mathcal{E}}_\theta$,

$$\begin{aligned}
\mathbb{E}\left[Z_i^2 \mid Z_i^2 > z\right] &= \frac{\mathbb{E}\left[Z_i^2 \cdot \mathbb{1}\left\{Z_i^2 > z\right\}\right]}{\mathbb{P}\left(Z_i^2 > z\right)} \\
&\underset{(a)}{=} \frac{\sqrt{\frac{2z}{\pi}} \cdot e^{-\frac{1}{2}z} + 2\Phi\left(-\sqrt{z}\right)}{2\Phi\left(-\sqrt{z}\right)} \\
&= \frac{\sqrt{\frac{z}{2\pi}} \cdot e^{-\frac{1}{2}z}}{\Phi\left(-\sqrt{z}\right)} + 1 \\
&\underset{(b)}{\leq} \frac{z^2}{z-1} + 1 \\
&\underset{(c)}{\leq} z + 3 \\
&\underset{(c)}{\leq} 4z,
\end{aligned}$$

where (a) uses (72), (73), (b) uses (74) and (c) uses $\ln\left(\frac{d^2T}{\delta}\right) \geq 2$

Plugging back to (71),

$$\mathbb{E}\left[\|\theta\| \mid \bar{\mathcal{E}}_\theta\right] \leq m + \sqrt{4\bar{\lambda}_{\mathbf{\Sigma}_*} d \sqrt{2\ln\left(\frac{d^2T}{\delta}\right)}} \leq \left(m + \sqrt{4\bar{\lambda}_{\mathbf{\Sigma}_*} \ln\left(\frac{d^2T}{\delta}\right)}\right)\sqrt{d}. \tag{75}$$

Plugging into (70) the proof follows. □

Finally, the regret incurred under the bad event,

$$\begin{aligned}
\mathbb{E}\left[\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\hat{\mu}_{\tau+1}, \hat{\mathbf{\Sigma}}_{\tau+1}, T-\tau\right)\mathbb{1}\left\{\bar{\mathcal{E}}\right\}\right] &\underset{(a)}{\leq} \mathbb{E}_\theta\left[\max_{\mu,\Sigma,\mathbf{A}_\tau}\left\{\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\mu, \mathbf{\Sigma}, T-\tau\right)\right\} \,\middle|\, \bar{\mathcal{E}}_\theta\right]\mathbb{E}\left[\mathbb{1}\left\{\bar{\mathcal{E}}\right\}\right] \\
&\leq \frac{9\delta}{dT}\mathbb{E}_\theta\left[\max_{\mu,\Sigma,\mathbf{A}_\tau}\left\{\mathrm{R}^*_{\mathrm{QB}_\tau}\left(\mu, \mathbf{\Sigma}, T-\tau\right)\right\} \,\middle|\, \bar{\mathcal{E}}_\theta\right] \\
&\underset{(b)}{\leq} \frac{9c_{\mathrm{bad}}\delta}{11\sqrt{d}},
\end{aligned} \tag{76}$$

where (a) uses that the events influence only the prior or the actions taken during the instance and (b) uses Lemma 11.

C.6 A Demonstration of Theorem 1

We demonstrate that using Theorem 1, a QB_τ algorithm with an adequate prior is a $(1+\alpha)$-*approximation* of KQB_τ, by presenting a case where k_1 is constant, using the following values of f_m, f_s, τ,

$$\tau = \max\left\{d, \frac{8a^2}{\lambda_{\mathbf{\Sigma}_\mathcal{A}}}\ln\left(\frac{d^2T}{\delta}\right)\right\}\ ;\ \|\hat{\mu}-\mu_*\| \leq \sqrt{f_m\delta} = \sqrt{d^2\delta}\ ;\ \left\|\hat{\mathbf{\Sigma}}-\mathbf{\Sigma}_*\right\|_{\mathrm{op}} \leq \sqrt{f_s\delta} = \sqrt{\delta}.$$

The value for τ ensures that $\mathcal{E}_v$ occurs with probability larger that $1-\frac{\delta}{dT}$ by Lemma 24. In order to find a valid value of δ, we first bound M (defined in Theorem 1),

$$M \leq \max\left\{\underbrace{3 + c_s^2\tau^2 f_s}_{\triangleq M_1}\ ,\ \underbrace{3 + 18c_\xi^2 c_s\left(f_m + \left(c_1 d + c_\xi^2 c_s/36\right)f_s\right)}_{\triangleq M_2}\right\}.$$

We begin by bounding M_1, M_2 separately,

$$\begin{aligned}
M_1 &\underset{(a)}{\leq} 3 + c_s^2\left(d^2 + \left(\frac{8a^2}{\lambda_{\mathbf{\Sigma}_\mathcal{A}}}\right)^2\ln^2\left(\frac{d^2T}{\delta}\right)\right) \\
&\underset{(b)}{\leq} \underbrace{3 + c_s^2 d^2 + 2c_s^2\left(\frac{8a^2}{\lambda_{\mathbf{\Sigma}_\mathcal{A}}}\right)^2\ln^2\left(d^2T\right)}_{\triangleq b_{M_1}} + \underbrace{2c_s^2\left(\frac{8a^2}{\lambda_{\mathbf{\Sigma}_\mathcal{A}}}\right)^2}_{\triangleq a_{M_1}}\ln^2\left(\frac{1}{\delta}\right), \\
M_2 &\underset{(c)}{\leq} 3 + 90\sigma^2\ln^2\left(\frac{dT}{\delta}\right)c_s\left(d^2 + \frac{4}{\lambda_{\mathbf{\Sigma}_*}}d + \frac{5\sigma^2 c_s}{36}\right) \\
&\underset{(b)}{\leq} \underbrace{3 + 180\sigma^2 c_s\left(d^2 + \frac{4d}{\lambda_{\mathbf{\Sigma}_*}} + \frac{5\sigma^2 c_s}{36}\right)\ln^2\left(dT\right)}_{\triangleq b_{M_2}} + \underbrace{180\sigma^2 c_s\left(d^2 + \frac{4d}{\lambda_{\mathbf{\Sigma}_*}} + \frac{5\sigma^2 c_s}{36}\right)}_{\triangleq a_{M_2}}\ln^2\left(\frac{1}{\delta}\right)
\end{aligned}$$

where (a) uses that $\max\{a^2, b^2\} \leq a^2 + b^2$, (b) uses the $(a+b)^2 \leq 2a^2 + 2b^2$ and (c) uses that $\ln\left(\frac{dT}{\delta}\right) \geq 1$ and $\ln\left(\frac{d^2T}{\delta}\right) \leq 2\ln\left(\frac{dT}{\delta}\right)$. Finally,

$$M \leq \max\left\{M_1, M_2\right\} \leq \left\{b_{M_1} + a_{M_1}\ln^2\left(\frac{1}{\delta}\right), b_{M_2} + a_{M_2}\ln^2\left(\frac{1}{\delta}\right)\right\} \triangleq b_M + a_M\ln^2\left(\frac{1}{\delta}\right).$$

Noticing that $a_M, b_M \in \tilde{\mathcal{O}}\left(d^2\right)$ and using the same derivations as in Appendix C.3 we choose,

$$\delta \triangleq \frac{1}{\left(8a_M^{1/2}\ln\left(4a_M^{1/2}\right)+2b_M^{1/2}\right)^2} \Rightarrow \begin{array}{l} \|\hat{\mu}-\mu_*\| \le \sqrt{d^2\delta} = \frac{d}{8a_M^{1/2}\ln\left(4a_M^{1/2}\right)+2b_M^{1/2}} \in \tilde{\mathcal{O}}(1) \\ \left\|\hat{\mathbf{\Sigma}}-\mathbf{\Sigma}_*\right\|_{\text{op}} \le \sqrt{\delta} = \frac{1}{8a_M^{1/2}\ln\left(4a_M^{1/2}\right)+2b_M^{1/2}} \in \tilde{\mathcal{O}}(1/d) \end{array}$$

Furthermore, for δ defined above,

$$k_1 = 12\sqrt{c_\xi^2 c_s}\sqrt{f_m\delta} + \left(c_s\tau + 12\sqrt{c_\xi^2 c_s c_1 d} + 2c_\xi^2 c_s\right)\sqrt{f_s\delta} \in \tilde{\mathcal{O}}(1),$$

The $\tilde{\mathcal{O}}$ notation indicates that k_1 is at most polylogarithmic in d, T. We can cancel this dependence by choosing $f_m = {}^{d^2}/_{k_1^2}, f_s = {}^{1}/_{k_1^2}$ instead, while preserving the same value of δ.

D PRIOR ESTIMATION ERROR

In the following section we prove that the prior formed by the MQB$_\tau$ algorithm meets the events of Theorem 1 with probability greater than $1 - \frac{8}{dnT}$.

D.1 Good Event Definition and Proof

The proof of Lemma 1 requires the events $\mathcal{E}_v$ to hold for every instance $j \in [n]$ where each event is denoted by $\mathcal{E}_{v_j}$. The MQB$_\tau$ version of the events, based on the events in (5), is defined as follows,

$$\begin{aligned} \mathcal{E}_{v(\text{MQB}_\tau)} &\triangleq \{\mathcal{E}_{v_j} \forall j \in [n]\}, \\ \mathcal{E}_{m(\text{MQB}_\tau)} &\triangleq \mathcal{E}_m \text{ with } \delta = \frac{1}{n-1}, \quad f_{m,n} = 3\left(\frac{2\sigma^2}{\lambda_{\mathbf{\Sigma}_\mathcal{A}} d} + \bar{\lambda}_{\mathbf{\Sigma}_*}\right)(d + \ln(dnT)), \\ \mathcal{E}_{s(\text{MQB}_\tau)} &\triangleq \mathcal{E}_s \text{ with } \delta = \frac{1}{n-1}, \quad f_{s,n} = 100^2\left(\frac{2\sigma^2}{\lambda_{\mathbf{\Sigma}_\mathcal{A}} d} + \bar{\lambda}_{\mathbf{\Sigma}_*}\right)^2(5d + 2\ln(dnT)), \\ \mathcal{E}_{n(\text{MQB}_\tau)} &\triangleq \{\mathcal{E}_{v(\text{MQB}_\tau)} \cap \mathcal{E}_{m(\text{MQB}_\tau)} \cap \mathcal{E}_{s(\text{MQB}_\tau)}\}. \end{aligned} \tag{77}$$

Lemma 1. *(MQB$_\tau$ conditions) For every instance $n > 10d + 4\ln(16dT)$, $\mathbb{P}\left(\mathcal{E}_{n(\text{MQB}_\tau)}\right) \ge 1 - {}^{8}/_{dnT}$.*

Proof. Using the union bound on Lemma 24 with $\delta = {}^{1}/_{N^2}$ for all instances up to the n_{th} instance,

$$\mathbb{P}\left(\mathcal{E}_{v(\text{MQB}_\tau)}\right) \ge 1 - \frac{1}{dNT}.$$

Lemma 15 and Lemma 22 for $n > 5d + 2\ln(dnT)$, with $\eta_n = \frac{1}{dnT}$ yield,

$$\mathbb{P}\left(\mathcal{E}_{m(\text{MQB}_\tau)}\right) \ge 1 - \frac{1}{dnT}; \quad \mathbb{P}\left(\mathcal{E}_{s(\text{MQB}_\tau)}\right) \ge 1 - \frac{6}{dnT}.$$

Using the union bound argument, for $n > 5d + 2\ln(dnT)$, $\mathbb{P}\left(\mathcal{E}_{n(\text{MQB}_\tau)}\right) \ge 1 - {}^{8}/_{dnT}$.

Next, using Lemma 38 with $a_1 = 2$, $b_1 = 5d + 2\ln(dT)$, we get that $n > 10d + 4\ln(16dT)$ implies $n > 5d + 2\ln(dnT)$. □

D.2 Mean Estimation Error

The mean estimation error originates from two different sources. First, after $n-1$ instances, the learner has interacted only with $n-1$ samples of the prior distribution $\mathcal{N}(\mu_*, \mathbf{\Sigma}_*)$. Second, at the end of each instance, she only has an estimator $\hat{\theta}_j$ for the true value of each sample θ_j. More formally,

$$\|\hat{\mu}_n - \mu_*\| = \left\|\frac{1}{n-1}\sum_{j=1}^{n-1}\left(\hat{\theta}_j - \mu_*\right)\right\| = \left\|\frac{1}{n-1}\sum_{j=1}^{n-1}\left(\underbrace{\hat{\theta}_j - \theta_j}_{\triangleq \rho_j} + \underbrace{\theta_j - \mu_*}_{\triangleq \Delta_j}\right)\right\|.$$

In order to bound the mean estimation error we first prove that each *inner instance error* ρ_j is unbiased and sub-Gaussian. Then, we show that a *single instance error* $\hat{\theta}_j - \mu_*$ is sub-Gaussian as well.

Lemma 12. *(Unbiasedness of the inner instance error)*

Under the event $\mathcal{E}_{v(\mathrm{MQB}_\tau)}$, for every instance $j \in [n]$,

$$\mathbb{E}\left[\rho_j\right] = \mathbb{E}\left[\hat{\theta}_j - \theta_j\right] = 0.$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\rho_j\right] &= \mathbb{E}\left[\hat{\theta}_j - \theta_j\right] \\
&= \mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1} \sum_{s=1}^{\tau} A_{j,s} x_{j,s} - \theta_j\right] \\
&= \mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1} \sum_{s=1}^{\tau} A_{j,s} \left(A_{j,s}^\top \theta_j + \xi_{j,s}\right) - \theta_j\right] \\
&= \mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1} \sum_{s=1}^{\tau} A_{j,s} \xi_{j,s}\right] \\
&= \sum_{s=1}^{\tau} \mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1} A_{j,s} \xi_{j,s}\right] \\
&\underset{(a)}{=} \sum_{s=1}^{\tau} \mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1} A_{j,s}\right] \mathbb{E}\left[\xi_{j,s}\right] \\
&\underset{(b)}{=} 0,
\end{aligned}$$

where (a) uses that the actions taken in the first τ time-steps are independent from the noise terms and (b) uses that $\xi_{j,s}$ is a zero-mean noise. □

Lemma 13. *(Sub-Gaussianity of the inner instance error)*

Under the event $\mathcal{E}_{v(\mathrm{MQB}_\tau)}$, for every instance $j \in [n]$, ρ_j is a $\sqrt{\frac{2\sigma^2}{\lambda_{\mathbf{\Sigma}_\mathcal{A}} d}}$ sub-Gaussian vector.

Proof. For any $s \in \mathbb{R}$ and $U \in \mathbb{R}^d$ s.t. $\|U\| = 1$

$$\begin{aligned}
\mathbb{E}\left[\exp\left(sU^\top \rho_j\right)\right] &= \mathbb{E}\left[\exp\left(sU^\top \mathbf{V}_{j,\tau}^{-1} \sum_{s=1}^{\tau} A_{j,s} \xi_{j,s}\right)\right] \\
&\underset{(a)}{=} \mathop{\mathbb{E}}_{\mathbf{A}_{j,\tau}}\left[\mathbb{E}\left[\exp\left(sU^\top \mathbf{V}_{j,\tau}^{-1} \sum_{s=1}^{\tau} A_{j,s} \xi_{j,s}\right)\right] \,\middle|\, \mathbf{A}_{j,\tau}\right] \\
&\underset{(b)}{\leq} \mathbb{E}\left[\exp\left(\frac{s^2\sigma^2}{2} \sum_{s=1}^{\tau} \left(U^\top \mathbf{V}_{j,\tau}^{-1} A_{j,s}\right)^2\right)\right] \\
&= \mathbb{E}\left[\exp\left(\frac{s^2\sigma^2}{2} \sum_{s=1}^{\tau} A_{j,s}^\top \mathbf{V}_{j,\tau}^{-1} U U^\top \mathbf{V}_{j,\tau}^{-1} A_{j,s}\right)\right] \\
&\underset{(c)}{=} \mathbb{E}\left[\exp\left(\frac{s^2\sigma^2}{2} \mathrm{Tr}\left(\sum_{s=1}^{\tau} A_{j,s} A_{j,s}^\top \mathbf{V}_{j,\tau}^{-1} U U^\top \mathbf{V}_{j,\tau}^{-1}\right)\right)\right] \\
&\underset{(c)}{=} \mathbb{E}\left[\exp\left(\frac{s^2\sigma^2}{2} U^\top \mathbf{V}_{j,\tau}^{-1} U\right)\right] \\
&\underset{(d)}{\leq} \exp\left(\frac{s^2}{2} \cdot \frac{2\sigma^2}{\lambda_{\mathbf{\Sigma}_\mathcal{A}} d}\right),
\end{aligned}$$

where (a) uses the law of total expectation, (b) uses that the actions taken in the first τ time-steps are independent from the noise terms, the MGF of a Gaussian variable and the law of total expectation, (c) uses the linearity and the product properties of the trace and (d) uses Lemma 26 and the event $\mathcal{E}_{v(\mathrm{MQB}_\tau)}$. □

Lemma 14. *(Sub-Gaussianity of the single instance error)*

Under the event $\mathcal{E}_{v(\mathrm{MQB}_\tau)}$, for every instance $j \in [n]$, $\hat{\theta}_j - \mu_*$ is a $\sqrt{\frac{2\sigma^2}{\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}} d} + \bar{\lambda}_{\mathbf{\Sigma}_*}}$ sub-Gaussian vector.

Proof. For any $s \in \mathbb{R}$ and $U \in \mathbb{R}^d$ s.t. $\|U\| = 1$

$$\begin{aligned}
\mathbb{E}\left[\exp\left(sU^\top\left(\hat{\theta}_j - \mu_*\right)\right)\right] &\underset{(a)}{=} \mathbb{E}\left[\exp\left(sU^\top \rho_j\right)\exp\left(sU^\top \Delta_j\right)\right] \\
&\underset{(b)}{=} \mathbb{E}\left[\exp\left(sU^\top \rho_j\right)\right]\mathbb{E}\left[\exp\left(sU^\top \Delta_j\right)\right] \\
&\underset{(c)}{\leq} \exp\left(\frac{s^2}{2}\cdot\frac{2\sigma^2}{\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}} d}\right)\exp\left(\frac{s^2 U^\top \mathbf{\Sigma}_* U}{2}\right) \\
&\underset{(d)}{\leq} \exp\left(\frac{s^2}{2}\left(\frac{2\sigma^2}{\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}} d} + \bar{\lambda}_{\mathbf{\Sigma}_*}\right)\right),
\end{aligned} \tag{78}$$

where (a) uses the definitions of ρ_j, Δ_j, (b) uses that the actions taken during the first τ time-steps are independent of the noise terms and θ_j, (c) uses the MGF of the Gaussian variable $U^\top \Delta_j$ and Lemma 13 and (d) uses Lemma 26, together with Assumption 2. □

The following lemma bounds the mean estimation error of MQB_τ algorithm with high probability.

Lemma 15. *(Mean estimation error)*

For every $\eta_n > 0$ and for every instance $n > 1$,

$$\mathbb{P}\left(\mathcal{E}_{m(\mathrm{MQB}_\tau)} \mid \mathcal{E}_{v(\mathrm{MQB}_\tau)}\right) \geq 1 - \eta_n.$$

Proof. We start by proving that $\hat{\mu}_n - \mu_*$ is a $\sqrt{\frac{2\sigma^2 + \underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}} \bar{\lambda}_{\mathbf{\Sigma}_*} d}{\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}} d(n-1)}}$ sub-Gaussian vector. For any $s \in \mathbb{R}$ and $U \in \mathbb{R}^d$ s.t. $\|U\| = 1$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(sU^\top\left(\hat{\mu}_n - \mu_*\right)\right)\right] &= \mathbb{E}\left[\exp\left(\frac{s}{n-1}U^\top \sum_{j=1}^{n-1}\left(\hat{\theta}_j - \mu_*\right)\right)\right] \\
&\underset{(a)}{=} \prod_{j=1}^{n-1}\mathbb{E}\left[\exp\left(\frac{s}{n-1}U^\top\left(\rho_j + \Delta_j\right)\right)\right] \\
&\underset{(b)}{\leq} \exp\left(\frac{s^2}{2}\cdot\frac{2\sigma^2 + \underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}}\bar{\lambda}_{\mathbf{\Sigma}_*} d}{\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}} d(n-1)}\right),
\end{aligned}$$

where (a) uses that the actions taken during the first τ time-steps are independent of the noise terms, θ_j and the inner instance errors of other instances and (b) uses the same steps as in (78).

From Lemma 28,

$$\mathbb{P}\left(\|\hat{\mu}_n - \mu_*\|^2 > \frac{2\sigma^2 + \underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}}\bar{\lambda}_{\mathbf{\Sigma}_*} d}{\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}} d(n-1)} \cdot \left(d + 2\sqrt{d\ln\left(1/\eta_n\right)} + 2\ln\left(1/\eta_n\right)\right)\right) \leq \eta_n$$

and using the inequality of arithmetic and geometric means, $\sqrt{d\ln\left(1/\eta_n\right)} \leq \left(d + \ln\left(1/\eta_n\right)\right)/2$, the proof follows. □

D.3 Covariance Estimation Error

In this section, we bound the estimation error of the estimated covariance matrix $\hat{\mathbf{\Sigma}}_n$ (Eq. (11)), WRT the true covariance matrix $\mathbf{\Sigma}_*$, under the operator norm.

D.3.1 Covariance Estimation Error Decomposition

Define,

$$\boldsymbol{\Sigma} \triangleq \frac{\sigma^2}{n-1}\sum_{j=1}^{n-1}\mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\right] + \boldsymbol{\Sigma}_*, \tag{79}$$

$$\begin{aligned}
&\left\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_*\right\|_{\text{op}} \\
&= \left\|\frac{1}{n-2}\sum_{j=1}^{n-1}\left(\hat{\theta}_j - \hat{\mu}_n\right)\left(\hat{\theta}_j - \hat{\mu}_n\right)^\top - \frac{\sigma^2}{n-1}\sum_{j=1}^{n-1}\mathbf{V}_{j,\tau}^{-1} - \boldsymbol{\Sigma}_*\right\|_{\text{op}} \\
&\leq \left\|\frac{1}{n-2}\sum_{j=1}^{n-1}\left(\hat{\theta}_j - \hat{\mu}_n\right)\left(\hat{\theta}_j - \hat{\mu}_n\right)^\top - \boldsymbol{\Sigma}\right\|_{\text{op}} + \left\|\frac{\sigma^2}{n-1}\left(\sum_{j=1}^{n-1}\mathbf{V}_{j,\tau}^{-1} - \sum_{j=1}^{n-1}\mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\right]\right)\right\|_{\text{op}} \\
&\underset{(a)}{=} \left\|\frac{1}{n-2}\sum_{j=1}^{n-1}\left(\hat{\theta}_j - \mu_*\right)\left(\hat{\theta}_j - \mu_*\right)^\top - \frac{n-1}{n-2}\left(\hat{\mu}_n - \mu_*\right)\left(\hat{\mu}_n - \mu_*\right)^\top - \boldsymbol{\Sigma}\right\|_{\text{op}} \\
&\qquad + \left\|\frac{\sigma^2}{n-1}\left(\sum_{j=1}^{n-1}\mathbf{V}_{j,\tau}^{-1} - \sum_{j=1}^{n-1}\mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\right]\right)\right\|_{\text{op}} \\
&\leq \underbrace{\frac{n-1}{n-2}\left\|\frac{1}{n-1}\sum_{j=1}^{n-1}\left(\hat{\theta}_j - \mu_*\right)\left(\hat{\theta}_j - \mu_*\right)^\top - \boldsymbol{\Sigma}\right\|_{\text{op}}}_{\text{Term A}} + \underbrace{\frac{n-1}{n-2}\left\|\frac{1}{n-1}\sum_{j=1}^{n-1}\left(\hat{\mu}_n - \mu_*\right)\left(\hat{\mu}_n - \mu_*\right)^\top - \frac{1}{n-1}\boldsymbol{\Sigma}\right\|_{\text{op}}}_{\text{Term B}} \\
&\qquad + \underbrace{\left\|\frac{\sigma^2}{n-1}\left(\sum_{j=1}^{n-1}\mathbf{V}_{j,\tau}^{-1} - \sum_{j=1}^{n-1}\mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\right]\right)\right\|_{\text{op}}}_{\text{Term C}},
\end{aligned} \tag{80}$$

where (a) uses $\hat{\mu}_n = \frac{1}{n-1}\sum_{j=1}^{n-1}\hat{\theta}_j$ and both inequalities come from the triangle inequality.

We bound each of the three terms separately, using Theorem 6.5 in Wainwright (2019) for terms A and B. For convenience, we state it here.

Lemma 16. *(Empirical covariance bounds, Theorem 6.5 in Wainwright (2019), constants were taken from Bastani et al. (2021)).*

For any row-wise σ sub-Gaussian random matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, the sample covariance matrix $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top$ satisfies the bound

$$\mathbb{P}\left(\left\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_{\text{op}} \geq 32\sigma^2 \cdot \max\left\{\sqrt{\frac{5d + 2\ln\left(\frac{2}{\delta}\right)}{n}}, \frac{5d + 2\ln\left(\frac{2}{\delta}\right)}{n}\right\}\right) \leq \delta \quad \forall\, 0 < \delta < 1.$$

The first part of Terms A and B in (80) are of the form $\frac{1}{n-1}\sum_{j=1}^{n-1} XX^\top$ for vector X equal to $\hat{\theta}_j - \mu_*$ and $\hat{\mu}_n - \mu_*$, respectively. Lemmas 14 and 15, proved during the mean estimation analysis (Appendix D.2), show that these vectors are sub-Gaussian. The proof of Lemma 16, implicitly requires that $\mathbb{E}\left[\hat{\boldsymbol{\Sigma}}\right] = \boldsymbol{\Sigma}$, i.e. the estimator $\hat{\boldsymbol{\Sigma}}$ is unbiased, which we prove next.

D.3.2 Lack of Bias of the Covariance Estimator

Lemma 17. *(Lack of bias of the covariance estimator - auxiliary)*

$$\mathbb{E}\left[\rho_j \rho_j^\top\right] = \sigma^2 \mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\right].$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\rho_j\rho_j^\top\right] &= \mathbb{E}\left[\left(\mathbf{V}_{j,\tau}^{-1}\sum_{s=1}^{\tau}A_{j,s}\xi_{j,s}\right)\left(\mathbf{V}_{j,\tau}^{-1}\sum_{k=1}^{\tau}A_{j,k}\xi_{j,k}\right)^\top\right]\\ &\underset{(a)}{=} \mathbb{E}\left[\left(\mathbf{V}_{j,\tau}^{-1}\sum_{s=1}^{\tau}A_{j,s}\xi_{j,s}\right)\left(\sum_{k=1}^{\tau}A_{j,k}^\top\mathbf{V}_{j,\tau}^{-1}\xi_{j,k}\right)\right]\\ &= \underbrace{\mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\sum_{s=1}^{\tau}A_{j,s}A_{j,s}^\top\mathbf{V}_{j,\tau}^{-1}\xi_{j,s}^2\right]}_{\text{same time-step}} + \underbrace{\mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\sum_{s=1}^{\tau}\sum_{k\neq s}^{\tau}A_{j,s}A_{j,k}^\top\mathbf{V}_{j,\tau}^{-1}\xi_{j,s}\xi_{j,k}\right]}_{\text{different time-steps}},\end{aligned} \tag{81}$$

where (a) uses that $\mathbf{V}_{j,\tau}$ is symmetric.

Analyzing the *same time-step*

$$\begin{aligned}\mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\sum_{s=1}^{\tau}A_{j,s}A_{j,s}^\top\mathbf{V}_{j,\tau}^{-1}\xi_{j,s}^2\right] &\underset{(a)}{=} \mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\sum_{s=1}^{\tau}A_{j,s}A_{j,s}^\top\mathbf{V}_{j,\tau}^{-1}\right]\mathbb{E}\left[\xi_{j,s}^2\right]\\ &= \sigma^2\mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\left(\sum_{s=1}^{\tau}A_{j,s}A_{j,s}^\top\right)\mathbf{V}_{j,\tau}^{-1}\right]\\ &= \sigma^2\mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\right].\end{aligned}$$

Analyzing the *different time-steps*

$$\mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\sum_{s=1}^{\tau}\sum_{k\neq s}^{\tau}A_{j,s}A_{j,k}^\top\mathbf{V}_{j,\tau}^{-1}\xi_{j,s}\xi_{j,k}\right] \underset{(a)}{=} \sum_{s=1}^{\tau}\sum_{k\neq s}^{\tau}\mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}A_{j,s}A_{j,k}^\top\mathbf{V}_{j,\tau}^{-1}\right]\mathbb{E}\left[\xi_{j,s}\right]\mathbb{E}\left[\xi_{j,k}\right] = 0,$$

where (a) uses that the actions during the first τ time-steps are independent of the noise terms.

Plugging back to (81) the proof follows. □

Lemma 18. *(Lack of bias of the covariance estimator - Term A)*

$$\mathbb{E}\left[\frac{1}{n-1}\sum_{j=1}^{n-1}\left(\hat{\theta}_j-\mu_*\right)\left(\hat{\theta}_j-\mu_*\right)^\top\right] = \mathbf{\Sigma}.$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\left(\hat{\theta}_j-\mu_*\right)\left(\hat{\theta}_j-\mu_*\right)^\top\right] &= \mathbb{E}\left[\hat{\theta}_j\hat{\theta}_j^\top\right] - \mathbb{E}\left[\hat{\theta}_j\right]\mu_*^\top - \mu_*\mathbb{E}\left[\hat{\theta}_j^\top\right] + \mu_*\mu_*^\top\\ &\underset{(a)}{=} \mathbb{E}\left[\left(\theta_j+\rho_j\right)\left(\theta_j+\rho_j\right)^\top\right] - \mu_*\mu_*^\top\\ &\underset{(b)}{=} \mathbb{E}\left[\theta_j\theta_j^\top\right] - \mu_*\mu_*^\top + \mathbb{E}\left[\rho_j\rho_j^\top\right]\\ &\underset{(c)}{=} \mathbf{\Sigma}_* + \sigma^2\mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\right],\end{aligned}$$

where (a) uses that $\hat{\theta}_j$ is unbiased i.e. $\mathbb{E}\left[\hat{\theta}_j\right] = \mu_*$, (b) uses that the actions taken during the first τ time-steps are independent of the noise terms and of θ_j and Lemma 12 and (c) uses Lemma 17.

Using the definition of $\mathbf{\Sigma}$ in (79), summing over the instances and dividing by $n-1$ the proof follows. □

Lemma 19. *(Lack of bias of the covariance estimator - Term B - auxiliary)*

$$\mathbb{E}\left[\hat{\mu}_n\hat{\mu}_n^\top\right] = \frac{\mathbf{\Sigma}}{n-1} + \mu_*\mu_*^\top.$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\hat{\mu}_n\hat{\mu}_n^\top\right] &= \mathbb{E}\left[\left(\frac{1}{n-1}\sum_{i=1}^{n-1}\hat{\theta}_i\right)\left(\frac{1}{n-1}\sum_{i=1}^{n-1}\hat{\theta}_i\right)^\top\right] \\
&= \left(\frac{1}{n-1}\right)^2\mathbb{E}\left[\left(\sum_{i=1}^{n-1}\theta_i+\rho_i\right)\left(\sum_{i=1}^{n-1}\theta_i+\rho_i\right)^\top\right] \\
&\underset{(a)}{=} \left(\frac{1}{n-1}\right)^2\mathbb{E}\left[\sum_{i=1}^{n-1}\theta_i\theta_i^\top+\sum_{i=1}^{n-1}\sum_{j\neq i}\theta_i\theta_j^\top+\sum_{i=1}^{n-1}\rho_i\rho_i^\top\right] \\
&\underset{(b)}{=} \frac{\mathbf{\Sigma}_*+\mu_*\mu_*^\top}{n-1}+\frac{n-2}{n-1}\mu_*\mu_*^\top+\frac{\sigma^2\sum_{i=1}^{n-1}\mathbb{E}\left[\mathbf{V}_{i,\tau}^{-1}\right]}{(n-1)^2} \\
&= \frac{\mathbf{\Sigma}_*}{n-1}+\mu_*\mu_*^\top+\frac{\sigma^2\sum_{i=1}^{n-1}\mathbb{E}\left[\mathbf{V}_{i,\tau}^{-1}\right]}{(n-1)^2},
\end{aligned}$$

where (a) uses Lemma 12 and that the actions taken during the first τ time-steps are independent of the noise terms, of θ_j and of other inner instance errors and (b) uses Lemma 17.

Using the definition of $\mathbf{\Sigma}$ in (79) the proof follows. □

Lemma 20. *(Lack of bias of the covariance estimator - Term B)*

$$\mathbb{E}\left[\left(\hat{\mu}_n-\mu_*\right)\left(\hat{\mu}_n-\mu_*\right)^\top\right]=\frac{\mathbf{\Sigma}}{n-1}.$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{\mu}_n-\mu_*\right)\left(\hat{\mu}_n-\mu_*\right)^\top\right] &= \mathbb{E}\left[\hat{\mu}_n\hat{\mu}_n^\top\right]-\mathbb{E}\left[\hat{\mu}_n\right]\mu_*^\top-\mu_*\mathbb{E}\left[\hat{\mu}_n^\top\right]+\mu_*\mu_*^\top \\
&= \mathbb{E}\left[\hat{\mu}_n\hat{\mu}_n^\top\right]-\mu_*\mu_*^\top \\
&\underset{(a)}{=} \frac{\mathbf{\Sigma}}{n-1}+\mu_*\mu_*^\top-\mu_*\mu_*^\top,
\end{aligned}$$

where (a) uses Lemma 19. □

Lemma 21. *(Covariance estimation error - Term C)*

For every instance $n>1$, $\delta>0$, with probability greater than $1-\delta$,

$$\left\|\frac{\sigma^2}{n-1}\left(\sum_{j=1}^{n-1}\mathbf{V}_{j,\tau}^{-1}-\sum_{j=1}^{n-1}\mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\right]\right)\right\|_{\text{op}}\leq 4\cdot\frac{\sigma^2}{\lambda_{\mathbf{\Sigma}_\mathcal{A}}d}\sqrt{\frac{2\ln d+2\ln\left(2/\delta\right)}{n-1}}.$$

Proof. The maximal eigenvalue of $\mathbf{V}_{j,\tau}^{-1}-\mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\right]$ is upper bounded by,

$$\begin{aligned}
\lambda_{\max}\left(\mathbf{V}_{j,\tau}^{-1}-\mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\right]\right) &\underset{(a)}{\leq} \lambda_{\max}\left(\mathbf{V}_{j,\tau}^{-1}\right)+\lambda_{\max}\left(-\mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\right]\right) \\
&\underset{(b)}{\leq} \lambda_{\max}\left(\mathbf{V}_{j,\tau}^{-1}\right) \\
&\underset{(c)}{\leq} \frac{2}{\lambda_{\mathbf{\Sigma}_\mathcal{A}}d},
\end{aligned}$$

where (a) uses Weyl's inequality, (b) uses that $\mathbf{V}_{j,\tau}$ is PD and (c) uses the event $\mathcal{E}_{v(\text{MQB}_\tau)}$.

Using Lemma 27 with $\mathbf{X}_j \triangleq \mathbf{V}_{j,\tau}^{-1} - \mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\right], \mathbf{A}_j^2 \triangleq \frac{4}{\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}}^2 d^2}\mathbf{I}$ and $s^2 = \frac{4(n-1)}{\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}}^2 d^2}$, we get,

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{j=1}^{n-1}\mathbf{V}_{j,\tau}^{-1} - \mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\right]\right) \geq \delta\right) \leq d\exp\left(-\frac{\left(\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}} d\right)^2 \delta^2}{32(n-1)}\right),$$

multiplying the RHS by 2 to achieve the spectral norm bound, rearranging and multiplying by $\sigma^2/n-1$ the proof follows. □

D.3.3 Covariance Estimation Error Bound

Lemma 22. *(Covariance estimation error)*

For every $\eta_n > 0$ and for every instance $n > 5d + 2\ln(1/\eta_n)$,

$$\mathbb{P}\left(\mathcal{E}_{s(\mathrm{MQB}_\tau)} \mid \mathcal{E}_{v(\mathrm{MQB}_\tau)}\right) \geq 1 - 6\eta_n.$$

Proof. Using Lemma 16 on terms A and B in (80), together with Lemma 21 on Term C and a union bound argument, we get that for $n > 5d + 2\ln(1/\eta_n)$, the three following expressions exist in probability greater than $1 - 6\eta_n$,

$$\left\|\frac{1}{n-1}\sum_{j=1}^{n-1}\left(\hat{\theta}_j - \mu_*\right)\left(\hat{\theta}_j - \mu_*\right)^\top - \mathbf{\Sigma}\right\|_{\mathrm{op}} \leq 32 \cdot \frac{2\sigma^2 + \underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}}\bar{\lambda}_{\mathbf{\Sigma}_*} d}{\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}} d} \cdot \sqrt{\frac{5d + 2\ln(1/\eta_n)}{n-1}},$$

$$\left\|\frac{1}{n-1}\sum_{j=1}^{n-1}\left(\hat{\mu}_n - \mu_*\right)\left(\hat{\mu}_n - \mu_*\right)^\top - \frac{\mathbf{\Sigma}}{n-1}\right\|_{\mathrm{op}} \leq 32 \cdot \frac{2\sigma^2 + \underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}}\bar{\lambda}_{\mathbf{\Sigma}_*} d}{\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}} d(n-1)} \cdot \sqrt{\frac{5d + 2\ln(1/\eta_n)}{n-1}},$$

$$\left\|\frac{\sigma^2}{n-1}\left(\sum_{j=1}^{n-1}\mathbf{V}_{j,\tau}^{-1} - \sum_{j=1}^{n-1}\mathbb{E}\left[\mathbf{V}_{j,\tau}^{-1}\right]\right)\right\|_{\mathrm{op}} \leq 4 \cdot \frac{\sigma^2}{\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}} d}\sqrt{\frac{2\ln d + 2\ln(1/\eta_n)}{n-1}}.$$

Plugging back to (80),

$$\begin{aligned}\left\|\hat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}_*\right\|_{\mathrm{op}} &\leq 32\sqrt{\frac{5d + 2\ln(1/\eta_n)}{n-1}}\left(\frac{n}{n-2}\left(\frac{2\sigma^2 + \underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}}\bar{\lambda}_{\mathbf{\Sigma}_*} d}{\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}} d}\right) + \frac{\sigma^2}{8\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}} d}\right) \\ &\leq 50\sqrt{\frac{5d + 2\ln(1/\eta_n)}{n-1}}\left(\frac{2\sigma^2}{\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}} d} + \bar{\lambda}_{\mathbf{\Sigma}_*}\right),\end{aligned}$$

where the last inequality uses that $\frac{n}{n-2} \leq 3/2$ for $n \geq 6$.

Using the covariance widening scheme suggested by Bastani et al. (2021) and proved in Lemma 25 we get,

$$\left\|\hat{\mathbf{\Sigma}}_n^{\mathrm{w}} - \mathbf{\Sigma}_*\right\|_{\mathrm{op}} \leq 100\sqrt{\frac{5d + 2\ln(1/\eta_n)}{n-1}}\left(\frac{2\sigma^2}{\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}} d} + \bar{\lambda}_{\mathbf{\Sigma}_*}\right), \quad \hat{\mathbf{\Sigma}}_n^{\mathrm{w}} \succeq \mathbf{\Sigma}_*.$$

□

E META ALGORITHM REGRET

In the following section we prove Theorem 2.

Theorem 2. *For $N_0 \leq N$, the MQB_τ N-instance relative regret is bounded by,*

$$\begin{aligned}&\sum_{n=1}^{N}\mathbb{E}\left[\mathrm{R}_{\mathrm{QB}_\tau}^{\mathrm{K}}\left(\hat{\mu}_n, \hat{\mathbf{\Sigma}}_n^{\mathrm{w}}, T\right)\right] \\ &\leq \underbrace{k_2\sqrt{N}\mathbb{E}\left[\mathrm{R}_{\mathrm{QB}_\tau}^{*}\left(\mu_{*,\tau+1}, \mathbf{\Sigma}_{*,\tau+1}, T-\tau\right)\right]}_{\substack{\text{regular}\\ \text{instances}}} + \underbrace{N_0\mathrm{R}_{\mathrm{sing}}}_{\substack{\text{exploration}\\ \text{instances}}},\end{aligned}$$

$$
\begin{aligned}
k_2 &\triangleq 24\sqrt{c_\xi^2 c_s}\sqrt{3\left(2\sigma^2/(\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}} d)+\bar{\lambda}_{\mathbf{\Sigma}_*}\right)}\sqrt{d+\ln\left(dNT\right)}+4c_{\mathrm{w}}\left[c_s\tau+12\sqrt{c_\xi^2 c_s c_1 d}+2c_\xi^2 c_s\right]\sqrt{5d+2\ln\left(dNT\right)},\\
N_0 &= \max\Big\{3, 4c_{\mathrm{w}}^2 c_s^2\tau^2\left(5d+2\ln\left(d(N+1)T\right)\right),\\
&\qquad 18c_\xi^2 c_s\left(3\left(2\sigma^2/(\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}} d)+\bar{\lambda}_{\mathbf{\Sigma}_*}\right)\left(d+\ln\left(d(N+1)T\right)\right)+4c_{\mathrm{w}}^2\left(c_1 d+c_\xi^2 c_s/36\right)\left(5d+2\ln\left(d(N+1)T\right)\right)\right)\Big\},
\end{aligned}
\tag{82}
$$

where τ defined in (8), c_{w} in (12) and the other constants are as in (21) with $\delta = 1/N$.

Proof.

$$
\begin{aligned}
\sum_{n=1}^{N}\mathbb{E}\left[\mathrm{R}^{\mathrm{K}}_{\mathrm{QB}_\tau}\left(\hat{\mu}_n,\hat{\mathbf{\Sigma}}^{\mathrm{w}}_n,T\right)\right] &= \underbrace{\sum_{n=1}^{N_0}\mathbb{E}\left[\mathrm{R}^{\mathrm{K}}_{\mathrm{QB}_\tau}\left(\hat{\mu}_n,\hat{\mathbf{\Sigma}}^{\mathrm{w}}_n,T\right)\right]}_{\text{exploration instances}}+\underbrace{\sum_{n=N_0+1}^{N}\mathbb{E}\left[\mathrm{R}^{\mathrm{K}}_{\mathrm{QB}_\tau}\left(\hat{\mu}_n,\hat{\mathbf{\Sigma}}^{\mathrm{w}}_n,T\right)\right]}_{\text{regular instances}}\\
&\underset{(a)}{\leq} N_0\mathrm{R}_{\exp}+\sum_{n=N_0+1}^{N}\left(\frac{\tilde{k}_2(n)}{2\sqrt{n-1}}\mathbb{E}\left[\mathrm{R}^{*}_{\mathrm{QB}_\tau}\left(\mu_{*,\tau+1},\mathbf{\Sigma}_{*,\tau+1},T-\tau\right)\right]+\frac{c_{\mathrm{bad}}}{\sqrt{d}(n-1)}\right)\\
&\underset{(a)}{\leq} N_0\mathrm{R}_{\exp}+k_2\sqrt{N}\mathbb{E}\left[\mathrm{R}^{*}_{\mathrm{QB}_\tau}\left(\mu_{*,\tau+1},\mathbf{\Sigma}_{*,\tau+1},T-\tau\right)\right],
\end{aligned}
\tag{83}
$$

where (a) uses Lemma 1 and Theorem 1 with $\tilde{k}_2(n) \triangleq 2k_1(n)/\sqrt{\delta(n)}$ for the regular instances and (b) uses $k_2 = \tilde{k}_2(N)$ and $\sum_{n=N_0+1}^{N}\frac{1}{\sqrt{n-1}} \leq \int_{N_0}^{N}\frac{1}{\sqrt{x-1}}dx \leq 2\sqrt{N}$. □

F AUXILIARY LEMMAS

Lemma 23. *(Matrix Chernoff, Theorem 5.1.1 in Tropp (2015))*

Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random, Hermitian matrices with common dimension d. Assume that,

$$0 \leq \lambda_{\min}\left(\mathbf{X}_k\right) \text{ and } \lambda_{\max}\left(\mathbf{X}_k\right) \leq L \text{ for each index } k.$$

Then for every $0 \leq t < 1$,

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_k \mathbf{X}_k\right) \leq t\lambda_{\min}\left(\sum_k \mathbb{E}\left[\mathbf{X}_k\right]\right)\right) \leq de^{-\frac{(1-t)^2\lambda_{\min}\left(\sum_k \mathbb{E}[\mathbf{X}_k]\right)}{2L}}.$$

Lemma 24. *(Minimum eigenvalue of Gram matrix)*

For $\tau = \max\left\{d, \frac{8a^2}{\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}}}\ln\left(\frac{d^2T}{\delta}\right)\right\}$, the probability of the event $\mathcal{E}_v$ is bounded by,

$$\mathbb{P}\left(\mathcal{E}_v\right) \geq 1-\frac{\delta}{dT}.$$

Proof.

$$\lambda_{\min}\left(\mathbf{V}_\tau\right) = \lambda_{\min}\left(\sum_{s=1}^{\tau}A_s A_s^\top\right)$$

From Theorem 5.1.1 in Tropp (2015), Lemma 23 with $t = \frac{1}{2}$, and Assumption 1 we get,

$$\mathbb{P}\left(\lambda_{\min}\left(\mathbf{V}_\tau\right) \leq \frac{\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}} d}{2}\right) \leq \mathbb{P}\left(\lambda_{\min}\left(\mathbf{V}_\tau\right) \leq \frac{\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}}\tau}{2}\right) \leq de^{-\frac{\underline{\lambda}_{\mathbf{\Sigma}_\mathcal{A}}\tau}{8a^2}} \leq \frac{\delta}{dT}.$$

□

Lemma 25. *($\hat{\mathbf{\Sigma}}^{\mathrm{w}} \succeq \mathbf{\Sigma}_*$)*

Let $\hat{\mathbf{\Sigma}}$ be a symmetric matrix and $\mathbf{\Sigma}_*$ be a PD matrix s.t. $\left\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_*\right\|_{\mathrm{op}} \leq s$. Define $\hat{\mathbf{\Sigma}}^{\mathrm{w}} = \hat{\mathbf{\Sigma}} + s \cdot I$, then

$$\hat{\mathbf{\Sigma}}^{\mathrm{w}} \succeq \mathbf{\Sigma}_* \quad \textit{and} \quad \mathbf{\Sigma}_*^{-1} \succeq \left(\hat{\mathbf{\Sigma}}^{\mathrm{w}}\right)^{-1}.$$

Proof.

$$\lambda_{\min}\left(\hat{\mathbf{\Sigma}}^{\mathrm{w}} - \mathbf{\Sigma}_*\right) \underset{(a)}{\geq} \lambda_{\min}\left(s \cdot \mathbf{I}\right) + \lambda_{\min}\left(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_*\right) \underset{(b)}{\geq} s - \left\|\mathbf{\Sigma}_* - \hat{\mathbf{\Sigma}}\right\|_{\mathrm{op}} \geq 0,$$

where (a) uses Weyl's inequality and (b) uses $\lambda_{\min}\left(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_*\right) = -\lambda_{\max}\left(\mathbf{\Sigma}_* - \hat{\mathbf{\Sigma}}\right) \geq -\left\|\mathbf{\Sigma}_* - \hat{\mathbf{\Sigma}}\right\|_{\mathrm{op}}$. Since $\left(\hat{\mathbf{\Sigma}}^{\mathrm{w}} - \mathbf{\Sigma}_*\right)$ is PSD, $\hat{\mathbf{\Sigma}}^{\mathrm{w}}$ is PD, thus using Lemma 33 the proof follows. □

Lemma 26. *(maximal eigenvalue inequality)*

Let A be a vector and $\mathbf{B}$ a PD matrix, then,

$$\|A\| \leq \sqrt{\lambda_{\max}\left(\mathbf{B}\right)}\,\|A\|_{\mathbf{B}^{-1}}.$$

Proof. Since $\mathbf{B}$ is symmetric we can use an eigenvalue decomposition $\mathbf{B} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$, where $\mathbf{\Lambda}$ is a diagonal matrix containing all the eigenvalues of $\mathbf{B}$ and $\mathbf{Q}, \mathbf{Q}^\top$ are orthonormal matrices. Then,

$$\begin{aligned}
\|A\| &= \sqrt{A^\top A} \\
&= \sqrt{A^\top \mathbf{B}^{-1/2}\mathbf{B}\mathbf{B}^{-1/2}A} \\
&= \sqrt{A^\top \mathbf{B}^{-1/2}\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top\mathbf{B}^{-1/2}A} \\
&\underset{(a)}{=} \sqrt{U^\top \mathbf{\Lambda} U} \\
&\underset{(b)}{\leq} \sqrt{\lambda_{\max}\left(\mathbf{B}\right)}\sqrt{U^\top U} \\
&= \sqrt{\lambda_{\max}\left(\mathbf{B}\right)}\sqrt{A^\top \mathbf{B}^{-1/2}\mathbf{Q}\mathbf{Q}^\top\mathbf{B}^{-1/2}A} \\
&\underset{(c)}{=} \sqrt{\lambda_{\max}\left(\mathbf{B}\right)}\sqrt{A^\top \mathbf{B}^{-1}A} \\
&= \sqrt{\lambda_{\max}\left(\mathbf{B}\right)}\,\|A\|_{\mathbf{B}^{-1}},
\end{aligned}$$

where (a) uses $U \triangleq \mathbf{Q}^\top\mathbf{B}^{-1/2}A$, (b) uses that $U^\top U$ is a sum of non negative elements and (c) uses that $\mathbf{Q}$ is orthonormal. □

Lemma 27. *(Matrix Hoffeding, Theorem 1.3 in Tropp (2012))*

Consider a finite sequence $\{\mathbf{X}_j\}$ of independent, random, self-adjoint matrices with dimension d, and let $\{\mathbf{A}_j\}$ be a sequence of fixed self-adjoint matrices. Assume that each random matrix satisfies,

$$\mathbb{E}\{\mathbf{X}_j\} = 0 \quad \textit{and} \quad \mathbf{X}_j^2 \preceq \mathbf{A}_j^2 \quad \textit{almost surely.}$$

Then for every $\delta \geq 0$,

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_j \mathbf{X}_j\right) \geq \delta\right) \leq d e^{-\frac{\delta^2}{8s^2}} \quad \textit{where} \quad s^2 \triangleq \left\|\sum_j \mathbf{A}_j^2\right\|_{\mathrm{op}}.$$

Lemma 28. *(Concentration bound sub-Gaussian vector, Theorem 1 in Hsu et al. (2012))*

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be a matrix, and let $\mathbf{\Sigma} = \mathbf{A}^\top\mathbf{A}$. Suppose that $X = (X_1, \ldots, X_n)$ is a random vector such that for $\mu = 0$ and some $\sigma \geq 0$,

$$\mathbb{E}\left[\exp\left(U^\top X\right)\right] \leq \exp\left(\frac{\|U\|^2\sigma^2}{2}\right),$$

for all $U \in \mathbb{R}^n$. *For all* $\delta > 0$,

$$\mathbb{P}\left(\|\mathbf{A}X\|^2 > \sigma^2\left(\mathrm{Tr}\left(\mathbf{\Sigma}\right) + 2\sqrt{\mathrm{Tr}\left(\mathbf{\Sigma}^2\right)\delta} + 2\left\|\mathbf{\Sigma}\right\|_{\mathrm{op}}\delta\right)\right) \leq e^{-\delta}.$$

Lemma 29. *(Gaussian tail bounds, section 7.1 in Feller (1968))*

Let z *be a standard normal variable. Then for any* $t > 0$,

$$\mathbb{P}\left(z > t\right) > \frac{1}{\sqrt{2\pi}}\left(\frac{1}{t} - \frac{1}{t^3}\right)e^{\frac{-t^2}{2}}, \tag{84}$$

$$\mathbb{P}\left(z > t\right) \leq \frac{1}{\sqrt{2\pi}}\frac{1}{t}e^{\frac{-t^2}{2}}. \tag{85}$$

Lemma 30. *(Same non-zero eigenvalues for* $\mathbf{AB}$ *and* $\mathbf{BA}$*, Theorem 1.3.22 in Horn and Johnson (2012))*

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ *and* $\mathbf{B} \in \mathbb{R}^{n\times m}$ *with* $m \leq n$. *Then the* n *eigenvalues of* $\mathbf{BA}$ *are the* m *eigenvalues of* $\mathbf{AB}$ *together with* $n - m$ *zeros. If* $m = n$ *and at least one of* $\mathbf{A}$ *or* $\mathbf{B}$ *is nonsingular, then* $\mathbf{AB}$ *and* $\mathbf{BA}$ *are similar.*

Lemma 31. *(Same eigenvalues for product of PSD matrices* $\lambda_i\left(\mathbf{AB}\right) = \lambda_i\left(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2}\right)$*)*

Let $\mathbf{A}, \mathbf{B}$ *PSD matrices. Then,*

$$\lambda_i\left(\mathbf{AB}\right) = \lambda_i\left(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2}\right) = \lambda_i\left(\mathbf{BA}\right) = \lambda_i\left(\mathbf{B}^{1/2}\mathbf{A}\mathbf{B}^{1/2}\right).$$

Proof. $\mathbf{AB} = \mathbf{A}^{1/2}\mathbf{A}^{1/2}\mathbf{B}$. Using Lemma 30 this matrix has the same eigenvalues as $\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2}$. In the same line of proof we get that the eigenvalues of $\mathbf{BA}$ and $\mathbf{B}^{1/2}\mathbf{A}\mathbf{B}^{1/2}$ are the same. Finally, by Lemma 30 $\mathbf{AB}$ and $\mathbf{BA}$ has the same eigenvalues. □

Lemma 32. *(Non negative trace for product of PSD matrices)*

Let $\mathbf{A}$ *and* $\mathbf{B}$ *be PSD matrices. Then,* $\mathrm{Tr}\left(\mathbf{AB}\right) \geq 0$.

Proof. From Lemma 31,

$$\mathrm{Tr}\left(\mathbf{AB}\right) = \mathrm{Tr}\left(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2}\right) = \sum_{j=1}^{d}\lambda_j\left(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2}\right) \geq 0,$$

where the inequality uses that the matrix $\mathbf{B}$ is PSD, hence $\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2}$ is PSD. □

Lemma 33. *(Corollary 7.7.4.(a) in Horn and Johnson (2012))*

Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ *PD matrices.* $\mathbf{A} \succeq \mathbf{B}$ *iff* $\mathbf{B}^{-1} \succeq \mathbf{A}^{-1}$.

Lemma 34. *(Corollary 7.7.4.(b) in Horn and Johnson (2012))*

Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ *be symmetric matrices. If* $\mathbf{A} \succ 0$, $\mathbf{B} \succeq 0$ *and* $\mathbf{A} \succeq \mathbf{B}$, *then* $\mathbf{A}^{1/2} \succeq \mathbf{B}^{1/2}$.

Lemma 35. *(Exercise VI.7.2 in Bhatia (1997) (page 182))*

Let $\mathbf{A}$ *and* $\mathbf{B}$ *be Hermitian matrices. If* $\lambda_{\min}\left(\mathbf{A}\right) + \lambda_{\min}\left(\mathbf{B}\right) > 0$, *then*

$$\prod_{j=1}^{d}\left(\lambda_j\left(\mathbf{A}\right) + \lambda_j\left(\mathbf{B}\right)\right) \leq \mathrm{Det}\left(\mathbf{A} + \mathbf{B}\right).$$

Lemma 36. *(*$\mathbf{B}$ *bound)*

$$\left\|\mathbf{B}\right\|_{\mathrm{op}} \leq \frac{\sigma^2}{\lambda^2_{\mathbf{\Sigma}_*}}\left\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_*\right\|_{\mathrm{op}}.$$

Proof.

$$\begin{aligned}
\|\mathbf{B}\|_{\text{op}} &= \left\|\sigma^2\left(\boldsymbol{\Sigma}_*^{-1} - \hat{\boldsymbol{\Sigma}}^{-1}\right)\right\|_{\text{op}} \\
&= \left\|\sigma^2\left(\hat{\boldsymbol{\Sigma}}^{-1}\left(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_*\right)\boldsymbol{\Sigma}_*^{-1}\right)\right\|_{\text{op}} \\
&\underset{(a)}{\leq} \sigma^2 \left\|\hat{\boldsymbol{\Sigma}}^{-1}\right\|_{\text{op}} \left\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_*\right\|_{\text{op}} \left\|\boldsymbol{\Sigma}_*^{-1}\right\|_{\text{op}} \\
&\underset{(b)}{\leq} \frac{\sigma^2}{\left(\lambda_{\min}\left(\boldsymbol{\Sigma}_*\right)\right)^2} \left\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_*\right\|_{\text{op}} \\
&\underset{(c)}{\leq} \frac{\sigma^2}{\lambda_{\boldsymbol{\Sigma}_*}^2} \left\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_*\right\|_{\text{op}},
\end{aligned}$$

where (a) uses sub-multiplicative norm properties, (b) uses that $\hat{\boldsymbol{\Sigma}} \succeq \boldsymbol{\Sigma}_*$ and Lemma 26, and (c) uses Assumption 2. □

Lemma 37. *(Exponent bound, Lemma 20 in Bastani et al. (2021))*

For any $x \in [0,1]$, $\quad e^x \leq 1 + 2x$.

Lemma 38. *(Log bound, Lemma A.2 in Shalev-Shwartz and Ben-David (2014))*

Let $a_1 \geq 1$ *and* $b_1 > 0$. *Then,* $\quad x \geq 4a_1 \ln(2a_1) + 2b_1 \Rightarrow x \geq a_1 \ln(x) + b_1$.