Continuous Mean-Covariance Bandits

Yihan Du  
IIIS, Tsinghua University  
Beijing, China  
duyh18@mails.tsinghua.edu.cn

Siwei Wang  
CST, Tsinghua University  
Beijing, China  
wangsw2020@mail.tsinghua.edu.cn

Zhixuan Fang  
IIIS, Tsinghua University, Beijing, China  
Shanghai Qi Zhi Institute, Shanghai, China  
zfang@mail.tsinghua.edu.cn

Longbo Huang∗  
IIIS, Tsinghua University  
Beijing, China  
longbohuang@mail.tsinghua.edu.cn

Abstract

Existing risk-aware multi-armed bandit models typically focus on risk measures of individual options such as variance. As a result, they cannot be directly applied to important real-world online decision making problems with correlated options. In this paper, we propose a novel Continuous Mean-Covariance Bandit (CMCB) model to explicitly take into account option correlation. Specifically, in CMCB, there is a learner who sequentially chooses weight vectors on given options and observes random feedback according to the decisions. The agent’s objective is to achieve the best trade-off between reward and risk, measured with option covariance. To capture different reward observation scenarios in practice, we consider three feedback settings, i.e., full-information, semi-bandit and full-bandit feedback. We propose novel algorithms with rigorous regret analysis, and provide nearly matching lower bounds to validate their optimalities (in terms of the number of timesteps $T$). The experimental results also demonstrate the superiority of our algorithms. To the best of our knowledge, this is the first work that considers option correlation in risk-aware bandits and explicitly quantifies how arbitrary covariance structures impact the learning performance. The novel analytical techniques we developed for exploiting the estimated covariance to build concentration and bounding the risk of selected actions based on sampling strategy properties can likely find applications in other bandit analysis and be of independent interests.

1 Introduction

The stochastic Multi-Armed Bandit (MAB) Auer et al. [2002a], Thompson [1933], Agrawal and Goyal [2012] problem is a classic online learning model, which characterizes the exploration-exploitation trade-off in decision making. Recently, due to the increasing requirements of risk guarantees in practical applications, the Mean-Variance Bandits (MVB) Sani et al. [2012], Vakili and Zhao [2016], Zhu and Tan [2020] which aim at balancing the rewards and performance variances have received extensive attention. While MVB provides a successful risk-aware model, it only considers discrete decision space and focuses on the variances of individual arms (assuming independence among arms).

However, in many real-world scenarios, a decision often involves multiple options with certain correlation structure, which can heavily influence risk management and cannot be ignored. For instance, in finance, investors can select portfolios on multiple correlated assets, and the investment risk is closely related to the correlation among the chosen assets. The well-known “risk diversification”
strategy Butler and Domian [1991] embodies the importance of correlation to investment decisions. In clinical trials, a treatment often consists of different drugs with certain ratios, and the correlation among drugs plays an important role in the treatment risk. Failing to handle the correlation among multiple options, existing MVB results cannot be directly applied to these important real-world tasks.

Witnessing the above limitation of existing risk-aware results, in this paper, we propose a novel Continuous Mean-Covariance Bandit (CMCB) model, which considers a set of options (base arms) with continuous decision space and measures the risk of decisions with the option correlation. Specifically, in this model, a learner is given $d$ base arms, which are associated with an unknown joint reward distribution with a mean vector and covariance. At each timestep, the environment generates an underlying random reward for each base arm according to the joint distribution. Then, the learner selects a weight vector of base arms and observes the rewards. The goal of the learner is to minimize the expected cumulative regret, i.e., the total difference of the reward-risk (mean-covariance) utilities between the chosen actions and the optimal action, where the optimal action is defined as the weight vector that achieves the best trade-off between the expected reward and covariance-based risk. To capture important observation scenarios in practice, we consider three feedback settings in this model, i.e., full-information (CMCB-FI), semi-bandit (CMCB-SB) and full-bandit (CMCB-FB) feedback, which vary from seeing rewards of all options to receiving rewards of the selected options to only observing a weighted sum of rewards.

The CMCB framework finds a wide range of real-world applications, including finance Markowitz et al. [1952], company operation McInnerney and Roberts [2004] and online advertising Schwartz et al. [2017]. For example, in stock markets, investors choose portfolios based on the observed prices of all stocks (full-information feedback), with the goal of earning high returns and meanwhile minimizing risk. In company operation, managers allocate investment budgets to several correlated business and only observe the returns of the invested business (semi-bandit feedback), with the objective of achieving high returns and low risk. In clinical trials, clinicians select a treatment comprised of different drugs and only observe an overall therapeutic effect (full-bandit feedback), where good therapeutic effects and high stability are both desirable.

For both CMCB-FI and CMCB-SB, we propose novel algorithms and establish nearly matching lower bounds for the problems, and contribute novel techniques in analyzing the risk of chosen actions and exploiting the covariance information. For CMCB-FB, we develop a novel algorithm which adopts a carefully designed action set to estimate the expected rewards and covariance, with non-trivial regret guarantees. Our theoretical results offer an explicit quantification of the influences of arbitrary covariance structures on learning performance, and our empirical evaluations also demonstrate the superior performance of our algorithms.

Our work differs from previous works on bandits with covariance Warmuth and Kuzmin [2006, 2012], Degene and Perchet [2016], Perrault et al. [2020] in the following aspects. (i) We consider the reward-risk objective under continuous decision space and stochastic environment, while existing works study either combinatorial bandits, where the decision space is discrete and risk is not considered in the objective, or adversarial online optimization. (ii) We do not assume a prior knowledge or direct feedback on the covariance matrix as in Warmuth and Kuzmin [2006, 2012], Degene and Perchet [2016]. (iii) Our results for full-information and full-bandit feedback implicitly characterize the impacts of arbitrary covariance structures, whereas prior results, e.g., Degene and Perchet [2016], Perrault et al. [2020], only focus on independent or positively-correlated cases. These differences pose new challenges in algorithm design and analysis, and demand new analytical techniques.

We summarize the main contributions as follows.

- We propose a novel risk-aware bandit model called continuous mean-covariance bandit (CMCB), which considers correlated options with continuous decision space, and characterizes the trade-off between reward and covariance-based risk. Motivated by practical reward observation scenarios, three feedback settings are considered under CMCB, i.e., full-information (CMCB-FI), semi-bandit (CMCB-SB) and full-bandit (CMCB-FB).

- We design an algorithm $\mathcal{MC}$-Empirical for CMCB-FI with a $O(\sqrt{T})$ regret, and develop a novel analytical technique to build a relationship on risk between chosen actions and the optimal one using properties of the sampling strategy. We also derive a nearly matching lower bound, by analyzing the gap between hindsight knowledge and available empirical information under a Bayesian environment.
• For CMCB-SB, we develop MC-UCB, an algorithm that exploits the estimated covariance information to construct confidence intervals and achieves a $\tilde{O}(\sqrt{T})$ regret. A regret lower bound is also established, by investigating the necessary regret paid to differentiate two well-chosen distinct instances.

• We propose a novel algorithm MC-ETE for CMCB-FB, which employs a well-designed action set to carefully estimate the reward means and covariance, and achieves a $\tilde{O}(T^{2/3})$ regret guarantee under the severely limited feedback.

To our best knowledge, our work is the first to explicitly characterize the influences of arbitrary covariance structures on learning performance in risk-aware bandits. Our results shed light into risk management in online decision making with correlated options. Due to space limitation, we defer all detailed proofs to the supplementary material.

2 Related Work

(Risk-aware Bandits) Sani et al. Sani et al. [2012] initiate the classic mean-variance paradigm Markowitz et al. [1952], Gasser et al. [2017] in bandits, and formulate the mean-variance bandit problem, where the learner plays a single arm each time and the risk is measured by the variances of individual arms. Vakili & Zhao Vakili and Zhao [2015, 2016] further study this problem under a different metric and complete the regret analysis. Zhu & Tan Zhu and Tan [2020] provide a Thompson Sampling-based algorithm for mean-variance bandits. In addition to variance, several works consider other risk criteria. The VaR measure is studied in David and Shimkin [2016], and CVaR is also investigated to quantify the risk in Galichet et al. [2013], Kagrecha et al. [2019]. Cassel et al. Cassel et al. [2018] propose a general risk measure named empirical distributions performance measure (EDPM) and present an algorithmic framework for EDPM. All existing studies on risk-aware bandits only consider discrete decision space and assume independence among arms, and thus they cannot be applied to our CMCB problem.

(Bandits with Covariance) In the stochastic MAB setting, while there have been several works Degenne and Perchet [2016], Perrault et al. [2020] on covariance, they focus on the combinatorial bandit problem without considering risk. Degenne & Perchet Degenne and Perchet [2016] study the combinatorial semi-bandits with correlation, which assume a known upper bound on the covariance, and design an algorithm with this prior knowledge of covariance. Perrault et al. Perrault et al. [2020] further investigate this problem without the assumption on covariance under the sub-exponential distribution framework, and propose an algorithm with a tight asymptotic regret analysis. In the adversarial setting, Warmuth & Kuzmin Warmuth and Kuzmin [2006, 2012] consider an online variance minimization problem, where at each timestep the learner chooses a weight vector and receives a covariance matrix, and propose the exponentiated gradient based algorithms. Our work differs from the above works in the following aspects: compared to Degenne and Perchet [2016], Perrault et al. [2020], we consider a continuous decision space instead of combinatorial space, study the reward-risk objective instead of only maximizing the expected reward, and investigate two more feedback settings other than the semi-bandit feedback. Compared to Warmuth and Kuzmin [2006, 2012], we consider the stochastic environment and in our case, the covariance cannot be directly observed and needs to be estimated.

3 Continuous Mean-Covariance Bandits (CMCB)

Here we present the formulation for the Continuous Mean-Covariance Bandits (CMCB) problem. Specifically, a learner is given $d$ base arms labeled $1, \ldots, d$ and a decision (action) space $D \subseteq \Delta_d$, where $\Delta_d = \{w \in \mathbb{R}^d : 0 \leq w_i \leq 1, \forall i \in [d], \sum_i w_i = 1\}$ denotes the probability simplex in $\mathbb{R}^d$. The base arms are associated with an unknown $d$-dimensional joint reward distribution with mean vector $\theta^*$ and positive semi-definite covariance matrix $\Sigma^*$, where $\Sigma^*_{ii} \leq 1$ for any $i \in [d]$ without loss of generality. For any action $w \in D$, which can be regarded as a weight vector placed on the base arms, the instantaneous reward-risk utility is given by the following mean-covariance function

$$f(w) = w^\top \theta^* - \rho w^\top \Sigma^* w,$$ (1)
where \( w^\top \theta^* \) denotes the expected reward, \( w^\top \Sigma^* w \) represents the risk, i.e., reward variance, and \( \rho > 0 \) is a risk-aversion parameter that controls the weight placed on the risk. We define the optimal action as \( w^* = \arg\max_{w \in D} f(w) \). Compared to linear bandits Abbasi-Yadkori et al. [2011], Kazerouni et al. [2017], the additional quadratic term in \( f(w) \) raises significant challenges in estimating the covariance, bounding the risk of chosen actions and deriving covariance-dependent regret bounds.

At each timestep \( t \), the environment generates an underlying (unknown to the learner) random reward vector \( \theta_t = \theta^* + \eta_t \) according to the joint distribution, where \( \eta_t \) is a zero-mean noise vector and it is independent among different timestep \( t \). Note that here we consider an additive vector noise to the parameter \( \theta^* \), instead of the simpler scalar noise added in the observation (i.e., \( y_t = w^\top \theta^* + \eta_t \)) as in linear bandits Abbasi-Yadkori et al. [2011], Kazerouni et al. [2017]. Our noise setting better models the real-world scenarios where distinct actions incur different risk, and enables us to explicitly quantify the correlation effects. Following the standard assumption in the bandit literature Locatelli et al. [2016], Degene and Perchet [2016], Zhu and Tan [2020], we assume the noise is sub-Gaussian, i.e., \( \forall u \in \mathbb{R}^d, \mathbb{E}[\exp(u^\top \eta_t)] \leq \exp(\frac{1}{2} u^\top \Sigma^* u) \), where \( \Sigma^* \) is unknown. The learner selects an action \( w_t \in D \) and observes the feedback according to a certain structure (specified later). For any time horizon \( T > 0 \), define the expected cumulative regret as

\[
\mathbb{E}[R(T)] = \sum_{t=1}^{T} \mathbb{E}[f(w^*) - f(w_t)].
\]

The objective of the learner is to minimize \( \mathbb{E}[R(T)] \). Note that our mean-covariance function Eq. (1) extends the popular mean-variance measure Sani et al. [2012], Vakili and Zhao [2016], Zhu and Tan [2020] to the continuous decision space.

In the following, we consider three feedback settings motivated by reward observation scenarios in practice, including (i) full-information (CMCF-FI), observing random rewards of all base arms after a pull, (ii) semi-bandit (CMCB-SB), only observing random rewards of the selected base arms, and (iii) full-bandit (CMCB-FB), only seeing a weighted sum of the random rewards from base arms. We will present the formal definitions of these three feedback settings in the following sections.

**Notations.** For action \( w \in D \), let \( I_w \) be a diagonal matrix such that \( I_{w,ii} = 1 \{ w_i > 0 \} \). For a matrix \( A \), let \( A_w = I_w A I_w \) and \( \Lambda_A \) be a diagonal matrix with the same diagonal as \( A \).

4 CMCB with Full-Information Feedback (CMCB-FI)

We start with CMCB with full-information feedback (CMCB-FI). In this setting, at each timestep \( t \), the learner selects \( w_t \in \Delta_d \) and observes the random reward \( \theta_{t,i} \) for all \( i \in [d] \). CMCB-FI provides an online learning model for the celebrated Markowitz Markowitz et al. [1952], Gasser et al. [2017] problem in finance, where investors select portfolios and can observe the prices of all stocks at the end of the trading days.

Below, we propose an efficient Mean-Covariance Empirical algorithm (MC-Empirical) for CMCB-FI, and provide a novel regret analysis that fully characterizes how an arbitrary covariance structure affects the regret performance. We also present a nearly matching lower bound for CMCB-FI to demonstrate the optimality of MC-Empirical.

4.1 Algorithm for CMCB-FI

Algorithm 1 shows the detailed steps of MC-Empirical. Specifically, at each timestep \( t \), we use the empirical mean \( \hat{\theta}_t \) and covariance \( \hat{\Sigma}_t \) to estimate \( \theta^* \) and \( \Sigma^* \), respectively. Then, we form \( f_t(w) = w^\top \hat{\theta}_t - \rho w^\top \hat{\Sigma}_t w \), an empirical mean-covariance function of \( w \in \Delta_d \), and always choose the action with the maximum empirical objective value.

Although MC-Empirical appears to be intuitive, its analysis is highly non-trivial due to covariance-based risk in the objective. In this case, a naive universal bound cannot characterize the impact of covariance, and prior gap-dependent analysis (e.g., Degene and Perchet [2016], Perrault et al. [2020]) cannot be applied to solve our continuous space analysis with gap approximating to zero. Instead, we develop two novel techniques to handle the covariance, including using the actual covariance to analyze the confidence region of the expected rewards, and exploiting the empirical information of
the sampling strategy to bound the risk gap between selected actions and the optimal one. Different from prior works Deguenne and Perchet [2016], Perrault et al. [2020], which assume a prior knowledge on covariance or only focus on the independent and positively-related cases, our analysis does not require extra knowledge of covariance and explicitly quantifies the effects of arbitrary covariance structures. The regret performance of MC-Empirical is summarized in Theorem 1.

**Theorem 1 (Upper Bound for CMCB-FI).** Consider the continuous mean-covariance bandits with full-information feedback (CMCB-FI). For any $T \geq 1 + \frac{\Sigma_{\max}^2}{w^\top \Sigma w}$, algorithm MC-Empirical (Algorithm 1) achieves an expected cumulative regret bounded by

$$O\left(\left(\min\left\{\sqrt{w^\top \Sigma w} + \rho^{-\frac{1}{2}}\sqrt{\theta_{\max}^* - \theta_{\min}^*}, \sqrt{\Sigma_{\max}^2}\right\} + \rho\right) \ln T \sqrt{d T}\right),$$

where $\theta_{\max}^* = \max_{i \in [d]} \theta_i^*$, $\theta_{\min}^* = \min_{i \in [d]} \theta_i^*$ and $\Sigma_{\max}^* = \max_{i \in [d]} \Sigma_{i,i}^*$.

**Proof sketch.** Let $D_t$ be the diagonal matrix which takes value $t$ at each diagonal entry. We first build confidence intervals for the expected rewards of actions and the covariance as $|w^\top \theta^* - w^\top \hat{\theta}_{t-1}| \leq p_t(w) \triangleq c_1 \sqrt{\beta_t \sqrt{|w^\top D_t^{-1}_t (\lambda \Sigma D_t^{-1}_t + \sum_{s=1}^{t-1} \Sigma_s) D_t^{-1}_t w|}}$ and $|\Sigma_{ij} - \hat{\Sigma}_{ij,t-1}| \leq q_t \triangleq c_2 \frac{\ln t}{\sqrt{t}}$. Here $\beta_t \triangleq \ln t + d \ln (1 + \lambda^{-1})$, $\lambda \triangleq \frac{w^\top \Sigma w}{\Sigma_{\max}^2}$, and $c_1$ and $c_2$ are positive constants. Then, we obtain the confidence interval of $f(w)$ as $|f_{t-1}(w) - f(w)| \leq \rho_t(w) \triangleq p_t(w) + \rho w^\top Q_t w$, where $Q_t$ is a matrix with all entries equal to $q_t$. Since algorithm MC-Empirical always plays the empirical best action, we have $f(w^*) - f(w_t) \leq f_{t-1}(w^*) + r_t(w^*) - f(w_t) \leq f_{t-1}(w_t) + r_t(w^*) - f(w_t) \leq r_t(w^*) + r_t(w_t)$. Plugging the definitions of $f(w)$ and $r_t(w)$, we have

$$-\Delta_{\theta^*} + \rho \left(w_t^\top \Sigma w_t - w^\top \Sigma w^*ight) \leq f(w^*) - f(w_t) \leq (a) \frac{\sqrt{\beta_t \sqrt{w^\top \Sigma w^* + w^\top \Sigma w_t}} + \rho \ln t}{\sqrt{t-1}},$$

where $\Delta_{\theta^*} = \theta_{\max}^* - \theta_{\min}^*$ and $c_3$ is a positive constant. Since our goal is to bound the regret $f(w^*) - f(w_t)$ and in inequality (a) only the $\sqrt{w^\top \Sigma w_t}$ term is a variable, the challenge falls on bounding $w_t^\top \Sigma w_t$. Note that the left-hand-side of Eq. (3) is linear with respect to $w_t^\top \Sigma w_t$ and the right-hand-side only contains $\sqrt{w_t^\top \Sigma w_t}$, Then, using the property of sampling strategy on $w_t$, i.e., Eq. (3), again, after some algebraic analysis, we obtain $w_t^\top \Sigma w_t \leq c_4 \left(w^\top \Sigma w^* + \frac{1}{\rho} \Delta_{\theta^*} + \frac{1}{\rho^2} \sqrt{t} \sqrt{w^\top \Sigma w^* + \frac{\ln t}{\sqrt{t}} + \frac{\rho}{t}}\right)$ for some constant $c_4$. Plugging it into inequality (a) and doing a summation over $t$, we obtain the theorem. \(\square\)

**Remark 1.** Theorem 1 fully characterizes how an arbitrary covariance structure impacts the regret bound. To see this, note that in Eq. (2), under the $\min$ operation, the first $\sqrt{w^\top \Sigma w}$-related term dominates under reasonable $\rho$, and shrinks from positive to negative correlation, which implies that the more the base arms are negatively (positively) correlate, the lower (higher) regret the learner suffers. The intuition behind is that the negative (positive) correlation diversifies (intensifies) the risk of estimation error and narrows (enlarges) the confidence region for the expected reward of an action, which leads to a reduction (an increase) of regret.

Also note that when $\rho = 0$, the CMCB-FI problem reduces to a $d$-armed bandit problem with full-information feedback, and Eq. (2) becomes $O(\sqrt{d \Sigma_{\max}^2 T})$. For this degenerate case, the

---

**Algorithm 1 MC-Empirical**

1. **Input:** Risk-aversion parameter $\rho > 0$,
2. Initialization: Pull action $w_t = (\frac{1}{\sqrt{3}}, \ldots, \frac{1}{\sqrt{3}})$, and observe $\theta_t = (\theta_1, \ldots, \theta_d)^\top$, $\hat{\theta}_t,i \leftarrow \theta_t,i, \forall i \in [d]$. $\Sigma_{t,i,j} = (\theta_{t,i} - \hat{\theta}_t,i)(\theta_{t,j} - \hat{\theta}_t,j), \forall i,j \in [d]$.
3. for $t = 2, 3, \ldots$ do
   4. $w_t = \arg\max_{w \in \Delta_d} (w^\top \hat{\theta}_{t-1} - \rho w^\top \Sigma w)\Sigma_{t-1} w$.
   5. Pull $w_t$, observe $\theta_t = (\theta_1, \ldots, \theta_d)^\top$.
   6. $\hat{\theta}_{t,i} \leftarrow \frac{1}{t} \sum_{s=1}^{t} \theta_{s,i}, \forall i \in [d]$. $\Sigma_{t,i,j} = \frac{1}{t} \sum_{s=1}^{t} (\theta_{s,i} - \hat{\theta}_{s,i})(\theta_{s,j} - \hat{\theta}_{s,j}), \forall i,j \in [d]$.
8. end for
Algorithm 2 MC-UCB

1: Input: $\rho > 0$, $c \in (0, \frac{1}{2}]$ and regularization parameter $\lambda > 0$.
2: Initialize: $\forall i \in [d]$, pull $e_i$ that has 1 at the $i$-th entry and 0 elsewhere. $\forall i,j \in [d], i \neq j$, pull $e_{ij}$ that has $\frac{1}{2}$ at the $i$-th and the $j$-th entries, and 0 elsewhere. Update $N_{ij}(d^2)$, $\forall i,j \in [d], \theta_{ij}$ and $\Sigma_{ij}$.
3: for $t = d^2 + 1, \ldots$
4: $\sum_{t,i,j} \leftarrow \sum_{t-1,i,j} - g_{ij}(t)$
5: $\hat{\Sigma}_{t,i,j} \leftarrow \sum_{t-1,i,j} + g_{ij}(t)$
6: $w_t \leftarrow \arg\max_{w \in \Delta^d} (w^T \hat{\Theta}_{t-1} + E_t(w) - \rho w^T \Sigma w)$
7: Pull $w_t$ and observe all $\theta_{t,i}$ s.t. $w_t,i > 0$
8: $J_{t,i} \leftarrow \mathbb{I}\{w_{t,i}, w_{t,j} > 0\}, \forall i,j \in [d]$
9: $N_{ij}(t) \leftarrow N_{ij}(t-1) + J_{t,i,j}, \forall i,j \in [d]$
10: $\hat{\theta}_{t,i} \leftarrow \frac{\sum_{s=1}^{t} J_{s,i} \theta_{s,i}}{N_{ij}(t)}, \forall i \in [d]$
11: $\hat{\Sigma}_{t,i,j} \leftarrow \frac{\sum_{s=1}^{t} J_{s,i,j} (\theta_{s,i} - \hat{\theta}_{s,i}) (\theta_{s,j} - \hat{\theta}_{s,j})}{N_{ij}(t)}, \forall i,j \in [d]$
12: end for

optimal gap-dependent regret is $O\left(\sum_{\Delta = \Delta^*} \frac{\sqrt{T}}{T}\right)$ for constant gap $\Delta > 0$. By setting $\Delta = \sqrt{\sum_{\Delta = \Delta^*} T}$ at this gap-dependent result, one obtains the optimal gap-independent regret $O\left(\sqrt{\sum_{\Delta = \Delta^*} T}\right)$. Hence, when $\rho = 0$, Eq. (2) still offers a tight gap-independent regret bound with respect to $T$.

4.2 Lower Bound for CMCB-FI

Now we provide a regret lower bound for CMCB-FI to corroborate the optimality of MC-Empirical.

Since CMCB-FI considers full-information feedback and continuous decision space where the reward gap $\Delta$ (between the optimal action and the nearest optimal action) approaches to zero, existing lower bound analysis for linear Dani et al. [2008a,b] or discrete Lai and Robbins [1985], Degene and Perchet [2016], Perrault et al. [2020] bandit problems cannot be applied to this problem.

To tackle this challenge, we contribute a new analytical procedure to establish the lower bound for continuous and full-information bandit problems from the Bayesian perspective. The main idea is to construct an instance distribution, where $\Theta^*$ is drawn from a well-chosen prior Gaussian distribution. After $t$ pulls the posterior of $\Theta^*$ is still Gaussian with a mean vector $u_t$ related to sample outcomes. Since the hindsight strategy simply selects the action which maximizes the mean-covariance function with respect to $\Theta^*$ while a feasible strategy can only utilize the sample information ($u_t$), we show that any algorithm must suffer $\Omega\left(\sqrt{T}\right)$ regret due to the gap between random $\Theta^*$ and its mean $u_t$.

Theorem 2 (Lower Bound for CMCB-FI). There exists an instance distribution of the continuous mean-covariance bandits with full-information feedback problem (CMCB-FI), for which any algorithm has an expected cumulative regret bounded by $\Omega\left(\sqrt{T}\right)$.

Remark 2. This parameter-free lower bound demonstrates that the regret upper bound of MC-Empirical (Theorem 1) is near-optimal with respect to $T$. Unlike discrete bandit problems Lai and Robbins [1985], Degene and Perchet [2016], Perrault et al. [2020] where the optimal regret is usually $\frac{\log T}{\Delta}$ for constant gap $\Delta > 0$, CMCB-FI has a continuous decision space with gap $\Delta \to 0$ and a polylogarithmic regret is not achievable in general.

5 CMCB with Semi-Bandit Feedback (CMCB-SB)

In many practical tasks, the learner may not be able to simultaneously select (place positive weights on) all options and observe full information. Instead, the weight of each option is usually lower bounded and cannot be arbitrarily small. As a result, the learner only selects a subset of options and obtains their feedback, e.g., company investments Diadin [2019] on multiple business.

Motivated by such tasks, in this section we consider the CMCB problem with semi-bandit feedback (CMCB-SB), where the decision space is a restricted probability simplex $\Delta^c = \{w \in \mathbb{R}^d : w_i = 0 or c \leq w_i \leq 1, \forall i \in [d] and \sum_i w_i = 1\}$ for some constant $0 < c \leq \frac{1}{2}$. In this scenario, at

\footnote{When $c > \frac{1}{2}$, the learner can only place all weight on one option, and the problem trivially reduces to the mean-variance bandit setting Sani et al. [2012], Zhu and Tan [2020]. In this case, our Theorem 3 still provides a tight gap-independent bound.}
timestep $t$, the learner selects $w_t \in \Delta_d^*$ and only observes the rewards $\{\theta_{t,i} : w_i \geq \epsilon\}$ from the base arms that are placed positive weights on. Below, we propose the Mean-Covariance Upper Confidence Bound algorithm (MC-UCB) for CMCB-SB, and provide a regret lower bound, which shows that MC-UCB achieves the optimal performance with respect to $T$.

5.1 Algorithm for CMCB-SB

Algorithm MC-UCB for CMCB-SB is described in Algorithm 2. The main idea is to use the optimistic covariance to construct a confidence region for the expected reward of an action and calculate an upper confidence bound of the mean-covariance function, and then select the action with the maximum optimistic mean-covariance value.

In Algorithm 2, $N_{ij}(t)$ denotes the number of times $w_{s,i}, w_{s,j} > 0$ occurs among timestep $s \in [t]$. $J_{t,ij}$ is an indicator variable that takes value 1 if $w_{t,i}, w_{t,j} > 0$ and 0 otherwise. $D_t$ is a diagonal matrix such that $D_{t,ii} = N_{ii}(t)$. In Line 2, we update the number of observations by $N_{ii}(d^2) \leftarrow 2d-1$ for all $i \in [d]$ and $N_{ij}(d^2) \leftarrow 2$ for all $i, j \in [d], i \neq j$ (due to the initialized $d^2$ pulls), and calculate the empirical mean $\theta_{d}^*$ and empirical covariance $\Sigma_{d}^*$ using the equations in Lines 10,11.

For any $t > 1$ and $i,j \in [d]$, we define the confidence radius of covariance $\Sigma_{ij}$ as $g_{ij}(t) \triangleq 16\left(\frac{3\ln t}{N_{ij}(t-1)} + \sqrt{\frac{48\ln^2 t}{N_{ij}(t-1)n(t-1)}} + \sqrt{\frac{36\ln^2 t}{N_{ij}(t-1)n(t-1)}}\right)$, and the confidence region for the expected reward $w^\top \theta^*$ of action $w$ as

$$E_t(w) \triangleq 2\beta(\delta_t) \left( w^\top D_t^{-1} \left( \lambda A_{\Sigma_t} D_t^{-1} + \sum_{s=1}^{t-1} \bar{\Sigma}_{s,w} \right) D_t^{-1} w \right),$$

where $\lambda > 0$ is the regularization parameter, $\beta(\delta_t) = \ln\left(\frac{1}{\delta_t}\right) + d \ln t + \frac{d}{2} \ln(1 + \frac{1}{\delta_t})$ is the confidence term and $\delta_t = \frac{1}{t \ln^2 t}$ is the confidence parameter. At each timestep $t$, algorithm MC-UCB calculates the upper confidence bound of $f(w)$ using $g_{ij}(t)$ and $E_t(w)$, and selects the action $w_t$ that maximizes this upper confidence bound. Then, the learner observes rewards $\theta_{t,i}$ with $w_{t,i} > 0$ and update the statistical information according to the feedback.

In regret analysis, unlike Degenne and Perchet [2016] which uses a universal upper bound to analyze confidence intervals, we incorporate the estimated covariance into the confidence region for the expected reward of an action, which enables us to derive tighter regret bound and explicitly quantify the impact of the covariance structure on algorithm performance. We also contribute a new technique for handling the challenge raised by having different numbers of observations among base arms, in order to obtain an optimal $O(\sqrt{T})$ regret (here prior gap-dependent analysis Degenne and Perchet [2016], Perrault et al. [2020] still cannot be applied to solve this continuous problem). Theorem 3 gives the regret upper bound of algorithm MC-UCB.

**Theorem 3** (Upper Bound for CMCB-SB). Consider the continuous mean-covariance bandits with semi-bandit feedback problem (CMCB-SB). Then, for any $T \geq 0$, algorithm MC-UCB (Algorithm 2) with regularization parameter $\lambda > 0$ has an expected cumulative regret bounded by

$$O\left( L(\lambda)(\|\Sigma^*\|_+ + d^2)d \ln^2 T \cdot T + \rho d \ln T \sqrt{T} \right),$$

where $L(\lambda) = (\lambda + 1)(\ln(1 + 1/\lambda^2) + 1)$ and $\|\Sigma^*\|_+ = \sum_{i,j \in [d]} (\Sigma^*_{ij} \vee 0)$ for any $i,j \in [d]$.

**Remark 3.** Theorem 3 captures the effects of covariance structures in CMCB-SB, i.e., positive correlation renders a larger $\|\Sigma^*\|_+$ factor than the negative correlation or independent case, since the covariance influences the rate of estimate concentration for the expected rewards of actions. The regret bound for CMCB-SB has a heavier dependence on $d$ than that for CMCB-FI. This matches the fact that semi-bandit feedback only reveals rewards of the queried dimensions, and provides less information than full-information feedback in terms of observable dimensions.

5.2 Lower Bound for CMCB-SB

In this subsection, we establish a lower bound for CMCB-SB, and show that algorithm MC-UCB achieves the optimal regret with respect to $T$ up to logarithmic factors.
We propose the Mean-Covariance Exploration-Then-Exploitation algorithm (MC-ETE) for CMCB-FB in Algorithm 3. Specifically, we first choose a design action set $\pi = \{v_1, \ldots, v_d\}$ which contains $d = d(d+1)/2$ actions and satisfies that $B_\pi^+ = (v_1^T, \ldots, v_d^T)$ and $C_\pi = (v_1^2, \ldots, v_d^2, 2v_1 v_{1,1}, \ldots, 2v_1 v_{1,d}, \ldots, v_{d,1}^2, \ldots, v_{d,d}^2)$ are of full column rank. We also denote their Moore-Penrose inverses by $B_\pi^{-1}$ and $C_\pi^{-1}$, and it holds that $B_\pi^{-1} B_\pi = I_{d \times d}$ and $C_\pi^{-1} C_\pi = I_{d \times d}$. There exist more than one feasible $\pi$, and for simplicity and good performance we choose $v_1, \ldots, v_d$ as standard basis vectors in $\mathbb{R}^d$ and $\{v_{d+1}, \ldots, v_{d^2}\}$ as the set of all $\binom{d^2}{d}$ vectors where each vector has two entries equal to $\frac{1}{2}$ and others equal to $0$.

In an exploration round (Lines 8-21), we pull the designed actions in $\pi$ and maintain their empirical rewards and variances. Through linear transformation by $B_\pi^+$ and $C_\pi^+$, we obtain the estimators of the expected rewards and covariance of base arms (Lines 16-17). When the estimation confidence is high enough, we exploit the attained information to select the empirical best action (Lines 5). Theorem 5 presents the regret guarantee of MC-ETE.

Theorem 5 (Upper Bound for CMCB-FB). Consider the continuous mean-covariance bandits with full-bandit feedback (CMCB-FB). Then, for any $T > 0$, algorithm MC-ETE (Algorithm 3)
achieves an expected cumulative regret bounded by

$$O \left( Z(\rho, \pi) \sqrt{d(\ln T + d^2)} \cdot T^{\frac{d}{2}} + d \Delta_{\max} \cdot T^{\frac{d}{2}} \right),$$

where $Z(\rho, \pi) = \max_{w \in \Delta_d} (\sqrt{\mathbf{w}^\top B_+^* \Sigma_\pi^* \mathbf{B}_+^\top \mathbf{w} + \rho \|C_\pi^+\|}), \Sigma_\pi^* = \text{diag}(v_1^\top \Sigma^* v_1, \ldots, v_d^\top \Sigma^* v_d), \|C_\pi^+\| = \max_{i \in [d]} \sum_{j \in [d]} C_{\pi,ij}^+ \text{ and } \Delta_{\max} = f(\mathbf{w}^*) - \min_{w \in \Delta_d} f(w)$.

Remark 5. The choice of $\pi$ will affect the regret factor $\Sigma_\pi^*$ contained in $Z(\rho, \pi)$. Under our construction, $\Sigma_\pi^*$ can be regarded as a uniform representation of covariance $\Sigma^*$, and thus our regret bound demonstrates how the learning performance is influenced by the covariance structure, i.e., negative (positive) correlation shrinks (enlarges) the factor and leads to a lower (higher) regret.

Discussion on the ETE strategy. In contrast to common ETE-type algorithms, MC-ETE requires novel analytical techniques in handling the transformed estimate concentration while preserving the covariance information in regret bounds. In analysis, we build a novel concentration using key matrices $B_+^*$ and $C_+^*$ to adapt to the actual covariance structure, and construct a super-martingale which takes the aggregate noise in an exploration round as analytical basis to prove the concentration. These techniques allow us to capture the correlations in the results, and are new compared to both the former FI/SB settings and covariance-related bandit literature Warmuth and Kuzmin [2012], Degenne and Perchet [2016], Perrault et al. [2020]. In fact, under the full-bandit feedback, it is highly challenging to estimate the covariance without a fixed exploration (i.e., ETE) strategy. Note that even for its simplified offline version, where one uses given (non-fixed) full-bandit data to estimate the covariance, there is no available solution in the statistics literature to our best knowledge. Hence, for such online tasks with severely limited feedback, ETE is the most viable strategy currently available, as used in many partial observation works Lin et al. [2014], Chaudhuri and Tewari [2016], Chen et al. [2018]. We remark that our contribution in this setting focuses on designing a practical solution and deriving regret guarantees which explicitly characterize the correlation impacts. The lower bound for CMCB-FB remains open, which we leave for future work.

7 Experiments

In this section, we present experimental results for our algorithms on both synthetic and real-world datasets. For the synthetic dataset, we set $\theta^* = [0.2, 0.3, 0.2, 0.2, 0.2]^\top$ and $\Sigma^*$ has all diagonal entries equal to 1 and all off-diagonal entries equal to $-0.05$. For the real-world dataset, we use an open dataset US Funds from Yahoo Finance on Kaggle Leone [2020], which provides financial data of 1680 ETF funds in 2010-2017. We select five funds and generate a stochastic distribution ($\theta^*$ and $\Sigma^*$) from the data of returns (since we study a stochastic bandit problem). For both datasets, we set $d = 5$ and $\rho \in \{0.1, 1.0\}$. The random reward $\theta_i$ is drawn i.i.d. from Gaussian distribution $\mathcal{N}(\theta^*, \Sigma^*)$. We perform 50 independent runs for each algorithm and show the average regret and 95% confidence interval across runs, with logarithmic y-axis for clarity of magnitude comparison.

(CMCB-FI) We compare our algorithm MC-Empirical with two algorithms OGD Hazan et al. [2007] and LinearFI. OGD (Online Gradient Descent) Hazan et al. [2007] is designed for general online convex optimization with also a $O(\sqrt{T})$ regret guarantee, but its result cannot capture the covariance impacts as ours. LinearFI is a linear adaption of MC-Empirical that only aims to maximize the expected rewards. Figures 1(a), 1(d) show that our MC-Empirical enjoys multiple orders of magnitude reduction in regret compared to the benchmarks, since it efficiently exploits the empirical observations to select actions and well handles the covariance-based risk. In particular, the performance superiority of MC-Empirical over OGD demonstrates that our sample strategy sufficiently utilize the observed information than conventional gradient descent based policy.

(CMCB-SB) For CMCB-SB, we compare MC-UCB with two adaptions of OLS-UCB Degenne and Perchet [2016] (state-of-the-art for combinatorial bandits with covariance), named MC-UCB-Γ and OLS-UCB-Γ. MC-UCB-Γ uses the confidence region with a universal covariance upper bound $\Gamma$, instead

3In some cases, since algorithms are doing similar procedures (e.g., in Figures 1(c), 1(f), the algorithms are exploring the designed actions) and have low performance variance, the confidence intervals are narrow and indistinguishable.
of the adapting one used in our MC-UCB. OLS-UCB-C directly adapts OLS-UCB Degene and Perchet [2016] to the continuous decision space and only considers maximizing the expected rewards in its objective. As shown in Figures 1(b), 1(e), MC-UCB achieves the lowest regret since it utilizes the covariance information to accelerate the estimate concentration. Due to lack of a covariance-adapting confidence interval, MC-UCB-Γ shows an inferior regret performance than MC-UCB, and OLS-UCB-C suffers the highest regret due to its ignorance of risk.

(CMCB-FB) We compare MC-ETE with two baselines, OGD-ETE, which adopts OGD Hazan et al. [2007] during the exploitation phase, and LinearFB, which only investigates the expected reward maximization. From Figures 1(c), 1(f), one can see that, MC-ETE achieves the best regret performance due to its effective estimation of the covariance-based risk and efficiency in exploitation. Due to the inefficiency of gradient descent based policy in utilizing information, OGD-ETE has a higher regret than MC-ETE, whereas LinearFB shows the worst performance owing to the unawareness of the risk.

8 Conclusion and Future Work

In this paper, we propose a novel continuous mean-covariance bandit (CMCB) model, which investigates the reward-risk trade-off measured by option correlation. Under this model, we consider three feedback settings, i.e., full-information, semi-bandit and full-bandit feedback, to formulate different real-world reward observation scenarios. We propose novel algorithms for CMCB with rigorous regret analysis, and provide lower bounds for the problems to demonstrate our optimality. We also present empirical evaluations to show the superior performance of our algorithms. To our best knowledge, this is the first work to fully characterize the impacts of arbitrary covariance structures on learning performance for risk-aware bandits. There are several interesting directions for future work. For example, how to design an adaptive algorithm for CMCB-FB is a challenging open problem, and the lower bound for CMCB-FB is also worth further investigation.
Acknowledgments and Disclosure of Funding

The work of Yihan Du and Longbo Huang is supported in part by the Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grant 2020AAA0108400 and 2020AAA0108403.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. Advances in Neural Information Processing Systems, 24:2312–2320, 2011.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In Conference on Learning Theory, pages 39–1, 2012.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 47(2-3):235–256, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. SIAM Journal on Computing, 32(1):48–77, 2002b.

Kirt C Butler and Dale L Domian. Risk, diversification, and the investment horizon. Journal of Portfolio Management, 17(3):41, 1991.

Asaf Cassel, Shie Mannor, and Assaf Zeevi. A general approach to multi-armed bandits under risk criteria. In Conference on Learning Theory, pages 1295–1306, 2018.

Sougata Chaudhuri and Ambuj Tewari. Phased exploration with greedy exploitation in stochastic combinatorial partial monitoring games. In Advances in Neural Information Processing Systems, pages 2433–2441. 2016.

Lixing Chen, Jie Xu, and Zhuo Lu. Contextual combinatorial multi-armed bandits with volatile arms and submodular reward. In Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018.

Thomas M Cover. Elements of information theory. John Wiley & Sons, 1999.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In Conference on Learning Theory, 2008a.

Varsha Dani, Sham M Kakade, and Thomas Hayes. The price of bandit information for online optimization. In Advances in Neural Information Processing Systems, volume 20. Curran Associates, Inc., 2008b.

Yahel David and Nahum Shimkin. Pure exploration for max-quantile bandits. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 556–571. Springer, 2016.

Rémy Degenne and Vianney Perchet. Combinatorial semi-bandit with known covariance. In Advances in Neural Information Processing Systems, pages 2972–2980, 2016.

A Diadin. Analysis of a trade company investment project. Business and Economics, page 104, 2019.

Nicolas Galichet, Michele Sebag, and Olivierテイタウド。探索と利得最大化との安全：リスクを伴うマルチアームバンドの安全。In Asian Conference on Machine Learning, pages 245–260, 2013.

Stephan M Gasser, Margarethe Rammerstorfer, and Karl Weinmayer. Markowitz revisited: Social portfolio engineering. European Journal of Operational Research, 258(3):1181–1190, 2017.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. Machine Learning, 69(2-3):169–192, 2007.

Anmol Kagrecha, Jayakrishnan Nair, and Krishna Jagannathan. Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards. Advances in Neural Information Processing Systems, 32:11272–11281, 2019.
Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems*, pages 3913–3922, 2017.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Stefano Leone. Dataset: US funds dataset from yahoo finance. Kaggle, 2020. https://www.kaggle.com/stefanoleone992/mutual-funds-and-etfs?select=ETFs.csv.

Tian Lin, Bruno Abrahao, Robert Kleinberg, John Lui, and Wei Chen. Combinatorial partial monitoring game with linear feedback and its applications. In *International Conference on Machine Learning*, pages 901–909, 2014.

Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In *International Conference on Machine Learning*, pages 1690–1698. PMLR, 2016.

Harry M Markowitz et al. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

Joanne M McInerney and Tim S Roberts. Online learning: Social interaction and the creation of a sense of community. *Journal of Educational Technology & Society*, 7(3):73–81, 2004.

Pierre Perrault, Michal Valko, and Vianney Perchet. Covariance-adapting algorithm for semi-bandits with application to sparse outcomes. In *Conference on Learning Theory*, pages 3152–3184, 2020.

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2012.

Eric M Schwartz, Eric T Bradlow, and Peter S Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

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

Sattar Vakili and Qing Zhao. Mean-variance and value at risk in multi-armed bandit problems. In *The 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1330–1335. IEEE, 2015.

Sattar Vakili and Qing Zhao. Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):1093–1111, 2016.

Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

Manfred K Warmuth and Dima Kuzmin. Online variance minimization. In *International Conference on Computational Learning Theory*, pages 514–528. Springer, 2006.

Manfred K Warmuth and Dima Kuzmin. Online variance minimization. *Machine Learning*, 87(1):1–32, 2012.

Qiuyu Zhu and Vincent YF Tan. Thompson sampling algorithms for mean-variance bandits. *International Conference on Machine Learning*, 2020.
Appendix

A Technical Lemmas

In this section, we introduce two technical lemmas which will be used in our analysis.

Lemmas 1 and 2 give the concentration guarantees of algorithm M-C-UCB for CMCB-SB, which sets up a foundation for the concentration guarantees in CMCB-FI.

Lemma 1 (Concentration of Covariance for CMCB-SB). Consider the CMCB-SB problem and algorithm M-C-UCB (Algorithm 2). Define the event

\[ G_t \triangleq \left\{ |\Sigma^*_{ij} - \hat{\Sigma}_{ij,t-1}| \leq 16 \left( \frac{3 \ln t}{N_{ij}(t-1)} \vee \frac{3 \ln t}{N_i(t-1)} \right) + \sqrt{\frac{61 \ln^2 t}{N_{ij}(t-1)N_i(t-1)} + \frac{36 \ln^2 t}{N_{ij}(t-1)N_j(t-1)}}, \forall i, j \in [d] \right\} \]

For any \( t \geq 2 \), we have

\[ \Pr[G_t] \geq 1 - \frac{10d^2}{t^2}. \]

Proof. According to Proposition 2 in Perrault et al. [2020], we have that for any \( t \geq 2 \) and \( i, j \in [d] \),

\[ \Pr \left[ |\Sigma^*_{ij} - \hat{\Sigma}_{ij,t-1}| \leq 16 \left( \frac{3 \ln t}{N_{ij}(t-1)} \vee \frac{3 \ln t}{N_i(t-1)} \right) + \sqrt{\frac{61 \ln^2 t}{N_{ij}(t-1)N_i(t-1)} + \frac{36 \ln^2 t}{N_{ij}(t-1)N_j(t-1)}} \right] \leq 1 - \frac{10}{t^2}. \]

Using a union bound on \( i, j \in [d] \), we obtain Lemma 1.

Lemma 2 (Concentration of Means for CMCB-SB). Consider the CMCB-SB problem and algorithm M-C-UCB (Algorithm 2). Let \( 0 < \lambda < 1 \), and define \( \delta_t = \frac{1}{t \ln^2 t} \) and \( \beta(\delta_t) = \ln(1/\delta_t) + d \ln \ln t + \frac{d}{2} \ln(1 + e/\lambda) \) for \( t \geq 2 \). Then, for any \( t \geq 2 \) and \( w \in \Delta_{\overline{\Sigma}} \), with probability at least \( 1 - \delta_t \), we have

\[ |w^T \theta^* - w^T \hat{\theta}_{t-1}| \leq \sqrt{2\beta(\delta_t)} \sqrt{w^T D_{t-1}^{-1} \left( \lambda \Lambda \Sigma^* \cdot D_{t-1} + \sum_{s=1}^{t-1} \Sigma^*_{s,w} \right) D_{t-1}^{-1} w}. \]

Further define \( E_t(w) = \sqrt{2\beta(\delta_t)} \sqrt{w^T D_{t-1}^{-1} \left( \lambda \Lambda \Sigma^* \cdot D_{t-1} + \sum_{s=1}^{t-1} \Sigma^*_{s,w} \right) D_{t-1}^{-1} w} \). Then, for any \( t \geq 2 \) and \( w \in \Delta_{\overline{\Sigma}} \), the event \( \mathcal{H}_t \triangleq \{ |w^T \theta^* - w^T \hat{\theta}_{t-1}| \leq E_t(w) \} \) satisfies \( \Pr[\mathcal{H}_t | G_t] \geq 1 - \delta_t \).

Proof. The proof of Lemma 2 follows the analysis procedure in Degenne and Perchet [2016]. Specifically, assuming that event \( G_t \) occurs, we have \( \hat{\Sigma}_{t,i,j} \geq \Sigma^*_{i,j} \) for any \( i, j \in [d] \) and

\[ E_t(w) \geq \sqrt{2\beta(\delta_t)} \sqrt{w^T D_{t-1}^{-1} \left( \lambda \Lambda \Sigma^* \cdot D_{t-1} + \sum_{s=1}^{t-1} \Sigma^*_{s,w} \right) D_{t-1}^{-1} w}. \]
Thus, to prove Lemma 2, it suffices to prove that

\[
\Pr \left[ |w^T \theta^* - w^T \hat{\theta}_{t-1}| > \sqrt{2\beta(\delta_t)} \sqrt{\sum_{i=1}^{t-1} \left( \lambda \Lambda_{\Sigma_i} D_{t-1} + \sum_{s=1}^{t-1} \Sigma_{w,s}^* \right) D_{t-1}^{-1} w} \right] \leq \delta_t. \tag{4}
\]

Recall that \( N_i(t) = \sum_{s=1}^{t} \mathbb{I}\{w_{s,i} \geq c\} \) and \( D_t \) is a diagonal matrix such that \( D_{t,ii} = N_i(t) \) for any \( t > 0 \). For any \( w \in \Delta_d^c \), let \( I_w \) denote the diagonal matrix such that \( I_{w,i} = 1 \) for any \( w_i \geq c \) and \( I_{jj} = 0 \) for any \( j \neq w_j = 0 \), and let \( \Sigma_w = I_w \Sigma^* I_w \). Let \( \epsilon_t \) be the vector such that \( \eta_t = (\Sigma^*)^{1/2} \epsilon_t \) for any \( t > 0 \).

Let \( D \) be a positive definite matrix such that \( D \preceq \lambda \Lambda_{\Sigma^*} D_{t-1} \). Then, we have for any \( w \in \Delta_d^c \) that

\[
|w^T \left( \theta^* - \hat{\theta}_{t-1} \right)| = |w^T D_{t-1}^{-1} \sum_{s=1}^{t-1} I_{w,s} (\Sigma^*)^{1/2} \epsilon_s|
\]

\[
= \left| -w^T D_{t-1}^{-1} \left( D + \sum_{s=1}^{t-1} \Sigma_{w,s} \right)^{1/2} \left( D + \sum_{s=1}^{t-1} \Sigma_{w,s} \right)^{-1/2} \sum_{s=1}^{t-1} I_{w,s} (\Sigma^*)^{1/2} \epsilon_s \right|
\]

\[
\leq \sqrt{w^T D_{t-1}^{-1} \left( D + \sum_{s=1}^{t-1} \Sigma_{w,s} \right) D_{t-1}^{-1} w \cdot \sum_{s=1}^{t-1} I_{w,s} (\Sigma^*)^{1/2} \epsilon_s \left\| (D + \sum_{s=1}^{t-1} \Sigma_{w,s})^{-1} \right\|_2}
\]

Let \( S_t = \sum_{s=1}^{t-1} I_{w,s} (\Sigma^*)^{1/2} \epsilon_s \), \( V_t = \sum_{s=1}^{t-1} \Sigma_{w,s} \), and \( I_{D+V_t} = \frac{1}{2} \| S_t \|^2_{(D+V_t)^{-1}} \). We get

\[
\left\| \sum_{s=1}^{t-1} I_{w,s} (\Sigma^*)^{1/2} \epsilon_s \right\|_{(D + \sum_{s=1}^{t-1} \Sigma_{w,s})^{-1}} = \| S_t \|_{(D+V_t)^{-1}} = \sqrt{2I_{D+V_t}}.
\]

Since \( D \preceq \lambda \Lambda_{\Sigma^*} D_{t-1} \), we have

\[
|w^T \left( \theta^* - \hat{\theta}_{t-1} \right)| \leq \left[ w^T D_{t-1}^{-1} \left( D + \sum_{s=1}^{t-1} \Sigma_{w,s} \right) D_{t-1}^{-1} w \right] \cdot \sqrt{2I_{D+V_t}}
\]

\[
\leq \lambda w^T D_{t-1}^{-1} \Lambda_{\Sigma^*} \lambda w^T + \lambda w^T D_{t-1}^{-1} \left( \sum_{s=1}^{t-1} \Sigma_{w,s}^* \right) D_{t-1}^{-1} w \cdot \sqrt{2I_{D+V_t}}
\]

\[
= w^T D_{t-1}^{-1} \left( \lambda \Lambda_{\Sigma^*} D_{t-1} + \sum_{s=1}^{t-1} \Sigma_{w,s}^* \right) D_{t-1}^{-1} w \cdot \sqrt{2I_{D+V_t}}
\]

Thus,

\[
\Pr \left[ |w^T \left( \theta^* - \hat{\theta}_{t-1} \right)| > \sqrt{2\beta(\delta_t)} \sqrt{\sum_{i=1}^{t-1} \left( \lambda \Lambda_{\Sigma_i} D_{t-1} + \sum_{s=1}^{t-1} \Sigma_{w,s} \right) D_{t-1}^{-1} w} \right] \leq \Pr \left[ \sqrt{w^T D_{t-1}^{-1} \left( \lambda \Lambda_{\Sigma_i} D_{t-1} + \sum_{s=1}^{t-1} \Sigma_{w,s} \right) D_{t-1}^{-1} w} \cdot \sqrt{2I_{D+V_t}} \right]
\]

\[
\leq \Pr \left[ \sqrt{w^T D_{t-1}^{-1} \left( \lambda \Lambda_{\Sigma_i} D_{t-1} + \sum_{s=1}^{t-1} \Sigma_{w,s} \right) D_{t-1}^{-1} w} \cdot \sqrt{2I_{D+V_t}} \right]
\]

14
\[ > \sqrt{2\beta(\delta_t)} \left[ w^\top D_{t-1}^{-1} \left( \lambda \Lambda_{\Sigma^*} D_{t-1} + \sum_{s=1}^{t-1} \Sigma^*_w s \right) D_{t-1}^{-1} w \right] \]

\[ = \Pr [I_{D+V_t} > \beta(\delta_t)] \]

Hence, to prove Eq. (4), it suffices to prove

\[ \Pr [I_{D+V_t} > \beta(\delta_t)] \leq \delta_t. \]  

(5)

To do so, we introduce some notions. Let \( J_t \) be the \( \sigma \)-algebra \( \sigma(w_1, \varepsilon_1, \ldots, w_{t-1}, \varepsilon_{t-1}, w_t) \). Let \( u \in \mathbb{R}^d \) be a multivariate Gaussian random variable with mean \( 0 \) and covariance \( D^{-1} \), which is independent of all the other random variables, and use \( \varphi(u) \) denote its probability density function. Define

\[ P_s^u = \exp \left( u^\top I_w \left( \Sigma^* \right)^{\frac{1}{2}} \varepsilon_s - \frac{1}{2} u^\top \Sigma^*_w u \right), \]

\[ M_t^u = \exp \left( u^\top S_t - \frac{1}{2} \| u \|_V^2 \right), \]

and

\[ M_t \triangleq \mathbb{E}_u [M_t^u] = \int_{\mathbb{R}^d} \exp \left( u^\top S_t - \frac{1}{2} \| u \|_V^2 \right) \varphi(u)du. \]

We have \( M_t^u = \Pi_{s=1}^{t-1} P_s^u \). In the following, we prove \( \mathbb{E}[M_t] \leq 1 \).

For any \( s > 0 \), according to the sub-Gaussian property, \( \eta_s = \left( \Sigma^* \right)^{\frac{1}{2}} \varepsilon_s \) satisfies

\[ \forall v \in \mathbb{R}^d, \mathbb{E} \left[ e^{v^\top \left( \Sigma^* \right)^{\frac{1}{2}} \varepsilon_s} \right] \leq e^{\frac{1}{2} v^\top \Sigma^* v}, \]

which is equivalent to

\[ \forall v \in \mathbb{R}^d, \mathbb{E} \left[ e^{v^\top \left( \Sigma^* \right)^{\frac{1}{2}} \varepsilon_s - \frac{1}{2} v^\top \Sigma^* v} \right] \leq 1. \]

Thus, we have

\[ \mathbb{E} \left[ P_s^u | J_s \right] = \mathbb{E} \left[ \exp \left( u^\top I_w \left( \Sigma^* \right)^{\frac{1}{2}} \varepsilon_s - \frac{1}{2} u^\top \Sigma^*_w u \right) | J_s \right] \leq 1. \]

Then, we can obtain

\[ \mathbb{E}[M_t^u | J_{t-1}] = \mathbb{E} \left[ \Pi_{s=1}^{t-1} P_s^u | J_{t-1} \right] = (\Pi_{s=1}^{t-1} P_s^u) \mathbb{E} \left[ P_s^u | J_{t-1} \right] \leq M_{t-1}^u, \]

which implies that \( M_t^u \) is a super-martingale and \( \mathbb{E}[M_t^u | u] \leq 1 \). Thus,

\[ \mathbb{E}[M_t] = \mathbb{E}_u [\mathbb{E}[M_t^u | u]] \leq 1. \]

According to Lemma 9 in Abbasi-Yadkori et al. [2011], we have

\[ M_t \triangleq \int_{\mathbb{R}^d} \exp \left( u^\top S_t - \frac{1}{2} \| u \|_V^2 \right) \varphi(u)du = \sqrt{\frac{\det D}{\det(D+V_t)}} \exp (I_{D+V_t}). \]
Thus,

\[ \mathbb{E} \left[ \sqrt{\frac{\det D}{\det(D + V_t)}} \exp \left( I_{D + V_t} \right) \right] \leq 1. \]

Now we prove Eq. (5). First, we have

\[
\Pr \left[ I_{D + V_t} > \beta(\delta_t) \right] = \Pr \left[ \sqrt{\frac{\det D}{\det(D + V_t)}} \exp \left( I_{D + V_t} \right) > \sqrt{\frac{\det D}{\det(D + V_t)}} \exp \left( \beta(\delta_t) \right) \right]
\]

\[
= \Pr \left[ M_t > \frac{1}{\sqrt{\det(I + D^{-\frac{1}{2}} V_t D^{-\frac{1}{2}})}} \exp \left( \beta(\delta_t) \right) \right]
\]

\[
\leq \frac{\mathbb{E}[M_t]}{\sqrt{\det(I + D^{-\frac{1}{2}} V_t D^{-\frac{1}{2}})}} \exp \left( \beta(\delta_t) \right)
\]

\[
\leq \sqrt{\det(I + D^{-\frac{1}{2}} V_t D^{-\frac{1}{2}})} \exp \left( \beta(\delta_t) \right)
\]

(6)

Then, for some constant \( \gamma > 0 \) and for any \( a = (a_1, \ldots, a_d) \in \mathbb{N}^d \), we define the set of timesteps \( \mathcal{K}_a \subseteq [T] \) such that

\[
t \in \mathcal{K}_a \iff \forall i \in d, \ (1 + \gamma )a_i \leq N_i(t) < (1 + \gamma )a_i + 1.
\]

Define \( D_a \) a diagonal matrix with \( D_a_{ii} = (1 + \gamma )a_i \).

Suppose \( t \in \mathcal{K}_a \) for some fixed \( a \). We have

\[
\frac{1}{1 + \gamma } D_t \leq D_a \leq D_t.
\]

Let \( D = \lambda \Lambda \Sigma \cdot D_a \geq \frac{\lambda}{1 + \gamma } \Lambda \Sigma \cdot D_t \). Then, we have

\[
D^{-\frac{1}{2}} V_t D^{-\frac{1}{2}} \leq \frac{1 + \gamma}{\lambda} D_t^{-\frac{1}{2}} \Lambda \Sigma^{-\frac{1}{2}} V_t \Lambda \Sigma^{-\frac{1}{2}} D_t^{-\frac{1}{2}},
\]

where matrix \( D_t^{-\frac{1}{2}} \Lambda \Sigma^{-\frac{1}{2}} V_t \Lambda \Sigma^{-\frac{1}{2}} D_t^{-\frac{1}{2}} \) has \( d \) ones on the diagonal. Since the determinant of a positive definite matrix is smaller than the product of its diagonal terms, we have

\[
\det(I + D^{-\frac{1}{2}} V_t D^{-\frac{1}{2}}) \leq \det(I + \frac{1 + \gamma}{\lambda} D_t^{-\frac{1}{2}} \Lambda \Sigma^{-\frac{1}{2}} V_t \Lambda \Sigma^{-\frac{1}{2}} D_t^{-\frac{1}{2}})
\]

\[
\leq \left(1 + \frac{1 + \gamma}{\lambda}\right)^d
\]

(7)

Let \( 0 < \lambda < 1 \) and \( \gamma = e - 1 \). Using Eqs. (6) and (7), \( \beta(\delta_t) = \ln(1/\delta_t) + d \ln \ln t + \frac{d}{2} \ln(1 + e/\lambda) = \ln(t \ln^2 t) + d \ln \ln t + \frac{d}{2} \ln(1 + e/\lambda) \), and a union bound over \( a \), we have

\[
\Pr \left[ I_{D + V_t} > \beta(\delta_t) \right] \leq \sum_a \Pr \left[ I_{D + V_t} > \beta(\delta_t) | t \in \mathcal{K}_a, D = \lambda \Lambda \Sigma \cdot D_a \right]
\]

\[
\leq \sum_a \frac{\sqrt{\det(I + D^{-\frac{1}{2}} V_t D^{-\frac{1}{2}})}}{\exp(\beta(\delta_t))}
\]

16
\[ \leq \left( \frac{\ln t}{\ln(1 + \gamma)} \right)^d \cdot \exp \left( \ln(t \ln^2 t) + d \ln t + \frac{d}{2} \ln(1 + \frac{\gamma}{\lambda}) \right) \]
\[ = (\ln t)^d \cdot \frac{(1 + \frac{\gamma}{\lambda})^{\frac{d}{2}}}{t \ln^2 t \cdot (\ln t)^d \cdot (1 + \frac{\gamma}{\lambda})^{\frac{d}{2}}} \]
\[ = \frac{1}{t \ln^2 t} = \delta_t \]

Thus, Eq. (5) holds and we complete the proof of Lemma 2.

**B Proof for CMCB-FI**

**B.1 Proof of Theorem 1**

In order to prove Theorem 1, we first have the following Lemmas 3 and 4, which are adaptions of Lemmas 1 and 2 to CMCB-FI.

**Lemma 3 (Concentration of Covariance for CMCB-FI).** Consider the CMCB-FI problem and algorithm \texttt{MC-Empirical} (Algorithm 1). For any \( t \geq 2 \), the event

\[ \mathcal{E}_t \triangleq \left\{ |\Sigma_{ij}^* - \hat{\Sigma}_{ij,t-1}| \leq 16 \left( \frac{3 \ln t}{t - 1} \lor \sqrt{\frac{3 \ln t}{t - 1}} \right) + \left( 6 + 4\sqrt{3} \right) \frac{\ln t}{t - 1}, \forall i, j \in [d] \right\} \]

satisfies

\[ \Pr[\mathcal{E}_t] \geq 1 - \frac{10d^2}{t^2} \]

**Proof.** In CMCB-FI, we have \( N_{ij}(t - 1) = t - 1 \) for any \( t \geq 2 \) and \( i, j \in [d] \). Then, Lemma 3 can be obtained by applying Lemma 1 with \( N_{ij}(t - 1) = t - 1 \) for any \( i, j \in [d] \).

**Lemma 4 (Concentration of Means for CMCB-FI).** Consider the CMCB-FI problem and algorithm \texttt{MC-Empirical} (Algorithm 1). Let \( 0 < \lambda < 1 \). Define \( \delta_t = \frac{1}{t \ln^2 t} \) and \( \beta(\delta_t) = \ln(1/\delta_t) + \ln t + \frac{\lambda}{2} \ln(1 + e/\lambda) \) for \( t \geq 2 \). Define \( E_t(w) = \sqrt{2\beta(\delta_t)} \sqrt{w^\top D_{t-1}^{-1} (\lambda \Sigma^*_{ij} D_{t-1} + \sum_{s=1}^{t-1} \Sigma^*_s) D_{t-1}^{-1} w} \). Then, for any \( t \geq 2 \) and \( w \in \Delta_d \), the event \( \mathcal{F}_t \triangleq \{ |w^\top \Theta^* - w^\top \hat{\Theta}_{t-1}| \leq E_t(w) \} \) satisfies \( \Pr[\mathcal{F}_t] \geq 1 - \delta_t \).

**Proof.** In CMCB-FI, \( D_t \) is a diagonal matrix such that \( D_{t,ii} = N_i(t) = t \). Then, Lemma 4 can be obtained by applying Lemma 2 with \( D_t = tI \) and that the union bound on the number of samples only needs to consider one dimension. Specifically, in the proof of Lemma 2, we replace the set of timesteps \( K_a \) with \( K_a \subseteq [T] \) for \( a \in \mathbb{N} \), which stands for

\[ t \in K_a \Leftrightarrow (1 + \gamma)^a \leq t < (1 + \gamma)^{a+1}. \]

This completes the proof.

Now we are ready to prove Theorem 1.

**Proof.** (Theorem 1) Let \( \Delta_t = f(w^*) - f(w_t) \), \( g(t) = 16 \left( \frac{3 \ln t}{t - 1} \lor \sqrt{\frac{3 \ln t}{t - 1}} \right) + \left( 6 + 4\sqrt{3} \right) \frac{\ln t}{t - 1} \) denote the confidence radius of covariance \( \Sigma^*_{ij} \) for any \( i, j \in [d] \), and \( G(t) \) be the matrix with all entries equal to \( g(t) \). For any \( w \in \Delta_{\tilde{\gamma}} \), define \( \tilde{f}_t(w) = w^\top \hat{\Theta}_t - \rho w^\top \hat{\Sigma}_t w \) and \( h_t(w) = E_t(w) + \rho w^\top G_t w \).
For any $t \geq 2$, suppose that event $E_t \cap F_t$ occurs. Then,
\[ |\hat{f}_{t-1}(w) - f(w)| \leq h_t(w). \]
Therefore, we have
\[ \Delta_t \leq |\hat{f}_{t-1}(w^*) - f(w^*)| + |\hat{f}_{t-1}(w_t) - f(w_t)|. \]
This is because if instead $\Delta_t > |\hat{f}_{t-1}(w^*) - f(w^*)| + |\hat{f}_{t-1}(w_t) - f(w_t)|$, we have
\[
\begin{align*}
& \hat{f}_{t-1}(w^*) - \hat{f}_{t-1}(w_t) \\
= & \hat{f}_{t-1}(w^*) - \hat{f}_{t-1}(w_t) + (f(w^*) - f(w_t)) - (f(w^*) - f(w_t)) \\
\geq & \Delta_t - (f(w^*) - \hat{f}_{t-1}(w^*)) - (\hat{f}_{t-1}(w_t) - f(w_t)) \\
\geq & \Delta_t - |(f(w^*) - \hat{f}_{t-1}(w^*))| - |(\hat{f}_{t-1}(w_t) - f(w_t))| \\
> & 0,
\end{align*}
\]
which contradicts the selection strategy of $w_t$ in algorithm MC-Empirical. Thus, we obtain
\[
\begin{align*}
\Delta_t \leq & |\hat{f}_{t-1}(w^*) - f(w^*)| + |\hat{f}_{t-1}(w_t) - f(w_t)| \\
\leq & h_t(w^*) + h_t(w_t) \\
= & E_i(w^*) + \rho w^\top G_i w^* + E_i(w_t) + \rho w_t^\top G_i w_t
\end{align*}
\]
(8)
Now, for any $w \in \Delta^*_t$, we have
\[
\begin{align*}
w^\top G_i w &= \sum_{i,j [d]} g(t) w_i w_j \\
&= g(t) \sum_{i,j [d]} w_i w_j \\
&= g(t) \left( \sum_i w_i \right)^2 \\
&= g(t)
\end{align*}
\]
and
\[
E_i(w) = \sqrt{2\beta(\delta_t)} \sqrt{w^\top D_i^{-1} \left( \lambda \Lambda_{\Sigma^*} + \sum_{a=1}^{t-1} \Sigma^* \right) D_i^{-1} w}
= \sqrt{2\beta(\delta_t)} \sqrt{\lambda w^\top D_i^{-1} \Lambda_{\Sigma^*} w + w^\top D_i^{-1} \left( \sum_{s=1}^{t-1} \Sigma^* \right) D_i^{-1} w}
= \sqrt{2\beta(\delta_t)} \sqrt{\frac{1}{t-1} \lambda w^\top \Lambda_{\Sigma^*} w + \frac{1}{t-1} w^\top \Sigma^* w}
\leq \sqrt{\frac{2\beta(\delta_t)}{t-1} \lambda \Sigma^* + w^\top \Sigma^* w},
\]
where $\Sigma^*_\text{max}$ denotes the maximum diagonal entry of $\Sigma^*$. 

18
For any $t \geq 7$, $\frac{3\ln t}{t-1} < \frac{3\ln t}{t-1}$ and $g(t) \leq (6 + 20\sqrt{3}) \frac{\ln t}{\sqrt{t-1}}$, Eq. (8) can be written as

$$\Delta_t \leq E_t(w^*) + \rho w^* \Sigma w^* + E_t(w_t) + \rho w_t^\top G_t w_t$$

$$\leq \sqrt{\frac{2\beta(\delta_t)}{t-1}} \left( \sqrt{\lambda \Sigma_{\max}^* + w_t^\top \Sigma w_t} + \sqrt{\lambda \Sigma_{\max}^* + w^* \Sigma w^*} \right) + 82\rho \frac{\ln t}{\sqrt{t-1}}$$

$$= \sqrt{\frac{2\beta(\delta_t)}{t-1}} \left( 2\sqrt{\lambda \Sigma_{\max}^* + w^* \Sigma w^*} \right) + 82\rho \frac{\ln t}{\sqrt{t-1}} + \sqrt{2\beta(\delta_t) \frac{\ln t}{\sqrt{t-1}} w_t^\top \Sigma w_t}$$

(9)

Next, we investigate the upper bound of $w_t^\top \Sigma w_t$. According to Eq. (8), we have that $w_t^\top \Sigma w_t$ satisfies

$$\Delta_t \leq h_t(w^*) + h_t(w_t)$$

(10)

In Eq. (10), we have

$$\Delta_t = f(w^*) - f(w_t) \geq \theta_{\min} - \rho w^* \Sigma^* w^* - \theta_{\max} + \rho w_t^\top \Sigma^* w_t$$

and

$$h_t(w^*) + h_t(w_t) = E_t(w^*) + \rho w^* \Sigma^* w^* + E_t(w_t) + \rho w_t^\top G_t w_t$$

$$\leq \sqrt{\frac{2\beta(\delta_t)}{t-1}} \left( 2\sqrt{\lambda \Sigma_{\max}^* + w_t^\top \Sigma \Sigma w_t} + \sqrt{\lambda \Sigma_{\max}^* + w^* \Sigma^* w^*} \right) + 82\rho \frac{\ln t}{\sqrt{t-1}}$$

Thus, $w_t^\top \Sigma^* w_t$ satisfies

$$\theta_{\min} - \rho w^* \Sigma^* w^* - \theta_{\max} + \rho w_t^\top \Sigma^* w_t \leq \sqrt{\frac{2\beta(\delta_t)}{t-1}} \left( 2\sqrt{\lambda \Sigma_{\max}^* + w_t^\top \Sigma \Sigma w_t} + \sqrt{\lambda \Sigma_{\max}^* + w^* \Sigma^* w^*} \right) + 82\rho \frac{\ln t}{\sqrt{t-1}}$$

Rearranging the terms, we have

$$\rho w_t^\top \Sigma w_t - \sqrt{\frac{2\beta(\delta_t)}{t-1}} \sqrt{w_t^\top \Sigma w_t} - \left( \theta_{\max} - \theta_{\min} + \rho w^* \Sigma w^* \right) \leq \sqrt{\frac{2\beta(\delta_t)}{t-1}} \left( 2\sqrt{\lambda \Sigma_{\max}^* + w^* \Sigma^* w^*} \right) + 82\rho \frac{\ln t}{\sqrt{t-1}}$$

(11)

Let $x = w_t^\top \Sigma^* w_t$ and $0 < \lambda < 1$. Define function

$$y(x) = \rho x - c_1 \sqrt{x} - c_2 \leq 0,$$

where $c_1 = \sqrt{\frac{2\beta(\delta_t)}{t-1}} > 0$ and $c_2 = \theta_{\max}^* - \theta_{\min}^* + \rho w^* \Sigma^* w^* + \sqrt{\frac{2\beta(\delta_t)}{t-1}} \left( 2\sqrt{\lambda \Sigma_{\max}^* + w^* \Sigma^* w^*} \right) + 82\rho \frac{\ln t}{\sqrt{t-1}} > 0$.

Now since

$$y(x) = \rho x - c_1 \sqrt{x} - c_2 = \rho \left( \sqrt{x} - \frac{c_1}{2\rho} \right)^2 - \frac{c_1^2}{4\rho} - c_2,$$

19
by letting \( y(x) \leq 0 \), we have

\[
x \leq \left( \frac{c_1}{2\rho} + \sqrt{\frac{c_1^2}{4\rho^2} + \frac{c_2}{\rho}} \right)^2
\]

\[
\leq \frac{c_1^2}{4\rho^2} + \frac{c_1^2}{4\rho^2} + \frac{2c_2}{\rho}
\]

\[
= \frac{c_1^2}{\rho^2} + \frac{2c_2}{\rho}
\]

Therefore

\[
w_t^\top \Sigma^* w_t \leq \frac{2\beta(\delta_t)}{t - 1} + \frac{2}{\rho} \left( \theta_{\max}^* - \theta_{\min}^* + \rho w^* \Sigma^* w^* + \sqrt{\frac{2\beta(\delta_t)}{t - 1}} \left( 2\sqrt{\lambda \Sigma_{\max}^*} + \sqrt{w^* \Sigma w^*} \right) \right)
\]

\[
+ 82\rho \ln t \sqrt{t - 1}
\]

\[
\leq 2w^* \Sigma^* w^* + \frac{2}{\rho} (\theta_{\max}^* - \theta_{\min}^*) + \frac{2}{\rho} \sqrt{\frac{2\beta(\delta_t)}{t - 1}} \left( 2\sqrt{\lambda \Sigma_{\max}^*} + \sqrt{w^* \Sigma w^*} \right) + 164 \ln t \sqrt{t - 1}
\]

Thus, we have that, \( w_t^\top \Sigma^* w_t \) satisfies

\[
w_t^\top \Sigma^* w_t \leq \min \left\{ 2w^* \Sigma^* w^* + \frac{2}{\rho} (\theta_{\max}^* - \theta_{\min}^*) + \frac{2}{\rho} \sqrt{\frac{2\beta(\delta_t)}{t - 1}} \left( 2\sqrt{\lambda \Sigma_{\max}^*} + \sqrt{w^* \Sigma w^*} \right) + 164 \ln t \sqrt{t - 1} + 1 + \frac{2\beta(\delta_t)}{t - 1}, w_{\max}^\top \Sigma^* w_{\max} \right\},
\]

(12)

where \( w_{\max}^\top \triangleq \arg\max_{w \in \Delta_d} w^\top \Sigma^* w \).

Below we discuss the two terms in Eq. (12) separately.

Case (i): Plugging the first term of the upper bound of \( w_t^\top \Sigma^* w_t \) in Eq. (12) into Eq. (9), we have that for \( t \geq t_0 \),

\[
\Delta_t \leq \sqrt{\frac{2\beta(\delta_t)}{t - 1}} \left( 2\sqrt{\lambda \Sigma_{\max}^*} + \sqrt{w^* \Sigma w^*} \right) + 82\rho \ln t \sqrt{t - 1} + \sqrt{\frac{2\beta(\delta_t)}{t - 1}} \cdot \sqrt{w_t^\top \Sigma^* w_t}
\]

\[
\leq \sqrt{\frac{2\beta(\delta_t)}{t - 1}} \left( 2\sqrt{\lambda \Sigma_{\max}^*} + \sqrt{w^* \Sigma w^*} \right) + 82\rho \ln t \sqrt{t - 1} + \sqrt{\frac{2\beta(\delta_t)}{t - 1}} \cdot \sqrt{\frac{2\beta(\delta_t)}{t - 1}} \left( 2\sqrt{\lambda \Sigma_{\max}^*} + \sqrt{w^* \Sigma w^*} \right) + 164 \ln t \sqrt{t - 1} + \frac{1}{\rho^2} \cdot \frac{2\beta(\delta_t)}{t - 1}
\]

\[
\left( \sqrt{2w^* \Sigma w^*} + \sqrt{\frac{2}{\rho} (\theta_{\max}^* - \theta_{\min}^*)} + \frac{2}{\rho} \sqrt{\frac{2\beta(\delta_t)}{t - 1}} \left( 2\sqrt{\lambda \Sigma_{\max}^*} + \sqrt{w^* \Sigma w^*} \right) \right)^\frac{1}{2}
\]

\[
\left( \sqrt{2w^* \Sigma w^*} + \sqrt{\frac{2}{\rho} (\theta_{\max}^* - \theta_{\min}^*)} + \frac{2}{\rho} \sqrt{\frac{2\beta(\delta_t)}{t - 1}} \left( 2\sqrt{\lambda \Sigma_{\max}^*} + \sqrt{w^* \Sigma w^*} \right) \right)^\frac{1}{2}
\]

20
According to Lemmas 3 and 4, for any \( t \geq 2 \), we bound the probability of event \( \neg (\mathcal{E}_t \cap \mathcal{F}_t) \) as follows.

\[
\Pr [\neg (\mathcal{E}_t \cap \mathcal{F}_t)] \leq \frac{10d^2}{t^2} + \frac{1}{t \ln^2 t} \\
\leq \frac{10d^2}{t \ln^2 t} + \frac{1}{t \ln^2 t} \\
= \frac{11d^2}{t \ln^2 t}
\]

Recall that

\[
\beta(\delta_t) = \ln(t \ln^2 t) + \ln \ln t + \frac{d}{2} \ln(1 + e/\lambda) = O(\ln t + d \ln(1 + \lambda^{-1})).
\]

Let \( t_0 = 7 \). For any horizon \( T \) which satisfies \( T \geq 1 + \lambda^{-1} \), summing over \( t = 1, \ldots, T \), we obtain the regret upper bound

\[
\mathbb{E}[\mathcal{R}(T)] = O(t_0) + \sum_{t=t_0}^{T} O(\Delta_{\text{max}} \cdot \Pr [\neg (\mathcal{E}_t \cap \mathcal{F}_t)] + \Delta_t \cdot \mathbb{1}(\mathcal{E}_t \cap \mathcal{F}_t))
\]

\[
= \sum_{t=t_0}^{T} O \left( \Delta_{\text{max}} \cdot \frac{d^2}{t \ln^2 t} + \Delta_t \cdot \mathbb{1}(\mathcal{E}_t \cap \mathcal{F}_t) \right)
\]

\[
= \sum_{t=t_0}^{T} O \left( \sqrt{\frac{\ln t + d \ln(1 + \lambda^{-1})}{t - 1}} \left( \sqrt{\lambda \Sigma_{\text{max}}^*} + \sqrt{w_{\text{max}}^* \Sigma^* w_{\text{max}}} + \frac{1}{\sqrt{\rho}} \sqrt{\theta_{\text{max}}^* - \theta_{\text{min}}^*} \right) + \rho \frac{\ln t}{\sqrt{t - 1}} \right)
\]

\[
= O \left( \ln T \sqrt{d T} \left( \sqrt{\lambda \Sigma_{\text{max}}^*} + \sqrt{w_{\text{max}}^* \Sigma^* w_{\text{max}}} + \frac{1}{\sqrt{\rho}} \sqrt{\theta_{\text{max}}^* - \theta_{\text{min}}^*} + \rho \right) \right)
\]

Case (ii): Plugging the second term of the upper bound of \( w_{\text{max}}^T \Sigma^* w_t \) in Eq. (12) into Eq. (9), we have that for \( t \geq t_0 \),

\[
\Delta_t \leq \frac{2 \beta(\delta_t)}{t - 1} \left( 2 \sqrt{\lambda \Sigma_{\text{max}}^*} + \sqrt{w_{\text{max}}^* \Sigma^* w_{\text{max}}} + \sqrt{w_{\text{max}}^* \Sigma^* w_{\text{max}}} \right) + 82\rho \frac{\ln t}{\sqrt{T - 1}}
\]

\[
\leq 2 \frac{2 \beta(\delta_t)}{t - 1} \left( \sqrt{\lambda \Sigma_{\text{max}}^*} + \sqrt{w_{\text{max}}^* \Sigma^* w_{\text{max}}} \right) + 82\rho \frac{\ln t}{\sqrt{T - 1}}
\]
For any horizon $T$ which satisfies $T \geq 1 + \lambda^{-1}$, summing over $t = 1, \ldots, T$, we obtain the regret upper bound

\[
\mathbb{E}[R(T)] = O(t_0) + \sum_{t=t_0}^{T} O(\Delta_{\max} \cdot \Pr[\neg(E_t \cap F_t)] + \Delta_t \cdot \mathbb{I}\{E_t \cap F_t\})
\]

\[
= \sum_{t=t_0}^{T} O\left(\Delta_{\max} \cdot \frac{d^2}{t \ln^2 t} + \sqrt{\frac{\beta(\delta_t)}{t - 1}} \left(\sqrt{\lambda \Sigma_{\max}^* + \sqrt{w_{\max}^* \Sigma_{\max}^* w_{\max}^*}} + \rho \frac{\ln t}{\sqrt{t - 1}}\right)\right)
\]

\[
= \sum_{t=t_0}^{T} O\left(\sqrt{\lambda \Sigma_{\max}^* + \sqrt{w_{\max}^* \Sigma_{\max}^* w_{\max}^*}} + \rho \frac{\ln t}{\sqrt{t - 1}}\right)
\]

Combining cases (i) and (ii), we can obtain

\[
\mathbb{E}[R(T)] = O\left(\ln T \sqrt{dT} \left(\min \left\{\sqrt{w^* \Sigma^* w^* + \rho^{-\frac{1}{2}} \sqrt{\theta^{*\min} - \theta^{*\min}}, \sqrt{w_{\max}^* \Sigma_{\max}^* w_{\max}^*}} + \sqrt{\lambda \Sigma_{\max}^* + \rho}\right)\right)\right)
\]

Let $\Sigma_{\max}^* = \max_{i \in [d]} \Sigma_{ii}^*$. Letting $\lambda = \frac{w^* \Sigma^* w^*}{\Sigma_{\max}^*}$ and $T \geq 1 + \frac{\Sigma_{\max}^*}{w^* \Sigma^* w^*}$, we obtain

\[
\mathbb{E}[R(T)] = O\left(\ln T \sqrt{dT} \left(\min \left\{\sqrt{w^* \Sigma^* w^* + \rho^{-\frac{1}{2}} \sqrt{\theta^{*\min} - \theta^{*\min}}, \sqrt{w_{\max}^* \Sigma_{\max}^* w_{\max}^*}} + \rho\right)\right)\right)
\]

\[
= O\left(\ln T \sqrt{dT} \left(\min \left\{\sqrt{w^* \Sigma^* w^* + \rho^{-\frac{1}{2}} \sqrt{\theta^{*\min} - \theta^{*\min}}, \sqrt{\Sigma_{\max}^*}} + \rho\right)\right)\right),
\]

which completes the proof.

\[\square\]

**B.2 Proof of Theorem 2**

In order to prove Theorem 2, we first analyze the offline problem of CMCB-FI. Suppose that the covariance matrix $\Sigma^*$ is positive definite.

**(Offline Problem of CMCB-FI)** We define the quadratic optimization $\text{QuadOpt}(\theta^*, \Sigma^*)$ as

\[
\min_{w} \quad f(w) = \rho w^\top \Sigma^* w - w^\top \theta^*
\]

s.t. $w_i \geq 0, \quad \forall i \in [d]$

\[
\sum_{i=1}^{d} w_i = 1
\]

and $w^*$ as the optimal solution to $\text{QuadOpt}(\theta^*, \Sigma^*)$. We consider the KKT condition for this quadratic optimization as follows:

\[
2\rho \Sigma^* w - \theta^* - u - v 1 = 0
\]

\[
w_i u_i = 0, \quad \forall i \in [d]
\]

\[
u_i \geq 0, \quad \forall i \in [d]
\]
The posterior of $\Sigma$ is still $\Sigma_i$ for the problem. It suffices to prove a regret lower bound of algorithm $\mathcal{A}$ for the problem, it suffices to prove a regret lower bound of algorithm $\mathcal{A}$ for the problem. We consider the case when $w^*$ and $w_t$ both lie in the interior of the $d$-dimensional probability simplex $\triangle_d$, i.e., $w^*_i > 0, \forall i \in [d]$ and $w_{t,i} > 0, \forall i \in [d]$. From Eq. (13), $w^*$ satisfies
\[
\frac{1}{2\rho} I^{-1} \theta^* + \frac{1 - \frac{1}{2\rho} I^{-1} \theta^*}{\|I^{-1}\|} I^{-1} 1 > 0.
\]
Rearranging the terms, we have
\[
\|\theta^*\|_1 - d \theta^* < 2\rho 1.
\]
which is equivalent to
\[
\begin{aligned}
\theta_2^* + \cdots + \theta_d^* - (d-1)\theta_1^* &< 2\rho \\
\vdots \\
\theta_1^* + \cdots + \theta_{d-1}^* - (d-1)\theta_d^* &< 2\rho 
\end{aligned}
\]  
(14)

Similarly, \( w_t \) satisfies
\[
\begin{aligned}
\mu_2^* + \cdots + \mu_d^* - (d-1)\mu_1 &< 2\rho \\
\vdots \\
\mu_1^* + \cdots + \mu_{d-1}^* - (d-1)\mu_d &< 2\rho 
\end{aligned}
\]  
(15)

We first derive a condition that makes \( \mu_t \) lie in the interior of \( \triangle_d \). Recall that \( \mu_t \sim \mathcal{N}(0, \frac{t}{t+\omega}I) \). Define event
\[
\mathcal{E}_t \triangleq \left\{ -3\sqrt{\frac{t}{(t+\omega)\omega}} \leq \mu_t \leq 3\sqrt{\frac{t}{(t+\omega)\omega}} \right\}.
\]
According to the 3 – \( \sigma \) principle for Gaussian distributions, we have
\[
\Pr [\mathcal{E}_t] \geq (99.7\%)^d.
\]
Conditioning on \( \mathcal{E}_t \), under which Eq. (15) hold, it suffices to let
\[
3(d-1)\sqrt{\frac{t}{(t+\omega)\omega}} - \left( -3(d-1)\sqrt{\frac{t}{(t+\omega)\omega}} \right) < \rho,
\]
which is equivalent to
\[
\left( \frac{1+\omega}{t}\right) \omega > \frac{36(d-1)^2}{\rho^2},
\]  
(16)

when \( t > 0, m > 0, t + \omega > 0 \). Let \( t_1 > 0 \) be the smallest timestep that satisfies Eq. (16). Thus, when \( \mathcal{E}_t \) occurs and \( t \geq t_1, \mu_t \) lie in the interior of \( \triangle_d \).

Next, we derive some condition that make \( \theta^* \) lie in the interior of \( \triangle_d \). Recall that \( \theta^* | \mu_t \sim \mathcal{N}(\mu_t, \frac{1}{(t+\omega)I}) \).

Fix \( \mu_t \), and then we define event
\[
\mathcal{F}_t \triangleq \left\{ -3\sqrt{\frac{t}{(t+\omega)\omega}} \leq \theta^* - \mu_t \leq 3\sqrt{\frac{t}{(t+\omega)\omega}} \right\}.
\]
According to the 3 – \( \sigma \) principle for Gaussian distributions, we have
\[
\Pr [\mathcal{F}_t] \geq (99.7\%)^d.
\]
Conditioning on \( \mathcal{F}_t \), in order to let Eq. (14) hold, it suffices to let
\[
3(d-1)\frac{1}{\sqrt{t+\omega}} - \left( -3(d-1)\frac{1}{\sqrt{t+\omega}} \right) < \rho,
\]
which is equivalent to
\[
t > \frac{36(d-1)^2}{\rho^2} - \omega,
\]  
(17)

when \( t + \omega > 0 \). Let \( t_2 > 0 \) be the smallest timestep that satisfies Eq. (17). Thus, when \( \mathcal{F}_t \) occurs and \( t \geq t_2, \theta^* \) lie in the interior of \( \triangle_d \).
Step (ii). Suppose that $E_i \cap F_i \cap G_i$ occurs and consider $t \geq \hat{t} \triangleq \max\{t_1, t_2\}$. Then, $w^*$ and $w_t$ both lie in the interior of $\Delta_d$, i.e., $w^*_{i} > 0, \forall i \in [d]$ and $w_{t,i} > 0, \forall i \in [d]$. We have

$$w^* = \frac{1}{2\rho}(\Sigma^*)^{-1} \theta^* + \frac{1 - \frac{\rho}{4}(\Sigma^*)^{-1} \theta^*}{\|\Sigma^*\|^{-1}} (\Sigma^*)^{-1} 1$$

$$w_t = \frac{1}{2\rho}(\Sigma^*)^{-1} \mu_t + \frac{1 - \frac{\rho}{4}(\Sigma^*)^{-1} \mu_t}{\|\Sigma^*\|^{-1}} (\Sigma^*)^{-1} 1$$

Let $\Delta_{t_i} \triangleq \mu_t - \theta^*$ and thus $\Delta_{t_i} | \mu_t \sim N(0, \frac{1}{t + \omega} I)$. Let $\Delta w_t \triangleq w_t - w^* = \frac{1}{2\rho}(\Sigma^*)^{-1} \Delta_{t_i} - \frac{\rho}{4}(\Sigma^*)^{-1} \Delta_{t_i} | \mu_t = \frac{1}{2\rho} \Delta_{t_i} - \frac{\rho}{4d} \|\Delta_{t_i}\| I$. Then, we have

$$f(w^*) - f(w_t) = f(w^*) - f(w^* + \Delta w_t)$$

$$= ((w^*)^\top \theta^* - \rho(w^*)^\top \Sigma^* w^*)$$

$$- ((w^*)^\top \theta^* + (\Delta w_t)^\top \theta^* - \rho(w^*)^\top \Sigma^* w^* - 2\rho(\Delta w_t)^\top \Sigma^* w^* - \rho(\Delta w_t)^\top \Sigma^* \Delta w_t)$$

$$= - (\Delta w_t)^\top \theta^* + 2\rho(\Delta w_t)^\top \Sigma^* w^* + \rho(\Delta w_t)^\top \Sigma^* \Delta w_t$$

$$= (\Delta w_t)^\top \nabla f(\theta^*) + \rho(\Delta w_t)^\top \Sigma^* \Delta w_t$$

$$\geq \rho(\Delta w_t)^\top \Sigma^* \Delta w_t$$

$$= \rho(\Delta w_t)^\top \Delta w_t$$

$$= \rho \left( \frac{1}{4\rho^2} (\Delta_{t_i})^\top \Delta_{t_i} - \frac{1}{4\rho^2 d^2} \|\Delta_{t_i}\|^2 d \right)$$

$$= \frac{1}{4\rho} \left( \sum_{i=1}^d \Delta_{t,i}^2 - \frac{\left( \sum_{i=1}^d \Delta_{t,i} \right)^2}{d} \right)$$

$$= \frac{1}{4\rho d} \sum_{1 \leq i < j \leq d} (\Delta_{t,i} - \Delta_{t,j})^2$$

Let $i_1 = 1, j_1 = 2, i_2 = 3, j_2 = 4, \ldots, i_{\lceil d/2 \rceil} = 2 \lceil d/2 \rceil - 1, j_{\lceil d/2 \rceil} = 2 \lceil d/2 \rceil$. For any $i_k, j_k (k \in \lceil d/4 \rceil)$, $\Delta_{t,i_k} - \Delta_{t,j_k} | \mu_t \sim N(0, \frac{1}{t + \omega})$ and they are mutually independent among $k$.

Fix $\mu_t$, and then we define event

$$G_{t_i} \triangleq \left\{ |\Delta_{t,i} - \Delta_{t,j}| \geq 0.3 \sqrt{\frac{2}{t + \omega}}, \forall k \in \lceil d/2 \rceil \right\}.$$ 

From the c.d.f. of Gaussian distributions, we have

$$Pr[G_{t_i}] \geq (75\%)^\lceil d/2 \rceil.$$

From this, we get

$$Pr[F_{t} \cap G_{t_i}] \geq 1 - Pr[F_{t}] - Pr[G_{t_i}]$$

$$\geq 1 - (1 - (99.7\%)^d) - (1 - (75\%)^{\lceil d/2 \rceil})$$

$$\geq (99.7\%)^d + (75\%)^{\lceil d/2 \rceil} - 1.$$

When $d \leq 18$, $Pr[F_{t} \cap G_{t_i}] \geq (99.7\%)^d + (75\%)^{\lceil d/2 \rceil} - 1 > 0$.  

25
Step (iii). We bound the expected regret by considering the event $\mathcal{E}_t \cap \mathcal{F}_t \cap \mathcal{G}_t$ and $t \geq \tilde{t}$. Specifically,

$$
\sum_{t=1}^{T} \mathbb{E}_{\mu_t \sim N(0, \frac{1}{(T-t+1)})} \left[ \mathbb{E}_{\theta^* | \mu_t \sim N(\mu_t, \frac{1}{T-t+1})} \left[ f(w^*) - f(w_t) \right] \right] \\
\geq \sum_{t=1}^{T} \mathbb{E}_{\mu_t \sim N(0, \frac{1}{(T-t+1)})} \left[ \mathbb{E}_{\theta^* | \mu_t \sim N(\mu_t, \frac{1}{T-t+1})} \left[ f(w^*) - f(w_t) \right] \right] \Pr[\mathcal{E}_t] \\
\geq \sum_{t=1}^{T} \mathbb{E}_{\mu_t \sim N(0, \frac{1}{(T-t+1)})} \left[ \mathbb{E}_{\theta^* | \mu_t \sim N(\mu_t, \frac{1}{T-t+1})} \left[ f(w^*) - f(w_t) \right] \right] \Pr[\mathcal{F}_t \cap \mathcal{G}_t] \Pr[\mathcal{E}_t] \\
\geq \sum_{t=1}^{T} \mathbb{E}_{\mu_t \sim N(0, \frac{1}{(T-t+1)})} \left[ \mathbb{E}_{\theta^* | \mu_t \sim N(\mu_t, \frac{1}{T-t+1})} \left[ \frac{1}{4\rho d} \sum_{1 \leq i < j \leq d} (\Delta \theta_{t,i} - \Delta \theta_{t,j})^2 \right] \Pr[\mathcal{F}_t \cap \mathcal{G}_t] \Pr[\mathcal{E}_t] \right] \\
\geq \sum_{t=1}^{T} \mathbb{E}_{\mu_t \sim N(0, \frac{1}{(T-t+1)})} \left[ \mathbb{E}_{\theta^* | \mu_t \sim N(\mu_t, \frac{1}{T-t+1})} \left[ \frac{1}{4\rho d} \cdot 0.3^2 \cdot \frac{2}{t + \omega} \right] \Pr[\mathcal{F}_t \cap \mathcal{G}_t] \Pr[\mathcal{E}_t] \right] \\
\geq \sum_{t=1}^{T} \mathbb{E}_{\mu_t \sim N(0, \frac{1}{(T-t+1)})} \left[ \mathbb{E}_{\theta^* | \mu_t \sim N(\mu_t, \frac{1}{T-t+1})} \left[ \frac{1}{4\rho d} \cdot 0.3^2 \cdot \frac{2}{t + \omega} \right] \Pr[\mathcal{F}_t \cap \mathcal{G}_t] \Pr[\mathcal{E}_t] \right] \\
\geq \sum_{t=1}^{T} 0.01125 \left[ (99.7%)^d + (75\%)^{\frac{7\%}{18}} - 1 \right] (99.7%)^d \\
= \frac{0.01125 \left[ (99.7%)^d + (75\%)^{\frac{7\%}{18}} - 1 \right] (99.7%)^d}{\rho} \ln \left( \frac{t + \omega}{t + \omega} \right).
$$

In the following, we consider an intrinsic bound of the expected regret and set the problem parameters to proper quantities. Since the expected regret is upper bounded by $\Delta_{\text{max}} T$ and

$$
\Delta_{\text{max}} = f_{\text{max}} - f_{\text{min}} \\
\geq f \left( \frac{1}{d} \right) - (\theta^*_{\text{min}} - \rho) \\
= \frac{1}{d} \sum_{i=1}^{d} \theta^*_{i} - \rho \frac{1}{d} - (\theta^*_{\text{min}} - \rho) \\
= \frac{1}{d} \sum_{i=1}^{d} \theta^*_{i} - \theta^*_{\text{min}} + \frac{d - 1}{d} \rho \\
> 0,
$$

we conclude that the expected regret is lower bounded by

$$
\min \left\{ \frac{0.01125 \left[ (99.7%)^d + (75\%)^{\frac{7\%}{18}} - 1 \right] (99.7%)^d}{\rho} \ln \left( \frac{t + \omega}{t + \omega} \right), \left( \frac{1}{d} \sum_{i=1}^{d} \theta^*_{i} - \theta^*_{\text{min}} + \frac{d - 1}{d} \rho \right) T \right\}.
$$

Choose $\omega = 36(d - 1)^2 T$ and $\rho = \frac{1}{\sqrt{T}}$. According to Eqs. (16) and (17), we have $t_1 = t_2 = \tilde{t} = 1$. Therefore, for $d \leq 18$, the expected regret is lower bounded by $\Omega(\sqrt{T})$. 

\[\Box\]
C Proof for CMCB-SB

C.1 Proof of Theorem 3

Proof. (Theorem 3) Denote $\Delta_t = f(w^*) - f(w_t)$, $\bar{f}_t(w) = w^\top \hat{\theta}_{t-1} + E_t(w) - \rho w^\top \Sigma_{t-1} w$ and $G_t$ the matrix whose $i,j$-th entry is $g_{ij}(t)$.

For $t \geq d^2 + 1$, suppose that event $G_t \cap H_t$ occurs. Define $h_t(w) \triangleq E_t(w) + \rho w^\top G_t w$ for any $w \in \Delta^c_t$. Then, we have

$$0 \leq \bar{f}_t(w) - f(w) \leq 2h_t(w).$$

According to the selection strategy of $w_t$, we obtain

$$f(w^*) - f(w_t) \leq \bar{f}_t(w^*) - f(w_t) \leq \bar{f}_t(w_t) - f(w_t) \leq 2h_t(w_t)$$

Thus, for any $T \geq d^2 + 1$, we have that

$$\sum_{t=d^2+1}^{T} (f(w^*) - f(w_t)) \leq 2 \sum_{t=d^2+1}^{T} h_t(w_t)$$

$$= 2 \sum_{t=d^2+1}^{T} \left( \sqrt{2\beta(t)} \sqrt{w^\top D^{-1}_{t-1}(\lambda \Lambda_{\Sigma_t} D_{t-1} + \sum_{s=1}^{t-1} \Sigma_{s,w_s}) D^{-1}_{t-1} w + \rho w^\top G_t w} \right)$$

$$\leq 2\sqrt{2\beta(t)} \sum_{t=d^2+1}^{T} \left( w^\top D^{-1}_{t-1}(\lambda \Lambda_{\Sigma_t} D_{t-1} + \sum_{s=1}^{t-1} \Sigma_{s,w_s}) D^{-1}_{t-1} w + 2\rho \sum_{t=d^2+1}^{T} w^\top G_t w \right)$$

$$\leq 2\sqrt{2\beta(t)} \sum_{t=d^2+1}^{T} \left( w^\top D^{-1}_{t-1}(\lambda \Lambda_{\Sigma_t} D_{t-1} + \sum_{s=1}^{t-1} \Sigma_{s,w_s}) D^{-1}_{t-1} w + 2\rho \sum_{t=d^2+1}^{T} w^\top G_t w \right)$$

$$+ 2\rho \sum_{t=d^2+1}^{T} w^\top G_t w$$

$$\leq 2\sqrt{2\beta(t)} \sqrt{T \left( \sum_{t=d^2+1}^{T} (w^\top D^{-1}_{t-1} \Lambda_{\Sigma_t} w) + \sum_{t=d^2+1}^{T} (w^\top D^{-1}_{t-1} \sum_{s=1}^{t-1} \Sigma_{s,w_s} D^{-1}_{t-1} w) \right) + 2\rho \sum_{t=d^2+1}^{T} w^\top G_t w}$$

$$\leq 2\sqrt{2\beta(t)} \sqrt{T \left( \sum_{t=d^2+1}^{T} (w^\top D^{-1}_{t-1} \Lambda_{\Sigma_t} w) + \sum_{t=d^2+1}^{T} (w^\top D^{-1}_{t-1} \sum_{s=1}^{t-1} \Sigma_{s,w_s} D^{-1}_{t-1} w) \right) + 2\rho \sum_{t=d^2+1}^{T} w^\top G_t w}$$

$$+ 2\rho \sum_{t=d^2+1}^{T} w^\top G_t w$$

27
Let \((\Sigma^*_t)^+ = \Sigma^*_t \vee 0\) for any \(i, j, \in [d]\). We first address \(\Gamma_3\). Since for any \(t \geq d^2 + 1\) and \(i, j \in [d]\),

\[
g_{ij}(t) = \frac{3 \ln t}{N_{ij}(t-1)} \sqrt{\frac{3 \ln t}{N_{ij}(t-1)}} + \frac{48 \ln^2 t}{N_{ij}(t-1)N_i(t-1)} + \sqrt{36 \ln^2 t}\]

\[
\leq 48 \frac{\ln t}{\sqrt{N_{ij}(t-1)}} + \frac{48 \ln^2 t}{N_{ij}(t-1)} + \frac{36 \ln^2 t}{N_{ij}(t-1)}
\]

\[
\leq 61 \frac{\ln t}{\sqrt{N_{ij}(t-1)}},
\]

we can bound \(\Gamma_3\) as follows

\[
\Gamma_3 = \sum_{t=d^2+1}^{T} w^T G_t w
\]

\[
= \sum_{t=d^2+1}^{T} \sum_{i,j \in [d]} g_{ij}(t) w_{t,i} w_{t,j}
\]

\[
\leq 61 \sum_{i,j \in [d]} \sum_{t=d^2+1}^{T} \frac{\ln t}{\sqrt{N_{ij}(t-1)}} w_{t,i} w_{t,j}
\]

\[
\leq 61 \ln T \sum_{i,j \in [d]} \sum_{t=d^2+1}^{T} \frac{w_{t,i} w_{t,j}}{\sqrt{\sum_{s=1}^{t-1} w_{s,i} w_{s,j}}}
\]

\[
\leq 61 \ln T \sum_{i,j \in [d]} \sum_{t=d^2+1}^{T} w_{t,i} w_{t,j} \sqrt{\frac{1}{\sum_{s=1}^{t-1} w_{s,i} w_{s,j}}}
\]

\[
\leq 122 \ln T \sum_{i,j \in [d]} \sqrt{\sum_{t=1}^{T} w_{t,i} w_{t,j}}
\]

\[
\leq 122 \ln T \sqrt{d^2 \sum_{i,j \in [d]} \sum_{t=1}^{T} w_{t,i} w_{t,j}}
\]

\[
= 122 \ln T \sqrt{\sum_{t=1}^{T} \sum_{i,j \in [d]} w_{t,i} w_{t,j}}
\]

\[
= 122 \ln T \sqrt{\sum_{t=1}^{T} \left( \sum_{i \in [d]} w_{t,i} \right)^2}
\]

\[
\leq 122 \ln T \sqrt{T}
\]

Next, we obtain a bound for \(\Gamma_1\).

\[
\Gamma_1 = \sum_{t=d^2+1}^{T} (w^T D_{t-1}^{-1} \Lambda_{\xi t} w)
\]
Finally, we bound $\Gamma_2$.

$$
\Gamma_2
= \sum_{t=d^2+1}^{T} \left( w_i^T D_{t-1}^{-1} \left( \sum_{s=1}^{t-1} \Sigma_{s,iw_s} \right) D_{t-1}^{-1} w_j \right)
= \sum_{i,j \in [d]} \sum_{t=d^2+1}^{T} \frac{\sum_{s=1}^{t-1} \Sigma_{s,ij} \mathbb{I}\{w_{s,i}, w_{s,j} > 0\}}{N_i(t-1)N_j(t-1)} w_{t,i}w_{t,j}
\leq \sum_{i,j \in [d]} \sum_{t=d^2+1}^{T} \frac{\sum_{s=1}^{t-1} \Sigma_{s,ij} \mathbb{I}\{w_{s,i}, w_{s,j} > 0\}}{N_{ij}^2(t-1)} w_{t,i}w_{t,j}
\leq \sum_{t=d^2+1}^{T} \sum_{i,j \in [d]} \frac{\sum_{s=1}^{t-1} \left( \Sigma_{ij}^* + 2g_{ij}(s) \right) \mathbb{I}\{w_{s,i}, w_{s,j} > 0\}}{N_{ij}^2(t-1)} w_{t,i}w_{t,j}
= \sum_{i,j \in [d]} \left( \sum_{t=d^2+1}^{T} \frac{\sum_{s=1}^{t-1} \mathbb{I}\{w_{s,i}, w_{s,j} > 0\}}{N_{ij}^2(t-1)} w_{t,i}w_{t,j} + \sum_{t=d^2+1}^{T} \frac{\mathbb{I}\{w_{s,i}, w_{s,j} > 0\}}{N_{ij}^2(t-1)} w_{t,i}w_{t,j} \right)
\leq \sum_{i,j \in [d]} \left( \Sigma_{ij}^* \sum_{t=d^2+1}^{T} \frac{1}{N_{ij}(t-1)} w_{t,i}w_{t,j} + \sum_{t=d^2+1}^{T} \frac{\mathbb{I}\{w_{s,i}, w_{s,j} > 0\}}{N_{ij}^2(t-1)} w_{t,i}w_{t,j} \right)
$$
Recall that $\beta(\delta_T) = \ln(T \ln^2 T) + d \ln \ln T + \frac{d}{2} \ln(1 + e/\lambda) = O(\ln T + d \ln \ln T + d \ln(1 + \lambda^{-1}))$. Combining the bounds of $\Gamma_1, \Gamma_2$ and $\Gamma_3$, we obtain

$$
\sum_{t=d^2+1}^{T} \left( (f(w^*) - f(w_t)) \right)
\leq 2\sqrt{2\beta(\delta_T)} \sqrt{T} \left[ \lambda \sum_{t=d^2+1}^{T} \left( w^\top D_{t-1}^{-1} \Lambda \Sigma w \right) + \sum_{t=d^2+1}^{T} \left( w^\top D_{t-1}^{-1} \sum_{s=1}^{t-1} \Sigma_{s,w} D_{t-1}^{-1} w \right) \right] + 2\rho \sum_{t=d^2+1}^{T} w^\top G_t w

\leq 2\sqrt{2\beta(\delta_T)} \sqrt{T} \left[ \lambda \sum_{i \in [d]} (\Sigma_{i}^*)^+ \ln T + 244 \ln T \frac{1}{\sqrt{\sum_{t=1}^{d^2} w_{t,i}}} \right] + 2\rho \sum_{i \in [d]} (\Sigma_{i}^*)^+ \ln T + 488 \ln T \frac{1}{\sqrt{\sum_{t=1}^{d^2} w_{t,i}}} )

\leq 2\sqrt{2\beta(\delta_T)} \sqrt{T} \left[ \lambda \ln T \sum_{i \in [d]} (\Sigma_{i}^*)^+ + \ln T \sum_{i,j \in [d]} (\Sigma_{i}^*)^+ \right] + 244 \rho d \ln T \sqrt{T}

\leq 2\sqrt{2\beta(\delta_T)} \sqrt{T} \left[ \lambda \ln T \sum_{i \in [d]} (\Sigma_{i}^*)^+ + \ln T \sum_{i,j \in [d]} (\Sigma_{i}^*)^+ + (244 \lambda + 488) d^2 \ln T + 244 \rho d \ln T \sqrt{T} \right]

= O \left( \sqrt{(\ln T + d \ln \ln T + d \ln(1 + \lambda^{-1}))} \cdot T \left( \lambda \ln T \sum_{i \in [d]} (\Sigma_{i}^*)^+ + \ln T \sum_{i,j \in [d]} (\Sigma_{i}^*)^+ + (\lambda + 1) d^2 \ln T + \rho d \ln T \sqrt{T} \right) \right) \cdot \sqrt{T} \left( \lambda \ln T \sum_{i \in [d]} (\Sigma_{i}^*)^+ + \ln T \sum_{i,j \in [d]} (\Sigma_{i}^*)^+ + (\lambda + 1) d^2 \ln T + \rho d \ln T \sqrt{T} \right)
where

According to Lemmas 1 and 2, for any $t$ \( \geq 2 \), we obtain the regret upper bound

\[
\mathbb{E}[\mathcal{R}(T)] = O(\Delta_{\text{max}}) + \sum_{t=2}^{T} O(\Delta_{\text{max}} \cdot \Pr[\neg (G_t \cap H_t)] + \Delta_t \cdot I\{G_t \cap H_t\})
\]

\[
= O(\Delta_{\text{max}}) + \sum_{t=2}^{T} O(\Delta_{\text{max}} \cdot \frac{d^2}{t \ln^2 t}) + O\left(\sqrt{d \ln T (\ln T + \ln(1 + \lambda^{-1})) T} \right.
\]

\[
= O\left(\sqrt{d \ln T (\ln T + \ln(1 + \lambda^{-1})) T} \sqrt{\lambda \sum_{i \in [d]} (\Sigma_{ii}^*)^+ + \sum_{i,j \in [d]} (\Sigma_{ij}^*)^+ + (\lambda + 1)d^2 + \rho d \ln T \sqrt{T}}\right)
\]

\[
= O\left(\sqrt{d (\ln(1 + \lambda^{-1}) + 1) \ln^2 T \cdot T} \sqrt{(\lambda + 1) \left(\sum_{i,j \in [d]} (\Sigma_{ij}^*)^+ + d^2\right) + \rho d \ln T \sqrt{T}}\right)
\]

\[
= O\left(\sqrt{\left(L(\lambda)(\|\Sigma^*\|_+ + d^2)d \ln^2 T \cdot T + \rho d \ln T \sqrt{T}\right)}\right)
\]

where $L(\lambda) = (\lambda + 1) (\ln(1 + \lambda^{-1}) + 1)$ and $\|\Sigma^*\|_+ = \sum_{i,j \in [d]} (\Sigma_{ij}^* \vee 0)$ for any $i, j \in [d]$. \hfill \Box

### C.2 Proof of Theorem 4

**Proof.** First, we construct some instances with $d \geq 4$, $\frac{2}{3} \leq c \leq \frac{1}{2}$, $\Sigma_* = I$ and $\theta_t \sim N(\theta^*, I)$.

Let $I_j$ be a random instance constructed as follows: we uniformly choose a dimension $J$ from $[d]$, and the expected reward vector $\theta^*_j$ has $\frac{1}{2} + \varepsilon$ on its $J$-th entry and $\frac{1}{2}$ elsewhere, where $\varepsilon \in (0, \frac{2}{3}]$ will be specified later. Let $I_u$ be a uniform instance,
where $\theta^*_i$ has all its entries to be $\frac{1}{2}$. Let $Pr_f[\cdot]$ and $Pr_r[\cdot]$ denote the probabilities under instances $I_f$ and $I_r$, respectively, and let $Pr_f[\cdot] = Pr_f[\cdot | J = j]$. Analogously, $E_f[\cdot]$, $E_r[\cdot]$ and $E_f[\cdot] = E_f[\cdot | J = j]$ denote the expectation operations.

Fix an algorithm $\mathcal{A}$. Let $S_t \in \{\mathbb{R} \cup \{\perp\}\}^d$ be a random variable vector denoting the observations at timestep $t$, obtained by running $\mathcal{A}$. Here $\perp$ denotes no observation on this dimension. Let $Q_\perp$ denote the distribution on support $\{\perp\}$ which takes value $\perp$ with probability 1.

In CMCB-SB, if $w_{t,i} > 0$, we can observe the reward on the $i$-th dimension, i.e., $S_{t,i} = \theta_{t,i}$; otherwise, if $w_{t,i} = 0$, we cannot get observation on the $i$-th dimension, i.e., $S_{t,i} = \perp$. Let $D_j$ be the distribution of observation sequence $S_1, \ldots, S_t$ under instance $I_t$, and $D_j = D_{j,j-1}$ is the distribution conditioned on $J = j$. Let $D_u$ be the distribution of observation sequence $S_1, \ldots, S_t$ under instance $I_u$. For any $i \in [d]$, let $N_i = \sum_{t=1}^T I\{w_{t,i} > 0\}$ be the number of pulls that has a positive weight on the $i$-th dimension, i.e., the number of observations on the $i$-th dimension.

Following the analysis procedure of Lemma A.1 in Auer et al. [2002b], we have

$$KL(D_j \| D_u) = \sum_{t=1}^T KL(D_u[S_t|S_1, \ldots, S_{t-1}]||D_j[S_t|S_1, \ldots, S_{t-1}])$$

$$= \sum_{t=1}^T \sum_{i=1}^d \left(Pr[w_{t,i} > 0] \cdot KL(N(\theta^*_i, 1)||N(\theta^*_i, 1)) + Pr[w_{t,i} = 0] \cdot KL(Q_\perp || Q_\perp)\right)$$

$$= \sum_{t=1}^T \left(Pr[w_{t,j} > 0] \cdot KL\left(N\left(\frac{1}{2}, 1\right)||N\left(\frac{1}{2} + \varepsilon, 1\right)\right) + \sum_{i \neq j}^d Pr[w_{t,i} > 0] \cdot KL\left(N\left(\frac{1}{2}, 1\right)||N\left(\frac{1}{2}, 1\right)\right)\right)$$

$$= \frac{1}{2} \varepsilon^2 \cdot \sum_{t=1}^T Pr[w_{t,j} > 0]$$

$$= \frac{1}{2} \varepsilon^2 E_u[N_j]$$

Here the first equality comes from the chain rule of entropy Cover [1999]. The second equality is due to that given $S_1, \ldots, S_{t-1}$, if $w_{t,i} > 0$, the conditional distribution of $S_i$ is $N(\theta^*_i, 1)$, where "," refers to the subscript of instances; otherwise, if $w_{t,i} = 0$, $S_i$ is $\perp$ deterministically. The third equality is due to that $\theta^*_i$ and $\theta^*_j$ only have one different entry on the $j$-th dimension.

Let $\| \cdot \|$ with subscript $TV$ denote the total variance distance, and $KL(\cdot || \cdot)$ denote the Kullback–Leibler divergence. Using Eq. (28) in the analysis of Lemma A.1 in Auer et al. [2002b] and Pinsker’s inequality, we have

$$E_j[N_j] \leq E_u[N_j] + T\|D_j - D_u\|_{TV}$$

$$\leq E_u[N_j] + T \sqrt{\frac{1}{2} KL(D_j || D_u)}$$

$$= E_u[N_j] + \frac{T \varepsilon}{2} \sqrt{E_u[N_j]}$$

Let $m = \lfloor \frac{d}{\varepsilon} \rfloor \leq \frac{d}{\varepsilon}$ denote the maximum number of positive entries for a feasible action, i.e., the maximum number of observations for a pull. Performing the above argument for all $j \in [d]$ and using $\sum_{j \in [d]} E_u[N_j] \leq mT$, we have

$$\sum_{j \in [d]} E_j[N_j] \leq \sum_{j \in [d]} E_u[N_j] + \frac{T \varepsilon}{2} \sum_{j \in [d]} \sqrt{E_u[N_j]}$$

32
\[
\leq mT + \frac{T\varepsilon}{2} \sqrt{d \sum_{j \in [d]} E_u[N_j]}
\]

and thus

\[
E_J[N_J] = \frac{1}{d} \sum_{j \in [d]} E_j[N_j] \leq \frac{mT}{d} + \frac{T\varepsilon}{2} \sqrt{\frac{mT}{d}}
\]

Letting \( \rho \leq \frac{\varepsilon}{2(1-c)} \), the expected reward (linear) term dominates \( f(w) \), and the best action \( w^* \) under \( I_j \) has the weight 1 on the \( J \)-th entry and 0 elsewhere.

Recall that \( m \leq \frac{1}{\varepsilon} \). For each pull that has no weight on the \( J \)-th entry, algorithm \( A \) must suffer a regret at least

\[
\left( \frac{1}{2} + \varepsilon - \rho \right) - \left( \frac{1}{2} - \rho \cdot \frac{1}{m} \right)
\geq \varepsilon - \frac{m-1}{m} \rho
\geq \varepsilon - \frac{m-1}{m} \cdot \frac{\varepsilon}{2(1-c)}
\geq \varepsilon - \frac{m-1}{m} \cdot \frac{\varepsilon}{2(1 - \frac{1}{m})}
= \frac{\varepsilon}{2}
\]

Thus, the regret is lower bounded by

\[
E[R(T)] \geq (T - E_J[N_J]) \cdot \frac{\varepsilon}{2}
\geq \left( T - \frac{mT}{d} - \frac{T\varepsilon}{2} \sqrt{\frac{mT}{d}} \right) \cdot \frac{\varepsilon}{2}
= \Omega \left( T\varepsilon - T\varepsilon^2 \sqrt{\frac{mT}{d}} \right),
\]

where the last equality is due to \( m \leq \frac{d}{T} \).

Letting \( \varepsilon = a_0 \sqrt{\frac{d}{Tm}} \) for small enough constant \( a_0 \), we obtain the regret lower bound \( \Omega(\sqrt{dT}) = \Omega(\sqrt{cdT}) \).

D Proof for CMCB-FB

D.1 Proof of Theorem 5

In order to prove Theorem 5, we first prove Lemmas 5 and 6, which give the concentrations of covariance and means for CMCB-FB, using different techniques than those for CMCB-SB (Lemmas 1 and 2) and CMCB-FI (Lemmas 3 and 4).
Lemma 5 (Concentration of Covariance for CMCB-FB). Consider the CMCB-FB problem and algorithm MC-ETE (Algorithm 3). For any $t > 0$, the event

$$\mathcal{M}_t \triangleq \left\{ |\Sigma^*_{ij} - \Sigma_{ij,t-1}| \leq 5\| C^+_\pi \| \sqrt{\frac{3 \ln t}{2N_\pi(t)}} \right\}$$

satisfies

$$\Pr[\mathcal{M}_t] \geq 1 - \frac{6d^2}{t^2},$$

where $\| C^+_\pi \| \triangleq \max_{i \in [d]} \left\{ \sum_{j \in [d]} |C^+_{\pi,ij}| \right\}$.

Proof. Let $\sigma = (\Sigma^*_{11}, \ldots, \Sigma^*_{dd}, \Sigma^*_{12}, \ldots, \Sigma^*_{1d}, \Sigma^*_{23}, \ldots, \Sigma^*_{d-1,d})^T \in \mathbb{R}^d$ denote the column vector that stacks the $\tilde{d}$ distinct entries in the covariance matrix $\Sigma^*$. Recall that the $d \times \tilde{d}$ matrix

$$C_\pi = \begin{bmatrix}
    w_{1,1}^2 & \cdots & w_{1,d}^2 & 2w_{1,1}w_{1,2} & 2w_{1,1}w_{1,3} & \cdots & 2w_{1,1}w_{1,d} \\
    w_{2,1}^2 & \cdots & w_{2,d}^2 & 2w_{2,1}w_{2,2} & 2w_{2,1}w_{2,3} & \cdots & 2w_{2,1}w_{2,d} \\
    \vdots & & \vdots & \vdots & \vdots & \cdots & \vdots \\
    w_{d,1}^2 & \cdots & w_{d,d}^2 & 2w_{d,1}w_{d,2} & 2w_{d,1}w_{d,3} & \cdots & 2w_{d,1}w_{d,d}
\end{bmatrix},$$

where $w_{i,j}$ denotes the $j$-th entry of portfolio vector $w_i$ in design set $W$. We use $C_{\pi,k}$ to denote the $k$-th row in matrix $C_\pi$.

We recall the feedback structure in algorithm MC-ETE as follows. At each timestep $t$, the learner plays an action $w_t \in \Delta_d$, and observes the full-bandit feedback $y_t = w_t^\top \theta_t$ with $\mathbb{E}[y_t] = w_t^\top \theta^*$ and $\text{Var}[y_t] = w_t^\top \Sigma w_t = \sum_{i \in [d]} w_{t,i}^2 \Sigma^*_{ii} + \sum_{i,j \in [d], i < j} 2w_{t,i}w_{t,j} \Sigma^*_{ij}$. Then, during exploration round $s$, where each action in design set $\pi = \{v_1, \ldots, v_d\}$ is pulled once, the full-bandit feedback $y_t$ has mean $y(\pi) \triangleq (v_1^\top \theta^*, \ldots, v_d^\top \theta^*)^T$ and variance $z(\pi) \triangleq (v_1^\top \Sigma^* v_1, \ldots, v_d^\top \Sigma^* v_d)^T = (C_\pi^T \sigma, \ldots, C_{\pi,d}^T \sigma)^T = C_\pi \sigma$. For any $t > 0$, denote $\tilde{y}_t$ the empirical mean of $y(\pi) \triangleq (v_1^\top \theta^*, \ldots, v_d^\top \theta^*)^T$ and $\tilde{z}_t$ the empirical variance of $z(\pi) \triangleq (C_\pi^T \sigma, \ldots, C_{\pi,d}^T \sigma)^T$.

Using the Chernoff-Hoeffding inequality for empirical variances (Lemma 1 in Sani et al. [2012]), we have that for any $t > 0$ and $k \in \tilde{d}$, with probability at least $1 - \frac{6d^2}{t^2}$,

$$|\tilde{z}_{t-1,k} - z_k| = |\tilde{z}_{t-1,k} - C_\pi^T \sigma| \leq 5 \sqrt{\frac{3 \ln t}{2N_\pi(t-1)}}.$$

Let $C_{\pi,i}$ denote the $i$-th row of matrix $C_\pi$ and $C_{\pi,ik}$ denote the $ik$-th entry of matrix $C_\pi$. Since $C_\pi^T C_\pi = I$, for any $i \in [\tilde{d}]$, we have

$$|\tilde{\sigma}_{t-1,i} - \sigma_i| = \left| (C_{\pi,i}^+)^T \tilde{z}_t - (C_{\pi,i}^+)^T C_\pi \sigma \right| \leq \sum_{k \in [d]} |C_{\pi,ik}^+| |\tilde{z}_{t-1,k} - C_{\pi,k} \sigma| \leq 5 \sum_{k \in [d]} |C_{\pi,ik}^+| \sqrt{\frac{3 \ln t}{2N_\pi(t)}}.$$
In addition, since $\sigma = (\Sigma^*_1, \ldots, \Sigma^*_{d,d}, \Sigma^*_1, \ldots, \Sigma^*_{d,d-1})^\top \in \mathbb{R}^d$ is the column vector stacking the distinct entries in $\Sigma^*$, we obtain the lemma.

**Lemma 6** (Concentration of Means for CMCB-FB), Consider the CMCB-FB problem and algorithm MC-ETE (Algorithm 3). Let $\delta_t > 0$, $\lambda > 0$, $\beta(\delta_t) = \ln(1/\delta_t) + \ln \ln t + \frac{d}{2} \ln(1 + e/\lambda)$ and $E_t(w) = \sqrt{\beta(\delta_t)} \sqrt{w^\top B^+_\pi D_{t-1}^\top (\lambda \Lambda \Sigma^*_t D_{t-1} + \sum_{s=1}^t \Sigma^*_s) D_{t-1}^{-1} (B^+_\pi)^\top w}$, where $\Sigma^*_t = \text{diag}(v^*_1, \Sigma^*_v v_1, \ldots, v^*_d, \Sigma^*_v v_d)$. Then, the event $N_t = \{||w^\top \theta^* - w^\top \hat{\theta}_{t-1}|| \leq E_t(w), \forall w \in D\}$ satisfies $\Pr[N_t] \geq 1 - \delta_t$.

**Proof.** Recall that the $d \times d$ matrix $B_\pi = [v_1^\top; \ldots; v^*_d]$, and $B^+_\pi$ is the Moore–Penrose pseudoinverse of $B_\pi$. Since $B_\pi$ is of full column rank, $B^+_\pi B_\pi = I$.

We recall the feedback structure in algorithm MC-ETE as follows. At each timestep $t$, the learner plays an action $w_t \in \triangle_d$, and observes the full-bandit feedback $y_t = w_t^\top \theta_t$ such that $\mathbb{E}[y_t] = w^\top \theta^*$ and $\text{Var}[y_t] = w^\top \Sigma^* w = \sum_{i,j \in [d]} w_i^2 v^*_j \Sigma^*_i,j + \sum_{i,j \in [d], i < j} 2 w_i w_j v^*_i v^*_j$. Then, during exploration round $s$, where each action in design set $\pi = \{v_1, \ldots, v_d\}$ is pulled once, the full-bandit feedback $y_s$ has mean $y(\pi) \triangleq (v^*_1, \ldots, v^*_d)^\top$ and variance $z(\pi) \triangleq (v^*_1, \Sigma^*_v v_1, \ldots, v^*_d, \Sigma^*_v v_d)^\top = (C^T_{\pi,1} \sigma, \ldots, C^T_{\pi,d} \sigma)^\top = C_{\pi} \sigma$. For any $t > 0$, $\hat{y}_t$ is the empirical mean of $y(\pi) \triangleq (v^*_1, \ldots, v^*_d)^\top$ and $\hat{z}_t$ is the empirical variance of $z(\pi) \triangleq (C^T_{\pi,1} \sigma, \ldots, C^T_{\pi,d} \sigma)^\top$.

Let $D_t$ be the $d \times d$ diagonal matrix such that $D_{t,ii} = N_\pi(t)$ for any $i \in [d]$ and $t > 0$. Let $\Sigma^*_t$ be a $d \times d$ diagonal matrix such that $\Sigma^*_{t,ii} = v^*_i \Sigma^*_v w_i$ for any $i \in [d]$, and thus $\Lambda \Sigma^*_t = \Sigma^*_t$. Let $\epsilon_t$ be the vector such that $\epsilon_t = (\Sigma^*)^{1/2} \epsilon_t$ for any timestep $t > 0$. Let $\zeta_s = (v_1^\top (\Sigma^*)^{1/2} \epsilon_{s,1}, \ldots, v^*_d (\Sigma^*)^{1/2} \epsilon_{s,d})^\top$, where $(\Sigma^*)^{1/2} \epsilon_{s,k}$ denotes the noise of the $k$-th sample in the $s$-th exploration round.

Note that, the following analysis of $\|w^\top (\theta^* - \hat{\theta}_{t-1})\|$ and the constructions (definitions) of the noise $\zeta_s$, matrices $S_t$, $V_t$ and super-martingale $M_t$ are different from those in CMCB-SB (Lemmas 1 and 2) and CMCB-FI (Lemmas 3 and 4).

For any $w \in \triangle_d$, we have
\[
\begin{align*}
\|w^\top (\theta^* - \hat{\theta}_{t-1})\| & = |w^\top (B^+_{\pi} B_{\pi} \theta^* - B^+_{\pi} \hat{y}_{t-1})| \\
& = |w^\top B^+_{\pi} (y(\pi) - \hat{y}_{t-1})| \\
& = -w^\top B^+_{\pi} D_{t-1}^{-1} \sum_{s=1}^{N_\pi(t-1)} \zeta_s \\
& \leq \sqrt{w^\top B^+_{\pi} D_{t-1}^{-1} (D + \sum_{s=1}^{N_\pi(t-1)} \Sigma^*_s)^{1/2} (D + \sum_{s=1}^{N_\pi(t-1)} \Sigma^*_s)^{-1/2} \sum_{s=1}^{N_\pi(t-1)} \zeta_s} \\
& \leq \sqrt{3 \ln t / 2N_\pi(t)}
\end{align*}
\]
Let \( S_t = \sum_{s=1}^{N_{s}(t-1)} \zeta_s \), \( V_t = \sum_{s=1}^{N_{s}(t-1)} \Sigma^*_s \) and \( I_{D+V_t} = \frac{1}{2} \| S_t \|^2_{(D+V_t)^{-1}} \). Then, we have

\[
\| \sum_{s=1}^{N_{s}(t-1)} \zeta_s \|_{(D+\sum_{s=1}^{N_{s}(t-1)} \Sigma^*_s)^{-1}} = \| S_t \|_{(D+V_t)^{-1}} = \sqrt{2I_{D+V_t}}.
\]

Since \( D \leq \lambda \Sigma^*_s D_{t-1} \), we get

\[
|w^\top (\theta^* - \hat{\theta}_{t-1})| \leq \sqrt{2\beta(\delta_t)} \sqrt{w^\top B^+_s D_{t-1}^{-1} (\lambda \Sigma^*_s D_{t-1} + \sum_{s=1}^{N_s(t-1)} \Sigma^*_s) D_{t-1}^{-1} (B^+_s)^\top w}.
\]

Thus,

\[
\Pr \left[ |w^\top (\theta^* - \hat{\theta}_{t-1})| > \sqrt{2\beta(\delta_t)} \sqrt{w^\top B^+_s D_{t-1}^{-1} (\lambda \Sigma^*_s D_{t-1} + \sum_{s=1}^{N_s(t-1)} \Sigma^*_s) D_{t-1}^{-1} (B^+_s)^\top w} \right]
\leq \Pr \left[ \sqrt{w^\top B^+_s D_{t-1}^{-1} (\lambda \Sigma^*_s D_{t-1} + \sum_{s=1}^{N_s(t-1)} \Sigma^*_s) D_{t-1}^{-1} (B^+_s)^\top w} > \sqrt{2\beta(\delta_t)} \sqrt{w^\top B^+_s D_{t-1}^{-1} (\lambda \Sigma^*_s D_{t-1} + \sum_{s=1}^{N_s(t-1)} \Sigma^*_s) D_{t-1}^{-1} (B^+_s)^\top w} \right]
= \Pr \left[ I_{D+V_t} > \beta(\delta_t) \right]
\]

Hence, to prove Eq. (4), it suffices to prove

\[
\Pr \left[ I_{D+V_t} > \beta(\delta_t) \right] \leq \delta_t.
\] (18)

To prove Eq. (18), we introduce some notions. Let \( u \in \mathbb{R}^d \) be a multivariate Gaussian random variable with mean \( \theta \) and covariance \( D^{-1} \), which is independent of all the other random variables and we use \( \varphi(u) \) denote its probability density function. Let

\[
P^u_s = \exp \left( u^\top \zeta_s - \frac{1}{2} u^\top \Sigma^*_s u \right),
\]

\[
M^u_t \triangleq \exp \left( u^\top S_t - \frac{1}{2} \| u \|^2_{V_t} \right),
\]

and

\[
M_t \triangleq E_u[M^u_t] = \int_{\mathbb{R}^d} \exp \left( u^\top S_t - \frac{1}{2} \| u \|^2_{V_t} \right) \varphi(u) du,
\]

36
where \( s = 1, \ldots, N(t-1) \) is the index of exploration round. We have \( M_t^u = \Pi_{s=1}^{N(t-1)} P_s^u \). In the following, we prove \( \mathbb{E}[M_t] \leq 1 \).

For any timestep \( t \), \( \eta_t = (\Sigma^*)^{-\frac{1}{2}} \varepsilon_t \) is \( \Sigma^* \)-sub-Gaussian and \( \eta_t \) is independent among different timestep \( t \). Then, \( \zeta_s = (v_1^T (\Sigma^*)^{-\frac{1}{2}} \varepsilon_{s,1}, \ldots, v_d^T (\Sigma^*)^{-\frac{1}{2}} \varepsilon_{s,d})^T \) is \( \Sigma_s^* \)-sub-Gaussian. According to the sub-Gaussian property, \( \zeta_s \) satisfies

\[
\forall \mathbf{v} \in \mathbb{R}^d, \mathbb{E} \left[ e^{\mathbf{v}^T \zeta_s} \right] \leq e^{\frac{1}{2} \mathbf{v}^T \Sigma_s^* \mathbf{v}},
\]

which is equivalent to

\[
\forall \mathbf{v} \in \mathbb{R}^d, \mathbb{E} \left[ e^{\mathbf{v}^T \zeta_s - \frac{1}{2} \mathbf{v}^T \Sigma_s^* \mathbf{v}} \right] \leq 1.
\]

Let \( \mathcal{J}_s \) be the \( \sigma \)-algebra \( \sigma(W, \zeta_1, \ldots, W, \zeta_{s-1}, \pi) \). Thus, we have

\[
\mathbb{E} [P_s^u | \mathcal{J}_s] = \mathbb{E} \left[ \exp \left( \mathbf{u}^T \zeta_s - \frac{1}{2} \mathbf{u}^T \Sigma_s^* \mathbf{u} \right) \mathbf{u}^T \right] \leq 1.
\]

Then, we can obtain

\[
\mathbb{E}[M_t^u | \mathcal{J}_{N(t-1)}] = \mathbb{E} \left[ \Pi_{s=1}^{N(t-1)} P_s^u | \mathcal{J}_{N(t-1)} \right] = \left( \Pi_{s=1}^{N(t-2)} P_s^u \right) \mathbb{E} \left[ P_s^u | \mathcal{J}_{N(t-1)} \right] \leq M_{t-1}^u,
\]

which implies that \( M_t^u \) is a super-martingale and \( \mathbb{E}[M_t^u | \mathbf{u}] \leq 1 \). Thus,

\[
\mathbb{E}[M_t] = \mathbb{E}_u[\mathbb{E}[M_t^u | \mathbf{u}]] \leq 1.
\]

According to Lemma 9 in Abbasi-Yadkori et al. [2011], we have

\[
M_t \triangleq \int_{\mathbb{R}^d} \exp \left( \mathbf{u}^T S_t - \frac{1}{2} \| \mathbf{u} \|^2_{V_t} \right) \varphi(\mathbf{u}) d\mathbf{u} = \sqrt{\frac{\det D}{\det(D + V_t)}} \exp \left( I_{D+V_t} \right).
\]

Thus,

\[
\mathbb{E} \left[ \sqrt{\frac{\det D}{\det(D + V_t)}} \exp \left( I_{D+V_t} \right) \right] \leq 1.
\]

Now we prove Eq. (18). First, we have

\[
\text{Pr} \left[ I_{D+V_t} > \beta(\delta_t) \right] = \text{Pr} \left[ \sqrt{\frac{\det D}{\det(D + V_t)}} \exp \left( I_{D+V_t} \right) > \sqrt{\frac{\det D}{\det(D + V_t)}} \exp \left( \beta(\delta_t) \right) \right] = \text{Pr} \left[ M_t > \frac{1}{\sqrt{\det(I + D^{-\frac{1}{2}} V_t D^{-\frac{1}{2}})}} \exp \left( \beta(\delta_t) \right) \right]
\]

\[
\leq \frac{\mathbb{E}[M_t]}{\sqrt{\det(I + D^{-\frac{1}{2}} V_t D^{-\frac{1}{2}})}} \exp \left( \beta(\delta_t) \right)
\]

\[
\leq \sqrt{\frac{\det(I + D^{-\frac{1}{2}} V_t D^{-\frac{1}{2}})}{\exp(\beta(\delta_t))}}
\]

(19)
Then, for some constant $\gamma > 0$ and for any $a \in \mathbb{N}$, we define the set of timesteps $\mathcal{K}_a \subseteq [T]$ such that

$$t \in \mathcal{K}_a \iff (1 + \gamma)^a \leq N_\pi(t - 1) < (1 + \gamma)^{a+1}.$$  

Define $D_a$ as a diagonal matrix such that $D_{a,ii} = (1 + \gamma)^a$, $\forall i \in \tilde{d}$. Suppose $t \in \mathcal{K}_a$ for some fixed $a$. Then, we have

$$\frac{1}{1 + \gamma} D_t \preceq D_a \preceq D_t.$$  

Let $D = \lambda \Sigma_x D_a \succeq \frac{\lambda}{1 + \gamma} \Sigma_x D_t$. Then, we have

$$D^{-\frac{1}{2}} V_t D^{-\frac{1}{2}} \preceq \frac{1 + \gamma}{\lambda} D_t^{-\frac{1}{2}} \Sigma_x^{-\frac{1}{2}} V_t \Sigma_x^{-\frac{1}{2}} D_t^{-\frac{1}{2}},$$

where matrix $D_t^{-\frac{1}{2}} \Sigma_x^{-\frac{1}{2}} V_t \Sigma_x^{-\frac{1}{2}} D_t^{-\frac{1}{2}}$ has $\tilde{d}$ ones on the diagonal. Since the determinant of a positive definite matrix is smaller than the product of its diagonal terms, we have

$$\det(I + D^{-\frac{1}{2}} V_t D^{-\frac{1}{2}}) \leq \det(I + \frac{1 + \gamma}{\lambda} D_t^{-\frac{1}{2}} \Sigma_x^{-\frac{1}{2}} V_t \Sigma_x^{-\frac{1}{2}} D_t^{-\frac{1}{2}}) \leq \left(1 + \frac{1 + \gamma}{\lambda}\right)^{\tilde{d}}.$$  

Let $\gamma = e - 1$. Using Eqs. (19) and (20), $\beta(\delta_t) = \ln(1/\delta_t) + \ln t + \frac{\tilde{d}}{2} \ln(1 + \xi/\lambda) = \ln(t \ln t) + \ln t + \frac{\tilde{d}}{2} \ln(1 + \xi)$ and a union bound over $a$, we have

$$\Pr \left[ I_{D + V_t} > \beta(\delta_t) \right] \leq \sum_a \Pr \left[ I_{D + V_t} > \beta(\delta_t) \mid t \in \mathcal{K}_a, D = \lambda \Sigma_x D_a \right] \leq \sum_a \sqrt{\frac{\det(I + D^{-\frac{1}{2}} V_t D^{-\frac{1}{2}})}{\exp(\beta(\delta_t))}} \leq \frac{\ln t}{\ln(1 + \gamma)} \cdot \exp \left(\frac{(1 + \frac{1 + \gamma}{\lambda})^{\frac{\tilde{d}}{2}}}{\ln(t \ln t) + \ln t + \frac{\tilde{d}}{2} \ln(1 + \xi)}\right) = \frac{\ln t}{\ln(1 + \gamma)} \cdot \exp \left(\frac{(1 + \frac{\xi}{\lambda})^{\frac{\tilde{d}}{2}}}{\ln(1 + \xi)}\right) = \frac{1}{\ln^2 t} \cdot \ln t \cdot (1 + \frac{\xi}{\lambda})^{\frac{\tilde{d}}{2}} = \frac{1}{\delta_t}.$$  

Thus, Eq. (18) holds and we complete the proof of Lemma 2. $\square$

Now, we give the proof of Theorem 5.

Proof. (Theorem 5) First, we bound the number of exploration rounds up to time $T$. Let $\psi(t) = t^{\frac{\tilde{d}}{2}}/d$. According to the condition of exploitation (Line 4 in Algorithm 3), we have that if at timestep $t$ algorithm MC-ETE starts an exploration round, then $t$ satisfies

$$N_\pi(t - 1) \leq \psi(t).$$
Let \( t_0 \) denote the timestep at which algorithm \( \text{MC-ETE} \) starts the last exploration round. Then, we have
\[
N_\pi(t_0 - 1) \leq \psi(t_0)
\]
and thus
\[
N_\pi(T) = N_\pi(t_0) = N_\pi(t_0 - 1) + 1 \\
\leq \psi(t_0) + 1 \\
\leq \psi(T) + 1.
\]

Next, for each timestep \( t \), we bound the estimation error of \( f(w) \). For any \( t > 0 \), let \( \Delta_t \triangleq f(w^*) - f(w_t) \) and \( \hat{f}_t(w) \triangleq w^\top \hat{\theta}_t - \rho w^\top \Sigma w \). Suppose that event \( \mathcal{M}_t \cap \mathcal{N}_t \) occurs. Then, according to Lemmas 5 and 6, we have that for any \( w \in \triangle_d \),
\[
\left| f(w) - \hat{f}_{t-1}(w) \right| \leq \sqrt{2\beta(\delta_t)} \left( w^\top B^+_\pi D^{-1}_{t-1} \left( \lambda \Lambda \Sigma^*_\pi D_{t-1} + \sum_{s=1}^{N_\pi(t-1)} \Sigma^*_{\pi} \right) D^{-1}_{t-1} (B^+_\pi) \right)^\top w \\
+ 5\rho \|C^+_\pi\| \sqrt{\frac{3 \ln t}{2N_\pi(t-1)}}
\leq \sqrt{2 \ln(t \ln^2 t) + \ln \ln t + \frac{\tilde{d}}{2} \ln(1 + \frac{e}{\lambda})} \cdot \sqrt{w^\top B^+_\pi D^{-1}_{t-1} (\lambda \Lambda \Sigma^*_\pi + \Sigma^*_\pi) (B^+_\pi) \right)^\top w \\
+ 5\rho \|C^+_\pi\| \sqrt{\frac{3 \ln t}{2N_\pi(t-1)}}
\leq 7 \sqrt{\ln t + \frac{\tilde{d}}{2} \ln(1 + \frac{e}{\lambda})} \cdot \left( \sqrt{w^\top B^+_\pi (\lambda \Lambda \Sigma^*_\pi + \Sigma^*_\pi) (B^+_\pi) \right)^\top w + \rho \|C^+_\pi\| \right) \cdot \sqrt{\frac{1}{N_\pi(t-1)}}
\]

Let \( Z(\rho, \pi) \triangleq \max_{w \in \triangle_d} \left( \sqrt{w^\top B^+_\pi (\lambda \Lambda \Sigma^*_\pi + \Sigma^*_\pi) (B^+_\pi) \right)^\top w + \rho \|C^+_\pi\| \right) \). Then, we have
\[
\left| f(w) - \hat{f}_{t-1}(w) \right| \leq 7 \cdot Z(\rho, \pi) \sqrt{\frac{\ln t + \tilde{d} \ln(1 + e/\lambda)}{N_\pi(t-1)}}.
\]

Let \( \mathcal{L}^{\text{exploit}}(t) \) denote the event that algorithm \( \text{MC-ETE} \) does the exploitation at timestep \( t \). For any \( t > 0 \), if \( \mathcal{L}^{\text{exploit}}(t) \) occurs, we have \( N_\pi(t-1) > \psi(t) \). Thus
\[
\left| f(w) - \hat{f}_{t-1}(w) \right| \leq 7 \cdot Z(\rho, \pi) \sqrt{\frac{\ln t + \tilde{d} \ln(1 + e/\lambda)}{\psi(t)}} \cdot \frac{\psi(t)}{N_\pi(t-1)}
\leq 7 \cdot Z(\rho, \pi) \sqrt{\frac{d \left( \ln t + \tilde{d} \ln(1 + e/\lambda) \right)}{t}} \cdot t^{-\frac{1}{2}}
\]

According to Lemmas 5.6, for any \( t \geq 2 \), we bound the probability of event \( \neg(\mathcal{M}_t \cap \mathcal{N}_t) \) by
\[
\Pr \left[ \neg(\mathcal{M}_t \cap \mathcal{N}_t) \right] \leq \frac{6d^2}{t^2} + \frac{1}{t \ln^2 t}
\]

39
The expected regret of algorithm $\mathcal{MC}$-ETE can be divided into two parts, one due to exploration and the other due to exploitation. Then, we can obtain

$$E[R(T)] \leq N_\pi(T) \cdot \tilde{d} \Delta_{\max} + \sum_{t=1}^{T} E[\Delta_t | L^{\text{exploit}}(t)]$$

$$\leq (\psi(T) + 1) \cdot \tilde{d} \Delta_{\max} + \sum_{t=1}^{T} \left( E[\Delta_t | L^{\text{exploit}}(t), M_t \cap N_t] + E[\Delta_t | L^{\text{exploit}}(t), \neg(M_t \cap N_t)] \cdot \Pr[\neg(M_t \cap N_t)] \right)$$

$$= O \left( \left( \frac{T^\frac{3}{d} + 1}{d} \right) \cdot \tilde{d} \Delta_{\max} + \sum_{t=1}^{T} \left( Z(\rho, \pi) \sqrt{d(\ln t + \tilde{d} \ln(1 + \rho/\lambda)) \cdot t^{-\frac{1}{2}} + \Delta_{\max} \cdot \frac{d^2}{t \ln^2 t}} \right) \right)$$

$$= O \left( T^{\frac{3}{2}} d \Delta_{\max} + d^2 \Delta_{\max} + Z(\rho, \pi) \sqrt{d(\ln T + d^2 \ln(1 + \lambda^{-1})) \cdot T^{\frac{3}{2}} + d^2 \Delta_{\max}} \right)$$

$$= O \left( Z(\rho, \pi) \sqrt{d(\ln T + d^2 \ln(1 + \lambda^{-1})) \cdot T^{\frac{3}{2}} + d \Delta_{\max} \cdot T^{\frac{3}{2}}} \right)$$

Choosing $\lambda = \frac{1}{2}$ and using $\Lambda \Sigma^* = \Sigma^*_\lambda$, we obtain

$$E[R(T)] = O \left( Z(\rho, \pi) \sqrt{d(\ln T + d^2) \cdot T^{\frac{3}{2}} + d \Delta_{\max} \cdot T^{\frac{3}{2}}} \right),$$

where $Z(\rho, \pi) = \max_{w \in \Delta_d} \left( \sqrt{w^\top B^*_\pi \Sigma^*_\lambda (B^*_\pi)^\top w + \rho \| C^*_\pi \|} \right)$. \qed