Robust Contextual Portfolio Optimization with Gaussian Mixture Models

Yijie Wang[1], Grani A. Hanasusanto[1], and Chin Pang Ho[2]

[1] *Graduate Program in Operations Research and Industrial Engineering, The University of Texas at Austin,* {yijie-wang,grani.hanasusanto}@utexas.edu

[2] *School of Data Science, City University of Hong Kong,* clint.ho@cityu.edu.hk

Abstract

We consider the portfolio optimization problem with contextual information that is available to better quantify and predict the uncertain returns of assets. Motivated by the regime modeling techniques for the finance market, we consider the setting where both the uncertain returns and the contextual information follow a Gaussian Mixture (GM) distribution. This problem is shown to be equivalent to a nominal portfolio optimization problem where the means and the covariance matrix are adjusted by the contextual information. We then apply robust optimization and propose the robust contextual portfolio optimization problem, which reduces the sensitivity of model parameters used in the Gaussian Mixture Model (GMM). A tractable formulation is derived to approximate the solution of the robust contextual portfolio optimization problem. We conduct a numerical experiment in the US equity markets, and the results demonstrate the advantage of our proposed model against other benchmark methods.

Keywords: Portfolio optimization, Contextual optimization, Robust optimization

1 Introduction

The portfolio optimization problem, one of the most important problems in computational finance, has been studied by the research community and the industry for many years (Benati and Rizzi, 2007; Birge, 2007; DeMiguel et al., 2009a; Konno and Yamazaki, 1991; Markowitz, 1952; Perold, 1984; Rockafellar et al., 2000). Classical models for the portfolio optimization problem, such as

the Markowitz model (Markowitz, 1952), assume that exact information of the considered assets (e.g., means and variances of the returns) is available. This modeling requirement brings profound difficulty in practice because small estimation errors would result in poor decisions (Britten-Jones, 1999; Chopra and Ziemba, 2013). To achieve satisfactory performance, one would need to control and reduce the estimation error with an unrealistically large amount of data. Moreover, the return of an asset is often affected by various external factors that are usually not considered in the portfolio optimization model, such as economic situation, governmental policy, business cycle, etc. (Eugene and French, 1992; Flannery and Protopapadakis, 2002; Grinblatt et al., 1995). These factors are typically not considered in classical models. As a result, the unconditional distribution of returns does not well present the actual uncertainty of return in any given investment period.

In recent years, robust optimization has been a popular approach to address the aforementioned issue of the sensitivity of model parameters (Ben-Tal et al., 2006, 2009; Bertsimas and Sim, 2004; Pflug and Wozabal, 2007). In contrast to the classical approach, the robust optimization scheme constructs an uncertainty set of asset returns, and optimizes for the best portfolio allocation in view of the worst-case returns from within the set (DeMiguel and Nogales, 2009; Fabozzi et al., 2007; Goldfarb and Iyengar, 2003; Kakouris and Rustem, 2014; Rujeerapaiboon et al., 2016; Ye et al., 2012; Zymler et al., 2011). For readers interested in the robust portfolio selection problem, we refer to a recent comprehensive review by Ghahtarani et al. (2022). There is also a rich literature on distributionally robust portfolio optimization, with diverse choices of distributional ambiguity set, e.g., moment-based (Zhu and Fukushima, 2009; Nguyen et al., 2021a), event-wise (Chen et al., 2020), optimal-transportation (Blanchet et al., 2022) and Wasserstein ambiguity sets (Pflug and Wozabal, 2007; Blanchet et al., 2021).

Even though these approaches can provide reliable performance in practice, the decisions are often conservative. As mentioned above, since many external factors could influence the model parameters, the uncertainty set would have to be unnecessarily large in order to contain the underlying model parameter scenarios in different contexts. Therefore, the output decision would be overly conservative, as it safely anticipates a wide range of asset returns.

To overcome this challenge, we propose incorporating contextual information into the portfolio optimization problem in this paper. Contextual decision making under uncertainty has been a popular topic in recent years (Athey et al., 2019; Ban and Rudin, 2019; Bertsimas and McCord,

2019; Bertsimas and Kallus, 2020; Kallus and Mao, 2022; Kannan et al., 2020a; Sen and Deng, 2018) as it takes into consideration the interrelationships between the external uncertain factors (that are not included in the optimization model) and the uncertain parameters of the optimization model. By considering the external factors, commonly referred to as side information (Srivastava et al., 2021) or contextual information (Pagnoncelli et al., 2022), the decision-maker is able to faithfully adapt her optimization model to the given context. This approach allows the decision-maker to better quantify the uncertain model parameters with additional information that are available before making her decisions (Kannan et al., 2020b; Nguyen et al., 2021b). While there exist various non-parametric approaches to estimate the interrelationships between the side information and the uncertain parameters (Nguyen et al., 2021b; Srivastava et al., 2021), they may suffer from high variance in estimation compared to the parametric approaches, which is often preferable when one has domain knowledge about the distribution of the uncertain parameters.

The Gaussian Mixture Model (GMM) is a powerful and flexible parametric modeling framework for financial markets as it can adequately approximate skewed return distributions and elegantly model different market regimes (Akgiray and Booth, 1987; Arditti, 1967; Ball and Torous, 1983; Seyfi et al., 2021). There is a long history of studies observing that the underlying distribution of asset returns is skewed (Beedles, 1986; Fabozzi et al., 2005; Neuberger, 2012; Popova et al., 2007) and has a heavier tail than that of a Gaussian distribution (Cont, 2001; Fama, 1965; Praetz and Wilson, 1978; Zi-Yi, 2017). Thus, GMM emerges as a natural choice for bridging the gap between these empirical observations and the classical portfolio optimization models. Indeed, when the number of mixtures equals one, the GMM reduces to the classical model where asset returns follow a Gaussian distribution (Markowitz, 1952); on the other hand, when the number of mixtures coincides with the number of samples, the model is equivalent to the Kernel Density Estimation (KDE) method with Gaussian kernels (Epanechnikov, 1969; Liu et al., 2022; Silverman, 2018). For the real-world financial markets, Kon (1984) observes that a handful number of Gaussian mixtures can accurately approximate asset returns. The GMM has also drawn significant attention due to its straightforward interpretation of market regimes. Indeed, asset returns behave differently under different market regimes (Ang and Bekaert, 2004). For example, Ang and Bekaert (2002) and Campbell et al. (2002) observe that the correlations among asset returns increase during bear markets. In this case, the GMM can immediately build such regime-dependent structures by fitting different regimes with

different clusters (Gupta and Dhingra, 2012; Rydén et al., 1998). Botte and Bao (2021) are the first to incorporate side information with asset returns into the GMM. Intuitively, equity markets have the tendency to change their behavior over time as the macroeconomic shifts, which creates regimes. Their empirical review demonstrates that the equity markets returns, along with several economic indices such as interest rate, inflation rate, etc., can be modeled into several clusters using the GMM. They also apply economic research to the fitted model and find that each cluster aptly represents a specific regime in history, e.g., prosperity, crisis, and inflation. However, although this empirical study inspires the use of GMM with side information to estimate the current regime, it remains unclear how one could integrate this approach into portfolio optimization.

Motivated by this empirical observation, we propose a new robust contextual portfolio optimization problem. Our model assumes that both the uncertain returns and side information follow a Gaussian Mixture (GM) distribution, and it leverages robust optimization to reduce the sensitivity of model parameters in the fitted Gaussian Mixture model to provide reliable decisions. We remark that the same modeling assumption in which the uncertain returns follow a GM distribution has been studied by Buckley et al. (2008), where the authors consider a two-component GMM and analyze several objectives such as the Markowitz mean-variance, the Sharpe ratio, and an exponential utility. Hentati-Kaffel and Prigent (2014) study the optimal portfolio under arbitrary utility functions. The numerical experiment on historical data suggests that the GMM model leads to significantly different portfolios compared with those obtained from a Gaussian return model. Robust portfolio optimization with two-component GMMs is studied by Gambacciani and Paolella (2017), who propose an approach for estimating asset returns using a fast new variation of the minimum covariance determinant (MCD) method. Arabacı and Kocuk (2020) derive formulations for the robust portfolio optimization problems under the assumption that the stock returns follow a two-component GM distribution. Shi and Kim (2021) explore different coherent risk measures and show that the mean-risk portfolio optimization problem with GMMs admits a closed-form solution by fixing the location and skewness parameters. Recently, Luxenberg and Boyd (2022) investigate the portfolio optimization problem with exponential utility under the GM return assumption. They show that the problem admits a convex reformulation and can be solved efficiently using the off-the-shelf solvers. However, none of these papers have considered exploiting contextual information and using robust optimization under the generic GMM setting to improve the performance of portfolios.

We summarize the main contributions of the paper:

1. We consider the contextual portfolio optimization problem, which assumes both the uncertain returns and side information follows a GM distribution. We show that this problem can be reformulated as a classical portfolio optimization problem where the mean vector and the covariance matrix are adjusted according to the value of side information.

2. We devise a robust counterpart to the original problem to mitigate the unfavorable effect of parameter estimation errors in the Gaussian Mixture Model. We show that under certain assumptions, the solution to the robust contextual portfolio optimization problem offers attractive out-of-sample performance. While the problem is computationally challenging, we derive a tractable conservative approximation in second-order cone programming, which can be solved efficiently using the off-the-shelf solvers.

3. To demonstrate the practical viability of our proposed method, we numerically examine the performance of the Robust Contextual Gaussian Mixture model. Compared with the benchmark methods, our proposed approach achieves a higher average return and a better annualized Sharpe ratio in the out-of-sample test.

The remainder of the paper is organized as follows. Section 2 proposes the contextual portfolio optimization problem where the uncertain returns and the side information follow a GM distribution. Section 3 develops the robust counterpart of the contextual portfolio optimization problem and derives its tractable conservative approximation. We conduct experiments in Section 4 and provide some concluding remarks in Section 5. Some technical proofs are deferred to the appendix.

Notation We use bold lowercase and uppercase letters for a vector and a matrix, respectively. All Random variables are designated by a tilde sign (e.g., $\tilde{\boldsymbol{\xi}}$), while their realizations are denoted without tildes (e.g., $\boldsymbol{\xi}$). The set of all positive definite and positive semidefinite matrices in $\mathbb{R}^{n\times n}$ are denoted as $\mathbb{S}_{++}^n$ and $\mathbb{S}_{+}^n$, respectively. The probability simplex in $\mathbb{R}_{+}^K$ is denoted as Δ_K. Unless otherwise specified, $\|\boldsymbol{A}\|$ denotes the spectral norm of matrix $\boldsymbol{A}$, and $\|\boldsymbol{v}\|$ denotes the Euclidean norm of vector $\boldsymbol{v}$.

2 Contextual Portfolio Optimization under Gaussian Mixtures Model

We define $\tilde{\boldsymbol{r}} \in \mathbb{R}^n$ to be the random returns of n assets in a specific period and $\tilde{\boldsymbol{s}} \in \mathbb{R}^d$ to be the signals or side information observed at the beginning of the period. In this paper, we are interested in solving the contextual portfolio optimization problem

$$\min_{\boldsymbol{w} \in \Delta_n} \mathbb{E}_{\mathbb{G}}[-\tilde{\boldsymbol{r}}^\top \boldsymbol{w} \mid \tilde{\boldsymbol{s}} = \boldsymbol{s}] + \eta \mathbb{V}_{\mathbb{G}}[\tilde{\boldsymbol{r}}^\top \boldsymbol{w} \mid \tilde{\boldsymbol{s}} = \boldsymbol{s}], \tag{1}$$

where the decision variables $\boldsymbol{w} \in \mathbb{R}^n$ correspond to the allocations to the considered assets. Intuitively, problem (1) aims to maximize the conditional expectation of portfolio returns while ensuring the portfolio risk, captured by the conditional variance, is small. The parameter $\eta \in \mathbb{R}_+$ controls the level of risk aversion of the decision maker, and the subscript $\mathbb{G}$ signifies that $(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{s}})$ jointly follows a Gaussian Mixtures (GM) distribution $\mathbb{G}$ with K components. That is,

$$(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{s}}) \sim \mathbb{G}\left(\{\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k, p^k\}_{k=1}^K\right),$$

where $\boldsymbol{\mu}^k = (\boldsymbol{\mu}_r^k, \boldsymbol{\mu}_s^k) \in \mathbb{R}^{n+d}$ denotes the mean vector of the k-th component of GMM. Here, the subscripts r and s are used to indicate that $\boldsymbol{\mu}_r^k$ and $\boldsymbol{\mu}_s^k$ are the mean vectors of $\tilde{\boldsymbol{r}}$ and $\tilde{\boldsymbol{s}}$ in the k-th component, respectively; we adopt these subscripts for the other variables in similar manner. The covariance matrix of the k-th component is denoted by

$$\boldsymbol{\Sigma}^k = \begin{bmatrix} \boldsymbol{\Sigma}_{rr}^k & \boldsymbol{\Sigma}_{rs}^k \\ \boldsymbol{\Sigma}_{sr}^k & \boldsymbol{\Sigma}_{ss}^k \end{bmatrix} \in \mathbb{S}_{++}^{(n+d)\times(n+d)}.$$

The mixture weights are represented by $\boldsymbol{p} \in \Delta_K$ where $p^k \in [0,1]$ represents the weight of the k-th component. We further define the k-th precision matrix, which is the inverse of the k-th covariance matrix, as

$$\boldsymbol{\Psi}^k := (\boldsymbol{\Sigma}^k)^{-1} = \begin{bmatrix} \boldsymbol{\Psi}_{rr}^k & \boldsymbol{\Psi}_{rs}^k \\ \boldsymbol{\Psi}_{sr}^k & \boldsymbol{\Psi}_{ss}^k \end{bmatrix} \in \mathbb{S}_{++}^{(n+d)\times(n+d)}.$$

Given that the random vector $(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{s}})$ follows a GM distribution, the following lemma shows that the conditional distribution of $\tilde{\boldsymbol{r}}$ given $\tilde{\boldsymbol{s}} = \boldsymbol{s}$ is also a GM distribution.

Lemma 1 (Conditional Gaussian Mixture Distribution)**.** *Consider a random vector $(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{s}}) \in \mathbb{R}^{n+d}$ governed by the GM distribution $\mathbb{G}(\{\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k, p^k\}_{k=1}^K)$ for some $\{\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k, p^k\}_{k=1}^K$. Conditioned on*

$\tilde{\boldsymbol{s}} = \boldsymbol{s}$, *we have*

$$\tilde{\boldsymbol{r}} \sim \mathbb{G}\left(\left\{\boldsymbol{\mu}_{r|s}^k, \boldsymbol{\Sigma}_{r|s}^k, p_{r|s}^k\right\}_{k=1}^K\right),$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{r|s}^k &= \boldsymbol{\mu}_r^k + \boldsymbol{\Sigma}_{rs}^k(\boldsymbol{\Sigma}_{ss}^k)^{-1}(\boldsymbol{s} - \boldsymbol{\mu}_s^k), \\
\boldsymbol{\Sigma}_{r|s}^k &= \boldsymbol{\Sigma}_{rr}^k - \boldsymbol{\Sigma}_{rs}^k(\boldsymbol{\Sigma}_{ss}^k)^{-1}\boldsymbol{\Sigma}_{sr}^k = (\boldsymbol{\Psi}_{rr}^k)^{-1}, \text{ and} \\
p_{r|s}^k &= \frac{p^k \mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}_s^k, \boldsymbol{\Sigma}_{ss}^k\right)}{\sum_{j=1}^K p^j \mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}_s^j, \boldsymbol{\Sigma}_{ss}^j\right)}.
\end{aligned} \tag{2}$$

Proof. We first consider the case where the random vectors $\tilde{\boldsymbol{r}}$ and $\tilde{\boldsymbol{s}}$ are jointly Gaussian (i.e., $K = 1$) with the density function $\mathcal{N}((\boldsymbol{r}, \boldsymbol{s})|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The marginal distributions of $\tilde{\boldsymbol{r}}$ and $\tilde{\boldsymbol{s}}$ are $\mathcal{N}(\boldsymbol{r}|\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_{rr})$ and $\mathcal{N}(\boldsymbol{s}|\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_{ss})$, respectively. The density function of the conditional distribution of $\tilde{\boldsymbol{r}}$ given $\boldsymbol{s}$ is (Bishop and Nasrabadi, 2006, Section 2.3.2, Equations (2.94)-(2.98))

$$p(\boldsymbol{r}|\boldsymbol{s}) = \mathcal{N}\left(\boldsymbol{r}|\boldsymbol{\mu}_{r|s}, \boldsymbol{\Sigma}_{r|s}\right),$$

where $\boldsymbol{\mu}_{r|s} = \boldsymbol{\mu}_r + \boldsymbol{\Sigma}_{rs}(\boldsymbol{\Sigma}_{ss})^{-1}(\boldsymbol{s} - \boldsymbol{\mu}_s)$ and $\boldsymbol{\Sigma}_{r|s} = \boldsymbol{\Sigma}_{rr} - \boldsymbol{\Sigma}_{rs}(\boldsymbol{\Sigma}_{ss})^{-1}\boldsymbol{\Sigma}_{sr}$.

We now extend this result to the case of GMM, where the marginal distribution of $\tilde{\boldsymbol{s}}$ is

$$p(\boldsymbol{s}) = \sum_{k=1}^K p^k \mathcal{N}(\boldsymbol{s}|\boldsymbol{\mu}_s^k, \boldsymbol{\Sigma}_{ss}^k).$$

Here, the conditional density function becomes

$$p(\boldsymbol{r}|\boldsymbol{s}) = \frac{p(\boldsymbol{r}, \boldsymbol{s})}{p(\boldsymbol{s})} = \sum_{k=1}^K \frac{p^k \mathcal{N}\left((\boldsymbol{r}, \boldsymbol{s})|\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k\right)}{\sum_{j=1}^K p^j \mathcal{N}(\boldsymbol{s}|\boldsymbol{\mu}_s^j, \boldsymbol{\Sigma}_{ss}^j)} = \sum_{k=1}^K \frac{p^k \mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}_s^k, \boldsymbol{\Sigma}_{ss}^k\right)}{\sum_{j=1}^K p^j \mathcal{N}(\boldsymbol{s}|\boldsymbol{\mu}_s^j, \boldsymbol{\Sigma}_{ss}^j)} \mathcal{N}\left(\boldsymbol{r} \mid \boldsymbol{\mu}_{r|s}^k, \boldsymbol{\Sigma}_{r|s}^k\right).$$

Thus, this is a GMM with components $\mathcal{N}(\boldsymbol{r} \mid \boldsymbol{\mu}_{r|s}^k, \boldsymbol{\Sigma}_{r|s}^k)$, $k \in [K]$, and mixture probabilities

$$\sum_{k=1}^K \frac{p^k \mathcal{N}\left(\boldsymbol{s}|\boldsymbol{\mu}_s^k, \boldsymbol{\Sigma}_{ss}^k\right)}{\sum_{j=1}^K p^j \mathcal{N}(\boldsymbol{s}|\boldsymbol{\mu}_s^j, \boldsymbol{\Sigma}_{ss}^j)} \quad \forall k \in [K],$$

which completes the proof. □

From the above lemma, we know that the conditional distribution of the vector $\tilde{\boldsymbol{r}}$ also follows a GM distribution, given any observed contextual information $\boldsymbol{s}$. This leads to our first main result.

Theorem 1. *Let the conditional parameters $p_{r|s}^k$, $\boldsymbol{\mu}_{r|s}^k$ and $\boldsymbol{\Sigma}_{r|s}^k$ be defined in* (2)*, and let $\boldsymbol{\Omega}_{r|s}^k = \left(\boldsymbol{\Sigma}_{r|s}^k + (\boldsymbol{\mu}_{r|s}^k - \boldsymbol{\mu}_{r|s})(\boldsymbol{\mu}_{r|s}^k - \boldsymbol{\mu}_{r|s})^\top\right)$. Given $\tilde{\boldsymbol{s}} = \boldsymbol{s}$, the conditional mean vector and covariance matrix of $\tilde{\boldsymbol{r}}$ are*

$$\boldsymbol{\mu}_{r|s} = \sum_{k=1}^K p_{r|s}^k \boldsymbol{\mu}_{r|s}^k \text{ and } \boldsymbol{\Omega}_{r|s} = \sum_{k=1}^K p_{r|s}^k \boldsymbol{\Omega}_{r|s}^k,$$

respectively. Hence, the contextual portfolio optimization problem (1) *is equivalent to the quadratic program*

$$\min_{\boldsymbol{w}\in\Delta_n} \sum_{k=1}^{K} p_{r|s}^k \left(-\boldsymbol{w}^\top \boldsymbol{\mu}_{r|s}^k + \eta \cdot \boldsymbol{w}^\top \boldsymbol{\Omega}_{r|s}^k \boldsymbol{w} \right). \tag{3}$$

Proof. The mean portfolio return in (1) can be written as

$$\mathbb{E}_{\mathbb{G}}[\tilde{\boldsymbol{r}}^\top \boldsymbol{w} \mid \tilde{\boldsymbol{s}} = \boldsymbol{s}] = \sum_{k=1}^{K} p_{r|s}^k \boldsymbol{w}^\top \boldsymbol{\mu}_{r|s}^k.$$

Meanwhile, the variance term can be reformulated as

$$\begin{aligned}
\mathbb{V}_{\mathbb{G}}[\tilde{\boldsymbol{r}}^\top \boldsymbol{w} \mid \tilde{\boldsymbol{s}} = \boldsymbol{s}] &= \mathbb{E}_{\mathbb{G}}[(\tilde{\boldsymbol{r}}^\top \boldsymbol{w})^2 \mid \tilde{\boldsymbol{s}} = \boldsymbol{s}] - \mathbb{E}_{\mathbb{G}}[\tilde{\boldsymbol{r}}^\top \boldsymbol{w} \mid \tilde{\boldsymbol{s}} = \boldsymbol{s}]^2 \\
&= \boldsymbol{w}^\top \mathbb{E}_{\mathbb{G}}[\tilde{\boldsymbol{r}}\tilde{\boldsymbol{r}}^\top \mid \tilde{\boldsymbol{s}} = \boldsymbol{s}]\boldsymbol{w} - \boldsymbol{w}^\top \boldsymbol{\mu}_{r|s} \boldsymbol{\mu}_{r|s}^\top \boldsymbol{w} \\
&= \boldsymbol{w}^\top \left(\sum_{k=1}^{K} p_{r|s}^k \left(\boldsymbol{\Sigma}_{r|s}^k + \boldsymbol{\mu}_{r|s}^k {\boldsymbol{\mu}_{r|s}^k}^\top \right) \right) \boldsymbol{w} - \boldsymbol{w}^\top \boldsymbol{\mu}_{r|s} \boldsymbol{\mu}_{r|s}^\top \boldsymbol{w} \\
&= \boldsymbol{w}^\top \left(\sum_{k=1}^{K} p_{r|s}^k \left(\boldsymbol{\Sigma}_{r|s}^k + (\boldsymbol{\mu}_{r|s}^k - \boldsymbol{\mu}_{r|s})(\boldsymbol{\mu}_{r|s}^k - \boldsymbol{\mu}_{r|s})^\top \right) \right) \boldsymbol{w} \\
&= \boldsymbol{w}^\top \boldsymbol{\Omega}_{r|s} \boldsymbol{w},
\end{aligned}$$

where the fourth equality holds since $\boldsymbol{\mu}_{r|s} = \sum_{k=1}^{K} p_{r|s}^k \boldsymbol{\mu}_{r|s}^k$ and the fifth equality holds because $\boldsymbol{\Omega}_{r|s} = \sum_{k=1}^{K} p_{r|s}^k \boldsymbol{\Omega}_{r|s}^k$. Thus, the claim follows. □

Given perfect information on the means and covariances of the asset returns, the contextual portfolio optimization problem constitutes a tractable convex quadratic optimization problem. However, in practice, the exact values of the underlying GM distribution parameters are not available to the portfolio manager and typically have to be estimated using the empirical-based GMM learning algorithms. While the empirical-based estimators may work well on the training dataset, they often fail to achieve an acceptable out-of-sample performance as they do not carefully consider the possible estimation errors from the learning algorithm. In the next section, we propose a robust counterpart of the contextual portfolio optimization problem that mitigates the adverse effect of the estimation errors and produces reliable decisions.

3 Robust Contextual Portfolio Optimization

In the empirical risk minimization (ERM) setting, decision makers naively adopt the empirical estimators $\hat{p}^k$, $\hat{\boldsymbol{\mu}}^k$, and $\hat{\boldsymbol{\Sigma}}^k$ from the GM learning algorithm to compute the empirical conditional means $\hat{\boldsymbol{\mu}}^k_{r|s}$, covariances $\hat{\boldsymbol{\Sigma}}^k_{r|s}$ and probabilities $\hat{p}^k_{r|s}$, $\forall k \in [K]$. Then those empirical conditional estimations are plugged into (3), which yields the empirical GMM portfolio optimization problem

$$\min_{\boldsymbol{w}\in\Delta_n} \sum_{k=1}^{K} p^k_{r|s} \left(-\boldsymbol{w}^\top \hat{\boldsymbol{\mu}}^k_{r|s} + \eta \cdot \boldsymbol{w}^\top \hat{\boldsymbol{\Omega}}^k_{r|s} \boldsymbol{w} \right), \tag{4}$$

where $\hat{\boldsymbol{\mu}}_{r|s} = \sum_{k=1}^{K} \hat{p}^k_{r|s} \hat{\boldsymbol{\mu}}^k_{r|s}$ and $\hat{\boldsymbol{\Omega}}^k_{r|s} = \hat{\boldsymbol{\Sigma}}^k_{r|s} + (\hat{\boldsymbol{\mu}}^k_{r|s} - \hat{\boldsymbol{\mu}}_{r|s})(\hat{\boldsymbol{\mu}}^k_{r|s} - \hat{\boldsymbol{\mu}}_{r|s})^\top$. Though the ERM method is easy to implement, it suffers from the notorious overfitting issue and may incur extremely poor performance in the out-of-sample test. In this paper, we address the unfavorable data overfitting effects using the idea of Robust Optimization (RO). In contrast to the ERM scheme, the RO approach does not impose the exact specifications of mean vectors, covariance matrices, or mixture probabilities of the GMM. Instead, it considers an uncertainty set $\mathcal{Y}$ that contains all plausible parameter estimations consistent with the historical observations, with the goal of obtaining an optimal portfolio strategy that minimizes the worst-case mean-variance objective function. In particular, we consider the robust counterpart of (3) given by

$$\min_{\boldsymbol{w}\in\Delta_n} \sup_{\{p^k,\boldsymbol{\mu}^k,\boldsymbol{\Sigma}^k\}_{k=1}^K \in \mathcal{Y}} \sum_{k=1}^{K} p^k_{r|s} \left(-\boldsymbol{w}^\top \boldsymbol{\mu}^k_{r|s} + \eta \cdot \boldsymbol{w}^\top \boldsymbol{\Omega}^k_{r|s} \boldsymbol{w} \right), \tag{5}$$

where

$$\mathcal{Y} \coloneqq \left\{ \{p^k, \boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k\}_{k=1}^K : \begin{array}{ll} |p^k - \hat{p}^k| \le \epsilon_p,\ \|\boldsymbol{\mu}^k - \hat{\boldsymbol{\mu}}^k\| \le \epsilon_\mu,\ \|\boldsymbol{\Sigma}^k - \hat{\boldsymbol{\Sigma}}^k\| \le \epsilon_\Sigma & \forall k \in [K] \\ \boldsymbol{p} \in \Delta_K,\ \boldsymbol{\mu}^k \in \mathbb{R}^{n+d},\ \boldsymbol{\Sigma}^k \in \mathbb{S}_+^{(n+d)\times(n+d)} & \forall k \in [K] \end{array} \right\}.$$

Therefore, the model is immunized against detrimental estimation errors of the model parameters in the nominal problem (3). In this paper, we assume that there exists an algorithm that can compute the radii of the norm balls in $\mathcal{Y}$ so that the unknown true parameters $\{\mu^{k\star}, \Sigma^{k\star}, p^{k\star}\}_{k=1}^K$ reside in $\mathcal{Y}$ with high probability.

Assumption 1. *Given N samples drawn i.i.d. from the true GM distribution $\mathbb{G}^\star \left(\{\boldsymbol{\mu}^{k\star}, \boldsymbol{\Sigma}^{k\star}, p^{k\star}\}_{k=1}^K \right)$, there exists an algorithm that outputs an estimation $\hat{\mathbb{G}} \left(\{\hat{\boldsymbol{\mu}}^k, \hat{\boldsymbol{\Sigma}}^k, \hat{p}^k\}_{k=1}^K \right)$ satisfying $|p^{k\star} - \hat{p}^k| \le \epsilon_p$, $\|\boldsymbol{\mu}^{k\star} - \hat{\boldsymbol{\mu}}^k\|_2 \le \epsilon_\mu$ and $\|\boldsymbol{\Sigma}^{k\star} - \hat{\boldsymbol{\Sigma}}^k\| \le \epsilon_\Sigma$ with probability $1-\delta$, where the radii ϵ_p, ϵ_μ, and ϵ_Σ and the tolerance δ may depend on the GMM parameters and the number of samples. In addition, for any*

component in the true GM distribution, the smallest eigenvalue of its covariance matrix is bounded below by a positive constant $\underline{\alpha}$*, i.e.,* $\underline{\alpha} \preceq \mathbf{\Sigma}^{k\star}$ $\forall k \in [K]$.

Since the paper focuses on tractable formulations of robust portfolio optimization with GMM rather than its statistical performance guarantees, the technical detail behind this assumption is beyond our scope. Nevertheless, we highlight several relevant results as follows. For mixtures $\Omega(\sqrt{\log K})$-separated spherical Gaussians, it is shown in Kwon and Caramanis (2020) that with proper initialization and $N \geq \tilde{O}((n+d)/\epsilon^2)$, the Expectation Maximization (EM) algorithm converges in $T = O(\log(1/\epsilon))$ iterations, where at the T-th iteration the estimates $\hat{p}^k$, $\hat{\mu}^k$, $(\hat{\sigma}^k)^2\mathbb{I}$, are accurate to within $\epsilon_p = \max_{k\in[K]} p^{k\star}\epsilon$, $\epsilon_\mu = \max_{k\in[K]} \sigma^{k\star}\epsilon$, $\epsilon_\Sigma = (\max_{k\in[K]} \sigma^{k\star})^2\epsilon/\sqrt{n+d}$, respectively, with a tolerance level δ that depends polynomially in T, $n+d$, and K. Note that instead of requiring $N \geq \tilde{O}((n+d)/\epsilon^2)$, we can decrease ϵ with N while keeping the confidence level $1-\delta$. When there are only $K = 2$ components, Hardt and Price (2015) show there exists an algorithm with polynomial sample complexity that learns arbitrary mixtures of Gaussians without any separation condition. Otherwise, the best known results for learning general mixtures of K Gaussians with polynomial sample complexity is derived in Sanjeev and Kannan (2001) with $\tilde{\Omega}((n+d)^{1/4})$ separation condition and in Kannan et al. (2005); Achlioptas and McSherry (2005) with $\tilde{\Omega}((\text{poly}(K))$ separation condition.

While problem (5) is an intuitive model that would provide a robust allocation for the contextual portfolio optimization problem, solving (5) is computationally challenging. The following lemma and theorem shows that one can compute an upper bound of (5) tractably using second-order cone programming.

Lemma 2 (Upper bounds)**.** *Consider* $\{\boldsymbol{\mu}^k_{r|s}, \mathbf{\Sigma}^k_{r|s}, \mathbf{\Omega}^k_{r|s}\}_{k=1}^K$ *and* $\{\hat{\boldsymbol{\mu}}^k_{r|s}, \hat{\mathbf{\Sigma}}^k_{r|s}, \hat{\mathbf{\Omega}}^k_{r|s}\}_{k=1}^K$ *defined in Theorem 1 and equation* (4)*. Define* $\alpha_k = \max\{\underline{\alpha}, \lambda(\hat{\mathbf{\Sigma}}^k)_{\min} - \epsilon_\Sigma\}$*,* $\beta_k = \|\hat{\mathbf{\Sigma}}^k\| + \epsilon_\Sigma$*, and* $\gamma_k = \|\boldsymbol{s} - \hat{\boldsymbol{\mu}}^k_s\| + \epsilon_\mu$*, for every* $k \in [K]$*. We have, for any* $\{p^k, \boldsymbol{\mu}^k, \mathbf{\Sigma}^k\}_{k=1}^K \in \mathcal{Y}$*,*

$$\|\boldsymbol{\mu}^k_{r|s} - \hat{\boldsymbol{\mu}}^k_{r|s}\| \leq \rho^k_\mu, \;\; \|\mathbf{\Sigma}^k_{r|s} - \hat{\mathbf{\Sigma}}^k_{r|s}\| \leq \left(\frac{\beta_k}{\alpha_k}\right)^2 \epsilon_\Sigma, \;\; \textit{and} \;\; \left\|\mathbf{\Omega}^k_{r|s} - \hat{\mathbf{\Omega}}^k_{r|s}\right\| \leq \rho^k_\Sigma,$$

where

$$\rho_\mu^k = \left(\frac{\beta_k}{\alpha_k} + 1\right)\epsilon_\mu + \frac{\alpha_k + \beta_k}{\alpha_k^2}\left(\|\hat{\boldsymbol{\mu}}\|_s^k + \|\boldsymbol{s}\|\right)\epsilon_\Sigma,$$

$$\rho_p^k = (p^k + \epsilon_p)\left(\frac{\epsilon_\Sigma\gamma_k^2 + \epsilon_\mu\gamma_k}{2\alpha_k^2\sqrt{(2\pi\alpha_k)^d}} + \frac{|\hat{\boldsymbol{\Sigma}}_{ss}^k + \epsilon_\Sigma \boldsymbol{I}_d| - |\hat{\boldsymbol{\Sigma}}_{ss}^k|}{2\alpha_k^d}\mathcal{N}\left(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k\right)\right) + \epsilon_p\mathcal{N}\left(\boldsymbol{s}|\hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k\right), \text{ and}$$

$$\rho_\Sigma^k = \left(\frac{\beta_k}{\alpha_k}\right)^2\epsilon_\Sigma + \left(\rho_\mu^k + \sum_{\ell\in[K]}(\hat{p}_{r|s}^\ell + \rho_p^\ell)\rho_\mu^\ell + \rho_p^\ell\|\hat{\boldsymbol{\mu}}_{r|s}^\ell\|\right)\left(2\left\|\hat{\boldsymbol{\mu}}_{r|s}^k - \hat{\boldsymbol{\mu}}_{r|s}\right\| + \left(\rho_\mu^k + \sum_{\ell\in[K]}(\hat{p}_{r|s}^\ell + \rho_p^\ell)\rho_\mu^\ell + \rho_p^\ell\|\hat{\boldsymbol{\mu}}_{r|s}^\ell\|\right)\right).$$

Proof of Lemma 2. Based on Lemma 9, we know that if $\|\boldsymbol{\mu}^k - \hat{\boldsymbol{\mu}}^k\|_2 \leq \epsilon_\mu$ then we have $\|\boldsymbol{\mu}_{r|s}^k - \hat{\boldsymbol{\mu}}_{r|s}^k\| \leq \rho_\mu^k$. Furthermore, from the lemma, if $\|\boldsymbol{\Sigma}^k - \hat{\boldsymbol{\Sigma}}^k\| \leq \epsilon_\Sigma$ then $\|\boldsymbol{\Sigma}_{r|s}^k - \hat{\boldsymbol{\Sigma}}_{r|s}^k\| \leq \left(\frac{\beta_k}{\alpha_k}\right)^2\epsilon_\Sigma$. Combining this result with Lemma 11, which provides an upper bound on the term

$$\left\|(\boldsymbol{\mu}_{r|s}^k - \boldsymbol{\mu}_{r|s})(\boldsymbol{\mu}_{r|s}^k - \boldsymbol{\mu}_{r|s})^\top - (\hat{\boldsymbol{\mu}}_{r|s}^k - \hat{\boldsymbol{\mu}}_{r|s})(\hat{\boldsymbol{\mu}}_{r|s}^k - \hat{\boldsymbol{\mu}}_{r|s})^\top\right\|,$$

we get

$$\left\|\boldsymbol{\Sigma}_{r|s}^k + (\boldsymbol{\mu}_{r|s}^k - \boldsymbol{\mu}_{r|s})(\boldsymbol{\mu}_{r|s}^k - \boldsymbol{\mu}_{r|s})^\top - \hat{\boldsymbol{\Omega}}_{r|s}^k\right\| \leq \rho_\Sigma^k.$$

Thus, the claim follows. □

Theorem 2 (Conservative Reformulation)**.** *Consider the setting in Lemma 2 and let* $\hat{\varphi}_k = \hat{p}^k\mathcal{N}\left(\boldsymbol{s}|\hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k\right)$. *The optimal value of the second-order cone program*

$$\begin{array}{ll}
\inf & \nu \\
\text{s.t.} & \boldsymbol{w} \in \Delta_n,\ \boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}_+^K,\ \boldsymbol{\tau} \in \mathbb{R}^K,\ \nu \in \mathbb{R} \\
& -\boldsymbol{w}^\top\hat{\boldsymbol{\mu}}_{r|s}^k + \rho_\mu^k\|\boldsymbol{w}\| + \eta\cdot\boldsymbol{w}^\top\left(\hat{\boldsymbol{\Omega}}_{r|s}^k + \rho_\Sigma^k\boldsymbol{I}_n\right)\boldsymbol{w} \leq \tau_k \quad \forall k \in [K] \\
& \boldsymbol{\tau} - \boldsymbol{u}_1 + \boldsymbol{u}_2 - \nu\mathbf{e} \leq \mathbf{0} \\
& \boldsymbol{u}_1^\top(\rho_p\mathbf{e} + \hat{\boldsymbol{\varphi}}) + \boldsymbol{u}_2^\top(\rho_p\mathbf{e} - \hat{\boldsymbol{\varphi}}) \leq 0
\end{array} \tag{6}$$

constitutes an upper bound to problem (5)*.*

Proof of Theorem 2. Since $p_{r|s}^k$ is non-negative for every $k \in [K]$, the objective function of problem (5) is upper bounded by

$$\sup_{\boldsymbol{p}\in\mathcal{Y}_p}\sum_{k=1}^K p_{r|s}^k\left(\sup_{\substack{\|\boldsymbol{\mu}_{r|s}^k - \hat{\boldsymbol{\mu}}_{r|s}^k\| \leq \rho_\mu^k \\ \|\boldsymbol{\Omega}_{r|s}^k - \hat{\boldsymbol{\Omega}}_{r|s}^k\| \leq \rho_\Sigma^k}} -\boldsymbol{w}^\top\boldsymbol{\mu}_{r|s}^k + \eta\cdot\boldsymbol{w}^\top\boldsymbol{\Omega}_{r|s}^k\boldsymbol{w}\right),$$

where $\mathcal{Y}_p \coloneqq \{\boldsymbol{p} \in \Delta_K : \{p^k, \boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k\}_{k=1}^K \in \mathcal{Y}\}$ is the projection of the uncertainty set $\mathcal{Y}$ onto the $\boldsymbol{p}$ axes.

We first deal with the inner optimization problems. For the k-th problem, its dual problem can be derived as

$$\begin{array}{ll} \inf & -\boldsymbol{w}^\top \hat{\boldsymbol{\mu}}_{r|s}^k + \rho_\mu \|\boldsymbol{w}\| + \eta \left(\left\langle \boldsymbol{Y}_1^k + \boldsymbol{Y}_2^k, \rho_\Sigma \boldsymbol{I}_n \right\rangle + \left\langle \boldsymbol{Y}_1^k - \boldsymbol{Y}_2^k, \hat{\boldsymbol{\Omega}}_{r|s}^k \right\rangle \right) \\ \text{s.t.} & \boldsymbol{Y}_1^k, \boldsymbol{Y}_2^k \in \mathbb{S}_+^{n \times n} \\ & \begin{bmatrix} \boldsymbol{Y}_1^k - \boldsymbol{Y}_2^k & \boldsymbol{w} \\ \boldsymbol{w}^\top & 1 \end{bmatrix} \succeq \mathbf{0}. \end{array} \tag{7}$$

Strong duality holds between the primal and dual pair because problem (7) has nonempty interior. In addition, it can be verified that $(\boldsymbol{Y}_1^k, \boldsymbol{Y}_2^k) = (\boldsymbol{w}\boldsymbol{w}^\top, \mathbf{0})$ is optimal to problem (7); see Lemma 6. Thus, the semidefinite program can be solved analytically. Let τ_k denote the optimal value of problem (7):

$$\tau_k = -\boldsymbol{w}^\top \hat{\boldsymbol{\mu}}_{r|s}^k + \rho_\mu^k \|\boldsymbol{w}\| + \eta \cdot \boldsymbol{w}^\top \left(\hat{\boldsymbol{\Omega}}_{r|s}^k + \rho_\Sigma^k \boldsymbol{I}_n \right) \boldsymbol{w}. \tag{8}$$

Therefore, problem (5) is upper bounded by

$$\min_{\boldsymbol{w} \in \Delta_n} \sup_{\boldsymbol{p} \in \mathcal{Y}_p} \sum_{k=1}^K p_{r|s}^k \tau_k.$$

Plugging the definition of $p_{r|s}^k$, we obtain the following maximization problem

$$\begin{array}{ll} \sup & \displaystyle\sum_{k=1}^K \frac{p^k \mathcal{N}\left(\boldsymbol{s} | \boldsymbol{\mu}_s^k, \boldsymbol{\Sigma}_{ss}^k\right) \tau_k}{\sum_{j=1}^K p^j \mathcal{N}\left(\boldsymbol{s} | \boldsymbol{\mu}_s^j, \boldsymbol{\Sigma}_{ss}^j\right)} \\ \text{s.t.} & p^k \in \Delta_K, \boldsymbol{\mu}^k \in \mathbb{R}^{d+s}, \boldsymbol{\Sigma}^k \in \mathbb{S}_+^{(n+d)\times(n+d)} \qquad \forall k \in [K] \\ & |p^k - \hat{p}^k| \le \epsilon_p, \|\boldsymbol{\mu}^k - \hat{\boldsymbol{\mu}}^k\|_2 \le \epsilon_\mu, \|\boldsymbol{\Sigma}^k - \hat{\boldsymbol{\Sigma}}^k\| \le \epsilon_\Sigma \quad \forall k \in [K]. \end{array} \tag{9}$$

Now we define a new variable $\varphi_k = p^k \mathcal{N}\left(\boldsymbol{s} | \boldsymbol{\mu}_s^k, \boldsymbol{\Sigma}_{ss}^k\right)$ and its empirical estimator $\hat{\varphi}_k = \hat{p}^k \mathcal{N}\left(\boldsymbol{s} | \hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k\right)$ for each $k \in [K]$. Based on Lemma 7, setting

$$\rho_p^k = (p^k + \epsilon_p) \left(\frac{(\|\boldsymbol{s} - \hat{\boldsymbol{\mu}}_s^k\| + \epsilon_\mu)\left(\frac{\epsilon_\Sigma}{2\alpha^2}(\|\boldsymbol{s} - \hat{\boldsymbol{\mu}}_s^k\| + \epsilon_\mu) + \frac{\epsilon_\mu}{\alpha}\right)}{\sqrt{(2\pi\alpha)^d}} + \frac{|\hat{\boldsymbol{\Sigma}}_{ss}^k + \epsilon_\Sigma \boldsymbol{I}| - |\hat{\boldsymbol{\Sigma}}_{ss}^k|}{2\alpha^d} \mathcal{N}\left(\boldsymbol{x} | \hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k\right) \right) + \epsilon_p \mathcal{N}\left(\boldsymbol{s} | \hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k\right),$$

the above optimization problem is upper bounded by

$$\sup_{\boldsymbol{\varphi} \in \mathbb{R}_+^K} \left\{ \sum_{k=1}^K \frac{\varphi_k \tau_k}{\sum_{j=1}^K \varphi_j} : \|\boldsymbol{\varphi} - \hat{\boldsymbol{\varphi}}\|_\infty \le \rho_0 \right\}. \tag{10}$$

Problem (10) is also known as a linear-fractional program (Bajalinov, 2003). Since the feasible region is non-empty and bounded, we can apply the Charnes-Cooper transformation (Charnes and Cooper, 1962) with

$$\boldsymbol{\ell} = \frac{\boldsymbol{\varphi}}{\mathbf{e}^\top \boldsymbol{\varphi}},\ t = \frac{1}{\mathbf{e}^\top \boldsymbol{\varphi}},$$

which reformulates the linear-fractional program (10) as an equivalent linear program

$$\begin{array}{ll} \sup & \boldsymbol{\tau}^\top \boldsymbol{\ell} \\ \text{s.t.} & \boldsymbol{\ell} \in \mathbb{R}^K_+,\ t \in \mathbb{R}_+ \\ & \boldsymbol{\ell} \leq (\rho_0 \mathbf{e} + \hat{\boldsymbol{\varphi}})\, t \\ & -\boldsymbol{\ell} \leq (\rho_0 \mathbf{e} - \hat{\boldsymbol{\varphi}})\, t \\ & \mathbf{e}^\top \boldsymbol{\ell} = 1. \end{array} \tag{11}$$

Dualizing this problem leads to the following minimization problem with the same optimal value:

$$\begin{array}{ll} \inf & \nu \\ \text{s.t.} & \boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^K_+, \nu \in \mathbb{R} \\ & \boldsymbol{\tau} - \boldsymbol{u}_1 + \boldsymbol{u}_2 - \nu \mathbf{e} \leq \mathbf{0} \\ & \boldsymbol{u}_1^\top (\rho_0 \mathbf{e} + \hat{\boldsymbol{\varphi}}) + \boldsymbol{u}_2^\top (\rho_0 \mathbf{e} - \hat{\boldsymbol{\varphi}}) \leq 0. \end{array} \tag{12}$$

Here, strong linear programming duality holds because problem (11) is feasible. For each $\tau_k \in [K]$ in problem (11), its optimal value can be obtained by solving the corresponding optimization problem (7). Combining the minimization problems yields the desired reformulation (6), which completes the proof. □

We remark that when the radii ϵ_p, ϵ_μ, and ϵ_Σ are brought down to 0, the upper bound (6) reduces to the true robust model (5), which is equivalent to the deterministic model (3) under the empirical estimates $\hat{p}^k_{\boldsymbol{r}|\boldsymbol{s}}$, $\hat{\boldsymbol{\mu}}^k_{\boldsymbol{r}|\boldsymbol{s}}$ and $\hat{\boldsymbol{\Sigma}}^k_{\boldsymbol{r}|\boldsymbol{s}}$, $k \in [K]$. Thus, our proposed approximation (6) is not overly conservative—it will become more accurate as we observe more samples and we decrease the radii accordingly with N. Under Assumption 1, the optimal solution of the approximation also enjoys the following out-of-sample performance guarantee.

Corollary 1 (Out-of-sample Guarantee)**.** *Let $\hat{\nu}$ and $\hat{\boldsymbol{w}}$ be respectively the optimal objective value and solution of the second-order cone program* (6)*. Then, with probability at least $1 - \delta$, we have*

$$\mathbb{E}_{\mathbb{G}^\star}[-\tilde{\boldsymbol{r}}^\top \hat{\boldsymbol{w}} \mid \tilde{\boldsymbol{s}} = \boldsymbol{s}] + \eta \mathbb{V}_{\mathbb{G}^\star}[\tilde{\boldsymbol{r}}^\top \hat{\boldsymbol{w}} \mid \tilde{\boldsymbol{s}} = \boldsymbol{s}] \leq \hat{\nu}.$$

Proof. We define the set of GM distributions whose parameters are close to the empirical estimates as

$$\hat{\mathcal{G}} \coloneqq \left\{ \{p^k, \boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k\}_{k=1}^K : \begin{array}{ll} |p^k - \hat{p}^k| \leq \epsilon_p,\ \|\boldsymbol{\mu}^k - \hat{\boldsymbol{\mu}}^k\| \leq \epsilon_\mu,\ \|\boldsymbol{\Sigma}^k - \hat{\boldsymbol{\Sigma}}^k\| \leq \epsilon_\Sigma & \forall k \in [K] \\ \boldsymbol{p} \in \Delta_K,\ \boldsymbol{\mu}^k \in \mathbb{R}^{n+d},\ \boldsymbol{\Sigma}^k \in \mathbb{S}_+^{(n+d)\times(n+d)} & \forall k \in [K] \end{array} \right\}.$$

By Assumption 1, we have $\mathbb{G}^\star \left(\{\boldsymbol{\mu}^{k\star}, \boldsymbol{\Sigma}^{k\star}, p^{k\star}\}_{k=1}^K \right) \in \mathcal{G}$ with probability $1-\delta$.

Observe now that for any fixed allocation $\boldsymbol{w}$, one can rewrite the objective function of the robust problem (5) as the *distributionally* robust model

$$\sup_{\mathbb{G}\in\hat{\mathcal{G}}} \mathbb{E}_{\mathbb{G}}[-\tilde{\boldsymbol{r}}^\top \boldsymbol{w} \mid \tilde{\boldsymbol{s}} = \boldsymbol{s}] + \eta \mathbb{V}_{\mathbb{G}}[\tilde{\boldsymbol{r}}^\top \boldsymbol{w} \mid \tilde{\boldsymbol{s}} = \boldsymbol{s}],$$

which is upper bounded by the optimal value of the conservative approximation (6) with fixed $\boldsymbol{w}$. Thus, the inequality

$$\mathbb{E}_{\mathbb{G}^\star}[-\tilde{\boldsymbol{r}}^\top \hat{\boldsymbol{w}} \mid \tilde{\boldsymbol{s}} = \boldsymbol{s}] + \eta \mathbb{V}_{\mathbb{G}^\star}[\tilde{\boldsymbol{r}}^\top \hat{\boldsymbol{w}} \mid \tilde{\boldsymbol{s}} = \boldsymbol{s}] \leq \sup_{\mathbb{G}\in\hat{\mathcal{G}}} \mathbb{E}_{\mathbb{G}}[-\tilde{\boldsymbol{r}}^\top \hat{\boldsymbol{w}} \mid \tilde{\boldsymbol{s}} = \boldsymbol{s}] + \eta \mathbb{V}_{\mathbb{G}}[\tilde{\boldsymbol{r}}^\top \hat{\boldsymbol{w}} \mid \tilde{\boldsymbol{s}} = \boldsymbol{s}],$$

holds with probability $1-\delta$. The claim then follows since the right-hand side is upper bounded by $\hat{\nu}$. □

4 Numerical Experiments

In this section, we present the numerical experiments and examine the performance of our proposed Robust Contextual Gaussian Mixture Model (RCGMM) along with several benchmark methods. All models are implemented in Python 3.7 with package CVXPY 1.1.0 and solved by MOSEK 9.2 (MOSEK ApS, 2019). All experiments were run on a 2.2GHz Intel Core i7 CPU laptop with 8GB RAM.

Historical Returns and Side Information: We download the S&P100 constituents data from Jan 2006 to Dec 2019 from *yfinance*[1]. To ensure the portfolio construction setup is stable and consistent, we only consider stocks listed in S&P100 during the entire period. This leaves us with 76 stocks in the portfolio. We conduct quarterly trading in the experiment and consider the quarterly stock percentage returns as our response variables $\tilde{\boldsymbol{r}}$. The reason for adopting this trading frequency is that most of the macroeconomic indicators are revealed quarterly. For the side

[1] https://pypi.org/project/yfinance/

information, we utilize 6 popular and publicly available macro indices: 1) US GDP growth rate, 2) US CPI, 3) US Interest rate, 4) Dollar Index, 5) US Unemployment rate, and 6) US Industrial Production Index. All macro data can be downloaded from *Economic Research: Federal Research Bank of St. Louis*[2].

Benchmark Methods: In the numerical experiment, we compare the following methods:

1. Equally-Weighted (EW) model: The EW portfolio allocates an equal weight to every asset when they are rebalanced. This method is also known as the $1/n$-portfolio and a detailed analysis can be found in DeMiguel et al. (2009b).

2. Mean-Variance (MV) model: The MV model, proposed by Markowitz (1952), is one of the best-known portfolio selection method. The model solves the optimization problem

$$\min_{\boldsymbol{w}\in\Delta_n} \mathbb{E}_{\hat{\mathbb{P}}}[-\tilde{\boldsymbol{r}}^\top \boldsymbol{w}] + \eta \cdot \mathbb{V}_{\hat{\mathbb{P}}}[\tilde{\boldsymbol{r}}^\top \boldsymbol{w}],$$

   where $\hat{\mathbb{P}}$ denotes the empirical distribution.

3. Conditional Mean-Variance (CMV) model: The CMV model assumes Gaussian returns, and solves the following optimization problem:

$$\min_{\boldsymbol{w}\in\Delta_n} \mathbb{E}_{\hat{\mathbb{P}}}\left[-\tilde{\boldsymbol{r}}^\top \boldsymbol{w} | \tilde{\boldsymbol{s}} = \boldsymbol{s}\right] + \eta \cdot \mathbb{V}_{\hat{\mathbb{P}}}\left[\tilde{\boldsymbol{r}}^\top \boldsymbol{w} | \tilde{\boldsymbol{s}} = \boldsymbol{s}\right].$$

   It incorporates side information into the MV model, which can also be regarded as the Contextual Gaussian Mixture model with $K = 1$.

4. Regularized Nadaraya-Watson (RNW) model: the NW regression method (Nadaraya, 1964; Watson, 1964) is a non-parametric regression scheme which approximates the conditional expectation with

$$\mathbb{E}\left[-\tilde{\boldsymbol{r}}^\top \boldsymbol{w} | \tilde{\boldsymbol{s}} = \boldsymbol{s}\right] \approx \hat{\mathbb{E}}\left[-\tilde{\boldsymbol{r}}^\top \boldsymbol{w} | \tilde{\boldsymbol{s}} = \boldsymbol{s}\right] = \frac{\sum_{i=1}^N \mathcal{K}\left(\frac{\boldsymbol{s}-\hat{\boldsymbol{s}}_i}{h}\right)(-\hat{\boldsymbol{r}}_i^\top \boldsymbol{w})}{\sum_{i=1}^N \mathcal{K}\left(\frac{\boldsymbol{s}-\hat{\boldsymbol{s}}_i}{h}\right)},$$

   where $\hat{\mathbb{E}}[-\tilde{\boldsymbol{r}}^\top \boldsymbol{w} | \tilde{\boldsymbol{s}} = \boldsymbol{s}]$ is the Nadaraya-Watson estimator, $\mathcal{K}$ is a prescribed kernel function and $h > 0$ is the bandwidth parameter of the kernel. This method does not rely on any specific distribution assumptions, such as Gaussian, on the asset returns. In addition, the model can

[2]https://research.stlouisfed.org

be efficiently robustified by introducing a conditional standard deviation term (Srivastava et al., 2021). Employing the result from (Srivastava et al., 2021, Corollary 3), the regularized NW problem can be reformulated as a second-order cone program. We implement this second-order cone program as a benchmark method for the experiment.

5. Robust Contextual Gaussian Mixture model (RCGMM): This is our proposed method, where the portfolio allocation is the solution of problem (6).

Experiment Setup: We first divide the dataset into a training set containing the stock prices from January 1, 2006, to January 1, 2016, and a test set containing data from January 1, 2016, to December 31, 2019. We next calculate the stocks' quarterly return using the first trading day's opening price each season. For example, the first stock trading day in the first quarter of 2006 is January 2, and the first trading day in the second quarter is April 3. Then, we divide the opening price on April 3 by the opening price on January 2 as the quarterly percentage return. Based on this principle, we obtain 55 quarterly return samples in total—40 for training and 15 for testing. In addition, for each historical quarterly return sample, we combine the previous season's macro indices as the corresponding signals or side information vector $\boldsymbol{s}$.

To tune the parameters, we split the first 4/5 of the training dataset into a subtraining set and take the remainder 1/5 as the validation set. We first tune the parameter η using the MV model in $[0.01, 10]$ on a logarithm search grid with 7 equidistant points. We pick the one that maximizes the Sharpe ratio evaluated using the validation set and apply the selected η to the CMV, NWR, and RCGMM models. Subsequently, we tune the number of clusters K and the radii of the uncertainty sets for the RCGMM model. We apply the *sklearn* built-in GMM learning algorithm to fit the GM distribution. We adopt a cross-validation method to select the best number of clusters from the set $\{1, 2, 3, 4, 5\}$. The criterion used to measure goodness-of-fit is the log-likelihood. After obtaining the best number of clusters, we fix the value of K and proceed to tune for the radii of the uncertainty sets. To avoid determining too many hyperparameters, we only tune the values of ρ_μ, ρ_Σ and keep ρ_p to be zero. We first set ρ_Σ to zero and tune for the best ρ_μ in $[0.01, 1]$ on a logarithm search grid with 7 equidistant points. Then we set ρ_μ to its best value and tune for ρ_Σ on the same search grid. For the RNW model, we adopt the exponential kernel for $\mathcal{K}$ and tune the bandwidth h from $\{1, 5, 10, 25, 50, 100\}$.

Experiment Results: Figure 1 depicts the performance of different strategies on the testing

data from the first quarter of 2016 to the third quarter of 2019. The y-axis represents the cumulative return, while the x-axis represents the date. Each grid stands for 3 months (one quarter), and transactions are only conducted at the beginning of the quarter. We do not consider transaction costs in the experiment as our trading frequency is low. Meanwhile, Table 1 reports the average return and the annualized Sharpe ratio of the different portfolio allocation strategies. To calculate the Sharpe ratio, we record the realized return in the test dataset for each quarter as r_i. Since we have four seasons in a year, the annualized Sharpe ratio is consequently computed by

$$\text{annualized Sharpe ratio} = \sqrt{4} \times \text{mean}\{r_i\}_{i\in[T]}/\text{std}\{r_i\}_{i\in[T]}.$$

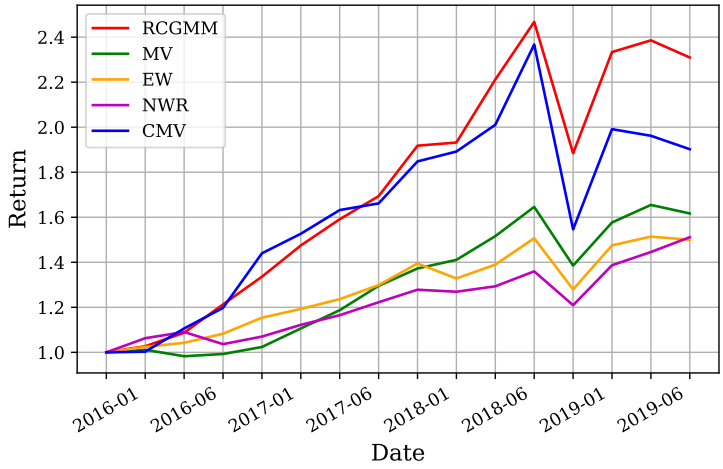

Figure 1. Out-of-sample Performance.

Model	EW	MV	CMV	NWR	RCGMM
Average Return	1.029	1.034	1.054	1.029	1.063
Annualized Sharpe ratio	0.914	1.034	0.796	1.082	1.232

Table 1. Statistics of different models

We remark that the RCGMM model outperforms its competitors in terms of both the average return and annualized Sharpe ratio. In addition, from the realized return curve plotted in Figure 1, we observe that the performance of the RCGMM is both profitable and stable. In contrast, although the CMV model achieves good performance before the third quarter of 2018, it yields a significant

retracement afterward. Meanwhile, we notice that the volatility of the MV method is lower than the CMV method. An avid reader might be interested in why the CMV model tends to make more 'aggressive' allocation decisions than the unconditional MV model when they have the same η. An intuitive explanation is that including side information will reduce the value of the variance term. Recall that the conditional variance is given by

$$\boldsymbol{\Sigma}_{r|s} = \quad \boldsymbol{\Sigma}_{rr} - \boldsymbol{\Sigma}_{rs}(\boldsymbol{\Sigma}_{ss})^{-1}\boldsymbol{\Sigma}_{sr}.$$

Thus, for any portfolio allocation $\boldsymbol{w}$, we have

$$\begin{aligned}\boldsymbol{w}^\top \boldsymbol{\Sigma}_{r|s}\boldsymbol{w} =& \boldsymbol{w}^\top \boldsymbol{\Sigma}_{rr}\boldsymbol{w} - \boldsymbol{w}^\top \boldsymbol{\Sigma}_{rs}(\boldsymbol{\Sigma}_{ss})^{-1}\boldsymbol{\Sigma}_{sr}\boldsymbol{w} \\ &= \boldsymbol{w}^\top \boldsymbol{\Sigma}_{rr}\boldsymbol{w} - \boldsymbol{x}^\top(\boldsymbol{\Sigma}_{ss})^{-1}\boldsymbol{x} \\ &\leq \boldsymbol{w}^\top \boldsymbol{\Sigma}_{rr}\boldsymbol{w},\end{aligned}$$

where $\boldsymbol{x} = \boldsymbol{\Sigma}_{sr}\boldsymbol{w}$ and the last inequality comes from the fact that $(\boldsymbol{\Sigma}_{ss})^{-1} \in \mathbb{S}_{++}^{d\times d}$. Thus, contextual information decreases the value of the variance term, and the model will become more confident in making certain decisions. This makes the overfitting effect more severe when the empirical estimators are used to approximate the conditional covariance matrix. Also, this 'aggressive' strategy may incur larger volatility in the out-of-sample test. From Table 1, we observe that while the CMV leads to a higher average return than the MV model, its Sharpe ratio gets smaller. Therefore, there is merit in including robustness and market regime shifts consideration using our proposed Robust Contextual Gaussian Mixture Model.

5 Conclusion

In this paper, we presented a new robust contextual optimization framework for portfolio optimization. Inspired by the regime modeling technique used for modeling financial markets, our framework models the uncertain returns of considered assets and the side information to follow a GMM. We derived a tractable conservative approximation for the robust optimization problem as a second-order cone program, which can be solved efficiently using off-the-shelf optimization software. Experimental results demonstrated the significant advantage of our approach over the state-of-the-art in exploiting the side information and alleviating the estimation errors from the empirical estimates. Our research opens up several promising directions for future research, such

as specialized computational schemes for robust contextual portfolio optimization and dynamic portfolio optimization using GMMs.

Acknowledgments

The work was partially supported by the National Science Foundation grants No. 1752125 and 2153606, and by the City University of Hong Kong (Project No. 9610481 and 7005688).

References

Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pages 458–469. Springer, 2005.

Vedat Akgiray and G Geoffrey Booth. Compound distribution models of stock returns: An empirical comparison. *Journal of Financial Research*, 10(3):269–280, 1987.

Andrew Ang and Geert Bekaert. International asset allocation with regime shifts. *The review of financial studies*, 15(4):1137–1187, 2002.

Andrew Ang and Geert Bekaert. How regimes affect asset allocation. *Financial Analysts Journal*, 60(2): 86–99, 2004.

Polen Arabacı and Burak Kocuk. Robust portfolio optimization models when stock returns are a mixture of normals. In *INFORMS International Conference on Service Science*, pages 419–430. Springer, 2020.

Fred D Arditti. Risk and the required return on equity. *The Journal of Finance*, 22(1):19–36, 1967.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47 (2):1148–1178, 2019.

Erik B Bajalinov. *Linear-fractional programming theory, methods, applications and software*, volume 84. Springer Science & Business Media, 2003.

Clifford A Ball and Walter N Torous. A simplified jump process for common stock returns. *Journal of Financial and Quantitative analysis*, 18(1):53–65, 1983.

Gah-Yi Ban and Cynthia Rudin. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108, 2019.

William L Beedles. Asymmetry in australian equity returns. *Australian Journal of Management*, 11(1):1–12, 1986.

Aharon Ben-Tal, Stephen Boyd, and Arkadi Nemirovski. Extending scope of robust optimization: Comprehensive robust counterparts of uncertain problems. *Mathematical Programming*, 107(1):63–89, 2006.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.

Stefano Benati and Romeo Rizzi. A mixed integer linear programming formulation of the optimal mean/value-at-risk portfolio problem. *European Journal of Operational Research*, 176(1):423–434, 2007.

Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66 (3):1025–1044, 2020.

Dimitris Bertsimas and Christopher McCord. From predictions to prescriptions in multistage optimization problems. *arXiv preprint arXiv:1904.11637*, 2019.

Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations research*, 52(1):35–53, 2004.

John R Birge. Optimization methods in dynamic portfolio management. *Handbooks in operations research and management science*, 15:845–865, 2007.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Jose Blanchet, Lin Chen, and Xun Yu Zhou. Distributionally robust mean-variance portfolio selection with wasserstein distances. *Management Science*, 2021.

Jose Blanchet, Karthyek Murthy, and Fan Zhang. Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes. *Mathematics of Operations Research*, 47(2):1500–1529, 2022.

Alex Botte and Doris Bao. A machine learning approach to regime modeling. *Two Sigma Street Review*, 2021.

Mark Britten-Jones. The sampling error in estimates of mean-variance efficient portfolio weights. *The Journal of Finance*, 54(2):655–671, 1999.

Ian Buckley, David Saunders, and Luis Seco. Portfolio optimization when asset returns have the gaussian mixture distribution. *European Journal of Operational Research*, 185(3):1434–1461, 2008.

Rachel Campbell, Kees Koedijk, and Paul Kofman. Increased correlation in bear markets. *Financial Analysts Journal*, 58(1):87–94, 2002.

Abraham Charnes and William W Cooper. Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, 9(3-4):181–186, 1962.

Zhi Chen, Melvyn Sim, and Peng Xiong. Robust stochastic optimization made easy with rsome. *Management Science*, 66(8):3329–3339, 2020.

Vijay K Chopra and William T Ziemba. The effect of errors in means, variances, and covariances on optimal portfolio choice. In *Handbook of the fundamentals of financial decision making: Part I*, pages 365–373. World Scientific, 2013.

Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative finance*, 1(2):223, 2001.

Victor DeMiguel and Francisco J Nogales. Portfolio selection with robust estimation. *Operations research*, 57(3):560–577, 2009.

Victor DeMiguel, Lorenzo Garlappi, Francisco J Nogales, and Raman Uppal. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management science*, 55 (5):798–812, 2009a.

Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The review of Financial studies*, 22(5):1915–1953, 2009b.

Vassiliy A Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158, 1969.

Fama Eugene and Kenneth French. The cross-section of expected stock returns. *Journal of Finance*, 47(2): 427–465, 1992.

Frank J Fabozzi, Svetlozar T Rachev, and Christian Menn. *Fat-tailed and skewed asset return distributions: implications for risk management, portfolio selection, and option pricing.* John Wiley & Sons, 2005.

Frank J Fabozzi, Petter N Kolm, Dessislava A Pachamanova, and Sergio M Focardi. Robust portfolio optimization. *The Journal of portfolio management*, 33(3):40–48, 2007.

Eugene F Fama. The behavior of stock-market prices. *The journal of Business*, 38(1):34–105, 1965.

Mark J Flannery and Aris A Protopapadakis. Macroeconomic factors do influence aggregate stock returns. *The review of financial studies*, 15(3):751–782, 2002.

Marco Gambacciani and Marc S Paolella. Robust normal mixtures for financial portfolio allocation. *Econometrics and Statistics*, 3:91–111, 2017.

Alireza Ghahtarani, Ahmed Saif, and Alireza Ghasemi. Robust portfolio selection problems: a comprehensive review. *Operational Research*, pages 1–62, 2022.

Donald Goldfarb and Garud Iyengar. Robust portfolio selection problems. *Mathematics of operations research*, 28(1):1–38, 2003.

Mark Grinblatt, Sheridan Titman, and Russ Wermers. Momentum investment strategies, portfolio performance, and herding: A study of mutual fund behavior. *The American economic review*, pages 1088–1105, 1995.

Aditya Gupta and Bhuwan Dhingra. Stock market prediction using hidden markov models. In *2012 Students Conference on Engineering and Systems*, pages 1–4. IEEE, 2012.

Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two Gaussians. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 753–760, 2015.

Rania Hentati-Kaffel and Jean-Luc Prigent. Portfolio optimization within mixture of distributions. 2014.

Iakovos Kakouris and Berç Rustem. Robust portfolio optimization with copulas. *European Journal of Operational Research*, 235(1):28–37, 2014.

Nathan Kallus and Xiaojie Mao. Stochastic optimization forests. *Management Science*, 2022.

Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. In *International conference on computational learning theory*, pages 444–457. Springer, 2005.

Rohit Kannan, Güzin Bayraksan, and James R Luedtke. Data-driven sample average approximation with covariate information. *Optimization Online. URL: http://www. optimization-online. org/DB HTML/2020/07/7932. html*, 2020a.

Rohit Kannan, Güzin Bayraksan, and James R Luedtke. Residuals-based distributionally robust optimization with covariate information. *arXiv preprint arXiv:2012.01088*, 2020b.

Stanley J Kon. Models of stock returns—a comparison. *The Journal of Finance*, 39(1):147–165, 1984.

Hiroshi Konno and Hiroaki Yamazaki. Mean-absolute deviation portfolio optimization model and its applications to tokyo stock market. *Management science*, 37(5):519–531, 1991.

Jeongyeol Kwon and Constantine Caramanis. The EM algorithm gives sample-optimality for learning mixtures of well-separated gaussians. In *Conference on Learning Theory*, pages 2425–2487. PMLR, 2020.

Wei Liu, Li Yang, and Bo Yu. Kernel density estimation based distributionally robust mean-cvar portfolio optimization. *Journal of Global Optimization*, pages 1–25, 2022.

Eric Luxenberg and Stephen Boyd. Portfolio construction with gaussian mixture returns and exponential utility via convex optimization. *arXiv preprint arXiv:2205.04563*, 2022.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, March 1952. URL `https://www.jstor.org/stable/2975974`.

MOSEK ApS. *MOSEK Optimizer API for Python 9.2.10*, 2019. URL `https://docs.mosek.com/9.2/pythonapi/index.html`.

Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

Anthony Neuberger. Realized skewness. *The Review of Financial Studies*, 25(11):3423–3455, 2012.

Viet Anh Nguyen, Soroosh Shafieezadeh Abadeh, Damir Filipović, and Daniel Kuhn. Mean-covariance robust risk measurement. *arXiv preprint arXiv:2112.09959*, 2021a.

Viet Anh Nguyen, Fan Zhang, Jose Blanchet, Erick Delage, and Yinyu Ye. Robustifying conditional portfolio decisions via optimal transport. *arXiv preprint arXiv:2103.16451*, 2021b.

Bernardo K Pagnoncelli, Domingo Ramírez, Hamed Rahimian, and Arturo Cifuentes. A synthetic data-plus-features driven approach for portfolio optimization. *Computational Economics*, pages 1–18, 2022.

Andre F Perold. Large-scale portfolio optimization. *Management science*, 30(10):1143–1160, 1984.

Georg Pflug and David Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442, 2007.

Ivilina Popova, David P Morton, Elmira Popova, and Jot Yau. Optimizing benchmark-based portfolios with hedge funds. *The Journal of Alternative Investments*, 10(1):35–55, 2007.

Peter Praetz and Edward JG Wilson. The distribution of stock market returns: 1958-1973. *Australian Journal of Management*, 3(1):79–90, 1978.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2: 21–42, 2000.

Napat Rujeerapaiboon, Daniel Kuhn, and Wolfram Wiesemann. Robust growth-optimal portfolios. *Management Science*, 62(7):2090–2109, 2016.

Tobias Rydén, Timo Teräsvirta, and Stefan Åsbrink. Stylized facts of daily return series and the hidden markov model. *Journal of applied econometrics*, 13(3):217–244, 1998.

Arora Sanjeev and Ravi Kannan. Learning mixtures of arbitrary Gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257, 2001.

Suvrajeet Sen and Yunxiao Deng. Learning enabled optimization: Towards a fusion of statistical learning and stochastic programming. *INFORMS Journal on Optimization (submitted)*, 2018.

Seyed Mohammad Sina Seyfi, Azin Sharifi, and Hamidreza Arian. Portfolio value-at-risk and expected-shortfall using an efficient simulation approach based on gaussian mixture model. *Mathematics and Computers in Simulation*, 190:1056–1079, 2021.

Xiang Shi and Young Shin Kim. Coherent risk measures and normal mixture distributions with applications in portfolio optimization. *International Journal of Theoretical and Applied Finance*, 24(04):2150019, 2021.

Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

Prateek R Srivastava, Yijie Wang, Grani A Hanasusanto, and Chin Pang Ho. On data-driven prescriptive analytics with side information: A regularized nadaraya-watson approach. *arXiv preprint arXiv:2110.04855*, 2021.

Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

Kai Ye, Panos Parpas, and Berç Rustem. Robust portfolio optimization: a conic programming approach. *Computational Optimization and Applications*, 52(2):463–481, 2012.

Shushang Zhu and Masao Fukushima. Worst-case conditional value-at-risk with application to robust portfolio management. *Operations research*, 57(5):1155–1168, 2009.

Guo Zi-Yi. Heavy-tailed distributions and risk management of equity market tail events. *Journal of Risk and Control*, 4(1), 2017.

Steve Zymler, Berç Rustem, and Daniel Kuhn. Robust portfolio optimization with derivative insurance guarantees. *European Journal of Operational Research*, 210(2):410–424, 2011.

Appendix A: Proofs of Auxiliary Results

Lemma 3. *Suppose matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are strictly positive definite with $\alpha\boldsymbol{I} \preceq \boldsymbol{A}$, $\alpha\boldsymbol{I} \preceq \boldsymbol{B}$. If $\|\boldsymbol{B}-\boldsymbol{A}\| \leq \epsilon$, then the difference of their inverse is bounded by*

$$\|\boldsymbol{A}^{-1}-\boldsymbol{B}^{-1}\| \leq \frac{\epsilon}{\alpha^2}$$

Proof. Observing that

$$\boldsymbol{A}^{-1}-\boldsymbol{B}^{-1} = \boldsymbol{A}^{-1}(\boldsymbol{B}-\boldsymbol{A})\boldsymbol{B}^{-1},$$

we have

$$\|\boldsymbol{A}^{-1}-\boldsymbol{B}^{-1}\| = \|\boldsymbol{A}^{-1}(\boldsymbol{B}-\boldsymbol{A})\boldsymbol{B}^{-1}\| \leq \|\boldsymbol{A}^{-1}(\boldsymbol{B}-\boldsymbol{A})\|\|\boldsymbol{B}^{-1}\| \leq \|\boldsymbol{A}^{-1}\|\|(\boldsymbol{B}-\boldsymbol{A})\|\|\boldsymbol{B}^{-1}\| \leq \frac{\epsilon}{\alpha^2},$$

where the inequalities comes from the fact that spectral norm is submultiplicative. □

Lemma 4. *For any vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, the spectral norm of the difference of their outer products is bounded by*

$$\|\boldsymbol{a}\boldsymbol{a}^\top - \boldsymbol{b}\boldsymbol{b}^\top\| \leq \|\boldsymbol{a}-\boldsymbol{b}\|(\|\boldsymbol{a}\|+\|\boldsymbol{b}\|).$$

Proof. We have

$$\begin{aligned}
\|\boldsymbol{a}\boldsymbol{a}^\top - \boldsymbol{b}\boldsymbol{b}^\top\| &= \sup_{\|\boldsymbol{x}\|=1} \|(\boldsymbol{a}\boldsymbol{a}^\top - \boldsymbol{b}\boldsymbol{b}^\top)\boldsymbol{x}\| \\
&\leq \sup_{\|\boldsymbol{x}\|=1} \|\boldsymbol{a}\boldsymbol{a}^\top\boldsymbol{x} - \boldsymbol{a}\boldsymbol{b}^\top\boldsymbol{x}\| + \sup_{\|\boldsymbol{x}\|=1} \|\boldsymbol{a}\boldsymbol{b}^\top\boldsymbol{x} - \boldsymbol{b}\boldsymbol{b}^\top\boldsymbol{x}\| \\
&\leq \sup_{\|\boldsymbol{x}\|=1} \|\boldsymbol{a}\||(\boldsymbol{a}-\boldsymbol{b})^\top\boldsymbol{x}| + \sup_{\|\boldsymbol{x}\|=1} \|\boldsymbol{a}-\boldsymbol{b}\||\boldsymbol{b}^\top\boldsymbol{x}| \\
&= \|\boldsymbol{a}\|\|\boldsymbol{a}-\boldsymbol{b}\| + \|\boldsymbol{a}-\boldsymbol{b}\|\|\boldsymbol{b}\|,
\end{aligned}$$

where the last equality follows from the definition of dual norm. Thus, the claim follows. □

Lemma 5. *Suppose $\boldsymbol{\mu}, \hat{\boldsymbol{\mu}} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}} \in \mathbb{S}_{++}$ satisfying $\alpha\boldsymbol{I} \preceq \boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}} \preceq \beta\boldsymbol{I}$. If*

$$\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|_2 \leq \epsilon_\mu, \|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}\| \leq \epsilon_\Sigma,$$

then we have

$$\begin{aligned}
&\left|\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) - \mathcal{N}\left(\boldsymbol{x}|\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}\right)\right| \\
\leq &\frac{\left(\rho_2/2 + \epsilon_\mu\|\hat{\boldsymbol{\Sigma}}^{-1}\|\right)\sum_{i=1}^d(|x_i-\hat{\mu}_i|+\rho)^2}{\sqrt{((2\pi\alpha)^d}} + \frac{|\hat{\boldsymbol{\Sigma}}+\epsilon_\Sigma\boldsymbol{I}|-|\hat{\boldsymbol{\Sigma}}|}{2\sqrt{(2\pi\alpha^2)^d|\hat{\boldsymbol{\Sigma}}|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})\right),
\end{aligned}$$

where $\rho = \frac{\epsilon_\Sigma \|\hat{\boldsymbol{\Sigma}}^{-1}\|^2}{1-\epsilon_\Sigma \|\hat{\boldsymbol{\Sigma}}^{-1}\|}$.

Proof. Without loss of generality, we sort the eigenvalues of $\hat{\boldsymbol{\Sigma}}$ in decreasing order as $\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_d$. Recall that the normal density function is given by

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}},$$

where π is known as the circular constant. Thus, we can rewrite the density function by its definition and obtain

$$\begin{aligned}
&\left|\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) - \mathcal{N}\left(\boldsymbol{x}|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\right)\right| \\
=&\left|\frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} - \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})\right)}{\sqrt{(2\pi)^d|\hat{\boldsymbol{\Sigma}}|}}\right| \\
=&\frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}||\hat{\boldsymbol{\Sigma}}|}}\left|\sqrt{|\hat{\boldsymbol{\Sigma}}|}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) - \sqrt{|\boldsymbol{\Sigma}|}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})\right)\right| \\
\leq&\frac{\left|\sqrt{|\hat{\boldsymbol{\Sigma}}|}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) - \sqrt{|\boldsymbol{\Sigma}|}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})\right)\right|}{\sqrt{(2\pi\alpha)^d|\hat{\boldsymbol{\Sigma}}|}}
\end{aligned}$$

Notice that the value of the denominator is given by data, and we only need to bound the numerator. Applying triangle inequality yields

$$\begin{aligned}
&\left|\sqrt{|\hat{\boldsymbol{\Sigma}}|}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) - \sqrt{|\boldsymbol{\Sigma}|}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})\right)\right| \\
\leq&\left|\sqrt{|\hat{\boldsymbol{\Sigma}}|}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) - \sqrt{|\hat{\boldsymbol{\Sigma}}|}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})\right)\right| + \\
&\quad\left|\sqrt{|\hat{\boldsymbol{\Sigma}}|}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})\right) - \sqrt{|\boldsymbol{\Sigma}|}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})\right)\right| \\
=&\sqrt{|\hat{\boldsymbol{\Sigma}}|}\left|\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) - \exp\left(-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})\right)\right| + \\
&\quad\left|\sqrt{|\boldsymbol{\Sigma}|} - \sqrt{|\hat{\boldsymbol{\Sigma}}|}\right|\exp\left(-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})\right). \qquad (13)
\end{aligned}$$

The above expression involves two absolute terms, and we then derive upper bounds for them. For the first term, since $\hat{\boldsymbol{\Sigma}}^{-1}, \hat{\boldsymbol{\Sigma}}^{-1} \in \mathbb{S}_{++}$, the products $-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) < 0$ and $-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}) < 0$. In addition, one can verify that the exponential function $f(a) = \exp(a)$ is

Lipschitz continuous with constant 1 when $a < 0$. Thus, we have

$$\begin{aligned}
&\left|\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) - \exp\left(-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})\right)\right| \\
\leq &\left|\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) - \left(-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})\right)\right| \\
\leq &\left|\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) - \left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\right| + \\
&\qquad\left|\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) - \left(-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})\right)\right| \\
= &\frac{1}{2}\left|(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{\Sigma}^{-1}-\hat{\boldsymbol{\Sigma}}^{-1})(\boldsymbol{x}-\boldsymbol{\mu})\right| + \frac{1}{2}\left|(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})^\top\hat{\boldsymbol{\Sigma}}^{-1}(2\boldsymbol{x}-\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\right|.
\end{aligned}$$

Based on Lemma 3, we know that $\|(\boldsymbol{\Sigma}^{-1}-\hat{\boldsymbol{\Sigma}}^{-1})\| \leq \frac{\epsilon_\Sigma}{\alpha^2}$. This spectral norm constraint can be equivalently written as

$$-\frac{\epsilon_\Sigma}{\alpha^2}\boldsymbol{I} \preceq \boldsymbol{\Sigma}^{-1}-\hat{\boldsymbol{\Sigma}}^{-1} \preceq \frac{\epsilon_\Sigma}{\alpha^2}\boldsymbol{I}.$$

This result further implies that

$$\begin{aligned}
&\frac{1}{2}\left|(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{\Sigma}^{-1}-\hat{\boldsymbol{\Sigma}}^{-1})(\boldsymbol{x}-\boldsymbol{\mu})\right| + \frac{1}{2}\left|(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})^\top\hat{\boldsymbol{\Sigma}}^{-1}(2\boldsymbol{x}-\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\right| \\
\leq &\frac{\epsilon_\Sigma}{2\alpha^2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{I}(\boldsymbol{x}-\boldsymbol{\mu}) + \frac{1}{2}\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\|\|\hat{\boldsymbol{\Sigma}}^{-1}(2\boldsymbol{x}-\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\| \\
\leq &\frac{\epsilon_\Sigma}{2\alpha^2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{I}(\boldsymbol{x}-\boldsymbol{\mu}) + \frac{1}{2}\epsilon_\mu\|\hat{\boldsymbol{\Sigma}}^{-1}\|\|(2\boldsymbol{x}-\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\| \\
\leq &\frac{\epsilon_\Sigma}{2\alpha^2}(\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|+\epsilon_\mu)^2 + \epsilon_\mu\|\hat{\boldsymbol{\Sigma}}^{-1}\|(\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|+\epsilon_\mu) \\
= &\frac{\epsilon_\Sigma\gamma^2}{2\alpha^2} + \frac{\epsilon_\mu\gamma}{\alpha} \\
= &\frac{\epsilon_\Sigma\gamma^2+\epsilon_\mu\gamma}{2\alpha^2},
\end{aligned}$$

where we define $\gamma = \|\boldsymbol{x}-\hat{\boldsymbol{\mu}}\|+\epsilon_\mu$. For the second term, we notice that $\exp\left(-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})\right)$ is determined by the GMM learning algorithm; thus, we only need to determine an upper bound

for $\left|\sqrt{|\boldsymbol{\Sigma}|}-\sqrt{|\hat{\boldsymbol{\Sigma}}|}\right|$. Applying algebraic transformations yields

$$\begin{aligned}
\left|\sqrt{|\boldsymbol{\Sigma}|}-\sqrt{|\hat{\boldsymbol{\Sigma}}|}\right| =& \frac{\left||\boldsymbol{\Sigma}|-|\hat{\boldsymbol{\Sigma}}|\right|}{\sqrt{|\boldsymbol{\Sigma}|}+\sqrt{|\hat{\boldsymbol{\Sigma}}|}} \\
\leq& \frac{1}{2\sqrt{\alpha^d}} \max\left\{\prod_{i=1}^d(\hat{\lambda}_i+\epsilon_\Sigma)-\prod_{i=1}^d\hat{\lambda}_i,\ \prod_{i=1}^d\hat{\lambda}_i-\prod_{i=1}^d\max\left\{\hat{\lambda}_i-\epsilon_\Sigma,\alpha\right\}\right\} \\
\leq& \frac{1}{2\sqrt{\alpha^d}} \max\left\{\prod_{i=1}^d(\hat{\lambda}_i+\epsilon_\Sigma)-\prod_{i=1}^d\hat{\lambda}_i,\ \prod_{i=1}^d\hat{\lambda}_i-\prod_{i=1}^d(\hat{\lambda}_i-\epsilon_\Sigma)\right\} \\
\leq& \frac{1}{2\sqrt{\alpha^d}} \left(\prod_{i=1}^d(\hat{\lambda}_i+\epsilon_\Sigma)-\prod_{i=1}^d\hat{\lambda}_i\right) \\
=& \frac{1}{2\sqrt{\alpha^d}} \left(|\hat{\boldsymbol{\Sigma}}+\epsilon_\Sigma\boldsymbol{I}|-|\hat{\boldsymbol{\Sigma}}|\right)
\end{aligned}$$

In summary, we conclude that

$$\begin{aligned}
&\left|\mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right)-\mathcal{N}\left(\boldsymbol{x}|\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}\right)\right| \\
\leq& \frac{1}{\sqrt{(2\pi\alpha)^d}}\frac{\epsilon_\Sigma\gamma^2+\epsilon_\mu\gamma}{2\alpha^2}+\frac{|\hat{\boldsymbol{\Sigma}}+\epsilon_\Sigma\boldsymbol{I}|-|\hat{\boldsymbol{\Sigma}}|}{2\sqrt{(2\pi\alpha^2)^d|\hat{\boldsymbol{\Sigma}}|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}})\right) \\
=& \frac{\epsilon_\Sigma\gamma^2+\epsilon_\mu\gamma}{2\alpha^2\sqrt{(2\pi\alpha)^d}}+\frac{|\hat{\boldsymbol{\Sigma}}+\epsilon_\Sigma\boldsymbol{I}|-|\hat{\boldsymbol{\Sigma}}|}{2\alpha^d}\mathcal{N}\left(\boldsymbol{x}|\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}\right),
\end{aligned}$$

which coincides with the result in Lemma 5. □

Lemma 6. *Suppose $\boldsymbol{w}\in\mathbb{R}^n$, $\boldsymbol{\Omega}\in\mathbb{S}_+^{n\times n}$ and $\rho\geq 0$, then $(\boldsymbol{Y}_1,\boldsymbol{Y}_2)=(\boldsymbol{w}\boldsymbol{w}^\top,\boldsymbol{0})$ is optimal to the following semidefinite program:*

$$\begin{array}{ll}
\inf & \langle\boldsymbol{Y}_1+\boldsymbol{Y}_2,\rho\boldsymbol{I}_n\rangle+\langle\boldsymbol{Y}_1-\boldsymbol{Y}_2,\boldsymbol{\Omega}\rangle \\
\text{s.t.} & \boldsymbol{Y}_1,\boldsymbol{Y}_2\in\mathbb{S}_+^{n\times n} \\
& \begin{bmatrix}\boldsymbol{Y}_1-\boldsymbol{Y}_2 & \boldsymbol{w}\\ \boldsymbol{w}^\top & 1\end{bmatrix}\succeq\boldsymbol{0}.
\end{array}$$

Proof of Lemma 6. It can be verified that $\boldsymbol{Y}_1=\boldsymbol{w}\boldsymbol{w}^\top$ and $\boldsymbol{Y}_2=\boldsymbol{0}$ are feasible to the semidefinite program. In addition, for any $\boldsymbol{Y}_1,\boldsymbol{Y}_2\succeq\boldsymbol{0}$, the semidefinite constraint in (7) can be equivalently written as

$$\boldsymbol{Y}_1-\boldsymbol{Y}_2\succeq\boldsymbol{w}\boldsymbol{w}^\top.$$

Therefore, we have

$$\begin{aligned}
&\langle \boldsymbol{Y}_1 + \boldsymbol{Y}_2, \rho \boldsymbol{I}_n \rangle + \langle \boldsymbol{Y}_1 - \boldsymbol{Y}_2, \boldsymbol{\Omega} \rangle \\
\geq &\left\langle \boldsymbol{w}\boldsymbol{w}^\top + 2\boldsymbol{Y}_2, \rho \boldsymbol{I}_n \right\rangle + \left\langle \boldsymbol{w}\boldsymbol{w}^\top, \boldsymbol{\Omega} \right\rangle \\
\geq &\left\langle \boldsymbol{w}\boldsymbol{w}^\top, \rho \boldsymbol{I}_n + \boldsymbol{\Omega} \right\rangle,
\end{aligned}$$

where the last inequality attained if and only if $\boldsymbol{Y}_1 = \boldsymbol{w}\boldsymbol{w}^\top$ and $\boldsymbol{Y}_2 = \boldsymbol{0}$. Thus, the claim follows. □

Lemma 7. *When* $\rho_0^k \geq (p^k + \epsilon_p) \left(\frac{\epsilon_\Sigma \gamma_k^2 + \epsilon_\mu \gamma_k}{2\alpha_k^2 \sqrt{(2\pi\alpha_k)^d}} + \frac{|\hat{\boldsymbol{\Sigma}}_{ss}^k + \epsilon_\Sigma \boldsymbol{I}| - |\hat{\boldsymbol{\Sigma}}_{ss}^k|}{2\alpha_k^d} \mathcal{N}\left(\boldsymbol{x} | \hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k \right) \right) + \epsilon_p \mathcal{N}\left(\boldsymbol{s} | \hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k \right)$, *the optimal value of problem* (9) *is upper bounded by problem* (10)*, where* $\alpha_k = \max\{\underline{\alpha}, \lambda(\hat{\boldsymbol{\Sigma}}^k)_{\min} - \epsilon_\Sigma\}$ *and* $\beta_k = \|\hat{\boldsymbol{\Sigma}}^k\| + \epsilon_\Sigma$.

Proof of Lemma 7. Based on Assumption 1 and Lemma 5, we know that $\|p^k - \hat{p}^k\| \leq \epsilon_\pi$ and $\left| \mathcal{N}\left(\boldsymbol{s} | \boldsymbol{\mu}_s^k, \boldsymbol{\Sigma}_{ss}^k \right) - \mathcal{N}\left(\boldsymbol{s} | \hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k \right) \right| \leq \frac{\epsilon_\Sigma \gamma_k^2 + \epsilon_\mu \gamma_k}{2\alpha_k^2 \sqrt{(2\pi\alpha_k)^d}} + \frac{|\hat{\boldsymbol{\Sigma}}_{ss}^k + \epsilon_\Sigma \boldsymbol{I}| - |\hat{\boldsymbol{\Sigma}}_{ss}^k|}{2\alpha_k^d} \mathcal{N}\left(\boldsymbol{x} | \hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k \right)$. For notational simplicity, we first define $r = \frac{\epsilon_\Sigma \gamma_k^2 + \epsilon_\mu \gamma_k}{2\alpha_k^2 \sqrt{(2\pi\alpha_k)^d}} + \frac{|\hat{\boldsymbol{\Sigma}}_{ss}^k + \epsilon_\Sigma \boldsymbol{I}| - |\hat{\boldsymbol{\Sigma}}_{ss}^k|}{2\alpha_k^d} \mathcal{N}\left(\boldsymbol{x} | \hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k \right)$. Noticing that the terms $p^k, \hat{p}^k$, $\mathcal{N}\left(\boldsymbol{s} | \boldsymbol{\mu}_s^k, \boldsymbol{\Sigma}_{ss}^k \right)$, and $\mathcal{N}\left(\boldsymbol{s} | \hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k \right)$ in problem (9) are all non-negative, we have

$$\begin{aligned}
&\left| p^k \mathcal{N}\left(\boldsymbol{s} | \boldsymbol{\mu}_s^k, \boldsymbol{\Sigma}_{ss}^k \right) - \hat{p}^k \mathcal{N}\left(\boldsymbol{s} | \hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k \right) \right| \\
\leq &\max \Big\{ (\hat{p}^k + \epsilon_\pi) \left(\mathcal{N}\left(\boldsymbol{s} | \hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k \right) + r \right) - \hat{p}^k \mathcal{N}\left(\boldsymbol{s} | \hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k \right), \\
&\qquad \hat{p}^k \mathcal{N}\left(\boldsymbol{s} | \hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k \right) - (\hat{p}^k - \epsilon_\pi) \left(\mathcal{N}\left(\boldsymbol{s} | \hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k \right) - r \right) \Big\} \\
= &\max \left\{ \epsilon_p \mathcal{N}\left(\boldsymbol{s} | \hat{\boldsymbol{\mu}}_s, \hat{\boldsymbol{\Sigma}}_{ss}^k \right) + \hat{p}^k r + \epsilon_\pi r, \epsilon_p \mathcal{N}\left(\boldsymbol{s} | \hat{\boldsymbol{\mu}}_s, \hat{\boldsymbol{\Sigma}}_{ss}^k \right) + \hat{p}^k r - \epsilon_\pi r \right\} \\
= &\epsilon_p \mathcal{N}\left(\boldsymbol{s} | \hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k \right) + \hat{p}^k r + \epsilon_\pi r \\
= &(p^k + \epsilon_p) \left(\frac{\epsilon_\Sigma \gamma_k^2 + \epsilon_\mu \gamma_k}{2\alpha_k^2 \sqrt{(2\pi\alpha_k)^d}} + \frac{|\hat{\boldsymbol{\Sigma}}_{ss}^k + \epsilon_\Sigma \boldsymbol{I}| - |\hat{\boldsymbol{\Sigma}}_{ss}^k|}{2\alpha_k^d} \mathcal{N}\left(\boldsymbol{x} | \hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k \right) \right) + \epsilon_p \mathcal{N}\left(\boldsymbol{s} | \hat{\boldsymbol{\mu}}_s^k, \hat{\boldsymbol{\Sigma}}_{ss}^k \right),
\end{aligned}$$

which completes the proof. □

Lemma 8. *Suppose* $\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\| \leq \epsilon_\Sigma$*, then we have*

$$\|\boldsymbol{\Sigma}_{rs}^k - \hat{\boldsymbol{\Sigma}}_{rs}^k\| \leq \epsilon_\Sigma, \ \|\boldsymbol{\Sigma}_{ss}^k - \hat{\boldsymbol{\Sigma}}_{ss}^k\| \leq \epsilon_\Sigma$$

Proof of Lemma 8. By the definition of spectral norm, we have

$$\begin{aligned}
\|\boldsymbol{\Sigma}_{rs}^k - \hat{\boldsymbol{\Sigma}}_{rs}^k\| &= \sup_{\|\boldsymbol{x}\|=1} \|(\boldsymbol{\Sigma}_{rs}^k - \hat{\boldsymbol{\Sigma}}_{rs}^k)\boldsymbol{x}\| \\
&\leq \sup_{\|\boldsymbol{y}\|=1} \left\| \begin{pmatrix} \boldsymbol{\Sigma}_{rs}^k - \hat{\boldsymbol{\Sigma}}_{rs}^k \\ \boldsymbol{\Sigma}_{rr}^k - \hat{\boldsymbol{\Sigma}}_{rr}^k \end{pmatrix} \boldsymbol{y} \right\| \\
&= \sup_{\|\boldsymbol{y}\|=1} \left\| (\boldsymbol{\Sigma}^k - \hat{\boldsymbol{\Sigma}}^k) \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{y} \end{pmatrix} \right\| \\
&\leq \sup_{\|\boldsymbol{z}\|=1} \|(\boldsymbol{\Sigma}^k - \hat{\boldsymbol{\Sigma}}^k)\boldsymbol{z}\| \\
&= \|\boldsymbol{\Sigma}^k - \hat{\boldsymbol{\Sigma}}^k\| \\
&\leq \epsilon_\Sigma.
\end{aligned}$$

The derivation for $\|\boldsymbol{\Sigma}_{ss}^k - \hat{\boldsymbol{\Sigma}}_{ss}^k\|$ follows the same strategy, and we omit for brevity. □

Lemma 9. *Suppose* $\|\boldsymbol{\mu}^k - \hat{\boldsymbol{\mu}}^k\|_2 \leq \epsilon_\mu$, $\|\boldsymbol{\Sigma}^k - \hat{\boldsymbol{\Sigma}}^k\| \leq \epsilon_\Sigma$. *By setting* $\alpha_k = \max\{\underline{\alpha}, \lambda(\hat{\boldsymbol{\Sigma}}^k)_{\min} - \epsilon_\Sigma\}$, $\beta_k = \|\hat{\boldsymbol{\Sigma}}^k\| + \epsilon_\Sigma$, *we have*

$$\begin{aligned}
\|\boldsymbol{\mu}_{r|s}^k - \hat{\boldsymbol{\mu}}_{r|s}^k\| &\leq \left(\frac{\beta_k}{\alpha_k} + 1\right)\epsilon_\mu + \frac{\alpha_k + \beta_k}{\alpha_k^2}\left(\|\hat{\boldsymbol{\mu}}\|_s^k + \|\boldsymbol{s}\|\right)\epsilon_\Sigma, \\
\|\boldsymbol{\Sigma}_{r|s}^k - \hat{\boldsymbol{\Sigma}}_{r|s}^k\| &\leq \left(\frac{\beta_k}{\alpha_k}\right)^2 \epsilon_\Sigma.
\end{aligned}$$

Proof. We first derive the error bound for the conditional mean.

$$\begin{aligned}
\|\boldsymbol{\mu}_{r|s}^k - \hat{\boldsymbol{\mu}}_{r|s}^k\| =& \|\boldsymbol{\mu}_r^k + \boldsymbol{\Sigma}_{rs}^k(\boldsymbol{\Sigma}_{ss}^k)^{-1}(\boldsymbol{s} - \boldsymbol{\mu}_s^k) - \hat{\boldsymbol{\mu}}_r^k - \hat{\boldsymbol{\Sigma}}_{rs}^k(\hat{\boldsymbol{\Sigma}}_{ss}^k)^{-1}(\boldsymbol{s} - \hat{\boldsymbol{\mu}}_s^k)\| \\
\leq& \|\boldsymbol{\mu}_r^k - \hat{\boldsymbol{\mu}}_r^k\| + \|\boldsymbol{\Sigma}_{rs}^k(\boldsymbol{\Sigma}_{ss}^k)^{-1}(\boldsymbol{s} - \boldsymbol{\mu}_s^k) - \hat{\boldsymbol{\Sigma}}_{rs}^k(\hat{\boldsymbol{\Sigma}}_{ss}^k)^{-1}(\boldsymbol{s} - \hat{\boldsymbol{\mu}}_s^k)\| \\
\leq& \epsilon_\mu + \|\boldsymbol{s}\| \cdot \|\boldsymbol{\Sigma}_{rs}^k(\boldsymbol{\Sigma}_{ss}^k)^{-1} - \hat{\boldsymbol{\Sigma}}_{rs}^k(\hat{\boldsymbol{\Sigma}}_{ss}^k)^{-1}\| + \|\boldsymbol{\Sigma}_{rs}^k(\boldsymbol{\Sigma}_{ss}^k)^{-1}\boldsymbol{\mu}_s^k - \hat{\boldsymbol{\Sigma}}_{rs}^k(\hat{\boldsymbol{\Sigma}}_{ss}^k)^{-1}\hat{\boldsymbol{\mu}}_s^k\|
\end{aligned}$$

where the last inequality comes from the fact that the norm of the sub-vector is less than the norm of the whole vector. For the second term, we have

$$\begin{aligned}
\|\boldsymbol{\Sigma}_{rs}^k(\boldsymbol{\Sigma}_{ss}^k)^{-1} - \hat{\boldsymbol{\Sigma}}_{rs}^k(\hat{\boldsymbol{\Sigma}}_{ss}^k)^{-1}\| \leq& \|\boldsymbol{\Sigma}_{rs}^k(\boldsymbol{\Sigma}_{ss}^k)^{-1} - \hat{\boldsymbol{\Sigma}}_{rs}^k(\boldsymbol{\Sigma}_{ss}^k)^{-1}\| + \|\hat{\boldsymbol{\Sigma}}_{rs}^k(\boldsymbol{\Sigma}_{ss}^k)^{-1} - \hat{\boldsymbol{\Sigma}}_{rs}^k(\hat{\boldsymbol{\Sigma}}_{ss}^k)^{-1}\| \\
\leq& \|(\boldsymbol{\Sigma}_{ss}^k)^{-1}\| \cdot \|\boldsymbol{\Sigma}_{rs}^k - \hat{\boldsymbol{\Sigma}}_{rs}^k\| + \|\hat{\boldsymbol{\Sigma}}_{rs}^k\| \cdot \|(\boldsymbol{\Sigma}_{ss}^k)^{-1} - (\hat{\boldsymbol{\Sigma}}_{ss}^k)^{-1}\| \\
\leq& \frac{\epsilon_\Sigma}{\alpha_k} + \frac{\beta_k \epsilon_\Sigma}{\alpha_k^2} \\
=& \frac{\alpha_k + \beta_k}{\alpha_k^2}\epsilon_\Sigma,
\end{aligned}$$

where the third inequality comes from Lemma 3. Next, we employ this result to the third term and obtain

$$\begin{aligned}
&\|\boldsymbol{\Sigma}_{rs}^k(\boldsymbol{\Sigma}_{ss}^k)^{-1}\boldsymbol{\mu}_s^k - \hat{\boldsymbol{\Sigma}}_{rs}^k(\hat{\boldsymbol{\Sigma}}_{ss}^k)^{-1}\hat{\boldsymbol{\mu}}_s^k\| \\
\leq&\|\boldsymbol{\Sigma}_{rs}^k(\boldsymbol{\Sigma}_{ss}^k)^{-1}\boldsymbol{\mu}_s^k - \boldsymbol{\Sigma}_{rs}^k(\boldsymbol{\Sigma}_{ss}^k)^{-1}\hat{\boldsymbol{\mu}}_s^k\| + \|\hat{\boldsymbol{\Sigma}}_{rs}^k(\hat{\boldsymbol{\Sigma}}_{ss}^k)^{-1}\hat{\boldsymbol{\mu}}_s^k - \hat{\boldsymbol{\Sigma}}_{rs}^k(\hat{\boldsymbol{\Sigma}}_{ss}^k)^{-1}\hat{\boldsymbol{\mu}}_s^k\| \\
\leq&\|\boldsymbol{\Sigma}_{rs}^k(\boldsymbol{\Sigma}_{ss}^k)^{-1}\| \cdot \|\boldsymbol{\mu}_s^k - \hat{\boldsymbol{\mu}}_s^k\| + \|\boldsymbol{\Sigma}_{rs}^k(\boldsymbol{\Sigma}_{ss}^k)^{-1} - \hat{\boldsymbol{\Sigma}}_{rs}^k(\hat{\boldsymbol{\Sigma}}_{ss}^k)^{-1}\| \cdot \|\hat{\boldsymbol{\mu}}_s^k\| \\
\leq&\frac{\beta_k}{\alpha_k}\epsilon_\mu + \frac{\alpha_k+\beta_k}{\alpha_k^2}\epsilon_\Sigma\|\hat{\boldsymbol{\mu}}_s^k\|.
\end{aligned}$$

Combining the results for the second and third term, we have

$$\|\boldsymbol{\mu}_{r|s}^k - \hat{\boldsymbol{\mu}}_{r|s}^k\| \leq \left(\frac{\beta_k}{\alpha_k}+1\right)\epsilon_\mu + \frac{\alpha_k+\beta_k}{\alpha_k^2}\left(\|\hat{\boldsymbol{\mu}}\|_s^k + \|\boldsymbol{s}\|\right)\epsilon_\Sigma.$$

We then derive the error bound for the conditional covariance. Recall that $\boldsymbol{\Psi}^k = \boldsymbol{\Sigma}^k$, thus, we can invoke Lemma 3 and obtain

$$\|\boldsymbol{\Psi}^k - \hat{\boldsymbol{\Psi}}^k\| \leq \frac{\epsilon_\Sigma}{\alpha_k^2}.$$

By Lemma 8, we know that the norm of a submatrix is less than the whole matrix, i.e.,

$$\|\boldsymbol{\Psi}_{rr}^k - \hat{\boldsymbol{\Psi}}_{rr}^k\| \leq \frac{\epsilon_\Sigma}{\alpha_k^2}.$$

By Lemma 1, the conditional covariance matrix is equivalent to the inverse of $(\boldsymbol{\Psi}_{rr}^k)^{-1}$. Hence, we apply Lemma 3 again and obtain

$$\|\boldsymbol{\Sigma}_{r|s}^k - \hat{\boldsymbol{\Sigma}}_{r|s}^k\| \leq \left(\frac{\beta_k}{\alpha_k}\right)^2\epsilon_\Sigma.$$

This completes the proof. □

Lemma 10. *Suppose* $|p_{r|s}^k - \hat{p}_{r|s}^k| \leq \rho_p^k$ *and* $\|\boldsymbol{\mu}_{r|s}^k - \hat{\boldsymbol{\mu}}_{r|s}^k\| \leq \rho_\mu^k$ *for all* $k \in [K]$. *Then,*

$$\left\|\sum_{k\in[K]} p_{r|s}^k\boldsymbol{\mu}_{r|s}^k - \sum_{k\in[K]}\hat{p}_{r|s}^k\hat{\boldsymbol{\mu}}_{r|s}^k\right\| \leq \sum_{k\in[K]}(\hat{p}_{r|s}^k+\rho_p^k)\rho_\mu^k + \rho_p^k\|\hat{\boldsymbol{\mu}}_{r|s}^k\|.$$

Proof. We have

$$\begin{aligned}
\left\|\sum_{k\in[K]} p^k_{r|s}\boldsymbol{\mu}^k_{r|s} - \sum_{k\in[K]} \hat{p}^k_{r|s}\hat{\boldsymbol{\mu}}^k_{r|s}\right\| &\leq \left\|\sum_{k\in[K]} p^k_{r|s}\boldsymbol{\mu}^k_{r|s} - \sum_{k\in[K]} p^k_{r|s}\hat{\boldsymbol{\mu}}^k_{r|s}\right\| + \left\|\sum_{k\in[K]} p^k_{r|s}\hat{\boldsymbol{\mu}}^k_{r|s} - \sum_{k\in[K]} \hat{p}^k_{r|s}\hat{\boldsymbol{\mu}}^k_{r|s}\right\| \\
&\leq \left\|\sum_{k\in[K]} p^k_{r|s}(\boldsymbol{\mu}^k_{r|s} - \hat{\boldsymbol{\mu}}^k_{r|s})\right\| + \left\|\sum_{k\in[K]} (p^k_{r|s} - \hat{p}^k_{r|s})\hat{\boldsymbol{\mu}}^k_{r|s}\right\| \\
&\leq \sum_{k\in[K]} \left(p^k_{r|s}\rho^k_\mu + \rho^k_p\|\hat{\boldsymbol{\mu}}^k_{r|s}\|\right) \\
&\leq \sum_{k\in[K]} (\hat{p}^k_{r|s} + \rho^k_p)\rho^k_\mu + \rho^k_p\|\hat{\boldsymbol{\mu}}^k_{r|s}\|.
\end{aligned}$$

Thus, the claim follows. □

Lemma 11. *Suppose* $|p^k_{r|s} - \hat{p}^k_{r|s}| \leq \rho^k_p$ *and* $\|\boldsymbol{\mu}^k_{r|s} - \hat{\boldsymbol{\mu}}^k_{r|s}\| \leq \rho^k_\mu$ *for all* $k \in [K]$. *Define* $\boldsymbol{\mu}_{r|s} = \sum_{k\in[K]} p^k_{r|s}\boldsymbol{\mu}^k_{r|s}$ *and* $\hat{\boldsymbol{\mu}}_{r|s} = \sum_{k\in[K]} \hat{p}^k_{r|s}\hat{\boldsymbol{\mu}}^k_{r|s}$. *Then, we have*

$$\begin{aligned}
&\left\|(\boldsymbol{\mu}^k_{r|s} - \boldsymbol{\mu}_{r|s})(\boldsymbol{\mu}^k_{r|s} - \boldsymbol{\mu}_{r|s})^\top - (\hat{\boldsymbol{\mu}}^k_{r|s} - \hat{\boldsymbol{\mu}}_{r|s})(\hat{\boldsymbol{\mu}}^k_{r|s} - \hat{\boldsymbol{\mu}}_{r|s})^\top\right\| \\
&\leq \left(\rho^k_\mu + \sum_{\ell\in[K]} (\hat{p}^\ell_{r|s} + \rho^\ell_p)\rho^\ell_\mu + \rho^\ell_p\|\hat{\boldsymbol{\mu}}^\ell_{r|s}\|\right)\left(2\left\|\hat{\boldsymbol{\mu}}^k_{r|s} - \hat{\boldsymbol{\mu}}_{r|s}\right\| + \left(\rho^k_\mu + \sum_{\ell\in[K]} (\hat{p}^\ell_{r|s} + \rho^\ell_p)\rho^\ell_\mu + \rho^\ell_p\|\hat{\boldsymbol{\mu}}^\ell_{r|s}\|\right)\right).
\end{aligned}$$

Proof. From Lemma 4, we get

$$\begin{aligned}
&\left\|(\boldsymbol{\mu}^k_{r|s} - \boldsymbol{\mu}_{r|s})(\boldsymbol{\mu}^k_{r|s} - \boldsymbol{\mu}_{r|s})^\top - (\hat{\boldsymbol{\mu}}^k_{r|s} - \hat{\boldsymbol{\mu}}_{r|s})(\hat{\boldsymbol{\mu}}^k_{r|s} - \hat{\boldsymbol{\mu}}_{r|s})^\top\right\| \\
&\leq \left\|(\boldsymbol{\mu}^k_{r|s} - \boldsymbol{\mu}_{r|s}) - (\hat{\boldsymbol{\mu}}^k_{r|s} - \hat{\boldsymbol{\mu}}_{r|s})\right\|\left(\left\|(\boldsymbol{\mu}^k_{r|s} - \boldsymbol{\mu}_{r|s})\right\| + \left\|\hat{\boldsymbol{\mu}}^k_{r|s} - \hat{\boldsymbol{\mu}}_{r|s}\right\|\right).
\end{aligned}$$

Applying Lemma 10, we obtain

$$\left\|(\boldsymbol{\mu}^k_{r|s} - \boldsymbol{\mu}_{r|s}) - (\hat{\boldsymbol{\mu}}^k_{r|s} - \hat{\boldsymbol{\mu}}_{r|s})\right\| \leq \rho^k_\mu + \sum_{\ell\in[K]} (\hat{p}^\ell_{r|s} + \rho^\ell_p)\rho^\ell_\mu + \rho^\ell_p\|\hat{\boldsymbol{\mu}}^\ell_{r|s}\|.$$

The claim then follows from upper bounding $\left\|(\boldsymbol{\mu}^k_{r|s} - \boldsymbol{\mu}_{r|s})\right\|$ with $\left\|\hat{\boldsymbol{\mu}}^k_{r|s} - \hat{\boldsymbol{\mu}}_{r|s}\right\| + \rho^k_\mu + \sum_{\ell\in[K]} (\hat{p}^\ell_{r|s} + \rho^\ell_p)\rho^\ell_\mu + \rho^\ell_p\|\hat{\boldsymbol{\mu}}^\ell_{r|s}\|$. □