Adversarial Analytics

Thèse N° 9731

Présentée le 20 septembre 2019

au Collège du management de la technologie
Chaire analyse de risque et optimisation
Programme doctoral en management de la technologie

pour l'obtention du grade de Docteur ès Sciences

par

Viet Anh NGUYEN

Acceptée sur proposition du jury

Prof. L. Lambertini, présidente du jury
Prof. D. Kuhn, directeur de thèse
Prof. M. Sim, rapporteur
Prof. E. Delage, rapporteur
Prof. D. Filipovic, rapporteur

2019

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Acknowledgements

Finishing the Ph.D. thesis is an emotional moment in my life. It gives me a once in a lifetime period to dwell on my past, reflect about my present and realize that this thesis would not be made possible without the support and encouragement I have received from many people during the past 6 years.

I would like to thank Dr. Daniel D. Kuhn for his excruciatingly detailed comments which have pushed the discoveries of the results in this thesis to the extreme limits. Above everything, you have been an example of the highest caliber “Swiss-quality” researcher, advisor and mentor: you give me freedom to dream and pursue my ideas, and give me excellent support when dealing with the technical difficulties of the problems. You are the role model that will tire me out for the years to come in case I want to surpass. To one degree or another, we share the same passion for the mountains and I have to admit that the best research discussions we had actually took place on the slopes of Saint Luc when you literally tore the results of this thesis apart and then re-assembled them together.

I am also indebted to my supervisor, Dr. Peyman Mohajerin Esfahani, for his continuity of dedication which attains the limit of superiority that anyone can wish for. Through these years, we have constantly enjoyed challenging our own understanding of (convex) optimization, sharing ups and downs, and tricking one another into high risk zero-to-low return research projects.

I am deeply grateful to the thesis committee members: Dr. Luisa Lambertini, Dr. Melvyn Sim, Dr. Erick Delage and Dr. Damir Filipović. I would like to extend my sincere gratitude to Dr. Erick Delage for his careful verification of many results in this thesis.

My research at EPFL has been funded from different sources. I would like to thank Dr. Michael Lehning from the CRYOS laboratory at EPFL for his generous funding and for the provision of the solar datasets that have ignited my research interests on the inverse covariance estimation. I acknowledge and thank the entire community of the College of Management at EPFL (professors, administrative staff and IT team) for their contribution to the progress of my dissertation.

I feel very fortunate to collaborate with many excellent researchers throughout my stay at EPFL: Soroosh, Manchung, Christos, Dr. Pierre Pinson and Dr. Wolfram Wiesemann. You are surprisingly supportive and optimistic towards my randomly-generated research ideas.

I would like to thank all current and former members of the RAO group with whom I have shared enlightening discussions: Napat, Angelos, Grani, Cagil, Bahar, Dirk, Kilian, Soroosh, Tobias and Trevor. Had it been without the peer pressure from you, I would have been more relaxed when I needed to think deeply about my models. In particular, I would like to thank

Napat for always offering me drinks; Angelos for introducing me to the best pizza restaurant in Montréal, the memory of which remains elusive; Soroosh for awesome burger nights and Trevor for joining me in many innocent randomized acts.

Friends from the Cube group have always been welcoming from my very first days in Lausanne, special thanks to Olov and Primerov who transformed my life into a much better version with mountaineering and skiing, and most importantly encouraged me to live my life to the fullest. The Bochet 29 group is constant source of fun, especially with Florian, Jannik and Andreas. I would like to personally thank Andreas for many exciting discussions which were finally crystallized into two NeurIPS papers. Thanks to Giovanni and Peppe who peacefully converted me into a Juventus fan and acclimatized me to the Italian community in EPFL with Monica, Fabiana, Claudia and Giada. I would also like to thank the Vietnamese community in Lausanne (VNLausanne) and the PhD friends from CDM and Gerzensee, especially Maryam, Corinne and Roman.

My sister and my brother-in-law have been always supporting me through all the difficulties I encountered during my studies, they have sacrificed a lot to give me the peace of mind in order to concentrate and pursue my research. Bông and Dua brighten up my everyday and make me positively rethink about my goals in life. Thank you very much.

The research leading up to this thesis encountered many challenging technical difficulties that have exhausted the energy of both me and Peyman. In October 2017, after almost 2 years working on the inverse covariance estimation problem, I was still not satisfied with the results: the reformulation at that time looked over-complicated, and I relentlessly blamed myself for not understanding the (seemingly simple) problem well enough. I constantly had the ideas of giving up and quitting my PhD. It was Hien's continuous encouragement during these rough periods that kept me on track and gave me the courage to attempt a fresh look into the problem, and magically, only 3 months later, the paper was completely re-written with an insightful and elegant proof. Your love and support have seen me through.

I would not be here today without my parents. It is extremely difficult for me to find an appropriate word to express my love for you. My studies have also been a long journey for you, both emotionally and physically. I will never say "I love you" enough.

Shall we all be happy!

Saint Sulpice, 6 September 2019 Viet Anh Nguyen

Abstract

Adversarial learning is an emergent technique that provides better security to machine learning systems by deliberately protecting them against specific vulnerabilities of the learning algorithms. Many adversarial learning problems can be cast equivalently as distributionally robust optimization problems that hedge against the least favorable probability distribution in a certain ambiguity set.

The main objectives of this thesis center around the development of novel analytics toolboxes using advanced probability and statistics machinery under the distributionally robust optimization/adversarial learning framework. Using a type-2 Wasserstein ambiguity set and its Gelbrich hull, which constitutes a conservative outer approximation, we propose new solutions with strong performance guarantees to several problems in statistical learning and risk management, while at the same time mitingating the curse of dimensionality inherent to these problems.

The first chapter proposes a distributionally robust inverse covariance estimator that minimizes the worst-case Stein's loss. The optimal estimator admits a closed-form representation and exhibits many desirable properties, none of which are imposed ad hoc but arise naturally from the distributionally robust optimization approach. The optimal estimator is closely related to a nonlinear eigenvalue shrinkage estimator. For this reason we refer to it as the Wasserstein shrinkage estimator. Furthermore, the Wasserstein shrinkage estimator can also be interpreted as a robust maximum likelihood estimator.

The second chapter proposes a distributionally robust minimum mean square error estimator. Under a mild assumption on the nominal distribution of the uncertain data, we show that the optimal estimator is an affine function of the observations, which can be constructed efficiently using a first-order optimization method to solve the underlying semidefinite program.

The third chapter studies distributionally robust risk measures under the Gelbrich hull ambiguity set, which is an outer approximation of the Wasserstein ambiguity set. We prove that the robustified Gelbrich risk of many popular law-invariant risk measures admit a closed form expression. The result is extended to provide tractable reformulations for the worst-case expected loss as well as the value-at-risk of nonlinear portfolios.

Zusammenfassung

Gegnerisches Lernen (adversarial learning) bezeichnet eine Methodik des Maschinellen Lernens, die gezielt die Schwachstellen der lernenden Systeme absichert. Viele dieser Methoden können als ein robustes Optimierungsproblem bezüglich Verteilungen (distributionally robust optimization problem) formuliert werden, welche sich gegenüber der ungünstigsten Wahrscheinlichkeitsverteilung einer spezifischen Unsicherheitsmenge absichern.

Das Hauptziel dieser Dissertation ist die Entwicklung von neuen analytischen Techniken zur Berechnung und Beschreibung der erwähnten Optimierungsprobleme anhand von statistischen und wahrscheinlichkeitstheoretischen Ansätzen. Wir stellen neue, mit starken Garantien ausgestattete Lösungen zu verschiedenen Problemen des Statistischen Lernens und des Risiko-Managements vor. Diese Lösungen werden mit Hilfe einer Wasserstein typ-2 Unsicherheitsmenge und ihrer sogenannten Gelbrich-Hülle, d.h. einer konservative Approximation der Unsicherheitsmenge, hergeleitet. Gleichzeitig zeigen wir, dass diese Lösungen den sogenannten Fluch der Dimensionen (curse of dimensionality) elegant umgehen.

Im ersten Kapitel beschreiben wir einen gegen die zugrundeliegende Wahrscheinlichkeitsverteilung robusten Schätzer für die inverse Kovarianz, welcher die Steinsche Verlustfunktion im ungünstigsten Fall minimiert. Wir zeigen, dass dieser optimale Schätzer analytisch berechnet warden kann und viele begehrenswerte Eigenschaften besitzt. Diese Eigenschaften entstammen der robusten Betrachtungsweise und werden in keiner Weise als Bedingung vorausgesetzt werden. Der beschriebene optimale Schätzer ist eng verwandt mit dem nichtlinearen Schrumpfschätzer (shrinkage estimator). Deshalb nennen wir ihn Wasserstein Schrumpfschätzer. Der Wasserstein Schrumpfschätzer kann des Weiteren als ein robuster Maximum-Likelihood-Schätzer interpretiert werden.

Das zweite Kapitel befasst sich mit einem robusten Schätzer zur minimalen mittleren quadratischen Abweichung. Unter schwachen Annahmen über die nominale Verteilung unbekannter Daten zeigen wir, dass dass der optimale Schätzer durch eine affine Funktion der Beobachtungen beschrieben wird, welche mit Hilfe einer Optimierungsmethode erster Ordnung für das zugrundeliegende Semidefinite Programm effizient berechnet werden kann.

Das dritte Kapitel untersucht robuste Risikomaße, bezüglich aller Verteilungen in einer Gelbrich-Hülle als Unsicherheitsmenge. Wir beweisen, dass die robuste Version vieler bekannter und verteilungsinvarianter Risikomaße eine geschlossene Darstellung hat. Wir verallgemeinern dieses Ergebnis, um eine effizient lösbare Formulierung des erwarteten Verlustes im ungünstigsten Fall und des “Werts-im-Risiko” (Value-at-Risk) eines nichtlinearen Portfolios herzuleiten.

Contents

List of Figures

List of Tables

Introduction

The beginning is the most important part of the work.
— Plato

Many decision problems in science, engineering and economics are affected by uncertain parameters ξ with probability distribution $\mathbb{P}$. Usually, the complex decision making process is mathematically captured as an optimization model whose aim is to find the best decision which minimizes a certain risk index. The risk index radically depends on the decision chosen and the distribution $\mathbb{P}$ of the random vector ξ. Unfortunately, $\mathbb{P}$ is fundamentally unknown in practice, and we lack an important input parameter for the decision problem. However, $\mathbb{P}$ may be indirectly observable through *training samples* drawn independently from $\mathbb{P}$. In addition, some structural properties of $\mathbb{P}$ may be known. For example, if the random vector represents a vector of uncertain prices, then $\mathbb{P}$ must be supported on the nonnegative orthant. Alternatively, $\mathbb{P}$ may be known to display certain symmetry or unimodality properties, or it may even be known to belong to some parametric distribution family.

If the true distribution $\mathbb{P}$ is unknown, it could be replaced with a *nominal distribution* $\widehat{\mathbb{P}}$ estimated from the training samples in the decision problem. Note that unlike $\mathbb{P}$, the nominal distribution $\widehat{\mathbb{P}}$ is accessible as it is constructed from observable quantities. Therefore, the decision problems under the nominal distribution $\widehat{\mathbb{P}}$ are at least in principle solvable. On the downside, even if the most sophisticated statistical tools are deployed, the nominal distribution $\widehat{\mathbb{P}}$ will invariably differ from the unknown true distribution $\mathbb{P}$ that generated the training samples. Moreover, if $\widehat{\mathbb{P}}$ is used instead of $\mathbb{P}$, the solutions of the decision problem under the nominal distribution $\widehat{\mathbb{P}}$ are likely to inherit any estimation errors in $\widehat{\mathbb{P}}$. In the context of financial portfolio theory it has even been observed that estimation errors in the input parameters of an optimization problem are often amplified by the optimization [27, 116]. To make things worse, one can generally show that even if the distributional input parameters of a decision problem are unbiased, the optimization results tend to be optimistically biased. Thus, implementing the optimal decisions leads to disappointment in out-of-sample tests. In decision analysis this phenomenon is sometimes termed the *optimizer's curse* [157], and in stochastic optimization it is referred to as the *optimization bias* [37, 153].

Ideally, to mitigate this issue, one should construct an estimator $\widehat{\mathbb{P}}$ that is close to the unknown true distribution $\mathbb{P}$ with high confidence. Unfortunately, the accuracy of the nominal distribution $\widehat{\mathbb{P}}$ cannot be increased beyond some fundamental limit by tuning the estimator. The only remaining option to reduce the estimation error is to increase the sample size, which may be expensive or impossible. Indeed, additional training samples may only become available in the future. Thus, the optimizer's curse is fundamental and cannot be eliminated. However, once the potential to improve the estimator $\widehat{\mathbb{P}}$ is exhausted, it may still be possible to alleviate the optimizer's curse by altering the decision problems directly. Specifically, we propose here to robustify these problems against the uncertainty about the true distribution $\mathbb{P}$. Distributional uncertainty is often referred to as *ambiguity* or *Knightian uncertainty* and is conveniently captured by an *ambiguity set*, that is, an uncertainty set in the space of probability distributions.

This thesis consider the distributionally robust decision problem that seeks decisions having minimum risk under the most adverse distributions in the ambiguity set. Intuitively, the distributionally robust decision problem can thus be viewed as a zero-sum game, where the decision-maker first selects a decision with the goal to minimize the risk, in response to which some fictitious adversary or 'nature' selects a distribution from within the ambiguity set with the goal to maximize the risk. The hope is that by minimizing the worst-case risk, we actually push down the risk under *all* distributions in the ambiguity set—in particular under the unknown true distribution $\mathbb{P}$, which is contained in the ambiguity set if the radius is large enough. Thus, there is reason to hope that the solutions of distributionally robust optimization problems with carefully calibrated ambiguity sets display low out-of-sample risk.

Recently, the Wasserstein distance emerges as an appealing option to construct the ambiguity set [100]. The distributionally robust optimization problems in general and distributionally robust optimization problems with Wasserstein ambiguity sets in particular are attractive for a multitude of diverse reasons.

- **Fidelity:** Distributionally robust models are more 'honest' than their nominal counterparts as they acknowledge the presence of distributional uncertainty. They also benefit from information about the type and magnitude of the estimation errors, which is conveniently encoded in the geometry and size of the ambiguity set.
- **Managing expectations:** Due to the optimizer's curse, the solutions of nominal decision

problems equipped with noisy estimators $\widehat{\mathbb{P}}$ display an optimistic in-sample risk, which cannot be realized out of sample. In contrast, the solutions of distributionally robust decision problems are guaranteed to display an out-of-sample risk that falls below the worst-case optimal risk whenever the ambiguity set contains the unknown true distribution. Thus, nominal decision problems over-promise and under-deliver, while distributionally robust decision problems under-promise and over-deliver.

- **Computational tractability:** The distributionally robust problem can often be reformulated as (or tightly approximated by) finite convex programs that are solvable in polynomial time using off-the-shelf numerical solvers.

- **Performance guarantees:** For judiciously calibrated ambiguity sets, it can be proven that the worst-case optimal risk for any fixed sample size provides an upper confidence bound on the out-of-sample risk attained by the optimizers of the distributionally robust decision problem (finite sample guarantee) and that the optimizers of the distributionally robust decision problem converge almost surely to an optimizer with perfect knowledge of $\mathbb{P}$ as the number of training samples tends to infinity (asymptotic guarantee).

- **Regularization by robustification:** The optimizer's curse is reminiscent of overfitting phenomena that plague most statistical learning models. One can show that distributionally robust learning models equipped with a Wasserstein ambiguity set are often equivalent to regularized learning models that minimize the sum of a nominal objective and a norm term that penalizes hypothesis complexity. Similarly, one can show that some distributionally robust maximum likelihood estimation models produce shrinkage estimators. Thus, Wasserstein distributional robustness offers new probabilistic interpretations for popular regularization techniques. The empirical success of regularization methods in statistics fuels hope that Wasserstein distributionally robust models can effectively combat the optimizer's curse across many application areas.

- **Anticipating black swans:** If uncertainty is modeled by the empirical distribution, then the nominal decision problem evaluates the admissible loss functions only at the training samples. However, possible future uncertainty realizations that differ from all training samples but could have devastating consequences ('black swans') are ignored. If the empirical distribution may be perturbed within a Wasserstein ball with a positive radius, on the other hand, then (possibly small amounts of) probability mass can be moved anywhere in the support set of the random vector. Thus, the Wasserstein distributionally robust decision problem faithfully anticipates the possibility of black swans. We emphasize that all distributions in a Kullback-Leibler divergence ball must be absolutely continuous with respect to the nominal distribution, which implies that the corresponding distributionally robust decision problems ignore the possibility of black swans.

- **Axiomatic justification:** Adopting a distributionally robust approach when facing uncertainty can be justified axiomatically. Under mild technical conditions, the decisions must be ranked by their corresponding worst-case risk over the ambiguity set of $\mathbb{P}$ [38, Theorem 12].

- **Optimality principle:** Data-driven optimization aims to use the training data directly to construct an estimator for the risk index under the unknown true distribution $\mathbb{P}$ (a predictor) and a decision that minimizes this predictor (a prescriptor) without the detour of constructing an estimator for $\mathbb{P}$. It has been shown that optimal predictors and the corresponding prescriptors can be constructed by solving a meta-optimization model that minimizes the in-sample risk of the predictor-prescriptor pairs subject to constraints guaranteeing that the in-sample risk is actually attainable out of sample. It has been shown that this meta-optimization problem admits a unique solution: the best predictor-prescriptor pair is obtained by solving a distributionally robust optimization problem over all distributions in some neighborhood of the empirical distribution [132, Theorem 7]. Thus, if one aims to transform training data to decisions, it is in some precise sense optimal to do this by solving a data-driven distributionally robust optimization problem.

Distributionally robust optimization models with Wasserstein ambiguity sets were introduced in [137]. Reformulations of these models as nonconvex optimization problems as well as initial attempts to solve these problems via algorithms from global optimization are reported in [175] and [136, § 7.1]. Convex reformulations and approximations were discovered in [118, 178] and significantly generalized in [16, 68].

Ideas from distributionally robust optimization also permeate several other areas of statistics and machine learning, ranging from hypothesis testing [69], inverse optimization [119] to classification and regression [151, 150]. At the technical level, distributionally robust optimization is the backbone of various adversarial learning techniques such as generative adversarial networks [5, 76] and auto-encoders [164].

Contributions and Structure of the Thesis

The main contributions of this thesis are divided into three self-contained chapters organized in the chronological order of discovery.

In Chapter 1 we revisit the problem of estimating the inverse covariance matrix of a random vector from i.i.d. sampled data from the lens of distributionally robust optimization. Using a coherent robustification of the maximum likelihood estimation problem for Gaussian random vector using the Wasserstein type-2 ambiguity set, we will show that the optimal robust estimator is a nonlinear shrinkage estimator. We also develop a Newton-type numerical method to solve the robust maximum likelihood estimation problem when there are conditional independence constraints in the estimation process. The content of this chapter is condensed in the following paper.

(i) V.A. Nguyen, D. Kuhn, and P. Mohajerin Esfahani. *Distributionally Robust Inverse Covariance Estimation: The Wasserstein Shrinkage Estimator.* Minor revision at Operations Research - Resubmitted.

In Chapter 2 we revisit the fundamental problem of minimum mean square error estimation. We show that if the nominal distribution is a Gaussian distribution, the optimal estimator is an affine function of the observations. Moreover, we prove that the optimal estimator can be recovered from the optimal solution of a linear semidefinite program. The result can be further extended to the case of elliptical nominal distribution. For large scale estimation problem we develop a first-order Frank-Wolfe method which enjoys linear convergence guarantees. The content of this chapter is presented in the following paper.

(ii) V.A. Nguyen, S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani. *Bridging Bayesian and Minimax Mean Square Error Estimation via Wasserstein Distributionally Robust Optimization*. Working paper.

In Chapter 3 we introduce the Gelbrich hull that exploits only the information about the first two moments of the nominal distribution and it is provably the superset of the Wasserstein type-2 ambiguity set. We propose a principled approach to construct the Gelbrich risk which conservatively approximates the risk under the Wasserstein ambiguity set. We show that the Gelbrich risk can be reformulated as a finite convex optimization, and in specific case, admits an analytical expression. The content of this chapter is exhibited in the following paper.

(iii) V.A. Nguyen, S. Shafieezadeh-Abadeh, D. Filipović and D. Kuhn. *Distributionally Robust Risk Measures with Structured Ambiguity Sets*. Working paper.

Statement of Originality

I hereby certify that this thesis is the result of my own work, where some parts are the result of collaborations with my supervisor Dr. Daniel Kuhn, my supervisor Dr. Peyman Mohajerin Esfahani, as well as my scientific collaborators: Dr. Damir Filipović and Soroosh Shafieezadeh-Abadeh. No other person's work has been used without due acknowledgment.

The enchanting charms of this sublime science reveal themselves in all their beauty only to those who have the courage to go deeply into it.

— Carl Friedrich Gauss

1 Distributionally Robust Inverse Covariance Estimation: The Wasserstein Shrinkage Estimator

> It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.
>
> — Albert Einstein

We introduce a distributionally robust maximum likelihood estimation model with a Wasserstein ambiguity set to infer the inverse covariance matrix of a p-dimensional Gaussian random vector from n independent samples. The proposed model minimizes the worst case (maximum) of Stein's loss across all normal reference distributions within a prescribed Wasserstein distance from the normal distribution characterized by the sample mean and the sample covariance matrix. We prove that this estimation problem is equivalent to a semidefinite program that is tractable in theory but beyond the reach of general purpose solvers for practically relevant problem dimensions p. In the absence of any prior structural information, the estimation problem has an analytical solution that is naturally interpreted as a nonlinear shrinkage estimator. Besides being invertible and well-conditioned even for $p > n$, the new shrinkage estimator is rotation-equivariant and preserves the order of the eigenvalues of the sample covariance matrix. These desirable properties are not imposed *ad hoc* but emerge naturally from the underlying distributionally robust optimization model. Finally, we develop a sequential quadratic approximation algorithm for efficiently solving the general estimation problem subject to conditional independence constraints typically encountered in Gaussian graphical models.

1.1 Introduction

The covariance matrix $\Sigma \triangleq \mathbb{E}_{\mathbb{P}}[(\xi - \mathbb{E}_{\mathbb{P}}[\xi])(\xi - \mathbb{E}_{\mathbb{P}}[\xi])^\top]$ of a random vector $\xi \in \mathbb{R}^p$ governed by a distribution $\mathbb{P}$ collects basic information about the spreads of all individual components and the linear dependencies among all pairs of components of ξ. The inverse Σ^{-1} of the covariance matrix is called the *precision matrix*. This terminology captures the intuition that a large spread reflects a low precision and vice versa. While the covariance matrix appears in the *formulations* of many problems in engineering, science and economics, it is often the precision matrix that emerges in their *solutions*. For example, the optimal classification rule in linear discriminant analysis [60], the optimal investment portfolio in Markowitz' celebrated mean-variance model [115] or the optimal array vector of the beamforming problem in signal processing [48] all depend on the precision matrix. Moreover, the optimal fingerprint method used to detect a multivariate climate change signal blurred by weather noise requires knowledge of the climate vector's precision matrix [142].

1.1.1 Background on Precision Matrix Estimation

If the distribution $\mathbb{P}$ of ξ is known, then the covariance matrix Σ and the precision matrix Σ^{-1} can at least principally be calculated in closed form. In practice, however, $\mathbb{P}$ is never known and only indirectly observable through n independent training samples $\widehat{\xi}_1, \dots, \widehat{\xi}_n$ from $\mathbb{P}$. In this setting, Σ and Σ^{-1} need to be estimated from the training data. Arguably the simplest estimator for Σ is the sample covariance matrix $\widehat{\Sigma} \triangleq \frac{1}{n}\sum_{i=1}^n (\widehat{\xi}_i - \widehat{\mu})(\widehat{\xi}_i - \widehat{\mu})^\top$, where $\widehat{\mu} \triangleq \frac{1}{n}\sum_{i=1}^n \widehat{\xi}_i$ stands for the sample mean. Note that $\widehat{\mu}$ and $\widehat{\Sigma}$ simply represent the actual mean and covariance matrix of the uniform distribution on the training samples. For later convenience, $\widehat{\Sigma}$ is defined here without Bessel's correction and thus constitutes a biased estimator.[1] Moreover, as a sum of n rank-1 matrices, $\widehat{\Sigma}$ is rank deficient in the big data regime ($p > n$). In this case, $\widehat{\Sigma}$ cannot be inverted to obtain a precision matrix estimator, which is often the actual quantity of interest.

If ξ follows a normal distribution with unknown mean μ and precision matrix $X \succ 0$, which we will assume throughout the rest of the paper, then the log-likelihood function of the training data can be expressed as

$$\begin{aligned}\widehat{\mathscr{L}}(\mu, X) &\triangleq -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log\det X - \frac{1}{2}\sum_{i=1}^n (\widehat{\xi}_i - \mu)^\top X (\widehat{\xi}_i - \mu) \\ &= -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log\det X - \frac{n}{2}\operatorname{Tr}\left[\widehat{\Sigma}X\right] - \frac{n}{2}(\widehat{\mu} - \mu)^\top X (\widehat{\mu} - \mu). \qquad (1.1)\end{aligned}$$

Note that $\widehat{\mathscr{L}}(\mu, X)$ is strictly concave in μ and X [18, Chapter 7] and depends on the training samples only through the sample mean and the sample covariance matrix. It is clear from the last expression that $\widehat{\mathscr{L}}(\mu, X)$ is maximized by $\mu^\star = \widehat{\mu}$ for any fixed X. The maximum likelihood estimator $X^\star$ for the precision matrix is thus obtained by maximizing $\widehat{\mathscr{L}}(\widehat{\mu}, X)$ over all $X \succ 0$,

[1] An elementary calculation shows that $\mathbb{E}_{\mathbb{P}^n}[\widehat{\Sigma}] = \frac{n-1}{n}\Sigma$.

which is tantamount to solving the convex program

$$\inf_{X \succ 0} -\log\det X + \mathrm{Tr}\left[\widehat{\Sigma} X\right]. \tag{1.2}$$

If $\widehat{\Sigma}$ is rank deficient, which necessarily happens for $p > n$, then problem (1.2) is unbounded. Indeed, expressing the sample covariance matrix as $\widehat{\Sigma} = R\Lambda R^\top$ with R orthogonal and $\Lambda \succeq 0$ diagonal, we may set $X_k = R\Lambda_k R^\top$ for any $k \in \mathbb{N}$, where $\Lambda_k \succ 0$ is the diagonal matrix with $(\Lambda_k)_{ii} = 1$ if $\lambda_i > 0$ and $(\Lambda_k)_{ii} = k$ if $\lambda_i = 0$. By construction, the objective value of X_k in (1.2) tends to $-\infty$ as k grows. If $\widehat{\Sigma}$ is invertible, on the other hand, then the first-order optimality conditions can be solved analytically, showing that the minimum of problem (1.2) is attained at $X^\star = \widehat{\Sigma}^{-1}$. This implies that maximum likelihood estimation under normality simply recovers the sample covariance matrix but fails to yield a precision matrix estimator for $p > n$.

Adding an ℓ_1-regularization term to its objective function guarantees that problem (1.2) has a unique minimizer $X^\star \succ 0$, which constitutes a proper (invertible) precision matrix estimator [84]. Moreover, as the ℓ_1-norm represents the convex envelope of the cardinality function on the unit hypercube, the ℓ_1-norm regularized maximum likelihood estimation problem promotes sparse precision matrices that encode interpretable Gaussian graphical models [6, 65]. Indeed, under the given normality assumption one can show that $X_{ij} = 0$ if and only if the random variables ξ_i and ξ_j are conditionally independent given $\{\xi_k\}_{k \notin \{i,j\}}$ [102]. The sparsity pattern of the precision matrix X thus captures the conditional independence structure of ξ.

In theory, the ℓ_1-norm regularized maximum likelihood estimation problem can be solved in polynomial time via modern interior point algorithms. In practice, however, scalability to high dimensions remains challenging due to the problem's semidefinite nature, and larger problem instances must be addressed with special-purpose methods such as the Newton-type QUIC algorithm [84].

Instead of penalizing the ℓ_1-norm of the precision matrix, one may alternatively penalize its inverse X^{-1} with the goal of promoting sparsity in the covariance matrix and thus controlling the *marginal* independence structure of ξ [15]. Despite its attractive statistical properties, this alternative model leads to a hard non-convex and non-smooth optimization problem, which can only be solved approximately.

By the Fisher-Neyman factorization theorem, $\widehat{\Sigma}$ is a sufficient statistic for the true covariance matrix Σ of a normally distributed random vector, that is, $\widehat{\Sigma}$ contains the same information about Σ as the entire training dataset. Without any loss of generality, we may thus focus on estimators that depend on the data only through $\widehat{\Sigma}$. If neither the covariance matrix Σ nor the precision matrix Σ^{-1} are known to be sparse and if there is no prior information about the orientation of their eigenvectors, it is reasonable to restrict attention to *rotation equivariant* estimators. A precision matrix estimator $\widehat{X}(\widehat{\Sigma})$ is called rotation equivariant if $\widehat{X}(R\widehat{\Sigma}R^\top) = R\widehat{X}(\widehat{\Sigma})R^\top$ for any rotation matrix R. This definition requires that the estimator for the rotated data coincides with the rotated estimator for the original data. One can show that rotation equivariant estimators have the same eigenvectors as the sample covariance matrix (see, *e.g.*, [134, Lemma 5.3] for a simple proof) and are thus uniquely determined by their eigenvalues. Hence, imposing rotation equivariance reduces the degrees of freedom from $p(p+1)/2$ to p. Using an entropy loss function introduced in [90], Stein was the first to demonstrate that superior covariance estimators in the sense of statistical decision theory can be constructed by shrinking the eigenvalues of the sample covariance matrix [159, 160]. Unfortunately, his proposed shrinkage transformation may alter the order of the eigenvalues and even undermine the positive semidefiniteness of the resulting estimator when $p > n$, which necessitates an *ad hoc* correction step involving an isotonic regression. Various refinements of this approach are reported in [43, 79, 176] and the references therein, but most of these works focus on the low-dimensional case when $n \geq p$.

Jensen's inequality suggests that the largest (smallest) eigenvalue of the sample covariance matrix $\widehat{\Sigma}$ is biased upwards (downwards), which implies that $\widehat{\Sigma}$ tends to be ill-conditioned [169]. This effect is most pronounced for $\Sigma \approx I$. A promising shrinkage estimator for the covariance matrix is thus obtained by forming a convex combination of the sample covariance matrix and the identity matrix scaled by the average of the sample eigenvalues [105]. If its convex weights are chosen optimally in view of the Frobenius risk, the resulting shrinkage estimator can be shown to be both well-conditioned and more accurate than $\widehat{\Sigma}$. Alternative shrinkage targets include the constant correlation model, which preserves the sample variances but equalizes all pairwise correlations [104], the single index model, which assumes that each random variable is explained by one systematic and one idiosyncratic risk factor [103], or the diagonal matrix of the sample eigenvalues [166] etc.

The *linear* shrinkage estimators described above are computationally attractive because evaluating convex combinations is cheap. Computing the corresponding precision matrix estimators requires a matrix inversion and is therefore more expensive. We emphasize that linear shrinkage estimators for the precision matrix itself, obtained by forming a cheap convex combination of the inverse sample covariance matrix and a shrinkage target, are not available in the big data regime when $p > n$ and $\widehat{\Sigma}$ fails to be invertible.

More recently, insights from random matrix theory have motivated a new rotation equivariant shrinkage estimator that applies an individualized shrinkage intensity to every sample eigenvalue [106]. While this *nonlinear* shrinkage estimator offers significant improvements over linear shrinkage, its evaluation necessitates the solution of a hard nonconvex optimization problem, which becomes cumbersome for large values of p. Alternative nonlinear shrinkage estimators can be obtained by imposing an upper bound on the condition number of the covariance matrix in the underlying maximum likelihood estimation problem [173].

Alternatively, multi-factor models familiar from the arbitrage pricing theory can be used to approximate the covariance matrix by a sum of a low-rank and a diagonal component, both of which have only few free parameters and are thus easier to estimate. Such a dimensionality reduction leads to stable estimators [28, 58].

1.1.2 Problem Statement and Contributions

This paper endeavors to develop a principled approach to precision matrix estimation, which is inspired by recent advances in distributionally robust optimization [39, 73, 172]. For the sake of argument, assume that the true distribution of ξ is given by $\mathbb{P} = \mathcal{N}(\mu_0, \Sigma_0)$, where $\Sigma_0 \succ 0$. If μ_0 and Σ_0 were known, the quality of some estimators μ and X for μ_0 and Σ_0^{-1}, respectively, could conveniently be measured by Stein's loss [90]

$$\begin{aligned} L(X,\mu) &\triangleq -\log\det(\Sigma_0 X) + \operatorname{Tr}\left[\Sigma_0 X\right] + (\mu_0-\mu)^\top X(\mu_0-\mu) - p \\ &= -\log\det X + \mathbb{E}_{\mathbb{P}}\left[(\xi-\mu)^\top X(\xi-\mu)\right] - \log\det\Sigma_0 - p, \end{aligned} \tag{1.3}$$

which is reminiscent of the log-likelihood function (1.1). It is easy to verify that Stein's loss is nonnegative for all $\mu \in \mathbb{R}^p$ and $X \in \mathbb{S}_+^p$ and vanishes only at the true mean $\mu = \mu_0$ and the true precision matrix $X = \Sigma_0^{-1}$. Of course, we cannot minimize Stein's loss directly because $\mathbb{P}$ is unknown. As a naïve remedy, one could instead minimize an approximation of Stein's loss obtained by removing the (unknown but irrelevant) normalization constant $-\log\det\Sigma_0 - p$ and replacing $\mathbb{P}$ in (1.3) with the empirical distribution $\widehat{\mathbb{P}}_n = \mathcal{N}(\widehat{\mu}, \widehat{\Sigma})$. However, in doing so we simply recover the standard maximum likelihood estimation problem, which is unbounded for $p > n$ and outputs the sample mean and the inverse sample covariance matrix for $p \le n$. This motivates us to robustify the empirical loss minimization problem by exploiting that $\widehat{\mathbb{P}}_n$ is close to $\mathbb{P}$ in Wasserstein distance.

Definition 1.1 (Wasserstein distance)**.** *The type-2 Wasserstein distance between two arbitrary distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ on $\mathbb{R}^p$ with finite second moments is defined as*

$$\mathbb{W}(\mathbb{P}_1,\mathbb{P}_2) \triangleq \inf_{\Pi}\left\{\left(\int_{\mathbb{R}^p\times\mathbb{R}^p} \left\|\xi_1-\xi_2\right\|^2 \Pi(\mathrm{d}\xi_1,\mathrm{d}\xi_2)\right)^{\frac{1}{2}} : \begin{array}{l} \Pi \textit{ is a joint distribution of } \xi_1 \textit{ and } \xi_2 \\ \textit{with marginals } \mathbb{P}_1 \textit{ and } \mathbb{P}_2\textit{, respectively} \end{array}\right\}.$$

The squared Wasserstein distance between $\mathbb{P}_1$ and $\mathbb{P}_2$ can be interpreted as the cost of moving the distribution $\mathbb{P}_1$ to the distribution $\mathbb{P}_2$, where $\|\xi_1 - \xi_2\|^2$ quantifies the cost of moving unit mass from ξ_1 to ξ_2.

A central limit type theorem for the Wasserstein distance between empirical normal distributions implies that $n \cdot \mathbb{W}(\widehat{\mathbb{P}}_n, \mathbb{P})^2$ converges weakly to a quadratic functional of independent normal random variables as the number n of training samples tends to infinity [143, Theorem 2.3]. We may thus conclude that for every $\eta \in (0,1)$ there exists $q(\eta) > 0$ such that $\mathbb{P}^n[\mathbb{W}(\widehat{\mathbb{P}}_n, \mathbb{P}) \le q(\eta) n^{-\frac{1}{2}}] \ge 1-\eta$ for all n large enough. In the following we denote by $\mathcal{N}^p$ the family of all normal distributions on $\mathbb{R}^p$ and by

$$\mathbb{B}_\rho = \{\mathbb{Q} \in \mathcal{N}^p : \mathbb{W}(\mathbb{Q}, \widehat{\mathbb{P}}_n) \le \rho\}$$

the ambiguity set of all normal distributions whose Wasserstein distance to $\widehat{\mathbb{P}}_n$ is at most $\rho \ge 0$. Note that $\mathbb{B}_\rho$ depends on the unknown true distribution $\mathbb{P}$ only through the training data and, for $\rho \ge q(\eta) n^{-\frac{1}{2}}$, contains $\mathbb{P}$ with confidence $1-\eta$ asymptotically as n tends to infinity. It is thus natural to formulate a *distributionally robust* estimation problem for the precision matrix that minimizes Stein's loss—modulo an irrelevant normalization constant—in the worst case across all reference distributions $\mathbb{Q} \in \mathbb{B}_\rho$.

$$\mathscr{J}(\widehat{\mu}, \widehat{\Sigma}) \triangleq \inf_{\mu \in \mathbb{R}^p,\, X \in \mathscr{X}} \left\{ -\log\det X + \sup_{\mathbb{Q} \in \mathbb{B}_\rho} \mathbb{E}_{\mathbb{Q}} \left[(\xi - \mu)^\top X (\xi - \mu) \right] \right\} \tag{1.4}$$

Here, $\mathscr{X} \subseteq \mathbb{S}^p_{++}$ denotes the set of admissible precision matrices. In the absence of any prior structural information, the only requirement is that X be positive semidefinite and invertible, in which case $\mathscr{X} = \mathbb{S}^p_{++}$. Known conditional independence relationships impose a sparsity pattern on X, which is easily enforced through linear equality constraints in $\mathscr{X}$. By adopting a worst-case perspective, we hope that the minimizers of (1.4) will have low Stein's loss with respect to all distributions in $\mathbb{B}_\rho$ including the unknown true distribution $\mathbb{P}$. As Stein's loss with respect to the empirical distribution is proportional to the log-likelihood function (1.1), problem (1.4) can also be interpreted as a robust maximum likelihood estimation problem that hedges against perturbations in the training samples. As we will show below, this robustification is tractable and has a regularizing effect.

Recently it has been discovered that distributionally robust optimization models with Wasserstein ambiguity sets centered at *discrete* distributions on $\mathbb{R}^p$ (and *without* any normality restrictions) are often equivalent to tractable convex programs [118, 178]. Extensions of these results to general Polish spaces are reported in [16, 68]. The explicit convex reformulations of Wasserstein distributionally robust models have not only facilitated efficient solution procedures but have also revealed insightful connections between distributional robustness and regularization in machine learning. Indeed, many classical regularization schemes of supervised learning such as the Lasso method can be explained by a Wasserstein distributionally robust model. This link was first discovered in the context of logistic regression [151] and later extended to other popular regression and classification models [16, 150] and even to generative adversarial networks in deep learning [67].

Model (1.4) differs fundamentally from all existing distributionally robust optimization models in that the ambiguity set contains only normal distributions. As the family of normal distributions fails to be closed under mixtures, the ambiguity set is thus nonconvex. In the remainder of the paper we devise efficient solution methods for problem (1.4), and we investigate the properties of the resulting precision matrix estimator.

The main contributions of this paper can be summarized as follows.

- Leveraging an analytical formula for the Wasserstein distance between two normal distributions derived in [72], we prove that the distributionally robust estimation problem (1.4) is equivalent to a tractable semidefinite program—despite the nonconvex nature of the underlying ambiguity set.
- We prove that problem (1.4) and its unique minimizer depend on the training data only

through $\widehat{\Sigma}$ (but not through $\widehat{\mu}$), which is reassuring because $\widehat{\Sigma}$ is a sufficient statistic for the precision matrix.

- In the absence of any structural information, we demonstrate that problem (1.4) has an analytical solution that is naturally interpreted as a nonlinear shrinkage estimator. Indeed, the optimal precision matrix estimator shares the eigenvectors of the sample covariance matrix, and as the radius ρ of the Wasserstein ambiguity set grows, its eigenvalues are shrunk towards 0 while preserving their order. At the same time, the condition number of the optimal estimator steadily improves and eventually converges to 1 even for $p > n$. These desirable properties are not enforced *ex ante* but emerge naturally from the underlying distributionally robust optimization model.

- In the presence of conditional independence constraints, the semidefinite program equivalent to (1.4) is beyond the reach of general purpose solvers for practically relevant problem dimensions p. We thus devise an efficient sequential quadratic approximation method reminiscent of the QUIC algorithm [84], which can solve instances of problem (1.4) with $p \lesssim 10^4$ on a standard PC.

- We derive an analytical formula for the extremal distribution that attains the supremum in (1.4).

An important aspect of the distributionally robust estimation problem (1.4) is the choice of the radius $\rho \geq 0$ of the ambiguity set $\mathbb{B}_\rho$. Ideally, this hyperparameter should be tuned so as to minimize the distance between the precision matrix estimator $X^\star(\rho)$ that solves (1.4) and the unknown true precision matrix Σ^{-1}. While this paper was under review, Blanchet and Si managed to prove that if distances in $\mathbb{S}_+^p$ are measured via Stein's loss function and there are no conditional independence constraints, then the Wasserstein radius that minimizes the *expected* distance between the estimator $X^\star(\rho)$ and the true precision matrix Σ^{-1} scales linearly with the sample size as n^{-1}, where the proportionality constant is a function of the true covariance matrix that is known in closed form [17, Theorem 1]. This result is surprising vis-à-vis the central limit type theorem [143, Theorem 2.3], which suggests a canonical square root scaling of the form $n^{-\frac{1}{2}}$. In practice, ρ should be calibrated in view of the training samples $\widehat{\xi}_1, \ldots, \widehat{\xi}_n$, for example via cross-validation using application-specific performance measures. Concrete examples of different cross-validation schemes are described in Section 1.6.

The paper is structured as follows. Section 1.2 demonstrates that the distributionally robust estimation problem (1.4) admits an exact reformulation as a tractable semidefinite program. Section 1.3 derives an analytical solution of this semidefinite program in the absence of any structural information, while Section 1.4 develops an efficient sequential quadratic approximation algorithm for the problem with conditional independence constraints. The extremal distribution that attains the worst-case expectation in (1.4) is characterized in Section 1.5, and numerical experiments based on synthetic and real data are reported in Section 1.6.

Notation. For any $A \in \mathbb{R}^{p\times p}$ we use $\mathrm{Tr}\left[A\right]$ to denote the trace and $\|A\| = \sqrt{\mathrm{Tr}\left[A^\top A\right]}$ to denote the Frobenius norm of A. By slight abuse of notation, the Euclidean norm of $v \in \mathbb{R}^p$ is also denoted by $\|v\|$. Moreover, I stands for the identity matrix. Its dimension is usually evident from the context. For any $A, B \in \mathbb{R}^{p\times p}$, we use $\langle A, B\rangle = \mathrm{Tr}\left[A^\top B\right]$ to denote the inner product and $A \otimes B \in \mathbb{R}^{p^2\times p^2}$ to denote the Kronecker product of A and B. The space of all symmetric matrices in $\mathbb{R}^{p\times p}$ is denoted by $\mathbb{S}^p$. We use $\mathbb{S}^p_+$ ($\mathbb{S}^p_{++}$) to represent the cone of symmetric positive semidefinite (positive definite) matrices in $\mathbb{S}^p$. For any $A, B \in \mathbb{S}^p$, the relation $A \succeq B$ ($A \succ B$) means that $A - B \in \mathbb{S}^p_+$ ($A - B \in \mathbb{S}^p_{++}$).

1.2 Tractable Reformulation

Throughout this paper we assume that the random vector $\xi \in \mathbb{R}^p$ is normally distributed. This is in line with the common practice in statistics and in the natural and social sciences, whereby normal distributions are routinely used to model random vectors whose distributions are unknown. The normality assumption is often justified by the central limit theorem, which suggests that random vectors influenced by many small and unrelated disturbances are approximately normally distributed. Moreover, the normal distribution maximizes entropy across all distributions with given first- and second-order moments, and as such it constitutes the least prejudiced distribution compatible with a given mean vector and covariance matrix.

1.2.1 Preliminaries

In order to facilitate rigorous statements, we first provide a formal definition of normal distributions.

Definition 1.2 (Normal distributions)**.** *We say that $\mathbb{P}$ is a normal distribution on $\mathbb{R}^p$ with mean $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{S}^p_+$, that is, $\mathbb{P} = \mathcal{N}(\mu, \Sigma)$, if $\mathbb{P}$ is supported on* $\mathrm{supp}(\mathbb{P}) = \{\mu + Ev : v \in \mathbb{R}^k\}$, *and if the density function of $\mathbb{P}$ with respect to the Lebesgue measure on* $\mathrm{supp}(\mathbb{P})$ *is given by*

$$\varrho_{\mathbb{P}}(\xi) \triangleq \frac{1}{\sqrt{(2\pi)^k \det(D)}} e^{-(\xi-\mu)^\top E D^{-1} E^\top (\xi-\mu)},$$

where $k = \mathrm{rank}(\Sigma)$, $D \in \mathbb{S}^k_{++}$ is the diagonal matrix of the positive eigenvalues of Σ, and $E \in \mathbb{R}^{p\times k}$ is the matrix whose columns correspond to the orthonormal eigenvectors of the positive eigenvalues of Σ. The family of all normal distributions on $\mathbb{R}^p$ is denoted by $\mathcal{N}^p$, while the subfamily of all distributions in $\mathcal{N}^p$ with zero means and arbitrary covariance matrices is denoted by $\mathcal{N}^p_0$.

Definition 1.2 explicitly allows for degenerate normal distributions with rank deficient covariance matrices.

The normality assumption also has distinct computational advantages. In fact, while the Wasserstein distance between two generic distributions is only given implicitly as the solution of a mass transportation problem, the Wasserstein distance between two normal distributions is known in closed form. It can be expressed explicitly as a function of the mean vectors and covariance matrices of the two distributions.

Proposition 1.3 (Givens and Shortt [72, Proposition 7])**.** *The type-2 Wasserstein distance between two normal distributions $\mathbb{P}_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $\mathbb{P}_2 = \mathcal{N}(\mu_2, \Sigma_2)$ with $\mu_1, \mu_2 \in \mathbb{R}^p$ and $\Sigma_1, \Sigma_2 \in \mathbb{S}_+^p$ amounts to*

$$\mathbb{W}(\mathbb{P}_1, \mathbb{P}_2) = \sqrt{\left\|\mu_1 - \mu_2\right\|^2 + \mathrm{Tr}\left[\Sigma_1\right] + \mathrm{Tr}\left[\Sigma_2\right] - 2\,\mathrm{Tr}\left[\sqrt{\sqrt{\Sigma_2}\Sigma_1\sqrt{\Sigma_2}}\right]}.$$

If $\mathbb{P}_1$ and $\mathbb{P}_2$ share the same mean vector (*e.g.*, if $\mu_1 = \mu_2 = 0$), then the Wasserstein distance $\mathbb{W}(\mathbb{P}_1, \mathbb{P}_2)$ reduces to a function of the covariance matrices Σ_1 and Σ_2 only, thereby inducing a metric on the cone $\mathbb{S}_+^p$.

Definition 1.4 (Induced metric on $\mathbb{S}_+^p$)**.** *Let $\mathbb{W}_S : \mathbb{S}_+^p \times \mathbb{S}_+^p \to \mathbb{R}_+$ be the metric on $\mathbb{S}_+^p$ induced by the type-2 Wasserstein metric on the family of normal distributions with equal means. Thus, for all $\Sigma_1, \Sigma_2 \in \mathbb{S}_+^p$ we set*

$$\mathbb{W}_S(\Sigma_1, \Sigma_2) \triangleq \sqrt{\mathrm{Tr}\left[\Sigma_1\right] + \mathrm{Tr}\left[\Sigma_2\right] - 2\,\mathrm{Tr}\left[\sqrt{\sqrt{\Sigma_2}\Sigma_1\sqrt{\Sigma_2}}\right]}.$$

The definition of $\mathbb{W}_S$ implies via Proposition 1.3 that $\mathbb{W}(\mathbb{P}_1, \mathbb{P}_2) = \mathbb{W}_S(\Sigma_1, \Sigma_2)$ for all $\mathbb{P}_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $\mathbb{P}_2 = \mathcal{N}(\mu_2, \Sigma_2)$ with $\mu_1 = \mu_2$. Thanks to its interpretation as the restriction of $\mathbb{W}$ to the space of normal distributions with a fixed mean, it is easy to verify that $\mathbb{W}_S$ is symmetric and positive definite and satisfies the triangle inequality. In other words, $\mathbb{W}_S$ inherits the property of being a metric from $\mathbb{W}$.

Corollary 1.5 (Commuting covariance matrices)**.** *If $\Sigma_1, \Sigma_2 \in \mathbb{S}_+^p$ commute ($\Sigma_1\Sigma_2 = \Sigma_2\Sigma_1$), then the induced Wasserstein distance $\mathbb{W}_S$ simplifies to the trace norm between the square roots of Σ_1 and Σ_2, that is,*

$$\mathbb{W}_S(\Sigma_1, \Sigma_2) = \left\|\sqrt{\Sigma_1} - \sqrt{\Sigma_2}\right\|.$$

Proof. The commutativity of Σ_1 and Σ_2 implies that $\sqrt{\Sigma_2}\Sigma_1\sqrt{\Sigma_2} = \Sigma_1\Sigma_2$, whereby

$$\mathbb{W}_S(\Sigma_1, \Sigma_2) = \sqrt{\mathrm{Tr}\left[\Sigma_1\right] + \mathrm{Tr}\left[\Sigma_2\right] - 2\,\mathrm{Tr}\left[\sqrt{\Sigma_1\Sigma_2}\right]} = \sqrt{\mathrm{Tr}\left[\left(\sqrt{\Sigma_1} - \sqrt{\Sigma_2}\right)^2\right]} = \left\|\sqrt{\Sigma_1} - \sqrt{\Sigma_2}\right\|.$$

Thus, the claim follows. □

Proposition 1.3 reveals that the Wasserstein distance between any two (possibly degenerate) normal distributions is finite. In contrast, the Kullback-Leibler divergence between degenerate and non-degenerate normal distributions is infinite.

Remark 1.6 (Kullback-Leibler divergence between normal distributions)**.** *A simple calculation* *shows that the Kullback-Leibler divergence from* $\mathbb{P}_2 = \mathcal{N}(\mu_2, \Sigma_2)$ *to* $\mathbb{P}_1 = \mathcal{N}(\mu_1, \Sigma_1)$ *amounts to*

$$D_{\mathrm{KL}}(\mathbb{P}_1 \| \mathbb{P}_2) = \frac{1}{2}\left[(\mu_2 - \mu_1)^\top \Sigma_2^{-1}(\mu_2 - \mu_1) + \mathrm{Tr}\left[\Sigma_1 \Sigma_2^{-1}\right] - p - \log\det \Sigma_1 + \log\det \Sigma_2\right]$$

whenever $\mu_1, \mu_2 \in \mathbb{R}^p$ *and* $\Sigma_1, \Sigma_2 \in \mathbb{S}^p_{++}$. *If either* $\mathbb{P}_1$ *or* $\mathbb{P}_2$ *is degenerate (that is, if* Σ_1 *is singular and* Σ_2 *invertible or vice versa), then* $\mathbb{P}_1$ *fails to be absolutely continuous with respect to* $\mathbb{P}_2$*, which implies that* $D_{\mathrm{KL}}(\mathbb{P}_1 \| \mathbb{P}_2) = \infty$. *Moreover, from the above formula it is easy to verify that* $D_{\mathrm{KL}}(\mathbb{P}_1 \| \mathbb{P}_2)$ *diverges if either* Σ_1 *or* Σ_2 *tends to a singular matrix.*

In the big data regime ($p > n$) the sample covariance matrix $\widehat{\Sigma}$ is singular even if the samples are drawn from a non-degenerate normal distribution $\mathbb{P} = \mathcal{N}(\mu, \Sigma)$ with $\Sigma \in \mathbb{S}^p_{++}$. In this case, the Kullback-Leibler distance between the empirical distribution $\widehat{\mathbb{P}} = \mathcal{N}(\widehat{\mu}, \widehat{\Sigma})$ and $\mathbb{P}$ is infinite, and thus $\widehat{\mathbb{P}}$ and $\mathbb{P}$ are perceived as maximally dissimilar despite their intimate relation. In contrast, their Wasserstein distance is finite.

1.2.2 Precision Matrix Estimation when the Mean Vector is Known

Before investigating the general problem (1.4), we first address a simpler problem variant where the true mean μ_0 of ξ is known to vanish. Thus, we temporarily assume that ξ follows $\mathcal{N}(0, \Sigma_0)$. In this setting, it makes sense to focus on the modified ambiguity set $\mathbb{B}^0_\rho \triangleq \{\mathbb{Q} \in \mathcal{N}^p_0 : \mathbb{W}(\mathbb{Q}, \widehat{\mathbb{P}}) \le \rho\}$, which contains all normal distributions with zero mean that have a Wasserstein distance of at most $\rho \ge 0$ from the empirical distribution $\widehat{\mathbb{P}} = \mathcal{N}(0, \widehat{\Sigma})$. Under these assumptions, the estimation problem (1.4) thus simplifies to

$$\mathcal{J}(\widehat{\Sigma}) \triangleq \inf_{X \in \mathcal{X}} \left\{ -\log\det X + \sup_{\mathbb{Q} \in \mathbb{B}^0_\rho} \mathbb{E}_{\mathbb{Q}}[\langle \xi\xi^\top, X \rangle] \right\}. \tag{1.5}$$

We are now ready to state the first main result of this section.

Theorem 1.7 (Convex reformulation)**.** *For any fixed* $\rho > 0$ *and* $\widehat{\Sigma} \succeq 0$*, the simplified distributionally robust estimation problem* (1.5) *is equivalent to*

$$\mathcal{J}(\widehat{\Sigma}) = \left\{ \begin{array}{ll} \inf\limits_{X, \gamma} & -\log\det X + \gamma\left(\rho^2 - \mathrm{Tr}\left[\widehat{\Sigma}\right]\right) + \gamma^2 \langle (\gamma I - X)^{-1}, \widehat{\Sigma} \rangle \\ \text{s.t.} & \gamma I \succ X \succ 0, \quad X \in \mathcal{X}. \end{array} \right. \tag{1.6}$$

Moreover, the optimal value function $\mathcal{J}(\widehat{\Sigma})$ *is continuous in* $\widehat{\Sigma} \in \mathbb{S}_+$.

The proof of Theorem 1.7 relies on several auxiliary results. A main ingredient to derive the convex program (1.6) is a reformulation of the worst-case expectation function $g : \mathbb{S}_+ \times \mathbb{S}_+ \to \mathbb{R}$ defined through

$$g(\widehat{\Sigma}, X) \triangleq \sup_{\mathbb{Q} \in \mathbb{B}^0_\rho} \mathbb{E}_{\mathbb{Q}}[\langle \xi\xi^\top, X \rangle]. \tag{1.7}$$

In Proposition 1.9 below we will demonstrate that $g(\widehat{\Sigma}, X)$ is continuous and coincides with the optimal value of an explicit semidefinite program, a result which depends on the following preparatory lemma.

Lemma 1.8 (Continuity properties of partial infima)**.** *Consider a function $\varphi : \mathscr{E} \times \Gamma \to \mathbb{R}$ on two normed spaces $\mathscr{E}$ and Γ, and define the partial infimum with respect to γ as $\Phi(\varepsilon) \triangleq \inf_{\gamma \in \Gamma} \varphi(\varepsilon, \gamma)$ for every $\varepsilon \in \mathscr{E}$.*

(i) If $\varphi(\varepsilon,\gamma)$ is continuous in ε at $\varepsilon_0 \in \mathscr{E}$ for every $\gamma \in \Gamma$, then $\Phi(\varepsilon)$ is upper-semicontinuous at ε_0.

(ii) If $\varphi(\varepsilon,\gamma)$ is calm from below at $\varepsilon_0 \in \mathscr{E}$ uniformly in $\gamma \in \Gamma$, that is, if there exists a constant $L \geq 0$ such that $\varphi(\varepsilon,\gamma) - \varphi(\varepsilon_0,\gamma) \geq -L\|\varepsilon_0 - \varepsilon\|$ for all $\gamma \in \Gamma$, then $\Phi(\varepsilon)$ is lower-semicontinuous at ε_0.

Proof. As for assertion (i), we have

$$\begin{aligned}\limsup_{\varepsilon\to\varepsilon_0} \Phi(\varepsilon) &= \inf_{\delta>0} \sup_{\|\varepsilon-\varepsilon_0\|\leq\delta} \Phi(\varepsilon) = \inf_{\delta>0} \sup_{\|\varepsilon-\varepsilon_0\|\leq\delta} \inf_{\gamma\in\Gamma} \varphi(\varepsilon,\gamma)\\ &\leq \inf_{\gamma\in\Gamma} \inf_{\delta>0} \sup_{\|\varepsilon-\varepsilon_0\|\leq\delta} \varphi(\varepsilon,\gamma) = \inf_{\gamma\in\Gamma} \limsup_{\varepsilon\to\varepsilon_0} \varphi(\varepsilon,\gamma) = \inf_{\gamma\in\Gamma} \varphi(\varepsilon_0,\gamma) = \Phi(\varepsilon_0),\end{aligned}$$

where the inequality follows from interchanging the infimum and supremum operators, while the penultimate equality in the last line relies on the continuity assumption. As for assertion (ii), note that

$$\begin{aligned}\liminf_{\varepsilon\to\varepsilon_0} \Phi(\varepsilon) &= \sup_{\delta>0} \inf_{\|\varepsilon-\varepsilon_0\|\leq\delta} \Phi(\varepsilon) = \sup_{\delta>0} \inf_{\|\varepsilon-\varepsilon_0\|\leq\delta} \inf_{\gamma\in\Gamma} \varphi(\varepsilon,\gamma) = \sup_{\delta>0} \inf_{\gamma\in\Gamma} \inf_{\|\varepsilon-\varepsilon_0\|\leq\delta} \varphi(\varepsilon,\gamma)\\ &\geq \sup_{\delta>0} \inf_{\gamma\in\Gamma} \inf_{\|\varepsilon-\varepsilon_0\|\leq\delta} \big(\varphi(\varepsilon_0,\gamma) - L\|\varepsilon_0-\varepsilon\|\big) = \sup_{\delta>0} \inf_{\gamma\in\Gamma} \big(\varphi(\varepsilon_0,\gamma) - L\delta\big)\\ &= \inf_{\gamma\in\Gamma} \varphi(\varepsilon_0,\gamma) = \Phi(\varepsilon_0),\end{aligned}$$

where the inequality in the second line holds due to the calmness assumption. □

Proposition 1.9 (Worst-case expectation function)**.** *For any fixed $\rho > 0$, $\widehat{\Sigma} \succeq 0$ and $X \succ 0$, the worst-case expectation $g(\widehat{\Sigma}, X)$ defined in* (1.7) *coincides with the optimal value of the tractable semidefinite program*

$$\begin{array}{cl}\inf\limits_{\gamma} & \gamma\big(\rho^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\big) + \gamma^2\big\langle(\gamma I - X)^{-1}, \widehat{\Sigma}\big\rangle \\ \mathrm{s.t.} & \gamma I \succ X.\end{array} \tag{1.8}$$

Moreover, the optimal value function $g(\widehat{\Sigma}, X)$ is continuous in $(\widehat{\Sigma}, X) \in \mathbb{S}^p_+ \times \mathbb{S}^p_{++}$.

Proof. Using the definitions of the worst-case expectation $g(\widehat{\Sigma}, X)$ and the ambiguity set $\mathbb{B}^0_\rho$, we find

$$g(\widehat{\Sigma}, X) = \sup_{\mathbb{Q}\in\mathbb{B}^0_\rho} \left\langle \mathbb{E}_{\mathbb{Q}}[\xi\xi^\top], X\right\rangle = \sup_{S\in\mathbb{S}^p_+} \left\{\left\langle S, X\right\rangle : \mathbb{W}_S(S,\widehat{\Sigma}) \le \rho\right\},$$

where the second equality holds because the metric $\mathbb{W}_S$ on $\mathbb{S}^p_+$ is induced by the type-2 Wasserstein metric $\mathbb{W}$ on $\mathcal{N}^p_0$, meaning that there is a one-to-one correspondence between distributions $\mathbb{Q} \in \mathcal{N}^p_0$ with $\mathbb{W}(\mathbb{Q},\widehat{\mathbb{P}}) \le \rho$ and covariance matrices $S \in \mathbb{S}^p_+$ with $\mathbb{W}_S(S,\widehat{\Sigma}) \le \rho$. The continuity of $g(\widehat{\Sigma}, X)$ thus follows from Berge's maximum theorem [11, pp. 115–116], which applies because $\left\langle S, X\right\rangle$ and $\mathbb{W}_S(S,\widehat{\Sigma})$ are continuous in $(S,\widehat{\Sigma},X) \in \mathbb{S}^p_+ \times \mathbb{S}^p_+ \times \mathbb{S}^p_{++}$, while $\{S \in \mathbb{S}^p_+ : \mathbb{W}_S(S,\widehat{\Sigma}) \le \rho\}$ is nonempty and compact for every $\widehat{\Sigma} \in \mathbb{S}^p_+$ and $\rho > 0$.

By the definition of the induced metric $\mathbb{W}_S$ we then obtain

$$g(\widehat{\Sigma}, X) = \sup_{S\in\mathbb{S}^p_+} \left\{\left\langle S, X\right\rangle : \mathrm{Tr}\left[\widehat{\Sigma}\right] + \mathrm{Tr}\left[S\right] - 2\,\mathrm{Tr}\left[\sqrt{\widehat{\Sigma}^{\frac{1}{2}} S \widehat{\Sigma}^{\frac{1}{2}}}\right] \le \rho^2\right\}. \tag{1.9}$$

To establish the equivalence between (1.8) and (1.9), we first assume that $\widehat{\Sigma} \succ 0$. The generalization to rank deficient sample covariance matrices will be addressed later. By dualizing the explicit constraint in (1.9) and introducing the constant matrix $M = \widehat{\Sigma}^{\frac{1}{2}}$, which inherits invertibility from $\widehat{\Sigma}$, we find

$$\begin{aligned} g(\widehat{\Sigma}, X) &= \sup_{S\in\mathbb{S}^p_+} \inf_{\gamma\ge 0} \left\langle S, X - \gamma I\right\rangle + 2\gamma\left\langle \sqrt{MSM}, I\right\rangle + \gamma\left(\rho^2 - \mathrm{Tr}\left[\widehat{\Sigma}\right]\right) \\ &= \inf_{\gamma\ge 0} \sup_{S\in\mathbb{S}^p_+} \left\langle S, X - \gamma I\right\rangle + 2\gamma\left\langle \sqrt{MSM}, I\right\rangle + \gamma\left(\rho^2 - \mathrm{Tr}\left[\widehat{\Sigma}\right]\right) \\ &= \inf_{\gamma\ge 0} \left\{\gamma\left(\rho^2 - \mathrm{Tr}\left[\widehat{\Sigma}\right]\right) + \sup_{B\in\mathbb{S}^p_+} \left\{\left\langle B^2, M^{-1}(X-\gamma I)M^{-1}\right\rangle + 2\gamma\left\langle B, I\right\rangle\right\}\right\}. \end{aligned} \tag{1.10}$$

Here, the first equality exploits the identity $\mathrm{Tr}\left[A\right] = \langle A, I\rangle$ for any $A \in \mathbb{R}^{p\times p}$, the second equality follows from strong duality, which holds because $\widehat{\Sigma}$ constitutes a Slater point for problem (1.9) when $\rho > 0$, and the third equality relies on the substitution $B \leftarrow \sqrt{MSM}$, which implies that $S = M^{-1}B^2M^{-1}$. Introducing the shorthand $\Delta = M^{-1}(X - \gamma I)M^{-1}$ allows us to simplify the inner maximization problem over B in (1.10) to

$$\sup_{B\in\mathbb{S}^p_+} \left\{\left\langle B^2, \Delta\right\rangle + 2\gamma\left\langle B, I\right\rangle\right\}. \tag{1.11}$$

If $\Delta \not\prec 0$, then (1.11) is unbounded. To see this, denote by $\overline{\lambda}(\Delta)$ the largest eigenvalue of Δ and by $\overline{v}$ a corresponding eigenvector. If $\overline{\lambda}(\Delta) > 0$, then the objective value of $B_k = k\cdot\overline{v}\,\overline{v}^\top \succeq 0$ in (1.11) grows quadratically with k. If $\overline{\lambda}(\Delta) = 0$, then $\gamma > 0$ for otherwise $X \preceq 0$ contrary to our assumption, and thus the objective value of B_k in (1.11) grows linearly with k. In both cases (1.11) is indeed unbounded.

If $\Delta \prec 0$, then (1.11) becomes a convex optimization problem that can be solved analytically. Indeed, the objective function of (1.11) is minimized by $B^\star = -\gamma\Delta^{-1}$, which satisfies the first-order optimality condition

$$B\Delta + \Delta B + 2\gamma I = 0 \tag{1.12}$$

and is strictly feasible in (1.11) because $\Delta \prec 0$. Moreover, as (1.12) is naturally interpreted as a continuous Lyapunov equation, its solution $B^\star$ can be shown to be unique; see, *e.g.*, [83, Theorem 12.5]. We may thus conclude that $B^\star$ is the unique maximizer of (1.11) and that the maximum of (1.11) amounts to $-\gamma^2 \operatorname{Tr}\left[\Delta^{-1}\right]$.

Adding the constraint $\gamma I \succ X$ to the outer minimization problem in (1.10), thus excluding all values of γ for which $\Delta \not\prec 0$ and the inner supremum is infinite, and replacing the optimal value of the inner maximization problem with $-\gamma^2 \operatorname{Tr}\left[\Delta^{-1}\right] = \gamma^2\big\langle(\gamma I - X)^{-1}, \widehat{\Sigma}\big\rangle$ yields (1.8). This establishes the claim for $\widehat{\Sigma} \succ 0$.

In the second part of the proof, we show that the claim remains valid for rank deficient sample covariance matrices. To this end, we denote the optimal value of problem (1.8) by $g'(\widehat{\Sigma}, X)$. From the first part of the proof we know that $g'(\widehat{\Sigma}, X) = g(\widehat{\Sigma}, X)$ for all $\widehat{\Sigma}, X \in \mathbb{S}^p_{++}$. We also know that $g(\widehat{\Sigma}, X)$ is continuous in $(\widehat{\Sigma}, X) \in \mathbb{S}^p_+ \times \mathbb{S}^p_{++}$. It remains to be shown that $g'(\widehat{\Sigma}, X) = g(\widehat{\Sigma}, X)$ for all $\widehat{\Sigma} \in \mathbb{S}^p_+$ and $X \in \mathbb{S}^p_{++}$.

Fix any $\widehat{\Sigma} \in \mathbb{S}^p_+$ and $X \in \mathbb{S}^p_{++}$, and note that $\widehat{\Sigma} + \varepsilon I \succ 0$ for every $\varepsilon > 0$. Defining the intervals $\mathscr{E} = \mathbb{R}_+$ and $\Gamma = \{\gamma \in \mathbb{R} : \gamma I \succ X\}$ as well as the auxiliary functions

$$\Phi(\varepsilon) = g'(\widehat{\Sigma} + \varepsilon I, X) \quad \text{and} \quad \varphi(\varepsilon, \gamma) = \gamma\left(\rho^2 - \operatorname{Tr}\left[\widehat{\Sigma} + \varepsilon I\right]\right) + \gamma^2\big\langle(\gamma I - X)^{-1}, \widehat{\Sigma} + \varepsilon I\big\rangle,$$

it follows from (1.8) that

$$\Phi(\varepsilon) = \inf_{\gamma \in \Gamma} \varphi(\varepsilon, \gamma) \quad \forall \varepsilon \in \mathscr{E}.$$

One can show via Lemma 1.8 that $\Phi(\varepsilon)$ is continuous at $\varepsilon = 0$. Indeed, $\varphi(\varepsilon, \gamma)$ is linear and thus continuous in ε for every $\gamma \in \Gamma$, which implies via Lemma 1.8(a) that $\Phi(\varepsilon)$ is upper-semicontinuous at $\varepsilon = 0$. Moreover, $\varphi(\varepsilon, \gamma)$ is calm from below at $\varepsilon = 0$ with $L = 0$ uniformly in $\gamma \in \Gamma$ because

$$\varphi(\varepsilon, \gamma) - \varphi(0, \gamma) = \gamma \operatorname{Tr}\left[(I - \gamma^{-1}X)^{-1} - I\right]\varepsilon \geq 0 \quad \forall \gamma \in \Gamma.$$

Here, the inequality holds for all $\gamma \in \Gamma$ due to the conditions $I \succ \gamma^{-1}X \succ 0$, which are equivalent to $0 \prec I - \gamma^{-1}X \prec I$ and imply $(I - \gamma^{-1}X)^{-1} \succ I$. Lemma 1.8(b) thus ensures that $\Phi(\varepsilon)$ is lower-semicontinuous at $\varepsilon = 0$. In summary, we conclude that $\Phi(\varepsilon)$ is indeed continuous at $\varepsilon = 0$.

Combining the above results, we find

$$g(\widehat{\Sigma}, X) = \lim_{\varepsilon \to 0^+} g(\widehat{\Sigma} + \varepsilon I, X) = \lim_{\varepsilon \to 0^+} g'(\widehat{\Sigma} + \varepsilon I, X) = \lim_{\varepsilon \to 0^+} \Phi(\varepsilon) = \Phi(0) = g'(\widehat{\Sigma}, X),$$

where the five equalities hold due to the continuity of $g(\widehat{\Sigma}, X)$ in $\widehat{\Sigma}$, the fact that $g(\widehat{\Sigma}, X) =$

$g'(\widehat{\Sigma}, X)$ for all $\widehat{\Sigma} \succ 0$, the definition of $\Phi(\varepsilon)$, the continuity of $\Phi(\varepsilon)$ at $\varepsilon = 0$ and once again from the definition of $\Phi(\varepsilon)$, respectively. The claim now follows because $\widehat{\Sigma} \in \mathbb{S}_+^p$ and $X \in \mathbb{S}_{++}^p$ were chosen arbitrarily. □

We have now collected all necessary ingredients for the proof of Theorem 1.7.

Proof of Theorem 1.7. By Proposition 1.9, the worst-case expectation in (1.5) coincides with the optimal value of the semidefinite program (1.8). Substituting this semidefinite program into (1.5) yields (1.6). Note that the condition $X \succ 0$, which ensures that $\log\det X$ is well-defined, is actually redundant because it is implied by the constraint $X \in \mathscr{X}$. Nevertheless, we make it explicit in (1.6) for the sake of clarity.

It remains to show that $\mathscr{J}(\widehat{\Sigma})$ is continuous. To this end, we first construct bounds on the minimizers of (1.6) that vary continuously with $\widehat{\Sigma}$. Such bounds can be constructed from any feasible decision (X_0, γ_0). Assume without loss of generality that $\gamma_0 > p/\rho^2$, and denote by $f_0(\widehat{\Sigma})$ the objective value of (X_0, γ_0) in (1.6), which constitutes a linear function of $\widehat{\Sigma}$. Moreover, define two continuous auxiliary functions

$$\overline{x}(\widehat{\Sigma}) \triangleq \frac{f_0(\widehat{\Sigma}) - p(1 - \log\gamma_0)}{\rho^2 - p\gamma_0^{-1}} \qquad \text{and} \qquad \underline{x}(\widehat{\Sigma}) \triangleq \frac{e^{-f_0(\widehat{\Sigma})}}{\overline{x}(\widehat{\Sigma})^{p-1}}, \tag{1.13}$$

which are strictly positive because $\gamma_0 > p/\rho^2$. Clearly, the infimum of problem (1.6) is determined only by feasible decisions (X, γ) with an objective value of at most $f_0(\widehat{\Sigma})$. All such decisions satisfy

$$\begin{aligned} f_0(\widehat{\Sigma}) &\geq -\log\det X + \gamma\rho^2 + \gamma\big\langle (I - \gamma^{-1}X)^{-1} - I, \widehat{\Sigma}\big\rangle \geq -\log\det X + \gamma\rho^2 \\ &\geq -p\log\gamma + \gamma\rho^2 \geq (\rho^2 - \gamma_0^{-1}p)\gamma + p(1 - \log\gamma_0), \end{aligned} \tag{1.14}$$

where the second and third inequalites exploit the estimates $(I - \gamma^{-1}X)^{-1} \succ I$ and $\det X \leq \det(\gamma I) = \gamma^p$, respectively, which are both implied by the constraint $\gamma I \succ X \succ 0$, and the last inequality holds because $\log\gamma \leq \log\gamma_0 + \gamma_0^{-1}(\gamma - \gamma_0)$ for all $\gamma > 0$. By rearranging the above inequality and recalling the definition of $\overline{x}(\widehat{\Sigma})$, we thus find $\gamma \leq \overline{x}(\widehat{\Sigma})$, which in turn implies that $X \prec \gamma I \preceq \overline{x}(\widehat{\Sigma}) I$.

Denoting by $\{x_i\}_{i \leq p}$ the eigenvalues of the matrix X and setting $x_{\min} = \min_{i \leq p} x_i$, we further find

$$f_0(\widehat{\Sigma}) \geq -\log\det X = -\log\left(\prod_{i=1}^{p} x_i\right) \geq -\log\left(x_{\min}\,\overline{x}(\widehat{\Sigma})^{p-1}\right) = -\log x_{\min} - (p-1)\log\overline{x}(\widehat{\Sigma}),$$

where the first inequality follows from (1.14), while the second inequality is based on overestimating all but the smallest eigenvalue of X by $\overline{x}(\widehat{\Sigma})$. By rearranging the above inequality and recalling the definition of $\underline{x}(\widehat{\Sigma})$, we thus find $x_{\min} \geq \underline{x}(\widehat{\Sigma})$, which in turn implies that $X \succeq \underline{x}(\widehat{\Sigma}) I$.

The above reasoning shows that the extra constraint $\underline{x}(\widehat{\Sigma})I \preceq X \preceq \overline{x}(\widehat{\Sigma})I$ has no impact on (1.6), that is,

$$\begin{aligned}\mathscr{J}(\widehat{\Sigma}) &= \left\{\begin{array}{ll}\inf\limits_{X} & -\log\det X + \inf\limits_{\gamma}\{\gamma\left(\rho^2 - \operatorname{Tr}\left[\widehat{\Sigma}\right]\right) + \gamma^2\langle(\gamma I - X)^{-1}, \widehat{\Sigma}\rangle : \gamma I \succ X\} \\ \text{s.t.} & X \in \mathscr{X}, \quad \underline{x}(\widehat{\Sigma})I \preceq X \preceq \overline{x}(\widehat{\Sigma})I.\end{array}\right. \\ &= \left\{\begin{array}{ll}\inf\limits_{X} & -\log\det X + g(\widehat{\Sigma}, X) \\ \text{s.t.} & X \in \mathscr{X}, \quad \underline{x}(\widehat{\Sigma})I \preceq X \preceq \overline{x}(\widehat{\Sigma})I,\end{array}\right.\end{aligned}$$

where the second equality follows from Proposition 1.9. The continuity of $\mathscr{J}(\widehat{\Sigma})$ now follows directly from Berge's maximum theorem [11, pp. 115–116], which applies due to the continuity of $g(\widehat{\Sigma}, X)$ established in Proposition 1.9, the compactness of the feasible set and the continuity of $\underline{x}(\widehat{\Sigma})$ and $\overline{x}(\widehat{\Sigma})$. □

An immediate consequence of Theorem 1.7 is that the simplified estimation problem (1.5) is equivalent to an explicit semidefinite program and is therefore in principle computationally tractable.

Corollary 1.10 (Tractability)**.** *For any fixed* $\rho > 0$ *and* $\widehat{\Sigma} \succeq 0$, *the simplified distributionally robust estimation problem* (1.5) *is equivalent to the tractable semidefinite program*

$$\mathscr{J}(\widehat{\Sigma}) = \left\{\begin{array}{ll}\inf\limits_{X,Y,\gamma} & -\log\det X + \gamma\left(\rho^2 - \operatorname{Tr}\left[\widehat{\Sigma}\right]\right) + \operatorname{Tr}\left[Y\right] \\ \text{s.t.} & \begin{bmatrix} Y & \gamma\widehat{\Sigma}^{\frac{1}{2}} \\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & \gamma I - X\end{bmatrix} \succeq 0 \\ & \gamma I \succ X \succ 0, \quad Y \succeq 0, \quad X \in \mathscr{X}.\end{array}\right. \tag{1.15}$$

Proof. We know from Theorem 1.7 that the estimation problem (1.5) is equivalent to the convex program (1.6). As X represents a decision variable instead of a parameter, however, problem (1.6) fails to be a semidefinite program per se. Indeed, its objective function involves the nonlinear term $h(X,\gamma) \triangleq \gamma^2\langle(\gamma I - X)^{-1}, \widehat{\Sigma}\rangle$, which is interpreted as ∞ outside of its domain $\{(X,\gamma) \in \mathbb{S}_+ \times \mathbb{R} : \gamma I \succ X\}$. However, $h(X,\gamma)$ constitutes a matrix fractional function as described in [18, Example 3.4] and thus admits the semidefinite reformulation

$$\begin{aligned} h(X,\gamma) &= \inf_{t}\{t \,:\, \gamma I \succ X, \quad \gamma^2\langle(\gamma I - X)^{-1}, \widehat{\Sigma}\rangle \le t\} \\ &= \inf_{Y,t}\left\{t \,:\, \gamma I \succ X, \quad Y \succeq \gamma^2\widehat{\Sigma}^{\frac{1}{2}}(\gamma I - X)^{-1}\widehat{\Sigma}^{\frac{1}{2}}, \quad \operatorname{Tr}\left[Y\right] \le t\right\} \\ &= \inf_{Y}\left\{\operatorname{Tr}\left[Y\right] \,:\, \gamma I \succ X, \quad \begin{bmatrix} Y & \gamma\widehat{\Sigma}^{\frac{1}{2}} \\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & \gamma I - X\end{bmatrix} \succeq 0\right\},\end{aligned}$$

where the second equality holds because $A \succeq B$ implies $\operatorname{Tr}\left[A\right] \ge \operatorname{Tr}\left[B\right]$, while the third equality follows from a standard Schur complement argument; see, *e.g.*, [18, Appendix A.5.5]. Thus, $h(X,\gamma)$ is representable as the optimal value of a parametric semidefinite program whose objective and constraint functions are jointly convex in the auxiliary decision variable Y and the parameters X and γ. The postulated reformulation (1.15) is then obtained by substituting the last expression into (1.6). □

1.2.3 Joint Estimation of the Mean Vector and the Precision Matrix

Now that we have derived a tractable semidefinite reformulation for the simplified estimation problem (1.5), we are ready to address the generic estimation problem (1.4), which does *not* assume knowledge of the mean and is robustified against all distributions in the ambiguity set $\mathbb{B}_\rho$ *without* mean constraints.

Theorem 1.11 (Sufficiency of $\widehat{\Sigma}$)**.** *For any fixed $\rho > 0$, $\widehat{\mu} \in \mathbb{R}^p$ and $\widehat{\Sigma} \in \mathbb{S}_+^p$, the general distributionally robust estimation problem* (1.4) *is equivalent to the optimization problem* (1.6) *and the tractable semidefinite program* (1.15)*. Moreover, the optimal value function $\mathscr{J}(\widehat{\mu}, \widehat{\Sigma})$ is constant in $\widehat{\mu}$ and continuous in $\widehat{\Sigma}$.*

Proof. By Proposition 1.3, the optimal value of the estimation problem (1.4) can be expressed as

$$\begin{aligned}\mathscr{J}(\widehat{\mu}, \widehat{\Sigma}) &= \inf_{\mu,\, X \in \mathcal{X}} -\log\det X + \left\{ \begin{array}{cl} \sup\limits_{\mu',\, S \succeq 0} & (\mu' - \mu)^\top X (\mu' - \mu) + \langle S, X\rangle \\ \text{s.t.} & \operatorname{Tr}[S] + \operatorname{Tr}[\widehat{\Sigma}] - 2\operatorname{Tr}\Big[\sqrt{\widehat{\Sigma}^{\frac{1}{2}} S \widehat{\Sigma}^{\frac{1}{2}}}\Big] \le \rho^2 - \|\mu' - \widehat{\mu}\|^2 \end{array} \right. \\ &= \inf_{\mu,\, X \in \mathcal{X}} -\log\det X + \sup_{\mu' : \|\mu' - \widehat{\mu}\| \le \rho} (\mu' - \mu)^\top X (\mu' - \mu) \\ &\qquad\qquad + \inf_{\gamma : \gamma I \succ X} \gamma\left(\rho^2 - \|\mu' - \widehat{\mu}\|^2 - \operatorname{Tr}[\widehat{\Sigma}]\right) + \gamma^2 \langle (\gamma I - X)^{-1}, \widehat{\Sigma}\rangle.\end{aligned}$$

Here, the second equality holds because the Wasserstein constraint is infeasible unless $\|\mu' - \widehat{\mu}\| \le \rho$ and because the maximization problem over S, which constitutes an instance of (1.9) with $\rho^2 - \|\mu' - \widehat{\mu}\|^2$ instead of ρ^2, can be reformulated as a minimization problem over γ thanks to Proposition 1.9. By the minimax theorem [13, Proposition 5.5.4], which applies because μ' ranges over a compact ball and because $X - \gamma I \prec 0$, we may then interchange the maximization over μ' with the minimization over γ to obtain

$$\begin{aligned}\mathscr{J}(\widehat{\mu}, \widehat{\Sigma}) = \inf_{\substack{\mu,\, X \in \mathcal{X}, \\ \gamma : \gamma I \succ X}} -\log\det X + \sup_{\mu' : \|\mu' - \widehat{\mu}\| \le \rho} & (\mu' - \mu)^\top X (\mu' - \mu) \\ & + \gamma\left(\rho^2 - \|\mu' - \widehat{\mu}\|^2 - \operatorname{Tr}[\widehat{\Sigma}]\right) + \gamma^2 \langle (\gamma I - X)^{-1}, \widehat{\Sigma}\rangle.\end{aligned}$$

Using the minimax theorem [13, Proposition 5.5.4] once again to interchange the minimization over μ with the maximization over μ' yields

$$\begin{aligned}\mathcal{J}(\widehat{\mu},\widehat{\Sigma}) = &\inf_{\substack{X\in\mathcal{X},\\ \gamma:\gamma I\succ X}} \; -\log\det X + \sup_{\mu':\|\mu'-\widehat{\mu}\|\le\rho} \; \inf_{\mu} (\mu'-\mu)^\top X(\mu'-\mu) \\ &\qquad\qquad + \gamma\left(\rho^2 - \|\mu'-\widehat{\mu}\|^2 - \operatorname{Tr}\big[\widehat{\Sigma}\big]\right) + \gamma^2\big\langle(\gamma I - X)^{-1},\widehat{\Sigma}\big\rangle \\ = &\inf_{\substack{X\in\mathcal{X},\\ \gamma:\gamma I\succ X}} \; -\log\det X + \gamma\left(\rho^2 - \operatorname{Tr}\big[\widehat{\Sigma}\big]\right) + \gamma^2\big\langle(\gamma I - X)^{-1},\widehat{\Sigma}\big\rangle,\end{aligned}$$

where the second equality holds because μ' is the unique optimal solution of the innermost minimization problem over μ, while $\widehat{\mu}$ is the unique optimal solution of the maximization problem over μ'. Thus, the general estimation problem (1.4) is equivalent to (1.6), and $\mathcal{J}(\widehat{\mu},\widehat{\Sigma})$ is manifestly constant in $\widehat{\mu}$. Theorem 1.7 further implies that $\mathcal{J}(\widehat{\mu},\widehat{\Sigma})$ is continuous in $\widehat{\Sigma}$, while Corollary 1.10 implies that (1.4) is equivalent to the tractable semidefinite program (1.15). These observations complete the proof. □

Theorem 1.11 asserts that the general estimation problem (1.4) is equivalent to the simplified estimation problem (1.5), which is based on the hypothesis that the mean of ξ is known to vanish. Theorem 1.11 further reveals that the general estimation problem (1.4) as well as its (unique) optimal solution depend on the training data only through the sample covariance matrix $\widehat{\Sigma}$. This is reassuring because $\widehat{\Sigma}$ is known to be a sufficient statistic for the precision matrix. As solving (1.4) is tantamount to solving (1.5), it suffices to devise solution procedures for the simplified estimation problem (1.5) or its equivalent reformulations (1.6) and (1.15).

We emphasize that the strictly convex log-determinant term in the objective of (1.15) is supported by state-of-the-art interior point solvers for semidefinite programs such as SDPT3 [168]. In principle, problem (1.15) can therefore be implemented directly in MATLAB using the YALMIP interface [112], for instance. In spite of its theoretical tractability, however, the semidefinite program (1.15) quickly becomes excruciatingly large, and direct solution with a general purpose solver becomes impracticable already for moderate values of p. This motivates us to investigate practically relevant special cases in which the estimation problem (1.5) can be solved either analytically (Section 1.3) or numerically using a dedicated fast Newton-type algorithm (Section 1.4).

1.3 Analytical Solution without Sparsity Information

If we have no prior information about the precision matrix, it is natural to set $\mathcal{X} = \mathbb{S}^p_{++}$. In this case, the distributionally robust estimation problem (1.5) can be solved in quasi-closed form.

Theorem 1.12 (Analytical solution without sparsity information)**.** *If $\rho > 0$, $\mathcal{X} = \mathbb{S}^p_{++}$ and $\widehat{\Sigma} \in \mathbb{S}^p_+$ admits the spectral decomposition $\widehat{\Sigma} = \sum_{i=1}^p \lambda_i v_i v_i^\top$ with eigenvalues λ_i and corresponding orthonormal eigenvectors v_i, $i \le p$, then the unique minimizer of* (1.5) *is given by*

$X^\star = \sum_{i=1}^p x_i^\star v_i v_i^\top$, where

$$x_i^\star = \gamma^\star \left[1 - \frac{1}{2}\left(\sqrt{\lambda_i^2(\gamma^\star)^2 + 4\lambda_i\gamma^\star} - \lambda_i\gamma^\star\right)\right] \qquad \forall i \le p \tag{1.16a}$$

and $\gamma^\star > 0$ is the unique positive solution of the algebraic equation

$$\left(\rho^2 - \frac{1}{2}\sum_{i=1}^p \lambda_i\right)\gamma - p + \frac{1}{2}\sum_{i=1}^p \sqrt{\lambda_i^2\gamma^2 + 4\lambda_i\gamma} = 0. \tag{1.16b}$$

Proof. We first demonstrate that the algebraic equation (1.16b) admits a unique solution in $\mathbb{R}_+$. For ease of exposition, we define $\varphi(\gamma)$ as the left-hand side of (1.16b). It is easy to see that $\varphi(0) = -p < 0$ and $\lim_{\gamma\to\infty}\varphi(\gamma)/\gamma = \rho^2$, which implies that $\varphi(\gamma)$ grows asymptotically linearly with γ at slope $\rho^2 > 0$. By the intermediate value theorem, we may thus conclude that the equation (1.16b) has a solution $\gamma^\star > 0$.

As $\lambda_i\gamma + 2 > \sqrt{\lambda_i^2\gamma^2 + 4\lambda_i\gamma}$, the derivative of $\varphi(\gamma)$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}\varphi(\gamma) = \rho^2 + \frac{1}{2}\sum_{i=1}^p \lambda_i\left(\frac{\lambda_i\gamma + 2}{\sqrt{\lambda_i^2\gamma^2 + 4\lambda_i\gamma}} - 1\right) > 0,$$

whereby $\varphi(\gamma)$ is strictly increasing in $\gamma \in \mathbb{R}_+$. Thus, the solution $\gamma^\star$ is unique. The positive slope of $\varphi(\gamma)$ further implies via the implicit function theorem that $\gamma^\star$ changes continuously with $\lambda_i \in \mathbb{R}_+$, $i \le p$.

In analogy to Proposition 1.9, we prove the claim first under the assumption that $\widehat{\Sigma} \succ 0$ and postpone the generalization to rank deficient sample covariance matrices. Focussing on $\widehat{\Sigma} \succ 0$, we will show that $(X^\star, \gamma^\star)$ is feasible and optimal in (1.6). By Theorem 1.7, this will imply that $X^\star$ is feasible and optimal in (1.5).

As $\gamma^\star > 0$ and $\widehat{\Sigma} \succ 0$, which means that $\lambda_i > 0$ for all $i \le p$, an elementary calculation shows that

$$2 > \sqrt{\lambda_i^2(\gamma^\star)^2 + 4\lambda_i\gamma^\star} - \lambda_i\gamma^\star > 0 \quad\iff\quad 1 > 1 - \frac{1}{2}\left(\sqrt{\lambda_i^2(\gamma^\star)^2 + 4\lambda_i\gamma^\star} - \lambda_i\gamma^\star\right) > 0.$$

Multiplying the last inequality by $\gamma^\star$ proves that $\gamma^\star > x_i^\star > 0$ for all $i \le p$, which in turn implies that $\gamma^\star I \succ X^\star \succ 0$. Thus, $(X^\star, \gamma^\star)$ is feasible in (1.6), and $X^\star$ is feasible in (1.5).

To prove optimality, we denote by $f(X, \gamma)$ the objective function of problem (1.6) and note that its gradient with respect to X vanishes at $(X^\star, \gamma^\star)$. Indeed, we have

$$\begin{aligned}\nabla_X f(X^\star,\gamma^\star) &= -(X^\star)^{-1} + (\gamma^\star)^2(\gamma^\star I - X^\star)^{-1}\widehat{\Sigma}(\gamma^\star I - X^\star)^{-1} \\ &= \sum_{i=1}^p \left((\gamma^\star)^2(\gamma^\star - x_i^\star)^{-2}\lambda_i - (x_i^\star)^{-1}\right) v_i v_i^\top \\ &= \sum_{i=1}^p \frac{(\gamma^\star)^2 x_i^\star \lambda_i - (\gamma^\star - x_i^\star)^2}{(\gamma^\star - x_i^\star)^2 x_i} v_i v_i^\top = 0,\end{aligned}$$

where the first equality exploits the basic rules of matrix calculus (see, *e.g.*, [12, p. 631]), the second equality holds because $\widehat{\Sigma}$ and X share the same eigenvectors v_i, $i \le p$, and the last equation follows from the identity

$$(\gamma^\star)^2 x_i^\star \lambda_i = (\gamma^\star - x_i^\star)^2 \quad \forall i \le p, \tag{1.17}$$

which is a direct consequence of the definitions of $\gamma^\star$ and $x_i^\star$, $i \le p$, in (1.16). Similarly, the partial derivative of $f(X,\gamma)$ with respect to γ vanishes at $(X^\star,\gamma^\star)$, too. In fact, we have

$$\begin{aligned}\frac{\partial}{\partial\gamma} f(X^\star,\gamma^\star) &= \rho^2 - \mathrm{Tr}\left[\widehat{\Sigma}\right] + 2\gamma^\star\,\mathrm{Tr}\left[(\gamma^\star I - X^\star)^{-1}\widehat{\Sigma}\right] - (\gamma^\star)^2\mathrm{Tr}\left[(\gamma^\star I - X^\star)^{-1}\widehat{\Sigma}(\gamma^\star I - X^\star)^{-1}\right] \\ &= \rho^2 - \sum_{i=1}^p \lambda_i \left(1 - \frac{2\gamma^\star}{\gamma^\star - x_i^\star} + \frac{(\gamma^\star)^2}{(\gamma^\star - x_i^\star)^2}\right) = \rho^2 - \sum_{i=1}^p \frac{(x_i^\star)^2}{(\gamma^\star - x_i^\star)^2}\lambda_i \\ &= \frac{1}{(\gamma^\star)^2}\left(\rho^2(\gamma^\star)^2 - \sum_{i=1}^p x_i^\star\right) = 0,\end{aligned}$$

where the second equality expresses $\widehat{\Sigma}$ and X in terms of their respective spectral decompositions, the fourth equality holds due to (1.17), and the last equality follows from the observation that $\rho^2(\gamma^\star)^2 = \sum_{i=1}^p x_i^\star$. In summary, we have shown that $(X^\star,\gamma^\star)$ satisfies the first-order optimality conditions of the convex optimization problem (1.6), which ensures that $X^\star$ is optimal in (1.5).

Consider now any (possibly singular) sample covariance matrix $\widehat{\Sigma} \in \mathbb{S}_+^p$. As $\gamma^\star > 0$, similar arguments as in the first part of the proof show that $\gamma^\star \ge x_i^\star > 0$ for all $i \le p$, which in turn implies that $\gamma^\star I \succeq X^\star \succ 0$. Moreover, if $\widehat{\Sigma}$ has at least one zero eigenvalue, it is easy to see that $\gamma^\star I \not\succ X^\star$, in which case $(X^\star,\gamma^\star)$ fails to be feasible in (1.6). However, $X^\star$ remains feasible and optimal in (1.5). To see this, consider the invertible sample covariance matrix $\widehat{\Sigma} + \varepsilon I \succ 0$ for some $\varepsilon > 0$, and denote by $(X^\star(\varepsilon), \gamma^\star(\varepsilon))$ the corresponding minimizer of problem (1.6) as constructed in (1.16). As the solution of the algebraic equation (1.16b) depends continuously on the eigenvalues of the sample covariance matrix, we conclude that the auxiliary variable $\gamma^\star(\varepsilon)$ and—by virtue of (1.16a)—the estimator $X^\star(\varepsilon)$ are both continuous in $\varepsilon \in \mathbb{R}_+$. Thus, we find

$$\mathscr{J}(\widehat{\Sigma}) = \lim_{\varepsilon\to 0^+} \mathscr{J}(\widehat{\Sigma} + \varepsilon I) = \lim_{\varepsilon\to 0^+} -\log\det X^\star(\varepsilon) + g(\widehat{\Sigma} + \varepsilon I, X^\star(\varepsilon)) = -\log\det X^\star + g(\widehat{\Sigma}, X^\star),$$

where the first equality follows from the continuity of $\mathscr{J}(\widehat{\Sigma})$ established in Theorem 1.7, the second equality holds because $X^\star(\varepsilon)$ is the optimal estimator corresponding to the sample covariance matrix $\widehat{\Sigma}+\varepsilon I \succ 0$ in problem (1.5), and the third equality follows from the continuity of $g(\widehat{\Sigma}, X)$ established in Proposition 1.9 and the fact that $\lim_{\varepsilon\to 0^+} X^\star(\varepsilon) = X^\star \succ 0$. Thus, $X^\star$ is indeed optimal in (1.5). The strict convexity of $-\log\det X$ further implies that $X^\star$ is unique. This observation completes the proof. □

Remark 1.13 (Properties of $X^\star$). *The optimal distributionally robust estimator $X^\star$ identified in Theorem 1.12 commutes with the sample covariance matrix $\widehat{\Sigma}$ because both matrices share the same eigenbasis. Moreover, the eigenvalues of $X^\star$ are obtained from those of $\widehat{\Sigma}$ via a nonlinear transformation that depends on the size ρ of the ambiguity set. We emphasize that all eigenvalues of $X^\star$ are positive for every $\rho > 0$, which implies that $X^\star$ is invertible. These insights suggest that $X^\star$ constitutes a nonlinear shrinkage estimator, which enjoys the rotation equivariance property (when all data points are rotated by $R \in \mathbb{R}^{p\times p}$, then $X^\star$ changes to $RX^\star R^\top$).*

Theorem 1.12 characterizes the optimal solution of problem (1.5) in quasi-closed form up to the spectral decomposition of $\widehat{\Sigma}$ and the numerical solution of equation (1.16b). By [131, Theorem 1.1], the eigenvalues of $\widehat{\Sigma}$ can be computed to within an absolute error ε in $\mathscr{O}(p^3)$ arithmetic operations. Moreover, as its left-hand side is increasing in $\gamma^\star$, equation (1.16b) can be solved reliably via bisection or by the Newton-Raphson method. The following lemma provides a priori bounds on $\gamma^\star$ that can be used to initialize the bisection interval.

Lemma 1.14 (Bisection interval). *For $\rho > 0$, the unique solution of* (1.16b) *satisfies $\gamma^\star \in [\gamma_{\min}, \gamma_{\max}]$, where*

$$\gamma_{\min} = \frac{p^2\lambda_{\max} + 2p\rho^2 - p\sqrt{p^2\lambda_{\max}^2 + 4p\rho^2\lambda_{\max}}}{2\rho^4} > 0, \qquad \gamma_{\max} = \min\left\{\frac{p}{\rho^2}, \frac{1}{\rho}\sqrt{\sum_{i=1}^{p}\frac{1}{\lambda_i}}\right\}, \tag{1.18}$$

and $\lambda_{\max}$ denotes the maximum eigenvalue of $\widehat{\Sigma}$.

Proof. By the definitions of $\gamma^\star$ and $x_i^\star$ in (1.16) we have $\lambda_i x_i^\star = (\gamma^\star - x_i^\star)^2/(\gamma^\star)^2 < 1$, which implies that $x_i^\star \le \frac{1}{\lambda_i}$. Using (1.16) one can further show that $(\gamma^\star)^2 = \frac{1}{\rho^2}\sum_{i=1}^p x_i^\star \le \frac{1}{\rho^2}\sum_{i=1}^p \frac{1}{\lambda_i}$, which is equivalent to $\gamma^\star \le \frac{1}{\rho}(\sum_{i=1}^p \frac{1}{\lambda_i})^{\frac{1}{2}}$. Note that this upper bound on $\gamma^\star$ is finite only if $\lambda_i > 0$ for all $i \le p$. To derive an upper bound that is universally meaningful, we denote the left-hand side of (1.16b) by $\varphi(\gamma)$ and note that $\rho^2\gamma - p \le \varphi(\gamma)$ for all $\gamma \ge 0$. This estimate implies that $\gamma^\star \le \frac{p}{\rho^2}$. Thus, we find $\gamma^\star \le \min\{\frac{p}{\rho^2}, \frac{1}{\rho}(\sum_{i=1}^p \frac{1}{\lambda_i})^{\frac{1}{2}}\} = \gamma_{\max}$.

To derive a lower bound on $\gamma^\star$, we set $\lambda_{\max} = \max_{i\le p}\lambda_i$ and observe that

$$\varphi(\gamma) \le \rho^2\gamma - p + \sum_{i=1}^p \sqrt{\lambda_i\gamma} \le \rho^2\gamma - p + p\sqrt{\lambda_{\max}\gamma},$$

where the first inequality holds because $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for all $a, b \ge 0$. As the unique positive zero of the right-hand side, $\gamma_{\min}$ provides a nontrivial lower bound on $\gamma^\star$. Thus, the claim follows. □

Lemma 1.14 implies that $\gamma^\star$ can be computed via the standard bisection algorithm to within an absolute error of ε in $\log_2((\gamma_{\max}-\gamma_{\min})/\varepsilon) = \mathcal{O}(\log_2 p)$ iterations. As evaluating the left-hand side of (1.16b) requires only $\mathcal{O}(p)$ arithmetic operations, the computational effort for constructing $X^\star$ is largely dominated by the cost of the spectral decomposition of the sample covariance matrix.

Remark 1.15 (Numerical stability). *If both $\gamma^\star$ and λ_i are large numbers, then formula* (1.16a) *for $x_i^\star$ becomes numerically unstable. A mathematically equivalent but numerically more robust reformulation of* (1.16a) *is*

$$x_i^\star = \gamma^\star \left(1 - \frac{2}{1 + \sqrt{1 + \frac{4}{\lambda_i \gamma^\star}}} \right).$$

In the following we investigate the impact of the Wasserstein radius ρ on the optimal Lagrange multiplier $\gamma^\star$ and the corresponding optimal estimator $X^\star$.

Proposition 1.16 (Sensitivity analysis). *Assume that the eigenvalues of $\widehat{\Sigma}$ are sorted in ascending order, that is, $\lambda_1 \le \cdots \le \lambda_p$. If $\gamma^\star(\rho)$ denotes the solution of* (1.16b), *and $x_i^\star(\rho)$, $i \le p$, represent the eigenvalues of $X^\star$ defined in* (1.16a), *which makes the dependence on $\rho > 0$ explicit, then the following assertions hold:*

(i) *$\gamma^\star(\rho)$ decreases with ρ, and $\lim_{\rho\to\infty} \gamma^\star(\rho) = 0$;*

(ii) *$x_i^\star(\rho)$ decreases with ρ, and $\lim_{\rho\to\infty} x_i^\star(\rho) = 0$ for all $i \le p$;*

(iii) *the eigenvalues of $X^\star$ are sorted in descending order, that is, $x_1^\star(\rho) \ge \cdots \ge x_p^\star(\rho)$ for every $\rho > 0$;*

(vi) *the condition number $x_1^\star(\rho)/x_p^\star(\rho)$ of $X^\star$ decreases with ρ, and $\lim_{\rho\to\infty} x_1^\star(\rho)/x_p^\star(\rho) = 1$.*

Proof. As the left-hand side of (1.16b) is strictly increasing in ρ, it is clear that $\gamma^\star(\rho)$ decreases with ρ. Moreover, the a priori bounds on $\gamma^\star(\rho)$ derived in Lemma 1.14 imply that

$$0 \le \lim_{\rho\to\infty} \gamma^\star(\rho) \le \lim_{\rho\to\infty} \frac{p}{\rho^2} = 0.$$

Thus, assertion (i) follows. Next, by the definition of the eigenvalue $x_i^\star$ in (1.16a), we have

$$\frac{\partial x_i^\star}{\partial \gamma^\star} = 1 + \lambda_i \gamma^\star - \frac{1}{2}\left(\sqrt{\lambda_i^2(\gamma^\star)^2 + 4\lambda_i\gamma^\star} + \frac{\lambda_i^2(\gamma^\star)^2 + 2\lambda_i\gamma^\star}{\sqrt{\lambda_i^2(\gamma^\star)^2 + 4\lambda_i\gamma^\star}} \right) = 1 + \lambda_i\gamma^\star - \frac{\lambda_i^2(\gamma^\star)^2 + 3\lambda_i\gamma^\star}{\sqrt{\lambda_i^2(\gamma^\star)^2 + 4\lambda_i\gamma^\star}}.$$

Elementary algebra indicates that $(1+z)\sqrt{z^2+4z} \geq z^2+3z$ for all $z \geq 0$, whereby the right-hand side of the above expression is strictly positive for every $\lambda_i \geq 0$ and $\gamma^\star \geq 0$. We conclude that $x_i^\star$ grows with $\gamma^\star$ and, by the monotonicity of $\gamma^\star(\rho)$ established in assertion (i), that $x_i^\star(\rho)$ decreases with ρ. As $\gamma^\star(\rho)$ drops to 0 for large ρ and as the continuous function (1.16a) evaluates to 0 at $\gamma^\star = 0$, we thus find that $x_i^\star(\rho)$ converges to 0 as ρ grows. These observations establish assertion (ii). As for assertion (iii), use (1.16a) to express the i-th eigenvalue of $X^\star$ as $x_i^\star = 1 - \frac{1}{2}\psi(\lambda_i)$, where the auxiliary function $\psi(\lambda) = \sqrt{\lambda^2(\gamma^\star)^2 + 4\lambda\gamma^\star} - \lambda\gamma^\star$ is defined for all $\lambda \geq 0$. Note that $\psi(\lambda)$ is monotonically increasing because

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\psi(\lambda) = \frac{\lambda(\gamma^\star)^2 + 2\gamma^\star}{\sqrt{\lambda^2(\gamma^\star)^2 + 4\lambda\gamma^\star}} - \gamma^\star = \gamma^\star\left(\frac{\lambda\gamma^\star + 2}{\sqrt{\lambda^2(\gamma^\star)^2 + 4\lambda\gamma^\star}} - 1\right) > 0.$$

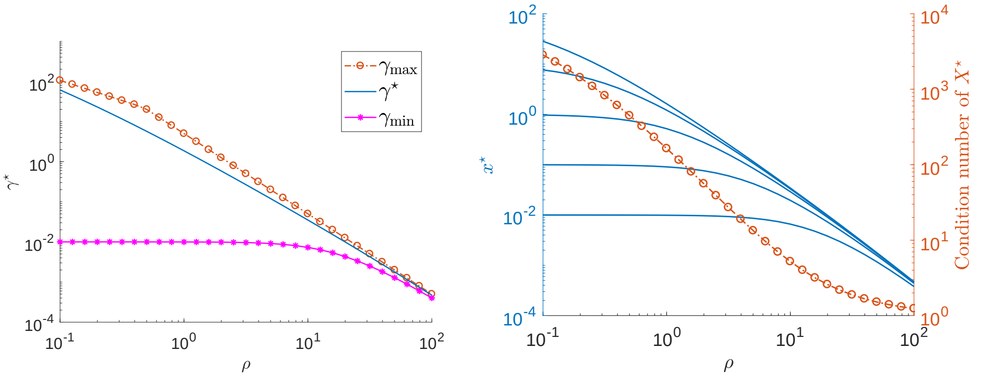
As $\lambda_{i+1} \geq \lambda_i$ for all $i < p$, we thus have $\psi(\lambda_{i+1}) \geq \psi(\lambda_i)$, which in turn implies that $x_{i+1}^\star \leq x_i^\star$. Hence, assertion (iii) follows. As for assertion (iv), note that by (1.16a) the condition number of $X^\star$ is given by

$$\frac{x_1^\star(\rho)}{x_p^\star(\rho)} = \frac{1 - \frac{1}{2}\left(\sqrt{\lambda_1^2\gamma^\star(\rho)^2 + 4\lambda_1\gamma^\star(\rho)} - \lambda_1\gamma^\star(\rho)\right)}{1 - \frac{1}{2}\left(\sqrt{\lambda_p^2\gamma^\star(\rho)^2 + 4\lambda_p\gamma^\star(\rho)} - \lambda_p\gamma^\star(\rho)\right)}.$$

The last expression converges to 1 as ρ tends to infinity because $\gamma^\star(\rho)$ vanishes asymptotically due to assertion (i). A tedious but straightforward calculation using (1.16a) shows that $\frac{\partial}{\partial\gamma^\star}\log(x_1^\star/x_p^\star) > 0$, which implies via the monotonicity of the logarithm that $x_1^\star/x_p^\star$ increases with $\gamma^\star$. As $\gamma^\star(\rho)$ decreases with ρ by virtue of assertion (i), we may then conclude that the condition number $x_1^\star(\rho)/x_p^\star(\rho)$ decreases with ρ. □

Figure 1.1 visualizes the dependence of $\gamma^\star$ and $X^\star$ on the Wasserstein radius ρ in an example where $p = 5$ and the eigenvalues of $\widehat{\Sigma}$ are given by $\lambda_i = 10^{i-3}$ for $i \leq 5$. Figure 1.1a displays $\gamma^\star$ as well as its a priori bounds $\gamma_{\min}$ and $\gamma_{\max}$ derived in Lemma 1.14. Note first that $\gamma^\star$ drops monotonically to 0 for large ρ, which is in line with Proposition 1.16(i). As $\gamma^\star$ represents the Lagrange multiplier of the Wasserstein constraint, which limits the size of the ambiguity set to ρ, this observation indicates that the worst-case expectation (1.7) displays a decreasing marginal increase in ρ. Figure 1.1b visualizes the eigenvalues $x_i^\star$, $i \leq 5$, as well as the condition number of $X^\star$. Note that all eigenvalues are monotonically shrunk towards 0 and that their order is preserved as ρ grows, which provides empirical support for Propositions 1.16(ii) and 1.16(iii), while the condition number decreases monotonically to 1, which corroborates Proposition 1.16(iv).

In summary, we have shown that $X^\star$ constitutes a nonlinear shrinkage estimator that is rotation equivariant, positive definite and well-conditioned. Moreover, $(X^\star)^{-1}$ preserves the order of the eigenvalues of $\widehat{\Sigma}$. We emphasize that neither the interpretation of $X^\star$ as a shrinkage estimator nor any of its desirable properties—most notably the improvement of its condition number with ρ—were dictated *ex ante*. Instead, these properties arose naturally from an intuitively appealing distributionally robust estimation scheme. In contrast, existing estimation



(a) Lagrange multiplier $\gamma^\star$ and its a priori bounds $\gamma_{\min}$ and $\gamma_{\max}$ from Lemma 1.14.

(b) Eigenvalues (left axis) and condition number (round marker - right axis) of $X^\star$.

Figure 1.1 – Dependence of the Lagrange multiplier $\gamma^\star$ (left panel) as well as the eigenvalues $x_i^\star$, $i \leq 5$, and the condition number $x_5^\star / x_1^\star$ of the optimal estimator $X^\star$ (right panel) on ρ.

schemes sometimes impose *ad hoc* constraints on condition numbers; see, *e.g.*, [173]. On the downside, as $X^\star$ shares the same eigenbasis as the sample covariance matrix $\widehat{\Sigma}$, it does not prompt a new robust principal component analysis. We henceforth refer to $X^\star$ as the *Wasserstein shrinkage estimator*.

1.4 Numerical Solution with Sparsity Information

We now investigate a more general setting where $\mathscr{X}$ may be a strict subset of $\mathbb{S}^p_{++}$, which captures a prescribed conditional independence structure of ξ. Specifically, we assume that there exists $\mathscr{E} \subseteq \{1,\ldots,p\}^2$ such that the random variables ξ_i and ξ_j are conditionally independent given $\xi_{-\{i,j\}}$ for any pair $(i,j) \in \mathscr{E}$, where $\xi_{-\{i,j\}}$ represents the truncation of the random vector ξ without the components ξ_i and ξ_j. It is well known that if ξ follows a normal distribution with covariance matrix $S \succ 0$ and precision matrix $X = S^{-1}$, then ξ_i and ξ_j are conditionally independent given $\xi_{-(i,j)}$ if and only if $X_{ij} = 0$. This reasoning forms the basis of the celebrated Gaussian graphical models, see, *e.g.*, [102]. Any prescribed conditional independence structure of ξ can thus conveniently be captured by the feasible set

$$\mathscr{X} = \{X \in \mathbb{S}^p_{++} : X_{ij} = 0 \quad \forall (i,j) \in \mathscr{E}\}.$$

We may assume without loss of generality that $\mathscr{E}$ inherits symmetry from X, that is, $(i,j) \in \mathscr{E} \implies (j,i) \in \mathscr{E}$. In Section 1.3 we have seen that the robust maximum likelihood estimation problem (1.5) admits an analytical solution when $\mathscr{E} = \emptyset$. In the general case, analytical tractability is lost. Indeed, if $\mathscr{E} \neq \emptyset$, then even the nominal estimation problem obtained by setting $\rho = 0$ requires numerical solution [34]. In this section we develop a Newton-type algorithm to solve (1.5) in the presence of prior conditional independence information. For the sake of consistency, we will refer to the optimal solution of problem (1.5) as the *Wasserstein* *shrinkage estimator* even in the presence of sparsity constraints.

Remark 1.17 (Conditional independence information in $\mathbb{B}_\rho$). *We emphasize that our proposed estimation model accounts for the prescribed conditional independence structure only in the feasible set $\mathcal{X}$ but not in the ambiguity set $\mathbb{B}_\rho$. Otherwise, the ambiguity set would have to be redefined as*

$$\mathbb{B}_\rho = \left\{ \mathbb{Q} \in \mathcal{N}_0^p \; : \; \mathbb{W}(\mathbb{Q}, \widehat{\mathbb{P}}) \leq \rho, \; (\mathbb{E}_\mathbb{Q}[\xi\xi^\top]^{-1})_{ij} = 0 \quad \forall (i,j) \in \mathcal{E} \right\}.$$

While conceptually attractive, this new ambiguity set is empty even for some $\rho > 0$ because the inverse sample covariance matrix $\widehat{\Sigma}^{-1}$ violates the prescribed conditional independence relationships with probability 1.

Recall from Theorem 1.7 that the estimation problem (1.5) is equivalent to the convex program (1.6) and that the optimal value of (1.6) depends continuously on $\widehat{\Sigma} \in \mathbb{S}_+^p$. In the remainder of this section we may thus assume without much loss of generality that $\widehat{\Sigma} \succ 0$. Otherwise, we can replace $\widehat{\Sigma}$ with $\widehat{\Sigma} + \varepsilon I$ for some small $\varepsilon > 0$ without significantly changing the estimation problem's solution. Inspired by [130, 84], we now develop a sequential quadratic approximation algorithm for solving problem (1.6) with sparsity information. Note that the set $\mathcal{X}$ of feasible precision matrices typically fixes many entries to zero, thus reducing the effective problem dimension and making a second-order algorithm attractive even for large instances of (1.6).

The proposed algorithm starts at $X_0 = I$ and at some $\gamma_0 > 1$, which are trivially feasible in (1.6). In each iteration the algorithm moves from the current iterate (X_t, γ_t) along a feasible descent direction, which is constructed from a quadratic approximation of the objective function of problem (1.6). A judiciously chosen step size guarantees that the next iterate (X_{t+1}, γ_{t+1}) remains feasible and has a better (lower) objective value; see Algorithm 1. The construction of the descent direction relies on the following lemma.

Lemma 1.18 (Fact 7.4.9 in [12]). *For any $A, B \in \mathbb{R}^{p \times p}$ and $X \in \mathbb{S}^p$, we have*

$$\operatorname{Tr}\left[AXBX\right] = \operatorname{vec}(X)^\top (B \otimes A^\top) \operatorname{vec}(X).$$

Proposition 1.19 (Descent direction). *Fix $(X, \gamma) \in \mathbb{S}_{++}^p \times \mathbb{R}_{++}$ with $\gamma I \succ X$, and define the orthogonal projection $P : \mathbb{R}^{p^2+1} \to \mathbb{R}^{p^2+1}$ through $(Pz)_k = 0$ if $k = p(j-1) + i$ for some $(i,j) \in \mathcal{E}$; $= \frac{1}{2} z_{p(j-1)+i} + \frac{1}{2} z_{p(i-1)+j}$ if $k = p(j-1) + i$ for some $i, j \leq p$ with $(i,j) \notin \mathcal{E}$; $= z_k$ if $k = p^2 + 1$. Moreover, define $G \triangleq I - \frac{X}{\gamma}$,*

$$H \triangleq \begin{bmatrix} X^{-1} \otimes X^{-1} + \frac{2}{\gamma} G^{-1} \widehat{\Sigma} G^{-1} \otimes G^{-1} & -\frac{1}{\gamma^2} \operatorname{vec}(G^{-1}[XG^{-1}\widehat{\Sigma} + \widehat{\Sigma} G^{-1} X] G^{-1}) \\ -\frac{1}{\gamma^2} \operatorname{vec}(G^{-1}[XG^{-1}\widehat{\Sigma} + \widehat{\Sigma} G^{-1} X] G^{-1})^\top & \frac{2}{\gamma^3} \operatorname{Tr}[G^{-1} X G^{-1} \widehat{\Sigma} G^{-1} X] \end{bmatrix} \in \mathbb{S}^{p^2+1}$$

and

$$g \triangleq \begin{bmatrix} \operatorname{vec}(G^{-1}\widehat{\Sigma}G^{-1} - X^{-1}) \\ \rho^2 + \operatorname{Tr}[G^{-1}\widehat{\Sigma}(I - \frac{1}{\gamma}G^{-1}X) - \widehat{\Sigma}] \end{bmatrix} \in \mathbb{R}^{p^2+1}.$$

Then, the unique solution $(\Delta_X^\star, \Delta_\gamma^\star) \in \mathbb{S}^p \times \mathbb{R}$ *of the linear system*

$$PH\Big((\operatorname{vec}(\Delta_X^\star)^\top, \Delta_\gamma^\star)^\top + g\Big) = 0 \quad \textit{and} \quad (\Delta_X^\star)_{ij} = 0 \quad \forall (i,j) \in \mathscr{E} \tag{1.19}$$

represents a feasible descent direction for the optimization problem (1.6) *at* (X, γ).

Proof. We first expand the objective function of problem (1.6) around $(X,\gamma) \in \mathbb{S}^p_{++} \times \mathbb{R}_{++}$ with $\gamma I \succ X$. By the rules of matrix calculus, the second-order Taylor expansion of the negative log-determinant is given by

$$-\log\det(X + \Delta_X) = -\log\det(X) - \operatorname{Tr}\left[X^{-1}\Delta_X\right] + \frac{1}{2}\operatorname{Tr}\left[X^{-1}\Delta_X X^{-1}\Delta_X\right] + \mathscr{O}(\|\Delta_X\|^3)$$

for $\Delta_X \in \mathbb{S}^p$, see also [18, page 644]. Moreover, by using a geometric series expansion, we obtain

$$\begin{aligned} \left(I - \frac{X + \Delta_X}{\gamma + \Delta_\gamma}\right)^{-1} &= \left(I - \frac{X+\Delta_X}{\gamma}\left(1 - \frac{\Delta_\gamma}{\gamma} + \frac{\Delta_\gamma^2}{\gamma^2} + \mathscr{O}(\|\Delta_\gamma\|^3)\right)\right)^{-1} \\ &= \left(I - \frac{X}{\gamma} + \frac{X\Delta_\gamma}{\gamma^2} - \frac{X\Delta_\gamma^2}{\gamma^3} - \frac{\Delta_X}{\gamma} + \frac{\Delta_X\Delta_\gamma}{\gamma^2} + \mathscr{O}(\|(\Delta_X,\Delta_\gamma)\|^3)\right)^{-1} \end{aligned}$$

for $\Delta_\gamma \in \mathbb{R}$. Expanding the matrix inverse as a Neumann series and setting $G = I - \frac{X}{\gamma}$, which is invertible because $\gamma I \succ X$, the above expression can be reformulated as

$$\begin{aligned} &G^{-\frac{1}{2}}\left(I + \tfrac{G^{-\frac{1}{2}}XG^{-\frac{1}{2}}\Delta_\gamma}{\gamma^2} - \tfrac{G^{-\frac{1}{2}}XG^{-\frac{1}{2}}\Delta_\gamma^2}{\gamma^3} - \tfrac{G^{-\frac{1}{2}}\Delta_X G^{-\frac{1}{2}}}{\gamma} + \tfrac{G^{-\frac{1}{2}}\Delta_X G^{-\frac{1}{2}}\Delta_\gamma}{\gamma^2} + \mathscr{O}(\|(\Delta_X,\Delta_\gamma)\|^3)\right)^{-1} G^{-\frac{1}{2}} \\ &= G^{-1} - \frac{G^{-1}XG^{-1}\Delta_\gamma}{\gamma^2} + \frac{G^{-1}XG^{-1}\Delta_\gamma^2}{\gamma^3} + \frac{G^{-1}\Delta_X G^{-1}}{\gamma} - \frac{G^{-1}\Delta_X G^{-1}\Delta_X G^{-1}\Delta_\gamma}{\gamma^2} + \frac{G^{-1}XG^{-1}XG^{-1}\Delta_\gamma^2}{\gamma^4} \\ &\quad + \frac{G^{-1}\Delta_X G^{-1}\Delta_X G^{-1}}{\gamma^2} - \frac{G^{-1}XG^{-1}\Delta_X G^{-1}\Delta_\gamma}{\gamma^3} - \frac{G^{-1}\Delta_X G^{-1}XG^{-1}\Delta_\gamma}{\gamma^3} + \mathscr{O}(\|(\Delta_X,\Delta_\gamma)\|^3). \end{aligned}$$

Thus, the second-order Taylor expansion of the last term in the objective function of (1.6) is given by

$$
\begin{aligned}
&(\gamma+\Delta_\gamma)^2\,\mathrm{Tr}\left[\left((\gamma+\Delta_\gamma)I-(X+\Delta_X)\right)^{-1}\widehat{\Sigma}\right]=(\gamma+\Delta_\gamma)\,\mathrm{Tr}\left[\left(I-\frac{X+\Delta_X}{\gamma+\Delta_\gamma}\right)^{-1}\widehat{\Sigma}\right]\\
&=\gamma\,\mathrm{Tr}\left[G^{-1}\widehat{\Sigma}\right]+\Delta_\gamma\,\mathrm{Tr}\left[G^{-1}\widehat{\Sigma}(I-\frac{1}{\gamma}G^{-1}X)\right]+\frac{\Delta_\gamma^2}{\gamma^3}\,\mathrm{Tr}\left[G^{-1}XG^{-1}\widehat{\Sigma}G^{-1}X\right]\\
&\quad+\mathrm{Tr}\left[G^{-1}\widehat{\Sigma}G^{-1}\Delta_X\right]-\frac{\Delta_\gamma}{\gamma^2}\,\mathrm{Tr}\left[G^{-1}\widehat{\Sigma}G^{-1}\Delta_XG^{-1}X+G^{-1}\widehat{\Sigma}G^{-1}XG^{-1}\Delta_X\right]\\
&\quad+\frac{1}{\gamma}\,\mathrm{Tr}\left[G^{-1}\Delta_XG^{-1}\widehat{\Sigma}G^{-1}\Delta_X\right]+\mathscr{O}(\|(\Delta_X,\Delta_\gamma)\|^3),
\end{aligned}
$$

where the second equality follows from the Taylor expansion of the matrix inverse derived above. Using Lemma 1.18, the objective function of (1.6) is thus representable as

$$
\begin{aligned}
&-\log\det(X+\Delta_X)+(\gamma+\Delta_\gamma)\left(\rho^2-\mathrm{Tr}\left[\widehat{\Sigma}\right]\right)+(\gamma+\Delta_\gamma)^2\,\mathrm{Tr}\left[\left((\gamma+\Delta_\gamma)I-(X+\Delta_X)\right)^{-1}\widehat{\Sigma}\right]\\
&\qquad=c+g^\top(\mathrm{vec}(\Delta_X)^\top,\Delta_\gamma)^\top+\frac{1}{2}(\mathrm{vec}(\Delta_X)^\top,\Delta_\gamma)H(\mathrm{vec}(\Delta_X)^\top,\Delta_\gamma)^\top+\mathscr{O}(\|(\Delta_X,\Delta_\gamma)\|^3)
\end{aligned}
$$

for some $c\in\mathbb{R}$, where the gradient $g\in\mathbb{R}^p$ and the Hessian $H\in\mathbb{S}^p$ are defined as in the proposition statement. A feasible descent direction for problem (1.6) is thus obtained by solving the auxiliary quadratic program

$$
\begin{array}{cl}
\min\limits_{\Delta_X,\Delta_\gamma} & g^\top(\mathrm{vec}(\Delta_X)^\top,\Delta_\gamma)^\top+\frac{1}{2}(\mathrm{vec}(\Delta_X)^\top,\Delta_\gamma)H(\mathrm{vec}(\Delta_X)^\top,\Delta_\gamma)^\top\\
\mathrm{s.t.} & \Delta_X\in\mathbb{S}^p,\ (\Delta_X)_{ij}=0\quad\forall(i,j)\in\mathscr{E}
\end{array}
\tag{1.20}
$$

Note that (1.20) has a unique minimizer because H is positive definite. Indeed, we have

$$
\begin{aligned}
&\frac{4}{\gamma^4}\mathrm{vec}(G^{-1}XG^{-1}\widehat{\Sigma}G^{-1})^\top\left(X^{-1}\otimes X^{-1}+\frac{2}{\gamma}G^{-1}\widehat{\Sigma}G^{-1}\otimes G^{-1}\right)^{-1}\mathrm{vec}(G^{-1}XG^{-1}\widehat{\Sigma}G^{-1})\\
<&\frac{4}{\gamma^4}\mathrm{vec}(G^{-1}XG^{-1}\widehat{\Sigma}G^{-1})^\top\left(\frac{2}{\gamma}G^{-1}\widehat{\Sigma}G^{-1}\otimes G^{-1}\right)^{-1}\mathrm{vec}(G^{-1}XG^{-1}\widehat{\Sigma}G^{-1})\\
=&\frac{2}{\gamma^3}\mathrm{vec}(G^{-1}XG^{-1}\widehat{\Sigma}G^{-1})^\top\left(G\widehat{\Sigma}^{-1}G\otimes G\right)\mathrm{vec}(G^{-1}XG^{-1}\widehat{\Sigma}G^{-1})\\
=&\frac{2}{\gamma^3}\,\mathrm{Tr}\left[G^{-1}XG^{-1}\widehat{\Sigma}G^{-1}X\right],
\end{aligned}
$$

where the inequality holds because $X\otimes X$ is positive definite and $G^{-1}XG^{-1}\widehat{\Sigma}G^{-1}\neq0$, the first equality follows from [12, Proposition 7.1.7], which asserts that $(A\otimes B)^{-1}=A^{-1}\otimes B^{-1}$ for any $A,B\in\mathbb{S}^p_{++}$, and the second equality follows from Lemma 1.18. The above derivation shows that the Schur complement of the positive definite block $X^{-1}\otimes X^{-1}+\frac{2}{\gamma}G^{-1}\widehat{\Sigma}G^{-1}\otimes G^{-1}$ in H is a positive number, which in turn implies that the Hessian H is positive definite. In the following, we denote the unique minimizer of (1.20) by $(\Delta_X^\star,\Delta_\gamma^\star)$. As $\Delta_X=0$ and $\Delta_\gamma=0$ is feasible in (1.20), it is clear that the objective value of $(\Delta_X^\star,\Delta_\gamma^\star)$ is nonpositive. In fact, as $H\succ0$, the minimum of (1.20) is negative unless $g=0$. Thus, $(\Delta_X^\star,\Delta_\gamma^\star)$ is a feasible descent direction.

Note that P defined in the proposition statement represents the orthogonal projection on the linear space

$$\mathcal{Z} = \left\{ z = (\text{vec}(\Delta_X)^\top, \Delta_\gamma)^\top \in \mathbb{R}^{p^2+1} : \Delta_X \in \mathbb{S}^p, \quad (\Delta_X)_{ij} = 0 \quad \forall (i,j) \in \mathcal{E} \right\}.$$

Indeed, it is easy to verify that $P^2 = P = P^\top$ because the range and the null space of P correspond to $\mathcal{Z}$ and its orthogonal complement, respectively. The quadratic program (1.20) is thus equivalent to

$$\min_{z \in \mathcal{Z}} \left\{ g^\top z + \frac{1}{2} z^\top H z \right\} = \min_{z \in \mathbb{R}^{p^2+1}} \left\{ g^\top z + \frac{1}{2} z^\top H z : Pz = z \right\}.$$

The minimizer $z^\star$ of the last reformulation and the optimal Lagrange multiplier $\mu^\star$ associated with its equality constraint correspond to the unique solution of the Karush-Kuhn-Tucker optimality conditions

$$Hz^\star + g + (I-P)\mu^\star = 0, \ (1-P)z^\star = 0 \quad \Longleftrightarrow \quad P(Hz^\star + g) = 0, \ (1-P)z^\star = 0,$$

which are mainfestly equivalent to (1.19). Thus, the claim follows. □

Given a descent direction $(\Delta_X^\star, \Delta_\gamma^\star)$ at a feasible point (X, γ), we use a variant of Armijo's rule [126, Section 3.1] to choose a step size $\alpha > 0$ that preserves feasibility of the next iterate $(X + \alpha\Delta_X^\star, \gamma + \alpha\Delta_\gamma^\star)$ and ensures a sufficient decrease of the objective function. Specifically, for a prescribed line search parameter $\sigma \in (0, \frac{1}{2})$, we set the step size α to the largest number in $\{\frac{1}{2^m}\}_{m \in \mathbb{Z}_+}$ satisfying the following two conditions:

(C1) Feasibility: $(\gamma + \alpha\Delta_\gamma^\star)I \succ X + \alpha\Delta_X^\star \succ 0$;

(C2) Sufficient decrease: $f(X+\alpha\Delta_X^\star, \gamma+\alpha\Delta_\gamma^\star) \le f(X,\gamma) + \sigma\alpha\delta$, where $\delta = g^\top(\text{vec}(\Delta_X^\star)^\top, \Delta_\gamma^\star)^\top < 0$, and g is defined as in Propostion 1.19.

Notice that the sparsity constraints are automatically satisfied at the next iterate thanks to the construction of the descent direction $(\Delta_X^\star, \Delta_\gamma^\star)$ in (1.19). Algorithm 1 repeats the procedure outlined above until $\|g\|$ drops below a given tolerance (10^{-3}) or until the iteration count exceeds a given threshold (10^2). Throughout the numerical experiments in Section 1.6 we set $\sigma = 10^{-4}$, which is the value recommended in [126].

Algorithm 1 Sequential quadratic approximation algorithm

```
Input: Sample covariance matrix $\widehat{\Sigma} \succ 0$, Wasserstein radius $\rho > 0$,
       line search parameter $\sigma \in (0, \frac{1}{2})$.
  Initialize $X_0 = I$ and $\gamma_0 > 1$, and set $t \leftarrow 0$
  while stopping criterion is violated do
    Find the descent direction $(\Delta_X^\star, \Delta_\gamma^\star)$ at $(X,\gamma) = (X_t, \gamma_t)$ by solving (1.19);
    Find the largest step size $\alpha_t \in \{\frac{1}{2^m}\}_{m \in \mathbb{Z}_+}$ satisfying (C1) and (C2);
    Set $X_{t+1} = X_t + \alpha_t \Delta_X^\star$, $\gamma_{t+1} = \gamma_t + \alpha_t \Delta_\gamma^\star$;
    Set $t \leftarrow t+1$;
  end while
Output: $X_t$
```

Remark 1.20 (Steepest descent algorithm)**.** *The computation of the descent direction in Proposition 1.19 requires second-order information. It is easy to verify that Proposition 1.19 remains valid if the Hessian H is replaced with the identity matrix, in which case the sequential quadratic approximation algorithm reduces to the classical steepest descent algorithm [126, Chapter 3].*

The next proposition establishes that Algorithm 1 converges to the unique minimizer of problem (1.6).

Proposition 1.21 (Convergence)**.** *Assume that* $\widehat{\Sigma} \succ 0$*,* $\rho > 0$ *and* $\sigma \in (0, \frac{1}{2})$*. For any initial feasible solution* (X_0, γ_0)*, the sequence* $\{(X_t, \gamma_t)\}_{t \in \mathbb{Z}_+}$ *generated by Algorithm 1 converges to the unique minimizer* $(X^\star, \gamma^\star)$ *of problem* (1.6)*. Moreover, the sequence converges locally quadratically.*

Proof. Denote by $f(X,\gamma)$ the objective function of problem (1.6), and define

$$\mathscr{C} \triangleq \left\{(X,\gamma) \in \mathscr{X} \times \mathbb{R}_+ : f(X,\gamma) \le f(X_0,\gamma_0),\ 0 \prec X \prec \gamma I\right\}$$

as the set of all feasible solutions that are at least as good as the initial solution (X_0, γ_0). The proof of Theorem 1.7 implies that $\underline{x} I \preceq X \preceq \overline{x} I$ and $\underline{x} \le \gamma \le \overline{x}$ for all $(X,\gamma) \in \mathscr{C}$, where the strictly positive constants $\underline{x}$ and $\overline{x}$ are defined as in (1.13). Note that, as $\widehat{\Sigma}$ is fixed in this proof, the dependence of $\underline{x}$ and $\overline{x}$ on $\widehat{\Sigma}$ is notationally suppressed to avoid clutter. Thus, $\mathscr{C}$ is bounded. Moreover, as $\widehat{\Sigma} \succ 0$, it is easy to verify $f(X,\gamma)$ tends to infinity if the smallest eigenvalue of X approaches 0 or if the largest eigenvalue of X approaches γ. The continuity of $f(X,\gamma)$ then implies that $\mathscr{C}$ is closed. In summary, we conclude that $\mathscr{C}$ is compact.

By the definition of $f(X,\gamma)$ in (1.6), any $(X,\gamma) \in \mathscr{C}$ satisfies

$$\begin{aligned} 0 &\le f(X_0,\gamma_0) + \log\det(X) - \gamma\left(\rho^2 - \mathrm{Tr}\left[\widehat{\Sigma}\right]\right) - \gamma\left\langle (I - \gamma^{-1}X)^{-1}, \widehat{\Sigma}\right\rangle \\ &\le f(X_0,\gamma_0) + p\log(\overline{x}) + \overline{x}\,\mathrm{Tr}\left[\widehat{\Sigma}\right] - \underline{x}\,\lambda_{\min}\,\mathrm{Tr}\left[(I - \gamma^{-1}X)^{-1}\right], \end{aligned}$$

where $\lambda_{\min}$ denotes the smallest eigenvalue of $\widehat{\Sigma}$, which is positive by assumption. Thus, we have

$$\operatorname{Tr}\left[(I-\gamma^{-1}X)^{-1}\right] \le \frac{1}{\underline{x}\lambda_{\min}}\left(f(X_0,\gamma_0)+p\log(\overline{x})+\overline{x}\operatorname{Tr}\left[\widehat{\Sigma}\right]\right),$$

which implies that the eigenvalues of $I-\frac{X}{\gamma}$ are uniformly bounded away from 0 on $\mathscr{C}$. More formally, there exists $c_0>0$ with $I-\frac{X}{\gamma}\succ c_0 I$ for all $(X,\gamma)\in\mathscr{C}$. As the objective function $f(X,\gamma)$ is smooth wherever it is defined, its gradient and Hessian constitute continuous functions on $\mathscr{C}$. Moreover, as $f(X,\gamma)$ is strictly convex on the compact set $\mathscr{C}$, the eigenvalues of its Hessian matrix are uniformly bounded away from 0. This implies that the inverse Hessian matrix and the descent direction $(\Delta_X^\star,\Delta_\gamma^\star)$ constructed in Proposition 1.19 are also continuous on $\mathscr{C}$. Hence, there exist $c_1,c_2>0$ such that $\Delta_X^\star\preceq c_1 I$ and $|\Delta_\gamma^\star|\le c_2$ uniformly on $\mathscr{C}$.

We conclude that any positive step size $\alpha<\underline{x}\min\left\{c_1^{-1},(c_1+c_2)^{-1}c_0\right\}$ satisfies the feasibility condition (C1) uniformly on $\mathscr{C}$ because $X+\alpha\Delta_X^\star\succ\left(\underline{x}-\alpha c_1\right)I\succeq 0$ and

$$(\gamma+\alpha\Delta_\gamma^\star)I\succeq X+c_0\underline{x}I+\alpha\left(\Delta_X^\star-\Delta_X^\star+\Delta_\gamma^\star I\right)\succeq X+c_0\underline{x}I+\alpha\left(\Delta_X^\star-(c_1+c_2)I\right)\succ X+\alpha\Delta_X^\star$$

for all $(X,\gamma)\in\mathscr{C}$. Moreover, by [167, Lemma 5(b)] there exists $\overline{\alpha}>0$ such that any positive step size $\alpha\le\overline{\alpha}$ satisfies the descent condition (C2) for all $(X,\gamma)\in\mathscr{C}$. In summary, there exists $m^\star\in\mathbb{Z}_+$ such that

$$\alpha^\star=\frac{1}{2^{m^\star}}<\min\left\{\overline{\alpha},\underline{x}\min\left\{c_1^{-1},(c_1+c_2)^{-1}c_0\right\}\right\}$$

satisfies both line search conditions (C1) and (C2) uniformly on $\mathscr{C}$. By induction, the iterates $\{(X_t,\gamma_t)\}_{t\in\mathbb{N}}$ generated by Algorithm 1 have nonincreasing objective values and thus all belong to $\mathscr{C}$, while the step sizes $\{\alpha_t\}_{t\in\mathbb{N}}$ generated by Algorithm 1 are all larger or equal to $\alpha^\star$. Hence, the algorithm's global convergence is guaranteed by [167, Theorem 1], while the local quadratic convergence follows from [84, Theorem 16]. □

Remark 1.22 (Refinements of Algorithm 1)**.** *For large values of p, computing and storing the exact Hessian matrix H from Proposition 1.19 is prohibitive. In this case, H can be approximated by a low-rank matrix as in the limited-memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) method without sacrificing global convergence [167]. Alternatively, one can resort to a coordinate descent method akin to the QUIC algorithm [84], in which case both the global and local convergence guarantees of Proposition 1.21 remain valid.*

Remark 1.23 (Learning the sparsity pattern)**.** *If the precision matrix is known to be sparse but has an unknown sparsity pattern, then one may set $\mathscr{X}=\mathbb{S}_{++}^d$ and add a weighted ℓ_1-regularization or Lasso term to the objective function of problem* (1.6) *in order to generate sparse Wasserstein shrinkage estimators. Different sparsity patterns can be obtained by tuning the weight of the Lasso term. The regularized nonlinear SDP* (1.6) *can then be solved with a variant of the QUIC algorithm [84]. Indeed, the gradient and the Hessian matrix of the smooth part of the objective function (which coincides with the robust version of Stein's loss function) can again be computed efficiently by leveraging Proposition 1.19. Details are omitted for brevity.*

1.5 Extremal Distributions

It is instructive to characterize the extremal distributions that attain the supremum in (1.7) for a given sample covariance matrix $\widehat{\Sigma}$ and a fixed candidate estimator X.

Theorem 1.24 (Extremal distributions). *For any $\widehat{\Sigma}, X \in \mathbb{S}^p_{++}$ and $\rho > 0$, the supremum in* (1.7) *is attained by the normal distribution $\mathbb{Q}^\star = \mathcal{N}(0, S^\star)$ with covariance matrix*

$$S^\star = (\gamma^\star)^2(\gamma^\star I - X)^{-1}\widehat{\Sigma}(\gamma^\star I - X)^{-1},$$

where $\gamma^\star$ is the unique solution with $\gamma^\star I \succ X$ of the following algebraic equation

$$\rho^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big] + 2\gamma^\star\,\mathrm{Tr}\big[(\gamma^\star I - X)^{-1}\widehat{\Sigma}\big] - (\gamma^\star)^2\,\mathrm{Tr}\big[(\gamma^\star I - X)^{-1}\widehat{\Sigma}(\gamma^\star I - X)^{-1}\big] = 0. \tag{1.21}$$

Proof. From Proposition 1.9 we know that the worst-case expectation problem (1.7) is equivalent to the semidefinite program (1.8). Note that the strictly convex objective function of (1.8) is bounded below by

$$\gamma\big(\rho^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\big) + \lambda_{\min}\gamma^2\,\mathrm{Tr}\big[(\gamma I - X)^{-1}\big],$$

where $\lambda_{\min}$ denotes the smallest eigenvalue of $\widehat{\Sigma}$. As $\lambda_{\min}$ is positive by assumption, the objective function of (1.8) tends to infinity as γ approaches the largest eigenvalue of X, in which case $\gamma I - X$ becomes singular. Thus, the unique optimal solution $\gamma^\star$ of (1.8) satisfies $\gamma^\star I \succ X$ and solves the first-order optimality condition (1.21).

Now we are ready to prove that $\mathbb{Q}^\star$ is both feasible and optimal in (1.7). By the formula for $S^\star$ in terms of $\gamma^\star$, $\widehat{\Sigma}$ and S and by using Definition 1.4 and Proposition 1.3, it is easy to verify that (1.21) is equivalent to

$$\mathrm{Tr}\big[S^\star\big] + \mathrm{Tr}\big[\widehat{\Sigma}\big] - 2\,\mathrm{Tr}\Big[\sqrt{\widehat{\Sigma}^{\frac{1}{2}} S^\star \widehat{\Sigma}^{\frac{1}{2}}}\Big] = \rho^2 \quad\Longleftrightarrow\quad \mathbb{W}_S(S^\star, \widehat{\Sigma}) = \mathbb{W}(\mathbb{Q}^\star, \widehat{\mathbb{P}}) = \rho,$$

which confirms that $\mathbb{Q}^\star$ is feasible in (1.7). Moreover, the objective value of $\mathbb{Q}^\star$ in (1.7) amounts to

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}^\star}[\langle \xi\xi^\top, X\rangle] &= \langle S^\star, X\rangle = (\gamma^\star)^2\big\langle(\gamma^\star I - X)^{-1}\widehat{\Sigma}(\gamma^\star I - X)^{-1}, X\big\rangle \\
&= (\gamma^\star)^2\big\langle(\gamma^\star I - X)^{-1}\widehat{\Sigma}(\gamma^\star I - X)^{-1}, (X - \gamma^\star I) + \gamma^\star I\big\rangle \\
&= -(\gamma^\star)^2\,\mathrm{Tr}\big[(\gamma^\star I - X)^{-1}\widehat{\Sigma}\big] + (\gamma^\star)^3\,\mathrm{Tr}\big[(\gamma^\star I - X)^{-1}\widehat{\Sigma}(\gamma^\star I - X)^{-1}\big] \\
&= \gamma^\star(\rho^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]) + (\gamma^\star)^2\big\langle(\gamma^\star I - X)^{-1}, \widehat{\Sigma}\big\rangle = g(\widehat{\Sigma}, X),
\end{aligned}$$

where the penultimate equality exploits (1.21), while the last equality follows from the optimality of $\gamma^\star$ in (1.8) and from Proposition 1.9. Thus, $\mathbb{Q}^\star$ is optimal in (1.7). □

In the absence of sparsity information (that is, if $\mathcal{X} = \mathbb{S}^p_{++}$), the unique minimizer $X^\star$ of problem (1.5) is available in closed form thanks to Theorem 1.12. In this case, the extremal distribution attaining the supremum in (1.7) at $X = X^\star$ can also be computed in closed form even if $\widehat{\Sigma}$ is rank deficient.

Corollary 1.25 (Extremal distribution for optimal estimator)**.** *Assume that* $\rho > 0$, $\mathcal{X} = \mathbb{S}^p_{++}$ *and* $\widehat{\Sigma} \in \mathbb{S}^p_+$ *admits the spectral decomposition* $\widehat{\Sigma} = \sum_{i=1}^p \lambda_i v_i v_i^\top$ *with eigenvalues* λ_i *and corresponding orthonormal eigenvectors* v_i, $i \le p$*. If* $(X^\star, \gamma^\star)$ *represents the unique solution of* (1.6) *given in Theorem 1.12, then the supremum in* (1.7) *at* $X = X^\star$ *is attained by the normal distribution* $\mathbb{Q}^\star = \mathcal{N}(0, S^\star)$ *with covariance matrix*

$$S^\star = \sum_{i=1}^p s_i^\star v_i v_i^\top, \quad \textit{where} \quad s_i^\star = \begin{cases} (\gamma^\star)^2 \lambda_i (\gamma^\star - x_i^\star)^{-2} & \textit{if } \lambda_i > 0, \\ (\gamma^\star)^{-1} & \textit{if } \lambda_i = 0. \end{cases}$$

Proof. If $\widehat{\Sigma} \succ 0$, the claim follows immediately by substituting the formula for $X^\star$ from Theorem 1.12 into the formula for $S^\star$ from Theorem 1.24. If $\widehat{\Sigma} \succeq 0$ is rank deficient, we consider the invertible sample covariance matrix $\widehat{\Sigma} + \varepsilon I \succ 0$ for some $\varepsilon > 0$, denote by $(X^\star(\varepsilon), \gamma^\star(\varepsilon))$ the corresponding minimizer of problem (1.6) as constructed in (1.16) and let $S^\star(\varepsilon)$ be the covariance matrix of the extremal distribution of problem (1.7) at $X = X^\star(\varepsilon)$. Using the same reasoning as in the proof of Theorem 1.12, one can show that $(X^\star(\varepsilon), \gamma^\star(\varepsilon))$ is continuous in $\varepsilon \in \mathbb{R}_+$ and converges to $(X^\star, \gamma^\star)$ as ε tends to 0. Similarly, $S^\star(\varepsilon)$ is continuous in $\varepsilon \in \mathbb{R}_+$ and converges to $S^\star$ as ε tends to 0. To see this, note that the eigenvalues $s_i^\star(\varepsilon)$, $i \le p$, of $S^\star(\varepsilon)$ satisfy

$$\begin{aligned} \lim_{\varepsilon \to 0^+} s_i^\star(\varepsilon) &= \lim_{\varepsilon \to 0^+} \frac{\gamma^\star(\varepsilon)^2 (\lambda_i + \varepsilon)}{(\gamma^\star(\varepsilon) - x_i^\star(\varepsilon))^2} \\ &= \lim_{\varepsilon \to 0^+} \frac{4(\lambda_i + \varepsilon)}{\left(\sqrt{(\lambda_i + \varepsilon)^2 \gamma^\star(\varepsilon)^2 + 4(\lambda_i + \varepsilon)\gamma^\star(\varepsilon)} - (\lambda_i + \varepsilon)\gamma^\star(\varepsilon)\right)^2} = s_i^\star \quad \forall i \le p, \end{aligned}$$

where the first equality follows from the first part of the proof, the second equality exploits (1.16a) and the third equality holds due to the definition of $s_i^\star$.

We are now armed to prove that $\mathbb{Q}^\star$ is both feasible and optimal in (1.7). Indeed, using the continuity of $S^\star(\varepsilon)$ and $\mathbb{W}_S(S_1, S_2)$ in their respective arguments, we find

$$\mathbb{W}(\mathbb{Q}^\star, \widehat{\mathbb{P}}) = \mathbb{W}_S(S^\star, \widehat{\Sigma}) = \lim_{\varepsilon \to 0^+} \mathbb{W}_S(S^\star(\varepsilon), \widehat{\Sigma} + \varepsilon I) = \rho,$$

where the last equality follows from the construction of $S^\star(\varepsilon)$ in the proof of Theorem 1.24. Thus, $\mathbb{Q}^\star$ is feasible in (1.7). Similarly, using the continuity of $S^\star(\varepsilon)$ and $X^\star(\varepsilon)$ in ε, we have

$$\mathbb{E}_{\mathbb{Q}^\star}[\langle \xi\xi^\top, X^\star \rangle] = \langle S^\star, X^\star \rangle = \lim_{\varepsilon \to 0^+} \langle S^\star(\varepsilon), X^\star(\varepsilon) \rangle = \lim_{\varepsilon \to 0^+} g(\widehat{\Sigma} + \varepsilon I, S^\star(\varepsilon)) = g(\widehat{\Sigma}, S^\star),$$

where the last two equalities follow from the construction of $S^\star(\varepsilon)$ in the proof of Theorem 1.24 and the continuity of $g(\widehat{\Sigma}, X)$ established in Proposition 1.9, respectively. Thus, $\mathbb{Q}^\star$ is optimal in (1.7). □

1.6 Numerical Experiments

To assess the statistical and computational properties of the proposed Wasserstein shrinkage estimator, we compare it against two state-of-the-art precision matrix estimators from the literature.

Definition 1.26 (Linear shrinkage estimator)**.** *Denote by* $\operatorname{diag}(\widehat{\Sigma})$ *the diagonal matrix of all sample variances. Then, the linear shrinkage estimator with mixing parameter* $\alpha \in [0,1]$ *is defined as*

$$X^\star = \left[(1-\alpha)\widehat{\Sigma} + \alpha \operatorname{diag}(\widehat{\Sigma})\right]^{-1}.$$

The linear shrinkage estimator uses the diagonal matrix of sample variances as the shrinkage target. Thus, the sample covariances are shrunk to zero, while the sample variances are preserved. We emphasize that the most prevalent shrinkage target is a scaled identity matrix [105]. The benefits of using $\operatorname{diag}(\widehat{\Sigma})$ instead are discussed in [148, § 2.4]. This particular linear shrinkage estimator can also be interpreted as the maximum a posteriori estimator based on an inverse Wishart prior [120, § 4.2.6]. Note that while $\widehat{\Sigma}$ is never invertible for $n < p$, $\operatorname{diag}(\widehat{\Sigma})$ is almost surely invertible whenever the true covariance matrix is invertible and $n > 1$. Thus, the linear shrinkage estimator is almost surely well-defined for all $\alpha > 0$. Moreover, it can be efficiently computed in $\mathcal{O}(p^3)$ arithmetic operations.

Definition 1.27 (ℓ_1-Regularized maximum likelihood estimator)**.** *The* ℓ_1*-regularized maximum likelihood estimator with penalty parameter* $\beta \geq 0$ *is defined as*

$$X^\star = \arg\min_{X \succeq 0} \left\{ -\log\det X + \langle \widehat{\Sigma}, X \rangle + \beta \sum_{i,j=1}^{p} |X_{ij}| \right\}.$$

Adding an ℓ_1-regularization term to the standard maximum likelihood estimation problem gives rise to sparse—and thus interpretable—estimators [6, 65]. The resulting semidefinite program can be solved with general-purpose interior point solvers such as SDPT3 or with structure-exploiting methods such as the QUIC algorithm, which enjoys a quadratic convergence rate and requires $\mathcal{O}(p^3)$ arithmetic operations per iteration [84]. In the remainder of this section we test the Wasserstein shrinkage, linear shrinkage and ℓ_1-regularized maximum likelihood estimators on synthetic and real datasets. All experiments are implemented in MATLAB, and the corresponding codes are included in the **W**asserstein **I**nverse Covariance **S**hrinkage **E**stimator (WISE) package available at https://www.github.com/nvietanh/wise.

Remark 1.28 (Bessel's correction)**.** *So far we used* $\mathcal{N}(\widehat{\mu}, \widehat{\Sigma})$ *as the nominal distribution, where the sample covariance matrix* $\widehat{\Sigma}$ *was identified with the (biased) maximum likelihood estimator. In practice, it is sometimes useful to use* $\widehat{\Sigma}/\kappa$ *as the nominal covariance matrix, where* $\kappa \in (0,1)$ *is a Bessel correction that removes the bias; see, e.g., Sections 1.6.2 and 1.6.2 below. Under the premise that* $\mathcal{X}$ *is a cone, it is easy to see that if* $(X^\star, \gamma^\star)$ *is optimal in* (1.15) *for a prescribed Wasserstein radius* ρ *and a scaled sample covariance matrix* $\widehat{\Sigma}/\kappa$*, then* $(\kappa X^\star, \kappa\gamma^\star)$ *is optimal* *in* (1.15) *for a scaled Wasserstein radius* $\sqrt{\kappa}\rho$ *and the original sample covariance matrix* $\widehat{\Sigma}$*. Thus, up to scaling, using a Bessel correction is tantamount to shrinking* ρ*.*

1.6.1 Experiments with Synthetic Data

Consider a $(p = 20)$-variate Gaussian random vector ξ with zero mean. The (unknown) true covariance matrix Σ_0 of ξ is constructed as follows. We first choose a density parameter $d \in \{12.5\%, 50\%, 100\%\}$. Using the legacy MATLAB 5.0 uniform generator initialized with seed 0, we then generate a matrix $C \in \mathbb{R}^{p\times p}$ with $\lfloor d \times p^2 \rfloor$ randomly selected nonzero elements, all of which represent independent Bernoulli random variables taking the values $+1$ or -1 with equal probabilities. Finally, we set $\Sigma_0 = (C^\top C + 10^{-3} I)^{-1} \succ 0$.

As usual, the quality of an estimator $X^\star$ for the precision matrix Σ_0^{-1} is evaluated using Stein's loss function

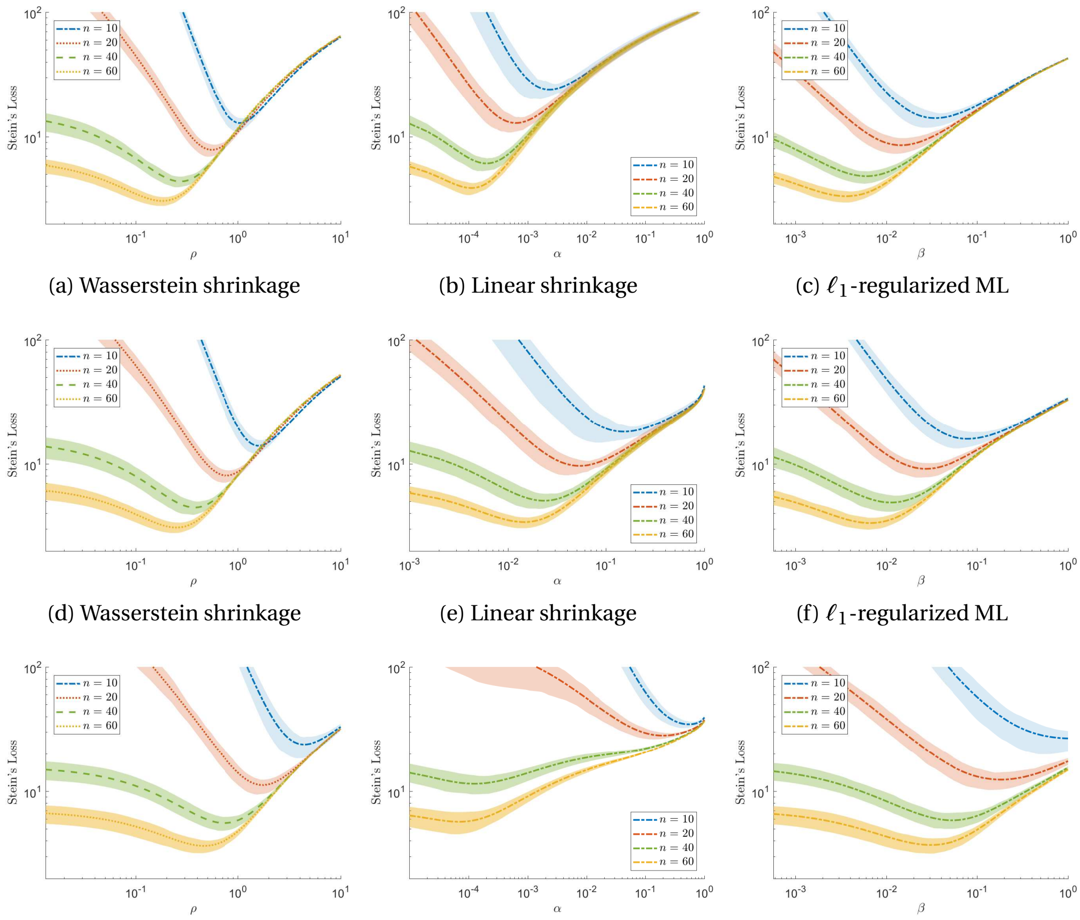
$$L(X^\star, \Sigma_0) = -\log\det(X^\star \Sigma_0) + \langle X^\star, \Sigma_0 \rangle - p,$$

which vanishes if $X^\star = \Sigma_0^{-1}$ and is strictly positive otherwise [90].

All simulation experiments involve 100 independent trials. In each trial, we first draw $n \in \{10, 20, 40, 60\}$ independent samples from $\mathcal{N}(0, \Sigma_0)$, which are used to compute the sample covariance matrix $\widehat{\Sigma}$ and the corresponding precision matrix estimators. Figure 1.2 shows Stein's loss of the Wasserstein shrinkage estimator without structure information for $\rho \in [10^{-2}, 10^1]$, the linear shrinkage estimator for $\alpha \in [10^{-5}, 10^0]$ and the ℓ_1-regularized maximum likelihood estimator for $\beta \in [5 \times 10^{-5}, 10^0]$. Lines represent averages, while shaded areas capture the tubes between the empirical 20% and 80% quantiles across all 100 trials. Note that all three estimators approach $\widehat{\Sigma}^{-1}$ when their respective tuning parameters tend to zero. As $\widehat{\Sigma}$ is rank deficient for $n < p = 20$, Stein's loss thus diverges for small tuning parameters when $n = 10$.

The best Wasserstein shrinkage estimator in a given trial is defined as the one that minimizes Stein's loss over all $\rho \geq 0$. The best linear shrinkage and ℓ_1-regularized maximum likelihood estimators are defined analogously. Figure 1.2 reveals that the best Wasserstein shrinkage estimators dominate the best linear shrinkage and—to a lesser extent—the best ℓ_1-regularized maximum likelihood estimators in terms of Stein's loss for all considered parameter settings. The dominance is more pronounced for small sample sizes. We emphasize that Stein's loss depends explicitly on the unknown true covariance matrix Σ_0. Thus, Figure 1.2 is not available in practice, and the optimal tuning parameters $\rho^\star$, $\alpha^\star$ and $\beta^\star$ cannot be computed exactly. The performance of different precision matrix estimators with *estimated* tuning parameters will be studied in Section 1.6.2.

For $d = 12.5\%$ and $d = 50\%$, the true precision matrix Σ_0^{-1} has many zeros, and prior knowledge of their positions could be used to improve estimator accuracy. To investigate this effect, we henceforth assume that the feasible set $\mathcal{X}$ correctly reflects a randomly selected portion of 50%, 75% or 100% of all zeros of Σ_0^{-1}, while $\mathcal{X}$ contains no (neither correct nor incorrect)



(a) Wasserstein shrinkage (b) Linear shrinkage (c) ℓ_1-regularized ML

(d) Wasserstein shrinkage (e) Linear shrinkage (f) ℓ_1-regularized ML

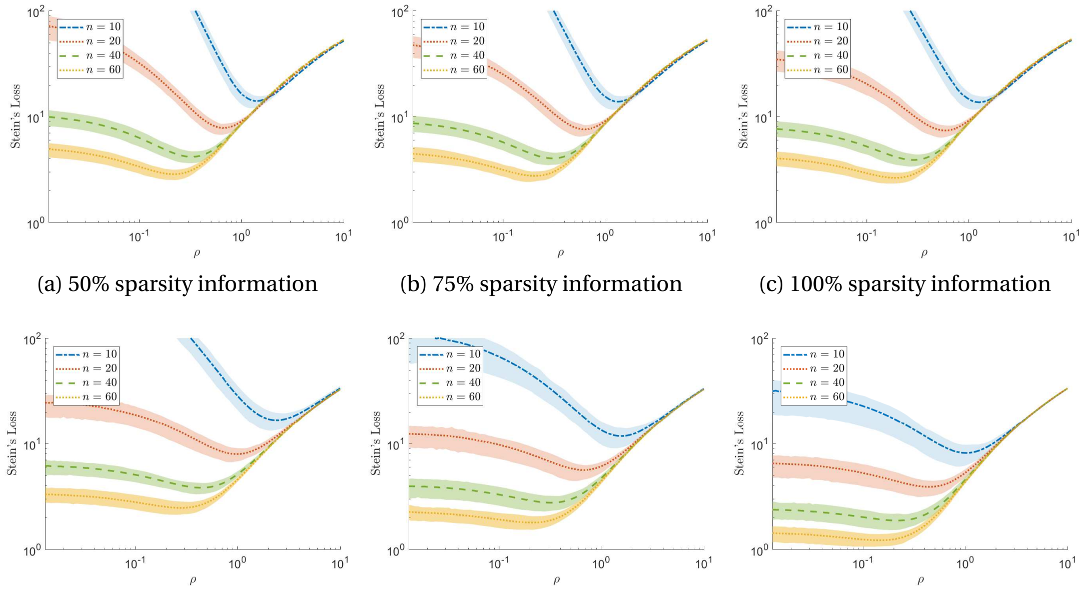
(g) Wasserstein shrinkage (h) Linear shrinkage (i) ℓ_1-regularized ML

Figure 1.2 – Stein's loss of the Wasserstein shrinkage, linear shrinkage and ℓ_1-regularized maximum likelihood estimators as a function of their respective tuning parameters for d = 100% (panels 1.2a–1.2c), d = 50% (panels 1.2d–1.2f) and d = 12.5% (panels 1.2g–1.2i).

information about the remaining zeros. In this setting, we construct the Wasserstein shrinkage estimator by solving problem (1.5) numerically.

Figure 1.3 shows Stein's loss of the Wasserstein shrinkage estimator with prior information for $\rho \in [10^{-2}, 10^1]$. Lines represent averages, while shaded areas capture the tubes between the empirical 20% and 80% quantiles across 100 trials. As expected, correct prior sparsity information improves estimator quality, and the more zeros are known, the better. Note that Σ_0^{-1} contains 21.5% zeros for d = 12.5% and 68% zeros for d = 50%.

In the last experiment, we investigate the Wasserstein radius $\rho^\star$ of the best Wasserstein shrinkage estimator without sparsity information. Figure 1.4 visualizes the average of $\rho^\star$ across 100 independent trials as a function of the sample size n. A standard regression analysis based on the data of Figure 1.4 reveals that $\rho^\star$ converges to zero approximately as $n^{-\kappa}$ with



(a) 50% sparsity information

(b) 75% sparsity information

(c) 100% sparsity information

(d) 50% sparsity information

(e) 75% sparsity information

(f) 100% sparsity information

Figure 1.3 – Stein's loss of the Wasserstein shrinkage estimator with 50%, 75% or 100% sparsity information as a function of the Wasserstein radius ρ for $d = 50\%$ (panels 1.3a–1.3c) and $d = 12.5\%$ (panels 1.3d–1.3f).

$\kappa \approx 61\%$ for $d = 12.5\%$, $\kappa \approx 66\%$ for $d = 50\%$ and $\kappa \approx 68\%$ for $d = 100\%$.

1.6.2 Experiments with Real Data

We now study the properties of the Wasserstein shrinkage estimator in the context of linear discriminant analysis, portfolio selection and the inference of solar irradiation patterns.

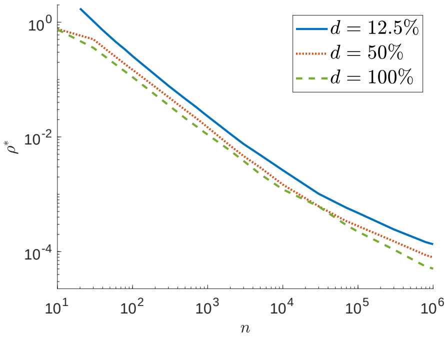


Figure 1.4 – Dependence of the the best Wasserstein radius $\rho^\star$ on the sample size n.

Linear Discriminant Analysis

Linear discriminant analysis aims to predict the class $y \in \mathscr{Y}$, $|\mathscr{Y}| < \infty$, of a feature vector $z \in \mathbb{R}^p$ under the assumption that the conditional distribution of z given y is normal with a class-dependent mean $\mu_y \in \mathbb{R}^p$ and class-*in*dependent covariance matrix $\Sigma_0 \in \mathbb{S}^p_{++}$ [82]. If all μ_y and Σ_0 are known, the maximum likelihood classifier $\mathscr{C} : \mathbb{R}^p \to \mathscr{Y}$ assigns z to a class that maximizes the likelihood of observing y, that is,

$$\mathscr{C}(z) \in \arg\min_{y \in \mathscr{Y}} (z - \mu_y)^\top \Sigma_0^{-1} (z - \mu_y). \tag{1.22}$$

In practice, however, the conditional moments are typically unknown and must be inferred from finitely many training samples $(\widehat{z}_i, \widehat{y}_i)$, $i \le n$. If we estimate μ_y by the sample average

$$\widehat{\mu}_y = \frac{1}{|\mathscr{I}_y|} \sum_{i \in \mathscr{I}_y} \widehat{x}_i\,,$$

where $\mathscr{I}_y = \{i \in \{1, \ldots, n\} : \widehat{y}_i = y\}$ records all samples in class y, then it is natural to define the residual feature vectors as $\widehat{\xi}_i = \widehat{z}_i - \widehat{\mu}_{\widehat{y}_i}$, $i \le n$. Accounting for Bessel's correction, the conditional distribution of $\widehat{\xi}_i$ given $\widehat{y}_i$ is normal with mean 0 and covariance matrix $(|\mathscr{I}_{\widehat{y}_i}| - 1)\,|\mathscr{I}_{\widehat{y}_i}|^{-1}\Sigma_0$. The marginal distribution of $\widehat{\xi}_i$ thus constitutes a mixture of $|\mathscr{Y}|$ normal distributions with mean 0, all of which share the same covariance matrix up to a scaling factor close to unity. As such, the residuals fail to be normally distributed. Moreover, due to their dependence on the sample means, the residuals are correlated. However, if each class accommodates many training samples, then the residuals can approximately be regarded as independent samples from $\mathscr{N}(0, \Sigma_0)$.

Irrespective of these complications, the sample covariance matrix

$$\widehat{\Sigma} = \frac{1}{n - |\mathscr{Y}|} \sum_{i=1}^{n} \widehat{\xi}_i \widehat{\xi}_i^\top$$

provides an unbiased estimator for Σ_0. Indeed, by the law of total expectation we have

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}[\widehat{\Sigma}] &= \frac{1}{n - |\mathscr{Y}|} \mathbb{E}_{\mathbb{P}} \left[\sum_{i=1}^{n} \mathbb{E}_{\mathbb{P}} \left[\widehat{\xi}_i \widehat{\xi}_i^\top \,\middle|\, \widehat{y}_i \right] \right] \\
&= \frac{1}{n - |\mathscr{Y}|} \sum_{y \in \mathscr{Y}} \mathbb{E}_{\mathbb{P}} \left[\sum_{i \in \mathscr{I}_y} \frac{|\mathscr{I}_{\widehat{y}_i}| - 1}{|\mathscr{I}_{\widehat{y}_i}|} \Sigma \right] = \frac{1}{n - |\mathscr{Y}|} \sum_{y \in \mathscr{Y}} (|\mathscr{I}_y| - 1)\Sigma_0 \; = \; \Sigma_0\,,
\end{aligned}$$

where $\mathbb{P}$ stands for the unknown true joint distribution of the residuals and class labels. In a data-driven setting, the ideal maximum likelihood classifier (1.22) is replaced with

$$\widehat{\mathscr{C}}(\xi) = \arg\min_{y \in \mathscr{Y}} (\xi - \widehat{\mu}_y)^\top X^\star (\xi - \widehat{\mu}_y)\,, \tag{1.23}$$

which depends on the raw data through the sample averages $\widehat{\mu}_y$, $y \in \mathscr{Y}$, and some precision matrix estimator $X^\star$. The possible choices for $X^\star$ include the Wasserstein shrinkage estimator without prior information, the linear shrinkage estimator and the ℓ_1-regularized maximum likelihood estimator, all of which depend on the data merely through $\widehat{\Sigma}$. Note that the naïve precision matrix estimator $\widehat{\Sigma}^{-1}$ exists only for $n > p$ and is therefore disregarded. All estimators depend on a scalar parameter (the Wasserstein radius ρ, the mixing parameter α or the penalty parameter β) that can be used to tune the performance of the classifier (1.23).

We test the classifier (1.23) equipped with different estimators $X^\star$ on two preprocessed datasets from [42]:

1. The "*colon cancer*" dataset contains 62 gene expression profiles, each of which involves 2,000 features and is classified either as normal tissue (NT) or tumor-affected tissue (TT). The data is split into a training dataset of 29 observations (9 in class NT and 20 in class TT) and a test dataset of 33 observations (13 in class NT and 20 in class TT).

2. The "*leukemia*" dataset contains 72 gene expression profiles, each of which involves 3,571 features and is classified either as acute lymphocytic leukemia (ALL) or acute myeloid leukemia (AML). The data is split into a training dataset of 38 observations (27 in class ALL and 11 in class AML) and a test dataset of 34 observations (20 in class ALL and 14 in class AML).

Classification is based solely on the first $p \in \{20, 40, 80, 100\}$ features of each gene expression profile. We use leave-one-out cross validation on the training data to tune the precision matrix estimator $X^\star$ with the goal to maximize the correct classification rate of the classifier (1.23). To keep the computational overhead manageable, we optimize the tuning parameters over the finite search grids

$$\rho \in \{10^{\frac{j}{20}-1} : j = 0, \ldots, 60\}, \quad \alpha \in \{10^{\frac{j}{20}-3} : j = 0, \ldots, 60\} \quad \text{and} \quad \beta \in \{10^{\frac{j}{20}-3} : j = 0, \ldots, 60\}.$$

We highlight that, in case of the ℓ_1-regularized maximum likelihood estimator, cross validation becomes computationally prohibitive for $p > 80$ even if the state-of-the-art QUIC routine is used [84] to solve the underlying semidefinite programs. In contrast, the Wasserstein and linear shrinkage estimators can be computed and tuned quickly even for $p \gg 100$. Once the optimal tuning parameters are found, we fix them and recalculate $X^\star$ on the basis of the entire training dataset. Finally, we substitute the resulting precision matrix estimator into the classifier (1.23) and evaluate its correct classification rate on the test dataset. The test results are reported in Table 1.1. We observe that the Wasserstein shrinkage estimator frequently outperforms the linear shrinkage and ℓ_1-regularized maximum likelihood estimators, especially for higher values of p.

Table 1.1 – Correct classification rate of the classifier (1.23) instantiated with different precision matrix estimators. The best result in each experiment is highlighted in bold.

	Colon cancer dataset				Leukemia dataset			
Estimator	$p = 20$	$p = 40$	$p = 80$	$p = 100$	$p = 20$	$p = 40$	$p = 80$	$p = 100$
Wasserstein	**72.73**	75.76	**78.79**	**75.76**	**73.53**	67.65	**91.18**	**91.18**
Linear	57.58	72.73	72.73	72.73	70.59	**70.59**	82.35	82.35
ℓ_1-regularized	**72.73**	**78.79**	**78.79**	72.73	70.59	64.71	82.35	82.35

Minimum Variance Portfolio Selection

Consider the minimum variance portfolio selection problem without short sale constraints [87]

$$\begin{array}{cl} \min\limits_{w \in \mathbb{R}^p} & w^\top \Sigma_0 w \\ \text{s.t.} & \mathbb{1}^\top w = 1\,, \end{array}$$

where the portfolio vector $w \in \mathbb{R}^p$ captures the percentage weights of initial capital allocated to p different assets with random returns, $\mathbb{1} \in \mathbb{R}^p$ stands for the vector of ones, and $\Sigma_0 \in \mathbb{S}_{++}^p$ denotes the covariance matrix of the asset returns. The objective represents the variance of the portfolio return, which is strictly convex in w thanks to the positive definiteness of Σ_0. The unique optimal solution of this portfolio selection problem is given by $w^\star = \Sigma_0^{-1}\mathbb{1}/\mathbb{1}^\top\Sigma_0^{-1}\mathbb{1}$. In practice, the unknown true precision matrix Σ_0^{-1} must be replaced with an estimator $X^\star$, which gives rise to the estimated minimum variance portfolio $\widehat{w}^\star = X^\star\mathbb{1}/\mathbb{1}^\top X^\star\mathbb{1}$.

A vast body of literature in finance focuses on finding accurate precision matrix estimators for portfolio construction, see, *e.g.*, [161, 40, 104, 75, 165]. In the following we compare the minimum variance portfolios based on the Wasserstein shrinkage estimator without structural information, the linear shrinkage estimator and ℓ_1-regularized maximum likelihood estimator on two preprocessed datasets from the Fama-French online data library:[2] the "*48 industry portfolios*" dataset (FF48) and the "*100 portfolios formed on size and book-to-market*" dataset (FF100). Recall that the estimators depend on the data only through the sample covariance matrix $\widehat{\Sigma}$, which is computed from the residual returns relative to the sample means and thus needs to account for Bessel's correction. The datasets both consist of monthly returns for the period from January 1996 to December 2016. The first 120 observations from January 1996 to December 2005 serve as the training dataset. The optimal tuning parameters that minimize the portfolio variance are estimated via leave-one-out cross validation on the training dataset using the finite search grids

$$\rho \in \{10^{\frac{j}{100}-2} : j = 0, \ldots, 200\}, \quad \alpha \in \{10^{\frac{j}{100}-2} : j = 0, \ldots, 200\} \quad \text{and} \quad \beta \in \{10^{\frac{j}{50}-4} : j = 0, \ldots, 200\}.$$

[2] See http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html (accessed January 2018)

The out-of-sample performance of the minimum variance portfolio corresponding to a particular precision matrix estimator is then evaluated using the rolling horizon method over the period from January 2006 to December 2016, where the sample covariance matrix needed as an input for the precision matrix is re-estimated every three months based on the most recent 120 observations (10 years), while the tuning parameter is kept fixed. The resulting out-of-sample mean, standard deviation and Sharpe ratio of the portfolio return are reported in Table 1.2. While the ℓ_1-regularized maximum likelihood estimator yields the portfolio with the lowest standard deviation for both datasets, the Wasserstein shrinkage estimator always generates the highest mean and, maybe surprisingly, the highest Sharpe ratio.

Table 1.2 – Standard deviation, mean and Sharpe ratio of the minimum variance portfolio based on different estimators. The best result in each experiment is highlighted in bold.

	FF48 dataset			FF100 dataset		
Estimator	std	mean	Sharpe	std	mean	Sharpe
Wasserstein shrinkage	3.146	**0.701**	**0.223**	3.518	**1.079**	**0.307**
Linear shrinkage	3.152	0.688	0.218	3.484	0.965	0.277
ℓ_1-regularized ML	**3.077**	0.668	0.217	**3.423**	1.010	0.295

Inference of Solar Irradiation Patterns

In the last experiment we aim to estimate the spatial distribution of solar irradiation in Switzerland using the "*surface incoming shortwave radiation*" (SIS) data provided by MeteoSwiss.[3] The SIS data captures the horizontal solar irradiation intensities in W/m^2 for pixels of size 1.6km by 2.3km based on the effective cloud albedo, which is derived from satellite imagery. The dataset spans 13 years from 2004 to 2016, with a total number of 4,749 daily observations. We deseasonalize the time series of each pixel as follows. First, we divide the original time series by a shifted sinusoid with a yearly period, whose baseline level, phase and amplitude are estimated via ordinary least squares regression. Next, we subtract unity. The resulting deseasonalized time series is viewed as the sample path of a zero mean Gaussian noise process. This approach relies on the assumption that the mean and the standard deviation of the original time series share the same seasonality pattern. It remains to estimate the joint distribution of the pixel-wise Gaussian white noise processes, which is fully determined by the precision matrix of the deseasonalized data. We estimate the precision matrix using the Wasserstein shrinkage, linear shrinkage and ℓ_1-regularized maximum likelihood estimators. As each pixel represents a geographical location and as the solar irradiation intensities at two distant pixels are likely to be conditionally independent given the intensities at all other pixels, it is reasonable to assume that the precision matrix is sparse; see also [36, 171]. Specifically, we assume here that the solar irradiation intensities at two pixels indexed by (i, j) and (i', j') are

[3] See http://www.meteoschweiz.admin.ch/data/products/2014/raeumliche-daten-globalstrahlung.html (accessed January 2018)

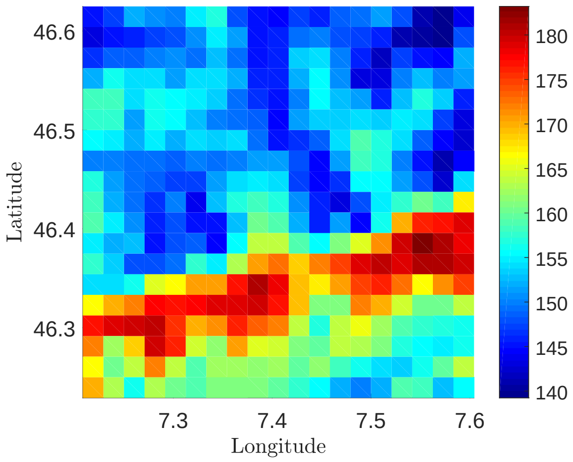


Figure 1.5 – Average solar irradiation intensities (W/m^2) for the Diablerets region in Switzerland.

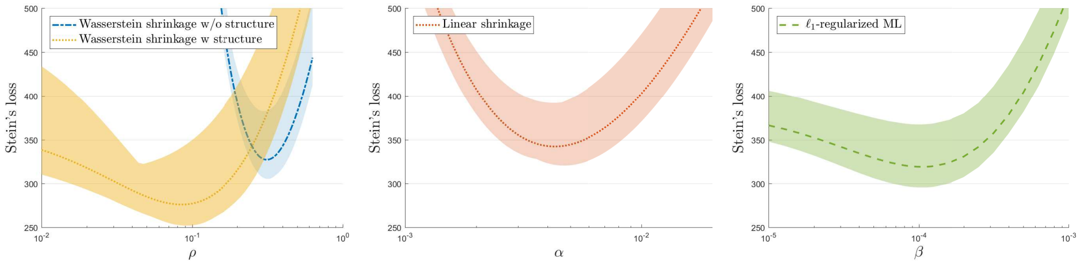
conditionally independent and that the corresponding entry of the precision matrix vanishes whenever $|i-i'|+|j-j'|>3$. This sparsity information can be used to enhance the basic Wasserstein shrinkage estimator.

Consider now the Diablerets region of Switzerland, which is described by a spatial matrix of 20×20 pixels. Thus, the corresponding precision matrix has dimension 400×400. The average daily solar irradiation intensities within the region of interest are visualized in Figure 1.5. We note that the sunshine exposure is highly variable due to the heterogeneous geographical terrain characterized by a high mountain range in the south intertwined with deep valleys in the north. In order to assess the quality of a specific precision matrix estimator, we use K-fold cross validation with $K=13$. The k-th fold comprises all observations of year k and is used to construct the estimator $X_k^\star$. The data of the remaining 12 years, without year k, are used to compute the empirical covariance matrix $\widehat{\Sigma}_{-k}$. The estimation error of $X_k^\star$ is then measured via Stein's loss

$$L(X_k^\star,\widehat{\Sigma}_{-k}) = -\log\det(X_k^\star\,\widehat{\Sigma}_{-k}) + \langle X_k^\star,\widehat{\Sigma}_{-k}\rangle - p.$$

We emphasize that here, in contrast to the experiment with synthetic data, $\widehat{\Sigma}_{-k}$ is used as a proxy for the unknown true covariance matrix Σ. Figure 1.6 shows Stein's loss of the Wasserstein shrinkage estimator with and without structure information for $\rho\in[10^{-2},10^{0}]$, the linear shrinkage estimator for $\alpha\in[10^{-3},2\times10^{-2}]$ and the ℓ_1-regularized maximum likelihood estimator for $\beta\in[10^{-5},10^{-3}]$. Lines represent averages, while shaded areas capture the tubes between the best- and worst-case loss realizations across all K folds.

The Wasserstein shrinkage estimator with structure information reduces the minimum average loss by 13.5% relative to the state-of-the-art ℓ_1-regularized maximum likelihood estimator. Moreover, the average runtimes for computing the different estimators amount to 51.84s for the Wasserstein shrinkage estimator with structural information (Algorithm 1), 0.08s for the Wasserstein shrinkage estimator without structural information (analytical formula and



(a) Wasserstein shrinkage

(b) Linear shrinkage

(c) ℓ_1-regularized ML

Figure 1.6 – Stein's loss of the Wasserstein shrinkage, linear shrinkage and ℓ_1-regularized maximum likelihood estimators as a function of their respective tuning parameters.

bisection algorithm), 0.01s for the linear shrinkage estimator (analytical formula) and 1493.61s for the ℓ_1-regularized maximum likelihood estimator (QUIC algorithm [84]).

2 Bridging Bayesian and Minimax Mean Square Error Estimation via Wasserstein Distributionally Robust Optimization

> This duality can be pursued further and is related to a duality between past and future and the notions of control and knowledge. Thus we may have knowledge of the past but cannot control it; we may control the future but have no knowledge of it.
>
> — Claude Shannon

We introduce a distributionally robust minimium mean square error estimation model with a Wasserstein ambiguity set to infer an unknown signal from a noisy measurement. The proposed model minimizes the worst-case (maximum) of the mean square expected loss across all reference distributions within a prescribed Wasserstein distance from a nominal distribution. We show that the proposed model can be conservatively approximated by the optimal value of a finite convex optimization problem. If the nominal distribution is elliptical, we prove that the optimal estimator is affine and can be recovered from the optimal solution of the related dual estimation problem whose reformulation is equivalent to a tractable semidefinite program. Finally, we develop a Frank-Wolfe algorithm for efficiently solving the robustified estimation problem in high dimensional settings, such as for image denoising using wavelet shrinkage.

2.1 Introduction

Consider the problem of estimating an unknown parameter $x \in \mathbb{R}^n$ based on a linear measurement $y \in \mathbb{R}^m$ corrupted by additive noise $w \in \mathbb{R}^m$. This setup is formalized through the linear measurement model

$$y = Hx + w, \tag{2.1}$$

where the observation matrix $H \in \mathbb{R}^{m\times n}$ is assumed to be known. We further assume that the distribution $\mathbb{P}_w$ of w has finite second moments and is independent of x. Thus, the conditional distribution $\mathbb{P}_{y|x}$ of y given x is obtained by shifting $\mathbb{P}_w$ by Hx. Note that the linear measurement model is fundamental for numerous applications in engineering (*e.g.*, linear systems theory [74, 127]), econometrics (*e.g.*, linear regression [162, 174], time series analysis [25, 80]), machine learning and signal processing (*e.g.*, Kalman filtering [97, 120, 129]) or information theory (*e.g.*, multiple-input multiple-output systems [32, 113]) etc.

An estimator of x given y is a measurable function $\psi : \mathbb{R}^m \to \mathbb{R}^n$ that grows at most linearly. Thus, there exists $C > 0$ such that $|\psi(y)| \leq C(1 + \|y\|)$ for all $y \in \mathbb{R}^m$. The function value $\psi(y)$ is the prediction of x based on the measurement y under the estimator ψ. In the following we denote the family of all estimators by $\mathscr{F}$. The quality of an estimator is measured by a risk function $\mathscr{R} : \mathscr{F} \times \mathbb{R}^n \to \mathbb{R}$, which quantifies the mismatch between the parameter x and its prediction $\psi(y)$. A popular risk function is the mean square error (MSE)

$$R(\psi, x) \triangleq \mathbb{E}_{\mathbb{P}_{y|x}}\left[\|x - \psi(y)\|^2\right],$$

which defines the estimation error as the expected squared Euclidean distance between $\psi(y)$ and x. If x was known, then $R(\psi, x)$ could be minimized directly, and the constant estimator $\psi^\star(y) \equiv x$ would be optimal. In practice, however, x is unobservable. Otherwise there would be no need to solve an estimation problem in the first place. With x unknown, it is impossible to minimize the MSE directly. The statistics literature proposes two complementary workarounds for this problem: the Bayesian approach and the minimax approach.

The Bayesian statistician treats x as a random vector governed by a *prior* distribution $\mathbb{P}_x$ that captures her beliefs about x before seeing y [114, § 1.2.4] and solves the minimum MSE (MMSE) estimation problem

$$\underset{\psi\in\mathscr{F}}{\text{minimize}}\ \mathbb{E}_{\mathbb{P}_x}\left[R(\psi, x)\right]. \tag{2.2}$$

If the distribution $\mathbb{P}_x$ of x has finite second moments, then (2.2) is solvable. In this case, the optimal estimator, which is usually termed the Bayesian MMSE estimator, is of the form $\psi^\star_{\mathscr{B}}(y) = \mathbb{E}_{\mathbb{P}_{x|y}}[x]$, where the conditional distribution $\mathbb{P}_{x|y}$ of x given y is obtained from $\mathbb{P}_x$ and $\mathbb{P}_{y|x}$ via Bayes' theorem. However, the Bayesian MMSE estimator suffers from two conceptual shortcomings. First, $\psi^\star_{\mathscr{B}}$ is highly sensitive to the prior distribution $\mathbb{P}_x$, which is troubling if the statistician has little confidence in her beliefs. Second, computing $\psi^\star_{\mathscr{B}}$ requires precise knowledge of the noise distribution $\mathbb{P}_w$, which is typically unobservable and thus uncertain at least to some extent. Moreover, $\psi^\star_{\mathscr{B}}$ may generically have a complicated functional form, and evaluating $\psi^\star_{\mathscr{B}}(y)$ to high precision for a particular measurement y (*e.g.*, via Monte Carlo simulation) may be computationally challenging if the dimension of x is high.

These shortcomings are mitigated if we restrict the space $\mathscr{F}$ of all measurable estimators in (2.2) to the space

$$\mathscr{A} \triangleq \left\{\psi \in \mathscr{F} \,:\, \exists A \in \mathbb{R}^{n\times m},\, b \in \mathbb{R}^n \text{ with } \psi(y) = Ay + b \;\forall y \in \mathbb{R}^m\right\} \tag{2.3}$$

of all *affine* estimators. In this case the distributions $\mathbb{P}_x$ and $\mathbb{P}_w$ need not be fully known. Instead, in order to evaluate the optimal affine estimator $\psi^\star_{\mathscr{A}}(y) = A^\star y + b^\star$, it is sufficient to know the mean vectors μ_x and μ_w as well as the covariance matrices Σ_x and Σ_w of the distributions $\mathbb{P}_x$ and $\mathbb{P}_w$, respectively. If $H\Sigma_x H^\top + \Sigma_w \succ 0$, which is the case if the noise covariance matrix has full rank, then the coefficients of the best affine estimator can be computed in closed form. Using (2.1) together with the independence of x and w one can show that

$$A^\star = \Sigma_x H^\top (H\Sigma_x H^\top + \Sigma_w)^{-1} \quad \text{and} \quad b^\star = \mu_x - A^\star(H\mu_x + \mu_w). \tag{2.4}$$

If the random vector (x, y) follows a normal distribution, then the best affine estimator is also optimal among all measurable estimators. In general, however, we do not know how much optimality is sacrificed by restricting attention to affine estimators. Moreover, the uncertainty about $\mathbb{P}_x$ and $\mathbb{P}_w$ transpires through to their first- and second-order moments. As the coefficients (2.4) tend to be highly sensitive to these moments, their uncertainty remains worrying.

The minimax approach models the statistician's prior knowledge concerning x via a convex closed uncertainty set $\mathscr{X} \subseteq \mathbb{R}^n$ as commonly used in robust optimization. The minimax MSE estimation problem is then formulated as a zero-sum game between the statistician, who selects the estimator $\psi \in \mathscr{F}$ with the goal to minimize the MSE, and nature, who chooses the parameter value $x \in \mathscr{X}$ with the goal to maximize the MSE.

$$\underset{\psi\in\mathscr{F}}{\text{minimize}} \; \max_{x\in\mathscr{X}} \; R(\psi, x) \tag{2.5a}$$

By construction, any minimizer $\psi^\star_{\mathscr{M}}$ of (2.5a) incurs the smallest possible estimation error under the worst parameter realization within the uncertainty set $\mathscr{X}$. For this reason $\psi^\star_{\mathscr{M}}$ is called a minimax estimator. Note that the MSE $R(\psi, x)$ generically displays a complicated non-concave dependence on x for any fixed ψ, which implies that nature's inner maximization problem in (2.5a) is usually non-convex. Thus, we should not expect the zero-sum game (2.5a) between the statistician and nature to admit a Nash equilibrium. However, the inner maximization problem can be convexified by allowing nature to play mixed (randomized) strategies, that is, by reformulating (2.5a) as the (equivalent) convex-concave saddle point problem

$$\underset{\psi \in \mathcal{F}}{\text{minimize}} \max_{\mathbb{Q}_x \in \mathcal{M}(\mathcal{X})} \mathbb{E}_{\mathbb{Q}_x}\left[R(\psi, x)\right], \tag{2.5b}$$

where $\mathcal{M}(\mathcal{X})$ stands for the family of all distributions supported on $\mathcal{X}$ with finite second-order moments. As $\mathbb{E}_{\mathbb{Q}_x}[R(\psi, x)]$ is convex in ψ for any fixed $\mathbb{Q}_x$ and concave (linear) in $\mathbb{Q}_x$ for any fixed ψ, while $\mathcal{F}$ and $\mathcal{M}(\mathcal{X})$ are both convex sets, the zero-sum game (2.5b) admits a Nash equilibrium $(\psi^\star_{\mathcal{M}}, \mathbb{Q}^\star_x)$ under mild technical conditions. Note that $\psi^\star_{\mathcal{M}}$ is again a minimax estimator. Moreover, $\psi^\star_{\mathcal{M}}$ is the statistician's best response to nature's choice $\mathbb{Q}^\star_x$ and vice versa. Using the terminology introduced above, this means that $\psi^\star_{\mathcal{M}}$ is the Bayesian MMSE estimator corresponding to the prior $\mathbb{Q}^\star_x$. For this reason, $\mathbb{Q}^\star_x$ is usually referred to as the *least favorable prior*. Even though the minimax approach exonerates the statistician from narrowing down her beliefs to a single prior distribution $\mathbb{Q}_x$, it still requires precise information about $\mathbb{P}_w$, which may not be available in practice. On the other hand, as it robustifies the estimator against *any* distribution on $\mathcal{X}$, the minimax approach is often regarded as overly pessimistic. Moreover, as in the case of the Bayesian MMSE estimation problem, $\psi^\star_{\mathcal{M}}$ may generically have a complicated functional form, and evaluating $\psi^\star_{\mathcal{M}}(y)$ to high precision may be computationally challenging if the dimension of x is high. A simple remedy to mitigate these computational challenges would be to restrict $\mathcal{F}$ to the family $\mathcal{A}$ of affine estimators. The loss of optimality incurred by this approximation for different choices of $\mathcal{X}$ is discussed in [95, § 4] and the references therein.

In this paper we bridge the Bayesian and the minimax approaches by leveraging tools from distributionally robust optimization. Specifically, we study distributionally robust estimation problems of the form

$$\underset{\psi \in \mathcal{F}}{\text{minimize}} \max_{\mathbb{Q}_x \in \mathcal{Q}_x} \mathbb{E}_{\mathbb{Q}_x}\left[R(\psi, x)\right], \tag{2.6}$$

where $\mathcal{Q}_x \subseteq \mathcal{M}(\mathbb{R}^n)$ is an *ambiguity set* of multiple (possibly infinitely many) plausible prior distributions of x. Note that if the ambiguity set collapses to the singleton $\mathcal{Q}_x = \{\mathbb{P}_x\}$ for some $\mathbb{P}_x \in \mathcal{M}(\mathbb{R}^n)$, then the distributionally robust estimation problem (2.6) reduces to the Bayesian MMSE estimation problem (2.2). Similarly, under the ambiguity set $\mathcal{Q}_x = \mathcal{M}(\mathcal{X})$ for some convex closed uncertainty set $\mathcal{X} \subseteq \mathbb{R}^n$, problem (2.6) reduces to the minimax mean square error estimation problem (2.5b). By providing considerable freedom in tailoring the ambiguity set $\mathcal{Q}_x$, the distributionally robust approach thus allows the statistician to reconcile the specificity of the Bayesian approach with the conservativeness of the minimax approach.

The estimation model (2.6) still relies on the premise that the noise distribution $\mathbb{P}_w$ is precisely known, and this assumption is not tenable in practice. However, nothing prevents us from further robustifying (2.6) against uncertainty in $\mathbb{P}_w$. To this end, we define $\mathcal{M}(\mathbb{R}^{n+m})$ as the family of all joint distributions of x and w with finite second-order moments. Moreover, we define the *average risk* $\mathscr{R}:\mathscr{F}\times\mathscr{M}(\mathbb{R}^{n+m})\to\mathbb{R}$ through

$$\mathscr{R}(\psi,\mathbb{P})\triangleq\mathbb{E}_{\mathbb{P}}[\|x-\psi(Hx+w)\|^2].$$

If $\mathbb{P}=\mathbb{P}_x\times\mathbb{P}_w$ for some marginal distributions $\mathbb{P}_x\in\mathscr{M}(\mathbb{R}^n)$ and $\mathbb{P}_w\in\mathscr{M}(\mathbb{R}^m)$, which implies that x and w are independent under $\mathbb{P}$, and if $\mathbb{P}_{y|x}$ is defined as $\mathbb{P}_w$ shifted by Hx, then $\mathscr{R}(\psi,\mathbb{P})=\mathbb{E}_{\mathbb{P}_x}[R(\psi,x)]$. Thus, the average risk $\mathscr{R}(\psi,\mathbb{P})$ corresponds indeed to the risk $R(\psi,x)$ averaged under the marginal distribution $\mathbb{P}_x$. In the remainder of this paper we will study generalized distributionally robust estimation problems of the form

$$\underset{\psi\in\mathscr{F}}{\text{minimize}}\ \sup_{\mathbb{Q}\in\mathbb{B}(\widehat{\mathbb{P}})}\mathscr{R}(\psi,\mathbb{Q}),\tag{2.7}$$

where the ambiguity set $\mathbb{B}(\widehat{\mathbb{P}})\subseteq\mathscr{M}(\mathbb{R}^{n+m})$ captures distributional uncertainty in both $\mathbb{P}_x$ and $\mathbb{P}_w$. Specifically, we will model $\mathbb{B}(\widehat{\mathbb{P}})$ as a set of factorizable distributions $\mathbb{Q}=\mathbb{Q}_x\times\mathbb{Q}_w$ close to a nominal distribution $\widehat{\mathbb{P}}=\widehat{\mathbb{P}}_x\times\widehat{\mathbb{P}}_w$ in the sense that $\mathbb{Q}_x$ and $\mathbb{Q}_w$ are close to $\widehat{\mathbb{P}}_x$ and $\widehat{\mathbb{P}}_w$ in Wasserstein distance, respectively.

Definition 2.1 (Wasserstein distance)**.** *For any $d\in\mathbb{N}$, the type-2 Wasserstein distance between two distributions $\mathbb{Q}_1,\mathbb{Q}_2\in\mathscr{M}(\mathbb{R}^d)$ is defined as*

$$\mathbb{W}(\mathbb{Q}_1,\mathbb{Q}_2)\triangleq\inf_{\pi\in\Pi(\mathbb{Q}_1,\mathbb{Q}_2)}\left(\int_{\mathbb{R}^d\times\mathbb{R}^d}\left\|\xi_1-\xi_2\right\|^2\pi(\mathrm{d}\xi_1,\mathrm{d}\xi_2)\right)^{\frac{1}{2}},$$

where $\Pi(\mathbb{Q}_1,\mathbb{Q}_2)$ denotes the set of all joint distributions or couplings $\pi\in\mathscr{M}(\mathbb{R}^d\times\mathbb{R}^d)$ of the random vectors $\xi_1\in\mathbb{R}^d$ and $\xi_2\in\mathbb{R}^d$ with marginal distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$, respectively.

The dependence of the Wasserstein distance on d is notationally suppressed to avoid clutter. Note that $\mathbb{W}(\mathbb{Q}_1,\mathbb{Q}_2)^2$ is naturally interpreted as the optimal value of a transportation problem that determines the minimum cost of moving the distribution $\mathbb{Q}_1$ to $\mathbb{Q}_2$, where the cost of moving a unit probability mass from ξ_1 to ξ_2 is given by the squared Euclidean distance $\|\xi_1-\xi_2\|^2$. For this reason, the optimization variable π is sometimes referred to as a transportation plan and the Wasserstein distance as the earth mover's distance.

Formally, we define the *Wasserstein ambiguity set* as

$$\mathbb{B}(\widehat{\mathbb{P}})\triangleq\left\{\mathbb{Q}_x\times\mathbb{Q}_w:\begin{array}{ll}\mathbb{Q}_x\in\mathscr{M}(\mathbb{R}^n), & \mathbb{W}(\mathbb{Q}_x,\widehat{\mathbb{P}}_x)\ \le\rho_x\\ \mathbb{Q}_w\in\mathscr{M}(\mathbb{R}^m), & \mathbb{W}(\mathbb{Q}_w,\widehat{\mathbb{P}}_w)\le\rho_w\end{array}\right\},\tag{2.8}$$

where $\widehat{\mathbb{P}}_x$ and $\widehat{\mathbb{P}}_w$ represent prescribed nominal distributions that could be constructed via statistical analysis or expert judgement, while the Wasserstein radii $\rho_x\ge 0$ and $\rho_w\ge 0$ constitute hyperparameters that quantify the statistician's uncertainty about the nominal distributions of x and w. We emphasize that the distributionally robust estimation model (2.7) generalizes all preceding models. Indeed, if $\rho_w=0$, then (2.7) reduces to the first distributionally robust model (2.6), which in turn encompasses both the MMSE estimation problem (2.2) (for $\rho_x=0$)

and the minimax estimation problem (2.5b) (for $\rho_x = \infty$) as special cases.

The distributionally robust estimation model (2.7) is conceptually attractive because the hyperparameters ρ_x and ρ_w allow the statistician to specify her level of trust in the nominal prior distribution $\widehat{\mathbb{P}}_x$ and the nominal noise distribution $\widehat{\mathbb{P}}_w$. In the remainder of the paper we will show that (2.7) is also computationally attractive. This is maybe surprising because mixtures of factorizable distributions are generally not factorizable, which implies that the Wasserstein ambiguity set $\mathbb{B}(\widehat{\mathbb{P}})$ is non-convex.

We remark that one could also work with an alternative ambiguity set of the form

$$\mathbb{B}'(\widehat{\mathbb{P}}) \triangleq \left\{\mathbb{Q}_x \times \mathbb{Q}_w : \mathbb{Q}_x \in \mathcal{M}(\mathbb{R}^n),\ \mathbb{Q}_w \in \mathcal{M}(\mathbb{R}^m),\ \mathbb{W}(\mathbb{Q}_x \times \mathbb{Q}_w, \widehat{\mathbb{P}}_x \times \widehat{\mathbb{P}}_w) \le \rho\right\}, \tag{2.9}$$

which involves only a single hyperparameter $\rho \ge 0$ and is therefore less expressive but maybe easier to calibrate than $\mathbb{B}(\widehat{\mathbb{P}})$. The following lemma is instrumental to understanding the relation between $\mathbb{B}(\widehat{\mathbb{P}})$ and $\mathbb{B}'(\widehat{\mathbb{P}})$.

Lemma 2.2 (Pythagoras' theorem for Wasserstein distances). *For any $\mathbb{Q}_x^1, \mathbb{Q}_x^2 \in \mathcal{M}(\mathbb{R}^n)$ and $\mathbb{Q}_w^1, \mathbb{Q}_w^2 \in \mathcal{M}(\mathbb{R}^m)$ we have*

$$\mathbb{W}(\mathbb{Q}_x^1 \times \mathbb{Q}_w^1, \mathbb{Q}_x^2 \times \mathbb{Q}_w^2)^2 = \mathbb{W}(\mathbb{Q}_x^1, \mathbb{Q}_x^2)^2 + \mathbb{W}(\mathbb{Q}_w^1, \mathbb{Q}_w^2)^2.$$

Proof. By the definition of the Wasserstein distance and by the standard Pythagorean theorem we have

$$\begin{aligned}
&\mathbb{W}(\mathbb{Q}_x^1 \times \mathbb{Q}_w^1, \mathbb{Q}_x^2 \times \mathbb{Q}_w^2)^2 \\
&= \inf_{\pi \in \Pi(\mathbb{Q}_x^1 \times \mathbb{Q}_w^1, \mathbb{Q}_x^2 \times \mathbb{Q}_w^2)} \int_{\mathbb{R}^{n+m} \times \mathbb{R}^{n+m}} \left\|x_1 - x_2\right\|^2 + \left\|w_1 - w_2\right\|^2 \pi(\mathrm{d}x_1, \mathrm{d}w_1, \mathrm{d}x_2, \mathrm{d}w_2) \\
&\le \inf_{\pi_x \in \Pi(\mathbb{Q}_x^1, \mathbb{Q}_x^2)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \left\|x_1 - x_2\right\|^2 \pi_x(\mathrm{d}x_1, \mathrm{d}x_2) + \inf_{\pi_w \in \Pi(\mathbb{Q}_w^1, \mathbb{Q}_w^2)} \int_{\mathbb{R}^m \times \mathbb{R}^m} \left\|w_1 - w_2\right\|^2 \pi_w(\mathrm{d}w_1, \mathrm{d}w_2) \\
&= \mathbb{W}(\mathbb{Q}_x^1, \mathbb{Q}_x^2)^2 + \mathbb{W}(\mathbb{Q}_w^1, \mathbb{Q}_w^2)^2,
\end{aligned}$$

where the inequality follows from the restriction to factorizable transportation plans of the form $\pi = \pi_x \times \pi_w$ for some $\pi_x \in \Pi(\mathbb{Q}_x^1, \mathbb{Q}_x^2)$ and $\pi_w \in \Pi(\mathbb{Q}_w^1, \mathbb{Q}_w^2)$. To prove the converse inequality, we define $\Pi_x(\mathbb{Q}_x^1, \mathbb{Q}_x^2)$ as the set of all joint distributions $\pi \in \mathcal{M}(\mathbb{R}^{n+m} \times \mathbb{R}^{n+m})$ of $(x_1, w_1) \in \mathbb{R}^{n+m}$ and $(x_2, w_2) \in \mathbb{R}^{n+m}$ under which x_1 and x_2 have marginal distributions $\mathbb{Q}_x^1$ and $\mathbb{Q}_x^2$, respectively. Similarly, we define $\Pi_w(\mathbb{Q}_w^1, \mathbb{Q}_w^2)$ as the set of all joint distributions $\pi \in \mathcal{M}(\mathbb{R}^{n+m} \times \mathbb{R}^{n+m})$ of $(x_1, w_1) \in \mathbb{R}^{n+m}$ and $(x_2, w_2) \in \mathbb{R}^{n+m}$ under which w_1 and w_2 have marginal distributions $\mathbb{Q}_w^1$ and $\mathbb{Q}_w^2$, respectively. Using this notation, we find

$$\begin{aligned}
\mathbb{W}(\mathbb{Q}_x^1 \times \mathbb{Q}_w^1, \mathbb{Q}_x^2 \times \mathbb{Q}_w^2)^2 \geq & \inf_{\pi \in \Pi(\mathbb{Q}_x^1 \times \mathbb{Q}_w^1, \mathbb{Q}_x^2 \times \mathbb{Q}_w^2)} \int_{\mathbb{R}^{n+m} \times \mathbb{R}^{n+m}} \left\| x_1 - x_2 \right\|^2 \pi(\mathrm{d}x_1, \mathrm{d}w_1, \mathrm{d}x_2, \mathrm{d}w_2) \\
& + \inf_{\pi \in \Pi(\mathbb{Q}_x^1 \times \mathbb{Q}_w^1, \mathbb{Q}_x^2 \times \mathbb{Q}_w^2)} \int_{\mathbb{R}^{n+m} \times \mathbb{R}^{n+m}} \left\| w_1 - w_2 \right\|^2 \pi(\mathrm{d}x_1, \mathrm{d}w_1, \mathrm{d}x_2, \mathrm{d}w_2) \\
\geq & \inf_{\pi \in \Pi_x(\mathbb{Q}_x^1, \mathbb{Q}_x^2)} \int_{\mathbb{R}^{n+m} \times \mathbb{R}^{n+m}} \left\| x_1 - x_2 \right\|^2 \pi(\mathrm{d}x_1, \mathrm{d}w_1, \mathrm{d}x_2, \mathrm{d}w_2) \\
& + \inf_{\pi \in \Pi_w(\mathbb{Q}_w^1, \mathbb{Q}_w^2)} \int_{\mathbb{R}^{n+m} \times \mathbb{R}^{n+m}} \left\| w_1 - w_2 \right\|^2 \pi(\mathrm{d}x_1, \mathrm{d}w_1, \mathrm{d}x_2, \mathrm{d}w_2) \\
= & \, \mathbb{W}(\mathbb{Q}_x^1, \mathbb{Q}_x^2)^2 + \mathbb{W}(\mathbb{Q}_w^1, \mathbb{Q}_w^2)^2,
\end{aligned}$$

where the first inequality exploits the superadditivity of the infimum operator, while the second inequality holds because $\Pi(\mathbb{Q}_x^1 \times \mathbb{Q}_w^1, \mathbb{Q}_x^2 \times \mathbb{Q}_w^2)$ contains both $\Pi_x(\mathbb{Q}_x^1, \mathbb{Q}_x^2)$ and $\Pi_w(\mathbb{Q}_w^1, \mathbb{Q}_w^2)$ as subsets. The equality in the last line follows from the observation that for any $\pi \in \Pi_x(\mathbb{Q}_x^1, \mathbb{Q}_x^2)$ the marginal distribution π_x of (x_1, x_2) is an element of $\Pi(\mathbb{Q}_x^1, \mathbb{Q}_x^2)$, and for any $\pi \in \Pi_w(\mathbb{Q}_x^1, \mathbb{Q}_x^2)$ the marginal distribution π_w of (w_1, w_2) is an element of $\Pi(\mathbb{Q}_w^1, \mathbb{Q}_w^2)$. Thus, the claim follows. □

If we denote the ambiguity sets (2.8) and (2.9) temporarily by $\mathbb{B}_{\rho_x, \rho_w}(\widehat{\mathbb{P}})$ and $\mathbb{B}'_\rho(\widehat{\mathbb{P}})$ in order to make their dependence on the hyperparameters explicit, then Lemma 2.2 implies that

$$\mathbb{B}'_\rho(\widehat{\mathbb{P}}) = \bigcup_{\rho_x^2 + \rho_w^2 \leq \rho^2} \mathbb{B}_{\rho_x, \rho_w}(\widehat{\mathbb{P}}).$$

This relation suggests that $\mathbb{B}'_\rho(\widehat{\mathbb{P}})$ could be substantially larger than $\mathbb{B}_{\rho_x, \rho_w}(\widehat{\mathbb{P}})$ for any fixed $\rho, \rho_x, \rho_w \geq 0$ with $\rho_x^2 + \rho_w^2 = \rho^2$ and thus lead to substantially more conservative estimators.

The key contributions of this paper can be summarized as follows.

- Using the subset of affine estimators and a superset of distributions prescribed by the Gelbrich distance, we propose a safe approximation for the generalized distributionally robust minimum mean square error estimation problem (2.7). We prove that this safe approximation is equivalent to a finite convex optimization problem, whose optimal solution is used to form the optimal affine estimator.

- If the nominal measure $\widehat{\mathbb{P}}$ is a Gaussian distribution, we propose in Section 2.2 a conservative approximation for the dual estimation problem (2.22) by confining the ambiguity set to Gaussian distributions. We prove that this conservative approximation is equivalent to a finite convex optimization problem, whose optimal solution is used to form the least favorable prior.

- If the nominal measure $\widehat{\mathbb{P}}$ is a Gaussian distribution, we consider in Section 2.3 the dual MMSE estimation problem. We show that this dual problem is equivalent to a finite convex optimization problem. We then prove in Section 2.4 that the optimal Wasserstein MMSE estimator that solves (2.7) is an affine estimator, and more interestingly, this optimal estimator is the Bayesian MMSE estimator corresponding to the least favorable prior, and can be recovered from the least favorable prior.

- In Section 2.5 we discuss the extension to the case when $\widehat{\mathbb{P}}$ is an elliptical distribution. Interestingly, with a slight modification of the required assumption, the results in Section 2.4 still hold and we can show that the optimality of the affine estimator is also retained.
- We further develop in Section 2.6 a decomposable Frank-Wolfe algorithm to solve the resulting nonlinear SDP programs with a linear convergence rate. We illustrate in Section 2.7 the performance of our proposed estimator on an image denoising application, and show the superior quality of the Wasserstein affine estimator.

The minimax MSE estimation problem (2.5a) also known as Chebyshev center [4] was first studied with ellipsoidal uncertainty set in [55, 56] and then extended to the case for which the matrix H has circulant pattern in [9]. Exact reformulation was presented in [8] when the uncertainty set $\mathbb{X}$ is the intersection of two ellipsoids. A semidefinite programming relaxation was later introduced in [54] to handle the case when $\mathbb{X}$ is the intersection of several ellipsoids. In this setting, the near-optimality of the affine estimators is proved in [96]. Distributionally robust MSE estimator have been studied in [57, 7, 53, 125] using uncertainty sets for the signal and noise covariance matrices and in [108, 109, 181, 182] using ambiguity sets defined via information divergences. In particular, restring to affine estimator, a minimax theorem using the uncertainty set over the covariance matrix space is established in [125]. Furthermore, the optimality of affine estimators is proved in [108, 109, 181, 182] for ambiguity sets constructed by information divergences.

The paper is structured as follows. Sections 2.2 and 2.3 develop conservative approximations for the primal and dual Wasserstein MMSE estimation problems, respectively, both of which are equivalent to tractable convex programs. Section 2.4 shows that if the nominal distribution is normal, then both approximations are exact and can be used to find a Nash equilibrium for the zero-sum game between the statistician and nature. Extensions to elliptical nominal distributions are discussed in Section 2.5. Section 2.6 develops an efficient Frank-Wolfe algorithm for the dual MMSE estimation problem, and Section 2.7.2 reports on numerical results.

Notation. For any $A \in \mathbb{R}^{d\times d}$ we use $\mathrm{Tr}\left[A\right]$ to denote the trace and $\|A\| = \sqrt{\mathrm{Tr}\left[A^\top A\right]}$ to denote the Frobenius norm of A. By slight abuse of notation, the Euclidean norm of $v \in \mathbb{R}^d$ is also denoted by $\|v\|$. Moreover, I_d stands for the identity matrix in $\mathbb{R}^{d\times d}$. For any $A, B \in \mathbb{R}^{d\times d}$, we use $\langle A, B\rangle = \mathrm{Tr}\left[A^\top B\right]$ to denote the trace inner product. The space of all symmetric matrices in $\mathbb{R}^{d\times d}$ is denoted by $\mathbb{S}^d$. We use $\mathbb{S}^d_+$ ($\mathbb{S}^d_{++}$) to represent the cone of symmetric positive semidefinite (positive definite) matrices in $\mathbb{S}^d$. For any $A, B \in \mathbb{S}^d$, the relation $A \succeq B$ ($A \succ B$) means that $A - B \in \mathbb{S}^d_+$ ($A - B \in \mathbb{S}^d_{++}$). The unique positive semidefinite square root of a matrix

$A \in \mathbb{S}_+^d$ is denoted by $A^{\frac{1}{2}}$. For any $A \in \mathbb{S}^d$, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of A, respectively.

2.2 The Gelbrich MMSE Estimation Problem

The distributionally robust estimation problem (2.7) poses two fundamental challenges. First, checking feasibility of the inner maximization problem in (2.7) requires computing the Wasserstein distances $\mathbb{W}(\widehat{\mathbb{P}}_x, \mathbb{Q}_x)$ and $\mathbb{W}(\widehat{\mathbb{P}}_w, \mathbb{Q}_w)$, which is #P-hard even if $\widehat{\mathbb{P}}_x$ and $\widehat{\mathbb{P}}_w$ are simple two-point distributions, while $\mathbb{Q}_x$ and $\mathbb{Q}_w$ are uniform distributions on hypercubes [163]. Efficient algorithms for computing Wasserstein distances are available only if both involved distributions are discrete [33, 135, 158], and analytical formulas are only known in exceptional cases (*e.g.*, if both distributions are Gaussian [72] or belong to the same family of elliptical distributions [71]). The second challenge is that the outer minimization problem in (2.7) constitutes an infinite-dimensional functional optimization problem. In order to bypass these computational challenges, we first seek a conservative approximation for (2.7) by relaxing the ambiguity set $\mathbb{B}(\widehat{\mathbb{P}})$ and restricting the feasible set $\mathscr{F}$. We begin by constructing an outer approximation for the ambiguity set. To this end, we introduce a new distance measure on the space of mean vectors and covariance matrices.

Definition 2.3 (Gelbrich distance)**.** *For any $d \in \mathbb{N}$, the Gelbrich distance between two tuples of mean vectors and covariance matrices $(\mu_1, \Sigma_1), (\mu_2, \Sigma_2) \in \mathbb{R}^d \times \mathbb{S}_+^d$ is defined as*

$$\mathbb{G}\left((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)\right) \triangleq \sqrt{\|\mu_1 - \mu_2\|^2 + \operatorname{Tr}\left[\Sigma_1 + \Sigma_2 - 2\left(\Sigma_2^{\frac{1}{2}} \Sigma_1 \Sigma_2^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]}.$$

The dependence of the Gelbrich distance on d is notationally suppressed in order to avoid clutter. One can show that $\mathbb{G}$ constitutes a metric on $\mathbb{R}^d \times \mathbb{S}_+^d$, that is, $\mathbb{G}$ is symmetric, non-negative, vanishes if and only if $(\mu_1, \Sigma_1) = (\mu_2, \Sigma_2)$ and satisfies the triangle inequality [72, pp. 239].

Proposition 2.4 (Commuting covariance matrices [72, p. 239])**.** *If $\mu_1, \mu_2 \in \mathbb{R}^d$ are identical and $\Sigma_1, \Sigma_2 \in \mathbb{S}_+^d$ commute ($\Sigma_1\Sigma_2 = \Sigma_2\Sigma_1$), then the Gelbrich distance simplifies to $\mathbb{G}\left((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)\right) = \|\sqrt{\Sigma_1} - \sqrt{\Sigma_2}\|$.*

While $\mathbb{G}$ itself is non-convex, we will see below that $\mathbb{G}^2$ is convex. Our interest in the Gelbrich distance stems mainly from the next proposition, which lower bounds the Wasserstein distance between two distributions in terms of their first- and second-order moments. We will later see that this bound becomes tight when $\mathbb{Q}_1$ and $\mathbb{Q}_2$ are normal or—more generally—elliptical distributions of the same type.

Proposition 2.5 (Moment bound on the Wasserstein distance [71, Theorem 2.1])**.** *For any distributions $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathscr{M}(\mathbb{R}^d)$ with mean vectors $\mu_1, \mu_2 \in \mathbb{R}^d$ and covariance matrices $\Sigma_1, \Sigma_2 \in$*

$\mathbb{S}_+^d$*, respectively, we have*

$$\mathbb{W}(\mathbb{Q}_1,\mathbb{Q}_2) \geq \mathbb{G}\big((\mu_1,\Sigma_1),(\mu_2,\Sigma_2)\big).$$

Proposition 2.5 prompts us to construct an outer approximation for the Wasserstein ambiguity set $\mathbb{B}(\widehat{\mathbb{P}})$ by using the Gelbrich distance. Specifically, we define the *Gelbrich ambiguity set* centered at $\widehat{\mathbb{P}} = \widehat{\mathbb{P}}_x \times \widehat{\mathbb{P}}_w$ as

$$\mathbb{G}(\widehat{\mathbb{P}}) \triangleq \left\{ \mathbb{Q}_x \times \mathbb{Q}_w : \begin{array}{l} \mathbb{Q}_x \in \mathscr{M}(\mathbb{R}^n),\ \ \mu_x = \mathbb{E}_{\mathbb{Q}_x}[x],\ \ \Sigma_x = \mathbb{E}_{\mathbb{Q}_x}[xx^\top] - \mu_x\mu_x^\top \\ \mathbb{Q}_w \in \mathscr{M}(\mathbb{R}^m),\ \mu_w = \mathbb{E}_{\mathbb{Q}_w}[w], \Sigma_w = \mathbb{E}_{\mathbb{Q}_w}[ww^\top] - \mu_w\mu_w^\top \\ \mathbb{G}((\mu_x,\Sigma_x),(\widehat{\mu}_x,\widehat{\Sigma}_x)) \leq \rho_x,\ \ \mathbb{G}((\mu_w,\Sigma_w),(\widehat{\mu}_w,\widehat{\Sigma}_w)) \leq \rho_w \end{array} \right\},$$

where $\widehat{\mu}_x$ and $\widehat{\mu}_w$ denote the mean vectors and $\widehat{\Sigma}_x$ and $\widehat{\Sigma}_w$ the covariance matrices of $\widehat{\mathbb{P}}_x$ and $\widehat{\mathbb{P}}_w$, respectively.

Corollary 2.6 (Relation between Gelbrich and Wasserstein ambiguity sets)**.** *For any* $\widehat{\mathbb{P}} = \widehat{\mathbb{P}}_x \times \widehat{\mathbb{P}}_w$ *with* $\widehat{\mathbb{P}}_x \in \mathscr{M}(\mathbb{R}^n)$ *and* $\widehat{\mathbb{P}}_w \in \mathscr{M}(\mathbb{R}^m)$ *we have* $\mathbb{B}(\widehat{\mathbb{P}}) \subseteq \mathbb{G}(\widehat{\mathbb{P}})$.

Proof. Select any $\mathbb{Q} = \mathbb{Q}_x \times \mathbb{Q}_w \in \mathbb{B}(\widehat{\mathbb{P}})$ and define μ_x and μ_w as the mean vectors and Σ_x and Σ_w as the covariance matrices of $\mathbb{Q}_x$ and $\mathbb{Q}_w$, respectively. By Proposition 2.5 we then have

$$\mathbb{G}((\mu_x,\Sigma_x),(\widehat{\mu}_x,\widehat{\Sigma}_x)) \leq \mathbb{W}(\mathbb{Q}_x,\widehat{\mathbb{P}}_x) \leq \rho_x \quad \text{and} \quad \mathbb{G}((\mu_w,\Sigma_w),(\widehat{\mu}_w,\widehat{\Sigma}_w)) \leq \mathbb{W}(\mathbb{Q}_w,\widehat{\mathbb{P}}_w) \leq \rho_w,$$

which in turn implies that $\mathbb{Q} \in \mathbb{G}(\widehat{\mathbb{P}})$. We may thus conclude that $\mathbb{B}(\widehat{\mathbb{P}}) \subseteq \mathbb{G}(\widehat{\mathbb{P}})$. □

By restricting $\mathscr{F}$ to the set $\mathscr{A}$ of all affine estimators while relaxing $\mathbb{B}(\widehat{\mathbb{P}})$ to the Gelbrich ambiguity set $\mathbb{G}(\widehat{\mathbb{P}})$, we obtain the following conservative approximation of the distributionally robust estimation problem (2.7).

$$\underset{\psi\in\mathscr{A}}{\text{minimize}} \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathscr{R}(\psi,\mathbb{Q}) \tag{2.10}$$

From now on we will call (2.7) and (2.10) the Wasserstein and Gelbrich MMSE estimation problems, and we will refer to their minimizers as Wasserstein and Gelbrich MMSE estimators, respectively. As the average risk $\mathscr{R}(\psi,\mathbb{Q})$ of a fixed affine estimator $\psi \in \mathscr{A}$ is convex and quadratic in the mean vector μ and affine in the covariance matrix Σ of the distribution $\mathbb{Q}$, the inner maximization problem in (2.10) is non-convex. Thus, one might suspect that the Gelbrich MMSE estimation problem is intractable. Below we will show, however, that (2.10) is equivalent to a finite convex program that can be solved in polynomial time. To this end, we first show that, under mild conditions, problem (2.10) is stable with respect to changes of its input parameters.

Proposition 2.7 (Regularity of the Gelbrich MMSE estimation problem)**.** *The Gelbrich MMSE estimation problem* (2.10) *enjoys the following regularity properties.*

(i) **Conservativeness:** *Problem* (2.10) *upper bounds the Wasserstein MMSE estimation problem* (2.7).

(ii) **Solvability:** *The minimum of* (2.10) *is attained if* $\widehat{\Sigma}_w \succ 0$ *or* $\rho_w > 0$.

(iii) **Stability:** *The minimum of* (2.10) *is continuous in* $(\rho_x, \rho_w, \widehat{\mu}_x, \widehat{\mu}_w, \widehat{\Sigma}_x, \widehat{\Sigma}_w)$ *if* $\widehat{\Sigma}_w \succ 0$ *or* $\rho_w > 0$.

Proof. The Gelbrich MMSE estimation problem (2.10) upper bounds the Wasserstein MMSE estimation problem (2.7) because $\mathcal{A} \subseteq \mathcal{F}$ and $\mathbb{G}(\widehat{\mathbb{P}}) \supseteq \mathbb{B}(\widehat{\mathbb{P}})$; see Corollary 2.6. Thus, assertion (i) follows. Recall now that any $\psi \in \mathcal{A}$ can be represented as $\psi(y) = Ay + b$ for some $A \in \mathbb{R}^{n\times m}$ and $b \in \mathbb{R}^n$. Moreover, for any distribution $\mathbb{Q} = \mathbb{Q}_x \times \mathbb{Q}_w \in \mathbb{G}(\widehat{\mathbb{P}})$, denote by μ_x and μ_w the mean vectors and by Σ_x and Σ_w the covariance matrices of $\mathbb{Q}_x$ and $\mathbb{Q}_w$, respectively. Hence, the objective function and the constraints of (2.10) depend on ψ and $\mathbb{Q}$ only through A, b, μ_x, μ_w, Σ_x and Σ_w. Indeed, the average risk of ψ under $\mathbb{Q}$ satisfies

$$\begin{aligned}\mathcal{R}(\psi, \mathbb{Q}) &= \mathbb{E}_{\mathbb{Q}}\left[\|x - A(Hx + w) - b\|^2\right] = \mathbb{E}_{\mathbb{Q}}\left[\|(I_n - AH)x - Aw - b\|^2\right] \\ &= \left\langle (I_n - AH)^\top (I_n - AH), \Sigma_x + \mu_x\mu_x^\top \right\rangle + \left\langle A^\top A, \Sigma_w + \mu_w\mu_w^\top \right\rangle + b^\top b \\ &\quad - 2\mu_x^\top (I_n - AH)^\top A\mu_w - 2b^\top((I_n - AH)\mu_x - A\mu_w) \triangleq f(A, b, \mu_x, \mu_w, \Sigma_x, \Sigma_w).\end{aligned}$$

Similarly, the constraints $\psi \in \mathcal{A}$ and $\mathbb{Q} \in \mathbb{G}(\widehat{\mathbb{P}})$ can be reformulated in terms of A, b, μ_x, μ_w, Σ_x and Σ_w. Thus, the Gelbrich MMSE estimation problem (2.10) is equivalent to

$$\begin{array}{lll} \inf\limits_{A,b} & \sup\limits_{\substack{\mu_x,\mu_w \\ \Sigma_x,\Sigma_w \succeq 0}} & f(A, b, \mu_x, \mu_w, \Sigma_x, \Sigma_w) \\ & \text{s.t.} & \mathbb{G}((\mu_x, \Sigma_x), (\widehat{\mu}_x, \widehat{\Sigma}_x))^2 \le \rho_x^2 \\ & & \mathbb{G}((\mu_w, \Sigma_w), (\widehat{\mu}_w, \widehat{\Sigma}_w))^2 \le \rho_w^2. \end{array} \tag{2.11}$$

Note that both sides of the two Gelbrich distance constraints have been squared without loss of generality. In the remainder we will use the shorthand $\theta \triangleq (\rho_x, \rho_w, \widehat{\mu}_x, \widehat{\mu}_w, \widehat{\Sigma}_x, \widehat{\Sigma}_w)$ to denote the vector of the problem's input parameters, which ranges over the set $\Theta \triangleq \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{S}_+^n \times \mathbb{S}_+^m$. As f is continuous, while the feasible set of the inner maximization problem is compact and depends continuously on θ, the optimal value of the minimax problem (2.11) is locally bounded in θ. Next, we introduce the continuous functions

$$\mu_x(\theta) \triangleq \widehat{\mu}_x, \quad \mu_w(\theta) \triangleq \widehat{\mu}_w, \quad \Sigma_x(\theta) = \widehat{\Sigma}_x \quad \text{and} \quad \Sigma_w(\theta) \triangleq \left(\widehat{\Sigma}_w^{\frac{1}{2}} + \frac{\rho_w}{\sqrt{m}} I_m\right)^2.$$

Trivially, we have $\mathbb{G}((\mu_x(\theta), \Sigma_x(\theta)), (\widehat{\mu}_x, \widehat{\Sigma}_x)) = 0$ for all $\theta \in \Theta$. Moreover, it is easy to see that $\widehat{\Sigma}_w$ and $\Sigma_w(\theta)$ commute. Proposition 2.4 thus implies that

$$\mathbb{G}((\mu_w(\theta), \Sigma_w(\theta)), (\widehat{\mu}_w, \widehat{\Sigma}_w)) = \left\|\Sigma_w^{\frac{1}{2}}(\theta) - \widehat{\Sigma}_w^{\frac{1}{2}}\right\| = \rho_w$$

for all $\theta \in \Theta$. This allows us to conclude that $\mu_x(\theta)$, $\mu_w(\theta)$, $\Sigma_x(\theta)$ and $\Sigma_w(\theta)$ are feasible in the inner maximization problem in (2.11). Assume now that $\widehat{\Sigma}_w \succ 0$ or $\rho_w > 0$, which implies that $\Sigma_w(\theta) \succ 0$. In this case f is strongly convex in the decision variables A and b of the outer minimization problem in (2.11), and

$$\left\{(A,b) \in \mathbb{R}^{n\times m} \times \mathbb{R}^n : f(A,b,\mu_x(\theta),\mu_w(\theta),\Sigma_x(\theta),\Sigma_w(\theta)) \leq \bar{f}\right\}$$

is compact for every $\bar{f} \in \mathbb{R}$. Thus, the minimax problem (2.11) satisfies all conditions of Lemma 2.32 in Appendix 2.8.1, which implies that both (2.11) and the equivalent Gelbrich MMSE estimation problem (2.10) are solvable and that their optimal value changes continuously with θ whenever $\widehat{\Sigma}_w \succ 0$ or $\rho_w > 0$. □

We are now ready to prove the main result of this section.

Theorem 2.8 (Gelbrich MMSE estimation problem)**.** *The Gelbrich MMSE estimation problem* (2.10) *is equivalent to the finite convex optimization problem*

$$\begin{array}{ll} \inf & \gamma_x\left(\rho_x^2 - \mathrm{Tr}[\widehat{\Sigma}_x]\right) + \gamma_x^2\langle [\gamma_x I_n - (I_n - AH)^\top (I_n - AH)]^{-1}, \widehat{\Sigma}_x\rangle \\ & \quad + \gamma_w\left(\rho_w^2 - \mathrm{Tr}[\widehat{\Sigma}_w]\right) + \gamma_w^2\langle (\gamma_w I_m - A^\top A)^{-1}, \widehat{\Sigma}_w\rangle \\ \text{s.t.} & A \in \mathbb{R}^{n\times m}, \quad \gamma_x, \gamma_w \in \mathbb{R}_+ \\ & \gamma_x I_n - (I_n - AH)^\top (I_n - AH) \succ 0, \quad \gamma_w I_m - A^\top A \succ 0. \end{array} \tag{2.12}$$

Moreover, if $\widehat{\Sigma}_w \succ 0$ *or* $\rho_w > 0$*, then* (2.12) *admits an optimal solution*[1] $A^\star$*, and the infimum of* (2.10) *is attained by the affine estimator* $\psi^\star(y) = A^\star y + b^\star$*, where* $b^\star = \widehat{\mu}_x - A^\star(H\widehat{\mu}_x + \widehat{\mu}_w)$*.*

Proof. Throughout this proof we denote by $\psi_{A,b} \in \mathscr{A}$ the affine estimator $\psi_{A,b}(y) = Ay + b$ corresponding to the sensitivity matrix $A \in \mathbb{R}^{n\times m}$ and the vector $b \in \mathbb{R}^n$ of intercepts. In the following we fix some $A \in \mathbb{R}^{n\times m}$ and define $K = I_n - AH$ in order to simplify the notation. By the definitions of the average risk $\mathscr{R}(\psi, \mathbb{Q})$ and the Gelbrich ball $\mathbb{G}(\widehat{\mathbb{P}})$, we then have

$$\inf_b \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathscr{R}(\psi_{A,b}, \mathbb{Q}) = \left\{ \begin{array}{lll} \inf\limits_b & \sup\limits_{\substack{\mu_x,\mu_w \\ \Sigma_x,\Sigma_w \succeq 0}} & \langle K^\top K, \Sigma_x + \mu_x\mu_x^\top\rangle + \langle A^\top A, \Sigma_w + \mu_w\mu_w^\top\rangle + b^\top b \\ & & \quad -2\mu_x^\top K^\top A\mu_w - 2b^\top(K\mu_x - A\mu_w) \\ & \text{s.t.} & \mathbb{G}((\mu_x,\Sigma_x),(\widehat{\mu}_x,\widehat{\Sigma}_x))^2 \leq \rho_x^2 \\ & & \mathbb{G}((\mu_w,\Sigma_w),(\widehat{\mu}_w,\widehat{\Sigma}_w))^2 \leq \rho_w^2. \end{array} \right. \tag{2.13}$$

The outer minimization problem in (2.13) is convex because the objective function of the minimax problem is convex in b for any fixed $(\mu_x, \mu_w, \Sigma_x, \Sigma_w)$ and because convexity is preserved under maximization. Moreover, the inner maximization problem in (2.13) is non-convex because its objective function is convex in (μ_x, μ_w). This observation prompts us to maximize

[1] We say that $A^\star$ solves (2.12) if adding the constraint $A = A^\star$ does not change the infimum of (2.12). Note that the infimum of the resulting problem over (γ_x, γ_w) may not be attained, *i.e.*, the existence of a solution $A^\star$ does not imply that (2.12) is solvable.

over (μ_x, μ_w) and (Σ_x, Σ_w) sequentially and to reformulate (2.13) as

$$\begin{aligned} \inf_{b} \sup_{\substack{\mu_x,\mu_w \\ \|\mu_x-\widehat{\mu}_x\|\leq\rho_x \\ \|\mu_w-\widehat{\mu}_w\|\leq\rho_w}} \quad \sup_{\Sigma_x,\Sigma_w\succeq 0} \quad & \left\langle K^\top K, \Sigma_x + \mu_x\mu_x^\top \right\rangle + \left\langle A^\top A, \Sigma_w + \mu_w\mu_w^\top \right\rangle + b^\top b \\ & -2\mu_x^\top K^\top A\mu_w - 2b^\top(K\mu_x - A\mu_w) \\ \text{s.t.} \quad & \mathbb{G}((\mu_x,\Sigma_x),(\widehat{\mu}_x,\widehat{\Sigma}_x))^2 \leq \rho_x^2 \\ & \mathbb{G}((\mu_w,\Sigma_w),(\widehat{\mu}_w,\widehat{\Sigma}_w))^2 \leq \rho_w^2. \end{aligned} \tag{2.14}$$

As $\|\mu_x - \widehat{\mu}_x\| \leq \mathbb{G}((\mu_x,\Sigma_x),(\widehat{\mu}_x,\widehat{\Sigma}_x))$ and as this inequality is tight for $\Sigma_x = \widehat{\Sigma}_x$, the extra constraint $\|\mu_x - \widehat{\mu}_x\| \leq \rho_x$ is actually redundant and merely ensures that the maximization problem over Σ_x remains feasible for any admissible choice of μ_x. An analogous statement holds for μ_w and Σ_w. By the definition of the Gelbrich distance, the innermost maximization problem over (Σ_x, Σ_w) in (2.14) admits the Lagrangian dual

$$\begin{aligned} \inf_{\gamma_x,\gamma_w\geq 0} \sup_{\Sigma_x,\Sigma_w\succeq 0} & \left\langle K^\top K, \Sigma_x + \mu_x\mu_x^\top \right\rangle + \left\langle A^\top A, \Sigma_w + \mu_w\mu_w^\top \right\rangle + b^\top b - 2\mu_x^\top K^\top A\mu_w - 2b^\top(K\mu_x - A\mu_w) \\ & + \gamma_x\left(\rho_x^2 - \|\mu_x - \widehat{\mu}_x\|^2 - \operatorname{Tr}\left[\Sigma_x + \widehat{\Sigma}_x - 2\left(\widehat{\Sigma}_x^{\frac{1}{2}}\Sigma_x\widehat{\Sigma}_x^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]\right) \\ & + \gamma_w\left(\rho_w^2 - \|\mu_w - \widehat{\mu}_w\|^2 - \operatorname{Tr}\left[\Sigma_w + \widehat{\Sigma}_w - 2\left(\widehat{\Sigma}_w^{\frac{1}{2}}\Sigma_w\widehat{\Sigma}_w^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]\right). \end{aligned} \tag{2.15}$$

Strong duality holds by [13, Proposition 5.5.4], which applies because the primal problem has a non-empty compact feasible set. Next, we observe that the inner maximization problem in (2.15) can be solved analytically by using Proposition 2.33 in the appendix, and thus the dual problem (2.15) is equivalent to

$$\begin{aligned} \inf_{\substack{\gamma_x,\gamma_w \\ \gamma_x I_n \succ K^\top K \\ \gamma_w I_m \succ A^\top A}} \quad & \left\langle K^\top K, \mu_x\mu_x^\top \right\rangle + \left\langle A^\top A, \mu_w\mu_w^\top \right\rangle + b^\top b - 2\mu_x^\top K^\top A\mu_w - 2b^\top(K\mu_x - A\mu_w) \\ & + \gamma_x\left(\rho_x^2 - \|\mu_x - \widehat{\mu}_x\|^2 - \operatorname{Tr}\left[\widehat{\Sigma}_x\right] + \gamma_x\left\langle(\gamma_x I_n - K^\top K)^{-1}, \widehat{\Sigma}_x\right\rangle\right) \\ & + \gamma_w\left(\rho_w^2 - \|\mu_w - \widehat{\mu}_w\|^2 - \operatorname{Tr}\left[\widehat{\Sigma}_w\right] + \gamma_w\left\langle(\gamma_w I_m - A^\top A)^{-1}, \widehat{\Sigma}_w\right\rangle\right). \end{aligned} \tag{2.16}$$

Substituting (2.16) back into (2.14) then allows us to reformulate the Gelbrich MMSE estimation problem (2.7) as

$$\begin{aligned} \inf_{b} \sup_{\substack{\mu_x,\mu_w \\ \|\mu_x-\widehat{\mu}_x\|\leq\rho_x \\ \|\mu_w-\widehat{\mu}_w\|\leq\rho_w}} \inf_{\substack{\gamma_x,\gamma_w \\ \gamma_x I_n \succ K^\top K \\ \gamma_w I_m \succ A^\top A}} & \left\langle K^\top K, \mu_x\mu_x^\top \right\rangle + \left\langle A^\top A, \mu_w\mu_w^\top \right\rangle + b^\top b - 2\mu_x^\top K^\top A\mu_w - 2b^\top(K\mu_x - A\mu_w) \\ & + \gamma_x\left(\rho_x^2 - \|\mu_x - \widehat{\mu}_x\|^2 - \operatorname{Tr}\left[\widehat{\Sigma}_x\right] + \gamma_x\left\langle(\gamma_x I_n - K^\top K)^{-1}, \widehat{\Sigma}_x\right\rangle\right) \\ & + \gamma_w\left(\rho_w^2 - \|\mu_w - \widehat{\mu}_w\|^2 - \operatorname{Tr}\left[\widehat{\Sigma}_w\right] + \gamma_w\left\langle(\gamma_w I_m - A^\top A)^{-1}, \widehat{\Sigma}_w\right\rangle\right). \end{aligned} \tag{2.17}$$

The infimum of the inner minimization problem over (γ_x, γ_w) in (2.17) is convex quadratic in b. Moreover, it is concave in (μ_x, μ_w) because $K^\top K - \gamma_x I_n \prec 0$ and $A^\top A - \gamma_w I_m \prec 0$ for any feasible choice of (γ_x, γ_w) and because concavity is preserved under minimization. Finally, the feasible set for (μ_x, μ_w) is convex and compact. By Sion's classical minimax theorem, we may therefore interchange the infimum over b with the supremum over (μ_x, μ_w). The minimization problem over b thus reduces to an unconstrained (strictly) convex quadratic program that has the unique optimal solution $b = K\mu_x - A\mu_w$. Substituting this expression back into (2.17) then yields

$$\begin{aligned} \sup_{\substack{\mu_x,\mu_w \\ \|\mu_x-\widehat{\mu}_x\|\leq\rho_x \\ \|\mu_w-\widehat{\mu}_w\|\leq\rho_w}} \inf_{\substack{\gamma_x,\gamma_w \\ \gamma_x I_n \succ K^\top K \\ \gamma_w I_m \succ A^\top A}} \; & \gamma_x\left(\rho_x^2 - \|\mu_x - \widehat{\mu}_x\|^2 - \operatorname{Tr}\left[\widehat{\Sigma}_x\right]\right) + \gamma_x^2\big\langle(\gamma_x I_n - K^\top K)^{-1}, \widehat{\Sigma}_x\big\rangle \\ & +\gamma_w\left(\rho_w^2 - \|\mu_w - \widehat{\mu}_w\|^2 - \operatorname{Tr}\left[\widehat{\Sigma}_w\right]\right) + \gamma_w^2\big\langle(\gamma_w I_m - A^\top A)^{-1}, \widehat{\Sigma}_w\big\rangle. \end{aligned} \tag{2.18}$$

It is easy to verify that the resulting maximization problem over (μ_x, μ_w) is solved by $\mu_x = \widehat{\mu}_x$ and $\mu_w = \widehat{\mu}_w$. Substituting the corresponding optimal value into (2.13) finally yields

$$\inf_b \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathscr{R}(\psi_{A,b}, \mathbb{Q}) = \left\{ \begin{aligned} \inf_{\substack{\gamma_x,\gamma_w \\ \gamma_x I_n \succ K^\top K \\ \gamma_w I_m \succ A^\top A}} \; & \gamma_x\left(\rho_x^2 - \operatorname{Tr}\left[\widehat{\Sigma}_x\right]\right) + \gamma_x^2\big\langle(\gamma_x I_n - K^\top K)^{-1}, \widehat{\Sigma}_x\big\rangle \\ & +\gamma_w\left(\rho_w^2 - \operatorname{Tr}\left[\widehat{\Sigma}_w\right]\right) + \gamma_w^2\big\langle(\gamma_w I_m - A^\top A)^{-1}, \widehat{\Sigma}_w\big\rangle. \end{aligned} \right.$$

From the above equation and the definition of K it is evident that the Gelbrich MMSE estimation problem

$$\inf_{\psi\in\mathscr{A}} \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathscr{R}(\psi, \mathbb{Q}) = \inf_{A,b} \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathscr{R}(\psi_{A,b}, \mathbb{Q}) \tag{2.19}$$

is indeed equivalent to the finite convex optimization problem (2.12).

Assume now that $\widehat{\Sigma}_w \succ 0$ or $\rho_w > 0$. In this case we know from Proposition 2.7 (ii) that the Gelbrich MMSE estimation problem (2.19) admits an optimal affine estimator $\psi^\star(y) = A^\star y + b^\star$ for some $A^\star \in \mathbb{R}^{n\times m}$ and $b^\star \in \mathbb{R}^m$. The reasoning in the first part of the proof then implies that $A^\star$ solves (2.12). Moreover, it implies that $b^\star$ is optimal in (2.13) when we fix $A = A^\star$. As (2.13) is equivalent to (2.17) and as the unique optimal solution of (2.17) for $A = A^\star$ is given by $b = \widehat{\mu}_x - A^\star(H\widehat{\mu}_x + \widehat{\mu}_w)$, we may finally conclude that

$$b^\star = \widehat{\mu}_x - A^\star(H\widehat{\mu}_x + \widehat{\mu}_w).$$

By reversing these arguments, one can further show that if $A^\star$ solves (2.12) and $b^\star$ is defined as above, then the affine estimator $\psi^\star(y) = A^\star y + b^\star$ is optimal in (2.19). This observation completes the proof. □

The strict semidefinite inequalities in (2.12) ensure that the inverse matrices in the objective function are well-defined. Using Schur complement arguments, the convex program (2.12) can be further simplified to a standard semidefinite program (SDP), which can be addressed with off-the-shelf solvers.

Corollary 2.9 (SDP reformulation). *The Gelbrich MMSE estimation problem* (2.10) *is equiva-*

lent to the SDP

$$
\begin{array}{cl}
\inf & \gamma_x\left(\rho_x^2-\mathrm{Tr}[\widehat{\Sigma}_x]\right)+\mathrm{Tr}[U_x]+\gamma_w\left(\rho_w^2-\mathrm{Tr}[\widehat{\Sigma}_w]\right)+\mathrm{Tr}[U_w]\\
\mathrm{s.t.} & A\in\mathbb{R}^{n\times m},\quad \gamma_x,\gamma_w\in\mathbb{R}_+\\
& U_x\in\mathbb{S}_+^n,\quad V_x\in\mathbb{S}_+^n,\quad U_w\in\mathbb{S}_+^m,\quad V_w\in\mathbb{S}_+^m\\
& \begin{bmatrix} U_x & \gamma_x\widehat{\Sigma}_x^{\frac{1}{2}}\\ \gamma_x\widehat{\Sigma}_x^{\frac{1}{2}} & V_x\end{bmatrix}\succeq 0,\quad \begin{bmatrix}\gamma_x I_n-V_x & I_n-H^\top A^\top\\ I_n-AH & I_n\end{bmatrix}\succeq 0\\
& \begin{bmatrix} U_w & \gamma_w\widehat{\Sigma}_w^{\frac{1}{2}}\\ \gamma_w\widehat{\Sigma}_w^{\frac{1}{2}} & V_w\end{bmatrix}\succeq 0,\quad \begin{bmatrix}\gamma_w I_m-V_w & A^\top\\ A & I_n\end{bmatrix}\succeq 0.
\end{array}
\tag{2.20}
$$

Proof. Define the extended real-valued function $h_w:\mathbb{R}^{n\times m}\times\mathbb{R}_+\to(-\infty,\infty]$ through

$$
h_w(A,\gamma_w)\triangleq\begin{cases}\gamma_w^2\langle(\gamma_w I_m-A^\top A)^{-1},\widehat{\Sigma}_w\rangle & \text{if } \gamma_w I_m-A^\top A\succ 0,\\ \infty & \text{otherwise.}\end{cases}
$$

If $\gamma_w I_m-A^\top A\succ 0$, then, we have

$$
\begin{aligned}
h_w(A,\gamma_w)&=\inf_{U_w\succeq 0}\left\{\mathrm{Tr}\left[U_w\right]\,:\,U_w\succeq\gamma_w^2\widehat{\Sigma}_w^{\frac{1}{2}}(\gamma_w I_m-A^\top A)^{-1}\widehat{\Sigma}_w^{\frac{1}{2}}\right\}\\
&=\inf_{U_w\succeq 0,V_w\succ 0}\left\{\mathrm{Tr}\left[U_w\right]\,:\,U_w\succeq\gamma_w^2\widehat{\Sigma}_w^{\frac{1}{2}}V_w^{-1}\widehat{\Sigma}_w^{\frac{1}{2}},\ \gamma_w I_m-A^\top A\succeq V_w\right\}\\
&=\inf_{U_w\succeq 0,V_w\succ 0}\left\{\mathrm{Tr}\left[U_w\right]\,:\,\begin{bmatrix}\gamma_w I_m-V_w & A^\top\\ A & I_n\end{bmatrix}\succeq 0,\ \begin{bmatrix}U_w & \gamma_x\widehat{\Sigma}_w^{\frac{1}{2}}\\ \gamma_w\widehat{\Sigma}_w^{\frac{1}{2}} & V_w\end{bmatrix}\succeq 0\right\},
\end{aligned}
\tag{2.21}
$$

where the first equality holds due to the cyclicity of the trace operator and because $U_w\succeq\bar{U}_w$ implies $\mathrm{Tr}\left[U_w\right]\geq\mathrm{Tr}\left[\bar{U}_w\right]$ for all $U_w,\bar{U}_w\succeq 0$, the second equality holds because $V_w\succeq\bar{V}_w$ is equivalent to $V_w^{-1}\preceq\bar{V}_w^{-1}$ for all $V_w,\bar{V}_w\succ 0$, and the last equality follows from standard Schur complement arguments; see, *e.g.*, [18, § A.5.5]. If $\gamma_w I_m-A^\top A\not\succ 0$, on the other hand, then the first matrix inequality in (2.21) implies that V_w must have at least one non-positive eigenvalue, which contradicts the constraint $V_w\succ 0$. The SDP (2.21) is therefore infeasible, and its infimum evaluates to ∞. Thus, $h_w(A,\gamma_w)$ coincides with the optimal value of the SDP (2.21) for all $A\in\mathbb{R}^{n\times m}$ and $\gamma_w\in\mathbb{R}_+$.

A similar SDP reformulation can be derived for the function $h_x:\mathbb{R}^{n\times m}\times\mathbb{R}_+\to(-\infty,\infty]$ defined through

$$
h_x(A,\gamma_x)\triangleq\begin{cases}\gamma_x^2\langle[\gamma_x I_n-(I_n-AH)^\top(I_n-AH)]^{-1},\widehat{\Sigma}_x\rangle & \text{if } \gamma_x I_n-(I_n-AH)^\top(I_n-AH)\succ 0,\\ \infty & \text{otherwise.}\end{cases}
$$

The claim now follows by substituting the SDP reformulations for $h_w(A,\gamma_w)$ and $h_x(A,\gamma_x)$ into (2.12). In doing so, we may relax the strict semidefinite inequalities $V_w\succ 0$ and $V_x\succ 0$ to weak inequalities $V_w\succeq 0$ and $V_x\succeq 0$, which amounts to taking the closure of the (non-empty) feasible set and does not change the infimum of problem (2.12). This observation completes the proof. □

Remark 2.10 (Numerical stability)**.** *The SDP* (2.20) *requires the square roots of the nominal covariance matrices as inputs. Unfortunately, iterative methods for computing matrix square roots often suffer from numerical instability in high dimensions. As a remedy, one may replace those matrix inequalities in* (2.20) *that involve* $\widehat{\Sigma}_x^{\frac{1}{2}}$ *and* $\widehat{\Sigma}_w^{\frac{1}{2}}$ *with*

$$\begin{bmatrix} U_x & \gamma_x \Lambda_x^\top \\ \gamma_x \Lambda_x & V_x \end{bmatrix} \succeq 0 \qquad \text{and} \qquad \begin{bmatrix} U_w & \gamma_w \Lambda_w^\top \\ \gamma_w \Lambda_w & V_w \end{bmatrix} \succeq 0,$$

where Λ_x *and* Λ_w *represent the lower triangular Cholesky factors of* $\widehat{\Sigma}_x$ *and* $\widehat{\Sigma}_w$*, respectively. Thus, we have* $\widehat{\Sigma}_x = \Lambda_x \Lambda_x^\top$ *and* $\widehat{\Sigma}_w = \Lambda_w \Lambda_w^\top$*. We emphasize that* Λ_x *and* Λ_w *can be computed reliably in high dimensions.*

2.3 The Dual Wasserstein MMSE Estimation Problem over Normal Priors

We now examine the dual Wasserstein MMSE estimation problem

$$\underset{\mathbb{Q} \in \mathbb{B}(\widehat{\mathbb{P}})}{\text{maximize}} \inf_{\psi \in \mathcal{F}} \mathcal{R}(\psi, \mathbb{Q}), \tag{2.22}$$

which is obtained from (2.7) by interchanging the order of minimization and maximization. Any maximizer $\mathbb{Q}^\star$ of this dual estimation problem, if it exists, will henceforth be called a *least favorable prior.* Unfortunately, problem (2.22) is generically intractable. Below we will demonstrate, however, that (2.22) becomes tractable if the nominal distribution $\widehat{\mathbb{P}}$ is normal.

Definition 2.11 (Normal distributions)**.** *We say that* $\mathbb{P}$ *is a normal distribution on* $\mathbb{R}^d$ *with mean* $\mu \in \mathbb{R}^d$ *and covariance matrix* $\Sigma \in \mathbb{S}_+^d$*, that is,* $\mathbb{P} = \mathcal{N}(\mu, \Sigma)$*, if* $\mathbb{P}$ *is supported on* $\text{supp}(\mathbb{P}) = \{\mu + Ev : v \in \mathbb{R}^k\}$*, and if the density function of* $\mathbb{P}$ *with respect to the Lebesgue measure on* $\text{supp}(\mathbb{P})$ *is given by*

$$\varrho_{\mathbb{P}}(\xi) \triangleq \frac{1}{\sqrt{(2\pi)^k \det(D)}} e^{-(\xi-\mu)^\top E D^{-1} E^\top (\xi - \mu)},$$

where $k = \text{rank}(\Sigma)$*,* $D \in \mathbb{S}_{++}^k$ *is the diagonal matrix of the positive eigenvalues of* Σ*, and* $E \in \mathbb{R}^{d \times k}$ *is the matrix whose columns correspond to the orthonormal eigenvectors of the positive eigenvalues of* Σ*.*

Definition 2.11 also accounts for degenerate normal distributions with singular covariance matrices. We now recall some basic properties of normal distributions that are crucial for the results of this paper.

Proposition 2.12 (Affine transformations [59, Theorem 2.16])**.** *If* $\xi \in \mathbb{R}^d$ *follows the normal distribution* $\mathcal{N}(\mu, \Sigma)$*, while* $A \in \mathbb{R}^{k \times d}$ *and* $b \in \mathbb{R}^k$*, then* $A\xi + b \in \mathbb{R}^k$ *follows the normal distribution* $\mathcal{N}(A\mu + b, A\Sigma A^\top)$*.*

Proposition 2.13 (Affine conditional expectations [22, Corollary 5]). *Assume that $\xi \in \mathbb{R}^d$ follows the normal distribution $\mathbb{P} = \mathcal{N}(\mu, \Sigma)$ and that*

$$\xi = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}, \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \text{and} \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

where $\xi_1, \mu_1 \in \mathbb{R}^{d_1}$, $\xi_2, \mu_2 \in \mathbb{R}^{d_2}$, $\Sigma_{11} \in \mathbb{R}^{d_1 \times d_1}$, $\Sigma_{22} \in \mathbb{R}^{d_2 \times d_2}$ and $\Sigma_{12} = \Sigma_{21}^\top \in \mathbb{R}^{d_1 \times d_2}$ for some $d_1, d_2 \in \mathbb{N}$ with $d_1 + d_2 = d$. Then, there exist $A \in \mathbb{R}^{d_1 \times d_2}$ and $b \in \mathbb{R}^{d_1}$ such that $\mathbb{E}_{\mathbb{P}}[\xi_1 | \xi_2] = A\xi_2 + b$ $\mathbb{P}$-almost surely.

Another useful but lesser known property of normal distributions is that their Wasserstein distances can be expressed analytically in terms of the distributions' first- and second-order moments.

Proposition 2.14 (Wasserstein distance between normal distributions [72, Proposition 7]). *The Wasserstein distance between two normal distributions $\mathbb{Q}_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $\mathbb{Q}_2 = \mathcal{N}(\mu_2, \Sigma_2)$ equals the Gelbrich distance between their mean vectors and covariance matrices, that is, $\mathbb{W}(\mathbb{Q}_1, \mathbb{Q}_2) = \mathbb{G}((\mu_1, \Sigma_1), (\mu_2, \Sigma_2))$.*

Assume now that the nominal distributions of the parameter $x \in \mathbb{R}^n$ and the noise $w \in \mathbb{R}^m$ are normal, that is, assume that $\widehat{\mathbb{P}}_x = \mathcal{N}(\widehat{\mu}_x, \widehat{\Sigma}_x)$ and $\widehat{\mathbb{P}}_w = \mathcal{N}(\widehat{\mu}_w, \widehat{\Sigma}_w)$. Thus, the joint nominal distribution $\widehat{\mathbb{P}} = \widehat{\mathbb{P}}_x \times \widehat{\mathbb{P}}_w$ is also normal, that is,

$$\widehat{\mathbb{P}} = \mathcal{N}(\widehat{\mu}, \widehat{\Sigma}) \qquad \text{where} \qquad \widehat{\mu} = \begin{bmatrix} \widehat{\mu}_x \\ \widehat{\mu}_w \end{bmatrix} \qquad \text{and} \qquad \widehat{\Sigma} = \begin{bmatrix} \widehat{\Sigma}_x & 0 \\ 0 & \widehat{\Sigma}_w \end{bmatrix}. \tag{2.23}$$

We highlight that normal distributions are natural candidates for $\widehat{\mathbb{P}}$. One reason for this is that the normal distribution has maximum entropy among all distributions with prescribed first- and second-order moments [32, § 12]. Therefore, it has appeal as the least prejudiced baseline model. Similarly, if the parameter x in (2.1) is normally distributed, then a normal distribution minimizes the mutual information between x and the observation y among all noise distributions with bounded variance [44, Lemma II.2]. In this sense, normally distributed noise renders the observations least informative. Conversely, if the noise in (2.1) is normally distributed, then a normal distribution maximizes the MMSE across all distributions of x with bounded variance [77, Proposition 15]. In this sense, normally distributed parameters are the hardest to estimate. Using normal nominal distributions thus amounts to adopting a worst-case perspective.

Armed with the fundamental results on normal distributions summarized above, we are now ready to address the dual Wasserstein MMSE estimation problem (2.22) with a normal nominal distribution. In analogy to Section 2.2, where we proposed the Gelbrich MMSE estimation problem as an easier conservative approximation for the original *primal* estimation problem (2.7), we will now construct an easier conservative approximation for the original *dual* estimation problem (2.22). To this end, we define the restricted ambiguity set

$$\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}}) \triangleq \left\{ \mathbb{Q}_x \times \mathbb{Q}_w \in \mathcal{M}(\mathbb{R}^n) \times \mathcal{M}(\mathbb{R}^m) : \begin{array}{l} \exists \Sigma_x \in \mathbb{S}^n_+,\ \Sigma_w \in \mathbb{S}^m_{++} \text{ with} \\ \mathbb{Q}_x = \mathcal{N}(\widehat{\mu}_x, \Sigma_x),\ \mathbb{Q}_w = \mathcal{N}(\widehat{\mu}_w, \Sigma_w), \\ \mathbb{W}(\mathbb{Q}_x, \widehat{\mathbb{P}}_x) \le \rho_x,\ \mathbb{W}(\mathbb{Q}_w, \widehat{\mathbb{P}}_w) \le \rho_w \end{array} \right\}.$$

By construction, $\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})$ contains all *normal* distributions $\mathbb{Q} = \mathbb{Q}_x \times \mathbb{Q}_w$ from within the original Wasserstein ambiguity set $\mathbb{B}(\widehat{\mathbb{P}})$ that have the same mean vector $(\widehat{\mu}_x, \widehat{\mu}_w)$ as the nominal distribution $\widehat{\mathbb{P}} = \widehat{\mathbb{P}}_x \times \widehat{\mathbb{P}}_w$, and where the covariance matrix of $\mathbb{Q}_w$ is strictly positive definite. Thus, we have $\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}}) \subseteq \mathbb{B}(\widehat{\mathbb{P}})$. Note also that $\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})$ is non-convex because mixtures of normal distributions usually fail to be normal.

By restricting the original Wasserstein ambiguity set $\mathbb{B}(\widehat{\mathbb{P}})$ to its subset $\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})$, we obtain the following conservative approximation for the dual Wasserstein MMSE estimation problem (2.22).

$$\underset{\mathbb{Q} \in \mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})}{\text{maximize}} \inf_{\psi \in \mathcal{F}} \mathcal{R}(\psi, \mathbb{Q}) \tag{2.24}$$

We will henceforth refer to (2.24) as the dual Wasserstein MMSE estimation problem *over normal priors*. The following main theorem shows that (2.24) is equivalent to a finite convex optimization problem.

Theorem 2.15 (Dual Wasserstein MMSE estimation problem over normal priors)**.** *Assume that the Wasserstein ambiguity set $\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})$ is centered at a normal distribution $\widehat{\mathbb{P}}$ of the form* (2.23)*. Then, the dual Wasserstein MMSE estimation problem over normal priors* (2.24) *is equivalent to the finite convex optimization problem*

$$\begin{array}{ll} \sup & \operatorname{Tr}\left[\Sigma_x - \Sigma_x H^\top \left(H\Sigma_x H^\top + \Sigma_w\right)^{-1} H\Sigma_x\right] \\ \text{s.t.} & \Sigma_x \in \mathbb{S}^n_+, \quad \Sigma_w \in \mathbb{S}^m_{++} \\ & \operatorname{Tr}\left[\Sigma_x + \widehat{\Sigma}_x - 2\left(\widehat{\Sigma}_x^{\frac{1}{2}} \Sigma_x \widehat{\Sigma}_x^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \le \rho_x^2 \\ & \operatorname{Tr}\left[\Sigma_w + \widehat{\Sigma}_w - 2\left(\widehat{\Sigma}_w^{\frac{1}{2}} \Sigma_w \widehat{\Sigma}_w^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \le \rho_w^2. \end{array} \tag{2.25}$$

If $\widehat{\Sigma}_w \succ 0$, then (2.25) *is solvable, and the maximizer denoted by $(\Sigma_x^\star, \Sigma_w^\star)$ satisfies $\Sigma_x^\star \succeq \lambda_{\min}(\widehat{\Sigma}_x) I_n$ and $\Sigma_w^\star \succeq \lambda_{\min}(\widehat{\Sigma}_w) I_m$. Moreover, the supremum of* (2.24) *is attained by the normal distribution $\mathbb{Q}^\star = \mathbb{Q}_x^\star \times \mathbb{Q}_w^\star$ defined through $\mathbb{Q}_x^\star = \mathcal{N}(\widehat{\mu}_x, \Sigma_x^\star)$ and $\mathbb{Q}_w^\star = \mathcal{N}(\widehat{\mu}_w, \Sigma_w^\star)$.*

Proof. If (x, w) is governed by a normal distribution $\mathbb{Q} \in \mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})$, then the linear transformation $(x, y) = (x, Hx + w)$ is also normally distributed by virtue of Proposition 2.12, and the average risk $\mathcal{R}(\psi, \mathbb{Q})$ is minimized by the Bayesian MMSE estimator $\psi^\star_{\mathcal{B}}(y) = \mathbb{E}_{\mathbb{P}_{x|y}}[x]$, which is affine due to Proposition 2.13. Thus, in the dual Wasserstein MMSE estimation problem with normal priors, the set $\mathcal{F}$ of *all* estimators may be restricted to the set $\mathcal{A}$ of all *affine* estimators without sacrificing optimality, that is,

$$\sup_{\mathbb{Q}\in\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})} \inf_{\psi\in\mathscr{F}} \mathscr{R}(\psi,\mathbb{Q}) = \sup_{\mathbb{Q}\in\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})} \inf_{\psi\in\mathscr{A}} \mathscr{R}(\psi,\mathbb{Q}). \tag{2.26}$$

As the average risk $\mathscr{R}(\psi,\mathbb{Q})$ of an affine estimator $\psi\in\mathscr{A}$ simply evaluates the expectation of a quadratic function in (x,w), it depends on $\mathbb{Q}$ only through its first and second moments. Moreover, as $\mathbb{Q}$ and $\widehat{\mathbb{P}}$ are normal distributions, their Wasserstein distance coincides with the Gelbrich distance between their mean vectors and covariance matrices; see Proposition 2.14. Thus, the maximization problem over $\mathbb{Q}$ on the right hand side of (2.26) can be recast as an equivalent maximization problem over the first and second moments of $\mathbb{Q}$. Specifically, by the definitions of $\mathscr{R}(\psi,\mathbb{Q})$ and $\mathbb{B}_{\phi}(\widehat{\mathbb{P}})$ we find

$$\sup_{\mathbb{Q}\in\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})} \inf_{\psi\in\mathscr{A}} \mathscr{R}(\psi,\mathbb{Q}) = \begin{cases} \displaystyle\sup_{\Sigma_x,\Sigma_w} \inf_{\substack{A,K\\ K=I_n-AH}} \inf_{b} & \left\langle K^\top K, \Sigma_x + \widehat{\mu}_x\widehat{\mu}_x^\top\right\rangle + \left\langle A^\top A, \Sigma_w + \widehat{\mu}_w\widehat{\mu}_w^\top\right\rangle + b^\top b \\ & \quad -2\widehat{\mu}_x^\top K^\top A\widehat{\mu}_w - 2b^\top(K\widehat{\mu}_x - A\widehat{\mu}_w) \\ \text{s.t.} & \mathbb{G}\left((\widehat{\mu}_x,\Sigma_x),(\widehat{\mu}_x,\widehat{\Sigma}_x)\right) \le \rho_x,\ \mathbb{G}\left((\widehat{\mu}_w,\Sigma_w),(\widehat{\mu}_w,\widehat{\Sigma}_w)\right) \le \rho_w \\ & \Sigma_x \succeq 0,\ \Sigma_w \succ 0, \end{cases}$$

where the auxiliary decision variable $K = I_n - AH$ has been introduced to simplify the objective function. The innermost minimization problem over b constitutes an unconstrained (strictly) convex quadratic program that has the unique optimal solution $b = K\widehat{\mu}_x - A\widehat{\mu}_w$. Substituting this minimizer back into the objective function of the above problem and recalling the definition of the Gelbrich distance then yields

$$\sup_{\mathbb{Q}\in\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})} \inf_{\psi\in\mathscr{A}} \mathscr{R}(\psi,\mathbb{Q}) = \begin{cases} \displaystyle\sup_{\Sigma_x,\Sigma_w} & \displaystyle\inf_{\substack{A,K\\ K=I_n-AH}} \quad \left\langle K^\top K, \Sigma_x\right\rangle + \left\langle A^\top A, \Sigma_w\right\rangle \\ \text{s.t.} & \operatorname{Tr}\left[\Sigma_x + \widehat{\Sigma}_x - 2\left(\widehat{\Sigma}_x^{\frac{1}{2}}\Sigma_x\widehat{\Sigma}_x^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \le \rho_x^2 \\ & \operatorname{Tr}\left[\Sigma_w + \widehat{\Sigma}_w - 2\left(\widehat{\Sigma}_w^{\frac{1}{2}}\Sigma_w\widehat{\Sigma}_w^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \le \rho_w^2 \\ & \Sigma_x \succeq 0,\ \Sigma_w \succ 0. \end{cases} \tag{2.27}$$

By using the equality $K = I_n - AH$ to eliminate K, the inner minimization problem in (2.27) can be reformulated as an unconstrained quadratic program in A. As $\Sigma_w \succ 0$, this quadratic program is strictly convex, and an elementary calculation reveals that its unique optimal solution is given by

$$A^\star = \Sigma_x H^\top \left(H\Sigma_x H^\top + \Sigma_w\right)^{-1}.$$

Substituting $A^\star$ as well as the corresponding auxiliary decision variable $K^\star = I_n - A^\star H$ into the objective function of (2.27) finally yields the postulated convex program (2.25).

Assume now that $\widehat{\Sigma}_w \succ 0$, and define

$$\mathscr{S}_x \triangleq \left\{\Sigma_x \in \mathbb{S}_+^n : \mathbb{G}\left((\widehat{\mu}_x,\Sigma_x),(\widehat{\mu}_x,\widehat{\Sigma}_x)\right) \le \rho_x\right\}$$

and

$$\mathscr{S}_w \triangleq \left\{\Sigma_w \in \mathbb{S}_+^m : \mathbb{G}\left((\widehat{\mu}_w, \Sigma_w), (\widehat{\mu}_w, \widehat{\Sigma}_w)\right) \le \rho_w\right\}.$$

Equations (2.26) and (2.27) imply that

$$\begin{aligned}\sup_{\mathbb{Q}\in\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})} \inf_{\psi\in\mathscr{F}} \mathscr{R}(\psi,\mathbb{Q}) &\le \sup_{\Sigma_x\in\mathscr{S}_x} \sup_{\Sigma_w\in\mathscr{S}_w} \inf_{\substack{A,K\\K=I_n-AH}} \left\langle K^\top K, \Sigma_x\right\rangle + \left\langle A^\top A, \Sigma_w\right\rangle \\ &= \sup_{\substack{\Sigma_x\in\mathscr{S}_x\\ \Sigma_x\succeq\lambda_{\min}(\widehat{\Sigma}_x)I_n}} \sup_{\substack{\Sigma_w\in\mathscr{S}_w\\ \Sigma_w\succeq\lambda_{\min}(\widehat{\Sigma}_w)I_m}} \inf_{\substack{A,K\\K=I_n-AH}} \left\langle K^\top K, \Sigma_x\right\rangle + \left\langle A^\top A, \Sigma_w\right\rangle, \end{aligned} \tag{2.28}$$

where the inequality holds because we relax the requirement that Σ_w be strictly positive definite, and the equality follows from applying Lemma 2.35 consecutively to each of the two maximization problems. If $\widehat{\Sigma}_w \succ 0$, then problem (2.28) constitutes a restriction of (2.27) and therefore provides also a lower bound on the dual Wasserstein MMSE estimation problem. In summary, we thus have

$$\sup_{\mathbb{Q}\in\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})} \inf_{\psi\in\mathscr{F}} \mathscr{R}(\psi,\mathbb{Q}) = \begin{cases} \sup\limits_{\Sigma_x,\Sigma_w} & \inf\limits_{\substack{A,K\\K=I_n-AH}} \left\langle K^\top K, \Sigma_x\right\rangle + \left\langle A^\top A, \Sigma_w\right\rangle \\ \text{s.t.} & \operatorname{Tr}\left[\Sigma_x + \widehat{\Sigma}_x - 2\left(\widehat{\Sigma}_x^{\frac{1}{2}}\Sigma_x\widehat{\Sigma}_x^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \le \rho_x^2 \\ & \operatorname{Tr}\left[\Sigma_w + \widehat{\Sigma}_w - 2\left(\widehat{\Sigma}_w^{\frac{1}{2}}\Sigma_w\widehat{\Sigma}_w^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \le \rho_w^2 \\ & \Sigma_x \succeq \lambda_{\min}(\widehat{\Sigma}_x)I_n,\ \Sigma_w \succeq \lambda_{\min}(\widehat{\Sigma}_w)I_m. \end{cases} \tag{2.29}$$

This reasoning implies that if $\widehat{\Sigma}_w \succ 0$, then the constraints $\Sigma_x \succeq \lambda_{\min}(\widehat{\Sigma}_x)I_n$ and $\Sigma_w \succeq \lambda_{\min}(\widehat{\Sigma}_w)I_m$ can be appended to problem (2.27) and, consequently, to problem (2.25) without altering their common optimal value. Problem (2.25) with the additional constraints $\Sigma_x \succeq \lambda_{\min}(\widehat{\Sigma}_x)I_n$ and $\Sigma_w \succeq \lambda_{\min}(\widehat{\Sigma}_w)I_m$ has a continuous objective function over a compact feasible set and is thus solvable. Any of its optimal solutions is also optimal in problem (2.25), which has no redundant constraints. Thus, problem (2.25) is solvable.

It remains to show that $\mathbb{Q}^\star$ as constructed in the theorem statement is optimal in (2.24). The feasibility of $(\Sigma_x^\star, \Sigma_w^\star)$ in (2.25) implies that $\mathbb{Q}^\star \in \mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})$, and thus $\mathbb{Q}^\star$ is feasible in (2.24). Moreover, we have

$$\sup_{\mathbb{Q}\in\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})} \inf_{\psi\in\mathscr{F}} \mathscr{R}(\psi,\mathbb{Q}) \ge \inf_{\psi\in\mathscr{F}} \mathscr{R}(\psi,\mathbb{Q}^\star) = \operatorname{Tr}\left[\Sigma_x^\star - \Sigma_x^\star H^\top \left(H\Sigma_x^\star H^\top + \Sigma_w^\star\right)^{-1} H\Sigma_x^\star\right], \tag{2.30}$$

where the equality follows from elementary algebra, recalling that the affine estimator $\psi(y) = A^\star y + b^\star$ with

$$A^\star = \Sigma_x^\star H^\top \left(H\Sigma_x^\star H^\top + \Sigma_w^\star\right)^{-1} \quad \text{and} \quad b^\star = \mu_x - A^\star(H\widehat{\mu}_x + \widehat{\mu}_w)$$

is the Bayesian MMSE estimator for the normal distribution $\mathbb{Q}^\star$. As the right hand side of (2.30) coincides with the maximum of (2.25) and as problem (2.25) is equivalent to the dual

Wasserstein MMSE estimation problem (2.24) over normal priors, we may thus conclude that the inequality in (2.30) is tight. Thus, we find

$$\sup_{\mathbb{Q}\in\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})} \inf_{\psi\in\mathcal{F}} \mathcal{R}(\psi,\mathbb{Q}) = \inf_{\psi\in\mathcal{F}} \mathcal{R}(\psi,\mathbb{Q}^\star),$$

which in turn implies that $\mathbb{Q}^\star$ is optimal in (2.24). This observation completes the proof. □

Remark 2.16 (Singular covariance matrices)**.** *A nonlinear SDP akin to* (2.25) *has been derived in [152] under the stronger assumption that the covariance matrix of the nominal distribution* $\widehat{\mathbb{P}}$ *is non-degenerate, which implies that* $\widehat{\Sigma}_x \succ 0$ *and* $\widehat{\Sigma}_w \succ 0$. *However, the weaker condition* $\widehat{\Sigma}_w \succ 0$ *is sufficient to ensure that the matrix inversion in the objective function of problem* (2.25) *is well-defined. Therefore, Theorem 2.15 remains valid if the nominal covariance matrix* $\widehat{\Sigma}_x$ *is singular, which occurs in many applications; see Section 2.7.2 for an example. On the other hand, it is common to require that* $\widehat{\Sigma}_w = \sigma^2 I_m$ *for some* $\sigma > 0$*, see, e.g., [23].*

Corollary 2.17 below asserts that the convex program (2.25) admits a canonical linear SDP reformulation. The proof is omitted as it relies on standard Schur complement arguments familiar from the proof of Corollary 2.9.

Corollary 2.17 (SDP reformulation)**.** *Assume that the Wasserstein ambiguity set* $\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})$ *is centered at a normal distribution* $\widehat{\mathbb{P}}$ *of the form* (2.23) *with noise covariance matrix* $\widehat{\Sigma}_w \succ 0$. *Then, the dual Wasserstein MMSE estimation problem* (2.24) *over normal priors is equivalent to the SDP*

$$\begin{array}{ll}
\max & \mathrm{Tr}\left[\Sigma_x\right] - \mathrm{Tr}\left[U\right] \\
\mathrm{s.t.} & \Sigma_x \in \mathbb{S}_+^n,\ \Sigma_w \in \mathbb{S}_+^m,\ V_x \in \mathbb{S}_+^n,\ V_w \in \mathbb{S}_+^m,\ U \in \mathbb{S}_+^n \\
& \begin{bmatrix} \widehat{\Sigma}_x^{\frac{1}{2}} \Sigma_x \widehat{\Sigma}_x^{\frac{1}{2}} & V_x \\ V_x & I_n \end{bmatrix} \succeq 0, \quad \begin{bmatrix} \widehat{\Sigma}_w^{\frac{1}{2}} \Sigma_w \widehat{\Sigma}_w^{\frac{1}{2}} & V_w \\ V_w & I_m \end{bmatrix} \succeq 0 \\
& \mathrm{Tr}\left[\Sigma_x + \widehat{\Sigma}_x - 2V_x\right] \le \rho_x^2, \quad \mathrm{Tr}\left[\Sigma_w + \widehat{\Sigma}_w - 2V_w\right] \le \rho_w^2 \\
& \begin{bmatrix} U & \Sigma_x H^\top \\ H\Sigma_x & H\Sigma_x H^\top + \Sigma_w \end{bmatrix} \succeq 0, \quad \Sigma_x \succeq \lambda_{\min}(\widehat{\Sigma}_x) I_n, \quad \Sigma_w \succeq \lambda_{\min}(\widehat{\Sigma}_w) I_m.
\end{array} \tag{2.31}$$

We emphasize that the lower bounds on Σ_x and Σ_w are redundant but have been made explicit in (2.31).

2.4 Nash Equilibrium and Optimality of Affine Estimators

If $\widehat{\mathbb{P}}$ is a normal distribution of the form (2.23), then we have

$$\inf_{\psi\in\mathcal{A}} \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathcal{R}(\psi,\mathbb{Q}) \ge \inf_{\psi\in\mathcal{F}} \sup_{\mathbb{Q}\in\mathbb{B}(\widehat{\mathbb{P}})} \mathcal{R}(\psi,\mathbb{Q}) \ge \sup_{\mathbb{Q}\in\mathbb{B}(\widehat{\mathbb{P}})} \inf_{\psi\in\mathcal{F}} \mathcal{R}(\psi,\mathbb{Q}) \ge \sup_{\mathbb{Q}\in\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})} \inf_{\psi\in\mathcal{F}} \mathcal{R}(\psi,\mathbb{Q}), \tag{2.32}$$

where the first inequality follows from the inclusions $\mathscr{A} \subseteq \mathscr{F}$ and $\mathbb{B}(\widehat{\mathbb{P}}) \subseteq \mathbb{G}(\widehat{\mathbb{P}})$, the second inequality exploits weak duality, and the last inequality holds due to the inclusion $\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}}) \subseteq \mathbb{B}(\widehat{\mathbb{P}})$. Note that the left-most minimax problem is the Gelbrich MMSE estimation problem (2.10) studied in Section 2.2, and the right-most maximin problem is the dual Wasserstein MMSE estimation problem (2.24) over normal priors studied in Section 2.3. We also highlight that these restricted primal and dual estimation problems sandwich the original Wasserstein estimation problems (2.7) and (2.22), which coincide with the second and third problems in (2.32), respectively. The following theorem asserts that all inequalities in (2.32) actually collapse to equalities.

Theorem 2.18 (Sandwich theorem)**.** *If* $\widehat{\mathbb{P}}$ *is a normal distribution of the form* (2.23)*, then the optimal values of the restricted primal and dual estimation problems* (2.10) *and* (2.24) *coincide, i.e.,*

$$\inf_{\psi\in\mathscr{A}} \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathscr{R}(\psi,\mathbb{Q}) = \sup_{\mathbb{Q}\in\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})} \inf_{\psi\in\mathscr{F}} \mathscr{R}(\psi,\mathbb{Q}).$$

Proof. By Theorem 2.8, the Gelbrich MMSE estimation problem (2.10) can be expressed as

$$\inf_{\psi\in\mathscr{A}} \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathscr{R}(\psi,\mathbb{Q}) = \left\{ \begin{array}{lll} \inf\limits_{\substack{A,K\\ K=I_n-AH}} & \inf\limits_{\substack{\gamma_x,\gamma_w\\ \gamma_x I_n \succ K^\top K\\ \gamma_w I_m \succ A^\top A}} & \gamma_x\left(\rho_x^2 - \mathrm{Tr}\left[\widehat{\Sigma}_x\right]\right) + \gamma_x^2\left\langle (\gamma_x I_n - K^\top K)^{-1}, \widehat{\Sigma}_x \right\rangle \\ & & \quad +\gamma_w\left(\rho_w^2 - \mathrm{Tr}\left[\widehat{\Sigma}_w\right]\right) + \gamma_w^2\left\langle (\gamma_w I_m - A^\top A)^{-1}, \widehat{\Sigma}_w \right\rangle, \end{array} \right.$$

where the auxiliary variable $K = I_n - AH$ has been introduced to highlight the problem's symmetries. Next, we introduce the feasible sets

$$\mathscr{S}_x \triangleq \left\{ \Sigma_x \in \mathbb{S}_+^n : \mathrm{Tr}\left[\Sigma_x + \widehat{\Sigma}_x - 2\left(\widehat{\Sigma}_x^{\frac{1}{2}} \Sigma_x \widehat{\Sigma}_x^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \le \rho_x^2 \right\}$$

and

$$\mathscr{S}_w \triangleq \left\{ \Sigma_w \in \mathbb{S}_+^m : \mathrm{Tr}\left[\Sigma_w + \widehat{\Sigma}_w - 2\left(\widehat{\Sigma}_w^{\frac{1}{2}} \Sigma_w \widehat{\Sigma}_w^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \le \rho_w^2 \right\},$$

both of which are convex and compact by virtue of Lemma 2.36. Using Proposition 2.34 (i) to reformulate the inner minimization problem over γ_x and γ_w, we then obtain

$$\begin{aligned} \inf_{\psi\in\mathscr{A}} \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathscr{R}(\psi,\mathbb{Q}) &= \inf_{\substack{A,K\\ K=I_n-AH}} \sup_{\substack{\Sigma_x\in\mathscr{S}_x\\ \Sigma_w\in\mathscr{S}_w}} \left\langle K^\top K, \Sigma_x \right\rangle + \left\langle A^\top A, \Sigma_w \right\rangle \\ &= \sup_{\Sigma_x\in\mathscr{S}_x} \sup_{\Sigma_w\in\mathscr{S}_w} \inf_{\substack{A,K\\ K=I_n-AH}} \left\langle K^\top K, \Sigma_x \right\rangle + \left\langle A^\top A, \Sigma_w \right\rangle \\ &= \sup_{\substack{\Sigma_x\in\mathscr{S}_x\\ \Sigma_x\succeq\lambda_{\min}(\widehat{\Sigma}_x) I_n}} \sup_{\substack{\Sigma_w\in\mathscr{S}_w\\ \Sigma_w\succeq\lambda_{\min}(\widehat{\Sigma}_w) I_m}} \inf_{\substack{A,K\\ K=I_n-AH}} \left\langle K^\top K, \Sigma_x \right\rangle + \left\langle A^\top A, \Sigma_w \right\rangle \\ &= \sup_{\mathbb{Q}\in\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})} \inf_{\psi\in\mathscr{F}} \mathscr{R}(\psi,\mathbb{Q}), \end{aligned}$$

where the second equality holds due to Sion's minimax theorem [156], and the third equality follows from Lemma 2.35 applied twice separately to Σ_x and Σ_w. The last equality has already been derived in the proof of Theorem 2.15; see Equation (2.29). Thus, the claim follows. □

Theorem 2.18 suggests that solving any of the restricted estimation problems is tantamount to solving both original primal and dual estimation problems. This intuition is formalized in the following corollary.

Corollary 2.19 (Nash equilibrium)**.** *If* $\widehat{\mathbb{P}}$ *is a normal distribution of the form* (2.23) *with* $\widehat{\Sigma}_w \succ 0$, *then the affine estimator* $\psi^\star$ *that solves* (2.10) *is optimal in the primal Wasserstein MMSE estimation problem* (2.7), *while the normal distribution* $\mathbb{Q}^\star$ *that solves* (2.24) *is optimal in the dual Wasserstein MMSE estimation problem* (2.22). *Moreover,* $\psi^\star$ *and* $\mathbb{Q}^\star$ *form a Nash equilibrium for the game between the statistician and nature, that is,*

$$\mathscr{R}(\psi^\star,\mathbb{Q}) \le \mathscr{R}(\psi^\star,\mathbb{Q}^\star) \le \mathscr{R}(\psi,\mathbb{Q}^\star) \quad \forall \psi \in \mathscr{F},\ \mathbb{Q} \in \mathbb{B}(\widehat{\mathbb{P}}). \tag{2.33}$$

Proof. As $\widehat{\Sigma}_w \succ 0$, the Gelbrich MMSE estimation problem (2.10) is solved by the affine estimator $\psi^\star$ defined in Theorem 2.8, and the dual Wasserstein MMSE estimation problem (2.22) over normal priors is solved by the normal distribution $\mathbb{Q}^\star$ defined in Theorem 2.15. Thus, we have

$$\begin{aligned}\mathscr{R}(\psi^\star,\mathbb{Q}^\star) \ge \inf_{\psi\in\mathscr{F}} \mathscr{R}(\psi,\mathbb{Q}^\star) &= \max_{\mathbb{Q}\in\mathbb{B}_{\mathscr{N}}(\widehat{\mathbb{P}})} \inf_{\psi\in\mathscr{F}} \mathscr{R}(\psi,\mathbb{Q}) \\ &= \min_{\psi\in\mathscr{A}} \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathscr{R}(\psi,\mathbb{Q}) = \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathscr{R}(\psi^\star,\mathbb{Q}) \ge \mathscr{R}(\psi^\star,\mathbb{Q}^\star),\end{aligned}$$

where the three equalities follow from the definition of $\mathbb{Q}^\star$, Theorem 2.18 and the definition of $\psi^\star$, respectively. As the left and the right hand sides of the above expression coincide, we may then conclude that

$$\mathscr{R}(\psi^\star,\mathbb{Q}^\star) = \mathscr{R}(\psi^\star,\mathbb{Q}) \le \mathscr{R}(\psi^\star,\mathbb{Q}^\star) \le \mathscr{R}(\psi,\mathbb{Q}^\star) \quad \forall \psi \in \mathscr{F},\ \mathbb{Q} \in \mathbb{G}(\widehat{\mathbb{P}}).$$

Moreover, as $\mathbb{B}(\widehat{\mathbb{P}}) \subseteq \mathbb{G}(\widehat{\mathbb{P}})$, the above relation implies (2.33).

It remains to be shown that $\psi^\star$ and $\mathbb{Q}^\star$ solve the primal and dual Wasserstein MMSE estimation problems (2.7) and (2.22), respectively. As for $\psi^\star$, we have

$$\sup_{\mathbb{Q}\in\mathbb{B}(\widehat{\mathbb{P}})} \mathscr{R}(\psi^\star,\mathbb{Q}) \le \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathscr{R}(\psi^\star,\mathbb{Q}) = \inf_{\psi\in\mathscr{A}} \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathscr{R}(\psi,\mathbb{Q}) = \inf_{\psi\in\mathscr{F}} \sup_{\mathbb{Q}\in\mathbb{B}(\widehat{\mathbb{P}})} \mathscr{R}(\psi,\mathbb{Q}),$$

where the first equality follows from the definition of $\psi^\star$, while the second equality exploits Theorem 2.18, which implies that all inequalities in (2.32) are in fact equalities. This reasoning shows that $\psi^\star$ is optimal in (2.7). The optimality of $\mathbb{Q}^\star$ in (2.22) can be proved similarly. Details are omitted for brevity. □

Corollary 2.19 implies that $\psi^\star$ can be viewed as a Bayesian estimator for the least favorable prior $\mathbb{Q}^\star$ and that $\mathbb{Q}^\star$ represents a worst-case distribution for the optimal estimator $\psi^\star$.

Next, we will argue that $\psi^\star$ can not only be constructed from the solution of the convex program (2.12), which is equivalent to the Gelbrich MMSE estimation problem (2.10), but also from the solution of the convex program (2.25), which is equivalent to the dual MMSE estimation problem (2.24) over normal priors. This alternative construction is useful because problem (2.25) is amenable to highly efficient first-order methods to be derived in Section 2.6.

Corollary 2.20 (Dual construction of the optimal estimator). *If $\widehat{\mathbb{P}}$ is a normal distribution of the form* (2.23) *with $\widehat{\Sigma}_w \succ 0$, and $(\Sigma_x^\star, \Sigma_w^\star)$ is a maximizer of* (2.24), *then the affine estimator $\psi^\star(y) = A^\star y + b^\star$ with*

$$A^\star = \Sigma_x^\star H^\top \left(H\Sigma_x^\star H^\top + \Sigma_w^\star\right)^{-1} \quad \textit{and} \quad b^\star = \widehat{\mu}_x - A^\star(H\widehat{\mu}_x + \widehat{\mu}_w) \tag{2.34}$$

solves the Wasserstein MMSE estimation problem (2.7).

Proof. Define $\psi^\star$ as the affine estimator that solves (2.10) and $\mathbb{Q}^\star$ as the normal distribution that solves (2.24). By Corollary 2.19, the second inequality in (2.33) holds for all admissible estimators $\psi \in \mathcal{F}$, which implies that $\psi^\star \in \arg\min_{\psi\in\mathcal{F}} \mathcal{R}(\psi, \mathbb{Q}^\star)$, that is, $\psi^\star$ solves the Bayesian MMSE estimation problem corresponding to $\mathbb{Q}^\star$. As any Bayesian MMSE estimator satisfies $\psi^\star(y) = \mathbb{E}_{\mathbb{Q}^\star_{x|y}}[x]$ for $\mathbb{Q}^\star$-almost all y and as $\Sigma_w^\star \succ 0$, we may use the known formulas for conditional normal distributions to conclude that the unique affine Bayesian MMSE estimator for $\mathbb{Q}^\star$ is of the form $\psi^\star(y) = A^\star y + b^\star$ with parameters defined as in (2.34). □

Remark 2.21 (Non-normal nominal distributions). *The results of this section imply that the optimal values of the finite convex programs* (2.12) *and* (2.25) *typically coincide even if the nominal distribution fails to be normal. To see this, assume that $\widehat{\mathbb{P}} = \widehat{\mathbb{P}}_x \times \widehat{\mathbb{P}}_w$, where $\widehat{\mathbb{P}}_x$ and $\widehat{\mathbb{P}}_w$ are arbitrary signal and noise distributions with finite mean vectors $\widehat{\mu}_x$ and $\widehat{\mu}_w$ and covariance matrices $\widehat{\Sigma}_x$ and $\widehat{\Sigma}_w$, respectively, and denote by $\widehat{\mathbb{P}}' = \widehat{\mathbb{P}}'_x \times \widehat{\mathbb{P}}'_w$ the normal distribution with the same first and second moments as $\widehat{\mathbb{P}}$. If $\widehat{\Sigma}_w \succ 0$, then the optimal values of* (2.12) *and* (2.25) *are equal by virtue of the Theorems 2.8, 2.15 and 2.18 applied to $\widehat{\mathbb{P}}'$. As* (2.12) *and* (2.25) *depend only on the first and second moments of $\widehat{\mathbb{P}}'$, their optimal values do not change if $\widehat{\mathbb{P}}'$ is replaced with $\widehat{\mathbb{P}}$. Therefore, the optimal values of* (2.12) *and* (2.25) *coincide for any nominal distribution $\widehat{\mathbb{P}}$ with finite first and second moments provided that $\widehat{\Sigma}_w \succ 0$. In this case, however, the minimum of the Gelbrich MMSE estimation problem* (2.10) *may strictly exceed the maximum of the dual Wasserstein MMSE estimation problem* (2.24) *over normal priors. Note that in this case the ambiguity set $\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})$ may even be empty. Moreover, while typically suboptimal for the original Wasserstein MMSE estimation problem* (2.7), *the affine estimator constructed in Corollary 2.20 remains optimal for the Gelbrich MMSE estimation problem* (2.10) *even if $\widehat{\mathbb{P}}$ fails to be normal.*

2.5 Elliptical Nominal Distributions

We will now show that the results of Sections 2.2–2.4 remain valid if $\widehat{\mathbb{P}}$ is an arbitrary elliptical (but maybe non-normal) distribution. To this end, we first review some basic results on elliptical distributions.

Definition 2.22 (Elliptical distributions). *The distribution $\mathbb{P}$ of $\xi \in \mathbb{R}^d$ is called elliptical if the characteristic function $\Phi_{\mathbb{P}}(t) \triangleq \mathbb{E}_{\mathbb{P}}[\exp(it^\top \xi)]$ of $\mathbb{P}$ is given by $\Phi_{\mathbb{P}}(t) = \exp(it^\top \mu)\phi(t^\top St)$ for some location parameter $\mu \in \mathbb{R}^d$, dispersion matrix $S \in \mathbb{S}_+^d$ and characteristic generator $\phi : \mathbb{R}_+ \to \mathbb{R}$. In this case we write $\mathbb{P} = \mathscr{E}_\phi^d(\mu, S)$. The class of all d-dimensional elliptical distributions with characteristic generator ϕ is denoted by $\mathscr{E}_\phi^d$.*

The class of elliptical distributions was introduced in [98] with the aim to generalize the family of normal distributions, which are obtained by setting the characteristic generator to $\phi(u) = e^{-u/2}$. We emphasize that, unlike the moment-generating function $M_{\mathbb{P}}(t) \triangleq \mathbb{E}_{\mathbb{P}}[\exp(t^\top \xi)]$, the characteristic function $\Phi_{\mathbb{P}}(t)$ is always finite for all $t \in \mathbb{R}^d$ even if some moments of $\mathbb{P}$ do not exist. Thus, Definition 2.22 is general enough to cover also heavy-tailed distributions with non-zero tail dependence coefficients [85]. Examples of elliptical distributions include the Laplace, logistic and t-distribution etc. Useful theoretical properties of elliptical distributions are discussed in [22, 59]. We also highlight that elliptical distributions are central to a wide spectrum of diverse applications ranging from genomics [141] and medical imaging [147] to finance [92, § 6.2.1], to name a few.

If the dispersion matrix $S \in \mathbb{S}_+^d$ has rank k, then there exists $\Lambda \in \mathbb{R}^{d\times k}$ with $S = \Lambda\Lambda^\top$, and there exists a generalized inverse $\Lambda^{-1} \in \mathbb{R}^{k\times d}$ with $\Lambda^{-1}\Lambda = I_k = \Lambda^\top(\Lambda^{-1})^\top$. One easily verifies that if $\xi \in \mathbb{R}^d$ follows an elliptical distribution $\mathbb{P} = \mathscr{E}_\phi^d(\mu, S)$, then $\tilde{\xi} \triangleq \Lambda^{-1}(\xi - \mu) \in \mathbb{R}^k$ follows the spherically symmetric distribution $\tilde{\mathbb{P}} = \mathscr{E}_\phi^d(0, I_k)$ with characteristic function $\Phi_{\tilde{\mathbb{P}}}(t) = \phi(\|t\|^2)$. Thus, the choice of the characteristic generator ϕ is constrained by the implicit condition that $\phi(\|t\|^2)$ must be an admissible characteristic function. For instance, the normalization of probability distributions necessitates that $\phi(0) = 1$, while the dominated convergence theorem implies that ϕ must be continuous etc. As any distribution is uniquely determined by its characteristic function, and as $\phi(\|t\|^2)$ depends only on the norm of t, the spherical distribution $\tilde{\mathbb{P}}$ is indeed invariant under rotations. This implies that $\mathbb{E}_{\tilde{\mathbb{P}}}[\tilde{\xi}] = 0$ and, via the linearity of the expectation, that $\mathbb{E}_{\mathbb{P}}[\xi] = \mu$ provided that $\tilde{\xi}$ and ξ are integrable, respectively. Thus, the location parameter μ of an elliptical distribution coincides with its mean vector whenever the mean exists. By the definition of the characteristic function, the covariance matrix of $\tilde{\mathbb{P}}$, if it exists, can be expressed as

$$\tilde{\Sigma} = -\nabla_t^2 \Phi_{\tilde{\mathbb{P}}}(t)\big|_{t=0} = -\nabla_t^2 \phi(\|t\|^2)\big|_{t=0} = -2\phi'(0) I_k,$$

where $\phi'(0)$ denotes the right derivative of $\phi(u)$ at $u = 0$. Hence, $\tilde{\Sigma}$ exists if and only if $\phi'(0)$ exists and is finite. Similarly, the covariance matrix of $\mathbb{P}$ is given by $\Sigma = -2\phi'(0)S$, if it exists [22, Theorem 4]. Below we will focus on elliptical distributions with finite first- and second-order moments (*i.e.*, we will only consider characteristic generators with $|\phi'(0)| < \infty$), and we will assume that $\phi'(0) = -\frac{1}{2}$, which ensures that the dispersion matrix S equals the covariance matrix Σ. The latter assumption does not restrict generality. In fact, changing the characteristic generator to $\phi(\frac{-u}{2\phi'(0)})$ and the dispersion matrix to $-2\phi'(0)S$ has no impact on the elliptical distribution $\mathbb{P}$ but matches the dispersion matrix S with the covariance matrix Σ.

The elliptical distributions inherit many desirable properties from the normal distributions but are substantially more expressive as they include also heavy- and light-tailed distributions. For example, any class of elliptical distributions with a common characteristic generator is closed under affine transformations and affine conditional expectations. Moreover, the Wasserstein distance between two elliptical distributions with the same characteristic generator equals the Gelbrich distance between their mean vectors and covariance matrices [71, Theorem 2.4]. Thus, the Propositions 2.12, 2.13 and 2.14 readily extend from the class of normal distributions to *any* class of elliptical distributions that share the same characteristic generator.

The above discussion suggests that the results of Sections 2.2–2.4 carry over almost verbatim to MMSE estimation problems involving elliptical nominal distributions. In the following we will therefore assume that

$$\widehat{\mathbb{P}} = \mathscr{E}_{\phi}^{n+m}(\widehat{\mu}, \widehat{\Sigma}) \qquad \text{with} \qquad \widehat{\mu} = \begin{bmatrix} \widehat{\mu}_x \\ \widehat{\mu}_w \end{bmatrix} \qquad \text{and} \qquad \widehat{\Sigma} = \begin{bmatrix} \widehat{\Sigma}_x & 0 \\ 0 & \widehat{\Sigma}_w \end{bmatrix}, \tag{2.35}$$

where ϕ denotes a prescribed characteristic generator. As the class of all elliptical distributions with characteristic generator ϕ is closed under affine transformations, the marginal distributions $\widehat{\mathbb{P}}_x$ and $\widehat{\mathbb{P}}_w$ of x and w under $\widehat{\mathbb{P}}$ are also elliptical distributions with the same characteristic generator ϕ.

Note that while the signal x and the noise w are uncorrelated under $\widehat{\mathbb{P}}$ irrespective of ϕ, they fail to be independent unless $\widehat{\mathbb{P}}$ is a normal distribution. When working with generic elliptical nominal distributions, we must therefore abandon any independence assumptions. Otherwise, the ambiguity set would be empty for small radii ρ_x and ρ_w. This insight prompts us to redefine the Wasserstein ambiguity set as

$$\mathbb{B}(\widehat{\mathbb{P}}) \triangleq \left\{ \mathbb{Q} \in \mathscr{M}(\mathbb{R}^{n+m}) : \mathbb{E}_{\mathbb{Q}}[xw^\top] = \mathbb{E}_{\mathbb{Q}}[x] \cdot \mathbb{E}_{\mathbb{Q}}[w]^\top,\ \mathbb{W}(\mathbb{Q}_x, \widehat{\mathbb{P}}_x) \le \rho_x,\ \mathbb{W}(\mathbb{Q}_w, \widehat{\mathbb{P}}_w) \le \rho_w \right\}, \tag{2.36}$$

which relaxes the independence condition in (2.8) and merely requires x and w to be uncorrelated. When using the new ambiguity set (2.36) to model the distributional uncertainty, we can again compute a Nash equilibrium between the statistician and nature by solving a tractable convex optimization problem.

Theorem 2.23 (Elliptical distributions)**.** *Assume that* $\widehat{\mathbb{P}}$ *is an elliptical distribution of the form* (2.35) *with characterisic generator* ϕ *and noise covariance matrix* $\widehat{\Sigma}_w \succ 0$*, and define the ambiguity set* $\mathbb{B}(\widehat{\mathbb{P}})$ *as in* (2.36)*. If* $(\Sigma_x^\star, \Sigma_w^\star)$ *solves the finite convex program* (2.24)*, then the affine estimator* $\psi^\star(y) = A^\star y + b^\star$ *with*

$$A^\star = \Sigma_x^\star H^\top \left(H \Sigma_x^\star H^\top + \Sigma_w^\star\right)^{-1} \quad \textit{and} \quad b^\star = \widehat{\mu}_x - A^\star (H\widehat{\mu}_x + \widehat{\mu}_w)$$

solves the Wasserstein MMSE estimation problem (2.7)*, while the elliptical distribution*

$$\mathbb{Q}^\star = \mathscr{E}_{\phi}^{n+m}(\widehat{\mu}, \Sigma^\star) \quad \textit{with} \quad \Sigma^\star = \begin{bmatrix} \Sigma_x^\star & 0 \\ 0 & \Sigma_{w^\star} \end{bmatrix}$$

solves the dual Wasserstein MMSE estimation problem (2.22). *Moreover, $\psi^\star$ and $\mathbb{Q}^\star$ form a Nash equilibrium for the game between the statistician and nature, that is,*

$$\mathscr{R}(\psi^\star,\mathbb{Q}) \le \mathscr{R}(\psi^\star,\mathbb{Q}^\star) \le \mathscr{R}(\psi,\mathbb{Q}^\star) \quad \forall \psi \in \mathscr{F},\ \mathbb{Q} \in \mathbb{B}(\widehat{\mathbb{P}}).$$

Proof. The proof replicates the arguments used to establish Theorems 2.8, 2.15 and 2.18 as well as Corollary 2.19 with obvious minor modifications. Details are omitted for brevity. □

Theorem 2.23 asserts that the optimal estimator depends only on the first and second moments of the nominal distribution $\widehat{\mathbb{P}}$ but *not* on its characteristic generator. Whether $\widehat{\mathbb{P}}$ displays heavier or lighter tails than a normal distribution has therefore no impact on the prediction of the signal. Note, however, that the characteristic generator of $\widehat{\mathbb{P}}$ determines the shape of the least favorable prior.

2.6 Numerical Solution of Wasserstein MMSE Estimation Problems

In this section, we first briefly review the basic setting of the Frank-Wolfe algorithm following with three different stepsize rules proposed in the literature. We then show that under additional regularity assumptions the proposed rule indeed enjoys a linear convergence rate. We further show that the dual Wasserstein estimation reformulation in (2.25) meets these requirements.

2.6.1 Frank-Wolfe Algorithm for Generic Convex Optimization Problems

Consider a generic convex minimization problem of the form

$$f^\star \triangleq \min_{s\in\mathscr{S}} f(s) \tag{2.37}$$

with a convex compact feasible set $\mathscr{S} \subseteq \mathbb{R}^d$ and a convex differentiable objective function $f : \mathscr{S} \to \mathbb{R}$. We assume that for each precision $\delta \in [0,1]$ we have access to an inexact oracle $F : \mathscr{S} \to \mathscr{S}$ that maps any $s \in \mathscr{S}$ to a δ-approximate solution of an auxiliary problem linearized around s. More precisely, we assume that

$$(F(s)-s)^\top \nabla f(s) \le \delta \min_{z\in\mathscr{S}} (z-s)^\top \nabla f(s). \tag{2.38}$$

Note that the minimum on the right hand side of (2.38) vanishes if and only if s solves the original problem (2.37). Otherwise, the minimum is strictly negative. If $\delta = 1$, then the oracle returns an exact mininizer of the linearized problem. If $\delta = 0$, on the other hand, then the oracle returns any solution that is weakly preferred to s in the linearized problem. Given an oracle satisfying (2.38), one can design a Frank-Wolfe algorithm whose iterates obey the recursion

$$s_{k+1} = s_k + \eta_k(F(s_k) - s_k) \quad \forall k \in \mathbb{N} \cup \{0\}, \tag{2.39}$$

where $s_0 \in \mathscr{S}$ is an arbitrary initial feasible solution, δ is a prescribed precision, and $\eta_k \in [0,1]$ is a stepsize that may depend on the current iterate s_k. The Frank-Wolfe algorithm was originally developed for quadratic programs [63] and later extended to general convex programs with differentiable objective functions and compact convex feasible sets [107, 41, 51, 49, 50]. Convergence guarantees for the Frank-Wolfe algorithm typically rely on the assumption that the gradient of f is Lipschitz continuous [107, 49, 50, 70, 64], that f has a bounded curvature constant [29, 89], or that the gradient of f is Hölder continuous [122].

Our convergence analysis will rely on the following regularity conditions.

Assumption 2.24 (Regularity conditions)**.**

(i) The objective function f is β-smooth for some $\beta > 0$, i.e.,

$$\|\nabla f(s_1) - \nabla f(s_2)\| \leq \beta \|s_1 - s_2\| \quad \forall s_1, s_2 \in \mathscr{S}.$$

(ii) The feasible set $\mathscr{S}$ is α-strongly convex for some $\alpha > 0$, i.e.,

$$\theta s_1 + (1-\theta)s_2 - \theta(1-\theta)\frac{\alpha}{2}\|s_1 - s_2\|^2 \frac{\nabla f(s_1)}{\|\nabla f(s_1)\|} \in \mathscr{S} \quad \forall s_1, s_2 \in \mathscr{S},\ \theta \in [0,1].$$

(iii) The objective function f is ε-steep for some $\varepsilon > 0$, i.e.,

$$\|\nabla f(s)\| \geq \varepsilon \quad \forall s \in \mathscr{S}.$$

Assumption 2.24 (ii) relaxes the standard strong convexity condition prevailing in the literature, which requires that the condition used here remains valid if the normalized gradient $\nabla f(s_1)/\|\nabla f(s_1)\|$ is replaced with any other vector in the Euclidean unit ball, see, *e.g.*, [94, Equation (25)]. We emphasize that our weaker condition does not invalidate the standard convergence proofs for the Frank-Wolfe algorithm but is necessary for our purposes because the feasible set of (2.25) fails to be strongly convex in the traditional sense.

In the following we will distinguish three variants of the Frank-Wolfe algorithm with different stepsize rules. The *vanilla Frank-Wolfe* algorithm employs the harmonically decaying static stepsize

$$\eta_k = \frac{2}{2+k},$$

which results in a sublinear $\mathcal{O}(1/k)$ convergence whenever Assumption 2.24 (i) holds [63, 51].

The *adaptive Frank-Wolfe* algorithm uses the stepsize

$$\eta_k = \min\left\{1, \frac{(s_k - F(s_k))^\top \nabla f(s_k)}{\beta \| s_k - F(s_k)\|^2}\right\}, \tag{2.40}$$

which adapts to the iterate s_k. If all of the Assumptions 2.24 (i)–(iii) hold, then the adaptive Frank-Wolfe algorithm enjoys a linear $\mathcal{O}(\theta^k)$ convergence guarantee, where $\theta \in (0,1)$ is an explicit function of the oracle precision δ, the strong convexity parameter α, the smoothness parameter β and the steepness parameter ε [107]. Note that the stepsize (2.40) is constructed as the unique solution of the univariate quadratic program

$$\min_{\eta \in [0,1]} \; f(s_k) - \eta (s_k - F(s_k))^\top \nabla f(s_k) + \frac{1}{2} \beta \eta^2 \, \| s_k - F(s_k)\|^2 ,$$

which minimizes a quadratic majorant of the objective function f along the line segment from s_k to $F(s_k)$.

The adaptive stepsize rule (2.40) has undergone further scrutiny in [70], where it was discovered that one may improve the algorithm's convergence behavior by replacing the global smoothness parameter β in (2.40) with an adaptive smoothness parameter β_k that captures the smoothness of f along the line segment from s_k to $F(s_k)$. This extra flexibility is useful because β_k can be chosen smaller than the unnecessarily conservative global smoothness parameter β and because β_k is easier to estimate than β, which may not even be accessible.

Following [133], we will henceforth only require that $\beta_k > 0$ satisfies the inequality

$$f\Big(s_k - \eta_k(\beta_k)\big(s_k - F(s_k)\big)\Big) \leq f(s_k) - \eta_k(\beta_k)\big(s_k - F(s_k)\big)^\top \nabla f(s_k) + \frac{1}{2}\beta_k \eta_k(\beta_k)^2 \big\| s_k - F(s_k)\big\|^2, \tag{2.41}$$

where $\eta_k(\beta_k)$ is defined as the adaptive stepsize (2.40) with β replaced by β_k. As it adapts both to s_k and β_k, we will from now on refer to $\eta_k = \eta_k(\beta_k)$ as the *fully* adaptive stepsize. The above discussion implies that (2.41) is always satisfiable if Assumption 2.24 (i) holds, in which case one may simply set β_k to the global smoothness parameter β. In practice, however, the inequality (2.41) is often satisfiable for much smaller values $\beta_k \ll \beta$ that may not even be related to the smoothness properties of the objective function. A close upper bound on the smallest $\beta_k > 0$ that satisfies (2.41) can be found efficiently via backtracking line search. Specifically, the *fully adaptive Frank-Wolfe* algorithm sets β_k to the smallest element of the discrete search space $\frac{\beta_{k-1}}{\zeta} \cdot \{1, \tau, \tau^2, \tau^3, \ldots\}$ that satisfies (2.41), where $\tau > 1$ and $\zeta > 1$ are prescribed line search parameters. A detailed description of the fully adaptive Frank-Wolfe algorithm in pseudocode is provided in Algorithm 2.

It has been shown in [133] that Algorithm 2 enjoys the same sublinear $\mathcal{O}(1/k)$ convergence guarantee as the vanilla Frank-Wolfe algorithm when Assumption 2.24 (i) holds. Below we will leverage techniques from [107, 70] to show that Algorithm 2 offers a linear convergence rate if all of the Assumptions 2.24 (i)–(iii) hold.

Algorithm 2 Fully adaptive Frank-Wolfe algorithm

```
Input: initial feasible point $s_0 \in \mathscr{S}$, initial smoothness parameter $\beta_{-1} > 0$
        line search parameters $\tau > 1$, $\zeta > 1$, initial iteration counter $k = 0$
  while stopping criterion is not met do
    solve the oracle subproblem to find $\tilde{s}_k = F(s_k)$
    set $d_k \leftarrow \tilde{s}_k - s_k$ and $g_k \leftarrow -d_k^\top \nabla f(s_k)$
    set $\beta_k \leftarrow \beta_{k-1}/\zeta$ and $\eta \leftarrow \min\{1, g_k/(\beta_k \|d_k\|^2)\}$
    while $f(s_k + \eta d_k) > f(s_k) - \eta g_k + \frac{\eta^2 \beta_k}{2}\|d_k\|^2$ do
      $\beta_k \leftarrow \tau \beta_k$ and $\eta \leftarrow \min\{1, g_k/(\beta_k \|d_k\|^2)\}$
    end while
    set $\eta_k \leftarrow \eta$ and $s_{k+1} \leftarrow s_k + \eta_k d_k$
    set $k \leftarrow k + 1$
  end while
Output: $s_k$
```

Theorem 2.25 (Linear convergence of the fully adaptive Frank-Wolfe algorithm)**.** *If Assumption 2.24 holds and* $\overline{\beta} \triangleq \max\{\tau\beta, \beta_{-1}\}$*, then Algorithm 2 enjoys the linear convergence guarantee*

$$f(s_k) - f^\star \le \max\left\{1 - \frac{\delta}{2}, 1 - \frac{(1-\sqrt{1-\delta})\alpha\varepsilon}{4\overline{\beta}}\right\}^k (f(s_0) - f^\star) \quad \forall k \in \mathbb{N}.$$

The proof of Theorem 2.25 relies on the following preparatory lemma.

Lemma 2.26 (Bounds on the surrogate duality gap)**.** *The surrogate duality gap* $g_k \triangleq -d_k^\top \nabla f(s_k)$ *corresponding to the search direction* $d_k \triangleq F(s_k) - s_k$ *admits the following lower bounds.*

(i) If the objective function f *is convex, then* $g_k \ge \delta(f(s_k) - f^\star)$*.*

(ii) If the feasible set $\mathscr{S}$ *is* α*-strongly convex in the sense of Assumption 2.24 (ii), then*

$$g_k \ge \frac{(1-\sqrt{1-\delta})\alpha}{2\delta}\|d_k\|^2 \|\nabla f(s_k)\|.$$

Proof. By the definition of g_k we have

$$g_k = \big(s_k - F(s_k)\big)^\top \nabla f(s_k) \ge \delta\big(s_k - s\big)^\top \nabla f(s_k) \qquad \forall s \in \mathscr{S}, \tag{2.42}$$

where the inequality follows from the defining property (2.38) of the inexact oracle with precision δ. Setting s in (2.42) to a global minimizer $s^\star$ of (2.37) then implies via the first-order convexity condition for f that

$$g_k \geq \delta\big(s_k - s^\star\big)^\top \nabla f(s_k) \geq \delta\big(f(s_k) - f^\star\big).$$

This observation establishes assertion (i). To prove assertion (ii), we set

$$s(\theta) = \theta F(s_k) + (1-\theta) s_k - \frac{\alpha}{2}\theta(1-\theta)\|F(s_k) - s_k\|^2 \frac{\nabla f(s_k)}{\|\nabla f(s_k)\|},$$

where α is the strong convexity parameter of the feasible set $\mathscr{S}$, and $\theta \in [0,1]$ is an arbitrary convex weight. By assumption 2.24 (ii) we have $s \in \mathscr{S}$. Moreover, setting s in (2.42) to $s(\theta)$ implies that

$$\begin{aligned} g_k &\geq \delta\Big(\theta\big(s_k - F(s_k)\big) + \frac{\alpha}{2}\theta(1-\theta)\|F(s_k) - s_k\|^2 \frac{\nabla f(s_k)}{\|\nabla f(s_k)\|}\Big)^\top \nabla f(s_k) \\ &= \delta\big(\theta g_k + \frac{\alpha}{2}\theta(1-\theta)\|F(s_k) - s_k\|\|\nabla f(s_k)\|\big) \quad \forall \theta \in [0,1]. \end{aligned}$$

Reordering the above inequality to bring g_k to the left hand side yields

$$g_k \geq \frac{\alpha}{2}\|F(s_k) - s_k\|^2 \|\nabla f(s_k)\| \frac{\delta\theta(1-\theta)}{1-\delta\theta} \quad \forall \theta \in [0,1]. \tag{2.43}$$

A tedious but straightforward calculation shows that the lower bound on the right hand side of (2.43) is maximized by $\theta^\star = (1-\sqrt{1-\delta})/\delta$. Assertion (ii) then follows by substituting $\theta^\star$ into (2.43). □

Proof of Theorem 2.25. By Assumption 2.24 (i) the function f is β-smooth, and thus one can show that

$$f(s_k + \eta d_k) \leq f(s_k) - \eta g_k + \frac{\eta^2 \beta}{2}\|d_k\|^2 \quad \forall \eta \in [0,1], \tag{2.44}$$

where the surrogate duality gap $g_k \geq 0$ and the search direction $d_k \in \mathbb{R}^d$ are defined as in Lemma 2.26. We emphasize that (2.44) holds in fact for all $\eta \in \mathbb{R}$. However, the next iterate $s_{k+1} = s_k + \eta d_k$ may be infeasible unless $\eta \in [0,1]$. The inequality (2.44) implies that any $\beta_k \geq \beta$ satisfies the condition of the inner while loop of Algorithm 2, and thus the loop must terminate at the latest after $\lceil \log(\zeta\beta/\beta_{-1})/\log(\tau)\rceil$ iterations, outputting a smoothness parameter β_k and a stepsize η_k that satisfy the inequality (2.41).

We henceforth denote by $h_k = f(s_k) - f^\star$ the suboptimality of the k-th iterate and note that

$$h_{k+1} = f(s_k + \eta_k d_k) - f(s_k) + h_k \leq -g_k + \frac{1}{2}\beta_k \eta_k^2 \|d_k\|^2 + h_k, \tag{2.45}$$

where the inequality exploits (2.41) and the definitions of g_k and d_k. In order to show that h_k decays geometrically, we distinguish the cases (i) $g_k/(\beta_k\|d_k\|^2) \geq 1$ and (ii) $g_k/(\beta_k\|d_k\|^2) < 1$. In case (i), the stepsize η_k defined in (2.41) satisfies $\eta_k = \min\{1, g_k/(\beta_k\|d_k\|^2)\} = 1$, and thus we have

$$h_{k+1} \leq \left(\frac{\beta_k \|d_k\|^2}{2g_k} - 1\right) g_k + h_k \leq -\frac{g_k}{2} + h_k \leq \left(1 - \frac{\delta}{2}\right) h_k, \tag{2.46}$$

where the first inequality follows from (2.45), while the third inequality holds due to Lemma 2.26 (i).

In case (ii), the step size satisfies $\eta_k = g_k / \beta_k \|d_k\|^2 < 1$, and thus we find

$$\begin{aligned} h_{k+1} &\leq -g_k + \frac{g_k^2}{2\beta_k \|d_k\|^2} + h_k \leq -\frac{g_k^2}{2\beta_k \|d_k\|^2} + h_k \leq \left(1 - \frac{\delta g_k}{2\beta_k \|d_k\|^2}\right) h_k \\ &\leq \left(1 - \frac{(1-\sqrt{1-\delta})\alpha}{4\beta_k} \|\nabla f(s_k)\|\right) h_k \leq \left(1 - \frac{(1-\sqrt{1-\delta})\alpha\varepsilon}{4\overline{\beta}}\right) h_k, \end{aligned} \tag{2.47}$$

where the first and the second inequalities follow from (2.45) and from multiplying $-g_k$ with $\eta_k < 1$, respectively, while the third and the fourth inequalities exploit Lemmas 2.26 (i) and 2.26 (ii), respectively. The last inequality in (2.47) holds because of Assumption 2.24 (iii) and because $\beta_k \leq \overline{\beta}$ for all $k \in \mathbb{N}$; see [133, Proposition 2]. By the estimates (2.46) and (2.47), the suboptimality of the current iterate decays at least by

$$\max\left\{1 - \frac{\delta}{2}, 1 - \frac{(1-\sqrt{1-\delta})\alpha\varepsilon}{4\overline{\beta}}\right\} < 1$$

in each iteration of the algorithm. This observation completes the proof. □

2.6.2 Frank-Wolfe Algorithm for Wasserstein MMSE Estimation Problems

We now use the fully adaptive Frank-Wolfe algorithm of Section 2.6.1 to solve the nonlinear SDP (2.25), which is equivalent to the dual Wasserstein MMSE estimation problem over normal priors. Recall from Corollary 2.20 that any solution of (2.25) can be used to construct both a least favorable prior and an optimal estimator that form a Nash equilibrium. Unlike the generic convex program (2.37), the specific nonlinear SDP (2.25) is a convex *maximization* problem, and thus Algorithm 2 is applicable only after some obvious minor modifications.

Throughout this section we assume that $\widehat{\Sigma}_w \succ 0$, which implies via Theorem 2.15 that the nonlinear SDP (2.25) is solvable and and can be reformulated more concisely as

$$\max_{\Sigma_x \in \mathcal{S}_x, \Sigma_w \in \mathcal{S}_w} f(\Sigma_x, \Sigma_w),$$

where the objective function $f : \mathcal{S}_x \times \mathcal{S}_w \to \mathbb{R}$ is defined through

$$f(\Sigma_x, \Sigma_w) \triangleq \operatorname{Tr}\left[\Sigma_x - \Sigma_x H^\top \left(H \Sigma_x H^\top + \Sigma_w\right)^{-1} H \Sigma_x\right],$$

and where the separate feasible sets for Σ_x and Σ_w are given by

$$\mathscr{S}_x \triangleq \left\{\Sigma_x \in \mathbb{S}_+^n : \operatorname{Tr}\left[\Sigma_x + \widehat{\Sigma}_x - 2\left(\widehat{\Sigma}_x^{\frac{1}{2}}\Sigma_x\widehat{\Sigma}_x^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \le \rho_x^2,\ \Sigma_x \succeq \lambda_{\min}(\widehat{\Sigma}_x) I_n\right\}$$

and

$$\mathscr{S}_w \triangleq \left\{\Sigma_w \in \mathbb{S}_+^m : \operatorname{Tr}\left[\Sigma_w + \widehat{\Sigma}_w - 2\left(\widehat{\Sigma}_w^{\frac{1}{2}}\Sigma_w\widehat{\Sigma}_w^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \le \rho_w^2,\ \Sigma_w \succeq \lambda_{\min}(\widehat{\Sigma}_w) I_m\right\},$$

respectively. One readily verifies that f is convex and differentiable. Moreover, Lemma 2.36 implies that both $\mathscr{S}_x$ and $\mathscr{S}_w$ are convex and compact. The oracle problem that linearizes the objective function of the nonlinear SDP around a fixed feasible solution $\Sigma_x \in \mathscr{S}_x$ and $\Sigma_w \in \mathscr{S}_w$ can thus be expressed concisely as

$$\max_{L_x \in \mathscr{S}_x, L_w \in \mathscr{S}_w} \big\langle L_x - \Sigma_x, D_x\big\rangle + \big\langle L_w - \Sigma_w, D_w\big\rangle,$$

where $D_x \triangleq \nabla_{\Sigma_x} f\big(\Sigma_x, \Sigma_w\big)$ and $D_w \triangleq \nabla_{\Sigma_w} f\big(\Sigma_x, \Sigma_w\big)$.

$$D_x = (I_n - \Sigma_x H^\top G^{-1} H)^\top (I_n - \Sigma_x H^\top G^{-1} H)$$

and

$$D_w = G^{-1} H \Sigma_x^2 H^\top G^{-1},$$

respectively, and where $G \triangleq H\Sigma_x H^\top + \Sigma_w$. Details on the calculation of the derivatives can be found in Appendix 2.8.2.

The oracle problem is separable in (L_x, L_w) and its approximate solution can be found independently as the optimal solutions of two separate sub-programs of similar structure. In fact, we are interested in finding a feasible point $\tilde{L}_i$ satisfying

$$\begin{array}{rll}\big\langle \tilde{L}_i - \Sigma_i, D_i\big\rangle \ge \delta \displaystyle\max_{L_i \in \mathbb{S}_+^d} & \big\langle L_i - \Sigma_i, D_i\big\rangle & \\ \text{s.t.} & \operatorname{Tr}\big[L_i + \widehat{\Sigma}_i - 2\big(\widehat{\Sigma}_i^{\frac{1}{2}} L_i \widehat{\Sigma}_i^{\frac{1}{2}}\big)^{\frac{1}{2}}\big] \le \rho_i^2, \quad L_i \succeq \lambda_{\min}(\widehat{\Sigma}_i) I_d\end{array} \tag{2.48}$$

for $i \in \{x, w\}$. In the earlier version of this paper, we construct an approximate additive oracle for the above maximization problem whose correctness was guaranteed by [152, Theorem 3.2]. Algorithm 3, however, constructs an approximate multiplicative oracle to ensure the linear convergence of the fully adaptive Frank-Wolfe Algorithm 2.

Theorem 2.27 (Direction-finding subproblem)**.** *For any fixed input* $(\Sigma_i, D_i, \widehat{\Sigma}_i, \rho_i, \delta)$ *with* $i \in \{x, w\}$, *Algorithm 3 returns a feasible point satisfying* (2.48).

Proof. Denote the feasible set of the optimization program in (2.48) by $\mathscr{S}_i$. Note that the objective function f is jointly concave in Σ_x and Σ_w. Therefore, by concavity of f, for any

Algorithm 3 Bisection algorithm to solve (2.48)

```
Input: covariance matrix $\Sigma \in \mathbb{S}^d_{++}$ and its gradient matrix $D \in \mathbb{S}^d_{+}$,
       Wasserstein center $\widehat{\Sigma} \in \mathbb{S}^d_{++}$ and radius $\rho > 0$,
       oracle parameter $\delta \in (0,1)$
  Denote the largest eigenvalue of $D$ by $\lambda_1$, let $v_1$ be an eigenvector of $\lambda_1$
  Set $LB \leftarrow \lambda_1(1 + (v_1^\top \widehat{\Sigma} v_1)^{\frac{1}{2}}/\rho)$, $UB \leftarrow \lambda_1(1 + \mathrm{Tr}\big[\widehat{\Sigma}\big]^{\frac{1}{2}}/\rho)$
  repeat
    Set $\gamma \leftarrow (UB + LB)/2$, $\tilde{L} \leftarrow \gamma^2(\gamma I_d - D)^{-1}\widehat{\Sigma}(\gamma I_d - D)^{-1}$
    if $\rho^2 - \big\langle \widehat{\Sigma}, \big(I_d - \gamma(\gamma I_d - D)^{-1}\big)^2 \big\rangle < 0$ then
      Set $LB \leftarrow \gamma$
    else
      Set $UB \leftarrow \gamma$
    end if
    Set $\Delta \leftarrow \big\langle L - \Sigma, D \big\rangle / \big(\gamma(\rho^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]) + \gamma^2\big\langle (\gamma I_d - D)^{-1}, \widehat{\Sigma} \big\rangle - \big\langle \Sigma, D \big\rangle\big)$
  until $\rho^2 - \big\langle \widehat{\Sigma}, \big(I_d - \gamma(\gamma I_d - D)^{-1}\big)^2 \big\rangle > 0$ and $\Delta \geq \delta$
Output: $\tilde{L}$
```

$(\Sigma_x, \Sigma_w) \in \mathscr{S}_x \times \mathscr{S}_w$ we have

$$\max_{L_i \in \mathscr{S}_i} \big\langle L_i - \Sigma_i, D_i \big\rangle \geq f_i^\star - f(\Sigma_x, \Sigma_w) \geq 0$$

for $i \in \{x, w\}$, where $f_x^\star = \max_{S_x \in \mathscr{S}_x} f(S_x, \Sigma_w)$ and $f_w^\star = \max_{S_w \in \mathscr{S}_w} f(\Sigma_x, S_w)$; see for example [88, Section 2.2] for a proof. By Proposition 2.34 (i) and (iii), the above maximization admits the dual

$$\inf_{\gamma_i > \lambda_{\max}(D_i)} \gamma_i\Big(\rho_i^2 + \big\langle \gamma_i(\gamma_i I - D_i)^{-1} - I, \widehat{\Sigma}_i \big\rangle\Big) - \big\langle \Sigma_i, D_i \big\rangle \geq 0 \tag{2.49}$$

and its optimal solution follows the form

$$L_i^\star = (\gamma_i^\star)^2(\gamma_i^\star I - D_i)^{-1}\widehat{\Sigma}_i(\gamma_i^\star I - D_i)^{-1},$$

where $\gamma_i^\star$ is the unique optimizer of the above minimization problem satisfying

$$\rho_i^2 - \big\langle \widehat{\Sigma}_i, \big(I - \gamma_i^\star(\gamma_i^\star I - D_i)^{-1}\big)^2 \big\rangle = 0.$$

It is proved in [152, Theorem 3.2] that the condition $\rho_i^2 - \big\langle \widehat{\Sigma}_i, \big(I - \gamma_i(\gamma_i I - D_i)^{-1}\big)^2 \big\rangle > 0$ ensures the feasibility of the output; thus, we only focus on proving that the output is a δ-approximate solution, that is, it satisfies (2.48). Notice that any γ in Algorithm 3 is feasible in (2.49) by construction, and thus, we have

$$\gamma(\rho^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]) + \gamma^2\big\langle (\gamma I_d - D)^{-1}, \widehat{\Sigma} \big\rangle - \big\langle \Sigma, D \big\rangle \geq \max_{L_i \in \mathscr{S}_i} \big\langle L_i - \Sigma_i, D_i \big\rangle \geq 0.$$

Therefore, the condition $\Delta \geq \delta$ implies that L satisfies (2.48), that is,

$$\Delta \geq \delta \implies \frac{\langle \tilde{L} - \Sigma_i, D_i \rangle}{\max_{S_i \in \mathcal{S}_i} \langle S_i - \Sigma_i, D_i \rangle} \geq \delta.$$

Note that both numerator and denominator of Δ are Lipschitz continuous in γ for the bisection interval, and therefore, the bisection Algorithm 2 will terminate in finite time. □

If $\widehat{\Sigma}$ is singular, one simple idea is to add κI_d with $\kappa > 0$ to the nominal matrix in (2.48) and obtain an approximate solution to the direction finding subproblem using Algorithm 3. The next proposition asserts that the dual estimation problem (2.25) satisfies the conditions of Assumption 2.24.

Proposition 2.28 (Properties of the dual estimation problem (2.25))**.** *The dual estimation problem* (2.25) *satisfies the following properties:*

(i) *Its objective function is β-smooth over its feasible set, with the smoothness constant β satisfying*

$$\beta = 2\lambda_{\min}^{-1}(\widehat{\Sigma}_w)\big(C + C\,\lambda_{\max}^{2}(H^\top H) + \lambda_{\max}(H^\top H)\big), \tag{2.50}$$

where $C = \min\left\{\lambda_{\min}^{-1}(H^\top H), \lambda_{\max}(H^\top H)\cdot\lambda_{\min}^{-2}(\widehat{\Sigma}_w)\cdot\big(\rho_x + \operatorname{Tr}\big[\widehat{\Sigma}_x\big]^{\frac{1}{2}}\big)^4\right\}$.

(ii) *Its feasible set is α-strongly convex with parameter*

$$\alpha = \min\left\{\frac{\lambda_{\min}^{\frac{5}{4}}(\widehat{\Sigma}_x)}{2\rho_x\big(\rho_x + \operatorname{Tr}\big[\widehat{\Sigma}_x\big]^{\frac{1}{2}}\big)^{\frac{7}{2}}}, \frac{\lambda_{\min}^{\frac{5}{4}}(\widehat{\Sigma}_w)}{2\rho_w\big(\rho_w + \operatorname{Tr}\big[\widehat{\Sigma}_w\big]^{\frac{1}{2}}\big)^{\frac{7}{2}}}\right\}.$$

(iii) *Its objective function satisfies the lower-bounded gradient condition over its feasible set. More specifically, we have*

$$\min\{\|D_x\|_F, \|D_w\|_F\} \geq \varepsilon,$$

where $\varepsilon = \min\{\varepsilon_x,\ \varepsilon_w\}$ *with*

$$\varepsilon_x = \left(\frac{\lambda_{\min}(\widehat{\Sigma}_w)}{\big(\rho_w + \operatorname{Tr}\big[\widehat{\Sigma}_w\big]^{\frac{1}{2}}\big)^2 + \big(\rho_x + \operatorname{Tr}\big[\widehat{\Sigma}_x\big]^{\frac{1}{2}}\big)^2\lambda_{\max}(H^\top H)}\right)^2,$$

$$\varepsilon_w = \lambda_{\max}(H^\top H)\left(\frac{\lambda_{\min}(\widehat{\Sigma}_x)}{\big(\rho_w + \operatorname{Tr}\big[\widehat{\Sigma}_w\big]^{\frac{1}{2}}\big)^2 + \lambda_{\min}(\widehat{\Sigma}_x)\lambda_{\max}(H^\top H)}\right)^2.$$

(iv) *Its feasible set has diameter upper bounded by*

$$D_{\mathcal{S}} = \big(\rho_x + \operatorname{Tr}\big[\widehat{\Sigma}_x\big]^{\frac{1}{2}}\big)^2 + \big(\rho_w + \operatorname{Tr}\big[\widehat{\Sigma}_w\big]^{\frac{1}{2}}\big)^2.$$

Proof. To establish the smoothness property of f in claim (i), it suffices to provide a uniform upper bound for the largest eigenvalue of the negative Hessian matrix $\mathscr{H}$ defined in Section 2.8.2 over the feasible set of problem (2.25). The maximum eigenvalue of $\mathscr{H}$ can be upper bounded by

$$\lambda_{\max}(\mathscr{H}) \leq \lambda_{\max}(\mathscr{H}_{xx}) + \lambda_{\max}(\mathscr{H}_{ww}) = 2\left(\lambda_{\max}(D_x) \cdot \lambda_{\max}\left(H^\top G^{-1} H\right) + \lambda_{\max}(D_w) \cdot \lambda_{\max}\left(G^{-1}\right)\right),$$

where the inequality follows from [12, Fact 5.12.20] and the equality follows from [12, Proposition 7.1.10]. In the sequel, we provide the upper bound for each individual term in the above expression.

Because $G = H\Sigma_x H^\top + \Sigma_w$, we have $G \succeq \lambda_{\min}(\Sigma_w) I_m \succeq \lambda_{\min}(\widehat{\Sigma}_w) I_m$ and thus

$$\lambda_{\max}(G^{-1}) = \sigma_{\max}(G^{-1}) \leq \lambda_{\min}^{-1}(\widehat{\Sigma}_w),$$

and furthermore,

$$\lambda_{\max}(H^\top G^{-1} H) \leq \sigma_{\max}(G^{-1}) \cdot \sigma_{\max}^2(H) = \lambda_{\min}^{-1}(\widehat{\Sigma}_w) \cdot \lambda_{\max}(H^\top H).$$

Because $D_w = G^{-1} H \Sigma_x^2 H^\top G^{-1}$, we find

$$\lambda_{\max}(D_w) \leq \lambda_{\max}^2(\Sigma_x) \cdot \lambda_{\max}^2(G^{-1}) \cdot \lambda_{\max}(H^\top H) \leq \left(\rho_x + \operatorname{Tr}\left[\widehat{\Sigma}_x\right]^{\frac{1}{2}}\right)^4 \cdot \lambda_{\min}^{-2}(\widehat{\Sigma}_w) \cdot \lambda_{\max}(H^\top H),$$

where the last inequality follows from the bound in Lemma 2.36. Finally, we bound $\lambda_{\max}(D_x)$ by

$$\begin{aligned}
\lambda_{\max}(D_x) &\leq \lambda_{\max}(I_n) + \lambda_{\max}(H^\top G^{-1} H \Sigma_x^2 H^\top G^{-1} H) \\
&= 1 + \sigma_{\max}^2(H^\top G^{-1} H \Sigma_x) \\
&\leq 1 + \lambda_{\max}^2(\Sigma_x) \cdot \lambda_{\max}^2(G^{-1}) \cdot \lambda_{\max}^2(H^\top H) \\
&\leq 1 + \left(\rho_x + \operatorname{Tr}\left[\widehat{\Sigma}_x\right]^{\frac{1}{2}}\right)^4 \cdot \lambda_{\min}^{-2}(\widehat{\Sigma}_w) \cdot \lambda_{\max}^2(H^\top H).
\end{aligned}$$

The upper bound on $\lambda_{\max}(D_x)$ and $\lambda_{\max}(D_w)$ can be strengthened whenever H is of full column rank by exploiting the relationship $G^2 \succ \left(H\Sigma_x H^\top\right)^2 \succeq \lambda_{\min}(H^\top H) \cdot H\Sigma_x^2 H^\top$, which in turn implies that

$$\begin{aligned}
\lambda_{\max}(D_w) &= \lambda_{\max}(G^{-1} H \Sigma_x^2 H^\top G^{-1}) \leq \lambda_{\min}(H^\top H)^{-1}, \\
\lambda_{\max}(D_x) &\leq \lambda_{\max}(I_n + H^\top G^{-1} H \Sigma_x^2 H^\top G^{-1} H) \leq 1 + \lambda_{\max}(H^\top H) \lambda_{\min}(H^\top H)^{-1}.
\end{aligned}$$

Combining all the inequalities, we conclude that

$$\lambda_{\max}(\mathscr{H}) \leq 2\lambda_{\min}^{-1}(\widehat{\Sigma}_w)\left(C + C\,\lambda_{\max}^2(H^\top H) + \lambda_{\max}(H^\top H)\right),$$

where the constant C admits the value

$$C = \min\left\{\lambda_{\min}^{-1}(H^\top H), \lambda_{\max}(H^\top H) \cdot \lambda_{\min}^{-2}(\widehat{\Sigma}_w) \cdot \left(\rho_x + \operatorname{Tr}\left[\widehat{\Sigma}_x\right]^{\frac{1}{2}}\right)^4\right\}.$$

The claim about the smoothness of f thus follows.

To prove claim (ii), it is sufficient to show strong convexity for $\theta = 0.5$. We denote by $\mathscr{S}_x$ and $\mathscr{S}_w$ the feasible sets of the variables Σ_x and Σ_w in problem (2.25), that is,

$$\mathscr{S}_x \triangleq \left\{\Sigma_x \in \mathbb{S}^n_+ : \Sigma_x \succeq \lambda_{\min}(\widehat{\Sigma}_x) I_n,\ \mathrm{Tr}\left[\Sigma_x + \widehat{\Sigma}_x - 2\big(\widehat{\Sigma}_x^{\frac{1}{2}}\Sigma_x\widehat{\Sigma}_x^{\frac{1}{2}}\big)^{\frac{1}{2}}\right] \le \rho_x^2\right\} \tag{2.51a}$$

and

$$\mathscr{S}_w \triangleq \left\{\Sigma_w \in \mathbb{S}^m_{++} : \Sigma_w \succeq \lambda_{\min}(\widehat{\Sigma}_w) I_m,\ \mathrm{Tr}\left[\Sigma_w + \widehat{\Sigma}_w - 2\big(\widehat{\Sigma}_w^{\frac{1}{2}}\Sigma_w\widehat{\Sigma}_w^{\frac{1}{2}}\big)^{\frac{1}{2}}\right] \le \rho_w^2\right\}, \tag{2.51b}$$

where the redundant constraints $\Sigma_x \succeq \lambda_{\min}(\widehat{\Sigma}_x) I_n$ and $\Sigma_w \succeq \lambda_{\min}(\widehat{\Sigma}_w) I_m$ has been made explicit in $\mathscr{S}_x$ and $\mathscr{S}_w$, respectively. Fix any $\Sigma_x \in \mathscr{S}_x$, $L_x \in \mathscr{S}_x$, $\Sigma_w \in \mathscr{S}_w$ and $L_w \in \mathscr{S}_w$. Let $E_x = \frac{1}{2}(\Sigma_x + L_x)$ and $E_w = \frac{1}{2}(\Sigma_w + L_w)$ be the corresponding midpoint, while D_x and D_w denote the components of the derivative of the objective function at (Σ_x, Σ_w). Let $Z_x = D_x / \|D_x\|_F \succeq 0$ and $Z_w = D_w / \|D_w\|_F \succeq 0$, where the positive semidefiniteness of Z_x and Z_w follows from the positive semidefiniteness of D_x and D_w, respectively. Consider now the point

$$(\tilde{L}_x, \tilde{L}_w) = \Big(E_x + \frac{\alpha}{8}\|\Sigma_x - L_x\|_F^2 Z_x, E_w + \frac{\alpha}{8}\|\Sigma_w - L_w\|_F^2 Z_w\Big).$$

Because Σ_x, L_x and Z_w are positive semidefinite, $\tilde{L}_w$ is also positive semidefinite by construction. Because $\lambda_{\min}(\widehat{\Sigma}_w) I_m \preceq \widehat{\Sigma}_w \preceq (\rho_w + \mathrm{Tr}\big[\widehat{\Sigma}_w\big]^{\frac{1}{2}})^2 I_m$, [14, Theorem 1] implies that the non-negative function $\Sigma_w \mapsto \mathrm{Tr}\big[\Sigma_w + \widehat{\Sigma}_w - 2(\widehat{\Sigma}_w^{\frac{1}{2}}\Sigma_w\widehat{\Sigma}_w^{\frac{1}{2}})^{\frac{1}{2}}\big]$ is $\lambda_{\min}^{\frac{1}{2}}(\widehat{\Sigma}_w)/[2(\rho_w + \mathrm{Tr}\big[\widehat{\Sigma}_w\big]^{\frac{1}{2}})^3]$-strongly convex and $(\rho_w + \mathrm{Tr}\big[\widehat{\Sigma}_w\big]^{\frac{1}{2}})/[2\lambda_{\min}^{\frac{3}{2}}(\widehat{\Sigma}_w)]$-smooth. One can now use the reasoning in the proof of [94, Theorem 12] to show that

$$\mathrm{Tr}\left[\tilde{L}_w + \widehat{\Sigma}_w - 2\big(\widehat{\Sigma}_w^{\frac{1}{2}}\tilde{L}_w\widehat{\Sigma}_w^{\frac{1}{2}}\big)^{\frac{1}{2}}\right] \le \rho_w^2.$$

Through an analogous argument, we have $\tilde{L}_x \succeq 0$ and

$$\mathrm{Tr}\left[\tilde{L}_x + \widehat{\Sigma}_x - 2\big(\widehat{\Sigma}_x^{\frac{1}{2}}\tilde{L}_x\widehat{\Sigma}_x^{\frac{1}{2}}\big)^{\frac{1}{2}}\right] \le \rho_x^2.$$

These results imply that $(\tilde{L}_x, \tilde{L}_w)$ is feasible for problem (2.25), and thus claim (ii) holds.

Finally we proceed to prove the lower-bounded gradient condition of f in claim (iii). Denote by $\mathrm{spec}(T)$ the spectrum, or the set of eigenvalues, of any square matrix T. Recall that we set $G = H\Sigma_x H^\top + \Sigma_w$. Let $T_1 = I_n - \Sigma_x H^\top G^{-1} H$ and $T_2 = H\Sigma_x H^\top G^{-1} = I_m - \Sigma_w G^{-1}$. Then the largest eigenvalue of $D_x = T_1^\top T_1$ satisfies

$$\lambda_{\max}(D_x) = \sigma_{\max}^2(T_1) \ge \max_{\lambda \in \mathrm{spec}(T_1)} |\lambda|^2 = \max_{\lambda \in \mathrm{spec}(T_2)} |1 - \lambda|^2 = \lambda_{\max}^2(\Sigma_w G^{-1}), \tag{2.52}$$

where $|\cdot|$ is the absolute value of a (possibly complex) number. The first inequality follows from Browne's theorem [12, Fact 5.11.21], the second equality holds because the nonzero spectrum of $\Sigma_x H^\top G^{-1} H$ is identical to the ones of $H\Sigma_x H^\top G^{-1}$ [12, Proposition 4.4.10], and the last equality follows from the fact that

$$\mathrm{spec}(T_2) = \mathrm{spec}((G-\Sigma_w)G^{-1}) = 1 - \mathrm{spec}(\Sigma_w G^{-1}).$$

Notice that all eigenvalues of $\Sigma_w G^{-1}$ are real because the nonzero spectrum of $\Sigma_w G^{-1}$ and $G^{-\frac{1}{2}}\Sigma_w G^{-\frac{1}{2}}$ are identical, and $G^{-\frac{1}{2}}\Sigma_w G^{-\frac{1}{2}}$ is a symmetric matrix with real eigenvalues.

To achieve a uniform lower bound on the largest eigenvalue of D_x, we are interested in the following optimization problem and its lower bounds

$$\begin{aligned}\inf_{\substack{\Sigma_x\in\mathscr{S}_x\\ \Sigma_w\in\mathscr{S}_w}} \lambda_{\max}(\Sigma_w G^{-1}) &\geq \inf_{\substack{\Sigma_x\in\mathscr{S}_x\\ \Sigma_w\in\mathscr{S}_w}} \lambda_{\max}\left(\Sigma_w(\Sigma_w + \lambda_{\max}(\Sigma_x)\lambda_{\max}(H^\top H)\, I_m)^{-1}\right)\\ &\geq \inf_{\Sigma_w\in\mathscr{S}_w} \frac{\lambda_{\min}(\Sigma_w)}{\lambda_{\max}(\Sigma_w) + \left(\rho_x + \mathrm{Tr}\left[\widehat{\Sigma}_x\right]^{\frac{1}{2}}\right)^2 \lambda_{\max}(H^\top H)}\\ &\geq \frac{\lambda_{\min}(\widehat{\Sigma}_w)}{\left(\rho_w + \mathrm{Tr}\left[\widehat{\Sigma}_w\right]^{\frac{1}{2}}\right)^2 + \left(\rho_x + \mathrm{Tr}\left[\widehat{\Sigma}_x\right]^{\frac{1}{2}}\right)^2 \lambda_{\max}(H^\top H)}. \end{aligned} \tag{2.53}$$

Combining (2.52) and (2.53), we find $\|D_x\|_F \geq \lambda_{\max}(D_x) \geq \delta_x$, where δ_x is defined in the lemma statement.

Using an analogous reasoning, a uniform lower bound on the largest eigenvalue of D_w can be obtained by

$$\begin{aligned}\lambda_{\max}(D_w)\cdot\lambda_{\max}(H^\top H) \geq \lambda_{\max}(H^\top D_w H) &= \sigma^2_{\max}(H^\top G^{-1} H\Sigma_x)\\ &\geq \lambda^2_{\max}(H\Sigma_x H^\top G^{-1}) = \left(1 - \lambda_{\min}(\Sigma_w G^{-1})\right)^2\\ &= \left(1 - \frac{1}{1+\lambda_{\max}(H\Sigma_x H^\top \Sigma_w^{-1})}\right)^2.\end{aligned}$$

A uniform lower bound for $\lambda_{\max}(H\Sigma_x H^\top \Sigma_w^{-1})$ can be obtained by

$$\inf_{\substack{\Sigma_x\in\mathscr{S}_x\\ \Sigma_w\in\mathscr{S}_w}} \lambda_{\max}(H\Sigma_x H^\top\Sigma_w^{-1}) \geq \inf_{\substack{\Sigma_x\in\mathscr{S}_x\\ \Sigma_w\in\mathscr{S}_w}} \frac{\lambda_{\min}(\Sigma_x)}{\lambda_{\max}(\Sigma_w)}\cdot\lambda_{\max}(H^\top H) \geq \frac{\lambda_{\min}(\widehat{\Sigma}_x)}{\left(\rho_w + \mathrm{Tr}\left[\widehat{\Sigma}_w\right]^{\frac{1}{2}}\right)^2}\cdot\lambda_{\max}(H^\top H),$$

where the last inequality follows from the fact that $\Sigma_w \preceq (\rho_w + \mathrm{Tr}\left[\widehat{\Sigma}_w\right]^{\frac{1}{2}})^2 I_m$ for any $\Sigma_w \in \mathscr{S}_w$ (see Lemma 2.36) and $\Sigma_x \succeq \lambda_{\min}(\widehat{\Sigma}_x) I_n$ for any $\Sigma_x \in \mathscr{S}_x$. If we define δ_w as in the Lemma statement, then $\|D_w\|_F \geq \lambda_{\max}(D_w) \geq \delta_w$.

Lemma 2.36 implies that both sets $\mathscr{S}_x$ and $\mathscr{S}_w$ defined in (2.51) are bounded, thus the diameter of the joint feasible set $\mathscr{S} = \mathscr{S}_x \times \mathscr{S}_w$ with respect to the Frobenius norm can be bounded by

$$\mathrm{diam}_{\|\cdot\|_F}(\mathscr{S}) \leq \sup_{\Sigma_x\in\mathscr{S}_x} \mathrm{Tr}\left[\Sigma_x\right] + \sup_{\Sigma_w\in\mathscr{S}_w} \mathrm{Tr}\left[\Sigma_w\right] \leq \left(\rho_x + \mathrm{Tr}\left[\widehat{\Sigma}_x\right]^{\frac{1}{2}}\right)^2 + \left(\rho_w + \mathrm{Tr}\left[\widehat{\Sigma}_w\right]^{\frac{1}{2}}\right)^2,$$

where the first inequality holds due to [12, Equation (9.2.16)] and the second inequality follows from the proof of Lemma 2.36. This completes the proof. □

2.7 Numerical Experiments

All experiments are run on an Intel XEON CPU with 3.40GHz clock speed and 16GB of RAM. All SDPs are solved with MOSEK 8 using the YALMIP interface [112]. In order to ensure the reproducibility of our experiments, we make all source codes available at https://github.com/sorooshafiee/WAE.

2.7.1 Convergence Behavior on Synthetic Problems

In the first experiment we aim to study the convergence behavior of the Frank-Wolfe algorithm and make a comparison to a commercial SDP solver. The experiment comprises 10 simulation runs. In each run we randomly generate two covariance matrices $\widehat{\Sigma}_x$ and $\widehat{\Sigma}_w$ as follows. First, we draw two matrices Q_x and Q_w from the standard normal distribution on $\mathbb{R}^{d\times d}$, and we denote by R_x and R_w the orthogonal matrices whose columns correspond to the orthonormal eigenvectors of $Q_x + Q_x^\top$ and $Q_w + Q_w^\top$, respectively. Then, we define $\widehat{\Sigma}_x = R^\star \Lambda^\star (R^\star)^\top$ and $\widehat{\Sigma}_w = R\Lambda R^\top$, where $\Lambda^\star$ and Λ are diagonal matrices whose main diagonals are sampled uniformly from $[1,10]^d$ and $[1,2]^d$, respectively. The distributionally robust MMSE estimator is obtained by solving (2.25) for $\rho_x = \rho_w = \sqrt{d}$ via the Frank-Wolfe algorithm from Section 2.6.

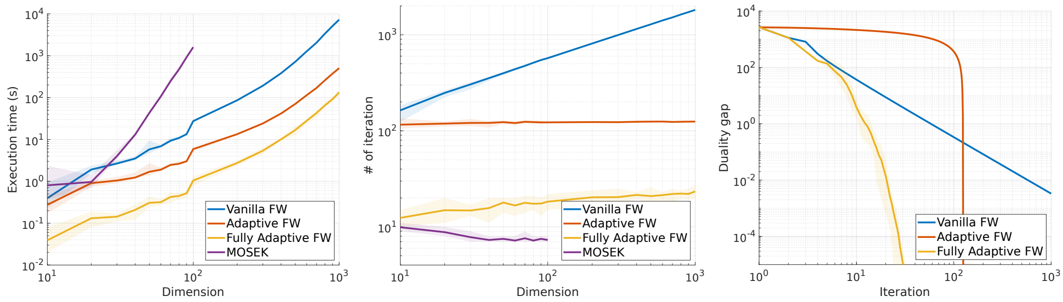
Figure 2.1a and Figure 2.1b compare the execution time and the number of required iterations to obtain the duality gap below 10^{-3} for different dimension d and solution approaches. In particular, we compare the solution time and iteration numbers when we solve the linear SDP (2.31) with MOSEK and the nonlinear SDP (2.25) with different variants of the Frank-Wolfe algorithm. Unfortunately, we fail to solve the linear SDP (2.31) using MOSEK with an out of memory error when the dimension d is larger than 100. Figure 2.1c compares the empirical rate of convergence for different variants of the Frank-Wolfe algorithm. As we observe, the fully adaptive version convergence to a very accurate solution after 20 iterations.

2.7.2 Image Denoising via Wasserstein Wavelet Shrinkage

In this section, we consider a wavelet shrinkage settings which has gained ubiquitous applications in image processing [46]. Suppose that our raw data $Y \in \mathbb{R}^{2^N}$ consists of noisy observations at 2^N regularly spaced points

$$Y_i = X_i + W_i, \quad i = 1,\dots,2^N,$$

of an unknown signal $X \in \mathbb{R}^{2^N}$ and W_i are i.i.d. noise following a normal distribution $\mathcal{N}(0,\sigma_w^2)$. The wavelet-based approach is centered around the assumption that the signal X is, or can be approximated by, a function with a small number of non-zeros wavelet coefficients. Under this



(a) Scaling of execution time (b) Scaling of iteration count (c) Convergence for $d = 1000$

Figure 2.1 – Convergence behavior of the Frank-Wolfe algorithm (shown are the average (solid line) and the range (shaded area) of the respective performance measures across 10 simulation runs)

assumption, the observation Y, the signal X and the noise W can be encoded onto the wavelet space using a periodic Discrete Wavelet Transform (DWT) represented by an operator $\mathscr{W}$. Under the operator $\mathscr{W}$, the linear observation system can be written in the wavelet coefficient space as

$$y = x + w, \tag{2.54}$$

where $y = \mathscr{W} Y \in \mathbb{R}^{2^N}$ are the wavelet coefficients of the noisy observation Y, $x = \mathscr{W} X \in \mathbb{R}^{2^N}$ are the wavelet coefficients of the true signal X and $w = \mathscr{W} W \in \mathbb{R}^{2^N}$ are the wavelet coefficients of the additive white noise. Because the periodic DWT $\mathscr{W}$ is an orthogonal operator, the DWT transforms a white noise into another white noise, thus the components of w are also i.i.d. random variables following a normal distribution $\mathcal{N}(0, \sigma_w^2)$.

Because of the orthogonality of the wavelet basis $\mathscr{W}$, the inverse DWT is its adjoint $\mathscr{W}^\top$. Thus, if $\widehat{x}$ is an estimate of the true wavelet coefficient of the signal x, there exists an estimate $\widehat{X}$ of the true signal X constructed from $\widehat{x}$ using the inverse DWT as $\widehat{X} = \mathscr{W}^\top \widehat{x}$. Furthermore, the Parseval's relation [114, Theorem 2.3] implies the isometric property $\|\widehat{x} - x\|_2 = \|\widehat{X} - X\|_{L_2}$. As a consequence, the MMSE estimator in the wavelet coefficient space leads to the MMSE estimator in the true signal space through the inverse wavelet transformation $\mathscr{W}^\top$. The readers are encouraged to refer to [114] for a comprehensive introduction on wavelet transform and its applications in signal processing.

Based on these properties, a wide class of signal, and especially image, denoising algorithms is based on the combination of the DWT and a wavelet coefficient shrinkage method. Denoising additive Gaussian white noise via thresholding wavelet coefficients was pioneered by [47], where the coefficients are compared to a given a threshold. Other examples include VisuShrink [47], RiskShrink [47], and SureShrink [45], where the thresholding operator is either hard or soft thresholding functions. Nevertheless, thresholding methods suffer from several drawbacks: (i) the threshold is usually selected in an *ad hoc* manner, (ii) soft thresholding provides a biased image, and (iii) hard thresholding yields an image with less bias but a higher variance. Shrinking the wavelet coefficients using a Bayes estimator can provide a better performance, and has been exploited in [23, 24, 117, 149]. The basic approach in Bayesian methods assumes that the local distribution of the wavelet coefficients are Gaussian with spatially varying variance. For example, the wavelet coefficients within subbands is assumed to follows an independent Gaussian distribution in [23, 149]. BayesShrink exploits this stochastic representation to introduce a data driven scheme for adaptive soft thresholding in different subbands [23].

Our Wasserstein denoising filter consists of the following three components:

1. A DWT to transform noisy data into wavelet coefficients and an inverse DWT to reconstruct the signal.
2. A local estimator of the covariance matrix for the signal and noise similar to [47, 45, 117].
3. A Wasserstein estimator, constructed using the Frank-Wolfe algorithm with an efficient implementation to exploit the structure of the denoising problem that can scale up for high resolution data.

Given a specific image $\widehat{Y} = (\widehat{Y}_i)_{i=1,\dots,2^N}$, one can apply the DWT to get the wavelet coefficients $\widehat{y} = \mathscr{W}\widehat{Y}$ and set the nominal noise distribution to $\widehat{\mathbb{P}}_w \sim \mathscr{N}(0, \widehat{\Sigma}_w)$ with $\widehat{\Sigma}_w = \widehat{\sigma}_w^2 I_{2^N}$, where the noise variance is estimated from the components $\widehat{y}_i$ contained in the subband HH_1 using a robust median estimator [47, 45] of the form

$$\widehat{\sigma}_w = \frac{\mathrm{Median}(\{|\widehat{y}_i|\}_{i\in \mathrm{HH}_1})}{0.6745} > 0.$$

Furthermore, an estimation of the moments information of the wavelet coefficients of the true signal x is proposed in [117] where the true image wavelet coefficients are modelled as realizations of a doubly stochastic process. In more details, x are assumed to be independent zero-mean Gaussian random variables with varying variances. These variances can be locally estimated as

$$(\widehat{\Sigma}_x)_{ii} = \max\left(0, \frac{1}{|\mathscr{M}_i|} \sum_{j\in\mathscr{M}_i} \widehat{y}_j^2 - \widehat{\sigma}_w^2\right)$$

for the i-th wavelet coefficient, where $\mathscr{M}_i$ is a set containing the indices of coefficients in the neighborhood of the i-th coefficient, and $|\mathscr{M}_i|$ is its cardinality. The reference distribution of the true image distribution is $\widehat{\mathbb{P}}_x \sim \mathscr{N}\left(0, \widehat{\Sigma}_x\right)$. Given the nominal distribution $\widehat{\mathbb{P}} = \widehat{\mathbb{P}}_x \times \widehat{\mathbb{P}}_w$, the classical MMSE estimator $\widehat{A} \in \mathbb{R}^{2^N \times 2^N}$ has the form

$$\widehat{A}_{ii} = \frac{(\widehat{\Sigma}_x)_{ii}}{(\widehat{\Sigma}_x)_{ii} + \widehat{\sigma}_w^2} \quad i = 1,\dots 2^N, \text{ and } \quad \widehat{A}_{ij} = 0 \quad i = 1,\dots,2^N, j = 1,\dots 2^N, i \neq j,$$

and the wavelet coefficient estimate of the true image are $\widehat{x}_i = \widehat{A}_{ii} \times \widehat{y}_i$ for any $i = 1,\dots 2^N$. We would like to emphasize that $\widehat{A}_{ii} < 1$ because $\widehat{\sigma}_w > 0$, thus the classical MMSE estimator already provides the shrinking effect for the wavelet coefficient estimation problem.

Alternatively, one may construct a denoising filter $\psi^\star_{\widehat{y}}$ by resorting to the distributionally robust MMSE estimator, which is equivalent to solving the following estimation problem

$$\inf_{\psi \in \mathscr{F}} \sup_{\mathbb{Q} \in \mathbb{B}(\widehat{\mathbb{P}}_{\widehat{y}})} \mathscr{R}(\psi, \mathbb{Q}). \tag{2.55}$$

The wavelet coefficients of the true image will be estimated by $\psi^\star_{\widehat{y}}(\widehat{y})$ and the denoised image is $\mathscr{W}^\top \psi^\star_{\widehat{y}}(\widehat{y})$. There are two imporant points that are worth to notice in this framework. First, the nominal distribution $\widehat{\mathbb{P}}_{\widehat{y}}$ in (2.55) is a Gaussian distribution distributions. Second, the estimator is *dependent* on the observed data $\widehat{y}$ through the construction of the ambiguity set $\mathbb{B}_\phi(\widehat{\mathbb{P}}_{\widehat{y}})$, and then applied to the same data $\widehat{y}$ to obtain the denoised coefficients. This is in stark constrast with the models studied in the previous sections where the ambiguity set, and thus the estimator, is *independent* of the data which will be used for testing. Indeed, the construction of the ambiguity set involves prior information which often come in the form of observations similar to the possible unrevealed test data. Nevertheless, for the image processing applications, any images can be unique which prohibits the initialization of a sensible reference distribution and the construction of a one-size-fits-all MMSE estimator for the denoising purpose. As a result, it is more plausible to look for an *ad hoc* estimator for each observed image, and the observed images should be used for both the construction and the application of the estimator. Using observed data to construct the reference distribution for the signal and noise has been widely applied in image processing and signal processing [57]. Through a similar argument, in a small sample setting, all of the data should be used for training and the error is also estimated on the same data [35, Section 1].

Because $\widehat{\Sigma}_w \succ 0$ by construction, Theorem 2.15 applies and the estimator can be found by solving program (2.25). There are two main difficulties that hinder the applications of the Frank-Wolfe algorithm introduced in Section 2.6. First, the reference covariance matrix $\widehat{\Sigma}_x$ of the signal may be low-rank, and thus the Frank-Wolfe algorithm is only guaranteed to converge sublinearly. Second, the high image resolution can slow down the convergence speed of the algorithm due to the exponential up-scaling of the dimension. Fortunately, the image denoising setup entails some advantages which can be thoroughly exploited to provide speedup to the numerical method: first, the matrix H in (2.54) is the identity matrix I_n, and second, the covariance matrices $\widehat{\Sigma}_x$ and $\widehat{\Sigma}_w$ are both diagonal by construction. In the rest of this section, we present an efficient implementation of the Frank-Wolfe algorithm that exploits the structure of the denoising setting to solve (2.25) in a high dimensional setting.

Lemma 2.29. *If D and $\widehat{\Sigma}$ share the same eigenbasis, then the optimal solution $L^\star$ of program* (2.48) *also admits the same eigenbasis for any $\rho \geq 0$.*

Proof. For any $\delta \geq 0$, let $\widehat{\Sigma}_\delta \triangleq \widehat{\Sigma} + \delta I_d \succ 0$, thus $\widehat{\Sigma} = \widehat{\Sigma}_0$ and $\widehat{\Sigma}_\delta$ shares the same eigenbasis with

$\widehat{\Sigma}$ for any $\delta \ge 0$. Let $L^\star(\delta)$ be the optimal solution of the perturbed program

$$L^\star(\delta) \triangleq \begin{cases} \arg \min\limits_{L \in \mathbb{S}_+^d} & \langle D, L \rangle \\ \quad\text{s.t.} & \mathrm{Tr}\left[L + \widehat{\Sigma}_\delta - 2\left(\widehat{\Sigma}_\delta^{\frac{1}{2}} L \widehat{\Sigma}_\delta^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \le \rho^2. \end{cases}$$

By Berge's maximum theorem [11, pp. 115-116], $L^\star(\delta)$ is an upper hemicontinuous correspondence, and thus is closed [11, Theorem 6, pp. 112]. This implies that as δ tends to 0, the limit point of $L^\star(\delta)$ belongs to the set of optimal solutions of input $\widehat{\Sigma}_0$ [11, Theorem 4, pp. 111]. Because for any $\delta > 0$, $L^\star(\delta)$ shares the same eigenbasis with D and $\widehat{\Sigma}_\delta$, see [123, Theorem 5.1], the limit point of $L^\star(\delta)$ also shares the same eigenbasis because the feasible set is closed thanks to Lemma 2.36. This implies that there exists an optimal solution of program (2.48) for input $(\widehat{\Sigma}, D)$ which shares the same eigenbasis with $\widehat{\Sigma}$ and D. The claim thus follows. □

Proposition 2.30. *If $\widehat{\Sigma}_x, \widehat{\Sigma}_w$ are diagonal matrices and $H = I_n$, then the optimal solution $(\Sigma_x^\star, \Sigma_w^\star)$ of program* (2.25) *are also diagonal matrices.*

Proof. Suppose that the Frank-Wolfe algorithm developed in Section 2.6 is used to solve program (2.25). Because the algorithm is initialized at $\widehat{\Sigma}_x$ and $\widehat{\Sigma}_w$ which are diagonal, it is straightforward to see that (D_x, D_w) are diagonal matrices from the definition. By applying Lemma 2.29, we conclude that $\Sigma_x^{(k)}$ and $\Sigma_w^{(k)}$ are also diagonal matrices at any iteration $k \in \mathbb{N}$. Because the Frank-Wolfe algorithm produces an approximate solution of problem (2.25) at any accuracy level, we thus can employ a continuity argument to conclude that there exists, in the limit as the accuracy level goes to zero, an optimal solution of program (2.25) which is diagonal. □

As a result of Lemma 2.30, the Frank-Wolfe algorithm can be modified to keep track of only the diagonal elements of the matrices, and thus, all matrix multiplications can be reduced to vector multiplications. The algorithm used for the wavelet coefficient shrinkage is hence more efficient in both memory usage and computational complexity. It has been widely observed that the wavelet coefficients are usually sparse, thus we are interested in constructing an estimator which has more prominent shrinking effect than the classical MMSE estimator. One such situation is highlighted in the following lemma.

Proposition 2.31. *If $\rho_x = 0$ and $\rho_w > 0$, then the Wasserstein MMSE estimator is equivalent to a shrinkage estimator of the wavelet coefficients, i.e., $A_{ii}^\star \le \widehat{A}_{ii}\; \forall i = 1, \ldots, 2^N$.*

Proof. Because the nominal distribution $\widehat{\mathbb{P}}$ is a Gaussian distribution of the form (2.23), Corollary 2.20 affirms that the optimal MMSE estimator can be recovered from the optimal solution $(\Sigma_x^\star, \Sigma_w^\star)$ of program (2.25). Because $\rho_x = 0$, we have $\Sigma_x^\star = \widehat{\Sigma}_x$. Because $(\Sigma_w^\star)_{ii} > (\widehat{\Sigma}_w)_{ii}$ for all i, we conclude that $A_{ii}^\star \le \widehat{A}_{ii}\; \forall i = 1, \ldots, 2^N$. □

Table 2.1 – Average PSNR over 10 independent trials

Noise	Variance	Noisy	RiskShrink	SureShrink	BayesShrink	MMSE	DRO
normal	$\sigma = 10$	27.78	31.04	32.19	32.88	33.58	33.66
	$\sigma = 15$	24.26	29.26	29.97	30.94	31.15	31.32
	$\sigma = 20$	21.76	28.00	28.46	29.64	29.36	29.58
	$\sigma = 25$	19.82	27.03	27.31	28.69	27.95	28.20
	$\sigma = 30$	18.24	26.30	26.39	27.93	26.70	27.08
logistic	$\sigma = 10$	27.78	31.07	32.16	32.87	33.40	33.53
	$\sigma = 15$	24.27	29.25	29.90	30.94	30.86	31.10
	$\sigma = 20$	21.77	27.96	28.35	29.66	29.04	29.32
	$\sigma = 25$	19.83	26.96	27.16	28.69	27.55	27.93
	$\sigma = 30$	18.24	26.18	26.19	27.96	26.29	26.80
laplace	$\sigma = 10$	27.78	31.05	32.05	32.79	33.14	33.34
	$\sigma = 15$	24.26	29.17	29.70	30.83	30.50	30.81
	$\sigma = 20$	21.77	27.82	28.03	29.50	28.52	28.95
	$\sigma = 25$	19.82	26.75	26.71	28.49	26.92	27.51
	$\sigma = 30$	18.23	25.90	25.71	27.73	25.62	26.33
t	$\sigma = 10$	27.77	30.77	31.82	32.49	32.68	32.98
	$\sigma = 15$	24.23	28.66	29.32	30.32	29.87	30.28
	$\sigma = 20$	21.77	27.21	27.66	28.92	27.89	28.54
	$\sigma = 25$	19.82	25.97	26.26	27.77	26.23	27.03
	$\sigma = 30$	18.23	24.91	25.12	26.79	24.83	25.73
pink	$\sigma = 10$	27.78	30.30	30.83	30.41	30.94	31.40
	$\sigma = 15$	24.27	27.83	27.90	27.56	27.79	28.59
	$\sigma = 20$	21.77	25.99	25.73	25.48	25.51	26.47
	$\sigma = 25$	19.82	24.82	24.13	23.84	23.72	24.77
	$\sigma = 30$	18.24	23.92	22.72	22.45	22.23	23.33
brown	$\sigma = 10$	27.78	27.94	28.30	28.05	28.22	28.43
	$\sigma = 15$	24.26	24.99	24.86	24.65	24.77	25.11
	$\sigma = 20$	21.77	22.85	22.41	22.23	22.32	22.71
	$\sigma = 25$	19.83	21.06	20.49	20.32	20.38	20.78
	$\sigma = 30$	18.25	19.72	18.96	18.80	18.85	19.29

We emphasize that our Wasserstein affine estimator is not an independent coefficient shrinkage method. Indeed, the optimal estimator will have the form

$$A^\star_{ii} = \frac{(\widehat{\Sigma}_x)_{ii}}{(\widehat{\Sigma}_x)_{ii} + (\Sigma^\star_w)_{ii}} \quad \forall i = 1, \ldots, 2^N,$$

where in general we have $(\Sigma^\star_w)_{ii} \neq (\Sigma^\star_w)_{jj}$ for $i \neq j$. In addition, $(\Sigma^\star_w)_{ii}$ and $(\Sigma^\star_w)_{jj}$ are dependent because they are extracted from the optimal solution $\Sigma^\star_w$, which in turns depends on all diagonal elements of $\widehat{\Sigma}_w$ and $\widehat{\Sigma}_x$.

To showcase the capability of our proposed model we consider several different cases where the noise does not follow Gaussian distributions. In specific, we consider the Laplace, Logistic, and t distributions to generate the random independent noise. We also consider the colored noise for the dependent noise. The color of a noise refers to its power spectrum with different colors of noise have significantly different properties: for example, as audio signals they will sound different to human ears, and as images they will have a visibly different texture. Many of these noises assume a signal with components at all frequencies, with a power spectral density per unit of bandwidth proportional to $1/f^{\beta}$ and hence they are examples of power-law noise. For instance, the spectral density of white noise is flat ($\beta = 0$), while pink noise has $\beta = 1$, and Brownian noise has $\beta = 2$. The denoising result of the corrupted Lena image is presented in Table 2.1.

2.8 Appendix

2.8.1 Auxiliary Results

In order to prove Proposition 2.7 in the main text, we establish here general conditions for the solvability and stability of parametric minimax problems.

Lemma 2.32 (Parametric minimax problems)**.** *Consider the parametric minimax problem*

$$J(\theta) \triangleq \inf_{u \in U(\theta)} \sup_{v \in V(\theta)} f(u, v, \theta),$$

where $\mathbb{U}$, $\mathbb{V}$ and Θ are metric spaces, $f : \mathbb{U} \times \mathbb{V} \times \Theta \to \mathbb{R}$ is a continuous function, $U : \Theta \rightrightarrows \mathbb{U}$ is a continuous multifunction and $V : \Theta \rightrightarrows \mathbb{V}$ is a compact-valued continuous multifunction. If there exist $\varepsilon, \delta, \bar{J} > 0$ and a continuous function $v : \Theta \cap \mathscr{B}_\delta(\theta_0) \to \mathbb{V}$ such that $v(\theta) \in V(\theta)$, $|J(\theta)| \le \bar{J}$ and $U_\varepsilon(\theta) \triangleq \{u \in U(\theta) : f(u, v(\theta), \theta) \le \bar{J} + \varepsilon\}$ is compact for every $\theta \in \Theta \cap \mathscr{B}_\delta(\theta_0)$, then the minimax problem is solvable for $\theta = \theta_0$, and $J(\theta)$ is continuous at $\theta = \theta_0$.

Proof. As $J(\theta)$ is finite and $\varepsilon > 0$, the outer minimization problem admits an ε-optimal solution $u_\varepsilon(\theta)$ for every parameter $\theta \in \Theta \cap \mathscr{B}_\delta(\theta_0)$. Thus, we have

$$f(u_\varepsilon(\theta), v(\theta), \theta) \le \sup_{v \in V(\theta)} f(u_\varepsilon(\theta), v, \theta) \le \inf_{u \in U(\theta)} \sup_{v \in V(\theta)} f(u, v, \theta) + \varepsilon \le \bar{J} + \varepsilon,$$

which implies that $u_\varepsilon(\theta) \in U_\varepsilon(\theta)$ for every $\theta \in \Theta \cap \mathscr{B}_\delta(\theta_0)$. Without loss of generality, we can thus restrict the feasible set $U(\theta)$ to $U_\varepsilon(\theta)$, that is, for any $\theta \in \Theta \cap \mathscr{B}_\delta(\theta_0)$ we may reformulate the minimax problem as

$$J(\theta) = \inf_{u \in U_\varepsilon(\theta)} \sup_{v \in V(\theta)} f(u, v, \theta).$$

Note that the optimal value function $F(u, \theta) \triangleq \sup_{v \in V(\theta)} f(u, v, \theta)$ of the inner maximization problem is continuous on $\mathbb{U} \times (\Theta \cap \mathscr{B}_\delta(\theta_0))$ by virtue of Berge's theorem [11, pp. 115–116], which applies because f is jointly continuous in all of its arguments, while V is a compact-valued continuous multifunction. For $\theta = \theta_0$, the reformulated minimax problem thus minimizes the continuous function $F(u,\theta_0)$ over all u in the compact set $U_\varepsilon(\theta_0)$ and is thus solvable.

The compact-valued multifunction U_ε is continuous on $\Theta \cap \mathscr{B}_\delta(\theta_0)$ by the continuity of $f(u, v(\theta), \theta)$. Applying Berge's theorem once again thus guarantees that $J(\theta) = \min_{u \in U_\varepsilon(\theta)} F(u,\theta)$ is continuous at $\theta = \theta_0$. □

In order to derive a tractable reformulation for the Gelbrich MMSE estimation problem studied in Section 2.2, we need to be able to solve nonlinear SDPs of the form

$$J^\star \triangleq \sup_{\Sigma \succeq 0} \; \langle D, \Sigma \rangle - \gamma \operatorname{Tr}\left[\Sigma - 2\big(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right] \tag{2.56}$$

parameterized by $D \in \mathbb{S}^d$, $\widehat{\Sigma} \in \mathbb{S}_+^d$ and $\gamma \in \mathbb{R}_+$. It is known that, under certain regularity conditions, problem (2.56) admits a unique optimal solution that is available in closed form [123]. In the following we review the construction of this optimal solution under slightly more general conditions.

Proposition 2.33 (Closed form solution of (2.56))**.** *For any $D \in \mathbb{S}^d$, $\widehat{\Sigma} \in \mathbb{S}_+^d$ and $\gamma \in \mathbb{R}_+$ the optimal value of the nonlinear SDP* (2.56) *is given by*

$$J^\star = \begin{cases} \gamma^2 \big\langle (\gamma I_d - D)^{-1}, \widehat{\Sigma} \big\rangle & \text{if } \gamma > \lambda_{\max}(D), \\ \liminf\limits_{\bar{\gamma} \downarrow \gamma} \bar{\gamma}^2 \big\langle (\bar{\gamma} I_d - D)^{-1}, \widehat{\Sigma} \big\rangle & \text{if } \gamma = \lambda_{\max}(D), \\ +\infty & \text{if } \gamma < \lambda_{\max}(D). \end{cases}$$

Moreover, problem (2.56) *is solved by* $\Sigma^\star = \gamma^2 (\gamma I_d - D)^{-1} \widehat{\Sigma} (\gamma I_d - D)^{-1}$ *whenever* $\gamma > \lambda_{\max}(D)$. *This solution is unique if* $\widehat{\Sigma} \succ 0$.

Proof. Assume first that $\gamma > \lambda_{\max}(D)$. Moreover, in order to simplify the proof, assume temporarily that $\widehat{\Sigma} \succ 0$. By applying the nonlinear variable transformation $B \leftarrow (\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}$, which implies that $\Sigma = \widehat{\Sigma}^{-\frac{1}{2}} B^2 \widehat{\Sigma}^{-\frac{1}{2}}$, we can reformulate problem (2.56) as

$$\begin{aligned} J^\star &= \sup_{B \succeq 0} \; \big\langle D, \widehat{\Sigma}^{-\frac{1}{2}} B^2 \widehat{\Sigma}^{-\frac{1}{2}} \big\rangle - \gamma \operatorname{Tr}\left[\widehat{\Sigma}^{-\frac{1}{2}} B^2 \widehat{\Sigma}^{-\frac{1}{2}} - 2B\right] \\ &= \sup_{B \succeq 0} \; \big\langle B^2, \widehat{\Sigma}^{-\frac{1}{2}} (D - \gamma I_d) \widehat{\Sigma}^{-\frac{1}{2}} \big\rangle + 2\gamma \big\langle B, I_d \big\rangle, \end{aligned}$$

where the second equality exploits the cyclicity of the trace operator. Introducing the auxiliary parameter $\Delta \triangleq \widehat{\Sigma}^{-\frac{1}{2}} (D - \gamma I_d) \widehat{\Sigma}^{-\frac{1}{2}}$, we can then rewrite the last maximization problem over B more concisely as

$$J^\star = \sup_{B \succeq 0} \; \big\langle B^2, \Delta \big\rangle + 2\gamma \big\langle B, I_d \big\rangle. \tag{2.57}$$

Note that (2.57) represents a convex maximization problem because $\gamma > \lambda_{\max}(D)$ and $\widehat{\Sigma} \succ 0$, which imply that $\Delta \prec 0$. Ignoring the positive semidefiniteness constraint on B, the objective function of (2.57) is uniquely minimized by the solution $B^\star = -\gamma\Delta^{-1}$ of the first-order optimality condition $B\Delta + \Delta B + 2\gamma I_d = 0$. Uniqueness follows from [83, Theorem 12.5]. As it is strictly positive definite, $B^\star$ thus uniquely solves (2.57), which in turn implies that $\Sigma^\star = \widehat{\Sigma}^{-\frac{1}{2}}(B^\star)^2\widehat{\Sigma}^{-\frac{1}{2}} = \gamma^2(\gamma I_d - D)^{-1}\widehat{\Sigma}(\gamma I_d - D)^{-1}$ uniquely solves (2.56). Substituting $\Sigma^\star$ back into the objective function of (2.56) further shows that $J^\star = \gamma^2\big\langle(\gamma I_d - D)^{-1}, \widehat{\Sigma}\big\rangle$.

Next, we will argue that the analytical formula for $J^\star$ in the regime $\gamma > \lambda_{\max}(D)$ remains valid even when $\widehat{\Sigma}$ is rank-deficient. To see this, define

$$J^\star(\widehat{\Sigma}) \triangleq \gamma^2\big\langle(\gamma I_d - D)^{-1}, \widehat{\Sigma}\big\rangle \quad \text{and} \quad \Sigma^\star(\widehat{\Sigma}) \triangleq \gamma^2(\gamma I_d - D)^{-1}\widehat{\Sigma}(\gamma I_d - D)^{-1}$$

as explicit continuous functions of the parameter $\widehat{\Sigma} \in \mathbb{S}_+^d$. Similarly, define the function

$$F(\Sigma, \widehat{\Sigma}) \triangleq \langle D, \Sigma\rangle - \gamma \operatorname{Tr}\big[\Sigma - 2\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\big],$$

which is jointly continuous in $\Sigma \in \mathbb{S}_+^d$ and $\widehat{\Sigma} \in \mathbb{S}_+^d$. We then have

$$J^\star(\widehat{\Sigma}) = \liminf_{\varepsilon\downarrow 0} J^\star(\widehat{\Sigma} + \varepsilon I_d) = \liminf_{\varepsilon\downarrow 0}\sup_{\Sigma\succeq 0} F(\Sigma, \widehat{\Sigma} + \varepsilon I_d) \geq \sup_{\Sigma\succeq 0} F(\Sigma, \widehat{\Sigma}) \geq F(\Sigma^\star(\widehat{\Sigma}), \widehat{\Sigma}) = J^\star(\widehat{\Sigma}),$$

where the first equality follows from the continuity of $J^\star(\widehat{\Sigma})$, while the second equality holds because $\widehat{\Sigma} + \varepsilon I_d \succ 0$ for every $\varepsilon > 0$ and because $J^\star(\widehat{\Sigma}') = \sup_{\Sigma\succ 0} F(\Sigma, \widehat{\Sigma}')$ for every $\widehat{\Sigma}' \succ 0$, which was established in the first part of the proof. The first inequality exploits the fact that a pointwise supremum of continuous functions is is lower semicontinuous, and the second inequality holds because $\Sigma^\star(\widehat{\Sigma} + \varepsilon I_d) \succ 0$ for every $\varepsilon > 0$. Finally, the last equality follows from elementary algebra. The above arguments imply that $J^\star(\widehat{\Sigma})$ and $\Sigma^\star(\widehat{\Sigma})$ represent the optimal value and an optimal solution of (2.56), respectively, even if $\widehat{\Sigma} \in \mathbb{S}_+^d$ is rank-deficient.

Assume next that $\gamma < \lambda_{\max}(D)$, and denote by $\overline{v} \in \mathbb{R}^d$ a normalized eigenvector of D corresponding to the eigenvalue $\lambda_{\max}(D)$. By optimizing only over matrices of the form $\Sigma = t\overline{v}\,\overline{v}^\top$ for some $t \geq 0$, we find

$$\begin{aligned} J^\star &\geq \sup_{t\geq 0}\; t\big\langle D - \gamma I_d, \overline{v}\,\overline{v}^\top\big\rangle + 2\sqrt{t}\gamma \operatorname{Tr}\big[\big(\widehat{\Sigma}^{\frac{1}{2}}\overline{v}\,\overline{v}^\top\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\big] \\ &= \sup_{t\geq 0}\; t\,(\lambda_{\max}(D) - \gamma) + 2\sqrt{t}\gamma \operatorname{Tr}\big[\big(\widehat{\Sigma}^{\frac{1}{2}}\overline{v}\,\overline{v}^\top\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\big] = \infty. \end{aligned}$$

Assume finally that $\gamma = \lambda_{\max}(D)$. To investigate this limiting case, note that the objective function of (2.56) is linear in γ, which implies that the optimal value of (2.56) is convex and lower semicontinuous in γ. Given the results for $\gamma \neq \lambda_{\max}(D)$, it is thus clear that for $\gamma = \lambda_{\max}(D)$ the optimal value of (2.56) must be given by $J^\star = \liminf_{\bar{\gamma}\downarrow\gamma}\bar{\gamma}^2\big\langle(\bar{\gamma} I_d - D)^{-1}, \widehat{\Sigma}\big\rangle$. This observation completes the proof. □

In order to derive search directions for the Frank-Wolfe algorithm developed in Section 2.6, we need to be able to solve constrained nonlinear SDPs of the form

$$\begin{array}{ll} \sup\limits_{\Sigma \succeq 0} & \langle D, \Sigma \rangle \\ \text{s.t.} & \operatorname{Tr}\left[\Sigma + \widehat{\Sigma} - 2\left(\widehat{\Sigma}^{\frac{1}{2}} \Sigma c \widehat{\Sigma}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \leq \rho^2 \end{array} \tag{2.58}$$

parameterized by $D \in \mathbb{S}^d$, $\widehat{\Sigma} \in \mathbb{S}_+^d$ and $\rho \in \mathbb{R}_+$. It is known that problem (2.58) admits a unique optimal solution that is available in quasi-closed form [123]. Below we review the construction of this optimal solution under more general conditions and uncover several previously unknown properties of this solution.

Proposition 2.34 (Quasi-closed form solution of (2.58))**.** *The following statements hold.*

(i) *If $D \in \mathbb{S}^d$, $\widehat{\Sigma} \in \mathbb{S}_+^d$ and $\rho \in \mathbb{R}_+$, then problem* (2.58) *is solvable, and its maximum matches that of the univariate convex minimization problem*

$$\inf_{\substack{\gamma \geq 0 \\ \gamma > \lambda_{\max}(D)}} \gamma\left(\rho^2 + \langle \gamma(\gamma I_d - D)^{-1} - I_d, \widehat{\Sigma} \rangle\right). \tag{2.59}$$

(ii) *If $D \neq 0$, $\widehat{\Sigma} \succ 0$ and $\rho > 0$, then problem* (2.59) *has a unique minimizer $\gamma^\star \in (\lambda_{\max}(D), \infty)$, and problem* (2.58) *is solved by $\Sigma^\star = {\gamma^\star}^2 (\gamma^\star I_d - D)^{-1} \widehat{\Sigma} (\gamma^\star I_d - D)^{-1}$.*

(iii) *If $D \succeq 0$, $D \neq 0$, $\widehat{\Sigma} \succ 0$ and $\rho > 0$, then $\gamma^\star$ is the unique solution of the algebraic equation*

$$\rho^2 - \langle \widehat{\Sigma}, (I_d - \gamma^\star (\gamma^\star I_d - D)^{-1})^2 \rangle = 0, \tag{2.60}$$

and $\Sigma^\star$ is the unique maximizer of (2.58). *Moreover, the Gelbrich distance constraint in* (2.58) *is binding at $\Sigma^\star$, and we have $\Sigma^\star \succ \lambda_{\min}(\widehat{\Sigma}) I_d$.*

Proof. As for assertion (i), note that the Lagrangian dual of (2.58) can be represented as

$$\inf_{\gamma \geq 0} \sup_{\Sigma \succeq 0} \langle D, \Sigma \rangle - \gamma \operatorname{Tr}\left[\Sigma + \widehat{\Sigma} - 2\left(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] + \gamma \rho^2. \tag{2.61}$$

Strong duality as well as primal solvability follow from [13, Proposition 5.5.4], which applies because the primal problem (2.58) has a continuous objective function and—by virtue of Lemma 2.36 below—a nonempty, compact and convex feasible set. The postulated reformulation (2.59) then follows immediately from replacing the supremum of the inner maximization problem in (2.61) with the analytical formula derived in Proposition 2.33. We emphasize that for $\gamma = \lambda_{\max}(D)$, depending on the problem data, the inner supremum in (2.61) may evaluate to any nonnegative real number or to $+\infty$. In order to avoid cumbersome case distinctions, we thus exclude the point $\gamma = \lambda_{\max}(D)$ from the feasible set of (2.59) without affecting the problem's infimum.

As for assertion (ii), note that $\Sigma = \widehat{\Sigma}$ represents a Slater point for the primal problem (2.58) because $\rho > 0$. Thus, the dual problem (2.61) is solvable by [13, Proposition 5.3.1]. To prove that (2.59) is also solvable, it remains to be shown that (2.61) does not attain its maximum at the boundary point $\gamma = \lambda_{\max}(D)$, which has been excluded from the feasible set of (2.59). This is the case, however, because of the assumption that $\widehat{\Sigma} \succ 0$ and $D \neq 0$, which ensures that the objective function value of $\gamma = \lambda_{\max}(D)$ in (2.61) amounts to $+\infty$. We may thus conclude that (2.59) admits a minimizer $\gamma^\star \in (\lambda_{\max}(D), \infty)$. This mimnimizer is unique because the objective function of (2.59) is strictly convex when $\widehat{\Sigma} \succ 0$. Finally, the Karush-Kuhn-Tucker optimality conditions [13, Proposition 5.3.2] imply that any solution of the primal problem (2.58) also solves the inner maximization problem in (2.61) at $\gamma = \gamma^\star$. The formula for $\Sigma^\star$ thus follows from Proposition 2.33.

To prove assertion (iii), note that the assumptions $D \succeq 0$ and $D \neq 0$ imply that $\gamma^\star > \lambda_{\max}(D) > 0$. Therefore, none of the constraints in (2.59) are binding at optimality. As the objective function of (2.59) is smooth and strictly convex, $\gamma^\star$ is thus uniquely determined by the first-order optimality condition (2.60), which forces the gradient of the objective function to vanish. The uniqueness of $\Sigma^\star$ follows from the uniqueness of $\gamma^\star$ and the uniqueness of the inner maximizer in (2.61); see Proposition 2.33. Moreover, as $\gamma^\star > 0$, the Gelbrich distance constraint in (2.59) is binding at $\Sigma^\star_\rho$ due to complementary slackness. Finally, we have

$$
\begin{aligned}
\frac{1}{\lambda_{\min}(\Sigma^\star)} &= \lambda_{\max}\big((\Sigma^\star)^{-1}\big) = \lambda_{\max}\Big(\Big(\gamma^{\star 2}(\gamma^\star I_n - D)^{-1}\widehat{\Sigma}(\gamma^\star I_d - D)^{-1}\Big)^{-1}\Big) \\
&= \lambda_{\max}\Big(\big(I_d - D/\gamma^\star)\widehat{\Sigma}^{-1}\big(I_d - D/\gamma^\star)\Big) \\
&\leq \lambda_{\max}(I_d - D/\gamma^\star)^2\, \lambda_{\max}(\widehat{\Sigma}^{-1}) < \lambda_{\max}(\widehat{\Sigma}^{-1}) = \frac{1}{\lambda_{\min}(\widehat{\Sigma})},
\end{aligned}
$$

where the strict inequality holds because $\gamma^\star > \lambda_{\max}(D) > 0$. We may thus conclude that the smallest eigenvalue of $\Sigma^\star$ exceeds the smallest eigenvalue of $\widehat{\Sigma}$. This observation concludes the proof. □

In Sections 2.3 and 2.4 we repeatedly encounter nonlinear SDPs of the form

$$
\begin{array}{ll}
\sup\limits_{\Sigma \succeq 0} & \inf\limits_{L \in \mathcal{C}} \ \langle L^\top L, \Sigma \rangle + f(L) \\
\text{s.t.} & \operatorname{Tr}\Big[\Sigma + \widehat{\Sigma} - 2\Big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\Big)^{\frac{1}{2}}\Big] \leq \rho^2
\end{array}
\tag{2.62}
$$

parameterized by $\widehat{\Sigma} \in \mathbb{S}^d_+$ and $\rho \in \mathbb{R}_+$, where $\mathcal{C} \subseteq \mathbb{R}^{\ell \times d}$ is a convex set and $f : \mathcal{C} \to \mathbb{R}$ a convex continuous function. Problem (2.62) is reminiscent of (2.58) but accommodates a nonlinear convex objective function. We do not attempt to characterize the maximizers of (2.62) for arbitrary choices of $\mathcal{C}$ and f, but we can prove that there is at least one well-behaved maximizer that is bounded away from 0.

Lemma 2.35 (Structural properties of the maximizers of (2.62))**.** *Assume that $\widehat{\Sigma} \in \mathbb{S}^d_+$ and $\rho \in \mathbb{R}_+$. If $\mathcal{C} \subseteq \mathbb{R}^{\ell \times d}$ is a nonempty convex set and $f : \mathcal{C} \to \mathbb{R}$ is a convex continuous function,* *then the nonlinear SDP* (2.62) *admits a maximizer* $\Sigma^\star \succeq \lambda_{\min}(\widehat{\Sigma}) I_d$.

Proof. Note that if $\rho = 0$ or $\lambda_{\min}(\widehat{\Sigma}) = 0$, then the claim holds trivially. Thus, we may henceforth assume without loss of generality that $\rho > 0$ and $\widehat{\Sigma} \succ 0$. Denoting the feasible set of (2.62) by

$$\mathscr{S} \triangleq \left\{ \Sigma \in \mathbb{S}_+^d : \operatorname{Tr}\left[\Sigma + \widehat{\Sigma} - 2\big(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right] \leq \rho^2 \right\},$$

we then find

$$\begin{aligned} \sup_{\Sigma \in \mathscr{S}} \inf_{L \in \mathscr{C}} \left\langle L^\top L, \Sigma \right\rangle + f(L) &= \inf_{L \in \mathscr{C}} \sup_{\Sigma \in \mathscr{S}} \left\langle L^\top L, \Sigma \right\rangle + f(L) \\ &= \inf_{L \in \mathscr{C}} \sup_{\substack{\Sigma \in \mathscr{S} \\ \Sigma \succeq \lambda_{\min}(\widehat{\Sigma}) I_d}} \left\langle L^\top L, \Sigma \right\rangle + f(L) = \sup_{\substack{\Sigma \in \mathscr{S} \\ \Sigma \succeq \lambda_{\min}(\widehat{\Sigma}) I_d}} \inf_{L \in \mathscr{C}} \left\langle L^\top L, \Sigma \right\rangle + f(L), \end{aligned}$$

where the first and the third equality follow from Sion's minimax theorem [156], which applies because $\left\langle L^\top L, \Sigma \right\rangle$ is convex and continuous in L for every fixed $\Sigma \succeq 0$ and because $\mathscr{S}$ is convex and compact by virtue of Lemma 2.36. The second equality follows readily from Proposition 2.34 (iii). The last maximization problem in the above expression has a solution $\Sigma^\star \succeq \lambda_{\min}(\widehat{\Sigma}) I_d$ because its feasible set is compact and its objective function is upper semicontinuous. Clearly, $\Sigma^\star$ also solves (2.62), and thus the claim follows. □

The proofs of Proposition 2.58 and Lemma 2.35 rely on the following auxiliary result.

Lemma 2.36 (Compactness of the feasible set [152, Lemma A.6])**.** *For any* $\widehat{\Sigma} \in \mathbb{S}_+^d$ *and* $\rho \in \mathbb{R}_+$, *the set*

$$\mathscr{S} \triangleq \left\{ \Sigma \in \mathbb{S}_+^d : \operatorname{Tr}\left[\Sigma + \widehat{\Sigma} - 2\big(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right] \leq \rho^2 \right\}$$

is convex and compact. Moreover, for any $\Sigma \in \mathscr{S}$ *we have* $\operatorname{Tr}\left[\Sigma\right] \leq \bar{\sigma}$ *and* $\Sigma \preceq \bar{\sigma} I_d$, *where* $\bar{\sigma} \triangleq (\rho + \operatorname{Tr}[\widehat{\Sigma}]^{\frac{1}{2}})^2$.

2.8.2 Taylor Expansion of the Objective Function

In this section, we provide the detailed steps leading to the Taylor expansion of the objective function f of the semidefinite program (2.25). We use $\otimes$ to denote the Kronecker product of two matrices [12, Definition 7.1.2]. We remind that

$$f(\Sigma_x, \Sigma_w) = \operatorname{Tr}\left[\Sigma_x - \Sigma_x H^\top \left(H \Sigma_x H^\top + \Sigma_w\right)^{-1} H \Sigma_x\right]$$

for any $\Sigma_x \in \mathbb{S}_+^n$, $\Sigma_w \in \mathbb{S}_+^m$. The following lemma serves as a useful tool to shorten the notations.

Lemma 2.37 ([12, Fact 7.4.9])**.** *For any matrices* A, B, C, D *of appropriate dimensions, we have*

$$\operatorname{Tr}\left[ABCD\right] = \operatorname{vec}(A)^\top (B \otimes D^\top) \operatorname{vec}(C^\top).$$

At any feasible solution (Σ_x, Σ_w) with $\Sigma_w \succ 0$, let $\Delta_x \in \mathbb{S}^n$, $\Delta_w \in \mathbb{S}^m$ be two symmetric perturbation matrices and let $G \triangleq H\Sigma_x H^\top + \Sigma_w \in \mathbb{S}_+^m$, the matrix inverse term of the objective function f can be approximated to the second order by

$$\begin{aligned}
&\left(H(\Sigma_x + \Delta_x)H^\top + \Sigma_w + \Delta_w\right)^{-1} \\
=&\left(G^{\frac{1}{2}}(I_m + G^{-\frac{1}{2}}(H\Delta_x H^\top + \Delta_w)G^{-\frac{1}{2}})G^{\frac{1}{2}}\right)^{-1} \\
=&G^{-\frac{1}{2}}(I_m + G^{-\frac{1}{2}}(H\Delta_x H^\top + \Delta_w)G^{-\frac{1}{2}})^{-1}G^{-\frac{1}{2}} \\
=&\, G^{-\frac{1}{2}}\left(I_m - G^{-\frac{1}{2}}(H\Delta_x H^\top + \Delta_w)G^{-\frac{1}{2}} + G^{-\frac{1}{2}}(H\Delta_x H^\top + \Delta_w)G^{-1}(H\Delta_x H^\top + \Delta_w)G^{-\frac{1}{2}} + \mathcal{O}(\|\Delta\|^3)\right)G^{-\frac{1}{2}} \\
=&\, G^{-1} - G^{-1}(H\Delta_x H^\top + \Delta_w)G^{-1} + G^{-1}(H\Delta_x H^\top + \Delta_w)G^{-1}(H\Delta_x H^\top + \Delta_w)G^{-1} + \mathcal{O}(\|\Delta\|^3),
\end{aligned}$$

where the third equality is the result of the second order approximation of matrix inversion [12, Proposition 9.4.13]. Hence, the second order approximation of the objective function f can be written as

$$\begin{aligned}
&f(\Sigma_x + \Delta_x, \Sigma_w + \Delta_w) \\
=&\operatorname{Tr}\left[\Sigma_x + \Delta_x - (\Sigma_x + \Delta_x)H^\top \left(H(\Sigma_x + \Delta_x)H^\top + \Sigma_w + \Delta_w\right)^{-1} H(\Sigma_x + \Delta_x)\right] \\
=&f(\Sigma_x, \Sigma_w) + \langle D_x, \Delta_x\rangle + \langle D_w, \Delta_w\rangle - \frac{1}{2}\begin{pmatrix}\operatorname{vec}(\Delta_x)\\ \operatorname{vec}(\Delta_w)\end{pmatrix}^\top \mathscr{H}\begin{pmatrix}\operatorname{vec}(\Delta_x)\\ \operatorname{vec}(\Delta_w)\end{pmatrix} + \mathcal{O}(\|\Delta\|^3),
\end{aligned}$$

where (D_x, D_w) is defined as

$$\begin{aligned}
D_x &= (I_n - \Sigma_x H^\top G^{-1}H)^\top (I_n - \Sigma_x H^\top G^{-1}H), \\
D_w &= G^{-1}H\Sigma_x^2 H^\top G^{-1},
\end{aligned}$$

and the negative Hessian matrix $\mathscr{H}$ is defined as

$$\mathscr{H} = \begin{bmatrix}\mathscr{H}_{xx} & \mathscr{H}_{xw}\\ \mathscr{H}_{xw}^\top & \mathscr{H}_{ww}\end{bmatrix}$$

with

$$\begin{aligned}
\mathscr{H}_{xx} &= 2D_x \otimes H^\top G^{-1}H \\
\mathscr{H}_{xw} &= H^\top G^{-1} \otimes (H^\top D_w - \Sigma_x H^\top G^{-1}) + (H^\top D_w - \Sigma_x H^\top G^{-1}) \otimes H^\top G^{-1} \\
\mathscr{H}_{ww} &= 2D_w \otimes G^{-1}
\end{aligned}$$

thanks to Lemma 2.37.

3 Distributionally Robust Risk Measure with Structured Ambiguity Sets

Truth is much too complicated to allow anything but approximations.
— John von Neumann

We formally introduce the Gelbrich hull, which is an outer approximation of the Wasserstein type-2 ambiguity set, and the corresponding Gelbrich risk, which is a conservative approximation of the Wasserstein risk. We provide theoretical insights about the Gelbrich hull and a decomposition of the Gelbrich risk as a two-layer optimization problem, which leads to a systematic approach to reformulate the Gelbrich risk. For a linear loss function, we show that the Gelbrich risk admits a closed-form expression for any family of consistent positive homogeneous and translation invariant risk measures. We provide the reformulation for the Gelbrich expected loss as a finite convex optimization problem for loss functions that can be expressed as the pointwise maximum of possibly non-convex quadratic functions. We further provide several extensions including the incorporation of support information and the generalization to the worst-case Value at Risk of nonlinear portfolios.

3.1 Introduction

Portfolio managers face the continuing challenges of optimally distributing their funds over a collection of possible assets to maximize the future return. Arguably, the seminal work of Markowitz [115] is one of the earliest mathematical models for portfolio optimization, in which the portfolio manager aims to maximize the expected return while at the same time minimizing the risk of a portfolio. While the risk of an investment in the Markowitz model is measured using its variance, it is well known that the stochastic return of an asset is rarely symmetric in practice, which renders the variance unsuitable to be used for portfolio allocation.

Numerous propositions for a more appropriate risk assessment have made research on downside risk measures a vibrant field spanning from economics to finance, with emerging applications in engineering, physics and operations research. The Value-at-Risk (VaR) of a portfolio, defined using the extremal quantile of its return distribution, has been serving as the industry standard to measure risk [93] and its used has been imposed by popular financial standards. Nevertheless, VaR suffers from a major theoretical pitfall which leads to heavy criticisms amid the 2008 financial crisis. VaR does not satisfy the subadditivity condition and thus it is not a convex risk measure. As a consequence, using VaR as a measure of risk may not necessarily promote portfolio diversification. In addition, portfolio allocation using VaR is notoriously challenging because VaR is not convex, and the resulting optimization problem that minimizes VaR is computationally intractable.

The Conditional Value-at-Risk (CVaR), which computes the average loss that exceeds the extremal quantile, has arisen as a favorable replacement of VaR. Contrary to VaR, CVaR is a convex risk measure and thus minimizing CVaR can be formulated as a convex optimization problem. Being a coherent risk measure, CVaR satisfies many desirable properties from the theoretical perspective. Moreover, CVaR is also a spectral risk measure and it can be expressed as the weighted average of returns and exhibits strong connection with risk aversion via its representation as an expectation of a piecewise linear utility function.

The expectation of a portfolio loss is also a spectral risk measure, despite the fact that it totally disregards the dispersion of the distribution. One can further convolve the expectation operator with a utility function to form the expected utility of the portfolio loss that better matches the risk attitude of portfolio managers.

The common problem with employing the distribution-based risk measures to assess the riskiness of an investment is that in reality precise knowledge of the underlying joint probability distribution of the asset returns are rarely available. In the best case, the portfolio managers have to leverage on historical data to estimate the distribution, which expose themselves to statistical risk. This problem has triggered the development of worst-case risk measures which are more robust to noise and misspecification of the asset returns' distribution. Instead of assessing the risk of a position with respect to a single distribution, the worst-case risk measure determines the worst possible risk over a set of candidate distributions that represents the portfolio managers' ambiguity with respect to the uncertain asset returns.

Worst-case risk measures where the ambiguity set is prescribed by the first two moments was originally proposed for VaR [52], CVaR [121, 26] and later on extended for spectral risk measure and coherent risk measure [110]. Worst-case expected loss for moment ambiguity set has also been widely studied in different contexts [39, 172]. Though intuitively appealing and admits closed form expression in many cases, estimating the first- and second-moments of the underlying random vector is difficult, or even impossible, when the available data is scarce. Alternatively, worst-case risk measures with (semi-)distance based ambiguity set was studied for Kullback-Leibler divergence [20] and phi-divergence [78]. The worst-case risk measures with Wasserstein ambiguity set was first studied in [138, 175] and was further extended to worst-case expected loss in [118, 16, 68, 178].

Motivated by these recent development, the starting point of this paper relies on the Wasserstein type-2 worst-case risk measures. While limited analytical solution exists for specific risk measure using the dual representation approach [175], this method will typically not generalize to a broader setting. For worst-case expected loss with the Wasserstein ambiguity set centered at an empirical distribution, tractable convex reformulations using duality techniques are available [179], nevertheless, these reformulations often result in an optimization problem that incurs a large number of auxiliary variables and constraints which prohibits its applications either in high-dimensional problems or in the big-data setting.

While calculating the Wasserstein type-2 distance between any two given distributions involves solving an optimization problem, there exists a generalized lower bound based on only the information about the first- and second-order moment of the two probability measures of interest [71]. Using this lower bound, which we term the Gelbrich distance, we construct a Gelbrich hull as a superset of the Wasserstein type-2 ambiguity set. As a consequence, instead of directly resolving the Wasserstein type-2 worst-case risk measures, this paper introduces the Gelbrich risk, a systematic conservative approximation of the Wasserstein risk. Throughout this paper, we demonstrate that the Gelbrich risk exhibits many nice properties, notably that its reformulation is more tractable than the reformulation for the original Wasserstein risk. Interestingly, we show that under certain conditions, the Gelbrich risk measure is a tight approximation of the Wasserstein risk measure.

The main contributions of this paper can be summarized as follows.

1. We introduce the Gelbrich hull, an outer approximation of the Wasserstein type-2 ambiguity set defined using the Gelbrich distance that takes into account only the information about the first two moments of the probability measures. We show that the Gelbrich hull is convex and closely related to the Chebyshev ambiguity set that is commonly used in the literature.

2. Using the Gelbrich approximation of the Wasserstein type-2 distance, we propose the Gelbrich risk which conservatively approximates the worst-case risk under the Wasser-

stein ambiguity set. We demonstrate structural properties of the Gelbrich risk, notably the decomposition of the Gelbrich risk as a two-layer optimization problem which facilitates a systematic reformulation of the Gelbrich risk as a finite dimensional convex optimization problem.

3. For a consistent family of law-invariant, positive homogeneous and translation invariant risk measures, we show that the corresponding Gelbrich risk of a linear loss admits a closed-form expression. This result paves the way for the generalization of the Gelbrich risk to any spectral risk measure and coherent risk measure.

4. We derive the semi-definite program reformulation for the Gelbrich expected loss for a wide range of loss functions that satisfy the quadratic growth condition. The reformulation of the Gelbrich expected loss typically involves a lower number of auxiliary variables and constraints, thus is supposed to be easier to solve than the Wasserstein expected loss.

5. We provide several extensions to generalize the Gelbrich risk by incorporating the support information, as well as deriving the Gelbrich Value-at-Risk for nonlinear portfolios.

The paper is structured as follows. Section 3.2 provides necessary background material for the construction of the Wasserstein risk. Section 3.3 proposes a principled approach to construct the Gelbrich hull that outer approximates the Wasserstein ambiguity set, and the related Gelbrich risk that safely approximates the Wasserstein risk. Section 3.4 studies the Gelbrich risk for linear (portfolio) loss, and Section 3.5 examines the Gelbrich expected loss for nonlinear loss function. Section 3.6 and 3.7 further expands the Gelbrich risk to different applications and to take into consideration the support information, and Section 3.8 concludes with a numerical application for the robust portfolio index tracking problem.

Notations. Throughout this paper, $\|\cdot\|$ denotes the 2-norm on the vector space. The space of all n-dimensional symmetric matrices is denoted by $\mathbb{S}^n$ and $\mathbb{S}^n_+$ ($\mathbb{S}^n_{++}$) denotes the space of symmetric positive semidefinite (definite, respectively) matrices. For any square matrix $A \in \mathbb{R}^{m\times m}$, the trace operator is defined as $\mathrm{Tr}\big[A\big] = \sum_{i=1}^n A_{ii}$, and for any $A \in \mathbb{S}^n$, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalue of A. For $N \in \mathbb{N}$, we set $[N] = \{1,\dots,N\}$.

3.2 Problem Statement

We provide in this section the preliminary elements that form the foundation for the theoretical development of the paper. We are equipped with a measurable space $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$, where $\mathscr{B}(\mathbb{R}^n)$ denotes the Borel algebra of $\mathbb{R}^n$, and a random vector ξ with values on $\mathbb{R}^n$. The set of all probability measures supported on $\mathbb{R}^n$ is denoted by $\mathscr{P}$. For a given probability measure $\mathbb{Q} \in \mathscr{P}$, the distribution of ξ under $\mathbb{Q}$ is uniquely determined by the cumulative distribution value $\mathbb{Q}(\xi \le \tau) = \mathbb{Q}(\xi_i \le \tau_i\ \forall i = 1,\dots,n)$ for any $\tau \in \mathbb{R}^n$. The first two moments of ξ under $\mathbb{Q}$ can be defined using the expectation operator as

$$\mathbb{E}_{\mathbb{Q}}[\xi] = \int_{\mathbb{R}^n} \xi \, \mathbb{Q}(\mathrm{d}\xi), \quad \text{and} \quad \mathbb{E}_{\mathbb{Q}}[\xi\xi^\top] = \int_{\mathbb{R}^n} \xi\xi^\top \, \mathbb{Q}(\mathrm{d}\xi).$$

We first review some common definitions of the distribution of ξ.

Definition 3.1 (Finite second moment distribution)**.** *The random vector ξ has finite second moment under $\mathbb{Q}$ if its expectation $\mathbb{E}_{\mathbb{Q}}[\xi]$ and its second moment matrix $\mathbb{E}_{\mathbb{Q}}[\xi\xi^\top]$ are both finite.*

Definition 3.2 (Symmetric distribution)**.** *The random vector ξ is (centrally) symmetric about $\mu \in \mathbb{R}^n$ under $\mathbb{Q}$ if for all vectors $\tau \in \mathbb{R}^n$, we have $\mathbb{Q}(\xi \geq \mu + \tau) = \mathbb{Q}(\xi \leq \mu - \tau)$.*

Definition 3.3 (Linear unimodal distribution)**.** *The random vector ξ is linearly unimodal about 0 under $\mathbb{Q}$ if for all vectors $w \in \mathbb{R}^n$, the cumulative distribution function of $w^\top\xi$ under $\mathbb{Q}$ is convex on $(-\infty, 0]$ and concave on $[0, +\infty)$*

Definition 3.4 (Elliptical distribution)**.** *The random vector ξ is (symmetric) elliptically distributed under $\mathbb{Q}$ if its characteristic function is given by*

$$\mathbb{E}_{\mathbb{Q}}[\exp(\sqrt{-1}\tau^\top\xi)] = \exp(\sqrt{-1}\tau^\top\mu)g(\tau^\top\Sigma\tau)$$

for some location parameter $\mu \in \mathbb{R}^n$, dispersion matrix $\Sigma \in \mathbb{S}^n_+$ and characteristic generator $g : \mathbb{R}_+ \to \mathbb{R}$.

The class of elliptical distributions generalizes common distributions such as the Gaussian distribution, the logistic distribution and the t-distribution [22, 59]. Through an appropriate normalization of the characteristic generator function ϕ, we assume without loss of generality that the covariance matrix of an elliptical distribution coincides with its dispersion matrix. Furthermore, because the characteristic function uniquely determines the distribution function, the tuple (ϕ, μ, Σ) uniquely identifies an elliptical distribution with generator function ϕ, mean vector μ and covariance matrix Σ, and this distribution is denoted by $\mathbb{Q} = \mathscr{P}_\phi(\mu, \Sigma)$. We denote by Φ the set of all possible characteristic generators of finite second-moment elliptical distributions. If $\phi \in \Phi$ is a characteristic generator of a unimodal (multimodal) elliptical distribution, then we say ϕ is a unimodal (multimodal, respectively) characteristic generator. Examples of unimodal elliptical distributions include Gaussian, t- and logistic distributions, while examples of multimodal elliptical distributions include a subclass of Kotz type and the multivariate Bessel type distributions.

Despite the fact that the true distribution of ξ is unknown, it is reasonable to assume that the decision maker possesses a certain belief about the true distribution of ξ. To this end, we consider the following hierarchy of ambiguity sets with structural information that captures the beliefs of the decision maker.

Definition 3.5 (Structural ambiguity sets)**.** *The hierarchy of structural ambiguity sets includes:*

1. *The ambiguity set $\mathscr{P}_2$ contains all probability measures supported on $\mathbb{R}^n$ with finite second moment.*

2. *The ambiguity set $\mathscr{P}_{\mathrm{S}}$ contains all probability measures in $\mathscr{P}_2$ under which ξ is symmetric about its mean.*

3. *The ambiguity set $\mathscr{P}_{\mathrm{SU}}$ contains all probability measures in $\mathscr{P}_2$ under which ξ is symmetric and linearly unimodal about its mean.*

4. *The ambiguity set $\mathscr{P}_\phi$ contains all probability measures under which ξ is elliptically distributed with characteristic generator function ϕ.*

Throughout this paper, we utilize σ as an index for the ambiguity set $\mathscr{P}_\sigma$, and σ may admit a value in $\mathscr{S} = \{2, \mathrm{S}, \mathrm{SU}\} \cup \Phi$. By construction, one can readily verify that $\mathscr{P}_{\phi_1} \cap \mathscr{P}_{\phi_2} = \emptyset$ for any elliptical characteristic generators $\phi_1 \neq \phi_2$. Furthermore, it holds that $\mathscr{P}_\phi \subset \mathscr{P}_{\mathrm{SU}} \subset \mathscr{P}_{\mathrm{S}} \subset \mathscr{P}_2$ for any unimodal characteristic generator $\phi \in \Phi$, and $\mathscr{P}_\phi \subset \mathscr{P}_{\mathrm{S}} \subset \mathscr{P}_2$ for any multimodal characteristic generator $\phi \in \Phi$. We note that only $\mathscr{P}_2$ is a convex set, while $\mathscr{P}_\sigma$ for $\sigma \in \{\mathrm{S}, \mathrm{SU}\} \cup \Phi$ are non-convex. Indeed, a mixture of two symmetric distributions is in general non-symmetric.

After choosing an information structure $\sigma \in \mathscr{S}$, it is natural for the decision maker to establish a *nominal distribution* $\widehat{\mathbb{P}} \in \mathscr{P}_\sigma$, which can be constructed from training samples in the data-driven setting, or from experts' belief in the Bayesian approach. Instead of making decision solely based on the nominal distribution $\widehat{\mathbb{P}}$, we assume that the decision maker will take into account an ambiguity set that contains probability measures in the neighborhood of $\widehat{\mathbb{P}}$. An attractive approach to construct the ambiguity set is to use the type-1 Wasserstein distance as a measure of dissimilarity between two distributions [118, 16, 68]. Recently, an emerging alternative is to use the type-2 Wasserstein distance in the ambiguity set which has revealed many interesting properties of the corresponding optimal solution [123, 152, 124].

Definition 3.6 (Type-2 Wasserstein distance)**.** *The type-*2 *Wasserstein distance between two probability measures $\mathbb{Q}$ and $\mathbb{Q}'$ on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ is defined as*

$$\mathbb{W}(\mathbb{Q}, \mathbb{Q}') = \left(\inf \mathbb{E}\left[\|X - X'\|^2\right]\right)^{\frac{1}{2}}, \tag{3.1}$$

where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^n$ and the infimum is taken over all joint distributions of n-dimensional random vectors X and X' with marginal distributions $\mathbb{Q}$ and $\mathbb{Q}'$, respectively.

We define the structured Wasserstein ambiguity set

$$\mathbb{B}_{\rho,\sigma}(\widehat{\mathbb{P}}) = \left\{\mathbb{Q} \in \mathscr{P}_\sigma : \mathbb{W}(\mathbb{Q}, \widehat{\mathbb{P}}) \leq \rho\right\}$$

as the ball of radius $\rho \geq 0$ in $\mathscr{P}_\sigma$ centered at the nominal distribution $\widehat{\mathbb{P}}$ with respect to the type-2 Wasserstein distance. Because $\widehat{\mathbb{P}} \in \mathscr{P}_\sigma$, the ball $\mathbb{B}_{\rho,\sigma}(\widehat{\mathbb{P}})$ is non-empty. We emphasize that $\mathbb{B}_{\rho,\sigma}(\widehat{\mathbb{P}})$ contains only probability measures that satisfy the information structure σ prescribed by the decision maker's beliefs about the distribution of ξ.

In this paper, we consider a loss function ℓ that maps any realization of the random vector $\xi \in \mathbb{R}^n$ to a real value $\ell(\xi)$. We denote by $\mathbb{L}$ the set of measurable loss functions from $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ to $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. For every $\mathbb{Q} \in \mathscr{P}_2$, we define a risk measure $\mathscr{R}_{\mathbb{Q}} : \mathbb{L} \to \mathbb{R} \cup \{+\infty\}$ that associates any loss function $\ell \in \mathbb{L}$ with a risk index $\mathscr{R}_{\mathbb{Q}}(\ell)$. Moreover, we assume that the family of risk measures $\{\mathscr{R}_{\mathbb{Q}}\}_{\mathbb{Q}\in\mathscr{P}_2}$ is consistent in the following sense.

Definition 3.7 (Consistent family of risk measures)**.** *The family of risk measures $\{\mathscr{R}_{\mathbb{Q}}\}_{\mathbb{Q}\in\mathscr{P}_2}$ is consistent if for any $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathscr{P}_2$, $\mathscr{R}_{\mathbb{Q}_1}(\ell_1) = \mathscr{R}_{\mathbb{Q}_2}(\ell_2)$ whenever the marginal distribution of $\ell_1(\xi)$ under $\mathbb{Q}_1$ and the marginal distribution of $\ell_2(\xi)$ under $\mathbb{Q}_2$ are equal.*

By definition, the consistency property implies that $\mathscr{R}_{\mathbb{Q}}$ is a law-invariant risk measure for any $\mathbb{Q} \in \mathscr{P}_2$. We note that the consistent property has been defined on $\mathscr{P}_2$, the most general set among all ambiguity sets in Definition 3.5, which in turns guarantees that the family of risk measures is consistent under any restriction $\mathscr{P}_\sigma$ for $\sigma \in \mathscr{S}$. Furthermore, many worst-case risk measures studied in the literature implicitly assumes that the family of risk measures is consistent, they include the worst-case Value-at-Risk [52, 121], worst-case Conditional Value-at-Risk [180], worst-case law-invariant coherent risk measure [110], the worst-case expected loss [39, 118], etc. We emphasize that there exists a systematic way to construct a consistent family of risk measures $\left\{\mathscr{R}_{\mathbb{Q}}\right\}_{\mathbb{Q}\in\mathscr{P}_2}$ as highlighted in the following remark.

Remark 3.8 (Construction of a consistent family of risk measures)**.** *There is a systematic way of constructing the family of risk measures $\{\mathscr{R}_{\mathbb{Q}}\}_{\mathbb{Q}\in\mathscr{P}_2}$ so that it satisfies the consistency property stated in Definition 3.7. First, fix a non-atomic probability measure $\mathbb{P}$ and a law-invariant risk measure $\mathscr{R}_{\mathbb{P}}$. For every probability space $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \mathbb{Q})$ for $\mathbb{Q} \in \mathscr{P}_2$, define a measurable function $X_{\mathbb{Q}} : \mathbb{R}^n \to \mathbb{R}^n$ such that $\mathbb{Q} = \mathbb{P} \circ X_{\mathbb{Q}}^{-1}$. The existence of $X_{\mathbb{Q}}$ is guaranteed because $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \mathbb{P})$ is atomless, in this case, $\mathbb{Q}$ is the pushforward measure of$\mathbb{P}$ under $X_{\mathbb{Q}}$. The risk measure $\mathscr{R}_{\mathbb{Q}}$ is constructed as follows for any $\ell \in \mathbb{L}$:*

$$\mathscr{R}_{\mathbb{Q}}(\ell) = \mathscr{R}_{\mathbb{P}}(\ell(X_{\mathbb{Q}})).$$

Using the structured Wasserstein ambiguity set, we define the *Wasserstein risk* of a loss function $\ell \in \mathbb{L}$ as the highest risk over all probability measures contained in $\mathbb{B}_{\rho,\sigma}(\widehat{\mathbb{P}})$, that is,

$$\mathscr{R}_{\rho,\sigma}(\widehat{\mathbb{P}}, \ell) = \sup_{\mathbb{Q}\in\mathbb{B}_{\rho,\sigma}(\widehat{\mathbb{P}})} \mathscr{R}_{\mathbb{Q}}(\ell). \tag{3.2}$$

Furthermore, given a subset of loss functions $\mathscr{L} \subseteq \mathbb{L}$, the *Wasserstein optimal risk* over $\mathscr{L}$ is defined as

$$\mathscr{R}_{\rho,\sigma}(\widehat{\mathbb{P}}, \mathscr{L}) = \inf_{\ell\in\mathscr{L}} \mathscr{R}_{\rho,\sigma}(\widehat{\mathbb{P}}, \ell), \tag{3.3}$$

which is a decision problem that searches for the loss function that minimizes the Wasserstein risk.

While being appealing at first glance, evaluating the Wasserstein risk in (3.2) is a challenging task because computing the type-2 Wassertein distance between two distributions is in general

#P-hard [163]. If the nominal distribution $\widehat{\mathbb{P}}$ is a discrete distribution supported on a finite number of points, problem (3.2) can be potentially reformulated as a finite dimensional optimization problem [179]. Unfortunately, these reformulations usually add a large number of auxiliary dual variables which prohibits the scalability of the reformulation approach as the number of atoms in the nominal distribution increases. In the remainder of this paper, we will explore a systematic approach to construct a conservative approximation of the Wasserstein risk, as well as the tractable reformulations of these approximation problems under specific settings.

3.3 Conservative Approximation of the Wasserstein Risk

3.3.1 Gelbrich Risk Approximation

A fundamental element of the Wasserstein risk evaluation problem (3.2) is the Wasserstein ball $\mathbb{B}_{\rho,\sigma}(\widehat{\mathbb{P}})$. In this section, we form an outer approximation of $\mathbb{B}_{\rho,\sigma}(\widehat{\mathbb{P}})$ using the moment information of the nominal distribution $\widehat{\mathbb{P}}$. Given a finite second-moment nominal measure $\widehat{\mathbb{P}} \in \mathscr{P}_2$, we denote by $\widehat{\mu} \in \mathbb{R}^n$ the mean vector and by $\widehat{\Sigma} \in \mathbb{S}_+^n$ the covariance matrix of ξ under $\widehat{\mathbb{P}}$. We consider the Gelbrich distance defined on the mean vector-covariance matrix space.

Definition 3.9 (Gelbrich distance)**.** *The Gelbrich distance between two tuples* $(\mu,\Sigma) \in \mathbb{R}^n \times \mathbb{S}_+^n$ *and* $(\widehat{\mu},\widehat{\Sigma}) \in \mathbb{R}^n \times \mathbb{S}_+^n$ *amounts to*

$$\mathbb{G}\big((\mu,\Sigma),(\widehat{\mu},\widehat{\Sigma})\big) \triangleq \sqrt{\|\mu-\widehat{\mu}\|_2^2 + \mathrm{Tr}\big[\Sigma+\widehat{\Sigma}-2\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\big]}.$$

One can readily show that $\mathbb{G}$ is non-negative, symmetric, vanishes if and only if $(\mu,\Sigma) = (\widehat{\mu},\widehat{\Sigma})$ and satisfies the triangle inequality, and thus $\mathbb{G}$ is a metric on $\mathbb{R}^n \times \mathbb{S}_+^n$ [72, pp. 239]. The Gelbrich distance provides a lower bound on the Wasserstein type-2 distance between probability distributions in terms of their mean vectors and covariance matrices as highlighted in the following theorem.

Theorem 3.10 (Gelbrich bound [71, Theorems 2.1 and 2.4])**.** *If the distributions* $\mathbb{Q}$ *and* $\mathbb{Q}'$ *have mean vectors* $\mu,\mu' \in \mathbb{R}^n$ *and covariance matrices* $\Sigma,\Sigma' \in \mathbb{S}_+^n$*, respectively, then*

$$\mathbb{W}(\mathbb{Q},\mathbb{Q}') \geq \mathbb{G}\left((\mu,\Sigma),(\mu',\Sigma')\right). \tag{3.4}$$

The bound is exact if $\mathbb{Q}$ *and* $\mathbb{Q}'$ *are elliptical distributions with the same characteristic generator.*

Using the Gelbrich distance as a measure of dissimilarity, we define the uncertainty set in the space of mean vectors and covariance matrices as

$$\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma}) = \big\{(\mu,\Sigma) \in \mathbb{R}^n \times \mathbb{S}_+^n : \mathbb{G}\big((\mu,\Sigma),(\widehat{\mu},\widehat{\Sigma})\big) \leq \rho\big\}.$$

Intuitively, $\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ contains all tuples of mean vectors and covariance matrix of a Gelbrich distance less than or equal to ρ from the center $(\widehat{\mu},\widehat{\Sigma})$. The uncertainty set $\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ is of interest because it covers the projection of the type-2 Wasserstein ball $\mathbb{B}_{\rho,\sigma}(\widehat{\mathbb{P}})$ onto the space of mean vectors and covariance matrices. Moreover, if ϕ is a unimodal characteristic generator and $\widehat{\mathbb{P}}=\mathscr{P}_\phi(\widehat{\mu},\widehat{\Sigma})$, then the uncertainty set $\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ actually coincides the projection of $\mathbb{B}_{\rho,\sigma}(\widehat{\mathbb{P}})$ for any $\sigma\in\{2,\mathrm{S},\mathrm{SU},\phi\}$.

Proposition 3.11 (Projection of $\mathbb{B}_{\rho,\sigma}(\widehat{\mathbb{P}})$ onto the mean-covariance space)**.** *If the nominal distribution $\widehat{\mathbb{P}}$ has mean vector $\widehat{\mu}\in\mathbb{R}^n$ and covariance matrix $\widehat{\Sigma}\in\mathbb{S}^n_+$, then for any $\sigma\in\mathscr{S}$,*

$$\left\{\left(\mathbb{E}_{\mathbb{Q}}[\xi],\mathbb{E}_{\mathbb{Q}}[(\xi-\mathbb{E}_{\mathbb{Q}}[\xi])(\xi-\mathbb{E}_{\mathbb{Q}}[\xi])^\top]\right):\mathbb{Q}\in\mathbb{B}_{\rho,\sigma}(\widehat{\mathbb{P}})\right\}\subseteq\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma}).$$

The inclusion becomes an equality if $\widehat{\mathbb{P}}=\mathscr{P}_\phi(\widehat{\mu},\widehat{\Sigma})$ is an elliptical distribution.

Proposition 3.11 follows immediately from Theorem 3.10. Furthermore, if $\widehat{\mathbb{P}}=\mathscr{P}_\phi(\widehat{\mu},\widehat{\Sigma})$ then any elliptical distribution $\mathbb{Q}=\mathscr{P}_\phi(\mu,\Sigma)$ with the same characteristic generator ϕ and with $\mathbb{W}(\mathbb{Q},\widehat{\mathbb{P}})\le\rho$ belongs to $\mathbb{B}_{\rho,2}(\widehat{\mathbb{P}})$.

A useful ambiguity set in the space of probability distributions is the *Gelbrich hull*, which is constructed as the pre-image of $\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ under the mean-covariance projection.

Definition 3.12 (Gelbrich hull)**.** *For any $\sigma\in\mathscr{S}$, the Gelbrich hull is given by*

$$\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})=\left\{\mathbb{Q}\in\mathscr{P}_\sigma:\left(\mathbb{E}_{\mathbb{Q}}[\xi],\mathbb{E}_{\mathbb{Q}}[(\xi-\mathbb{E}_{\mathbb{Q}}[\xi])(\xi-\mathbb{E}_{\mathbb{Q}}[\xi])^\top]\right)\in\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})\right\}.$$

By definition, $\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})$ contains all distributions supported on $\mathbb{R}^n$ satisfying the information structure σ whose mean vectors and covariance matrices fall into the uncertainty set $\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$. If we define $\mathscr{P}_\sigma(\mu,\Sigma)$ as the *structural Chebyshev ambiguity set* that contains all distributions on $\mathbb{R}^n$ satisfying the information structure σ with fixed mean vector $\mu\in\mathbb{R}^n$ and covariance matrix $\Sigma\in\mathbb{S}^n_+$, then the Gelbrich hull can also be expressed as

$$\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})=\bigcup_{(\mu,\Sigma)\in\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})}\mathscr{P}_\sigma(\mu,\Sigma).\tag{3.5}$$

From this representation it is evident that if $\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})$ contains a distribution $\mathbb{Q}$, then it contains *all* distributions on $\mathbb{R}^n$ that satisfy the information structure σ and have the same mean vector and covariance matrix as $\mathbb{Q}$. It is easy to verify that the Gelbrich hull $\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})$ provides an outer approximation for any Wasserstein ball $\mathbb{B}_{\rho,\sigma}(\widehat{\mathbb{P}})$. Indeed, if $\mathbb{B}_{\rho,\sigma}(\widehat{\mathbb{P}})$ contains a distribution $\mathbb{Q}$ with mean vector μ and covariance matrix Σ, then $(\mu,\Sigma)\in\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ by virtue of Proposition 3.11, which implies via (3.5) that $\mathbb{Q}\in\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})$. These insights culminate in the following theorem.

Theorem 3.13 (Gelbrich hull)**.** *If the nominal distribution $\widehat{\mathbb{P}}$ has mean vector $\widehat{\mu}\in\mathbb{R}^n$ and covariance matrix $\widehat{\Sigma}\in\mathbb{S}^n_+$, then we have $\mathbb{B}_{\rho,\sigma}(\widehat{\mathbb{P}})\subseteq\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})$ for any $\sigma\in\mathscr{S}$. If $\sigma=\phi$ and $\widehat{\mathbb{P}}=\mathscr{P}_\phi(\widehat{\mu},\widehat{\Sigma})$, then $\mathbb{B}_{\rho,\phi}(\widehat{\mathbb{P}})=\mathbb{G}_{\rho,\phi}(\widehat{\mu},\widehat{\Sigma})$.*

Proof. Pick any $\mathbb{Q} \in \mathbb{B}_{\rho,\sigma}(\widehat{\mathbb{P}})$ and define $\mu \in \mathbb{R}^n$ as the mean vector and $\Sigma \in \mathbb{S}_+^n$ as the covariance matrix of ξ under $\mathbb{Q}$. By Theorem 3.10, we find

$$\mathbb{G}\big((\mu,\Sigma),(\widehat{\mu},\widehat{\Sigma})\big) \le \mathbb{W}(\mathbb{Q},\widehat{\mathbb{P}}) \le \rho,$$

which implies that $\mathbb{Q} \in \mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})$. We may thus conclude that $\mathbb{B}_{\rho,\sigma}(\widehat{\mathbb{P}}) \subseteq \mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})$ for any $\sigma \in \mathscr{S}$.

When $\sigma = \phi$, pick any $\mathbb{Q} \in \mathbb{G}_{\rho,\phi}(\widehat{\mathbb{P}})$ and define $\mu \in \mathbb{R}^n$ as the mean vector and $\Sigma \in \mathbb{S}_+^n$ as the covariance matrix of ξ under $\mathbb{Q}$. By Theorem 3.10, we find

$$\mathbb{W}(\mathbb{Q},\widehat{\mathbb{P}}) = \mathbb{G}\big((\mu,\Sigma),(\widehat{\mu},\widehat{\Sigma})\big) \le \rho,$$

where the equality holds because both $\widehat{\mathbb{P}}$ and $\mathbb{Q}$ share the same characteristic generator ϕ. This implies that $\mathbb{Q} \in \mathbb{B}_{\rho,\phi}(\widehat{\mathbb{P}})$, and hence $\mathbb{B}_{\rho,\phi}(\widehat{\mathbb{P}}) = \mathbb{G}_{\rho,\phi}(\widehat{\mu},\widehat{\Sigma})$. This completes the proof. □

Theorem 3.13 shows that the Gelbrich hull provides an outer approximation for all Wasserstein balls $\mathbb{B}_{\rho,\sigma}(\widehat{\mathbb{P}})$ solely on the basis of mean and covariance information. Discarding all information about $\widehat{\mathbb{P}}$ beyond its first- and second-order moments can be seen as a compression of available information. This amounts to sacrificing higher-order moment information and may improve the tractability of the risk evaluation problem (3.2) and the distributionally robust decision problem (3.3). To show this, we define the *Gelbrich risk* as

$$\overline{\mathscr{R}}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma},\ell) = \sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})} \mathscr{R}_{\mathbb{Q}}(\ell) \tag{3.6}$$

and the *optimal Gelbrich risk* as

$$\overline{\mathscr{R}}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma},\mathscr{L}) = \inf_{\ell\in\mathscr{L}} \overline{\mathscr{R}}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma},\ell). \tag{3.7}$$

Theorem 3.13 immediately implies that the (optimal) Gelbrich risk provides an upper bound on the (optimal) Wasserstein risk.

Corollary 3.14 (Gelbrich risk)**.** *If the nominal distribution $\widehat{\mathbb{P}}$ has mean vector $\widehat{\mu} \in \mathbb{R}^n$ and covariance matrix $\widehat{\Sigma} \in \mathbb{S}_+^n$, then for any $\sigma \in \mathscr{S}$, we have*

$$\mathscr{R}_{\rho,\sigma}(\widehat{\mathbb{P}},\ell) \le \overline{\mathscr{R}}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma},\ell) \quad \forall \ell \in \mathscr{L} \qquad \text{and} \qquad \mathscr{R}_{\rho,\sigma}(\widehat{\mathbb{P}},\mathscr{L}) \le \overline{\mathscr{R}}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma},\mathscr{L}).$$

If $\sigma = \phi$ and $\widehat{\mathbb{P}} = \mathscr{P}_\phi(\widehat{\mu},\widehat{\Sigma})$, we have

$$\mathscr{R}_{\rho,\phi}(\widehat{\mathbb{P}},\ell) = \overline{\mathscr{R}}_{\rho,\phi}(\widehat{\mu},\widehat{\Sigma},\ell) \quad \forall \ell \in \mathscr{L} \qquad \text{and} \qquad \mathscr{R}_{\rho,\phi}(\widehat{\mathbb{P}},\mathscr{L}) = \overline{\mathscr{R}}_{\rho,\phi}(\widehat{\mu},\widehat{\Sigma},\mathscr{L}).$$

Unlike the mean vector $\mu = \mathbb{E}_{\mathbb{Q}}[\xi]$ and the second-order moment matrix $M = \mathbb{E}_{\mathbb{Q}}[\xi\xi^\top]$, both of which constitute linear functions of the underlying distribution $\mathbb{Q}$, the covariance matrix $\Sigma = M - \mu\mu^\top$ is nonlinear in $\mathbb{Q}$. The condition $(\mu,\Sigma) \in \mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ thus appears to be nonconvex in $\mathbb{Q}$. To gain a clearer understanding, it is instructive to introduce the uncertainty set $\mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})$ for (μ, M) induced by the uncertainty set $\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ for (μ,Σ), that is,

$$\mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma}) = \left\{(\mu, M) \in \mathbb{R}^n \times \mathbb{S}_+^n : (\mu, M - \mu\mu^\top) \in \mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})\right\}.$$

Maybe surprisingly, even though it is defined as the pre-image of a convex set under a *non*linear transformation, one can prove that $\mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})$ is convex, see Proposition 3.17.

The representation (3.5) of the Gelbrich hull as a union of Chebyshev ambiguity sets suggests that the Gelbrich risk of any fixed loss function $\ell(\xi)$ can be expressed as the optimal value of the following two-layer optimization problem

$$\overline{\mathscr{R}}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma},\ell) = \sup_{(\mu,\Sigma)\in\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \; \sup_{\mathbb{Q}\in\mathscr{P}_\sigma(\mu,\Sigma)} \mathscr{R}_{\mathbb{Q}}(\ell) \tag{3.8a}$$

$$= \sup_{(\mu,M)\in\mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})} \; \sup_{\mathbb{Q}\in\mathscr{P}_\sigma(\mu,M-\mu\mu^\top)} \mathscr{R}_{\mathbb{Q}}(\ell) \tag{3.8b}$$

Note that (3.8b) follows immediately from the definition of the uncertainty set $\mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})$ and the formula for the covariance matrix in terms of the mean vector and the second-order moment matrix. The inner problems in (3.8a) and (3.8b) both represent the same distributionally robust optimization problem over a Chebyshev ambiguity set but with different parameterizations. This problem can be viewed as an infinite-dimensional linear program over all probability distributions $\mathbb{Q}$ that satisfy the linear equality constraints $\mathbb{E}_{\mathbb{Q}}[\xi] = \mu$ and $\mathbb{E}_{\mathbb{Q}}[\xi\xi^\top] = M$. The outer problem in (3.8a) hedges against ambiguity in the mean vector and the covariance matrix, while the one in (3.8b) hedges against ambiguity in the first- and second-order moments. The formulation (3.8a) is conceptually appealing because of its connection to the Wasserstein distance and because it is more natural to characterize a distribution in terms of its mean vector and covariance matrix. The formulation (3.8b), on the other hand, is computationally attractive because it expresses the inner problem with linear constraints while the feasible set of the outer problem remains convex.

Remark 3.15 (Second layer of robustness)**.** *Distributionally robust optimization problems akin to* (3.8a) *and* (3.8b) *that accommodate a second layer of robustness to account for moment ambiguity have been investigated in [39, 52, 81, 121, 145, 184], among others. As the optimal value of the inner maximization problem is always concave in* (μ, M) *but typically nonconcave in* (μ,Σ)*, moment ambiguity has mostly been modeled through convex uncertainty sets for* (μ, M)*, thereby ensuring convexity of the outer maximization problem. For example, uncertainty sets that force* μ *to lie in an ellipsoid and* M *in the intersection of two positive semi-definite cones were studied in [39], while box-type uncertainty sets for* (μ, M) *were proposed in [121] and refined in [81, 184]. Convex uncertainty sets for* (μ,Σ) *were shown to render the outer maximization problems convex only in special cases,* e.g.*, when evaluating a worst-case value-at-risk of a linear or quadratic loss function [52, 145]. The convex uncertainty set* $\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ *for* (μ,Σ) *is remarkable because it leads to a second-layer maximization problem in* (3.8a) *that potentially admits a convex reformulation.* □

3.3.2 Convexity of the Gelbrich Hull

We now establish the convexity of the Gelbrich hull $\mathbb{G}_{\rho,\sigma}(\widehat{\mathbb{P}})$. To this end, we first prove a structural result certifying that the uncertainty set $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$, a finite dimensional subset on the space of mean vector and covariance matrix, is compact and convex.

Proposition 3.16 (Convexity of $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$)**.** *For any $\widehat{\mu}\in\mathbb{R}^n$, $\widehat{\Sigma}\in\mathbb{S}_+^n$ and $\rho\in\mathbb{R}_+$, $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ is compact and convex.*

Proof. Notice that $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ can be expressed as

$$\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})=\left\{(\mu,\Sigma)\in\mathbb{R}^n\times\mathbb{S}_+^n:\mathbb{G}\big((\mu,\Sigma),(\widehat{\mu},\widehat{\Sigma})\big)^2\le\rho^2\right\},$$

where the squared Gelbrich distance $\mathbb{G}\big((\mu,\Sigma),(\widehat{\mu},\widehat{\Sigma})\big)^2$ is a continuous, convex function of $(\mu,\Sigma)\in\mathbb{R}^n\times\mathbb{S}_+^n$. Hence, $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ is convex and closed. Furthermore, for any $(\mu,\Sigma)\in\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$, we have $\|\mu-\widehat{\mu}\|\le\rho$, which implies that μ is bounded. Moreover, [152, Lemma A.6] implies that for any $(\mu,\Sigma)\in\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$, we find $0\preceq\Sigma\preceq\big(\rho+\mathrm{Tr}\big[\widehat{\Sigma}\big]^{\frac{1}{2}}\big)^2 I$. We conclude that $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ is compact. □

Proposition 3.17 (Convexity of $\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})$)**.** *For any $\widehat{\mu}\in\mathbb{R}^n$, $\widehat{\Sigma}\in\mathbb{S}_+^n$ and $\rho\in\mathbb{R}_+$, $\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})$ is compact and convex.*

The proof of Proposition 3.17 requires the following preparatory lemma that establishes the equivalent formulation of the squared Gelbrich distance as the optimal value of a maximization problem.

Lemma 3.18. *For any $\mu,\widehat{\mu}\in\mathbb{R}^n$ and $\Sigma,\widehat{\Sigma}\in\mathbb{S}_+^n$, we have*

$$\mathbb{G}\big((\mu,\Sigma),(\widehat{\mu},\widehat{\Sigma})\big)^2=\left\{\begin{array}{ll}\sup & \mathrm{Tr}\big[(\Sigma+\mu\mu^\top)(I-A_{11})+(\widehat{\Sigma}+\widehat{\mu}\widehat{\mu}^\top)(I-A_{22})\big]+\mu^\top A_{11}\mu+\widehat{\mu}^\top A_{22}\widehat{\mu}-2\widehat{\mu}^\top\mu\\ \text{s.t.} & A_{11}\in\mathbb{S}_+^n,\ A_{22}\in\mathbb{S}_+^n,\ \begin{bmatrix}A_{11} & -I\\ -I & A_{22}\end{bmatrix}\succeq 0.\end{array}\right.$$

Proof. We first define the function $\mathbb{G}'$ for any $\mu,\widehat{\mu}\in\mathbb{R}^n$ and $M,\widehat{M}\in\mathbb{S}_+^n$ as

$$\mathbb{G}'\big((\mu,M),(\widehat{\mu},\widehat{M})\big)^2\triangleq\left\{\begin{array}{ll}\sup & \mathrm{Tr}\big[M(I-A_{11})+\widehat{M}(I-A_{22})\big]+\mu^\top A_{11}\mu+\widehat{\mu}^\top A_{22}\widehat{\mu}-2\widehat{\mu}^\top\mu\\ \text{s.t.} & A_{11}\in\mathbb{S}_+^n,\ A_{22}\in\mathbb{S}_+^n,\ \begin{bmatrix}A_{11} & -I\\ -I & A_{22}\end{bmatrix}\succeq 0.\end{array}\right. \tag{3.9}$$

From the proof of [72, Proposition 7], we can re-express the Gelbrich distance $\mathbb{G}$ as the optimal value of a minimization problem

$$
\begin{aligned}
\mathbb{G}\big((\mu,\Sigma),(\widehat{\mu},\widehat{\Sigma})\big)^2 &= \left\{ \begin{array}{ll} \min & \|\mu-\widehat{\mu}\|_2^2 + \mathrm{Tr}\left[\Sigma+\widehat{\Sigma}-2C\right] \\ \mathrm{s.t.} & C\in\mathbb{R}^{n\times n},\ \begin{bmatrix} \Sigma & C \\ C^\top & \widehat{\Sigma} \end{bmatrix} \succeq 0 \end{array}\right. \\
&= \left\{ \begin{array}{ll} \min & \mathrm{Tr}\left[\mu\mu^\top+\widehat{\mu}\widehat{\mu}^\top\right] + \mathrm{Tr}\left[\Sigma+\widehat{\Sigma}\right] - 2\,\mathrm{Tr}\left[C+\mu\widehat{\mu}^\top\right] \\ \mathrm{s.t.} & C\in\mathbb{R}^{n\times n},\ \begin{bmatrix} \Sigma & C \\ C^\top & \widehat{\Sigma} \end{bmatrix} \succeq 0 \end{array}\right. \\
&= \left\{ \begin{array}{ll} \min & \mathrm{Tr}\left[(\Sigma+\mu\mu^\top)+(\widehat{\Sigma}+\widehat{\mu}\widehat{\mu}^\top)\right] - 2\,\mathrm{Tr}\left[K\right] \\ \mathrm{s.t.} & K\in\mathbb{R}^{n\times n},\ \begin{bmatrix} \Sigma+\mu\mu^\top & K \\ K^\top & \widehat{\Sigma}+\widehat{\mu}\widehat{\mu}^\top \end{bmatrix} \succeq \begin{bmatrix} \mu \\ \widehat{\mu} \end{bmatrix} \begin{bmatrix} \mu \\ \widehat{\mu} \end{bmatrix}^\top, \end{array}\right.
\end{aligned}
$$

where the last equality follows from the change of variable $K \leftarrow C+\mu\widehat{\mu}^\top$. Using a weak duality argument, we have

$$
\begin{aligned}
\mathbb{G}\left((\mu,\Sigma),(\widehat{\mu},\widehat{\Sigma})\right)^2 &\geq \left\{ \begin{array}{ll} \sup & \mathrm{Tr}\left[(\Sigma+\mu\mu^\top)(I-A_{11})+(\widehat{\Sigma}+\widehat{\mu}\widehat{\mu}^\top)(I-A_{22})\right] + \mu^\top A_{11}\mu + \widehat{\mu}^\top A_{22}\widehat{\mu} - 2\widehat{\mu}^\top\mu \\ \mathrm{s.t.} & A_{11}\in\mathbb{S}_+^n,\ A_{22}\in\mathbb{S}_+^n,\ \begin{bmatrix} A_{11} & -I \\ -I & A_{22} \end{bmatrix} \succeq 0 \end{array}\right. \\
&= \mathbb{G}'\big((\mu,\Sigma+\mu\mu^\top),(\widehat{\mu},\widehat{\Sigma}+\widehat{\mu}\widehat{\mu}^\top)\big)^2,
\end{aligned}
$$

where the equality follows from the definition of $\mathbb{G}'$ in (3.9). Thus, $\mathbb{G}\big((\mu,\Sigma),(\widehat{\mu},\widehat{\Sigma})\big)^2$ constitutes an upper bound on $\mathbb{G}'\big((\mu,\Sigma+\mu\mu^\top),(\widehat{\mu},\widehat{\Sigma}+\widehat{\mu}\widehat{\mu}^\top)\big)^2$. We now proceed to prove that the inequality above holds as an equality. Denote momentarily $f_\Sigma(\widehat{\Sigma})$ as the optimal value of the problem

$$
f_\Sigma(\widehat{\Sigma}) = \left\{ \begin{array}{ll} \inf & \mathrm{Tr}\left[\Sigma A_{11}\right] + \mathrm{Tr}\left[\widehat{\Sigma} A_{22}\right] \\ \mathrm{s.t.} & A_{11}\in\mathbb{S}_+^n,\ A_{22}\in\mathbb{S}_+^n,\ \begin{bmatrix} A_{11} & -I \\ -I & A_{22} \end{bmatrix} \succeq 0 \end{array}\right. \tag{3.10}
$$

parametrized by $\widehat{\Sigma}$. By construction, for any $\mu,\widehat{\mu}\in\mathbb{R}^n$ and $\Sigma,\widehat{\Sigma}\in\mathbb{S}_+^n$, we have

$$
\begin{aligned}
\mathbb{G}'\big((\mu,\Sigma+\mu\mu^\top),(\widehat{\mu},\widehat{\Sigma}+\widehat{\mu}\widehat{\mu}^\top)\big)^2 &= \left\{ \begin{array}{ll} \sup & -\mathrm{Tr}\left[\Sigma A_{11}\right] - \mathrm{Tr}\left[\widehat{\Sigma} A_{22}\right] + \mathrm{Tr}\left[\Sigma+\widehat{\Sigma}\right] + \|\mu-\widehat{\mu}\|_2^2 \\ \mathrm{s.t.} & A_{11}\in\mathbb{S}_+^n,\ A_{22}\in\mathbb{S}_+^n,\ \begin{bmatrix} A_{11} & -I \\ -I & A_{22} \end{bmatrix} \succeq 0 \end{array}\right. \\
&= -f_\Sigma(\widehat{\Sigma}) + \mathrm{Tr}\left[\Sigma+\widehat{\Sigma}\right] + \|\mu-\widehat{\mu}\|_2^2,
\end{aligned}
$$

where the first equality follows from the definition of M and $\widehat{M}$ while the second equality holds due to the definition of $f_\Sigma(\widehat{\Sigma})$. By imposing the constraint $A_{11} = A_{22}^\dagger$, where $A_{22}^\dagger$ denotes the (Moore-Penrose) pseudoinverse of A_{22} into the infimum problem (3.10), we have the trivial upper bound on $f_\Sigma(\widehat{\Sigma})$ as

$$f_\Sigma(\widehat{\Sigma}) \le \left\{ \begin{array}{ll} \inf & \mathrm{Tr}\left[\Sigma A_{22}^\dagger\right] + \mathrm{Tr}\left[\widehat{\Sigma} A_{22}\right] \\ \mathrm{s.t.} & A_{22} \in \mathbb{S}_+^n,\ \begin{bmatrix} A_{22}^\dagger & -I \\ -I & A_{22} \end{bmatrix} \ge 0. \end{array} \right.$$

If $\widehat{\Sigma} \succ 0$, the above infimum problem over A_{22} admits the optimal value $\mathrm{Tr}\left[2\sqrt{\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}}\right]$ [128, Theorem 4]. As a consequence, for any $\widehat{\Sigma} \succ 0$, we have

$$\mathbb{G}'\big((\mu, \Sigma + \mu\mu^\top), (\widehat{\mu}, \widehat{\Sigma} + \widehat{\mu}\widehat{\mu}^\top)\big)^2 \ge \|\mu - \widehat{\mu}\|_2^2 + \mathrm{Tr}\left[\Sigma + \widehat{\Sigma} - 2\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right] = \mathbb{G}\big((\mu, \Sigma), (\widehat{\mu}, \widehat{\Sigma})\big)^2,$$

where the last equality is from the analytical expression of $\mathbb{G}$ in Definition 3.9. Combining with the upper bound established previously, we conclude that for any $\widehat{\Sigma} \succ 0$

$$\mathbb{G}\big((\mu, \Sigma), (\widehat{\mu}, \widehat{\Sigma})\big)^2 = \mathbb{G}'\big((\mu, \Sigma + \mu\mu^\top), (\widehat{\mu}, \widehat{\Sigma} + \widehat{\mu}\widehat{\mu}^\top)\big)^2.$$

To complete the proof, we now extend the equality above to the case when $\widehat{\Sigma}$ is singular. To this end, we first show that $f_\Sigma(\widehat{\Sigma})$ defined in (3.10) is continuous on $\mathbb{S}_+^n$. Fix $\widehat{\Sigma} \in \mathbb{S}_+^n$, it is clear that $\widehat{\Sigma} + \varepsilon I \succ 0$ for any $\varepsilon > 0$. Define the interval $\mathscr{E} = \mathbb{R}_+$ and the feasible set

$$\mathscr{A} = \left\{ A_{11} \in \mathbb{S}_+^n,\ A_{22} \in \mathbb{S}_+^n : \begin{bmatrix} A_{11} & -I \\ -I & A_{22} \end{bmatrix} \ge 0 \right\}.$$

In addition, define the auxiliary functions $\psi(\varepsilon, A) \triangleq \mathrm{Tr}\left[\Sigma A_{11}\right] + \mathrm{Tr}\left[\widehat{\Sigma} A_{22}\right] + \varepsilon \mathrm{Tr}\left[A_{22}\right]$ with $A = (A_{11}, A_{22})$, and $\Psi(\varepsilon) \triangleq f_\Sigma(\widehat{\Sigma} + \varepsilon I)$. It follows from the definition of $f_\Sigma(\widehat{\Sigma})$ in (3.10) that

$$\Psi(\varepsilon) = \inf_{A \in \mathscr{A}} \psi(\widehat{\Sigma} + \varepsilon I, A) \quad \forall \varepsilon \in \mathscr{E}.$$

We now show that $\Psi(\varepsilon)$ is continuous at $\varepsilon = 0$. Because ψ is linear and thus continuous in ε for any $A \in \mathscr{A}$, it implies that Ψ is upper-semicontinuous at $\varepsilon = 0$ [123, Lemma 2.7(a)]. Moreover, ψ is calm from below at $\varepsilon = 0$ uniformly over $A \in \mathscr{A}$ with the calmness constant 0 because

$$\psi(\varepsilon, A) - \psi(0, A) = \varepsilon \mathrm{Tr}\left[A_{22}\right] \ge 0 \quad \forall A \in \mathscr{A}.$$

We can assert that Ψ is lower-semicontinuous at $\varepsilon = 0$ [123, Lemma 2.7(b)]. This implies that Ψ is continuous at $\varepsilon = 0$, and as a consequence, $f_\Sigma(\widehat{\Sigma})$ is continuous over $\mathbb{S}_+^n$. Thus, for any singular covariance matrix $\widehat{\Sigma} \in \mathbb{S}_+^n$, we have

$$\begin{aligned}
\mathbb{G}\big((\mu,\Sigma),(\widehat{\mu},\widehat{\Sigma})\big)^2 &= \lim_{\varepsilon\downarrow 0} \mathbb{G}\big((\mu,\Sigma),(\widehat{\mu},\widehat{\Sigma}+\varepsilon I)\big)^2 \\
&= \lim_{\varepsilon\downarrow 0} \mathbb{G}'\big((\mu,\Sigma+\mu\mu^\top),(\widehat{\mu},\widehat{\Sigma}+\varepsilon I+\widehat{\mu}\widehat{\mu}^\top)\big)^2 \\
&= \lim_{\varepsilon\downarrow 0} -f_\Sigma(\widehat{\Sigma}+\varepsilon I) + \mathrm{Tr}\big[\Sigma+\widehat{\Sigma}\big] + \|\mu-\widehat{\mu}\|^2 \\
&= \lim_{\varepsilon\downarrow 0} -\Psi(\varepsilon) + \mathrm{Tr}\big[\Sigma+\widehat{\Sigma}\big] + \|\mu-\widehat{\mu}\|^2 \\
&= -\Psi(0) + \mathrm{Tr}\big[\Sigma+\widehat{\Sigma}\big] + \|\mu-\widehat{\mu}\|^2 \\
&= -f_\Sigma(\widehat{\Sigma}) + \mathrm{Tr}\big[\Sigma+\widehat{\Sigma}\big] + \|\mu-\widehat{\mu}\|^2 \\
&= \mathbb{G}'\big((\mu,\Sigma+\mu\mu^\top),(\widehat{\mu},\widehat{\Sigma}+\widehat{\mu}\widehat{\mu}^\top)\big)^2,
\end{aligned}$$

where the equalities come from the continuity of $\mathbb{G}$, the equivalence between $\mathbb{G}$ and $\mathbb{G}'$ when $\widehat{\Sigma} \succ 0$, the definition of $f_\Sigma(\widehat{\Sigma})$ in (3.10), the definition of Ψ, the continuity of Ψ at $\varepsilon = 0$, the definition of Ψ again and finally the definition of $\mathbb{G}'$, respectively. We now can finally conclude that for any $\mu, \widehat{\mu} \in \mathbb{R}^n$ and $\Sigma, \widehat{\Sigma} \in \mathbb{S}_+^n$

$$\mathbb{G}\big((\mu,\Sigma),(\widehat{\mu},\widehat{\Sigma})\big)^2 = \mathbb{G}'\big((\mu,\Sigma+\mu\mu^\top),(\widehat{\mu},\widehat{\Sigma}+\widehat{\mu}\widehat{\mu}^\top)\big)^2.$$

This completes the proof. □

Proof of Proposition 3.17. Let $\mathbb{G}'$ be defined as in (3.9). Thanks to the equivalence between $\mathbb{G}$ and $\mathbb{G}'$ established in Lemma 3.18, we find

$$\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma}) = \Big\{(\mu, M) \in \mathbb{R}^n \times \mathbb{S}_+^n : M - \mu\mu^\top \succeq 0,\ \mathbb{G}'\big((\mu,M),(\widehat{\mu},\widehat{\Sigma}+\widehat{\mu}\widehat{\mu}^\top)\big)^2 \le \rho^2\Big\}.$$

For a fixed $\widehat{\mu} \in \mathbb{R}^n$ and $\widehat{\Sigma} \in \mathbb{S}_+^n$, the squared-$\mathbb{G}'$ function is jointly convex over (μ, M) because it is the pointwise supremum of convex quadratic functions. Hence, $\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})$ is a convex set. The compactness of $\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})$ is a direct consequence of the compactness of $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ established in Proposition 3.16. □

We emphasize that $\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})$ is in general non-convex for $\sigma \in \{\mathrm{S},\mathrm{SU}\} \cup \Phi$ due to the non-convexity of the structural ambiguity set $\mathcal{P}_\sigma$. Moreover, the convexity of $\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})$ is not apparent from Definition 3.12, which introduces the Gelbrich hull as the pre-image of a convex set under a *non*linear transformation. Fortunately, by the definition of the induced uncertainty set $\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})$, the set $\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})$ can be equivalently expressed as

$$\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma}) = \big\{\mathbb{Q} \in \mathcal{P}_2 : \big(\mathbb{E}_\mathbb{Q}[\xi], \mathbb{E}_\mathbb{Q}[\xi\xi^\top]\big) \in \mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})\big\}.$$

Thus, the Gelbrich hull $\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})$ can be expressed as the pre-image of the convex set $\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})$ under a linear transformation, which shows that it is actually convex. This fact is summarized in the following corollary whose proof is omitted.

Corollary 3.19 (Convexity of the Gelbrich hull)**.** *For any $\widehat{\mu} \in \mathbb{R}^n$, $\widehat{\Sigma} \in \mathbb{S}_+^n$ and $\rho \in \mathbb{R}_+$, the Gelbrich hull $\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})$ is convex.*

3.3.3 Support functions of $\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ and $\mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})$

The uncertainty set $\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ can conveniently be used in classical robust optimization. Indeed, a robust constraint that requires a concave function $h(\mu,\Sigma)$ to be nonpositive for all $(\mu,\Sigma) \in \mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ can be reformulated as a convex constraint that involves the concave conjugate $h_\star$ of h and the support function of the uncertainty set $\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ [10, Theorem 2], that is,

$$h(\mu,\Sigma) \le 0 \quad \forall (\mu,\Sigma) \in \mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma}) \quad \Longleftrightarrow \exists\, q \in \mathbb{R}^n, Q \in \mathbb{S}^n \text{ such that } \delta^\star_{\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(q,Q) - h_\star(q,Q) \le 0,$$

This constraint is computationally tractable for many commonly used constraint functions because the support function of $\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ is conic representable.

Lemma 3.20 (Support function of $\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$)**.** *The support function of $\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ coincides with the optimal value of a tractable SDP, that is, for any $\rho \in \mathbb{R}_+$, $q \in \mathbb{R}^n$ and $Q \in \mathbb{S}^n$, we have*

$$\begin{array}{rcl} \delta^\star_{\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(q,Q) & = & \inf \quad \widehat{\mu}^\top q + \tau + \gamma\big(\rho^2 - \mathrm{Tr}\,[\widehat{\Sigma}]\big) + \mathrm{Tr}\,[Z] \\ & & \text{s.t.} \quad \gamma \in \mathbb{R}_+,\ \tau \in \mathbb{R}_+,\ Z \in \mathbb{S}_+^n \\ & & \qquad \begin{bmatrix} \gamma I - Q & \gamma \widehat{\Sigma}^{\frac{1}{2}} \\ \gamma \widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0, \quad \left\| \begin{pmatrix} \|q\| \\ \tau - \gamma \end{pmatrix} \right\| \le \tau + \gamma. \end{array}$$

Proof. Evaluating the support function of $\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ at $(q,Q) \in \mathbb{R}^n \times \mathbb{S}^n$ amounts to solving the finite convex program

$$\delta^\star_{\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(q,Q) = \left\{ \begin{array}{ll} \sup\limits_{\mu,\Sigma \succeq 0} & q^\top \mu + \mathrm{Tr}\,[Q\Sigma] \\ \text{s.t.} & \|\mu - \widehat{\mu}\|^2 + \mathrm{Tr}\,\big[\Sigma + \widehat{\Sigma} - 2\big(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\big] \le \rho^2. \end{array} \right. \tag{3.11}$$

Suppose that $\rho > 0$. Using duality arguments, we find

$$\begin{aligned} & \delta^\star_{\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(q,Q) \\ & = \sup_{\mu,\Sigma \succeq 0} \inf_{\gamma \ge 0} \quad q^\top \mu + \mathrm{Tr}\,[Q\Sigma] + \gamma \Big[\rho^2 - \|\mu - \widehat{\mu}\|^2 - \mathrm{Tr}\,\big[\Sigma + \widehat{\Sigma} - 2\big(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\big]\Big] \\ & = \inf_{\gamma \ge 0} \sup_{\mu,\Sigma \succeq 0} \quad q^\top \mu + \mathrm{Tr}\,[Q\Sigma] + \gamma \Big[\rho^2 - \|\mu - \widehat{\mu}\|^2 - \mathrm{Tr}\,\big[\Sigma + \widehat{\Sigma} - 2\big(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\big]\Big] \\ & = \inf_{\gamma \ge 0} \Big\{ \gamma\big(\rho^2 - \mathrm{Tr}\,[\widehat{\Sigma}]\big) + \sup_{\mu} \big\{ q^\top \mu - \gamma \|\mu - \widehat{\mu}\|^2 \big\} + \sup_{\Sigma \succeq 0} \Big\{ \mathrm{Tr}\,[(Q - \gamma I)\Sigma] + 2\gamma\, \mathrm{Tr}\,\big[\big(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\big] \Big\} \Big\}, \end{aligned} \tag{3.12}$$

where the second equality follows from strong duality, which holds because $\rho > 0$ and $(\widehat{\mu},\widehat{\Sigma})$ constitutes a Slater point for the primal problem. The subproblem in μ is a concave maximization problem, thus we can reformulate its optimal value by the epigraph formulation

$$\sup_{\mu} \{q^\top \mu - \gamma\|\mu - \widehat{\mu}\|^2\} \le \eta \iff \begin{bmatrix} \gamma I & \gamma\widehat{\mu} + \frac{q}{2} \\ (\gamma\widehat{\mu} + \frac{q}{2})^\top & \gamma\|\widehat{\mu}\|^2 + \eta \end{bmatrix} \succeq 0.$$

Due to its arrow-shaped structure, the above SDP constraint can be simplified to a second-order cone constraint, and thus we have

$$\sup_{\mu} \{q^\top \mu - \gamma\|\mu - \widehat{\mu}\|^2\} \le \eta \quad \iff \quad \eta \ge \widehat{\mu}^\top q + \tau, \ \left\| \begin{pmatrix} \|q\| \\ \tau - \gamma \end{pmatrix} \right\| \le \tau + \gamma, \ \tau \in \mathbb{R}_+.$$

The maximization subproblem over Σ can be solved analytically by using Proposition 3.54, and thus problem (3.12) is equivalent to

$$\begin{array}{cl} \inf\limits_{\gamma \ge 0, \tau \ge 0} & \widehat{\mu}^\top q + \tau + \gamma\big(\rho^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\big) + \gamma^2\,\mathrm{Tr}\big[(\gamma I - Q)^{-1}\widehat{\Sigma}\big] \\ \text{s.t.} & \gamma I \succ Q, \ \left\| \begin{pmatrix} \|q\| \\ \tau - \gamma \end{pmatrix} \right\| \le \tau + \gamma. \end{array} \tag{3.13}$$

Notice that the reformulation (3.13) is valid for $\rho \ge 0$ because $\delta^\star_{\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(q, Q)$ defined in (3.11) is continuous over $\rho \in \mathbb{R}_+$, and the optimal value of the infimum problem (3.13) is also continuous over $\rho \in \mathbb{R}_+$.

To complete the proof, we provide the semidefinite program reformulation for the nonlinear term in the objective function of (3.13). Define momentarily the extended real-valued function $f(\gamma) \triangleq \gamma^2\,\mathrm{Tr}\big[(\gamma I - Q)^{-1}\widehat{\Sigma}\big]$ over the domain $\{\gamma \in \mathbb{R}_+ : \gamma I \succ Q\}$, and $f(\gamma)$ is interpreted as ∞ outside this domain. For any γ such that $\gamma I \succ Q$, we can write

$$f(\gamma) = \min_{Z \succeq 0} \Big\{ \mathrm{Tr}\big[Z\big] \,:\, Z \succeq \gamma^2 \widehat{\Sigma}^{\frac{1}{2}} (\gamma I - Q)^{-1} \widehat{\Sigma}^{\frac{1}{2}} \Big\} = \min_{Z \succeq 0} \left\{ \mathrm{Tr}\big[Z\big] \,:\, \begin{bmatrix} \gamma I - Q & \gamma\widehat{\Sigma}^{\frac{1}{2}} \\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0 \right\},$$

where the first equality is from the cyclicity property of the trace operator and the fact that $A \succeq B$ implies that $\mathrm{Tr}\big[A\big] \ge \mathrm{Tr}\big[B\big]$, and the second equality is from the Schur complement argument [18, §A.5.5]. Thus, we find

$$\delta^\star_{\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(q, Q) = \left\{ \begin{array}{cl} \inf & \widehat{\mu}^\top q + \tau + \gamma\big(\rho^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\big) + \mathrm{Tr}\big[Z\big] \\ \text{s.t.} & \gamma \in \mathbb{R}_+, \ \tau \in \mathbb{R}_+, \ Z \in \mathbb{S}^n_+ \\ & \begin{bmatrix} \gamma I - Q & \gamma\widehat{\Sigma}^{\frac{1}{2}} \\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0, \ \left\| \begin{pmatrix} \|q\| \\ \tau - \gamma \end{pmatrix} \right\| \le \tau + \gamma, \ \gamma I \succ Q. \end{array} \right.$$

Because the objective function of the above program is continuous, we can replace the constraint $\gamma I \succ Q$ by $\gamma I \succeq Q$. Moreover, the constraint $\gamma I \succeq Q$ can be omitted because the first semidefinite constraint already implies that $\gamma I \succeq Q$. The proof is completed. □

The next lemma establishes the optimal mean vector and the covariance matrix that solves the optimization problem involved in the evaluation of the support function of $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$.

Lemma 3.21 (Extremal mean vectors and covariance matrices)**.** *Suppose that* $\widehat{\Sigma} \succ 0$ *and either*

- $q \neq 0$*, or*
- $\lambda_{\max}(Q) > 0$*, or*
- $Q = -LL^\top \neq 0$ *for some* $L \in \mathbb{R}^{n\times k}$ *and* $\rho^2 \leq \operatorname{Tr}\big[\widehat{\Sigma}L(L^\top L)^{-1}L^\top\big]$.

Then there exists $\gamma^\star \geq 0$ *with* $\gamma^\star I \succ Q$ *that solves the nonlinear algebraic equation*

$$\frac{\|q\|^2}{4\gamma^2} + \operatorname{Tr}\big[\widehat{\Sigma}\big(I - \gamma(\gamma I - Q)^{-1}\big)^2\big] = \rho^2, \tag{3.14a}$$

and the value $\delta^\star_{\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(q, Q)$ *is attained by* $(\mu^\star, \Sigma^\star) \in \mathbb{R}^n \times \mathbb{S}^n_{++}$ *satisfying*

$$\mu^\star = \widehat{\mu} + \frac{q}{2\gamma^\star}, \quad \Sigma^\star = \left(I - \frac{Q}{\gamma^\star}\right)^{-1} \widehat{\Sigma} \left(I - \frac{Q}{\gamma^\star}\right)^{-1}. \tag{3.14b}$$

Moreoever, if $Q \succeq 0$ *then we have* $\Sigma^\star \succeq \lambda_{\min}(\widehat{\Sigma}) I$.

Proof. It suffices to show that $(\mu^\star, \Sigma^\star)$ solves the maximization problem (3.11). If $\rho = 0$, the asymptotic value $\gamma^\star = +\infty$ solves (3.14a), thus we have $\mu^\star = \widehat{\mu}$, $\Sigma^\star = \widehat{\Sigma}$ and $(\mu^\star, \Sigma^\star)$ is trivially optimal for problem (3.11) when $\rho = 0$. It remains to show for the case $\rho > 0$.

First, consider the case where $q \neq 0$. From the proof of Lemma 3.20, evaluating the support function of $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ is tantamount to solving problem (3.13) which can be expressed in the equivalent form

$$\delta^\star_{\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(q, Q) = \left\{ \begin{array}{ll} \inf & \widehat{\mu}^\top q + \frac{\|q\|^2}{4\gamma} + \gamma\big(\rho^2 - \operatorname{Tr}[\widehat{\Sigma}]\big) + \gamma^2 \operatorname{Tr}\big[(\gamma I - Q)^{-1}\widehat{\Sigma}\big] \\ \text{s.t.} & \gamma \in \mathbb{R}_{++},\ \gamma I \succ Q \end{array} \right. \tag{3.15}$$

because $\|q\| \neq 0$. Denote momentarily the objective function of problem (3.15) as $f(\gamma)$. Because $\widehat{\Sigma} \succ 0$, the objective value of (3.15) evaluated at any feasible solution γ is lower bounded by

$$f(\gamma) \geq q^\top \widehat{\mu} + \frac{\|q\|^2}{4\gamma} + \gamma\big(\rho^2 - \operatorname{Tr}[\widehat{\Sigma}]\big) + \lambda_{\min}(\widehat{\Sigma})\gamma^2 \operatorname{Tr}\big[(\gamma I - Q)^{-1}\big]$$

and thus as γ approaches $\max\{0, \lambda_{\max}(Q)\}$, the objective value of (3.15) tends to ∞. Thus, the constraints $\gamma > 0$ and $\gamma I \succ Q$ in (3.15) becomes redundant. Moreover, the gradient of f can be written as

$$\begin{aligned} \nabla_\gamma f &= -\frac{\|q\|^2}{4\gamma^2} + \rho^2 - \operatorname{Tr}[\widehat{\Sigma}] + 2\gamma \operatorname{Tr}\big[(\gamma I - Q)^{-1}\widehat{\Sigma}\big] - \gamma^2 \operatorname{Tr}\big[(\gamma I - Q)^{-2}\widehat{\Sigma}\big] \\ &= \rho^2 - \frac{\|q\|^2}{4\gamma^2} - \operatorname{Tr}\big[\widehat{\Sigma}\big(I - \gamma(\gamma I - Q)^{-1}\big)^2\big]. \end{aligned}$$

This implies that if $\gamma^\star > \max\{0, \lambda_{\max}(Q)\}$ solves the nonlinear algebraic equation (3.14a), then $\gamma^\star$ is also the solution of the first-order optimality condition of problem (3.15), and thus $\gamma^\star$ is the optimal solution of (3.15). Furthermore, one can constate that as γ approaches $\max\{0, \lambda_{\max}(Q)\}$, $\nabla_\gamma f$ tends to $-\infty$, and as γ tends to infinity, $\nabla_\gamma f$ tends to $\rho^2 > 0$. This asserts that there exists a finite value $\gamma^\star$ that solves (3.14a).

In the sequel, we verify that $(\mu^\star, \Sigma^\star)$ defined in (3.14b) is the optimal solution of problem (3.11). Notice that

$$\mathbb{G}\big((\mu^\star, \Sigma^\star), (\widehat{\mu}, \widehat{\Sigma})\big)^2 = \frac{\|q\|^2}{4(\gamma^\star)^2} + \operatorname{Tr}\big[\widehat{\Sigma}(I - \gamma^\star(\gamma^\star I - Q)^{-1})^2\big] = \rho^2,$$

where the first equality follows from the definition of $(\mu^\star, \Sigma^\star)$ in (3.14b), and the second equality holds because $\gamma^\star$ solves (3.14a). As such, $(\mu^\star, \Sigma^\star)$ is feasible for the optimization problem (3.11). Furthermore, the objective value of $(\mu^\star, \Sigma^\star)$ in (3.11) amounts to

$$\begin{aligned}
q^\top \mu^\star + \operatorname{Tr}\big[Q\Sigma^\star\big] &= q^\top \widehat{\mu} + \frac{\|q\|^2}{2\gamma^\star} + (\gamma^\star)^2 \operatorname{Tr}\big[Q(\gamma^\star I - Q)^{-1}\widehat{\Sigma}(\gamma^\star I - Q)^{-1}\big] \\
&= q^\top \widehat{\mu} + \frac{\|q\|^2}{2\gamma^\star} + (\gamma^\star)^2 \operatorname{Tr}\big[(Q - \gamma^\star I + \gamma^\star I)(\gamma^\star I - Q)^{-1}\widehat{\Sigma}(\gamma^\star I - Q)^{-1}\big] \\
&= q^\top \widehat{\mu} + \frac{\|q\|^2}{2\gamma^\star} - (\gamma^\star)^2 \operatorname{Tr}\big[\widehat{\Sigma}(\gamma^\star I - Q)^{-1}\big] + (\gamma^\star)^3 \operatorname{Tr}\big[\widehat{\Sigma}(\gamma I - Q)^{-2}\big] \\
&= q^\top \widehat{\mu} + \frac{\|q\|^2}{2\gamma^\star} + \gamma^\star \left(\rho^2 - \frac{\|q\|^2}{4(\gamma^\star)^2} - \operatorname{Tr}\big[\widehat{\Sigma}\big]\right) + (\gamma^\star)^2 \operatorname{Tr}\big[(\gamma^\star I - Q)^{-1}\widehat{\Sigma}\big] \\
&= q^\top \widehat{\mu} + \frac{\|q\|^2}{4\gamma^\star} + \gamma^\star\big(\rho^2 - \operatorname{Tr}\big[\widehat{\Sigma}\big]\big) + (\gamma^\star)^2 \operatorname{Tr}\big[(\gamma^\star I - Q)^{-1}\widehat{\Sigma}\big] = \delta^\star_{\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})}(q, Q),
\end{aligned}$$

where the fourth equality is from the fact that $\gamma^\star > 0$ solves (3.14a) and thus the relationship

$$-(\gamma^\star)^2 \operatorname{Tr}\big[\widehat{\Sigma}(\gamma^\star I - Q)^{-1}\big] + (\gamma^\star)^3 \operatorname{Tr}\big[\widehat{\Sigma}(\gamma I - Q)^{-2}\big] = \gamma^\star \left(\rho^2 - \frac{\|q\|^2}{4(\gamma^\star)^2} - \operatorname{Tr}\big[\widehat{\Sigma}\big]\right) + (\gamma^\star)^2 \operatorname{Tr}\big[(\gamma^\star I - Q)^{-1}\widehat{\Sigma}\big]$$

holds. Finally, the last equality follows from the optimality of $\gamma^\star$ in (3.15). Thus, $(\mu^\star, \Sigma^\star)$ is optimal in (3.11).

In the second case, consider when $q = 0$ and $\lambda_{\max}(Q) > 0$. In this case, we can follow the steps as in the case of $q \neq 0$ almost verbatim to arrive at the same conclusion, the only difference is that now $\|q\| = 0$ and the related fractional term can be dropped from the calculation.

In the last case, consider when $q = 0$, $\lambda_{\max}(Q) \leq 0$ and $\rho^2 \leq \operatorname{Tr}\big[\widehat{\Sigma} L (L^\top L)^{-1} L^\top\big]$. If we use I_k to denote the k-by-k identity matrix, then we have

$$\begin{aligned}
\operatorname{Tr}\big[\widehat{\Sigma}(I - (I - \gamma^{-1}Q)^{-1})^2\big] &= \operatorname{Tr}\big[\widehat{\Sigma}(I - (I + \gamma^{-1}LL^\top)^{-1})^2\big] = \operatorname{Tr}\big[\widehat{\Sigma}(I - (I - L(\gamma I_k + L^\top L)^{-1}L^\top))^2\big] \\
&= \operatorname{Tr}\big[\widehat{\Sigma}(L(\gamma I_k + L^\top L)^{-1}L^\top)^2\big],
\end{aligned}$$

where the second equality utilizes the Woodbury matrix inversion formula [12, Corollary 2.8.8]. One can now verify that as γ approaches 0, $\nabla_\gamma f$ tends to $\rho^2 - \mathrm{Tr}\left[\widehat{\Sigma} L (L^\top L)^{-1} L^\top\right] \le 0$, thus there exists $\gamma^\star \in \mathbb{R}_{++}$ that solves the algebraic equation (3.14a).

To complete the proof, we notice that when $Q \succeq 0$ then $(I - (\gamma^\star)^{-1} Q)^{-1} \succeq I$ because $\gamma^\star I \succ Q$, and thus it is easy to verify that $\Sigma^\star$ defined in (3.14b) satisfies $\Sigma^\star \succeq \lambda_{\min}(\widehat{\Sigma}) I$. □

Thanks to its convexity established in Proposition 3.17, the uncertainty set $\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})$ can again conveniently be used in classical robust optimization. Indeed, a robust constraint that requires a concave function $h(\mu, M)$ to be nonpositive for all $(\mu, M) \in \mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})$ can be reformulated as a simple convex constraint involving the concave conjugate of $h(\mu, M)$ and the support function of $\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})$. This constraint is computationally tractable for many commonly used constraint functions because the support function of $\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})$ is SDP-representable.

Lemma 3.22 (Support function of $\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})$)**.** *The support function of $\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})$ coincides with the optimal value of a tractable SDP, that is, for any $\rho \in \mathbb{R}_+$, $q \in \mathbb{R}^n$ and $Q \in \mathbb{S}^n$, we have*

$$\begin{array}{rll} \delta^\star_{\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(q, Q) = & \inf & \gamma\big(\rho^2 - \|\widehat{\mu}\|^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\big) + z + \mathrm{Tr}\big[Z\big] \\ & \text{s.t.} & \gamma \in \mathbb{R}_+,\ z \in \mathbb{R}_+,\ Z \in \mathbb{S}^n_+ \\ & & \begin{bmatrix} \gamma I - Q & \gamma \widehat{\Sigma}^{\frac{1}{2}} \\ \gamma \widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0,\ \begin{bmatrix} \gamma I - Q & \gamma \widehat{\mu} + \frac{q}{2} \\ (\gamma \widehat{\mu} + \frac{q}{2})^\top & z \end{bmatrix} \succeq 0. \end{array} \tag{3.16}$$

Proof. Evaluating the support function of $\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})$ at $(q, Q) \in \mathbb{R}^n \times \mathbb{S}^n$ amounts to solving a maximization problem

$$\delta^\star_{\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(q, Q) = \left\{ \begin{array}{ll} \sup & \mu^\top q + \mathrm{Tr}\big[(\Sigma + \mu\mu^\top) Q\big] \\ \text{s.t.} & \mu \in \mathbb{R}^n,\ \Sigma \in \mathbb{S}^n_+ \\ & \|\mu - \widehat{\mu}\|^2 + \mathrm{Tr}\big[\Sigma + \widehat{\Sigma} - 2\big(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\big] \le \rho^2 \end{array} \right. \tag{3.17}$$

whose objective function is convex when $Q \succeq 0$. Consider now the optimization problem

$$\mathscr{J}(\varepsilon) \triangleq \left\{ \begin{array}{ll} \sup & \mu^\top q + \mathrm{Tr}\big[(\Sigma + \mu\mu^\top) Q\big] \\ \text{s.t.} & \mu \in \mathbb{R}^n,\ \Sigma \in \mathbb{S}^n_+ \\ & \|\mu - \widehat{\mu}\|^2 + \mathrm{Tr}\big[\Sigma + \widehat{\Sigma} - 2\big(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\big] \le \rho^2 + \varepsilon \\ & \|\mu - \widehat{\mu}\|^2 \le \rho^2, \end{array} \right.$$

parametrized by $\varepsilon \ge 0$, where we have made the dependence of $\mathscr{J}$ on $\widehat{\mu}, \widehat{\Sigma}$ and ρ implicit to avoid clutter. By construction, we have $\delta^\star_{\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(q, Q) \le \mathscr{J}(\varepsilon)$ for any $\varepsilon \ge 0$ and $\delta^\star_{\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(q, Q) = \mathscr{J}(0)$.

If $\rho > 0$ and $\varepsilon > 0$, we have

$$\begin{aligned}
\mathscr{J}(\varepsilon) &= \left\{ \begin{array}{lll} \sup\limits_{\mu:\|\mu-\widehat{\mu}\|\leq\rho^2} & \sup\limits_{\Sigma\succeq 0} & \mu^\top q + \mathrm{Tr}\left[(\Sigma+\mu\mu^\top)Q\right] \\ & \text{s.t.} & \mathrm{Tr}\left[\Sigma+\widehat{\Sigma}-2\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right] \leq \rho^2+\varepsilon-\|\mu-\widehat{\mu}\|^2 \end{array}\right. \\
&= \sup_{\mu:\|\mu-\widehat{\mu}\|\leq\rho^2} \sup_{\Sigma\succeq 0} \inf_{\nu\geq 0} \; \mu^\top q + \mathrm{Tr}\left[(\Sigma+\mu\mu^\top)Q\right] + \nu\big(\rho^2+\varepsilon-\|\mu-\widehat{\mu}\|^2 - \mathrm{Tr}\left[\Sigma+\widehat{\Sigma}-2\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right]\big) \\
&= \sup_{\mu:\|\mu-\widehat{\mu}\|\leq\rho^2} \inf_{\nu\geq 0} \sup_{\Sigma\succeq 0} \; \mu^\top q + \mathrm{Tr}\left[(\Sigma+\mu\mu^\top)Q\right] + \nu\big(\rho^2+\varepsilon-\|\mu-\widehat{\mu}\|^2 - \mathrm{Tr}\left[\Sigma+\widehat{\Sigma}-2\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right]\big) \\
&= \sup_{\mu:\|\mu-\widehat{\mu}\|\leq\rho^2} \inf_{\nu\geq 0} \; \Big\{\mu^\top q + \mu^\top Q\mu + \nu\big(\rho^2+\varepsilon-\|\mu-\widehat{\mu}\|^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\big) \\
&\qquad\qquad + \sup_{\Sigma\succeq 0}\Big\{\mathrm{Tr}\left[(Q-\nu I)\Sigma\right] + 2\nu\,\mathrm{Tr}\left[\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right]\Big\}\Big\}
\end{aligned}$$

where the third equality follows from strong duality which holds because $\varepsilon > 0$ and $\widehat{\Sigma}$ constitutes a Slater point for the primal problem. We can apply Proposition 3.54 to solve analytically the inner-most supremum problem over Σ

$$\begin{aligned}
\mathscr{J}(\varepsilon) &= \sup_{\mu:\|\mu-\widehat{\mu}\|\leq\rho^2} \inf_{\substack{\nu\geq 0\\ \nu I\succ Q}} \; \mu^\top q + \mu^\top Q\mu + \nu\big(\rho^2+\varepsilon-\|\mu-\widehat{\mu}\|^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\big) + \nu^2\,\mathrm{Tr}\left[(\nu I-Q)^{-1}\widehat{\Sigma}\right] \\
&= \inf_{\substack{\nu\geq 0\\ \nu I\succ Q}} \sup_{\mu:\|\mu-\widehat{\mu}\|\leq\rho^2} \; \mu^\top q + \mu^\top Q\mu + \nu\big(\rho^2+\varepsilon-\|\mu-\widehat{\mu}\|^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\big) + \nu^2\,\mathrm{Tr}\left[(\nu I-Q)^{-1}\widehat{\Sigma}\right],
\end{aligned}$$

where the second equality exploits Sion's minimax theorem [156, Corollary 3.3]. As the objective function of the resulting minimax problem is concave in μ for any fixed ν satisfying $\nu I \succ Q$ and as $\rho > 0$, the inner maximization problem constitutes a strictly feasible convex optimization problem. By strong duality, we have

$$\begin{aligned}
\mathscr{J}(\varepsilon) = \inf_{\substack{\nu\geq 0,\lambda\geq 0\\ \nu I\succ Q}} \Big\{(\nu+\lambda)\rho^2 + \nu(\varepsilon - \mathrm{Tr}\big[\widehat{\Sigma}\big]) + \nu^2\,\mathrm{Tr}\left[(\nu I-Q)^{-1}\widehat{\Sigma}\right] \\
+ \sup_{\mu}\left\{\mu^\top q + \mu^\top Q\mu - (\nu+\lambda)\|\mu-\widehat{\mu}\|^2\right\}\Big\}.
\end{aligned}$$

By introducing an epigraphical variable τ for the supremum over μ and by defining $\gamma = \nu+\lambda \geq 0$, we can rewrite $\mathscr{J}(\varepsilon)$ as

$$\mathscr{J}(\varepsilon) = \left\{ \begin{array}{ll} \inf & \gamma\rho^2 + \nu(\varepsilon - \mathrm{Tr}\big[\widehat{\Sigma}\big]) + \nu^2\,\mathrm{Tr}\left[(\nu I-Q)^{-1}\widehat{\Sigma}\right] + \tau \\ \text{s.t.} & \nu\in\mathbb{R}_+,\ \tau\in\mathbb{R}_+,\ \gamma\in\mathbb{R}_+ \\ & \begin{bmatrix} \gamma I - Q & \gamma\widehat{\mu}+\frac{q}{2} \\ (\gamma\widehat{\mu}+\frac{q}{2})^\top & \gamma\|\widehat{\mu}\|^2+\tau \end{bmatrix} \succeq 0,\ \gamma\geq\nu,\ \nu I\succ Q. \end{array}\right. \tag{3.19}$$

Let $\bar{\mathscr{J}}(\varepsilon)$ be the optimal value of the infimum program on the right hand side of (3.19). So far, we have shown that $\mathscr{J}(\varepsilon) = \bar{\mathscr{J}}(\varepsilon)$ for any $\varepsilon > 0$. We now establish the equality when $\varepsilon = 0$. By Berge's maximum principle [11, pp. 115–116], $\mathscr{J}(\varepsilon)$ is continuous over $\mathbb{R}_+$. Applying a similar argument as in the proof of [123, Theorem 2.8], we can show that $\bar{\mathscr{J}}(\varepsilon)$ is continuous over $\mathbb{R}_+$.

As such, we can conclude that

$$\mathscr{J}(0) = \lim_{\varepsilon \downarrow 0} \mathscr{J}(\varepsilon) = \lim_{\varepsilon \downarrow 0} \bar{\mathscr{J}}(\varepsilon) = \bar{\mathscr{J}}(0),$$

where the equalities are from the continuity of $\mathscr{J}$, the fact that $\mathscr{J}(\varepsilon) = \bar{\mathscr{J}}(\varepsilon)$ for any $\varepsilon > 0$, and the continuity of $\bar{\mathscr{J}}$. Because $\delta^\star_{\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(q,Q) = \mathscr{J}(0)$, we find for $\widehat{\Sigma} \succ 0$ and $\rho > 0$

$$\delta^\star_{\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(q,Q) = \left\{ \begin{array}{ll} \inf & \gamma\rho^2 - \nu \operatorname{Tr}\big[\widehat{\Sigma}\big] + \nu^2 \operatorname{Tr}\big[(\nu I - Q)^{-1}\widehat{\Sigma}\big] + \tau \\ \text{s.t.} & \nu \in \mathbb{R}_+,\ \tau \in \mathbb{R}_+,\ \gamma \in \mathbb{R}_+ \\ & \begin{bmatrix} \gamma I - Q & \gamma\widehat{\mu} + \frac{q}{2} \\ (\gamma\widehat{\mu} + \frac{q}{2})^\top & \gamma\|\widehat{\mu}\|^2 + \tau \end{bmatrix} \succeq 0,\ \gamma \geq \nu,\ \nu I \succ Q. \end{array} \right. \tag{3.20}$$

In the next step, we proceed to eliminate the variable ν from the above optimization problem. Problem (3.20) can be re-express in the following equivalent form

$$\begin{array}{ll} \inf & \gamma\rho^2 + \tau + f(\gamma) \\ \text{s.t.} & \tau \in \mathbb{R}_+,\ \gamma \in \mathbb{R}_+ \\ & \begin{bmatrix} \gamma I - Q & \gamma\widehat{\mu} + \frac{q}{2} \\ (\gamma\widehat{\mu} + \frac{q}{2})^\top & \gamma\|\widehat{\mu}\|^2 + \tau \end{bmatrix} \succeq 0,\ \gamma I \succ Q, \end{array} \tag{3.21}$$

where the function $f : \mathbb{R}_+ \to \mathbb{R}$ is momentarily defined as

$$f(\gamma) = \inf\left\{-\nu \operatorname{Tr}\big[\widehat{\Sigma}\big] + \nu^2 \operatorname{Tr}\big[(\nu I - Q)^{-1}\widehat{\Sigma}\big] \,:\, 0 \leq \nu \leq \gamma,\ \nu I \succ Q\right\}.$$

Suppose that $Q \neq 0$ and thus Q can be written using its eigenvalue decomposition $Q = LDL^\top$ for some diagonal matrix $D \in \mathbb{R}^{k\times k}$ with non-zero diagonal elements, and $L \in \mathbb{R}^{n\times k}$ satisfying $L^\top L = I_k$, where we denote by I_k the k-by-k identity matrix. Suppose further that $\gamma > 0$. For any $0 < \nu \leq \gamma$, we have

$$\nu(\nu I - Q)^{-1} = \big(I - \nu^{-1} LDL^\top\big)^{-1} = I + L(\nu D^{-1} - I_k)^{-1} L^\top,$$

where the second equality follows from the Woodbury matrix inversion formula [12, Corollary 2.8.8]. Thus, for any $0 < \nu \leq \gamma$, we have

$$-\nu \operatorname{Tr}\big[\widehat{\Sigma}\big] + \nu^2 \operatorname{Tr}\big[(\nu I - Q)^{-1}\widehat{\Sigma}\big] = \nu \operatorname{Tr}\big[(\nu D^{-1} - I_k)^{-1} L^\top \widehat{\Sigma} L\big].$$

Notice that the above formula is also valid when $\nu = 0$, in which case both sides of the above equation evaluate to 0 and the evaluation of the left hand side is understood in the limit as ν tends to 0. This implies that whenever $\gamma > 0$, we have

$$f(\gamma) = \inf\left\{\nu \operatorname{Tr}\big[(\nu D^{-1} - I_k)^{-1} L^\top \widehat{\Sigma} L\big] :\ 0 \leq \nu \leq \gamma,\ \nu I_k \succ D\right\}.$$

Denote momentarily by $g(\nu)$ the objective function of the above program. It is easy to verify that $g(v)$ is continuous over $[0,\gamma]$, and for any $v > \max\{0, \lambda_{\max}(D)\}$, the gradient of g satisfies

$$g'(v) = -\operatorname{Tr}\left[(vD^{-1} - I_k)^{-1} L^\top \widehat{\Sigma} L (vD^{-1} - I_k)^{-1}\right] < 0$$

which implies that given any γ which is feasible in problem (3.21), the optimal value $f(\gamma)$ is attained by the optimal solution $v^\star(\gamma) = \gamma$. Evaluating the objective function g at this optimal solution gives

$$f(\gamma) = -\gamma \operatorname{Tr}\left[\widehat{\Sigma}\right] + \gamma^2 \operatorname{Tr}\left[(\gamma I - Q)^{-1}\widehat{\Sigma}\right].$$

Notice that when $\gamma = 0$, the feasible set of v becomes a singleton and the above expression for $f(\gamma)$ holds trivially. Hence, we conclude that when $Q \neq 0$, problem (3.20) is equivalent to

$$\delta^\star_{\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(q,Q) = \left\{ \begin{array}{ll} \inf & \gamma\big(\rho^2 - \operatorname{Tr}\left[\widehat{\Sigma}\right]\big) + \gamma^2 \operatorname{Tr}\left[(\gamma I - Q)^{-1}\widehat{\Sigma}\right] + \tau \\ \text{s.t.} & \tau \in \mathbb{R}_+,\ \gamma \in \mathbb{R}_+ \\ & \begin{bmatrix} \gamma I - Q & \gamma\widehat{\mu} + \frac{q}{2} \\ (\gamma\widehat{\mu} + \frac{q}{2})^\top & \gamma\|\widehat{\mu}\|^2 + \tau \end{bmatrix} \succeq 0,\ \gamma I \succ Q. \end{array} \right. \tag{3.22}$$

One can readily verify that the reformulation (3.22) is also valid when $Q = 0$. Indeed, when $Q = 0$, problem (3.22) collapses into

$$\delta^\star_{\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(q,0) = \left\{ \begin{array}{ll} \inf & \gamma\rho^2 + \tau \\ \text{s.t.} & \tau \in \mathbb{R}_+,\ \gamma \in \mathbb{R}_+ \\ & \begin{bmatrix} \gamma I & \gamma\widehat{\mu} + \frac{q}{2} \\ (\gamma\widehat{\mu} + \frac{q}{2})^\top & \gamma\|\widehat{\mu}\|^2 + \tau \end{bmatrix} \succeq 0,\ \gamma > 0. \end{array} \right.$$

The optimal solution of this minimization problem can be shown to be $\gamma^\star = \|q\|/(2\rho)$ and $\tau^\star = \widehat{\mu}^\top q + \rho\|q\|/2$, with optimal value $\widehat{\mu}^\top q + \rho\|q\|$. This optimal value coincides with the value of the support function at $(q,0)$ evaluated using the definition of the support function

$$\delta^\star_{\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(q,0) = \sup\left\{\mu^\top q :\ \mu \in \mathbb{R}^n,\ \|\mu - \widehat{\mu}\|^2 \leq \rho^2\right\} = \widehat{\mu}^\top q + \rho\|q\|,$$

where the last equality follows from the property of the dual norm.

The extension of the reformulation (3.22) to the situation where $\rho \geq 0$ can be achieved by employing a similar continuity argument as in the proof of Lemma 3.20. The last step involves applying the Schur complement to reformulate the nonlinear term in the objective function of (3.22) as a linear semidefinite program constraint, and performing a variable substitution $z = \tau + \gamma\|\widehat{\mu}\|^2$. We thus find

$$\delta^\star_{\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(q,Q) = \left\{ \begin{array}{ll} \inf & \gamma\big(\rho^2 - \|\widehat{\mu}\|^2 - \operatorname{Tr}\left[\widehat{\Sigma}\right]\big) + z + \operatorname{Tr}\left[Z\right] \\ \text{s.t.} & \tau \in \mathbb{R}_+,\ \gamma \in \mathbb{R}_+,\ z \in \mathbb{R}_+,\ Z \in \mathbb{S}^n_+ \\ & \begin{bmatrix} \gamma I - Q & \gamma\widehat{\mu} + \frac{q}{2} \\ (\gamma\widehat{\mu} + \frac{q}{2})^\top & z \end{bmatrix} \succeq 0,\ \begin{bmatrix} \gamma I - Q & \gamma\widehat{\Sigma}^{\frac{1}{2}} \\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0,\ \gamma I \succ Q. \end{array} \right. \tag{3.23}$$

The last constraint $\gamma I \succ Q$ can be altered to $\gamma I \succeq Q$ because problem (3.23) has a continuous objective function, and this constraint $\gamma I \succeq Q$ can be dropped because the other semidefinite constraints of (3.23) necessarily imply that $\gamma I \succeq Q$. This completes the proof. □

As a parallel counterpart to Lemma 3.21, the next lemma establishes the result regarding the optimal mean vector and covariance matrix that solves the optimization problem involved in the evaluation of the support function of $\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})$.

Lemma 3.23 (Extremal mean vectors and covariance matrices). *Suppose that $\widehat{\Sigma} \succ 0$ and either*

- $\lambda_{\max}(Q) > 0$*, or*
- $\lambda_{\max}(Q) < 0$ *and* $\rho^2 \le \mathrm{Tr}\left[\widehat{\Sigma}\right] + \|\widehat{\mu} + Q^{-1}q/2\|^2$*, or*
- $Q = -LL^\top \neq 0$ *for some* $L \in \mathbb{R}^{n\times k}$ *and* $\rho^2 \le \mathrm{Tr}\left[\widehat{\Sigma}L(L^\top L)^{-1}L^\top\right]$.

Then there exists $\gamma^\star \ge 0$ with $\gamma^\star I \succ Q$ that solves the nonlinear algebraic equation

$$\|\widehat{\mu} - (\gamma I - Q)^{-1}(q/2 + \gamma\widehat{\mu})\|^2 + \mathrm{Tr}\left[\widehat{\Sigma}\left(I - \gamma(\gamma I - Q)^{-1}\right)^2\right] = \rho^2, \tag{3.24a}$$

and the value $\delta^\star_{\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(q, Q)$ is attained by $(\mu^\star, \Sigma^\star) \in \mathbb{R}^n \times \mathbb{S}^n_+$ satisfying

$$\mu^\star = (\gamma^\star I - Q)^{-1}\left(\gamma^\star\widehat{\mu} + \frac{q}{2}\right), \quad \Sigma^\star = \left(I - \frac{Q}{\gamma^\star}\right)^{-1}\widehat{\Sigma}\left(I - \frac{Q}{\gamma^\star}\right)^{-1}. \tag{3.24b}$$

Moreover, if $Q \succeq 0$ then we have $\Sigma^\star \succeq \lambda_{\min}(\widehat{\Sigma})I$.

Proof. It suffices to show that $(\mu^\star, \Sigma^\star)$ defined in (3.24b) is the maximizer of problem (3.17). If $\rho = 0$, the asymptotic value $\gamma^\star = +\infty$ solves (3.24a), thus we have $\mu^\star = \widehat{\mu}$, $\Sigma^\star = \widehat{\Sigma}$ and $(\mu^\star, \Sigma^\star)$ is trivially optimal for problem (3.17) when $\rho = 0$. It remains to prove for the case $\rho > 0$.

The proof of Lemma 3.22 implies that evaluating the support function of $\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})$ is equivalent to solving problem (3.23). Because $\widehat{\Sigma} \succ 0$, we can re-express the SDP (3.23) using the Schur complement reformulation of the SDP constraints as

$$\delta^\star_{\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(q, Q) = \left\{\begin{array}{ll} \inf & \gamma\left(\rho^2 - \|\widehat{\mu}\|^2 - \mathrm{Tr}\left[\widehat{\Sigma}\right]\right) + \gamma^2\,\mathrm{Tr}\left[(\gamma I - Q)^{-1}\widehat{\Sigma}\right] + (\gamma\widehat{\mu} + \frac{q}{2})^\top[\gamma I - Q]^{-1}(\gamma\widehat{\mu} + \frac{q}{2}) \\ \text{s.t.} & \gamma \in \mathbb{R}_+,\ \gamma I \succ Q. \end{array}\right. \tag{3.25}$$

Denote momentarily the objective function of (3.25) as $f(\gamma)$. The gradient of f for any feasible solution γ satisfies

$$\nabla_\gamma f = \rho^2 - \mathrm{Tr}\left[\widehat{\Sigma}\left(I - \gamma(\gamma I - Q)^{-1}\right)^2\right] - \left\|\widehat{\mu} - (\gamma I - Q)^{-1}\left(\gamma\widehat{\mu} + \frac{q}{2}\right)\right\|^2.$$

Thus if $\gamma^\star$ solves the nonlinear algebraic equation (3.24a), then $\gamma^\star$ also solves the first-order optimality condition of problem (3.25). This in turn implies that $\gamma^\star$ is the minimizer of problem (3.25).

First, consider the case where $\lambda_{\max}(Q) > 0$. One can verify that as γ approaches $\lambda_{\max}(Q)$, $\nabla_\gamma f$ tends to $-\infty$, and as γ tends to infinity, $\nabla_\gamma f$ tends to $\rho^2 > 0$. This asserts that there exists a finite value $\gamma^\star$ that solves (3.24a). We are now ready to show that $(\mu^\star, \Sigma^\star)$ is feasible and optimal in problem (3.17). Notice that

$$\mathbb{G}\big((\mu^\star,\Sigma^\star),(\widehat{\mu},\widehat{\Sigma})\big)^2 = \left\|(\gamma^\star I - Q)^{-1}\left(\gamma^\star\widehat{\mu}+\frac{q}{2}\right)-\widehat{\mu}\right\|^2 + \mathrm{Tr}\left[\widehat{\Sigma}\left(I-\gamma^\star(\gamma^\star I-Q)^{-1}\right)^2\right] = \rho^2,$$

where the first equality follows from substituting the value of $\mu^\star$ and $\Sigma^\star$ from (3.24b) into the definition of the Gelbrich distance $\mathbb{G}$, and the second equality is because $\gamma^\star$ solves (3.24a). This implies that $(\mu^\star,\Sigma^\star) \in \mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$. Finally, we show that $(\mu^\star,\Sigma^\star)$ will attain the value $\delta^\star_{\mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(q,Q)$. We find

$$\begin{aligned}
\mathrm{Tr}\left[Q\Sigma^\star\right] &= \mathrm{Tr}\left[Q(\gamma^\star)^2(\gamma^\star I-Q)^{-1}\widehat{\Sigma}(\gamma^\star I-Q)^{-1}\right] && (3.26a)\\
&= \mathrm{Tr}\left[(Q-\gamma^\star I+\gamma^\star I)(\gamma^\star)^2(\gamma^\star I-Q)^{-1}\widehat{\Sigma}(\gamma^\star I-Q)^{-1}\right]\\
&= -(\gamma^\star)^2\,\mathrm{Tr}\left[(\gamma^\star I-Q)^{-1}\widehat{\Sigma}\right]+(\gamma^\star)^3\,\mathrm{Tr}\left[(\gamma^\star I-Q)^{-1}\widehat{\Sigma}(\gamma^\star I-Q)^{-1}\right]\\
&= \gamma^\star\big(\rho^2-\|\widehat{\mu}-(\gamma^\star I-Q)^{-1}(\gamma^\star\widehat{\mu}+q/2)\|^2-\mathrm{Tr}\left[\widehat{\Sigma}\right]+\gamma^\star\,\mathrm{Tr}\left[(\gamma^\star I-Q)^{-1}\widehat{\Sigma}\right]\big) && (3.26b)\\
&= \gamma^\star\big(\rho^2-\|\widehat{\mu}-\mu^\star\|^2-\mathrm{Tr}\left[\widehat{\Sigma}\right]+\gamma^\star\,\mathrm{Tr}\left[(\gamma^\star I-Q)^{-1}\widehat{\Sigma}\right]\big) && (3.26c)\\
&= \gamma^\star\big(\rho^2-\|\widehat{\mu}\|^2-\mathrm{Tr}\left[\widehat{\Sigma}\right]\big)+(\gamma^\star)^2\,\mathrm{Tr}\left[(\gamma^\star I-Q)^{-1}\widehat{\Sigma}\right]+2\gamma^\star\widehat{\mu}^\top\mu^\star-\gamma^\star\|\mu^\star\|^2,
\end{aligned}$$

where equality (3.26a) is from the definition of $\Sigma^\star$ in (3.24b), equality (3.26b) follows from the fact that $\gamma^\star$ solves (3.24a) and thus we can write

$$\begin{aligned}
&-\gamma^\star\,\mathrm{Tr}\left[(\gamma^\star I-Q)^{-1}\widehat{\Sigma}\right]+(\gamma^\star)^2\,\mathrm{Tr}\left[(\gamma^\star I-Q)^{-1}\widehat{\Sigma}(\gamma^\star I-Q)^{-1}\right]\\
&\qquad= \rho^2-\|\widehat{\mu}-(\gamma^\star I-Q)^{-1}(q/2+\gamma^\star\widehat{\mu})\|^2-\mathrm{Tr}\left[\widehat{\Sigma}\right]+\gamma^\star\,\mathrm{Tr}\left[(\gamma^\star I-Q)^{-1}\widehat{\Sigma}\right].
\end{aligned}$$

Finally, equality (3.26c) is from the definition of $\mu^\star$ in (3.24b). We thus find

$$\begin{aligned}
&q^\top\mu^\star+\mathrm{Tr}\left[Q(\Sigma^\star+\mu^\star(\mu^\star)^\top)\right]\\
=&\gamma^\star\big(\rho^2-\|\widehat{\mu}\|^2-\mathrm{Tr}\left[\widehat{\Sigma}\right]\big)+(\gamma^\star)^2\,\mathrm{Tr}\left[(\gamma^\star I-Q)^{-1}\widehat{\Sigma}\right]+2(\gamma^\star\widehat{\mu}+q/2)^\top\mu^\star+(\mu^\star)^\top(Q-\gamma^\star I)\mu^\star\\
=&\gamma^\star\big(\rho^2-\|\widehat{\mu}\|^2-\mathrm{Tr}\left[\widehat{\Sigma}\right]\big)+(\gamma^\star)^2\,\mathrm{Tr}\left[(\gamma^\star I-Q)^{-1}\widehat{\Sigma}\right]+(\gamma^\star\widehat{\mu}+q/2)^\top(\gamma^\star I-Q)^{-1}(\gamma^\star\widehat{\mu}+q/2)
\end{aligned}$$

which equals the optimal value of the dual program (3.25) because $\gamma^\star$ is also the minimizer of problem (3.25). This implies that $(\mu^\star,\Sigma^\star)$ is optimal for the support function evaluation problem of $\mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})$.

Consider the case where $\lambda_{\max}(Q) < 0$ and $\rho^2 \le \mathrm{Tr}\left[\widehat{\Sigma}\right] + \|\widehat{\mu}+Q^{-1}q/2\|^2$. One can verify that as γ approaches 0, $\nabla_\gamma f$ tends to $\rho^2 - \mathrm{Tr}\left[\widehat{\Sigma}\right] - \|\widehat{\mu}+Q^{-1}q/2\|^2 \le 0$, and as γ tends to infinity, $\nabla_\gamma f$ tends to $\rho^2 > 0$. This asserts that there exists a finite value $\gamma^\star$ that solves (3.24a). Showing that $(\mu^\star,\Sigma^\star)$ is feasible and optimal in (3.17) follows verbatim from the first part of the proof.

In the last case, consider the case where $Q = -LL^\top \neq 0$ for some $L \in \mathbb{R}^{n\times k}$ and $\rho^2 \le \mathrm{Tr}\left[\widehat{\Sigma}\right]$. The gradient of f can be re-expressed as

$$\begin{aligned}\nabla_\gamma f &= \rho^2 - \mathrm{Tr}\left[\widehat{\Sigma}\left(I-\gamma(\gamma I+LL^\top)^{-1}\right)^2\right] - \left\|\widehat{\mu}-(\gamma I+LL^\top)^{-1}\left(\gamma\widehat{\mu}+\frac{q}{2}\right)\right\|^2\\ &\leq \rho^2 - \mathrm{Tr}\left[\widehat{\Sigma}\left(I-\gamma(\gamma I+LL^\top)^{-1}\right)^2\right].\end{aligned}$$

Using the same calculation as in the proof of Lemma 3.21, we can show that as γ tends to 0, $\nabla_\gamma f$ tends to a non-positive value. This justifies the existence of $\gamma^\star$ that solves equation (3.24a).

To show that $\Sigma^\star \succeq \lambda_{\min}(\widehat{\Sigma})I$ whenever $Q \succeq 0$, we can employ an analogous reasoning to the last part of the proof of Lemma 3.21. This completes the proof. □

To give an intuition for Lemma 3.23, note that the SDP (3.16) can be converted to an equivalent nonlinear program (NLP) in the single decision variable γ by using Schur complements to show that

$$z = (q+\gamma\widehat{\mu})^\top(\gamma I-Q)^{-1}(q+\gamma\widehat{\mu}) \quad \text{and} \quad Z = \gamma^2\,\widehat{\Sigma}^{\frac{1}{2}}(\gamma I-Q)^{-1}\widehat{\Sigma}^{\frac{1}{2}}$$

at optimality. The resulting NLP minimizes a strictly convex objective function that explodes as γ drops to $\lambda_{\max}(Q)$ or as γ tends to infinity. Equation (3.24a) represents its first-order optimality condition, whose unique solution $\gamma^\star$ can be computed efficiently to any precision via bisection or the Newton-Raphson method.

Remark 3.24 (On the existence of the solution of (3.24a)). *Consider the following one dimensional example with $\widehat{\Sigma} = 1$, $\widehat{\mu} = 0.5$, $q = 1$ and $Q = -1$. Evaluating the support function of $\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})$ is equivalent to solving*

$$\sup\left\{\mu - \|\mu\|^2 - \mathrm{Tr}\left[\Sigma\right] \;:\; \mu\in\mathbb{R},\; \Sigma\in\mathbb{R}_+,\; \|\mu-0.5\|^2 + \mathrm{Tr}\left[1+\Sigma-2\Sigma^{\frac{1}{2}}\right] \leq \rho^2\right\}.$$

If $\rho > 1$, we can easily verify that the optimal solution is $\mu^\star = 0.5$, $\Sigma^\star = 0$. In this case, the optimal solution $(\mu^\star,\Sigma^\star)$ lies strictly inside the feasible set prescribed by the Gelbrich distance constraint, and thus there is no solution to the first-order optimality condition. Indeed, (3.24a) *becomes*

$$\frac{1}{(\gamma+1)^2} = \rho^2,$$

which admits a solution only if $\rho^2 \leq 1$. This example also shows that the condition $q \neq 0$ is not sufficient to ensure the existence of $\gamma^\star$ that solves (3.24a).

3.4 Risk Measures of Linear Portfolios

We study in this section the family of consistent risk measures which enjoys widespread applications in risk management and finance [62]. We briefly review some basic properties of a risk measure $\mathscr{R}_{\mathbb{Q}}$.

Definition 3.25 (Properties of risk measures). *A risk measure $\mathscr{R}_{\mathbb{Q}} : \mathbb{L} \to \mathbb{R}\cup\{+\infty\}$ satisfies the condition of*

- ***translation invariance*** *if* $\mathscr{R}_{\mathbb{Q}}(\ell + \lambda) = \mathscr{R}_{\mathbb{Q}}(\ell) + \lambda$ *for any* $\ell \in \mathbb{L}$, $\lambda \in \mathbb{R}$.
- ***positive homogeneity*** *if* $\mathscr{R}_{\mathbb{Q}}(\lambda\ell) = \lambda\mathscr{R}_{\mathbb{Q}}(\ell)$ *for any* $\lambda \geq 0$ *and* $\ell \in \mathbb{L}$.
- ***monotonicty*** *if* $\mathscr{R}_{\mathbb{Q}}(\ell_1) \leq \mathscr{R}_{\mathbb{Q}}(\ell_2)$ *for any* $\ell_1, \ell_2 \in \mathbb{L}$ *and* $\ell_1 \leq \ell_2$ $\mathbb{Q}$*-almost surely.*
- ***convexity*** *if* $\mathscr{R}_{\mathbb{Q}}(\lambda\ell_1 + (1-\lambda)\ell_2) \leq \lambda\mathscr{R}_{\mathbb{Q}}(\ell_1) + (1-\lambda)\mathscr{R}_{\mathbb{Q}}(\ell_2)$ *for any* $\ell_1, \ell_2 \in \mathbb{L}, \lambda \in [0,1]$.

A risk measure $\mathscr{R}_{\mathbb{Q}}$ is monetary if it satisfies translation invariance and monotonicity. A monetary risk measure is convex if it additionally satisfies the convexity condition. A convex risk measure is coherent if it is also positive homogeneous.

We focus on a subset of loss functions that can be written as a linear function of the asset returns ξ of the form $\ell(\xi) = -w^\top \xi$ for some $w \in \mathbb{R}^n$. To facilitate a rigorous exposition, we define $\mathscr{D}$ as the (convex) set of cumulative distribution functions F on $\mathbb{R}$. More specifically, $\mathscr{D}$ contains all functions $F : \mathbb{R} \to [0,1]$, F is non-decreasing and right-continuous, $\lim_{t\downarrow -\infty} F(t) = 0, \lim_{t\uparrow +\infty} F(t) = 1$. For any information structure $\sigma \in \mathscr{S}$, we can in a straightforward manner define $\mathscr{D}_\sigma \subset \mathscr{D}$ as the space of *structural* distribution functions that satisfy the finite second moment (for $\sigma = 2$), symmetric (for $\sigma = \mathrm{S}$), symmetric and linearly unimodal (for $\sigma = \mathrm{SU}$), and the elliptical distribution with generator function ϕ (for $\sigma = \phi$) conditions correspondingly. Using similar notation, $\mathscr{D}_\sigma(m, s^2)$ will restrict $\mathscr{D}_\sigma$ to distribution functions whose underlying random variable has fixed mean $m \in \mathbb{R}$ and variance $s^2 \in \mathbb{R}_+$.

Under the assumption that the family of risk measures $\{\mathscr{R}_{\mathbb{Q}}\}_{\mathbb{Q}\in\mathscr{P}_\sigma}$ is consistent, there exists a *distributional risk measure* $\varrho : \mathscr{D}_\sigma \to \mathbb{R}$ defined over the space of cumulative distribution functions such that

$$\mathscr{R}_{\mathbb{Q}_1}(\ell_1) = \varrho(F_{\ell_1}^{\mathbb{Q}_1}) = \varrho(F_{\ell_2}^{\mathbb{Q}_2}) = \mathscr{R}_{\mathbb{Q}_2}(\ell_2) \quad \forall \mathbb{Q}_1, \mathbb{Q}_2 \in \mathscr{P}_\sigma,\ \forall \ell_1, \ell_2 \in \mathscr{L} \text{ such that } F_{\ell_1}^{\mathbb{Q}_1} = F_{\ell_2}^{\mathbb{Q}_2}.$$

In this case, we say that ϱ is translation invariant (monotone or positive homogeneous, respectively) if $\mathscr{R}_{\mathbb{Q}}$ is translation invariant (monotone or positive homogeneous, respectively) for any $\mathbb{Q} \in \mathscr{P}_\sigma$. We first establish an auxiliary projection result, pioneered by [140, 177], which binds the family of risk measures $\{\mathscr{R}_{\mathbb{Q}}\}_{\mathbb{Q}\in\mathscr{P}_\sigma}$ with its corresponding distribution function risk measure ϱ.

Proposition 3.26 (Univariate projection)**.** *For any* $\mu \in \mathbb{R}^n$ *and* $\Sigma \in \mathbb{S}_+^n$, *if* $\ell(\xi) = -w^\top\xi$ *for some* $w \in \mathbb{R}^n$, *then for any* $\sigma \in \mathscr{S}$, *we have*

$$\sup_{\mathbb{Q}\in\mathscr{P}_\sigma(\mu,\Sigma)} \mathscr{R}_{\mathbb{Q}}\left(-w^\top\xi\right) = \sup_{F\in\mathscr{D}_\sigma(-w^\top\mu,\, w^\top\Sigma w)} \varrho(F).$$

Proof. To simplify the notations, we use the shorthands $m = -w^\top\mu$ and $s^2 = w^\top\Sigma w \geq 0$. When $s = 0$, the distribution of $-w^\top\xi$ under any $\mathbb{Q} \in \mathscr{P}_\sigma(\mu, \Sigma)$ coincides with the distribution of a random variables with value $-w^\top\mu$ almost surely, and the claim holds trivially. It suffices thus to prove for the case when $s^2 > 0$.

In the first step, we prove the inclusion $\mathscr{D}_\sigma(m, s^2) \subseteq \{F^{\mathbb{Q}}_{-w^\top \xi} : \mathbb{Q} \in \mathscr{P}_\sigma(\mu, \Sigma)\}$. To this end, fix any $F \in \mathscr{D}_\sigma(m, s^2)$, we show that there exists a measure $\mathbb{Q} \in \mathscr{P}_\sigma(\mu, \Sigma)$ such that $F \equiv F^{\mathbb{Q}}_{-w^\top \xi}$. We prove this claim by construction. First, construct a probability space $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \nu_1)$ and a univariate random variable $\eta : \mathbb{R}^n \to \mathbb{R}$ such that $\nu_1(\eta \le t) = F(t)$, $\forall t \in \mathbb{R}$. Using the one-dimensional version of the mapping in the proof of [177, Theorem 1], we construct the auxiliary n-dimensional random vector $\zeta : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$\xi = -s^{-2}\Sigma w \eta + (I - s^{-2}\Sigma w w^\top)\zeta. \tag{3.27}$$

Construct another probability space $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \nu_n)$ such that under ν_n, the random vector ζ satisfies the information structure σ with mean vector μ and covariance matrix Σ. On the product probability space $(\mathbb{R}^n \times \mathbb{R}^n, \mathscr{B}(\mathbb{R}^n) \otimes \mathscr{B}(\mathbb{R}^n))$, we define the product measure $\nu_1 \times \nu_n$, the existence of which is guaranteed by the result of the Hahn-Kolmogorov theorem. On $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$, we can then deduce a probability measure $\mathbb{Q}$ from the product measure $\nu_1 \times \nu_n$ under which ξ satisfies the information structure σ with mean vector μ and covariance matrix Σ. The measure $\mathbb{Q}$ can be constructed using [30, Theorem 5.1.4] as

$$\begin{aligned}\mathbb{Q}(\xi \le \tau) &= \int_{\mathbb{R}} \nu_n\big((I - s^{-2}\Sigma w w^\top)\zeta \le \tau + s^{-2}\Sigma w t\big)\, \mathrm{d}\nu_1(\eta \le t)\\ &= \int_{\mathbb{R}} \nu_n\big((I - s^{-2}\Sigma w w^\top)\zeta \le \tau + s^{-2}\Sigma w t\big)\, \mathrm{d}F(t)\end{aligned}$$

for any $\tau \in \mathbb{R}^n$. By construction in (3.27), $-w^\top \xi = \eta$, and as a result, the distribution of $-w^\top \xi$ under $\mathbb{Q}$ is equivalent to F. From (3.27), it is easy to verify that under $\mathbb{Q}$, ξ satisfies the information structure σ with mean vector μ and covariance matrix Σ, so $\mathbb{Q} \in \mathscr{P}_\sigma(\mu, \Sigma)$. Thus for any distribution function $F \in \mathscr{D}_\sigma(m, s^2)$, there exists a measure $\mathbb{Q} \in \mathscr{P}_\sigma(\mu, \Sigma)$ such that $F \equiv F^{\mathbb{Q}}_{-w^\top \xi}$.

Next, we prove the inverse inclusion, i.e., we will show that $\{F^{\mathbb{Q}}_{-w^\top \xi} : \mathbb{Q} \in \mathscr{P}_\sigma(\mu, \Sigma)\} \subseteq \mathscr{D}_\sigma(m, s^2)$. Fix any measure $\mathbb{Q} \in \mathscr{P}_\sigma(\mu, \Sigma)$ and construct the cumulative distribution function F such that $F(t) = \mathbb{Q}(-w^\top \xi \le t)$ for any $t \in \mathbb{R}$. Because the information structure σ is preserved under linear combination, we find $F \in \mathscr{D}_\sigma$. Furthermore, by construction, F is the distribution function of a random variable with mean m and variance s^2. Thus, there exists $F \in \mathscr{D}_\sigma(\mu, s^2)$ such that $F \equiv F^{\mathbb{Q}}_{-w^\top \xi}$, so $\{F^{\mathbb{Q}}_{-w^\top \xi} : \mathbb{Q} \in \mathscr{P}_\sigma(\mu, \Sigma)\} \subseteq \mathscr{D}_\sigma(m, s^2)$ holds.

We have

$$\sup_{\mathbb{Q} \in \mathscr{P}_\sigma(\mu, \Sigma)} \mathscr{R}_{\mathbb{Q}}\big(-w^\top \xi\big) = \sup_{\mathbb{Q} \in \mathscr{P}_\sigma(\mu, \Sigma)} \varrho(F^{\mathbb{Q}}_{-w^\top \xi}) = \sup_{F \in \mathscr{D}_\sigma(-w^\top \mu, w^\top \Sigma w)} \varrho(F),$$

where the first equality follows from the consistency property of the family of risk measures, and the second equality is from the equivalence $\mathscr{D}_\sigma(-w^\top \mu, w^\top \Sigma w) = \{F^{\mathbb{Q}}_{-w^\top \xi} : \mathbb{Q} \in \mathscr{P}_\sigma(\mu, \Sigma)\}$ established previously. Noticing that the proof is valid for any arbitrary σ and as such the claim is proven whenever $s > 0$. This completes the proof. □

For a risk measure ϱ defined over distribution function space and an information structure σ, we define the *standard risk coefficient* $\alpha \in \mathbb{R}$ as the minimum uniform bound of the risk over all standard distribution functions in $\mathscr{D}_\sigma$ with mean 0 and variance 1

$$\alpha \triangleq \sup_{F \in \mathscr{D}_\sigma(0,1)} \varrho(F). \tag{3.28}$$

By construction, α depends only on the information structure σ and the risk measure ϱ. Moreover, the next result shows that under mild assumptions about ϱ, the worst-case risk over all distribution functions of arbitrary fixed mean $m \in \mathbb{R}$ and standard deviation $s \in \mathbb{R}_+$ can be re-expressed as a linear function of the standard risk coefficient α as the result of the following proposition.

Proposition 3.27 (Mean-Standard deviation reformulation)**.** *Suppose that the distributional risk measure ϱ is translation invariant and positive homogeneous. For any information structure $\sigma \in \mathscr{S}$, if the standard risk coefficient α defined in* (3.28) *is non-negative, then for any $(m, s) \in \mathbb{R} \times \mathbb{R}_+$, we have*

$$\sup_{F \in \mathscr{D}_\sigma(m,s^2)} \varrho(F) = m + \alpha s.$$

Proof. To simplify the notation, define the worst-case risk function $h_\sigma : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ as

$$h_\sigma(m,s) \triangleq \sup_{F \in \mathscr{D}_\sigma(m,s^2)} \varrho(F),$$

and thus the standard risk coefficient α defined in (3.28) satisfies $\alpha = h_\sigma(0,1)$. For any $F \in \mathscr{D}$ and $m \in \mathbb{R}$, we define the m-shifted function F_{+m} as $F_{+m}(t) \triangleq F(t-m)\ \forall t \in \mathbb{R}$. For any $F \in \mathscr{D}$ and $\lambda \in \mathbb{R}_{++}$, we define the λ-scaled function F_λ as $F_\lambda(t) \triangleq F(t/\lambda)\ \forall t \in \mathbb{R}$. For $\lambda = 0$, F_0 coincides with the Heaviside function, which is the cumulative distribution function of a random variable which equals 0 almost surely.

We first prove that h_σ inherits a form of translation invariance and positive homogeneity from ϱ. We have

$$h_\sigma(m,s) = \sup_{F \in \mathscr{D}_\sigma(m,s^2)} \varrho(F) = \sup_{F \in \mathscr{D}_\sigma(m,s^2)} \left\{\varrho(F_{-m}) + m\right\} = \sup_{\tilde{F} \in \mathscr{D}_\sigma(0,s^2)} \varrho(\tilde{F}) + m = h_\sigma(0,s) + m,$$

where the first and last equality come from the definition of h_σ, the second equality comes from translation invariance of the risk measure ϱ. For the third equality, notice that each $\tilde{F} \in \mathscr{D}_\sigma(0,s^2)$ is equivalent to a shifted distribution F_{-m} of some $F \in \mathscr{D}_\sigma(m,s^2)$, and if $F \in \mathscr{D}_\sigma(m,s^2)$ then $F_{-m} \in \mathscr{D}_\sigma(0,s^2)$. Furthermore, for any $\lambda > 0$ we have

$$h_\sigma(\lambda m, \lambda s) = \sup_{F \in \mathscr{D}_\sigma(\lambda m, \lambda^2 s^2)} \varrho(F) = \sup_{\tilde{F} \in \mathscr{D}_\sigma(m,s^2)} \varrho(\tilde{F}_\lambda) = \lambda \sup_{\tilde{F} \in \mathscr{D}_\sigma(m,s^2)} \varrho(\tilde{F}) = \lambda h_\sigma(m,s),$$

where the first and last equality come from the definition of h_σ and the third equality comes from positive homogeneity of the risk measure ϱ. For the second equality, notice that for any $\tilde{F} \in \mathscr{D}_\sigma(m, s^2)$, its scaled distribution $\tilde{F}_\lambda$ is equivalent to some $F \in \mathscr{D}_\sigma(\lambda m, \lambda^2 s^2)$, and any $F \in \mathscr{D}_\sigma(\lambda m, \lambda^2 s^2)$ is equivalent to the scaled distribution $\tilde{F}_\lambda$ of some $\tilde{F} \in \mathscr{D}_\sigma(m, s^2)$. For $\lambda = 0$, the above relationship holds trivially because ϱ is positive homogeneous and thus $\varrho(F_0) = 0$ by normalization.

In the second part of the proof, for any $\lambda > 0$ we find

$$\lambda\left(h_\sigma(0,s) + m\right) = \lambda h_\sigma(m,s) = h_\sigma(\lambda m, \lambda s) = h_\sigma(0, \lambda s) + \lambda m \implies h_\sigma(0,\lambda s) = \lambda h_\sigma(0,s).$$

Substituting $s = 1$ into the above equation gives $h_\sigma(0,\lambda) = \lambda h_\sigma(0,1) = \alpha\lambda$ which then implies $h_\sigma(0,s) = \alpha s$. The proof is now completed by noticing that when $\lambda = 1$, we have $h_\sigma(m,s) = h_\sigma(0,s) + m = \alpha s + m$. □

We are now ready to present the main theorem of this section which provides the tractable reformulation of the Gelbrich risk under some conditions of the risk measure family $\{\mathscr{R}_\mathbb{Q}\}_{\mathbb{Q}\in\mathscr{P}_\sigma}$.

Theorem 3.28 (Gelbrich risk of linear portfolios)**.** *Given any information structure $\sigma \in \mathscr{S}$, suppose that $\{\mathscr{R}_\mathbb{Q}\}_{\mathbb{Q}\in\mathscr{P}_\sigma}$ is a consistent family of translation invariant and positive homogeneous risk measures with the corresponding distributional risk measure ϱ, and that the standard risk coefficient α defined in* (3.28) *satisfies $0 \le \alpha < +\infty$. Then the Gelbrich risk of the linear portfolio loss $\ell(\xi) = -w^\top\xi$ admits the following closed-form expression*

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})} \mathscr{R}_\mathbb{Q}\left(-w^\top\xi\right) = -\widehat{\mu}^\top w + \alpha\sqrt{w^\top\widehat{\Sigma}w} + \rho\sqrt{1+\alpha^2}\|w\|. \tag{3.29}$$

If $\sigma = \phi$ and $\widehat{\mathbb{P}} = \mathscr{P}_\phi(\widehat{\mu},\widehat{\Sigma})$, then the Wasserstein risk of a linear portfolio is equal to the Gelbrich risk.

Proof. We can write the Gelbrich risk for a linear portfolio loss as

$$\begin{aligned}
\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})} \mathscr{R}_\mathbb{Q}(-w^\top\xi) &= \sup_{(\mu,\Sigma)\in\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \sup_{\mathbb{Q}\in\mathscr{P}_\sigma(\mu,\Sigma)} \mathscr{R}_\mathbb{Q}(-w^\top\xi) && \text{(3.30a)}\\
&= \sup_{(\mu,\Sigma)\in\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \sup_{\mathbb{Q}\in\mathscr{P}_\sigma(\mu,\Sigma)} \varrho\left(F^\mathbb{Q}_{-w^\top\xi}\right) && \text{(3.30b)}\\
&= \sup_{(\mu,\Sigma)\in\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \sup_{F\in\mathscr{D}_\sigma(-\mu^\top w, w^\top\Sigma w)} \varrho(F) && \text{(3.30c)}\\
&= \sup_{(\mu,\Sigma)\in\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})} -\mu^\top w + \alpha\sqrt{w^\top\Sigma w} && \text{(3.30d)}\\
&= \left\{\begin{array}{ll} \sup\limits_{\mu,\Sigma\succeq 0} & -\mu^\top w + \alpha\sqrt{w^\top\Sigma w} \\ \text{s.t.} & \|\mu-\widehat{\mu}\|^2 + \operatorname{Tr}\left[\Sigma + \widehat{\Sigma} - 2\left(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \le \rho^2, \end{array}\right.
\end{aligned}$$

where equality (3.30a) is the result of the decomposition (3.8a), equality (3.30b) utilizes the consistency property of the family of risk measures, equality (3.30c) and (3.30c) are the results of Proposition 3.26 and Proposition 3.27 respectively. The last equality follows from the definition of the uncertainty set $\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$.

Consider the case when $\rho > 0$ and $\alpha > 0$. Using a duality argument, the Gelbrich risk is equivalent to

$$\begin{aligned}
&\sup_{\mu,\Sigma\succeq 0}\inf_{\gamma\geq 0} \quad -\mu^\top w+\alpha\sqrt{w^\top\Sigma w}+\gamma\left[\rho^2-\|\mu-\widehat{\mu}\|^2-\mathrm{Tr}\left[\Sigma+\widehat{\Sigma}-2\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right]\right]\\
&=\inf_{\gamma\geq 0}\sup_{\mu,\Sigma\succeq 0} \quad -\mu^\top w+\alpha\sqrt{w^\top\Sigma w}+\gamma\left[\rho^2-\|\mu-\widehat{\mu}\|^2-\mathrm{Tr}\left[\Sigma+\widehat{\Sigma}-2\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right]\right] \qquad (3.31\text{a})\\
&=\inf_{\gamma\geq 0}\left\{\gamma\big(\rho^2-\mathrm{Tr}\,[\widehat{\Sigma}]\big)+\sup_{\mu,\Sigma\succeq 0}\left\{-\mu^\top w-\gamma\|\mu-\widehat{\mu}\|^2+\alpha\sqrt{w^\top\Sigma w}+\gamma\,\mathrm{Tr}\left[-\Sigma+2\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right]\right\}\right\}\\
&=\inf_{\gamma\geq 0}\left\{\gamma\big(\rho^2-\mathrm{Tr}\,[\widehat{\Sigma}]\big)+\sup_{\mu}\left\{-\mu^\top w-\gamma\|\mu-\widehat{\mu}\|^2\right\}+\sup_{\Sigma\succeq 0}\left\{\alpha\sqrt{w^\top\Sigma w}+\gamma\,\mathrm{Tr}\left[-\Sigma+2\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right]\right\}\right\}
\end{aligned}$$

where equality (3.31a) is from strong duality which holds because $(\widehat{\mu},\widehat{\Sigma})$ constitutes a Slater point for the convex primal problem (3.30). For the supremum subproblem over μ, we can employ an epigraphical formulation to show that for any $\gamma \geq 0$

$$\sup_{\mu}\left\{-\mu^\top w-\gamma\|\mu-\widehat{\mu}\|^2\right\}=-\widehat{\mu}^\top w+\frac{\|w\|^2}{4\gamma},$$

where for $\gamma = 0$, the expression on the right hand side is understood in the limit sense as γ tends to 0. The supremum subproblem over Σ is equivalent to a convex optimization problem

$$\begin{array}{cl}
\displaystyle\sup_{t,\Sigma} & \alpha t+\gamma\,\mathrm{Tr}\left[-\Sigma+2\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right]\\
\text{s.t.} & t\geq 0,\ \Sigma\succeq 0,\ t^2-w^\top\Sigma w\leq 0.
\end{array}$$

This optimization problem satisfies the Slater's condition. Suppose momentarily that $\widehat{\Sigma}\succ 0$. Using a duality argument, the above optimization problem is equivalent to

$$\begin{aligned}
&\sup_{t\geq 0,\Sigma\succeq 0}\inf_{\lambda\geq 0}\ \alpha t+\gamma\,\mathrm{Tr}\left[-\Sigma+2\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right]+\lambda(w^\top\Sigma w-t^2)\\
&=\inf_{\lambda\geq 0}\sup_{t\geq 0,\Sigma\succeq 0}\ \alpha t-\lambda t^2+\mathrm{Tr}\left[\Sigma(\lambda ww^\top-\gamma I)\right]+2\gamma\,\mathrm{Tr}\left[\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right] \qquad (3.32\text{a})\\
&=\inf_{\lambda\geq 0}\sup_{t\geq 0,B\succeq 0}\ \left\{\alpha t-\lambda t^2+2\gamma\,\mathrm{Tr}\,[B]+\mathrm{Tr}\,[B^2\Delta]\right\}, \qquad (3.32\text{b})
\end{aligned}$$

where equality (3.32a) follows from strong duality. In (3.32b), we have used the following change of variable $B^2 \leftarrow \widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}$ and thus we can rewrite $\Sigma=\widehat{\Sigma}^{-\frac{1}{2}}B^2\widehat{\Sigma}^{-\frac{1}{2}}$ which is valid thanks to the invertibility of $\widehat{\Sigma}$ and define $\Delta\triangleq\widehat{\Sigma}^{-\frac{1}{2}}(\lambda ww^\top-\gamma I)\widehat{\Sigma}^{-\frac{1}{2}}$. The inner maximization problem is separable in t and B, and the joint first order condition is

$$\left\{\begin{array}{rcl}
\alpha-2\lambda t^\star & = & 0\\
B^\star\Delta+\Delta B^\star+2\gamma I & = & 0.
\end{array}\right.$$

The first condition requires that $\lambda > 0$, and in this case it implies that $t^\star = \alpha/(2\lambda) > 0$. A necessary condition for the optimal value of B is that $0 \succ \Delta$, which in turns requires $\gamma I - \lambda w w^\top \succ 0$, or equivalently $\lambda < \gamma \|w\|^2$. Under that condition, the optimal value of B is $B^\star = -\gamma \Delta^{-1}$, which further can be shown to be unique [83, Theorem 12.5]. Problem (3.32b) can be reformulated as

$$\begin{aligned}\inf_{0<\lambda<\gamma\|w\|^{-2}} \left\{ \frac{\alpha^2}{4\lambda} + \gamma^2 \operatorname{Tr}\left[\widehat{\Sigma}^{\frac{1}{2}} (\gamma I - \lambda w w^\top)^{-1} \widehat{\Sigma}^{\frac{1}{2}}\right] \right\} &= \inf_{0<\lambda<\gamma\|w\|^{-2}} \left\{ \frac{\alpha^2}{4\lambda} + \gamma \operatorname{Tr}\left[\widehat{\Sigma}\right] + \frac{w^\top \widehat{\Sigma} w}{\frac{1}{\lambda} - \frac{1}{\gamma}\|w\|^2} \right\} \\ &= \gamma \operatorname{Tr}\left[\widehat{\Sigma}\right] + \frac{\alpha^2}{4} \frac{\|w\|^2}{\gamma} + \alpha \sqrt{w^\top \widehat{\Sigma} w},\end{aligned}$$

where the first equality is by applying the Sherman-Morrison formula [12, Corollary 2.8.8] to rewrite the matrix inversion. The optimal solution $\lambda^\star$ of the infimum problem satisfies

$$\frac{1}{\lambda^\star} = \frac{\|w\|^2}{\gamma} + \frac{2}{\alpha}\sqrt{w^\top \widehat{\Sigma} w}$$

which leads us to the second equality. As a consequence, Gelbrich risk admits an equivalent reformulation

$$\begin{aligned}\sup_{\mathbb{Q} \in \mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})} \mathscr{R}_{\mathbb{Q}}(-w^\top \xi) &= \inf_{\gamma>0} \left\{ -\widehat{\mu}^\top w + \alpha \sqrt{w^\top \widehat{\Sigma} w} + \gamma \rho^2 + \frac{1+\alpha^2}{4} \frac{\|w\|^2}{\gamma} \right\} &\quad (3.33a) \\ &= -\widehat{\mu}^\top w + \alpha \sqrt{w^\top \widehat{\Sigma} w} + \rho \sqrt{1+\alpha^2}\|w\|, &\quad (3.33b)\end{aligned}$$

where the optimal solution for γ in the infimum problem (3.33a) is $\gamma^\star = (2\rho)^{-1}\sqrt{1+\alpha^2}\|w\|$, and replacing $\gamma^\star$ into (3.33a) gives the expression (3.33b) when $\widehat{\Sigma} \succ 0$, $\alpha > 0$ and $\rho > 0$. One can readily verify that the reformulation (3.33b) is also valid under a more general condition where $\alpha \geq 0$ and $\rho \geq 0$.

We complete the proof by extending the reformulation (3.33b) to the case when $\widehat{\Sigma} \succeq 0$. Denote by $\mathscr{J}(\widehat{\Sigma})$ the Gelbrich risk parametrized by $\widehat{\Sigma}$, that is,

$$\mathscr{J}(\widehat{\Sigma}) \triangleq \left\{ \begin{array}{ll} \sup\limits_{\mu, \Sigma \succeq 0} & -\mu^\top w + \alpha \sqrt{w^\top \Sigma w} \\ \text{s.t.} & \|\mu - \widehat{\mu}\|^2 + \operatorname{Tr}\left[\Sigma + \widehat{\Sigma} - 2\big(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right] \leq \rho^2. \end{array} \right.$$

By Berge's maximum theorem [11, pp. 115-116], $\mathscr{J}$ is continuous over the domain $\mathbb{S}_+^n$. Let $\bar{\mathscr{J}}(\widehat{\Sigma}) = -\widehat{\mu}^\top w + \alpha \sqrt{w^\top \widehat{\Sigma} w} + \rho \sqrt{1+\alpha^2}\|w\|$, which is a continuous function over $\mathbb{S}_+^n$. Previously, we have shown that $\mathscr{J}(\widehat{\Sigma}) = \bar{\mathscr{J}}(\widehat{\Sigma})$ for any $\widehat{\Sigma} \succ 0$, and as such we can conclude that $\mathscr{J}(\widehat{\Sigma}) = \bar{\mathscr{J}}(\widehat{\Sigma})$ for any $\widehat{\Sigma} \in \mathbb{S}_+^n$ because both $\mathscr{J}$ and $\bar{\mathscr{J}}$ are continuous over $\mathbb{S}_+^n$.

We note that when $\sigma = \phi$, the Wasserstein risk equals to the Gelbrich risk, which is a direct consequence of Corollary 3.14. This completes the proof. □

When $\alpha < 0$, the reformulation of the Gelbrich risk measure involves solving a non-convex optimization problem, thus in this case the reformulation is not available. We can show that the standard risk coefficient α defined in (3.28) is non-negative whenever ϱ is a coherent risk measure.

Lemma 3.29 (Non-negative standard risk coefficient)**.** *If ϱ is a coherent distributionally risk measure, then the standard risk coefficient α defined in* (3.28) *is non-negative for any $\sigma \in \mathscr{S}$.*

Proof. Fix a distribution $F \in \mathscr{D}_\sigma(0,1)$ such that F is the distribution of $w^\top \xi$ under a symmetric probability measure $\mathbb{Q} \in \mathscr{P}_\sigma$. Denote by F_- the distribution of $-w^\top \xi$ under $\mathbb{Q}$. By construction, we have $F_- \in \mathscr{D}_\sigma(0,1)$, and because $\mathbb{Q}$ is a symmetric measure, we have $F \equiv F_-$. Let $\mathscr{R}_\mathbb{Q}$ be the risk measure constructed from ϱ (cf. Remark 3.8). We find

$$0 = \mathscr{R}_\mathbb{Q}(0) = \mathscr{R}_\mathbb{Q}(0.5w^\top\xi - 0.5w^\top\xi) \le 0.5\mathscr{R}_\mathbb{Q}(w^\top\xi) + 0.5\mathscr{R}_\mathbb{Q}(-w^\top\xi) = 0.5\varrho(F) + 0.5\varrho(F_-) = \varrho(F),$$

where the first equality follows from the normalization of the coherent risk measure, the inequality is due to the convexity of the risk measure, and the last equality follows from the fact that $F \equiv F_-$. The claim thus follows. □

Using Theorem 3.28, given that the portfolio manager's (ambiguous) risk attitude can be represented using a family of consistent risk measures which are translation invariant and positive homogeneous with standard risk coefficient $\alpha \ge 0$, the problem of finding a portfolio weight that minimizes the Gelbrich risk can be written as

$$\min_{w \in \mathscr{W}} \sup_{\mathbb{Q} \in \mathbb{G}_{\rho,\sigma}(\widehat{\mathbb{P}})} \mathscr{R}_\mathbb{Q}(-w^\top\xi) = \min_{w \in \mathscr{W}} \left\{ -\widehat{\mu}^\top w + \alpha\sqrt{w^\top \widehat{\Sigma} w} + \rho\sqrt{1+\alpha^2}\|w\| \right\}. \tag{3.34}$$

From Corollary 3.14, the optimal value of the above optimization problem provides a conservative approximation for the optimal Wasserstein risk of the best portfolio in $\mathscr{W}$. For any $\rho > 0$, the objective function of the optimization problem on the right hand side of (3.34) is convex in w, and thus problem (3.34) can be efficiently solved using off-the-shelf conic programming solvers provided that the set of feasible portfolio allocations $\mathscr{W}$ is representable using semidefinite constraints.

For a specific feasible set $\mathscr{W}$, the optimizer of the portfolio optimization problem that minimizes the Gelbrich risk (3.34) can be shown to converge to a $1/n$-uniform portfolio as the degree of ambiguity increases. This phenomenon was first formally proven in [138] for a specific class of convex risk measures under the Wasserstein ambiguity set.

Corollary 3.30. *Suppose that $\mathscr{W} = \{w \in \mathbb{R}^n_+ : e^\top w = 1\}$, where e is a vector of ones. As the level of ambiguity ρ tends to ∞, the optimal portfolio that minimizes the Gelbrich risk in problem* (3.34) *converges to the* $1/n$*-uniform portfolio.*

Proof. As ρ approaches ∞, the term $\rho\sqrt{1+\alpha^2}\|w\|$ dominates the other terms in the objective function of (3.34). Thus, the minimizer of problem (3.34) converges to the minimizer of the following problem

$$\min \{\|w\| : \ w \in \mathbb{R}_+^n, \ e^\top w = 1\},$$

which can be shown to be the uniform portfolio. This completes the proof. □

We now provide some examples of the Gelbrich risk corresponding to popular risk measures in the literature, along with its reformulation. For a probability measure $\mathbb{Q}$, the Value-at-Risk (VaR) at the risk level $\beta \in (0,1)$ of a portfolio loss $\ell(\xi)$ is defined as

$$\mathbb{Q}\text{-VaR}_\beta(\ell(\xi)) \triangleq \inf\left\{\tau \in \mathbb{R} : \mathbb{Q}(\tau \le \ell(\xi)) \le \beta\right\}.$$

Even though VaR is neither a coherent nor a convex risk measure because it does not satisfy the sub-additivity condition, VaR is still considered as an industry standard measure of risk [31]. Because VaR satisfies translation invariance and positive homogeneity, the Gelbrich risk for the family of VaR risk measures $\{\mathscr{R}_\mathbb{Q}\}_{\mathbb{Q}\in\mathscr{P}_\sigma}$, where $\mathscr{R}_\mathbb{Q}$ is $\mathbb{Q}$-VaR$_\beta$ for each $\mathbb{Q} \in \mathscr{P}_\sigma$ can be conveniently constructed using the results established previously in this section.

Lemma 3.31 (Gelbrich Value-at-Risk)**.** *For any $\beta \in (0,1)$ and $\sigma \in \mathscr{S}$, the Gelbrich VaR of a loss $\ell(\xi)$ is*

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-VaR}_\beta(\ell(\xi)) = \inf\left\{\tau \in \mathbb{R} : \sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}(\tau \le \ell(\xi)) \le \beta\right\}. \tag{3.35}$$

Moreover, the Gelbrich VaR for a linear portfolio loss $\ell(\xi) = -w^\top \xi$ admits the following expressions:

(i) Suppose that $\sigma = 2$. If $\beta \in (0,1)$, then $\alpha = \sqrt{(1-\beta)/\beta}$ and

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-VaR}_\beta(-w^\top\xi) = -\widehat{\mu}^\top w + \sqrt{\frac{1-\beta}{\beta}}\sqrt{w^\top\widehat{\Sigma}w} + \frac{\rho}{\sqrt{\beta}}\|w\|.$$

(ii) Suppose that $\sigma = \mathrm{S}$. If $\beta \in (0,\frac{1}{2})$, then $\alpha = \sqrt{1/(2\beta)}$ and

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\mathrm{S}}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-VaR}_\beta(-w^\top\xi) = -\widehat{\mu}^\top w + \sqrt{\frac{1}{2\beta}}\sqrt{w^\top\widehat{\Sigma}w} + \rho\sqrt{1+\frac{1}{2\beta}}\|w\|.$$

If $\beta \in [\frac{1}{2},1)$, then $\alpha = 0$ and

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\mathrm{S}}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-VaR}_\beta(-w^\top\xi) = -\widehat{\mu}^\top w + \rho\|w\|.$$

(iii) Suppose that $\sigma = \mathrm{SU}$. If $\beta \in (0,\frac{1}{2})$, then $\alpha = 2/(3\sqrt{2\beta})$ and

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\mathrm{SU}}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-VaR}_\beta(-w^\top\xi) = -\widehat{\mu}^\top w + \frac{2}{3}\sqrt{\frac{1}{2\beta}}\sqrt{w^\top\widehat{\Sigma}w} + \rho\sqrt{1+\frac{2}{9\beta}}\|w\|.$$

If $\beta \in [\frac{1}{2}, 1)$, then $\alpha = 0$ and

$$\sup_{\mathbb{Q} \in \mathbb{G}_{\rho,\mathrm{SU}}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-VaR}_\beta(-w^\top \xi) = -\widehat{\mu}^\top w + \rho \|w\|.$$

(iv) Suppose that $\sigma = \phi$. If $\beta \in (0, \frac{1}{2}]$, then $\alpha = F_\phi^{-1}(1-\beta)$ and

$$\sup_{\mathbb{Q} \in \mathbb{G}_{\rho,\phi}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-VaR}_\beta(-w^\top \xi) = -\widehat{\mu}^\top w + F_\phi^{-1}(1-\beta)\sqrt{w^\top \widehat{\Sigma} w} + \rho\sqrt{1 + (F_\phi^{-1}(1-\beta))^2}\|w\|.$$

where F_ϕ is the cumulative distribution function of a univariate standard elliptical distribution with mean 0, variance 1 and characteristic generator ϕ.

Proof. For each information structure $\sigma \in \{2, \mathrm{S}, \mathrm{SU}\}$, the corresponding value of the standard risk coefficient α can be found in [177, Proposition 1]. The analytical expression for each case is a direct application of Theorem 3.28.

For $\sigma = \phi$, the standard elliptical distribution with characteristic generator ϕ of mean 0 and variance 1 admits a well-defined cumulative density function F_ϕ, and α is defined using the $1-\beta$ quantile value $\alpha = F_\phi^{-1}(1-\beta)$. This completes the proof. □

Contrary to VaR, the Conditional Value-at-Risk (CVaR) is a coherent risk measure [144]. For a probability measure $\mathbb{Q}$, the CVaR at the risk level $\beta \in (0,1)$ for a portfolio loss $\ell(\xi)$ is defined as

$$\mathbb{Q}\text{-CVaR}_\beta(\ell(\xi)) \triangleq \inf_{\tau \in \mathbb{R}} \left\{ \tau + \frac{1}{\beta} \mathbb{E}_\mathbb{Q}\left[(\ell(\xi) - \tau)^+\right] \right\},$$

where $(\ell(\xi) - \tau)^+ = \max\{0, \ell(\xi) - \tau\}$. The next lemma establishes the analytical expressions for the Gelbrich CVaR for different information structures.

Lemma 3.32 (Gelbrich Conditional Value-at-Risk). *For any $\beta \in (0,1)$ and $\sigma \in \mathscr{S}$, the Gelbrich CVaR of a loss $\ell(\xi)$ is*

$$\sup_{\mathbb{Q} \in \mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-CVaR}_\beta(\ell(\xi)) = \inf_{\tau \in \mathbb{R}} \left\{ \tau + \frac{1}{\beta} \sup_{\mathbb{Q} \in \mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})} \mathbb{E}_\mathbb{Q}\left[(\ell(\xi) - \tau)^+\right] \right\}.$$

Moreover, the Gelbrich CVaR for a linear portfolio loss $\ell(\xi) = -w^\top \xi$ admits the following expressions:

(i) Suppose that $\sigma = 2$. If $\beta \in (0,1)$, then $\alpha = \sqrt{(1-\beta)/\beta}$ and

$$\sup_{\mathbb{Q} \in \mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-CVaR}_\beta(-w^\top \xi) = -\widehat{\mu}^\top w + \sqrt{\frac{1-\beta}{\beta}}\sqrt{w^\top \widehat{\Sigma} w} + \frac{\rho}{\sqrt{\beta}}\|w\|.$$

(ii) *Suppose that* $\sigma = \mathrm{S}$. *If* $\beta \in (0, \frac{1}{2}]$, *then* $\alpha = \sqrt{1/(2\beta)}$ *and*

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\mathrm{S}}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-CVaR}_{\beta}(-w^\top\xi) = -\widehat{\mu}^\top w + \sqrt{\frac{1}{2\beta}}\sqrt{w^\top\widehat{\Sigma}w} + \rho\sqrt{1+\frac{1}{2\beta}}\|w\|.$$

If $\beta \in [\frac{1}{2}, 1)$ *then* $\alpha = \sqrt{1-\beta}/(\sqrt{2}\beta)$ *and*

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\mathrm{S}}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-CVaR}_{\beta}(-w^\top\xi) = -\widehat{\mu}^\top w + \sqrt{\frac{1-\beta}{2\beta^2}}\sqrt{w^\top\widehat{\Sigma}w} + \rho\sqrt{1+\frac{1-\beta}{2\beta^2}}\|w\|.$$

(iii) *Suppose that* $\sigma = \mathrm{SU}$. *If* $\beta \in (0, \frac{1}{3}]$, *then* $\alpha = 2/(3\sqrt{\beta})$ *and*

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\mathrm{SU}}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-CVaR}_{\beta}(-w^\top\xi) = -\widehat{\mu}^\top w + \frac{2}{3\sqrt{\beta}}\sqrt{w^\top\widehat{\Sigma}w} + \rho\sqrt{1+\frac{4}{9\beta}}\|w\|.$$

If $\beta \in [\frac{1}{3}, \frac{2}{3}]$, *then* $\alpha = \sqrt{3}(1-\beta)$ *and*

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\mathrm{SU}}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-CVaR}_{\beta}(-w^\top\xi) = -\widehat{\mu}^\top w + \sqrt{3}(1-\beta)\sqrt{w^\top\widehat{\Sigma}w} + \rho\sqrt{1+3(1-\beta)^2}\|w\|.$$

If $\beta \in [\frac{2}{3}, 1)$ *then* $\alpha = 2\sqrt{1-\beta}/(3\beta)$ *and*

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\mathrm{SU}}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-CVaR}_{\beta}(-w^\top\xi) = -\widehat{\mu}^\top w + \frac{2\sqrt{1-\beta}}{3\beta}\sqrt{w^\top\widehat{\Sigma}w} + \rho\sqrt{1+\frac{4(1-\beta)}{9\beta^2}}\|w\|.$$

(iv) *Suppose that* $\sigma = \phi$, *where* $\phi(u) = \exp(-u/2)$ *represents the characteristic generator of Gaussian distribution. If* $\beta \in (0, \frac{1}{2})$, *then* $\alpha = (\beta\sqrt{2\pi})^{-1}\exp(-z_\beta^2/2)$, *where* z_β *is the upper* β*-percentile of the standard Gaussian distribution.*

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\phi}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-CVaR}_{\beta}(-w^\top\xi) = -\widehat{\mu}^\top w + \frac{\exp(-z_\beta^2/2)}{\sqrt{2\pi}\beta}\sqrt{w^\top\widehat{\Sigma}w} + \rho\sqrt{1+\frac{\exp(-z_\beta^2)}{2\pi\beta^2}}\|w\|.$$

Proof. For each information structure $\sigma \in \mathscr{S}$, the corresponding value of the standard risk coefficient α can be found in [177, Proposition 2]. The analytical expression for each case is a direct application of Theorem 3.28. □

The closed-form expression for the Gelbrich CVaR for other family of elliptical distributions also exists but it is case dependent. Next, we consider the family of spectral risk measure proposed by [1].

Definition 3.33 (Spectral risk measure [1]). *A distributional risk measure* ϱ_ψ *is called a spectral* *risk measure with spectrum $\psi \in \mathscr{A}$ if*

$$\varrho_\psi(F) = \int_0^1 \psi(\beta) F^{-1}(\beta) \mathrm{d}\beta,$$

where $F^{-1}(\beta) = \inf\{q : F(q) \geq \beta\}$ and $\mathscr{A}$ is the set of all possible spectra

$$\mathscr{A} \triangleq \left\{ \psi : [0,1) \to \mathbb{R}_+ \middle| \int_0^1 \psi(\beta) \mathrm{d}\beta = 1, \psi\text{: right continuous, monotonically nondecreasing} \right\}.$$

Lemma 3.34 (Gelbrich spectral risk)**.** *Suppose that $\sigma = 2$ and that the family of risk measures $\{\mathscr{R}_\mathbb{Q}\}_{\mathbb{Q}\in\mathscr{P}_2}$ admits a spectral distributional risk measure ϱ_ψ with spectrum ψ. If ψ is square integrable, then the Gelbrich spectral risk of a linear portfolio loss $\ell(\xi) = -w^\top \xi$ is*

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathscr{R}_\mathbb{Q}\left(-w^\top\xi\right) = -\widehat{\mu}^\top w + \alpha\sqrt{w^\top \widehat{\Sigma} w} + \rho\sqrt{1+\alpha^2}\|w\|, \tag{3.36}$$

where $\alpha = \sqrt{\int_0^1 \psi(\beta)^2 \mathrm{d}\beta - 1} \geq 0$.

Proof. Because a spectral risk measure is coherent [1], it is therefore translation invariant and positive homogeneous. Hence, we can apply a similar decomposition as (3.30) to have

$$\begin{aligned}
\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathscr{R}_\mathbb{Q}\left(-w^\top\xi\right) &= \sup_{(\mu,\Sigma)\in\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \sup_{F\in\mathscr{D}_2(-\mu^\top w, w^\top \Sigma w)} \varrho_\psi(F) \\
&= \sup_{(\mu,\Sigma)\in\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \left\{ -\mu^\top w + \alpha\sqrt{w^\top \Sigma w} \right\},
\end{aligned}$$

where the last equality is a direct application of [110, Theorem 2] with $\alpha = \sqrt{\int_0^1 \psi(\beta)^2 \mathrm{d}\beta - 1} \geq 0$. The reformulation of the last supremum problem follows the same procedure as the proof of Theorem 3.28. This completes the proof. □

The CVaR risk measure at the risk level $\beta \in (0,1)$ is also a spectral risk measure with spectrum $\psi(\beta) = \beta^{-1}\mathbb{1}_{[1-\beta,1)}(\beta)$, and thus $\int_0^1 \psi(\beta)^2 \mathrm{d}\beta = \beta^{-1}$. One can easily verify that the expression for the Gelbrich CVaR with $\sigma = 2$ in Lemma 3.32 coincides with that for the Gelbrich CVaR using the spectral risk property in Lemma 3.34.

Interestingly, there is a tight connection between a coherent risk measure and the family of spectral risk measures. More specifically, the Kusuoka representation dictates that any coherent risk measure can be expressed as a supremum risk of a subset of admissible spectral risk measures.

Definition 3.35 (Kusuoka representation [101, 154])**.** *Any distributional coherent risk measure ϱ admits the following representation*

$$\varrho(F) = \sup_{\psi\in\Psi} \varrho_\psi(F) \qquad \forall F \in \mathscr{D}_2,$$

where $\Psi \subseteq \mathscr{A}$ *is a set of admissible spectra.*

Using the Kusuoka representation, the Gelbrich risk can be generalized to any consistent family of coherent risk measures by the result of the following lemma.

Lemma 3.36 (Gelbrich coherent risk)**.** *Suppose that* $\sigma = 2$ *and that the family of risk measures* $\{\mathscr{R}_{\mathbb{Q}}\}_{\mathbb{Q}\in\mathscr{P}_2}$ *admits a coherent distributional risk measure* ϱ *with a Kusuoka representation using admissible spectra* Ψ*. If* ψ *is square integrable for all* $\psi \in \Psi$*, then the Gelbrich coherent risk of a linear porfolio loss* $\ell(\xi) = -w^\top \xi$ *is*

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\hat{\mu},\hat{\Sigma})} \mathscr{R}_{\mathbb{Q}}\left(-w^\top\xi\right) = -\hat{\mu}^\top w + \alpha\sqrt{w^\top\hat{\Sigma}w} + \rho\sqrt{1+\alpha^2}\|w\|, \tag{3.37}$$

where $\alpha = \sqrt{\sup_{\psi\in\Psi}\int_0^1 \psi(\beta)^2 \mathrm{d}\beta - 1} \geq 0$*.*

Proof. Because ϱ is translation invariant and positive homogeneous, we can apply a similar decomposition as (3.30) to have

$$\begin{aligned}\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\hat{\mu},\hat{\Sigma})} \mathscr{R}_{\mathbb{Q}}\left(-w^\top\xi\right) &= \sup_{(\mu,\Sigma)\in\mathscr{U}_\rho(\hat{\mu},\hat{\Sigma})}\ \sup_{F\in\mathscr{D}_2(-\mu^\top w, w^\top\Sigma w)} \varrho_\psi(F) \\ &= \sup_{(\mu,\Sigma)\in\mathscr{U}_\rho(\hat{\mu},\hat{\Sigma})} \left\{-\mu^\top w + \alpha\sqrt{w^\top\Sigma w}\right\},\end{aligned}$$

where the last equality is a direct application of [110, Theorem 3] with

$$\alpha = \sqrt{\sup_{\psi\in\Psi}\int_0^1 \psi(\beta)^2 \mathrm{d}\beta - 1} \geq 0.$$

Reformulating the last supremum problem as in the proof of Theorem 3.28 completes the proof. □

It is often instructive to construct the extremal probability measure that attains the optimal value of the Gelbrich risk measure. Given a portfolio allocation $w \in \mathbb{R}^n$, let $\mathbb{Q}^\star \in \mathscr{P}_\sigma$ be the extremal probability measure associated with the linear portfolio loss $\ell(\xi) = -w^\top\xi$, that is,

$$\mathbb{Q}^\star \triangleq \arg\max_{\mathbb{Q}\in\mathbb{G}_{\rho,\sigma}(\hat{\mu},\hat{\Sigma})} \mathscr{R}_{\mathbb{Q}}(\ell).$$

The next proposition characterizes $\mathbb{Q}^\star$ up to its first- and second-moment information.

Proposition 3.37 (Extremal mean vector and covariance matrix)**.** *Suppose that the conditions of Theorem 3.28 hold. The extremal distribution* $\mathbb{Q}^\star$ *that attains the Gelbrich risk has mean vector* $\mu^\star \in \mathbb{R}^n$ *and covariance matrix* $\Sigma^\star \in \mathbb{S}_+^n$*, where*

$$\mu^\star = \hat{\mu} - \frac{\rho}{\sqrt{1+\alpha^2}\|w\|}w, \ \textit{and}\ \Sigma^\star = \left(I + \frac{\rho\alpha w w^\top}{\sqrt{1+\alpha^2}\|w\|\sqrt{w^\top\hat{\Sigma}w}}\right)\hat{\Sigma}\left(I + \frac{\rho\alpha w w^\top}{\sqrt{1+\alpha^2}\|w\|\sqrt{w^\top\hat{\Sigma}w}}\right).$$

Furthermore, if $\sigma = \phi$ and $\widehat{\mathbb{P}} = \mathscr{P}_\phi(\widehat{\mu}, \widehat{\Sigma})$, then $\mathbb{Q}^\star = \mathscr{P}_\phi\left(\mu^\star, \Sigma^\star\right)$ attains both the Gelbrich risk and the Wasserstein risk.

Proof. From the proof of Theorem 3.28, we find

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})} \mathscr{R}_{\mathbb{Q}}\left(-w^\top \xi\right) = \sup_{(\mu,\Sigma)\in\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})} -\mu^\top w + \alpha\sqrt{w^\top \Sigma w} = -\widehat{\mu}^\top w + \alpha\sqrt{w^\top \widehat{\Sigma} w} + \rho\sqrt{1+\alpha^2}\|w\|,$$

thus it suffices to show that $(\mu^\star, \Sigma^\star)$ is the optimal solution of the optimization problem

$$\max_{(\mu,\Sigma)\in\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \left\{-\mu^\top w + \alpha\sqrt{w^\top \Sigma w}\right\}.$$

Indeed, by using the definition of the Gelbrich distance $\mathbb{G}$ in Definition 3.9, we have

$$\begin{aligned}
\mathbb{G}\big((\mu^\star, \Sigma^\star), (\widehat{\mu}, \widehat{\Sigma})\big)^2 &= \|\mu^\star - \widehat{\mu}\|^2 + \mathrm{Tr}\left[\widehat{\Sigma} + \Sigma^\star - 2\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma^\star\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right] \\
&= \frac{\rho^2}{1+\alpha^2} + \mathrm{Tr}\left[\widehat{\Sigma}\frac{\rho^2\alpha^2 w w^\top w w^\top}{(1+\alpha^2)\|w\|^2 w^\top \widehat{\Sigma} w}\right] \\
&= \frac{\rho^2}{1+\alpha^2} + \frac{\rho^2\alpha^2}{1+\alpha^2} = \rho^2,
\end{aligned}$$

where the second equality utilizes the definition of $\mu^\star$ and $\Sigma^\star$ in the statement of the proposition. This implies that $(\mu^\star, \Sigma^\star) \in \mathscr{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$. Furthermore, we find

$$\begin{aligned}
-(\mu^\star)^\top w + \alpha\sqrt{w^\top \Sigma^\star w} &= -\widehat{\mu}^\top w + \frac{\rho\|w\|}{\sqrt{1+\alpha^2}} + \alpha\sqrt{\left(\sqrt{w^\top \widehat{\Sigma} w} + \frac{\rho\alpha}{\sqrt{1+\alpha^2}}\|w\|\right)^2} \\
&= -\widehat{\mu}^\top w + \alpha\sqrt{w^\top \widehat{\Sigma} w} + \rho\sqrt{1+\alpha^2}\|w\|,
\end{aligned}$$

which coincides with the value of the Gelbrich risk.

If $\sigma = \phi$, then $\mathbb{Q}^\star = \mathscr{P}_\phi(\mu^\star, \Sigma^\star)$ belongs to the Gelbrich hull $\mathbb{G}_{\rho,\phi}(\widehat{\mu}, \widehat{\Sigma})$ and thus it is an extremal distribution for the Gelbrich risk. The claim for the Wasserstein risk holds trivially because $\mathbb{G}_{\rho,\phi}(\widehat{\mu}, \widehat{\Sigma}) = \mathbb{B}_{\rho,\phi}(\widehat{\mathbb{P}})$. □

Proposition 3.37 only characterizes the first two moments of the extremal distribution $\mathbb{Q}^\star$. Since the projection constructed in the proof of Proposition 3.26 is potentially many-to-one, there may exist multiple extremal distributions in the Gelbrich hull $\mathbb{G}_{\rho,\sigma}(\widehat{\mu}, \widehat{\Sigma})$ that share the same extremal mean vector and covariance matrix $(\mu^\star, \Sigma^\star)$ specified by Proposition 3.37. Once the information structure is elliptical $\sigma = \phi$, we can uniquely identify the extremal probability measure. We conclude this section with two remarks regarding the reformulation of the worst-case linear single chance constraint and the connection between the Gelbrich risk and robust optimization.

Remark 3.38 (Distributionally robust linear chance constraint)**.** *The Gelbrich VaR provides a straightforward reformulation for the worst-case linear chance constraint over the Gelbrich hull*

$\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})$ *as follows*

$$\left\{w \in \mathbb{R}^n : \inf_{\mathbb{Q}\in\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}(-w^\top \xi \le b) \ge 1-\beta\right\} = \left\{w \in \mathbb{R}^n : \sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-VaR}_\beta(-w^\top \xi) \le b\right\}$$

for any $\beta \in (0,1)$. *Replacing the reformulation of the Gelbrich VaR in Lemma 3.31 corresponding to the information structure* σ *provides the reformulation of the Gelbrich chance constrained feasible set.*

Remark 3.39 (Connection with robust optimization)**.** *The reformulation of the Gelbrich risk of a linear portfolio loss* (3.29) *is equivalent to the robust counterpart of*

$$\sup_{u\in\mathbb{U}} u^\top w,$$

where the uncertainty set $\mathbb{U}$ *is defined as*

$$\mathbb{U} = \left\{u \in \mathbb{R}^n : \exists u_1 \in \mathbb{R}^n, u_2 \in \mathbb{R}^n \text{ such that } \|u_1\| \le \alpha, \|u_2\| \le \rho\sqrt{1+\alpha^2}, u = \widehat{\Sigma}^{\frac{1}{2}} u_1 + u_2 + \widehat{\mu}\right\}.$$

Remark 3.40 (Factor model)**.** *For practical purposes, it is common to assume that the return of* k *stocks* $\eta \in \mathbb{R}^k$ *can be decomposed using a factor model with* $n \ll k$ *components as* $\eta = A\xi$ *for some* $A \in \mathbb{R}^{k\times n}$. *If the Wasserstein ball* $\mathbb{B}_{\rho,\sigma}(\widehat{\mathbb{P}})$ *and the Gelbrich hull* $\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})$ *are constructed over the factor* ξ, *the result of Theorem 3.28 can be applied in a straightforward manner to provide the reformulation for the Gelbrich risk of the linear portfolio* $\ell(\xi) = -w^\top A\xi$ *for some portfolio allocation* $w \in \mathbb{R}^k$ *as*

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})} \mathscr{R}_\mathbb{Q}\left(-w^\top A\xi\right) = -\widehat{\mu}^\top A^\top w + \alpha\sqrt{w^\top A\widehat{\Sigma}A^\top w} + \rho\sqrt{1+\alpha^2}\|A^\top w\|.$$

The portfolio optimization problem that searches for the optimal allocation $w \in \mathbb{R}^k$ *remains convex provided that the feasible portfolio allocation set is convex, and it is tractably solvable if the feasible portfolio allocation set is SDP representable.*

3.5 Reformulations of Worst-case Expected Loss Risks

In Section 3.4, we have studied thoroughly the Gelbrich risk of a linear loss function with different structural information σ. Throughout this section, we attempt to evaluate the risk of a *non*linear loss function. Towards this end, we restrict the setting to $\sigma = 2$ and consider the family of risk measures $\{\mathscr{R}_\mathbb{Q}\}_{\mathbb{Q}\in\mathscr{P}_2}$ where $\mathscr{R}_\mathbb{Q}$ is defined as the expect loss, that is,

$$\mathscr{R}_\mathbb{Q}(\ell) = \mathbb{E}_\mathbb{Q}[\ell(\xi)] \quad \forall \mathbb{Q} \in \mathscr{P}_2.$$

By construction, this family of expectation risk measures satisfies the consistency property in Definition 3.7. Following the results of Corollary 3.14, given the nominal distribution $\widehat{\mathbb{P}}$ with mean $\widehat{\mu} \in \mathbb{R}^n$ and covariance matrix $\widehat{\Sigma} \in \mathbb{S}^n_+$, the Wasserstein expected loss

$$\sup_{\mathbb{Q} \in \mathbb{B}_{\rho,2}(\widehat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)]$$

is upper bounded by the Gelbrich expected loss

$$\sup_{\mathbb{Q} \in \mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)]. \tag{3.38}$$

Consequently, this section will focus on the Gelbrich expected loss as a tractable conservative approximation of the Wasserstein expected loss. We can employ (3.8) to decompose the worst-case expected loss under the Gelbrich hull $\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})$ into two consecutive maximization problems

$$\sup_{\mathbb{Q} \in \mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] = \sup_{(\mu,\Sigma) \in \mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \sup_{\mathbb{Q} \in \mathscr{P}_2(\mu,\Sigma)} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)],$$

which offers a systematic approach to derive convex reformulations. A tractable SDP reformulation is available, for example, when the loss function $\ell(\xi)$ is a pointwise maximum of finitely many (possibly indefinite) quadratic functions.

Theorem 3.41 (Piecewise quadratic loss I). *Assume that $\ell(\xi) = \max_{j \in [J]}\{\xi^\top Q_j \xi + 2q_j^\top \xi + q_j^0\}$ with $Q_j \in \mathbb{S}^n$, $q_j \in \mathbb{R}^n$, and $q_j^0 \in \mathbb{R}$ for any $j \in [J]$ is a piecewise quadratic loss function. If $\sigma = 2$, $\widehat{\mu} \in \mathbb{R}^n$ and $\widehat{\Sigma} \in \mathbb{S}^n_+$, then for any $\rho \in \mathbb{R}_{++}$, the Gelbrich expected loss is equal to the optimal value of a tractable semidefinite program, that is,*

$$\sup_{\mathbb{Q} \in \mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] = \left\{ \begin{array}{ll} \inf & y_0 + \gamma\big(\rho^2 - \|\widehat{\mu}\|^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\big) + z + \mathrm{Tr}\left[Z\right] \\ \mathrm{s.t.} & \gamma \in \mathbb{R}_+,\ y_0 \in \mathbb{R},\ y \in \mathbb{R}^n,\ Y \in \mathbb{S}^n,\ z \in \mathbb{R}_+,\ Z \in \mathbb{S}^n_+ \\ & \begin{bmatrix} \gamma I - Y & \gamma \widehat{\Sigma}^{\frac{1}{2}} \\ \gamma \widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0,\ \begin{bmatrix} \gamma I - Y & \gamma\widehat{\mu} + y \\ (\gamma\widehat{\mu} + y)^\top & z \end{bmatrix} \succeq 0 \\ & \begin{bmatrix} Y - Q_j & y - q_j \\ y^\top - q_j^\top & y_0 - q_j^0 \end{bmatrix} \succeq 0 \quad \forall j \in [J]. \end{array} \right. \tag{3.39}$$

Proof. By applying the decomposition (3.8a), we find

$$\sup_{\mathbb{Q} \in \mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] = \sup_{(\mu,\Sigma) \in \mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \sup_{\mathbb{Q} \in \mathscr{P}_2(\mu,\Sigma)} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)].$$

We can appply [183, Lemma A.1] to construct the dual of the inner supremum problem as

$$\sup_{\mathbb{Q} \in \mathscr{P}_2(\mu,\Sigma)} \mathbb{E}_{\mathbb{Q}}\left[\ell(\xi)\right] \le \left\{ \begin{array}{ll} \inf & y_0 + 2y^\top \mu + \mathrm{Tr}\left[Y(\Sigma + \mu\mu^\top)\right] \\ \mathrm{s.t.} & y_0 \in \mathbb{R},\ y \in \mathbb{R}^n,\ Y \in \mathbb{S}^n \\ & y_0 + 2y^\top \xi + \xi^\top Y \xi \ge \ell(\xi) \qquad \forall \xi \in \mathbb{R}^n, \end{array} \right. \tag{3.40}$$

where the inequality is tight whenever $\Sigma \succ 0$ thanks to the strong duality result [86]. Let $\mathcal{Y}$ be the convex feasible set defined by

$$\begin{aligned}\mathscr{Y} &= \left\{y_0 \in \mathbb{R},\ y \in \mathbb{R}^n,\ Y \in \mathbb{S}^n : y_0 + 2y^\top \xi + \xi^\top Y \xi \geq \ell(\xi) \quad \forall \xi \in \mathbb{R}^n\right\} \\ &= \left\{ y_0 \in \mathbb{R},\ y \in \mathbb{R}^n,\ Y \in \mathbb{S}^n : \begin{bmatrix} Y - Q_j & y - q_j \\ y^\top - q_j^\top & y_0 - q_j^0 \end{bmatrix} \succeq 0 \quad \forall j \in [J] \right\},\end{aligned}$$

where the equality is obtained by simply substituting the expression of the loss function and formulating the quadratic constraints using semidefinite constraints. We now have

$$\begin{aligned}\sup_{\mathbb{Q} \in \mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{E}_{\mathbb{Q}}\left[\ell(\xi)\right] &= \sup_{(\mu,\Sigma) \in \mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \sup_{\mathbb{Q} \in \mathscr{P}_2(\mu,\Sigma)} \mathbb{E}_{\mathbb{Q}}\left[\ell(\xi)\right] && \text{(3.41a)} \\ &\leq \sup_{(\mu,\Sigma) \in \mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \inf_{(y_0,y,Y) \in \mathscr{Y}} y_0 + 2y^\top \mu + \mathrm{Tr}\left[Y(\Sigma + \mu\mu^\top)\right] && \text{(3.41b)} \\ &= \sup_{(\mu,M) \in \mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})} \inf_{(y_0,y,Y) \in \mathscr{Y}} y_0 + 2y^\top \mu + \mathrm{Tr}\left[YM\right] && \text{(3.41c)} \\ &= \inf_{(y_0,y,Y) \in \mathscr{Y}} \sup_{(\mu,M) \in \mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})} y_0 + 2y^\top \mu + \mathrm{Tr}\left[YM\right] && \text{(3.41d)} \\ &= \inf_{(y_0,y,Y) \in \mathscr{Y}} \left\{ y_0 + \sup_{(\mu,M) \in \mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})} 2y^\top \mu + \mathrm{Tr}\left[YM\right] \right\} \\ &= \inf_{(y_0,y,Y) \in \mathscr{Y}} \left\{ y_0 + \delta^\star_{\mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(2y, Y) \right\},\end{aligned}$$

where the inequality (3.41b) is from the weak duality result of the inner supremum problem established in (3.40), and the equality (3.41c) follows from the equivalence of the uncertainty set $\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ and $\mathscr{V}_\rho(\widehat{\mu},\widehat{M})$ with $\widehat{M} = \widehat{\Sigma} + \widehat{\mu}\widehat{\mu}^\top$. Finally, equality (3.41d) is due to Sion's minimax theorem [156, Corollary 3.3]. By replacing the reformulation of the support function for $\mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})$ evaluated at $(2y, Y)$ using Lemma 3.22, we have established that the infimum problem in (3.39) is the upper bound of the worst-case expected loss.

In the second part of the proof, we show that the bound is actually tight, which means the inequality in (3.41b) holds with equality. Denote momentarily by f the optimal value of the inner infimum in (3.41c), that is,

$$f(\mu, M) \triangleq \inf_{(y_0,y,Y) \in \mathscr{Y}} y_0 + 2y^\top \mu + \mathrm{Tr}\left[YM\right].$$

By definition, f is the pointwise infimum of upper semicontinuous functions, thus f is upper semicontinuous [3, Lemma 2.41]. Furthermore, by virtue of Proposition 3.17, $\mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})$ is a compact set. Thus the set of optimizers of the supremum problem (3.41c) is non-empty [3, Theorem 2.43] and (3.41c) can be written with the maximum operator as

$$\max_{(\mu,M) \in \mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})} f(\mu, M). \tag{3.42}$$

Denote by $(\mu^\star, M^\star)$ the optimal solution of (3.42). If $M^\star - \mu^\star(\mu^\star)^\top \succ 0$, then $(\mu^\star, M^\star -$

$\mu^\star(\mu^\star)^\top$) is feasible for problem (3.41a) and thus (3.41b) holds trivially as an equality. It now suffices to show the equality when $M^\star - \mu^\star(\mu^\star)^\top \not\succ 0$. For any $\rho > 0$, it is easy to show that there exists $(\bar\mu, \bar\Sigma) \in \mathcal{U}_\rho(\widehat\mu, \widehat\Sigma)$ such that $\bar\Sigma \succ 0$. Denote $\bar M = \bar\Sigma + \bar\mu\bar\mu^\top$. Consider the sequence $\{\mu_k, M_k\}_{k\in\mathbb{N}}$ defined as

$$\mu_k = \frac{1}{k}\bar\mu + \frac{k-1}{k}\mu^\star, \quad M_k = \frac{1}{k}\bar M + \frac{k-1}{k}M^\star.$$

Notice that the covariance matrix Σ_k associated with (μ_k, M_k) is positive definite for any $k \in \mathbb{N}$ because

$$\begin{aligned}\Sigma_k = M_k - \mu_k\mu_k^\top &= \frac{1}{k}(\bar\Sigma + \bar\mu\bar\mu^\top) + \frac{k-1}{k}(\Sigma^\star + \mu^\star(\mu^\star)^\top) - \left(\frac{1}{k}\bar\mu + \frac{k-1}{k}\mu^\star\right)\left(\frac{1}{k}\bar\mu + \frac{k-1}{k}\mu^\star\right)^\top \\ &= \frac{1}{k}\bar\Sigma + \frac{k-1}{k}\Sigma^\star + \frac{k-1}{k^2}(\bar\mu - \mu^\star)(\bar\mu - \mu^\star)^\top \succ 0.\end{aligned}$$

Because f is the pointwise infimum of linear, thus concave, functions, f is also concave. Thus we find

$$f(\mu_k, M_k) \geq \frac{1}{k} f(\bar\mu, \bar M) + \frac{k-1}{k} f(\mu^\star, M^\star).$$

As a result, we have

$$\begin{aligned} f(\mu^\star, M^\star) = \lim_{k\to\infty} \frac{1}{k} f(\bar\mu, \bar M) + \frac{k-1}{k} f(\mu^\star, M^\star) &\leq \lim_{k\to\infty} f(\mu_k, M_k) \\ &= \lim_{k\to\infty} \sup_{\mathbb{Q}\in\mathcal{P}_2(\mu_k, M_k - \mu_k\mu_k^\top)} \mathbb{E}_\mathbb{Q}[\ell(\xi)] \\ &\leq \sup_{(\mu,M)\in\mathcal{V}_\rho(\widehat\mu,\widehat\Sigma)} \sup_{\mathbb{Q}\in\mathcal{P}_2(\mu, M-\mu\mu^\top)} \mathbb{E}_\mathbb{Q}[\ell(\xi)], \end{aligned}$$

where the first equality is from the concavity of f, the second equality is from the strong duality because $\Sigma_k \succ 0$ [86], and the last inequality is from the definition of the supremum. This renders (3.41b) an equality and thus completes the proof. □

In order to construct an extremal distribution for the Gelbrich risk evaluation problem (3.6), it is expedient to derive the dual of the SDP (3.39).

Theorem 3.42 (Piecewise quadratic loss II). *If all conditions of Theorem 3.41 hold, then we have*

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat\mu,\widehat\Sigma)} \mathbb{E}_\mathbb{Q}[\ell(\xi)] = \left\{ \begin{array}{ll} \max & \sum_{j\in[J]} \mathrm{Tr}\left[Q_j\Theta_j\right] + 2q_j^\top\theta_j + q_j^0\alpha_j \\ \text{s.t.} & \mu \in \mathbb{R}^n,\ \Sigma \in \mathbb{S}_+^n,\ \alpha_j \in \mathbb{R}_+,\ \theta_j \in \mathbb{R}^n,\ \Theta_j \in \mathbb{S}_+^n \quad \forall j \in [J] \\ & \begin{bmatrix} \Theta_j & \theta_j \\ \theta_j^\top & \alpha_j \end{bmatrix} \succeq 0 \quad \forall j \in [J] \\ & \sum_{j\in[J]} \alpha_j = 1,\ \sum_{j\in[J]} \theta_j = \mu,\ \sum_{j\in[J]} \Theta_j = \Sigma + \mu\mu^\top,\ (\mu,\Sigma) \in \mathcal{U}_\rho(\widehat\mu,\widehat\Sigma). \end{array} \right. \tag{3.43}$$

Proof. Following the results of Theorem 3.41, we have the equivalence

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] = \left\{ \begin{array}{ll} \sup\limits_{(\mu,\Sigma)\in\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})} & \inf \quad y_0 + 2y^\top\mu + \mathrm{Tr}\left[Y(\Sigma+\mu\mu^\top)\right] \\ & \text{s.t.} \quad y_0\in\mathbb{R}, y\in\mathbb{R}^n, Y\in\mathbb{S}^n \\ & \qquad \begin{bmatrix} Y-Q_j & y-q_j \\ y^\top-q_j^\top & y_0-q_j^0 \end{bmatrix} \succeq 0 \quad \forall j\in[J]. \end{array} \right.$$

Consider the inner infimum subproblem. Notice that the Slater's condition holds for this infimum problem, and thus we can derive its SDP dual form as

$$\left\{ \begin{array}{ll} \sup & \sum\limits_{j\in[J]} \mathrm{Tr}\left[\Theta_j Q_j\right] + 2q_j^\top\theta_j + q_j^0\alpha_j \\ \text{s.t.} & \mu\in\mathbb{R}^n,\ \Sigma\in\mathbb{S}_+^n,\ \Theta_j\in\mathbb{S}_+^n,\ \theta_j\in\mathbb{R}^n,\ \alpha_j\in\mathbb{R} \quad \forall j\in[J] \\ & \sum\limits_{j\in[J]} \alpha_j = 1,\ \sum\limits_{j\in[J]} \theta_j = \mu,\ \sum\limits_{j\in[J]} \Theta_j = \mu\mu^\top + \Sigma \\ & \begin{bmatrix} \Theta_j & \theta_j \\ \theta_j^\top & \alpha_j \end{bmatrix} \succeq 0 \quad \forall j\in[J]. \end{array} \right.$$

Rejoining the two supremum operators completes the proof. □

Note that problem (3.43) has a continuous objective function as well as a compact feasible set and is therefore solvable. Any optimal solution $(\mu^\star,\Sigma^\star,\{\alpha_j^\star,\theta_j^\star,\Theta_j^\star\}_j)$ can principally be used to construct an extremal distribution $\mathbb{Q}^\star$ that attains the supremum in the Gelbrich expected loss problem (3.38). Specifically, for any $j\in[J]$ let $\mathbb{Q}_j^\star$ be any distribution supported on

$$\Xi_j = \left\{\xi\in\mathbb{R}^n : \xi^\top Q_j\xi + 2q_j^\top\xi + q_j^0 \geq \xi^\top Q_{j'}\xi + 2q_{j'}^\top\xi + q_{j'}^0\ \forall j'\neq j\right\}.$$

If $\alpha_j^\star > 0$, we impose the additional requirement that $\mathbb{Q}_j^\star$ has mean value $\theta_j^\star/\alpha_j^\star$ and second-order moment matrix $\Theta_j^\star/\alpha_j^\star$. Such a distribution is indeed guaranteed to exist. One can then show that the mixture distribution $\mathbb{Q}^\star = \sum_{j\in[J]}\alpha_j^\star\cdot\mathbb{Q}_j^\star$ is optimal in (3.38). By construction, this distribution $\mathbb{Q}^\star$ has mean vector $\mu^\star$ and covariance matrix $\Sigma^\star$. We emphasize that problem (3.43) can be reformulated as a tractable SDP because the uncertainty set $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ is SDP-representable. Thus, problem (3.43) can be solved in polynomial time. Even though the mixture components $\mathbb{Q}_j^\star$, $j\in[J]$, are guaranteed to exist, however, one can prove that it is NP-hard to construct them. In other words, even though it is easy to solve (3.43) and even though any solution of (3.43) gives rise to a solution $\mathbb{Q}^\star$ of the Gelbrich expected loss evaluation problem (3.38), constructing $\mathbb{Q}^\star$ remains hard.

While exactly computable in polynomial time, the Gelbrich expected loss of a piecewise quadratic loss function may only provide a loose upper bound on the Wasserstein expected loss, which is often the actual quantity of interest. One can prove, however, that the Gelbrich expected loss (3.38) coincides with the worst-case expected loss with respect to a type-2 Wasserstein ball $\mathbb{B}_{\rho,2}(\widehat{\mathbb{P}})$ if the loss function is quadratic and the nominal distribution is elliptical.

Theorem 3.43 (Indefinite quadratic loss I)**.** *Assume that $\ell(\xi) = \xi^\top Q\xi + 2q^\top\xi$ with $Q \in \mathbb{S}^n$ and $q \in \mathbb{R}^n$ is a quadratic loss function. If $\sigma = 2$, $\widehat{\mu} \in \mathbb{R}^n$ and $\widehat{\Sigma} \in \mathbb{S}^n_+$, then for any $\rho \in \mathbb{R}_+$, the Gelbrich expected loss is equal to the optimal value of a tractable semidefinite program, that is,*

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] = \left\{ \begin{array}{cl} \inf & \gamma\big(\rho^2 - \|\widehat{\mu}\|^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\big) + z + \mathrm{Tr}\left[Z\right] \\ \mathrm{s.t.} & \gamma \in \mathbb{R}_+,\ z \in \mathbb{R}_+,\ Z \in \mathbb{S}^n_+ \\ & \begin{bmatrix} \gamma I - Q & \gamma\widehat{\Sigma}^{\frac{1}{2}} \\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0,\ \begin{bmatrix} \gamma I - Q & \gamma\widehat{\mu} + q \\ (\gamma\widehat{\mu} + q)^\top & z \end{bmatrix} \succeq 0. \end{array} \right. \tag{3.44}$$

Moreover, if $\widehat{\mathbb{P}} = \mathcal{P}_\phi(\widehat{\mu},\widehat{\Sigma})$ is an elliptical distribution with mean vector $\widehat{\mu}$ and covariance matrix $\widehat{\Sigma}$, then

$$\sup_{\mathbb{Q}\in\mathbb{B}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] = \sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)].$$

Proof. By applying the decomposition of the Gelbrich ambiguity set (3.8b), we have

$$\begin{aligned} \sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{E}_{\mathbb{Q}}\left[\ell(\xi)\right] &= \sup_{(\mu,M)\in\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})} \sup_{\mathbb{Q}\in\mathcal{P}_2(\mu,M-\mu\mu^\top)} \mathbb{E}_{\mathbb{Q}}\left[\ell(\xi)\right] \\ &= \sup_{(\mu,M)\in\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})} 2q^\top\mu + \mathrm{Tr}\left[QM\right] = \delta^\star_{\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(2q,Q). \end{aligned}$$

Applying the reformulation of the support function of $\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})$ using Lemma 3.22 completes the reformulation (3.44).

If $\widehat{\mathbb{P}} = \mathcal{P}_\phi(\widehat{\mu},\widehat{\Sigma})$, we can bound the worst-case expected loss over the Wasserstein ball $\mathbb{B}_{\rho,2}(\widehat{\mathbb{P}})$ by restricting to the subspace of all elliptical distributions sharing the same generator function ϕ with $\widehat{\mathbb{P}}$ as

$$\sup_{\mathbb{Q}\in\mathbb{B}_{\rho,2}(\widehat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] \geq \sup_{\mathbb{Q}\in\mathbb{B}_{\rho,\phi}(\widehat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)].$$

The decomposition of the worst-case expected loss over the ambiguity set $\mathbb{B}_{\rho,\phi}(\widehat{\mathbb{P}})$ can be written as

$$\begin{aligned} \sup_{\mathbb{Q}\in\mathbb{B}_{\rho,\phi}(\widehat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] &= \sup_{(\mu,M)\in\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})} \sup_{\mathbb{Q}\in\mathcal{P}_\phi(\mu,M-\mu\mu^\top)} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] \\ &= \sup_{(\mu,M)\in\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})} 2q^\top\mu + \mathrm{Tr}\left[QM\right] = \delta^\star_{\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(2q,Q) = \sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{E}_{\mathbb{Q}}\left[\ell(\xi)\right], \end{aligned}$$

where the equalities are from the decomposition (3.8b), the fact that the expectation of a quadratic function depends only on the first and second moment of the distribution, the definition of the support function of $\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})$ and the reformulation established in the first part of the proof. From Corollary 3.14, we have

$$\sup_{\mathbb{Q}\in\mathbb{B}_{\rho,2}(\widehat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] \leq \sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)],$$

This shows that

$$\sup_{\mathbb{Q}\in\mathbb{B}_{\rho,2}(\widehat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] = \sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] = \delta^{\star}_{\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(2q,Q),$$

and all of them are equal to the optimal value of (3.44). The proof is completed. □

The SDP (3.44) is easily obtained from (3.39) by setting $J=1$ and noting that $Y=Q$, $y=q$ and $y_0=0$ at optimality. As usual, a discrete extremal distribution $\mathbb{Q}^\star$ for the Gelbrich expected loss evaluation problem (3.38) can be derived from the dual of the SDP (3.44). In the following we denote the mean vector and the covariance matrix of $\mathbb{Q}^\star$ by $\mu^\star$ and $\Sigma^\star$, respectively. As $\mathbb{Q}^\star \in \mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})$, and as the Gelbrich hull is constructed solely on the basis of first- and second-order moment information, *any* distribution with mean vector $\mu^\star$ and covariance matrix $\Sigma^\star$ belongs to $\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})$, too. Moreover, as $\ell(\xi)$ is quadratic, the risk $\mathscr{R}_{\mathbb{Q}^\star}(\ell)$ depends on $\mathbb{Q}^\star$ only through its first- and second-order moments. This implies that *any* distribution with mean vector $\mu^\star$ and covariance matrix $\Sigma^\star$ is optimal in the Wasserstein expected loss evaluation problem.

Consider now the problem of evaluating the worst-case expected loss of the quadratic loss function $\ell(\xi)$ over a Wasserstein ball $\mathbb{B}_{\rho,2}(\widehat{\mathbb{P}})$ centered at an elliptial nominal distribution $\widehat{\mathbb{P}} = \mathscr{P}_\phi(\widehat{\mu},\widehat{\Sigma})$. Theorem 3.10 ensures that all elliptical distributions in the Gelbrich hull with the same characteristic generator as $\widehat{\mathbb{P}}$ belong to the Wasserstein ball $\mathbb{B}_{\rho,2}(\widehat{\mathbb{P}})$. This implies that the special elliptical distribution $\mathbb{Q}^\star = \mathscr{P}_\phi(\mu^\star,\Sigma^\star)$ is feasible in the Wasserstein expected loss evaluation problem. Moreover, we have

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathscr{R}_{\mathbb{Q}}(\ell) = \mathscr{R}_{\mathbb{Q}^\star}(\ell) \le \sup_{\mathbb{Q}\in\mathbb{B}_{\rho,2}(\widehat{\mathbb{P}})} \mathscr{R}_{\mathbb{Q}}(\ell) \le \sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathscr{R}_{\mathbb{Q}}(\ell),$$

where the equality holds because $\mathbb{Q}^\star$ is optimal in the Gelbrich expected loss evaluation problem (3.38), while the two inequalities follow from the feasibility of $\mathbb{Q}^\star$ in the worst-case risk evaluation problem under the Wasserstein ball $\mathbb{B}_{\rho,2}(\widehat{\mathbb{P}})$ and Corollary 3.14, respectively. Thus, all inequalities in the above expression are exact, which implies that $\mathbb{Q}^\star$ is actually optimal in the Wasserstein expected loss evaluation problem.

We highlight that the SDP (3.44) derived in Theorem 3.43 for *elliptical* nominal distributions accommodates only two linear matrix inequalities. If we take a non-parametric approach with *empirical* nominal distributions, the reformulation of the worst-case expected loss for quadratic loss function entails the same number of linear matrix inequalities as the number of atoms of the nominal distribution, and may thus be considerably harder to solve [179]. Next, we show how $\mathbb{Q}^\star$ can be constructed from the optimality conditions of the SDP (3.44).

Corollary 3.44 (Indefinite quadratic loss II)**.** *0 Suppose that all conditions of Theorem 3.43 hold. If* $\widehat{\Sigma} \succ 0$ *and either*

- $\lambda_{\max}(Q) > 0$*, or*

- $\lambda_{\max}(Q) < 0$ *and* $\rho^2 \le \operatorname{Tr}\big[\widehat{\Sigma}\big] + \|\widehat{\mu} + Q^{-1} q\|^2$
- $Q = -LL^\top \neq 0$ *for some* $L \in \mathbb{R}^{n\times k}$ *and* $\rho^2 \le \operatorname{Tr}\big[\widehat{\Sigma} L (L^\top L)^{-1} L^\top\big]$.

Then there exists $\gamma^\star \ge 0$ *with* $\gamma^\star I \succ Q$ *that solves the nonlinear algebraic equation*

$$\|\widehat{\mu} - (\gamma I - Q)^{-1}(\gamma \widehat{\mu} + q)\|^2 + \operatorname{Tr}\big[\widehat{\Sigma}\big(I - \gamma(\gamma I - Q)^{-1}\big)^2\big] = \rho^2, \tag{3.46a}$$

and the Gelbrich expected loss is attained by any probability measure with mean vector $\mu^\star \in \mathbb{R}^n$ *and covariance matrix* $\Sigma^\star \in \mathbb{S}^n_+$ *satisfying*

$$\mu^\star = (\gamma^\star I - Q)^{-1}\big(\gamma^\star \widehat{\mu} + q\big), \quad \Sigma^\star = \left(I - \frac{Q}{\gamma^\star}\right)^{-1} \widehat{\Sigma} \left(I - \frac{Q}{\gamma^\star}\right)^{-1}. \tag{3.46b}$$

If $\widehat{\mathbb{P}} = \mathscr{P}_\phi(\widehat{\mu}, \widehat{\Sigma})$, *then the elliptical distribution* $\mathbb{Q}^\star = \mathscr{P}_\phi(\mu^\star, \Sigma^\star)$ *attains the worst-case expected loss under both the Wasserstein and Gelbrich ambiguity set.*

Proof. The proof of Theorem 3.43 implies that

$$\sup_{\mathbb{Q} \in \mathbb{G}_{\rho,2}(\widehat{\mu}, \widehat{\Sigma})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] = \delta^\star_{\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})}(2q, Q).$$

Applying Lemma 3.23, one can show that $(\mu^\star, \Sigma^\star)$ defined in (3.46b) belongs to $\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})$ and attains the value $\delta^\star_{\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})}(2q, Q)$. This implies that any distribution with mean vector $\mu^\star$ and covariance matrix $\Sigma^\star$ will attain the value of the Gelbrich expected loss.

If $\widehat{\mathbb{P}} = \mathscr{P}_\phi(\widehat{\mu}, \widehat{\Sigma})$, we can easily verify that $\mathbb{Q}^\star = \mathscr{P}_\phi(\mu^\star, \Sigma^\star)$ belongs to $\mathbb{B}_{\rho,2}(\widehat{\mathbb{P}})$, and hence $\mathbb{Q}^\star$ is the extremal distribution for the Wasserstein expected loss. □

We end this section by providing the reformulation of the Gelbrich expected loss when the loss function can be expressed as the infimum convolution of two quadratic functions. The infimum convolution representation of the loss function can be used to model a more complex penalization when facing outliers, see Corollary 3.53 for an application in robust regression.

Theorem 3.45 (Infimum convolution loss functions)**.** *Suppose that* $\ell(\xi)$ *is the inf-convolution of two quadratic functions parametrized by* $\theta \in \mathbb{R}^k$ *as*

$$\ell(\xi) = \inf_{\vartheta \in \mathbb{R}^k} \ell_1(\vartheta, \xi) + \ell_2(\theta - \vartheta, \xi),$$

where $\ell_j(\theta, \xi) = \xi^\top Q_j(\theta)\xi + 2q_j(\theta)^\top \xi + q_j^0(\theta)$ *for some* $Q_j(\theta) \in \mathbb{S}^n$, $q_j(\theta) \in \mathbb{R}^n$ *and* $q_j^0(\theta) \in \mathbb{R}$ *for any* $j \in \{1,2\}$. *If* $\sigma = 2$, $\widehat{\mu} \in \mathbb{R}^n$, *and* $\widehat{\Sigma} \in \mathbb{S}^n_+$, *then for any* $\rho \in \mathbb{R}_{++}$ *the Gelbrich expected loss is* *equal to the optimal value of a finite dimensional optimization problem, that is,*

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] = \left\{\begin{array}{ll} \inf & y_0+\gamma\big(\rho^2-\|\widehat{\mu}\|^2-\mathrm{Tr}\big[\widehat{\Sigma}\big]\big)+z+\mathrm{Tr}\big[Z\big] \\ \mathrm{s.t.} & \gamma\in\mathbb{R}_+,\ y_0\in\mathbb{R},\ y\in\mathbb{R}^n,\ Y\in\mathbb{S}^n,\ z\in\mathbb{R}_+,\ Z\in\mathbb{S}^n_+,\ \vartheta\in\mathbb{R}^k \\ & \begin{bmatrix} \gamma I-Y & \gamma\widehat{\Sigma}^{\frac{1}{2}} \\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix}\succeq 0,\ \begin{bmatrix} \gamma I-Y & \gamma\widehat{\mu}+y \\ (\gamma\widehat{\mu}+y)^\top & z \end{bmatrix}\succeq 0 \\ & \begin{bmatrix} Y-Q_1(\vartheta)-Q_2(\theta-\vartheta) & y-q_1(\vartheta)-q_2(\theta-\vartheta) \\ (y-q_1(\vartheta)-q_2(\theta-\vartheta))^\top & y_0-q_1^0(\vartheta)-q_2^0(\theta-\vartheta) \end{bmatrix}\succeq 0. \end{array}\right. \tag{3.47}$$

If $Q_j(\theta)$ and $q_j(\theta)$ are linear functions of θ, and $q_j^0(\theta)$ are convex quadratic functions of θ for $j\in\{1,2\}$, then problem (3.47) *can be further reformulated as a semi-definite program.*

Proof. Let $\mathscr{Y}$ be the convex feasible set defined by

$$\begin{aligned} \mathscr{Y} &= \left\{y_0\in\mathbb{R},\ y\in\mathbb{R}^n,\ Y\in\mathbb{S}^n : y_0+2y^\top\xi+\xi^\top Y\xi\geq\ell(\xi)\quad\forall\xi\in\mathbb{R}^n\right\} \\ &= \left\{y_0\in\mathbb{R},\ y\in\mathbb{R}^n,\ Y\in\mathbb{S}^n : \exists\vartheta\in\mathbb{R}^k \text{ s.t. } \begin{bmatrix} Y-Q_1(\vartheta)-Q_2(\theta-\vartheta) & y-q_1(\vartheta)-q_2(\theta-\vartheta) \\ (y-q_1(\vartheta)-q_2(\theta-\vartheta))^\top & y_0-q_1^0(\vartheta)-q_2^0(\theta-\vartheta) \end{bmatrix}\succeq 0\right\}, \end{aligned}$$

where the last equality is obtained by simply substituting the expression of the loss function $\ell(\xi)$ and formulating the quadratic constraints as semidefinite constraints using S-lemma [139]. Following the similar steps as in the proof of Theorem 3.41, we have

$$\begin{aligned} \sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] &= \sup_{(\mu,\Sigma)\in\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})}\ \sup_{\mathbb{Q}\in\mathscr{P}_2(\mu,\Sigma)} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] \\ &\leq \inf_{(y_0,y,Y)\in\mathscr{Y}} \left\{y_0+\delta^\star_{\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(2y,Y)\right\}, \end{aligned}$$

where the inequality can be further shown to hold as an equality using the same techniques as in the second part of the proof of Theorem 3.41. Substituting the formula for the support function of $\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})$ in Lemma 3.22 and the definition of $\mathscr{Y}$ into the above infimum problem gives the reformulation (3.47).

When $q_j(\theta)$ are linear functions of θ, and $Q_j(\theta)$, $q_j^0(\theta)$ are convex quadratic functions of θ for $j\in\{1,2\}$, the last constraint of (3.47) can be reformulated as semidefinite constraints using standard techniques [170, Section 2]. This finishes the proof. □

3.6 Extensions

We present in this section several extensions of the worst-case risk under the Gelbrich hull

3.6.1 Robust Mean-Variance Portfolios

In this section, we consider the situation where the consistent family of risk measures $\{\mathscr{R}_{\mathbb{Q}}\}_{\mathbb{Q}\in\mathscr{P}_\sigma}$ is related to the popular class of mean-variance risk measures. Given the mean vector $\mu \in \mathbb{R}^n$ and the covariance matrix $\Sigma \in \mathbb{S}_+^n$ of the asset returns, the risk $\mathscr{R}_{\mathbb{Q}}(\ell)$ of a linear portfolio loss $\ell(\xi) = -w^\top \xi$ can be written as a weighted combination of the mean and variance as

$$\mathscr{R}_{\mathbb{Q}}(-w^\top \xi) = \mathbb{E}_{\mathbb{Q}}[-w^\top \xi] + \alpha \mathbb{V}\mathrm{ar}_{\mathbb{Q}}(-w^\top \xi),$$

where $\mathbb{V}\mathrm{ar}_{\mathbb{Q}}(-w^\top \xi)$ denotes the variance of the loss $-w^\top \xi$ under the probability $\mathbb{Q}$. The mean-variance risk measure is the building block of modern portfolio theory introduced in [115], and it also arises in many other settings including but not limited to the entropic risk measure for Gaussian returns [61] and the log-optimal portfolio problem [99].

Because the mean-variance risk measure is not positive homogeneous, the results in Section 3.4 are not applicable to calculate the Gelbrich mean-variance risk. Nevertheless, for a linear portfolio loss $\ell = -w^\top \xi$, the Gelbrich mean-variance risk can be expressed using the two-layer decomposition (3.8a) as

$$\begin{aligned}\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})} \mathscr{R}_{\mathbb{Q}}(\ell) &= \sup_{(\mu,\Sigma)\in\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \ \sup_{\mathbb{Q}\in\mathscr{P}_\sigma(\mu,\Sigma)} \mathscr{R}_{\mathbb{Q}}(-w^\top \xi) \\ &= \sup_{(\mu,\Sigma)\in\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \left\{-w^\top \mu + \alpha w^\top \Sigma w\right\} = \delta^\star_{\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(-w, \alpha w w^\top).\end{aligned}$$

At this point, the reformulation of the support function of $\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ in Lemma 3.20 can be readily applied to reformulate the Gelbrich mean-variance risk as the optimal value of a semidefinite program. The next theorem asserts that under additional assumptions on α and $\widehat{\Sigma}$, the resulting optimization can be further reduced to a second-order cone program.

Theorem 3.46 (Gelbrich mean-variance risk). *Suppose that $\{\mathscr{R}_{\mathbb{Q}}\}_{\mathbb{Q}\in\mathscr{P}_\sigma}$ is a family of mean-variance risk measure with $\alpha > 0$. If $\widehat{\Sigma} \succ 0$, the Gelbrich mean-variance risk of a linear portfolio loss $\ell(\xi) = -w^\top \xi$ is equal to the optimal value of the following second order cone program*

$$\begin{array}{cl}\inf & \gamma\rho^2 - \widehat{\mu}^\top w + \frac{1}{4} y + \alpha z \\ \text{s.t.} & \gamma \in \mathbb{R}_+,\ y \in \mathbb{R}_+,\ z \in \mathbb{R}_+ \\ & \left\| \begin{pmatrix} 2\widehat{\Sigma}^{\frac{1}{2}} w \\ z + \alpha y - 1 \end{pmatrix} \right\| \leq z - \alpha y + 1,\ \left\| \begin{pmatrix} 2w \\ y - \gamma \end{pmatrix} \right\| \leq y + \gamma,\ \alpha y \leq 1.\end{array} \tag{3.48}$$

Proof. By applying Lemma 3.20, we can rewrite the Gelbrich mean-variance risk measure as

$$\begin{aligned}\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})} \mathscr{R}_{\mathbb{Q}}(-w^\top\xi) &= \delta^\star_{\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(-w,\alpha ww^\top)\\ &= \left\{\begin{array}{ll}\inf & -\widehat{\mu}^\top w+\tau+\gamma\big(\rho^2-\mathrm{Tr}\big[\widehat{\Sigma}\big]\big)+\mathrm{Tr}\left[Z\right]\\ \mathrm{s.t.} & \gamma\in\mathbb{R}_+,\ \tau\in\mathbb{R}_+,\ Z\in\mathbb{S}^n_+\\ & \begin{bmatrix}\gamma I-\alpha ww^\top & \gamma\widehat{\Sigma}^{\frac{1}{2}}\\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z\end{bmatrix}\succeq 0,\ \left\|\begin{pmatrix}\|w\|\\ \tau-\gamma\end{pmatrix}\right\|\le\tau+\gamma.\end{array}\right.\end{aligned}$$

When $\alpha>0$ and $\widehat{\Sigma}\succ 0$, we can show using a limit argument that the objective value of the above optimization problem tends to infinity as γ tends to $\alpha\|w\|^2\ge 0$. Thus, we can rewrite the above semidefinite program using the Schur complement as

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})} \mathscr{R}_{\mathbb{Q}}\left(-w^\top\xi\right)=\left\{\begin{array}{ll}\inf & -\widehat{\mu}^\top w+\frac{\|w\|^2}{4\gamma}+\gamma\big(\rho^2-\mathrm{Tr}\big[\widehat{\Sigma}\big]\big)+\gamma^2\,\mathrm{Tr}\left[(\gamma I-\alpha ww^\top)^{-1}\widehat{\Sigma}\right]\\ \mathrm{s.t.} & \gamma\in\mathbb{R}_+,\ \gamma>\alpha\|w\|^2.\end{array}\right.$$

Applying the Sherman-Morrison formula [12, Corollary 2.8.8] to re-express the matrix inversion term, we find for any $\gamma>\max\{0,\alpha\|w\|^2\}$

$$\gamma^2\,\mathrm{Tr}\left[(\gamma I-\alpha ww^\top)^{-1}\widehat{\Sigma}\right]=\gamma\,\mathrm{Tr}\big[\widehat{\Sigma}\big]+\frac{\alpha w^\top\widehat{\Sigma}w}{1-\frac{\alpha}{\gamma}\|w\|^2}.$$

Substituting this expression into the reformulation of the Gelbrich risk gives

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})} \mathscr{R}_{\mathbb{Q}}(-w^\top\xi)=\inf_{\gamma>\alpha\|w\|^2}\quad\gamma\rho^2-\widehat{\mu}^\top w+\frac{\|w\|^2}{4\gamma}+\alpha(1-\alpha\gamma^{-1}\|w\|^2)^{-1}w^\top\widehat{\Sigma}w.$$

By adding an auxiliary variable $y\ge 0$ coupled with the constraint $\|w\|^2/\gamma\le y\le\alpha^{-1}$, we have

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,\sigma}(\widehat{\mu},\widehat{\Sigma})} \mathscr{R}_{\mathbb{Q}}\left(-w^\top\xi\right)=\left\{\begin{array}{ll}\inf & \gamma\rho^2-\widehat{\mu}^\top w+\frac{y}{4}+\alpha(1-\alpha y)^{-1}w^\top\widehat{\Sigma}w\\ \mathrm{s.t.} & \|w\|^2/\gamma\le y\le\alpha^{-1},\ \gamma>0.\end{array}\right.$$

The last part of the proof involves rewriting the quadratic-over-linear term using the second order cone reformulation [111, Equation (8)] with the epigraphical variable z as

$$\frac{w^\top\widehat{\Sigma}w}{1-\alpha y}\le z\quad\Longleftrightarrow\quad\left\|\begin{pmatrix}2\widehat{\Sigma}^{\frac{1}{2}}w\\ z+\alpha y-1\end{pmatrix}\right\|\le z-\alpha y+1.$$

This completes the proof. □

To construct the extremal distribution for the Gelbrich mean-variance risk, we can directly use the result of Lemma 3.21 to characterize the first two moments of the extremal distribution. In addition, we can also conclude that when $\sigma=\phi$, the extremal distribution is unique because an elliptical distribution with generator function ϕ is uniquely defined by its mean vector and covariance matrix. These results are omitted.

3.6.2 Gelbrich Polyhedral Value-at-Risk

We now consider the problem of evaluating the Gelbrich VaR of a portfolio that consists of both stocks and their derivatives. A naïve approach is to treat the derivatives as usual stocks and proceed with the evaluation of the Gelbrich VaR of a linear portfolio loss as introduced in Section 3.4. However, because the Gelbrich risk depends only on the first- and second-moment of the joint stock-derivative return distribution, this approach fails to capture the inherent *non*linear dependency of the derivative returns on the stock returns. Thus, we consider an explicit portfolio model that consists of an allocation $w \in \mathcal{W} \subseteq \mathbb{R}^n$ over n stocks and a derivative allocation $x \in \mathcal{X} \subseteq \mathbb{R}^k_+$ over k derivatives. For simplicity, we consider only vanilla derivatives of the stocks, and in this case, the return of the derivatives can be written as a piecewise linear function of the stock return ξ. Given a mixed portfolio $(w, x) \in \mathcal{W} \times \mathcal{X}$, the portfolio loss can be written as a polyhedral function of ξ for some matrices $A, B \in \mathbb{R}^{k \times n}$ and vectors $a, b \in \mathbb{R}^k$ as

$$\ell(\xi) = -w^\top \xi - x^\top \max\{A\xi + a, B\xi + b\},$$

where the max operator is understood as the element-wise maximum of two vectors. We emphasize that the short selling of derivatives is prohibited in this model.

The Gelbrich VaR defined in (3.35) can be employed to evaluate the worst-case VaR of a polyhedral loss function over all possible probability measures contained in the Gelbrich hull $\mathbb{G}_{\rho,2}(\widehat{\mu}, \widehat{\Sigma})$. The next theorem demonstrates that the Gelbrich polyhedral VaR can be computed efficiently by solving a semidefinite program.

Theorem 3.47 (Gelbrich polyhedral VaR)**.** *Suppose that $\ell(\xi) = w^\top \xi + x^\top \max\{A\xi + a, B\xi + b\}$ for some matrices $A, B \in \mathbb{R}^{k \times n}$ and vectors $a, b \in \mathbb{R}^k$. For any $\rho \in \mathbb{R}_{++}$ and $\beta \in (0,1)$, the Gelbrich polyhedral Value-at-Risk is equivalent to the optimal value of a semidefinite program, that is,*

$$\sup_{\mathbb{Q} \in \mathbb{G}_{\rho,2}(\widehat{\mu}, \widehat{\Sigma})} \mathbb{Q}\text{-VaR}_\beta(\ell(\xi)) = \left\{ \begin{array}{cl} \inf & \tau \\ \text{s.t.} & y_0 \in \mathbb{R},\ y \in \mathbb{R}^n,\ Y \in \mathbb{S}^n_+,\ \gamma \in \mathbb{R}_+,\ z \in \mathbb{R}_+,\ Z \in \mathbb{S}^n_+ \\ & \tau \in \mathbb{R},\ \eta \in \mathbb{R}_+,\ t \in \mathbb{R}^k_+,\ v \in \mathbb{R}^k \\ & y_0 + \gamma\big(\rho^2 - \|\widehat{\mu}\|^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\big) + z + \mathrm{Tr}\big[Z\big] \leq \eta\beta \\ & \begin{bmatrix} \gamma I - Y & \gamma \widehat{\Sigma}^{\frac{1}{2}} \\ \gamma \widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0,\ \begin{bmatrix} \gamma I - Y & \gamma \widehat{\mu} + y \\ (\gamma \widehat{\mu} + y)^\top & z \end{bmatrix} \succeq 0,\ \begin{bmatrix} Y & y \\ y^\top & y_0 \end{bmatrix} \succeq 0 \\ & t \leq x,\ v = w + (A - B)^\top t + B^\top x \\ & \begin{bmatrix} Y & y \\ y^\top & y_0 \end{bmatrix} + \begin{bmatrix} 0 & v \\ v^\top & -\eta + 2(\tau + (a-b)^\top t + b^\top x) \end{bmatrix} \succeq 0. \end{array} \right. \tag{3.49}$$

Proof. For any $\tau \in \mathbb{R}$, define the following set

$$\mathcal{S}_\tau = \left\{\xi \in \mathbb{R}^n : \tau + w^\top \xi + x^\top \max\{A\xi + a, B\xi + b\} \leq 0\right\}.$$

By applying Lemma 3.55, we have

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\left(\tau \leq -w^\top \xi - x^\top \max\{A\xi + a, B\xi + b\}\right) = \sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}(\xi \in \mathscr{S}_\tau)$$

$$= \left\{ \begin{array}{ll} \inf & y_0 + \gamma\big(\rho^2 - \|\widehat{\mu}\|^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\big) + z + \mathrm{Tr}\big[Z\big] \\ \text{s.t.} & \gamma \in \mathbb{R}_+,\ y_0 \in \mathbb{R},\ y \in \mathbb{R}^n,\ Y \in \mathbb{S}^n,\ z \in \mathbb{R}_+,\ Z \in \mathbb{S}^n_+ \\ & \begin{bmatrix} \gamma I - Y & \gamma\widehat{\Sigma}^{\frac{1}{2}} \\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0,\ \begin{bmatrix} \gamma I - Y & \gamma\widehat{\mu} + y \\ (\gamma\widehat{\mu} + y)^\top & z \end{bmatrix} \succeq 0,\ \begin{bmatrix} Y & y \\ y^\top & y_0 \end{bmatrix} \succeq 0 \\ & y_0 + 2y^\top \xi + \xi^\top Y \xi \geq 1 \quad \forall \xi \in \mathscr{S}_\tau. \end{array} \right. \tag{3.50}$$

By applying the same argument in the proof of [184, Theorem 4.1], problem (3.50) is equivalent to the following semidefinite program

$$\begin{array}{ll} \inf & y_0 + \gamma\big(\rho^2 - \|\widehat{\mu}\|^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\big) + z + \mathrm{Tr}\big[Z\big] \\ \text{s.t.} & \gamma \in \mathbb{R}_+,\ y_0 \in \mathbb{R},\ y \in \mathbb{R}^n,\ Y \in \mathbb{S}^n,\ z \in \mathbb{R}_+,\ Z \in \mathbb{S}^n_+ \\ & \eta \in \mathbb{R}_+,\ t \in \mathbb{R}^k_+,\ v \in \mathbb{R}^k \\ & \begin{bmatrix} \gamma I - Y & \gamma\widehat{\Sigma}^{\frac{1}{2}} \\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0,\ \begin{bmatrix} \gamma I - Y & \gamma\widehat{\mu} + y \\ (\gamma\widehat{\mu} + y)^\top & z \end{bmatrix} \succeq 0,\ \begin{bmatrix} Y & y \\ y^\top & y_0 \end{bmatrix} \succeq 0, \\ & t \leq x,\ v = w + (A - B)^\top t + B^\top x \\ & \begin{bmatrix} Y & y \\ y^\top & y_0 \end{bmatrix} + \begin{bmatrix} 0 & \eta v \\ \eta v^\top & -1 + 2\eta(\tau + (a - b)^\top t + b^\top x) \end{bmatrix} \succeq 0 \end{array}$$

for all but one value of τ. Because the Gelbrich polyhedral VaR can be rewritten as

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-}\mathrm{VaR}_\beta(\ell(\xi)) = \inf\left\{ \tau \in \mathbb{R} : \sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}(\tau \leq \ell(\xi)) \leq \beta \right\},$$

we can re-express the Gelbrich polyhedral VaR as the optimal value of the following optimization problem

$$\begin{array}{ll} \inf & \tau \\ \text{s.t.} & \gamma \in \mathbb{R}_+,\ y_0 \in \mathbb{R},\ y \in \mathbb{R}^n,\ Y \in \mathbb{S}^n,\ z \in \mathbb{R}_+,\ Z \in \mathbb{S}^n_+ \\ & \tau \in \mathbb{R},\ \eta \in \mathbb{R}_+,\ t \in \mathbb{R}^k_+,\ v \in \mathbb{R}^k \\ & y_0 + \gamma\big(\rho^2 - \|\widehat{\mu}\|^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\big) + z + \mathrm{Tr}\big[Z\big] \leq \beta \\ & \begin{bmatrix} \gamma I - Y & \gamma\widehat{\Sigma}^{\frac{1}{2}} \\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0,\ \begin{bmatrix} \gamma I - Y & \gamma\widehat{\mu} + y \\ (\gamma\widehat{\mu} + y)^\top & z \end{bmatrix} \succeq 0,\ \begin{bmatrix} Y & y \\ y^\top & y_0 \end{bmatrix} \succeq 0, \\ & t \leq x,\ v = w + (A - B)^\top t + B^\top x \\ & \begin{bmatrix} Y & y \\ y^\top & y_0 \end{bmatrix} + \begin{bmatrix} 0 & \eta v \\ \eta v^\top & -1 + 2\eta(\tau + (a - b)^\top t + b^\top x) \end{bmatrix} \succeq 0. \end{array} \tag{3.51}$$

Problem (3.51) is non-convex because of the bilinear terms in the last constraint. It can be shown that any feasible solution of (3.51) with vanishing η-component will satisfy

$$y_0 + \gamma\big(\rho^2 - \|\widehat{\mu}\|^2 - \operatorname{Tr}\big[\widehat{\Sigma}\big]\big) + z + \operatorname{Tr}\big[Z\big] \geq 1,$$

which is in conflict with the constraint

$$y_0 + \gamma\big(\rho^2 - \|\widehat{\mu}\|^2 - \operatorname{Tr}\big[\widehat{\Sigma}\big]\big) + z + \operatorname{Tr}\big[Z\big] \leq \beta$$

for $\beta \in (0,1)$. This implies that any feasible solution of (3.51) will have $\eta > 0$. Thus, we can divide all constraints of problem (3.51) by η. Subsequently, we substitute η by $1/\eta$ and substitute $(\gamma, y_0, y, Y, z, Z)$ in (3.51) by $(\gamma/\eta, y_0/\eta, y/\eta, Y/\eta, z/\eta, Z/\eta)$. This completes the proof. □

The constraints in the semidefinite program reformulation (3.49) are linear in terms of (w, x), and thus the portfolio optimization problem that minimizes the Gelbrich polyhedral VaR can be formulated as a finite convex optimization problem provided that $\mathcal{W}$ and $\mathcal{X}$ are both convex, conic representable sets. While in this section we have assumed that there is no coupling constraint between the stock allocation feasible set $\mathcal{W}$ and the derivative allocation feasible set $\mathcal{X}$, incorporating this interaction is straightforward provided that the coupling constraints can be written using conic constraints.

3.6.3 Gelbrich Quadratic Value-at-Risk

The results of Section 3.6.2 rely fundamentally on the restriction that the stochastic returns of the derivatives can be modelled using a piecewise linear function of the stock returns ξ. In practice, an exotic derivative may exhibit strongly *non*linear dependence on the stock returns ξ, hence the polyhedral model in Section 3.6.2 is inadequate for assessing the risk of a complex portfolio position. We consider in this section an approximation of the portfolio return using a second-order Taylor expansion which is commonly known as the delta-gamma approximation [91]. Given a portfolio allocation $w \in \mathbb{R}^n$ over both stocks and derivatives, the loss of a portfolio is approximated using a quadratic function of ξ as

$$\ell(\xi) = -\theta(w) - \Delta(w)^\top \xi - \frac{1}{2}\xi^\top \Gamma(w)\xi,$$

where $\theta(w) \in \mathbb{R}$, $\Delta(w) \in \mathbb{R}^n$ and $\Gamma(w) \in \mathbb{S}^n$ are parameters derived from the current portfolio position w. We emphasize that ℓ is possibly a non-convex function of ξ and that we do not prohibit the short selling of derivative in this section.

Theorem 3.48 (Gelbrich quadratic VaR)**.** *Suppose that $\ell(\xi) = -\theta(w) - \Delta(w)^\top \xi - \frac{1}{2} w^\top \Gamma(w)\xi$ for some $\theta(w) \in \mathbb{R}$, $\Delta(w) \in \mathbb{R}^n$ and $\Gamma(w) \in \mathbb{S}^n$. For any $\rho \in \mathbb{R}_{++}$ and $\beta \in (0,1)$, the Gelbrich quadratic VaR of the quadratic loss function $\ell(\xi)$ is equivalent to the optimal value of a semidefinite* *program, that is,*

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-VaR}_\beta(\ell(\xi)) = \left\{\begin{array}{ll} \inf & \tau \\ \text{s.t.} & y_0 \in \mathbb{R},\ y\in\mathbb{R}^n,\ Y\in\mathbb{S}_+^n,\ \gamma\in\mathbb{R}_+,\ z\in\mathbb{R}_+,\ Z\in\mathbb{S}_+^n,\ \tau\in\mathbb{R},\ \eta\in\mathbb{R}_+ \\ & y_0+\gamma\big(\rho^2-\|\widehat{\mu}\|^2-\operatorname{Tr}\big[\widehat{\Sigma}\big]\big)+z+\operatorname{Tr}\big[Z\big]\le\eta\beta \\ & \begin{bmatrix}\gamma I-Y & \gamma\widehat{\Sigma}^{\frac{1}{2}}\\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z\end{bmatrix}\succeq 0,\ \begin{bmatrix}\gamma I-Y & \gamma\widehat{\mu}+y\\ (\gamma\widehat{\mu}+y)^\top & z\end{bmatrix}\succeq 0,\ \begin{bmatrix}Y & y\\ y^\top & y_0\end{bmatrix}\succeq 0 \\ & \begin{bmatrix}Y & y\\ y^\top & y_0\end{bmatrix}+\begin{bmatrix}\Gamma(w) & \Delta(w)\\ \Delta(w)^\top & -\eta+2(\tau+\theta(w))\end{bmatrix}\succeq 0. \end{array}\right. \tag{3.52}$$

Proof. For any $\tau\in\mathbb{R}$, define the following set

$$\mathscr{S}_\tau=\left\{\xi\in\mathbb{R}^n:\tau+\theta(w)+\Delta(w)^\top\xi+\frac{1}{2}\xi^\top\Gamma(w)\xi\le 0\right\}.$$

By applying Lemma 3.55, we have for any $\tau\in\mathbb{R}$

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})}\mathbb{Q}\left(\tau\le-\theta(w)-\Delta(w)^\top\xi-\frac{1}{2}\xi^\top\Gamma(w)\xi\right)=\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})}\mathbb{Q}(\xi\in\mathscr{S}_\tau)$$

$$=\left\{\begin{array}{ll} \inf & y_0+\gamma\big(\rho^2-\|\widehat{\mu}\|^2-\operatorname{Tr}\big[\widehat{\Sigma}\big]\big)+z+\operatorname{Tr}\big[Z\big] \\ \text{s.t.} & \gamma\in\mathbb{R}_+,\ y_0\in\mathbb{R},\ y\in\mathbb{R}^n,\ Y\in\mathbb{S}^n,\ z\in\mathbb{R}_+,\ Z\in\mathbb{S}_+^n \\ & \begin{bmatrix}\gamma I-Y & \gamma\widehat{\Sigma}^{\frac{1}{2}}\\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z\end{bmatrix}\succeq 0,\ \begin{bmatrix}\gamma I-Y & \gamma\widehat{\mu}+y\\ (\gamma\widehat{\mu}+y)^\top & z\end{bmatrix}\succeq 0,\ \begin{bmatrix}Y & y\\ y^\top & y_0\end{bmatrix}\succeq 0 \\ & y_0+2y^\top\xi+\xi^\top Y\xi\ge 1\quad\forall\xi\in\mathscr{S}_\tau. \end{array}\right. \tag{3.53}$$

By applying the S-lemma [139] to reformulate the semi-infinite constraint into an equivalent semidefinite constraint, problem (3.53) is equivalent to

$$\begin{array}{ll} \inf & y_0+\gamma\big(\rho^2-\|\widehat{\mu}\|^2-\operatorname{Tr}\big[\widehat{\Sigma}\big]\big)+z+\operatorname{Tr}\big[Z\big] \\ \text{s.t.} & \gamma\in\mathbb{R}_+,\ y_0\in\mathbb{R},\ y\in\mathbb{R}^n,\ Y\in\mathbb{S}^n,\ z\in\mathbb{R}_+,\ Z\in\mathbb{S}_+^n,\ \eta\in\mathbb{R}_+ \\ & \begin{bmatrix}\gamma I-Y & \gamma\widehat{\Sigma}^{\frac{1}{2}}\\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z\end{bmatrix}\succeq 0,\ \begin{bmatrix}\gamma I-Y & \gamma\widehat{\mu}+y\\ (\gamma\widehat{\mu}+y)^\top & z\end{bmatrix}\succeq 0,\ \begin{bmatrix}Y & y\\ y^\top & y_0\end{bmatrix}\succeq 0 \\ & \begin{bmatrix}Y & y\\ y^\top & y_0\end{bmatrix}+\begin{bmatrix}\eta\Gamma(w) & \eta\Delta(w)\\ \eta\Delta(w)^\top & -1+2\eta(\tau+\theta(w))\end{bmatrix}\succeq 0. \end{array}$$

The Gelbrich quadratic VaR is thus equivalent to the optimal value of the problem

$$
\begin{array}{ll}
\inf & \tau \\
\text{s.t.} & \gamma \in \mathbb{R}_+,\ y_0 \in \mathbb{R},\ y \in \mathbb{R}^n,\ Y \in \mathbb{S}^n,\ z \in \mathbb{R}_+,\ Z \in \mathbb{S}_+^n,\ \tau \in \mathbb{R},\ \eta \in \mathbb{R}_+ \\
& y_0 + \gamma\big(\rho^2 - \|\widehat{\mu}\|^2 - \mathrm{Tr}\,[\widehat{\Sigma}]\big) + z + \mathrm{Tr}\,[Z] \le \beta \\
& \begin{bmatrix} \gamma I - Y & \gamma \widehat{\Sigma}^{\frac{1}{2}} \\ \gamma \widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0,\ \begin{bmatrix} \gamma I - Y & \gamma \widehat{\mu} + y \\ (\gamma \widehat{\mu} + y)^\top & z \end{bmatrix} \succeq 0,\ \begin{bmatrix} Y & y \\ y^\top & y_0 \end{bmatrix} \succeq 0 \\
& \begin{bmatrix} Y & y \\ y^\top & y_0 \end{bmatrix} + \begin{bmatrix} \eta \Gamma(w) & \eta \Delta(w) \\ \eta \Delta(w)^\top & -1 + 2\eta(\tau + \theta(w)) \end{bmatrix} \succeq 0.
\end{array}
\tag{3.54}
$$

One can verify using an analogous reasoning as in the proof of Theorem 3.47 that no feasible solution of (3.54) has a vanishing η component. Dividing the constraints of (3.54) by $\eta > 0$ and perform an analogous variables substitution as in Theorem 3.47 completes the proof. □

For a delta-gamma approximation, $\theta(w)$, $\Delta(w)$ and $\Gamma(w)$ are all linear functions of the portfolio position w, as a consequence, the problem of finding the optimal portfolio that minimizes the Gelbrich quadratic VaR is a finite convex optimization problem provided that the feasible portfolio allocation set $\mathcal{W}$ is conic representable.

3.6.4 Product of Lognormal Random Variables

The product of random variables that arises in many applications related to statistics, physics and various branches of sciences [66] and it governs many multiplicative process that explains the evolution of economic indices and biological measurements [2]. In finance, product of random variables is widely used to model the discrete-time wealth level subject to stochastic returns. In this section we are interested in evaluating the tail probability of a product of non-negative but ambiguous random variables, which are heavily used to calculate the risk of bankruptcy or to stress test the financial system. Instead of modelling ambiguity using the Chebyshev ambiguity set that directly imposes the first- and second-moment information regarding the distribution of the random variables [146], we assume that the random variables follow an ambiguous log-normal distribution. As a consequence, to model the primitive random variables we restrict ourselves to the family of Gaussian distributions, which is a special instance of elliptical distribution with characteristic generator $\phi(u) = \exp(-u/2)$. For a given measure $\mathbb{Q} \in \mathscr{P}_\phi(\mu, \Sigma)$, ξ is normally distributed with mean vector $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{S}_+^n$, and we denote by $\exp(\xi)$ the n-dimensional lognormal distribution with its i-th element defined by $\exp(\xi)_i = \exp(\xi_i)\ \forall i = 1, \ldots, n$. The problem of evaluating the tail probability of the product of random variables is equivalent to verifying sequentially whether

$$
\mathbb{Q}\left(\prod_{i \in [n]} \exp(\xi_i) \le T \right) \le \beta
$$

is true for some threshold $T \in \mathbb{R}_{++}$ and some uncertainty level $\beta \in (0, \frac{1}{2})$. When the true probability measure is unknown, we instead resort to evaluating the worst-case tail probability under the Gelbrich hull centered around the nominal distribution $\widehat{\mathbb{P}}$ which is a Gaussian distribution with mean $\widehat{\mu} \in \mathbb{R}^n$ and covariance matrix $\widehat{\Sigma} \in \mathbb{S}_+^n$. As a consequence, we are interested in verifying whether

$$\sup_{\mathbb{Q} \in \mathbb{G}_{\rho,\phi}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\left(\prod_{i \in [n]} \exp(\xi_i) \leq T \right) \leq \beta \tag{3.55}$$

is true for different levels of T and β. The next proposition shows that verifying the above relationship is equivalent to verifying a simple inequality.

Proposition 3.49. *Suppose that ϕ is the characteristic generator of Gaussian distributions. For any $T \in \mathbb{R}_{++}$ and $\beta \in (0, \frac{1}{2})$, the relation* (3.55) *is true if and only if*

$$-\widehat{\mu}^\top e + F_\phi^{-1}(1-\beta)\sqrt{e^\top \widehat{\Sigma} e} + \rho\sqrt{n(1 + (F_\phi^{-1}(1-\beta))^2)} \leq -\log T,$$

where e is a vector of ones and F_ϕ^{-1} is the inverse cumulative distribution function of a standard Gaussian distribution.

Proof. Using the definition of the Gelbrich Value-at-Risk in Lemma 3.31, for any threshold $T \in \mathbb{R}_{++}$ we find

$$\begin{aligned}
\sup_{\mathbb{Q} \in \mathbb{G}_{\rho,\phi}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\left(\prod_{i \in [n]} \exp(\xi_i) \leq T \right) \leq \beta \iff & \sup_{\mathbb{Q} \in \mathbb{G}_{\rho,\phi}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\left(\frac{1}{T} \leq \prod_{i \in [n]} \exp(-\xi_i) \right) \leq \beta \\
\iff & \sup_{\mathbb{Q} \in \mathbb{G}_{\rho,\phi}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\left(-\log T \leq -e^\top \xi \right) \leq \beta \\
\iff & \sup_{\mathbb{Q} \in \mathbb{G}_{\rho,\phi}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-}\mathrm{VaR}_\beta(-e^\top \xi) \leq -\log T,
\end{aligned}$$

where e is a vector of ones. Because ϕ is the characteristic generator of Gaussian distribution and $\beta \in (0, \frac{1}{2})$, Lemma 3.31 part (iv) can be utilized to reformulate the last constraint as

$$-\widehat{\mu}^\top e + F_\phi^{-1}(1-\beta)\sqrt{e^\top \widehat{\Sigma} e} + \rho\sqrt{n(1 + (F_\phi^{-1}(1-\beta))^2)} \leq -\log T,$$

which completes the proof. □

3.7 Injecting Support Information

In many cases, the decision maker possesses additional information that ξ is supported on a strict subset $\Xi \subset \mathbb{R}^n$. This information can be used to alleviate the conservativeness of the distributionally robust solution. We denote by $\mathscr{P}(\Xi)$ the collection of all probability measures on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ with support Ξ. In an analogous counterpart of Definition 3.5 for measures with support information, for $\sigma \in \{2, \mathrm{S}, \mathrm{SU}\}$ we define

$$\mathscr{P}_\sigma(\Xi) \triangleq \mathscr{P}_\sigma \cap \mathscr{P}(\Xi)$$

as the set of probability measures supported on Ξ that satisfy the information structure σ. We emphasize that the structural information $\sigma = \phi$ typically requires that $\Xi = \mathbb{R}^n$, thus the case $\sigma = \phi$ will be omitted in this section.

Given a nominal measure $\widehat{\mathbb{P}}$ whose support is contained in Ξ, we define the *Wasserstein ambiguity set with support information* as the ball of radius $\rho \geq 0$ in $\mathscr{P}_\sigma(\Xi)$ centered at $\widehat{\mathbb{P}}$ with respect to the type-2 Wasserstein distance

$$\mathbb{B}_{\rho,\sigma}(\Xi, \widehat{\mathbb{P}}) \triangleq \left\{\mathbb{Q} \in \mathscr{P}_\sigma(\Xi) : \mathbb{W}(\mathbb{Q}, \widehat{\mathbb{P}}) \leq \rho\right\}.$$

The *Gelbrich hull with support information* $\mathbb{G}_{\rho,\sigma}(\Xi, \widehat{\mu}, \widehat{\Sigma})$ associated with a mean vector $\widehat{\mu} \in \mathbb{R}^n$ and a covariance matrix $\widehat{\Sigma} \in \mathbb{S}^n_+$ can be defined in an analogous way as in Definition 3.12 as

$$\mathbb{G}_{\rho,\sigma}(\Xi, \widehat{\mu}, \widehat{\Sigma}) \triangleq \left\{\mathbb{Q} \in \mathscr{P}_\sigma(\Xi) : \left(\mathbb{E}_\mathbb{Q}[\xi], \mathbb{E}_\mathbb{Q}[(\xi - \mathbb{E}_\mathbb{Q}[\xi])(\xi - \mathbb{E}_\mathbb{Q}[\xi])^\top]\right) \in \mathscr{U}_\rho(\widehat{\mu}, \widehat{\Sigma})\right\}$$

for $\sigma \in \{2, \mathrm{S}, \mathrm{SU}\}$. One can readily verify that $\mathbb{G}_{\rho,\sigma}(\Xi, \widehat{\mu}, \widehat{\Sigma})$ is a superset of $\mathbb{B}_{\rho,\sigma}(\Xi, \widehat{\mathbb{P}})$ whenever the mean vector and the covariance matrix of the nominal measure $\widehat{\mathbb{P}}$ are $\widehat{\mu}$ and $\widehat{\Sigma}$ respectively. This result is exhibited in the below lemma, whose proof is omitted because it is a straightforward extension of Theorem 3.13 to take into account the support information Ξ.

Lemma 3.50 (Gelbrich hull with support)**.** *If the nominal distribution $\widehat{\mathbb{P}}$ has mean vector $\widehat{\mu} \in \mathbb{R}^n$ and covariance matrix $\widehat{\Sigma} \in \mathbb{S}^n_+$, then we have $\mathbb{B}_{\rho,\sigma}(\Xi, \widehat{\mathbb{P}}) \subseteq \mathbb{G}_{\rho,\sigma}(\Xi, \widehat{\mu}, \widehat{\Sigma})$ for any $\sigma \in \{2, \mathrm{S}, \mathrm{SU}\}$.*

Furthermore, if we define the *Chebyshev ambiguity set with support information* that includes all probability measures supported on Ξ with mean vector $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{S}^n_+$ as

$$\mathscr{P}_\sigma(\Xi, \mu, \Sigma) \triangleq \mathscr{P}_\sigma(\Xi) \cap \mathscr{P}_\sigma(\mu, \Sigma),$$

then we can re-express the Gelbrich hull with support information $\mathbb{G}_{\rho,\sigma}(\Xi, \widehat{\mu}, \widehat{\Sigma})$ as the union of Chebyshev ambiguity sets with support information via

$$\mathbb{G}_{\rho,\sigma}(\Xi, \widehat{\mu}, \widehat{\Sigma}) = \bigcup_{(\mu,\Sigma) \in \mathscr{U}_\rho(\widehat{\mu}, \widehat{\Sigma})} \mathscr{P}_\sigma(\Xi, \mu, \Sigma).$$

Consequentially, the decomposition (3.8) remains valid with the additional support information, that is,

$$\begin{aligned}
\sup_{\mathbb{Q} \in \mathbb{G}_{\rho,\sigma}(\Xi, \widehat{\mu}, \widehat{\Sigma})} \mathscr{R}_\mathbb{Q}(\ell) &= \sup_{(\mu,\Sigma) \in \mathscr{U}_\rho(\widehat{\mu}, \widehat{\Sigma})} \; \sup_{\mathbb{Q} \in \mathscr{P}_\sigma(\Xi, \mu, \Sigma)} \mathscr{R}_\mathbb{Q}(\ell) && \text{(3.56a)} \\
&= \sup_{(\mu,M) \in \mathscr{V}_\rho(\widehat{\mu}, \widehat{\Sigma})} \; \sup_{\mathbb{Q} \in \mathscr{P}_\sigma(\Xi, \mu, M - \mu\mu^\top)} \mathscr{R}_\mathbb{Q}(\ell). && \text{(3.56b)}
\end{aligned}$$

The above decomposition is a fundamental building block for the reformulation of various Gelbrich risks under the Gelbrich hull with support information $\mathbb{G}_{\rho,\sigma}(\Xi,\widehat{\mu},\widehat{\Sigma})$. In the remainder of this section, we will focus on the case $\sigma=2$ and provide the reformulation and/or the approximation of the Gelbrich expected loss, as well as an extension to the approximation of joint linear chance constraints.

3.7.1 Gelbrich Expected Loss with Support Information

We provide in this section the reformulation of the Gelbrich expected loss with support information. For simplicity, we focus on the case where the support can be described as an intersection of (possibly non-convex) ellipsoids. The next proposition shows that the Gelbrich expected loss of a piecewise quadratic loss function can be conservatively approximated by the optimal value of a semidefinite optimization problem, and this approximation is exact when the support is an ellipsoid.

Proposition 3.51 (Approximation of Gelbrich expected loss)**.** *Suppose that the conditions of Theorem 3.41 hold, and assume furthermore that the support Ξ can be represented as the intersection of $I\geq 1$ quadratic constraints of the form*

$$\Xi=\left\{\xi\in\mathbb{R}^n:\xi^\top A_i\xi+2a_i^\top\xi+a_i^0\leq 0,\ A_i\in\mathbb{S}^n\quad\forall i\in[I]\right\},$$

and Ξ has a non-empty interior. If $\sigma=2$, $\widehat{\mu}\in\mathbb{R}^n$ and $\widehat{\Sigma}\in\mathbb{S}_+^n$, then for any $\rho\in\mathbb{R}_{++}$, the Gelbrich expected loss of a piecewise quadratic loss function admits the conservative approximation

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\Xi,\widehat{\mu},\widehat{\Sigma})}\mathbb{E}_{\mathbb{Q}}[\ell(\xi)]\leq\left\{\begin{array}{cl}\inf & y_0+\gamma\big(\rho^2-\|\widehat{\mu}\|^2-\mathrm{Tr}\big[\widehat{\Sigma}\big]\big)+z+\mathrm{Tr}\big[Z\big]\\ \mathrm{s.t.} & \gamma\in\mathbb{R}_+,\ \eta\in\mathbb{R}_+^{I\times J},\ y_0\in\mathbb{R},\ y\in\mathbb{R}^n,\ Y\in\mathbb{S}^n,\ z\in\mathbb{R}_+,\ Z\in\mathbb{S}_+^n\\ & \begin{bmatrix}\gamma I-Y & \gamma\widehat{\Sigma}^{\frac{1}{2}}\\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z\end{bmatrix}\succeq 0,\ \begin{bmatrix}\gamma I-Y & \gamma\widehat{\mu}+y\\ (\gamma\widehat{\mu}+y)^\top & z\end{bmatrix}\succeq 0\\ & \begin{bmatrix}Y-Q_j & y-q_j\\ y^\top-q_j^\top & y_0-q_j^0\end{bmatrix}+\sum_{i\in[I]}\eta_{ij}\begin{bmatrix}A_i & a_i\\ a_i^\top & a_i^0\end{bmatrix}\succeq 0\quad\forall j\in[J].\end{array}\right.$$

Furthermore, for $I=1$, the above inequality becomes an equality.

Proof. By applying the decomposition (3.56a), we find

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\Xi,\widehat{\mu},\widehat{\Sigma})}\mathbb{E}_{\mathbb{Q}}[\ell(\xi)]=\sup_{(\mu,\Sigma)\in\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})}\ \sup_{\mathbb{Q}\in\mathscr{P}_2(\Xi,\mu,\Sigma)}\mathbb{E}_{\mathbb{Q}}[\ell(\xi)].$$

We can appply [183, Lemma A.1] to construct the dual of the inner supremum problem as

$$\sup_{\mathbb{Q}\in\mathscr{P}_2(\Xi,\mu,\Sigma)}\mathbb{E}_{\mathbb{Q}}\left[\ell(\xi)\right]\leq\left\{\begin{array}{cl}\inf & y_0+2y^\top\mu+\mathrm{Tr}\left[Y(\Sigma+\mu\mu^\top)\right]\\ \mathrm{s.t.} & y_0\in\mathbb{R},\ y\in\mathbb{R}^n,\ Y\in\mathbb{S}^n\\ & y_0+2y^\top\xi+\xi^\top Y\xi\geq\ell(\xi)\qquad\forall\xi\in\Xi,\end{array}\right.$$

where the inequality is tight whenever $\Sigma \succ 0$ thanks to the strong duality result [86]. Let $\mathcal{Y}$ be the convex feasible set defined by

$$\mathcal{Y} = \left\{ y_0 \in \mathbb{R},\ y \in \mathbb{R}^n,\ Y \in \mathbb{S}^n : y_0 + 2y^\top \xi + \xi^\top Y \xi \geq \ell(\xi) \quad \forall \xi \in \Xi \right\}.$$

Because $\rho > 0$, we can follow closely the proof of Theorem 3.41 to show that

$$\begin{aligned}
\sup_{\mathbb{Q} \in \mathbb{G}_{\rho,2}(\Xi, \widehat{\mu}, \widehat{\Sigma})} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] &= \sup_{(\mu,\Sigma) \in \mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})} \ \sup_{\mathbb{Q} \in \mathcal{P}_2(\Xi, \mu, \Sigma)} \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] \\
&= \inf_{(y_0, y, Y) \in \mathcal{Y}} \left\{ y_0 + \delta^\star_{\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})}(2y, Y) \right\} \\
&= \left\{ \begin{array}{ll}
\inf & y_0 + \gamma\big(\rho^2 - \|\widehat{\mu}\|^2 - \operatorname{Tr}\big[\widehat{\Sigma}\big]\big) + z + \operatorname{Tr}\left[Z\right] \\
\text{s.t.} & \gamma \in \mathbb{R}_+,\ y_0 \in \mathbb{R},\ y \in \mathbb{R}^n,\ Y \in \mathbb{S}^n,\ z \in \mathbb{R}_+,\ Z \in \mathbb{S}^n_+ \\
& \begin{bmatrix} \gamma I - Y & \gamma \widehat{\Sigma}^{\frac{1}{2}} \\ \gamma \widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0,\ \begin{bmatrix} \gamma I - Y & \gamma \widehat{\mu} + y \\ (\gamma \widehat{\mu} + y)^\top & z \end{bmatrix} \succeq 0 \\
& (y_0, y, Y) \in \mathcal{Y}.
\end{array} \right.
\end{aligned}$$

The last part of the proof involves the reformulation of the feasible set $\mathcal{Y}$. Indeed, we can rewrite $\mathcal{Y}$ as

$$\mathcal{Y} = \left\{ y_0 \in \mathbb{R}, y \in \mathbb{R}^n, Y \in \mathbb{S}^n : y_0 + 2y^\top \xi + \xi^\top Y \xi \geq q_j^0 + q_j^\top \xi + \xi^\top Q_j \xi \quad \forall \xi \in \Xi,\ \forall j \in [J] \right\}.$$

The set $\mathcal{Y}$ can be outer approximated using the S-lemma [139] as

$$\mathcal{Y} \subseteq \left\{ y_0 \in \mathbb{R}, y \in \mathbb{R}^n, Y \in \mathbb{S}^n : \exists \eta \in \mathbb{R}_+^{I \times J} \text{ s.t. } \begin{bmatrix} Y - Q_j & y - q_j \\ y^\top - q_j^\top & y_0 - q_j^0 \end{bmatrix} + \sum_{i \in [I]} \eta_{ij} \begin{bmatrix} A_i & a_i \\ a_i^\top & a_i^0 \end{bmatrix} \succeq 0 \quad \forall j \in [J] \right\}.$$

The above approximation is tight if $I = 1$. This completes the proof. □

We remark that the support information can be injected into the reformulation of the Gelbrich expected loss for a quadratic loss function in Theorem 3.43 using an analogous reasoning, thus the details are omitted.

3.7.2 Conservative Approximation of Joint Chance Constraints

Suppose that we are interested in verifying whether the below joint chance constraint

$$\inf_{\mathbb{Q} \in \mathbb{G}_{\rho,2}(\Xi, \widehat{\mu}, \widehat{\Sigma})} \mathbb{Q}\left(q_j^0 + q_j^\top \xi \leq 0 \quad \forall j \in [J] \right) \geq 1 - \beta \tag{3.57}$$

is valid for some $q_j^0 \in \mathbb{R}$ and $q_j \in \mathbb{R}^n$ $\forall j \in [J]$ at a risk level $\beta \in (0,1)$. For a decision problem, q_j^0 and q_j are dependent on the decision variables. If $J = 1$ and $\Xi = \mathbb{R}^n$, we recover the single chance constraint problem without support information and the analytical expression of the Gelbrich VaR in Lemma 3.31 can be utilized to verify this chance constraint using Remark 3.38.

In this section, we consider the case when Ξ is a compact subset of $\mathbb{R}^n$ and $J \geq 2$.

Applying the same approach as in [183], we associate a scaling factor $\alpha_j \in \mathbb{R}_{++}$ to the j-th constraint in the joint chance constraint (3.57), and thus given a strictly positive vector $\alpha \in \mathbb{R}^J_{++}$, the joint chance constraint (3.57) is equivalent to an individual but *non*linear chance constraint

$$\inf_{\mathbb{Q} \in \mathbb{G}_{\rho,2}(\Xi,\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\left(\max_{j \in [J]} \left\{\alpha_j \left(q_j^0 + q_j^\top \xi\right)\right\} \leq 0\right) \geq 1 - \beta.$$

We resort to the CVaR conservative approximation of the above individual chance constraint that holds in the following sense: if

$$\sup_{\mathbb{Q} \in \mathbb{G}_{\rho,2}(\Xi,\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-}\mathrm{CVaR}_\beta\left(\max_{1 \leq j \leq J} \left\{\alpha_j \left(q_j^0 + q_j^\top \xi\right)\right\}\right) \leq 0 \tag{3.58}$$

is valid, then (3.57) hold. The CVaR approximation approach requires evaluating the CVaR of a polyhedral loss function over the Gelbrich hull with support information $\mathbb{G}_{\rho,2}(\Xi,\widehat{\mu},\widehat{\Sigma})$. Under minor additional assumptions on Ξ, the next proposition shows that this quantity can be safely approximated by the optimal value of a convex program.

Proposition 3.52. *Suppose that the compact support Ξ has non-empty interior and that Ξ can be represented as the intersection of $I \geq 1$ quadratic constraints of the form*

$$\Xi = \left\{\xi \in \mathbb{R}^n : \xi^\top A_i \xi + 2a_i^\top \xi + a_i^0 \leq 0,\ A_i \in \mathbb{S}^n \quad \forall i \in [I]\right\}.$$

If $\sigma = 2$, $\widehat{\mu} \in \mathbb{R}^n$ and $\widehat{\Sigma} \in \mathbb{S}^n_+$, then for any $\rho \in \mathbb{R}_{++}$ and $\alpha \in \mathbb{R}^J_{++}$, we have

$$\begin{aligned}
&\sup_{\mathbb{Q} \in \mathbb{G}_{\rho,2}(\Xi,\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-}\mathrm{CVaR}_\beta\left(\max_{j \in [J]} \left\{\alpha_j \left(q_j^0 + q_j^\top \xi\right)\right\}\right) \\
&\leq \left\{
\begin{array}{ll}
\inf & \tau + \beta^{-1}\big(y_0 + \gamma\big(\rho^2 - \|\widehat{\mu}\|^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\big) + z + \mathrm{Tr}\left[Z\right]\big) \\
\text{s.t.} & \tau \in \mathbb{R},\ \gamma \in \mathbb{R}_+,\ \eta \in \mathbb{R}_+^{I \times (J+1)},\ y_0 \in \mathbb{R},\ y \in \mathbb{R}^n,\ Y \in \mathbb{S}^n_+,\ z \in \mathbb{R}_+,\ Z \in \mathbb{S}^n_+ \\
& \begin{bmatrix} \gamma I - Y & \gamma \widehat{\Sigma}^{\frac{1}{2}} \\ \gamma \widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0,\ \begin{bmatrix} \gamma I - Y & \gamma \widehat{\mu} + y \\ (\gamma \widehat{\mu} + y)^\top & z \end{bmatrix} \succeq 0 \\
& \begin{bmatrix} Y & y - \frac{1}{2}\alpha_j q_j \\ (y - \frac{1}{2}\alpha_j q_j)^\top & y_0 - \alpha_j q_j^0 + \tau \end{bmatrix} + \sum_{i \in [I]} \eta_{ij} \begin{bmatrix} A_i & a_i \\ a_i^\top & a_i^0 \end{bmatrix} \succeq 0 \quad \forall j \in [J] \\
& \begin{bmatrix} Y & y \\ y^\top & y_0 \end{bmatrix} + \sum_{i \in [I]} \eta_{i0} \begin{bmatrix} A_i & a_i \\ a_i^\top & a_i^0 \end{bmatrix} \succeq 0.
\end{array}
\right.
\end{aligned}$$

Furthermore, if $I = 1$ then the above reformulation is tight.

Proof. For any $\alpha \in \mathbb{R}^J_{++}$, we find

$$
\begin{aligned}
&\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\Xi,\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}\text{-}\operatorname{CVaR}_\beta\left(\max_{j\in[J]}\left\{\alpha_j\left(q_j^0+q_j^\top\xi\right)\right\}\right)\\
&= \sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\Xi,\widehat{\mu},\widehat{\Sigma})} \inf_{\tau\in\mathbb{R}}\left\{\tau+\frac{1}{\beta}\mathbb{E}_{\mathbb{Q}}\left(\left[\max_{j\in[J]}\left\{\alpha_j\left(q_j^0+q_j^\top\xi\right)\right\}-\tau\right]^+\right)\right\} && \text{(3.59a)}\\
&= \inf_{\tau\in\mathbb{R}}\left\{\tau+\frac{1}{\beta}\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\Xi,\widehat{\mu},\widehat{\Sigma})}\mathbb{E}_{\mathbb{Q}}\left(\left[\max_{j\in[J]}\left\{\alpha_j\left(q_j^0+q_j^\top\xi\right)\right\}-\tau\right]^+\right)\right\}, && \text{(3.59b)}
\end{aligned}
$$

where equality (3.59a) utilizes the definition of the CVaR in [144]. Because Ξ is compact, $\mathbb{G}_{\rho,2}(\Xi,\widehat{\mu},\widehat{\Sigma})$ is weakly compact, and the interchange of the sup-inf operators in equality (3.59b) is justified by a version of the stochastic saddle point result [155, Proposition 3.1]. Let $\ell(\xi)$ be a loss function inside the expectation operator in (3.59b). More specifically, we have

$$
\ell(\xi)=\left[\max_{j\in[J]}\left\{\alpha_j\left(q_j^0+q_j^\top\xi\right)\right\}-\tau\right]^+=\max\left\{\max_{j\in[J]}\{\alpha_j(q_j^0+q_j^\top\xi)-\tau\},0\right\},
$$

which is equivalent to a pointwise maximum of $J+1$ quadratic loss functions of the form

$$
\ell_0(\xi)=0,\qquad \ell_j(\xi)=\alpha_j q_j^0-\tau+\alpha_j q_j^\top\xi\qquad \forall j\in[J].
$$

The Gelbrich expected loss in (3.59b) can be approximated using Proposition 3.51 as

$$
\begin{aligned}
&\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\Xi,\widehat{\mu},\widehat{\Sigma})}\mathbb{E}_{\mathbb{Q}}\left(\left[\max_{j\in[J]}\left\{\alpha_j\left(q_j^0+q_j^\top\xi\right)\right\}-\tau\right]^+\right)\\
&\le\left\{
\begin{array}{ll}
\inf & y_0+\gamma\left(\rho^2-\|\widehat{\mu}\|^2-\operatorname{Tr}[\widehat{\Sigma}]\right)+z+\operatorname{Tr}[Z]\\
\text{s.t.} & \gamma\in\mathbb{R}_+,\ \eta\in\mathbb{R}_+^{I\times J},\ y_0\in\mathbb{R},\ y\in\mathbb{R}^n,\ Y\in\mathbb{S}^n,\ z\in\mathbb{R}_+,\ Z\in\mathbb{S}_+^n\\
& \begin{bmatrix}\gamma I-Y & \gamma\widehat{\Sigma}^{\frac12}\\ \gamma\widehat{\Sigma}^{\frac12} & Z\end{bmatrix}\succeq 0,\ \begin{bmatrix}\gamma I-Y & \gamma\widehat{\mu}+y\\ (\gamma\widehat{\mu}+y)^\top & z\end{bmatrix}\succeq 0\\
& \begin{bmatrix}Y & y-\frac12\alpha_j q_j\\ (y-\frac12\alpha_j q_j)^\top & y_0-\alpha_j q_j^0+\tau\end{bmatrix}+\sum_{i\in[I]}\eta_{ij}\begin{bmatrix}A_i & a_i\\ a_i^\top & a_i^0\end{bmatrix}\succeq 0\quad\forall j\in[J]\\
& \begin{bmatrix}Y & y\\ y^\top & y_0\end{bmatrix}+\sum_{i\in[I]}\eta_{i0}\begin{bmatrix}A_i & a_i\\ a_i^\top & a_i^0\end{bmatrix}\succeq 0,
\end{array}
\right.
\end{aligned}
$$

and the above approximation is tight if $I=1$. The proof is completed by replacing the above approximation into (3.59b). □

Proposition (3.52) derives the CVaR approximation of the joint chance constraint for a given scaling vector $\alpha\in\mathbb{R}^J_{++}$. When the joint chance constraint is utilized as an elementary constraint of a decision problem, the parameters q_j^0 and q_j are dependent on the decision variables. In this case, one can further tighten the CVaR approximation using a sequential optimization procedure similar to [183, Algorithm 3.1] that optimizes sequentially over the scaling factor

$\alpha \in \mathbb{R}^J_{++}$ and the primitive decision variables.

3.8 Numerical Experiment: Index Tracking Portfolio Optimization

We consider the passive portfolio allocation strategy where the goal is to construct a portfolio of $n-1$ stocks that tracks the return of a pre-specified market index. Let $r \in \mathbb{R}^{n-1}$ be a random vector representing the stochastic return of $n-1$ stocks and $r_{\text{market}} \in \mathbb{R}$ is the random return of the market index that the portfolio manager aims to track. Define the random vector $\xi = [r^\top, r_{\text{market}}]^\top$, the nominal index tracking portfolio allocation problem can be formulated as

$$\inf_{w \in \mathcal{W}} \quad \mathbb{E}_{\widehat{\mathbb{P}}}\left[\ell(w^\top \xi)\right],$$

where the expectation is taken over the nominal joint probability measure $\widehat{\mathbb{P}}$ of ξ, and $\ell : \mathbb{R} \to \mathbb{R}_+$ is a loss function that penalizes the mismatch between the portfolio return and the market index. The feasible set $\mathcal{W}$ is confined to

$$\mathcal{W} = \left\{ w \in \mathbb{R}^n : \sum_{i \in [n-1]} w_i = 1,\ w_n = -1,\ w_i \geq 0 \quad \forall i = 1, \dots, n-1 \right\},$$

where w_i is the i-th element of the vector w. For simplicity, we assume that the support of the random vector ξ is $\Xi = \mathbb{R}^n$, and we emphasize that the injection of the support information is straightforward from the results of Section 3.7.

If we denote by $\widehat{\mu}$ the mean vector and by $\widehat{\Sigma}$ the covariance matrix of ξ under $\widehat{\mathbb{P}}$, the corresponding distributionally robust index tracking portfolio allocation problem under the Gelbrich hull $\mathbb{G}_{\rho,2}(\widehat{\mu}, \widehat{\Sigma})$ can be formulated as

$$\inf_{w \in \mathcal{W}} \quad \sup_{\mathbb{Q} \in \mathbb{G}_{\rho,2}(\widehat{\mu}, \widehat{\Sigma})} \quad \mathbb{E}_{\mathbb{Q}}\left[\ell(\xi^\top w)\right]. \tag{3.60}$$

Interestingly, index tracking problem (3.60) can be considered as a special instance of a broader class of distributionally robust regression problems where ℓ represents a (convex) regression loss function.

Corollary 3.53 (Gelbrich regression problems). *Suppose that ℓ is a convex loss function. The optimal value of the distributionally robust regression problem* (3.60) *equals the optimal value of the following optimization problem*

$$\begin{array}{cl} \inf & y_0 + \gamma\left(\rho^2 - \|\widehat{\mu}\|^2 - \operatorname{Tr}\left[\widehat{\Sigma}\right]\right) + z + \operatorname{Tr}\left[Z\right] \\ \text{s.t.} & w \in \mathcal{W},\ \gamma \in \mathbb{R}_+,\ y_0 \in \mathbb{R},\ y \in \mathbb{R}^n,\ Y \in \mathbb{S}^n,\ z \in \mathbb{R}_+,\ Z \in \mathbb{S}^n_+ \\ & \begin{bmatrix} \gamma I - Y & \gamma \widehat{\Sigma}^{\frac{1}{2}} \\ \gamma \widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0,\ \begin{bmatrix} \gamma I - Y & \gamma \widehat{\mu} + y \\ (\gamma \widehat{\mu} + y)^\top & z \end{bmatrix} \succeq 0 \\ & (y_0, y, Y) \in \mathcal{Y}(w), \end{array} \tag{3.61}$$

where $\mathcal{Y}(w)$ *is a parametrized feasible set*

$$\mathcal{Y}(w) = \left\{ y_0 \in \mathbb{R}, y \in \mathbb{R}^n, Y \in \mathbb{S}^n : y_0 + 2\xi^\top y + \xi^\top Y \xi \geq \ell(w^\top \xi) \quad \forall \xi \in \mathbb{R}^n \right\}.$$

For common regression loss functions listed below, $\mathcal{Y}(w)$ *is representable as semidefinite constraints, and problem* (3.61) *is a semidefinite program.*

(i) ***Robust regression.*** *If* ℓ *is the Huber loss function with target value* $\alpha \in \mathbb{R}$ *and robustness parameter* $\beta > 0$*, that is,*

$$\ell(w^\top \xi) = \begin{cases} \frac{1}{2}\left(w^\top \xi - \alpha\right)^2 & \text{if } |w^\top \xi - \alpha| \leq \beta, \\ \beta\left(|w^\top \xi - \alpha| - \frac{1}{2}\beta\right) & \text{otherwise,} \end{cases}$$

then the feasible set $\mathcal{Y}(w)$ *can be expressed as*

$$\mathcal{Y}(w) = \left\{ y_0 \in \mathbb{R}, y \in \mathbb{R}^n, Y \in \mathbb{S}^n_+ : \begin{array}{c} \exists \theta_1 \in \mathbb{R}, \theta_2 \in \mathbb{R}_+, \\ \begin{bmatrix} Y & y - \frac{\beta}{2} w \\ y^\top - \frac{\beta}{2} w^\top & y_0 - \theta_2 + \beta(\alpha + \theta_1) \end{bmatrix} \geq 0, \quad \begin{bmatrix} \theta_2 & \theta_1 \\ \theta_1 & 2 \end{bmatrix} \geq 0 \\ \begin{bmatrix} Y & y + \frac{\beta}{2} w \\ y^\top + \frac{\beta}{2} w^\top & y_0 - \theta_2 - \beta(\alpha + \theta_1) \end{bmatrix} \geq 0 \end{array} \right\}.$$

(ii) ***Support vector regression.*** *If* ℓ *is the* ϵ*-insensitive loss function with target* $\alpha \in \mathbb{R}$*, that is,* $\ell(w^\top \xi) = \max\left\{0, |w^\top \xi - \alpha| - \epsilon\right\}$*, then the feasible set* $\mathcal{Y}(w)$ *can be expressed as*

$$\mathcal{Y}(w) = \left\{ y_0 \in \mathbb{R}, y \in \mathbb{R}^n, Y \in \mathbb{S}^n_+ : \begin{bmatrix} Y & y - \frac{1}{2} w \\ y^\top - \frac{1}{2} w^\top & y_0 + \alpha + \epsilon \end{bmatrix} \geq 0, \quad \begin{bmatrix} Y & y + \frac{1}{2} w \\ y^\top + \frac{1}{2} w^\top & y_0 - \alpha + \epsilon \end{bmatrix} \geq 0 \right\}.$$

(iii) ***Quantile regression.*** *If* ℓ *is the pinball loss function with target* $\alpha \in \mathbb{R}$ *and parameter* $\beta \in [0,1]$*, that is,* $\ell(w^\top \xi) = \max\left\{-\beta(w^\top \xi - \alpha), (1-\beta)(w^\top \xi - \alpha)\right\}$*, then the feasible set* $\mathcal{Y}(w)$ *can be expressed as*

$$\mathcal{Y}(w) = \left\{ y_0 \in \mathbb{R}, y \in \mathbb{R}^n, Y \in \mathbb{S}^n_+ : \begin{bmatrix} Y & y - \frac{\beta}{2} w \\ y^\top - \frac{\beta}{2} w^\top & y_0 - \beta\alpha \end{bmatrix} \geq 0, \quad \begin{bmatrix} Y & y + \frac{(1-\beta)}{2} w \\ y^\top + \frac{(1-\beta)}{2} w^\top & y_0 + (1-\beta)\alpha \end{bmatrix} \geq 0 \right\}.$$

(iv) $\|\cdot\|_1$ ***loss.*** *If* ℓ *is a 1-norm loss function with target* $\alpha \in \mathbb{R}$*, that is,* $\ell(w^\top \xi) = \|w^\top \xi - \alpha\|_1$*, then the feasible set* $\mathcal{Y}(w)$ *can be expressed as*

$$\mathcal{Y}(w) = \left\{ y_0 \in \mathbb{R}, y \in \mathbb{R}^n, Y \in \mathbb{S}^n_+ : \begin{bmatrix} Y & y - \frac{1}{2} w \\ y^\top - \frac{1}{2} w^\top & y_0 + \alpha \end{bmatrix} \geq 0, \quad \begin{bmatrix} Y & y + \frac{1}{2} w \\ y^\top + \frac{1}{2} w^\top & y_0 - \alpha \end{bmatrix} \geq 0 \right\}.$$

(v) $\|\cdot\|_2^2$ ***loss.*** *If* ℓ *is a squared 2-norm loss function with target* $\alpha \in \mathbb{R}$*, that is,* $\ell(w^\top \xi) =$

$\| w^\top \xi - \alpha \|_2^2$, *then the feasible set* $\mathscr{Y}_x$ *can be expressed as*

$$\mathscr{Y}(w) = \left\{ y_0 \in \mathbb{R}, y \in \mathbb{R}^n, Y \in \mathbb{S}^n_+ : \exists M \in \mathbb{S}^n_+, \quad \begin{bmatrix} M & y + \alpha w \\ y^\top + \alpha w^\top & y_0 - \alpha^2 \end{bmatrix} \succeq 0, \quad \begin{bmatrix} Y - M & w \\ w^\top & 1 \end{bmatrix} \succeq 0 \right\}.$$

Proof. The reformulation (3.61) follows directly from the proof of Theorem 3.41 by noting that because ℓ is convex, $\mathscr{Y}(w)$ is also a convex feasible set. We now proceed to provide the reformulation of $\mathscr{Y}(w)$ for different loss functions.

Consider when ℓ is a Huber loss function. In this case, we can write ℓ using the inf-convolution formulation as

$$\begin{aligned} \ell(w^\top \xi) &= \inf_{\theta_1 \in \mathbb{R}} \frac{1}{2}\theta_1^2 + \beta |w^\top \xi - \alpha - \theta_1| \\ &= \inf_{\theta_1 \in \mathbb{R}} \max\left\{ \frac{1}{2}\theta_1^2 + \beta(w^\top \xi - \alpha - \theta_1),\ \frac{1}{2}\theta_1^2 + \beta(-w^\top \xi + \alpha + \theta_1) \right\}. \end{aligned}$$

The semi-infinite constraint defining the feasible set $\mathscr{Y}(w)$ can be reformulated as

$$\exists \theta_1 \in \mathbb{R} : \left\{ \begin{array}{ll} y_0 - \frac{1}{2}\theta_1^2 + \beta(\alpha + \theta_1) + (2y - \beta x)^\top \xi + \xi^\top Y \xi \geq 0 & \forall \xi \in \mathbb{R}^n \\ y_0 - \frac{1}{2}\theta_1^2 - \beta(\alpha + \theta_1) + (2y + \beta x)^\top \xi + \xi^\top Y \xi \geq 0 & \forall \xi \in \mathbb{R}^n, \end{array} \right.$$

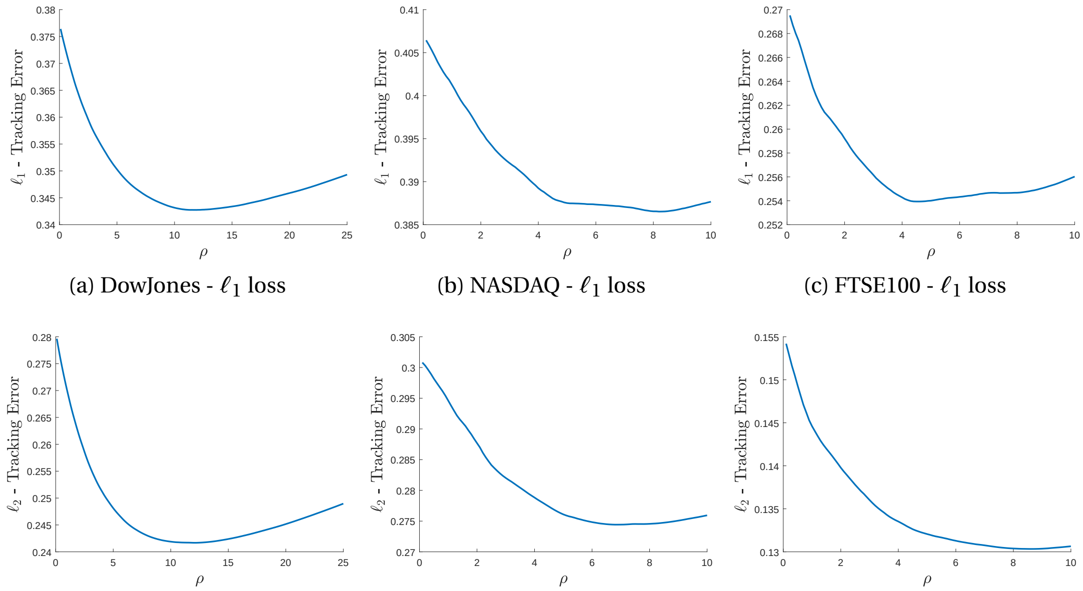
and hence $\mathscr{Y}(w)$ admits a conic representation

$$\mathscr{Y}(w) = \left\{ y_0 \in \mathbb{R}, y \in \mathbb{R}^n, Y \in \mathbb{S}^n_+ : \begin{array}{c} \exists \theta_1 \in \mathbb{R}, \\ \begin{bmatrix} Y & y - \frac{\beta}{2} w \\ y^\top - \frac{\beta}{2} w^\top & y_0 - \frac{1}{2}\theta_1^2 + \beta(\alpha + \theta_1) \end{bmatrix} \succeq 0, \\ \begin{bmatrix} Y & y + \frac{\beta}{2} w \\ y^\top + \frac{\beta}{2} w^\top & y_0 - \frac{1}{2}\theta_1^2 - \beta(\alpha + \theta_1) \end{bmatrix} \succeq 0, \end{array} \right\}.$$

which is nonlinear because of the quadratic terms in θ_1. In the last step, we replace the term $\frac{1}{2}\theta_1^2$ by an auxiliary variable $\theta_2 \in \mathbb{R}_+$ with an additional constraint $\theta_2 \geq \frac{1}{2}\theta_1^2$. Formulating this additional constraint as a semidefinite constraint completes the proof for the Huber loss.

The other loss functions can be trivially re-expressed as a pointwise maximium of quadratic functions, and the reformulation of $\mathscr{Y}(w)$ follows an analogous reasoning. The detailed proof is thus omitted. □

Because the portfolio manager wants to match the portfolio return as close as possible to the market return, the target parameter can be set to $\alpha = 0$. Furthermore, we will focus on the ℓ_1 and ℓ_2 loss to avoid the extra hyper-parameter of the loss function to be tuned. We use the standard dataset for index-tracking portfolio balancing [19] which includes the weekly returns for DowJones, NASDAQ100, FTSE100, the summary data about the dataset is shown in Table 3.1. We use a similar rolling horizon approach for portfolio balancing as in [19]: we use 52 weeks of return to estimate the sample average joint mean vector and



(a) DowJones - ℓ_1 loss

(b) NASDAQ - ℓ_1 loss

(c) FTSE100 - ℓ_1 loss

(d) DowJones - ℓ_2 loss

(e) NASDAQ - ℓ_2 loss

(f) FTSE100 - ℓ_2 loss

Figure 3.1 – Tracking error using ℓ_1 loss (panels 3.1a–3.1b) and ℓ_2 loss (panels 3.1d–3.1e)

covariance matrix of the assets and index returns. These estimates are used to construct the nominal distribution $\widehat{\mathbb{P}}$ for each level of ambiguity $\rho \in \{0.1 \times t, t = 0, \ldots, 250\}$ for the DowJones dataset and $\rho \in \{0.1 \times t, t = 0, \ldots, 100\}$ for the NASDAQ100 and FTSE100 dataset to find the optimal portfolio allocation that minimizes the worst-case index tracking expected error. The portfolio allocation is kept constant for 12 weeks to compute the weekly portfolio return and the corresponding mismatch between the portfolio return and the specified index.

Table 3.1 – Datasets for the experiments

Dataset Name	# of assets (n)	# of weeks (T)	Time interval	# of rebalancing
DowJones	28	1363	Feb 1990-Apr 2016	110
NASDAQ100	82	596	Nov 2004-Apr 2016	46
FTSE100	83	717	Jul 2002-Apr 2016	56

All experiments are run on an Intel XEON CPU with 3.40GHz clock speed and 16GB of RAM. All semidefinite programs are solved with MOSEK 8.1 using the YALMIP interface [112]. In order to ensure that our experiments are reproducible, the underlying source codes are accessible at https://github.com/nvietanh/GelbrichRM. Figure 3.1 shows the additional benefit of embracing the distributionally robust index tracking model (3.60) with the Gelbrich hull ambiguity set. For both the ℓ_1 and ℓ_2 loss function and throughout all three datasets, the optimal radius that minimizes the worst-case expectation of the misalignment between the portfolio return and the index target are all strictly positive.

3.9 Appendix

To derive certain results in this paper, we need to solve the nonlinear semidefinite program of the form

$$\sup_{\Sigma\succeq 0} \operatorname{Tr}\left[(Q-\gamma I)\Sigma\right] + 2\gamma \operatorname{Tr}\left[\left(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right].$$

The following result asserts that this optimization problem can be solved in quasi-closed form.

Proposition 3.54 ([124, Proposition A.2])**.** *For any $Q\in\mathbb{S}^n$, $\widehat{\Sigma}\in\mathbb{S}^n_+$ and $\gamma\in\mathbb{R}_+$ we have*

$$\sup_{\Sigma\succeq 0}\left\{\operatorname{Tr}\left[(Q-\gamma I)\Sigma\right] + 2\gamma \operatorname{Tr}\left[\left(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]\right\} = \begin{cases} \gamma^2 \operatorname{Tr}\left[(\gamma I-Q)^{-1}\widehat{\Sigma}\right] & \text{if } \gamma>\lambda_{\max}(Q),\\ \liminf\limits_{\bar\gamma\downarrow\gamma} \bar\gamma^2 \operatorname{Tr}\left[(\bar\gamma I-Q)^{-1}\widehat{\Sigma}\right] & \text{if } \gamma=\lambda_{\max}(Q),\\ +\infty & \text{if } \gamma<\lambda_{\max}(Q). \end{cases}$$

Moreover, the maximization problem is solved by $\Sigma^\star = \gamma^2(\gamma I-Q)^{-1}\widehat{\Sigma}(\gamma I-Q)^{-1}$ whenever $\gamma>\lambda_{\max}(Q)$. This solution is unique if $\widehat{\Sigma}\succ 0$.

Proving the results in Sections 3.6.2 and 3.6.3 requires the following lemma which is an extension of [21, Lemma 1].

Lemma 3.55. *Let $\mathscr{S}\subseteq\mathbb{R}^n$ be a measurable set (not necessarily convex). For any $\widehat{\mu}\in\mathbb{R}^n$, $\widehat{\Sigma}\in\mathbb{S}^n_+$ and $\rho\in\mathbb{R}_{++}$, let $\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})$ be the Gelbrich hull defined as in Definition 3.12. We have*

$$\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}(\xi\in\mathscr{S}) = \left\{\begin{array}{cl} \inf & y_0+\gamma\left(\rho^2-\|\widehat{\mu}\|^2-\operatorname{Tr}\left[\widehat{\Sigma}\right]\right)+z+\operatorname{Tr}\left[Z\right]\\ \text{s.t.} & \gamma\in\mathbb{R}_+,\ y_0\in\mathbb{R},\ y\in\mathbb{R}^n,\ Y\in\mathbb{S}^n,\ z\in\mathbb{R}_+,\ Z\in\mathbb{S}^n_+\\ & \begin{bmatrix}\gamma I-Y & \gamma\widehat{\Sigma}^{\frac{1}{2}}\\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z\end{bmatrix}\succeq 0,\ \begin{bmatrix}\gamma I-Y & \gamma\widehat{\mu}+y\\ (\gamma\widehat{\mu}+y)^\top & z\end{bmatrix}\succeq 0\\ & \begin{bmatrix}Y & y\\ y^\top & y_0\end{bmatrix}\succeq 0,\ y_0+2y^\top\xi+\xi^\top Y\xi\geq 1\quad \forall\xi\in\mathscr{S}.\end{array}\right.$$

Proof. Let $\mathbb{1}_{\mathscr{S}}(\xi)$ be the indicator function of the set $\mathscr{S}$, that is,

$$\mathbb{1}_{\mathscr{S}}(\xi) = \begin{cases} 1 & \text{if } \xi\in\mathscr{S},\\ 0 & \text{otherwise.}\end{cases}$$

Given a set $\mathscr{S}$, define momentarily the feasible set $\mathscr{Y}$ by

$$\begin{aligned}\mathscr{Y} &= \left\{y_0\in\mathbb{R},\ y\in\mathbb{R}^n,\ Y\in\mathbb{S}^n : y_0+2y^\top\xi+\xi^\top Y\xi\geq \mathbb{1}_{\mathscr{S}}(\xi)\quad \forall\xi\in\mathbb{R}^n\right\}\\ &= \left\{y_0\in\mathbb{R},\ y\in\mathbb{R}^n,\ Y\in\mathbb{S}^n : \begin{bmatrix}Y & y\\ y^\top & y_0\end{bmatrix}\succeq 0,\ y_0+2y^\top\xi+\xi^\top Y\xi\geq 1\quad \forall\xi\in\mathscr{S}\right\}.\end{aligned}$$

Notice that $\mathscr{Y}$ is a closed and convex set. We have

$$\begin{aligned}
\sup_{\mathbb{Q}\in\mathbb{G}_{\rho,2}(\widehat{\mu},\widehat{\Sigma})} \mathbb{Q}(\xi\in\mathscr{S}) &= \sup_{(\mu,\Sigma)\in\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \sup_{\mathbb{Q}\in\mathscr{P}_2(\mu,\Sigma)} \mathbb{E}_\mathbb{Q}[\mathbb{1}_\mathscr{S}(\xi)] && \text{(3.62a)}\\
&\leq \sup_{(\mu,\Sigma)\in\mathscr{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \inf_{(y_0,y,Y)\in\mathscr{Y}} y_0 + 2\mu^\top y + \mathrm{Tr}\left[(\Sigma+\mu\mu^\top)Y\right] && \text{(3.62b)}\\
&= \sup_{(\mu,M)\in\mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})} \inf_{(y_0,y,Y)\in\mathscr{Y}} y_0 + 2\mu^\top y + \mathrm{Tr}\left[MY\right]\\
&= \inf_{(y_0,y,Y)\in\mathscr{Y}} \sup_{(\mu,M)\in\mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})} y_0 + 2\mu^\top y + \mathrm{Tr}\left[MY\right] && \text{(3.62c)}\\
&= \inf_{(y_0,y,Y)\in\mathscr{Y}} y_0 + \delta^\star_{\mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})}(2y,Y),
\end{aligned}$$

where equality (3.62a) is from the two layer decomposition (3.8a) of the Gelbrich hull, and inequality (3.62b) is from the Isii's duality result [86]. Equality (3.62c) follows from the Sion's minimax theorem [156] which holds because $\mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})$ is compact by virtue of Proposition 3.17. The semidefinite program reformulation in the statement of the Lemma is obtained by utilizing the reformulation of the support function of $\mathscr{V}_\rho(\widehat{\mu},\widehat{\Sigma})$ in Lemma 3.22. In the last step, inequality (3.62b) can be proven to hold as an equality using an analogous argument as in the second part of the proof of Theorem 3.41. This completes the proof. □

Conclusions

We can only see a short distance ahead, but we can see plenty there that needs to be done.
— Alan Turing

This thesis contains three independent chapters which lay the foundations for distributionally robust optimization using Wasserstein type-2 ambiguity set along with various applications in statistical optimization, machine learning and risk assessment.

Chapter 1 studies the robustified maximum likelihood estimator for the inverse covariance matrix of a Gaussian random vectors. Using an ambiguity set that contains only Gaussian distributions, we show that the optimal estimator exists in closed-form and it belongs to the wider class of nonlinear shrinkage estimators. We demonstrate that the Wasserstein shrinkage estimator possesses many nice properties, all of which are not imposed adhoc but arise naturally from the framework of distributionally robust optimization. We develop a quadratic approximation numerical algorithm to solve the robustified estimation problem if additional constraints regarding the conditional independency among the components of the random vector are involved.

Chapter 2 studies the robustified estimator which aims to recover the true signal from noisy measurements with minimum mean square error. If *only* the center of the Wasserstein ambiguity set is supposed to be elliptical, we show that the optimal Wasserstein MMSE estimator is an affine function of the noisy measurements, and it can be recovered from the optimal solution of a semidefinite program if the nominal distribution of the noise is non-degenerate. We further develop a decomposable first-order Frank-Wolfe algorithm that converges linearly to the optimal solution of the semidefinite program.

Chapter 3 studies the Gelbrich hull, a superset of the Wasserstein ambiguity set, and the Gelbrich risk, a safe approximation of the Wasserstein risk. We show that the Gelbrich risk admits a two-layer decomposition that facilitates the reformulation of the Gelbrich risk. We prove that for a linear loss function, the Gelbrich risk of a family of consistent, positive homogeneous and translation invariance risk measures can be expressed in closed form. We also show that evaluating the Gelbrich expected loss is, under some regularity conditions, equivalent to solving a finite convex optimization problem.

The promising horizons for future research are highlighted below.

Wasserstein Statistical Estimation

This thesis examines the distributionaly robustification of two fundamental estimation problems in statistics, namely the maximum likelihood estimation problem and the minimum mean square error estimation problem. There is an abundance of other statistical estimation/inference problem where a coherent robustification can be applied in the same spirit of Chapter 1 and 2.

Parametric Family of Distributions

The backbone of this thesis is built upon the Gaussian distribution and its generalization to the family of elliptical distributions. A natural extension is to study other family of parametric distributions, a potential candidate is the family of exponential distributions of which the normal distribution is also an element. Family of discrete distributions, whose applications are ubiquitous, is another interesting subject for distributional robustification.

Approximation Schemes for Type-p Wasserstein Ambiguity Set

This thesis concentrates on Wasserstein type-2 distance and exploits its lower approximation using the Gelbrich distance. It is left unanswered whether a similar approximation can be obtained for the Wasserstein type-1 distance, and whether there can be a tight approximation for type-p Wasserstein distance with $p > 2$. The discovery of any lower bound of this type will lead to novel approximations of the Wasserstein risk and potentially enjoy widespread applications.

Applications in Various Problems

Apart from theoretical studies, the applications of the technical results in this thesis in operations management, healthcare and power systems is a large unexplored field. At the same time, there are a plethora of important adversarial learning tasks where the Wasserstein type-2 ambiguity set is left untouched. Developing application-tailored model using the results presented in this thesis, especially in Chapter 3, is an exciting future work.

Bibliography

[1] C. Acerbi, *Spectral measures of risk: A coherent representation of subjective risk aversion*, Journal of Banking & Finance, 26 (2002), pp. 1505–1518.

[2] J. Aitchison and J. Brown, *The Lognormal Distribution with Special Reference to Its Use in Economics*, Cambridge University Press, 1957.

[3] C. D. Aliprantis and K. C. Border, *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Springer, 2006.

[4] D. Amir, *Chebyshev centers and uniform convexity*, Pacific Journal of Mathematics, 77 (1978), pp. 1–6.

[5] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein generative adversarial networks*, in International Conference on Machine Learning, 2017, pp. 214–223.

[6] O. Banerjee, L. El Ghaoui, and A. d'Aspremont, *Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data*, Journal of Machine Learning Research, 9 (2008), pp. 485–516.

[7] A. Beck, A. Ben-Tal, and Y. C. Eldar, *Robust mean-squared error estimation of multiple signals in linear systems affected by model and noise uncertainties*, Mathematical Programming, 107 (2006), pp. 155–187.

[8] A. Beck and Y. C. Eldar, *Regularization in regression with bounded noise: A Chebyshev center approach*, SIAM Journal on Matrix Analysis and Applications, 29 (2007), pp. 606–625.

[9] A. Beck, Y. C. Eldar, and A. Ben-Tal, *Mean-squared error estimation for linear systems with block circulant uncertainty*, SIAM Journal on Matrix Analysis and Applications, 29 (2007), pp. 712–730.

[10] A. Ben-Tal, D. D. Hertog, and J. P. Vial, *Deriving robust counterparts of nonlinear uncertain inequalities*, Mathematical Programming, 149 (2015), pp. 265–299.

[11] C. Berge, *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity*, Courier Corporation, 1963.

[12] D. S. BERNSTEIN, *Matrix Mathematics: Theory, Facts, and Formulas*, Princeton University Press, 2009.

[13] D. P. BERTSEKAS, *Convex Optimization Theory*, Athena Scientific, 2009.

[14] R. BHATIA, T. JAIN, AND Y. LIM, *Strong convexity of sandwiched entropies and related optimization problems*, Reviews in Mathematical Physics, (2018).

[15] J. BIEN AND R. J. TIBSHIRANI, *Sparse estimation of a covariance matrix*, Biometrika, 98 (2011), pp. 807–820.

[16] J. BLANCHET AND K. MURTHY, *Quantifying distributional model risk via optimal transport*, Mathematics of Operations Research, 44 (2019), pp. 565–600.

[17] J. BLANCHET AND N. SI, *Optimal uncertainty size in distributionally robust inverse covariance estimation*, arXiv preprint arXiv:1901.07693, (2019).

[18] S. BOYD AND L. VANDENBERGHE, *Convex Optimization*, Cambridge University Press, 2004.

[19] R. BRUNI, F. CESARONE, A. SCOZZARI, AND F. TARDELLA, *Real-world datasets for portfolio selection and solutions of some stochastic dominance portfolio models*, Data in Brief, 8 (2016), pp. 858–862.

[20] G. CALAFIORE, *Ambiguous risk measures and optimal robust portfolios*, SIAM Journal on Optimization, 18 (2007), pp. 853–877.

[21] G. CALAFIORE, U. TOPCU, AND L. E. GHAOUI, *Parameter estimation with expected and residual-at-risk criteria*, Systems & Control Letters, 58 (2009), pp. 39–46.

[22] S. CAMBANIS, S. HUANG, AND G. SIMONS, *On the theory of elliptically contoured distributions*, Journal of Multivariate Analysis, 11 (1981), pp. 368–385.

[23] S. G. CHANG, B. YU, AND M. VETTERLI, *Adaptive wavelet thresholding for image denoising and compression*, IEEE Transactions on Image Processing, 9 (2000), pp. 1532–1546.

[24] S. G. CHANG, B. YU, AND M. VETTERLI, *Spatially adaptive wavelet thresholding with context modeling for image denoising*, IEEE Transactions on Image Processing, 9 (2000), pp. 1522–1531.

[25] C. CHATFIELD, *The Analysis of Time Series: An Introduction*, CRC Press, 2016.

[26] L. CHEN, S. HE, AND S. ZHANG, *Tight bounds for some risk measures, with applications to robust portfolio selection*, Operations Research, 59 (2011), pp. 847–865.

[27] V. CHOPRA AND W. ZIEMBA, *The effect of errors in means, variances, and covariances on optimal portfolio choice*, in The Kelly Capital Growth Investment Criterion: Theory and Practice, World Scientific Publishing, 2011, ch. 18, pp. 249–257.

[28] S. Y. CHUN, M. W. BROWNE, AND A. SHAPIRO, *Modified distribution-free goodness-of-fit test statistic*, Psychometrika, 83 (2018), pp. 48–66.

[29] K. L. CLARKSON, *Coresets, sparse greedy approximation, and the frank-wolfe algorithm*, ACM Transactions on Algorithms, 6 (2010), p. 63.

[30] D. L. COHN, *Measure Theory*, Springer, 2013.

[31] R. CONT, R. DEGUEST, AND G. SCANDOLO, *Robustness and sensitivity analysis of risk measurement procedures*, Quantitative Finance, 10 (2010), pp. 593–606.

[32] T. COVER AND J. THOMAS, *Elements of Information Theory*, John Wiley & Sons, 2012.

[33] M. CUTURI, *Sinkhorn distances: Lightspeed computation of optimal transport*, in Advances in Neural Information Processing Systems 26, 2013, pp. 2292–2300.

[34] J. DAHL, V. ROYCHOWDHURY, AND L. VANDENBERGHE, *Maximum likelihood estimation of Gaussian graphical models: Numerical implementation and topology selection.* UCLA Preprint, 2005.

[35] L. A. DALTON AND E. R. DOUGHERTY, *Bayesian minimum mean-square error estimation for classification error–Part I: Definition and the Bayesian MMSE error estimator for discrete classification*, IEEE Transactions on Signal Processing, 59 (2011), pp. 115–129.

[36] A. DAS, A. L. SAMPSON, C. LAINSCSEK, L. MULLER, W. LIN, J. C. DOYLE, S. S. CASH, E. HALGREN, AND T. J. SEJNOWSKI, *Interpretation of the precision matrix and its application in estimating sparse brain connectivity during sleep spindles from human electrocorticography recordings*, Neural Computation, 29 (2017), pp. 603–642.

[37] T. H. DE MELLO AND G. BAYRAKSAN, *Monte Carlo sampling-based methods for stochastic optimization*, Surveys in Operations Research and Management Science, 19 (2014), pp. 56–85.

[38] E. DELAGE, D. KUHN, AND W. WIESEMANN, *“Dice"-sion making under uncertainty: When can a random decision reduce risk?*, To appear in Management Science, (2018).

[39] E. DELAGE AND Y. YE, *Distributionally robust optimization under moment uncertainty with application to data-driven problems*, Operations Research, 58 (2010), pp. 595–612.

[40] V. DEMIGUEL AND F. J. NOGALES, *Portfolio selection with robust estimation*, Operations Research, 57 (2009), pp. 560–577.

[41] V. F. DEMYANOV AND A. M. RUBINOV, *Approximate Methods in Optimization Problems*, American Elsevier Publishing, 1970.

[42] M. DETTLING, *Bagboosting for tumor classification with gene expression data*, Bioinformatics, 20 (2004), pp. 3583–3593.

[43] D. K. DEY AND C. SRINIVASAN, *Estimation of a covariance matrix under Stein's loss*, Annals of Statistics, 13 (1985), pp. 1581–1591.

[44] S. N. DIGGAVI AND T. M. COVER, *The worst additive noise under a covariance constraint*, IEEE Transactions on Information Theory, 47 (2001), pp. 3072–3081.

[45] D. L. DONOHO AND I. M. JOHNSTONE, *Adapting to unknown smoothness via wavelet shrinkage*, Journal of the American Statistical Association, 90 (1995), pp. 1200–1224.

[46] D. L. DONOHO AND I. M. JOHNSTONE, *Minimax estimation via wavelet shrinkage*, Annals of Statistics, 26 (1998), pp. 879–921.

[47] D. L. DONOHO AND J. M. JOHNSTONE, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika, 81 (1994), pp. 425–455.

[48] L. DU, J. LI, AND P. STOICA, *Fully automatic computation of diagonal loading levels for robust adaptive beamforming*, IEEE Transactions on Aerospace and Electronic Systems, 46 (2010), pp. 449–458.

[49] J. C. DUNN, *Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals*, SIAM Journal on Control and Optimization, 17 (1979), pp. 187–211.

[50] ———, *Convergence rates for conditional gradient sequences generated by implicit step length rules*, SIAM Journal on Control and Optimization, 18 (1980), pp. 473–487.

[51] J. C. DUNN AND S. HARSHBARGER, *Conditional gradient algorithms with open loop step size rules*, Journal of Mathematical Analysis and Applications, 62 (1978), pp. 432–444.

[52] L. EL GHAOUI, M. OKS, AND F. OUSTRY, *Worst-case value-at-risk and robust portfolio optimization: A conic programming approach*, Operations Research, 51 (2003), pp. 543–556.

[53] Y. C. ELDAR, *Robust competitive estimation with signal and noise covariance uncertainties*, IEEE Transactions on Information Theory, 52 (2006), pp. 4532–4547.

[54] Y. C. ELDAR, A. BECK, AND M. TEBOULLE, *A minimax Chebyshev estimator for bounded error estimation*, IEEE Transactions on Signal Processing, 56 (2008), pp. 1388–1397.

[55] Y. C. ELDAR, A. BEN-TAL, AND A. NEMIROVSKI, *Linear minimax regret estimation of deterministic parameters with bounded data uncertainties*, IEEE Transactions on Signal Processing, 52 (2004), pp. 2177–2188.

[56] ———, *Robust mean-squared error estimation in the presence of model uncertainties*, IEEE Transactions on Signal Processing, 53 (2004), pp. 168–181.

[57] Y. C. ELDAR AND N. MERHAV, *A competitive minimax approach to robust estimation of random parameters*, IEEE Transactions on Signal Processing, 52 (2004), pp. 1931–1946.

[58] J. FAN, Y. FAN, AND J. LV, *High dimensional covariance matrix estimation using a factor model*, Journal of Econometrics, 147 (2008), pp. 186–197.

[59] K. FANG, S. KOTZ, AND K. NG, *Symmetric Multivariate and Related Distributions*, Chapman & Hall, 1990.

[60] R. A. FISHER, *The use of multiple measurements in taxonomic problems*, Annals of Eugenics, 7 (1936), pp. 179–188.

[61] H. FÖLLMER AND A. SCHIED, *Convex and coherent risk measures*, Working paper, (2008).

[62] H. FÖLLMER AND A. SCHIED, *Stochastic Finance. An Introduction in Discrete Time*, de Gruyter, 2008.

[63] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Naval Research Logistics, 3 (1956), pp. 95–110.

[64] R. M. FREUND AND P. GRIGAS, *New analysis and results for the Frank–Wolfe method*, Mathematical Programming, 155 (2016), pp. 199–230.

[65] J. FRIEDMAN, T. HASTIE, AND R. TIBSHIRANI, *Sparse inverse covariance estimation with the graphical lasso*, Biostatistics, 9 (2008), pp. 432–441.

[66] J. GALAMBOS AND I. SIMONELLI, *Products of Random Variables: Applications to Problems of Physics and to Arithmetical Functions*, Taylor & Francis, 2004.

[67] R. GAO, X. CHEN, AND A. KLEYWEGT, *Wasserstein distributional robustness and regularization in statistical learning*, arXiv preprint arXiv:1712.06050v2, (2016).

[68] R. GAO AND A. J. KLEYWEGT, *Distributionally robust stochastic optimization with Wasserstein distance*, arXiv preprint arXiv:1604.02199, (2016).

[69] R. GAO, L. XIE, Y. XIE, AND H. XU, *Robust hypothesis testing using Wasserstein uncertainty sets*, in Advances in Neural Information Processing Systems 31, 2018, pp. 7913–7923.

[70] D. GARBER AND E. HAZAN, *Faster rates for the Frank-Wolfe method over strongly-convex sets*, in International Conference on International Conference on Machine Learning, 2015, pp. 541–549.

[71] M. GELBRICH, *On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces*, Mathematische Nachrichten, 147 (1990), pp. 185–203.

[72] C. GIVENS AND R. SHORTT, *A class of Wasserstein metrics for probability distributions*, The Michigan Mathematical Journal, 31 (1984), pp. 231–240.

[73] J. GOH AND M. SIM, *Distributionally robust optimization and its tractable approximations*, Operations Research, 58 (2010), pp. 902–917.

[74] F. GOLNARAGHI AND B. KUO, *Automatic Control Systems*, McGraw-Hill Education, 2017.

[75] S. GOTO AND Y. XU, *Improving mean variance optimization through sparse hedging restrictions*, Journal of Financial and Quantitative Analysis, 50 (2015), p. 1415–1441.

[76] I. GULRAJANI, F. AHMED, M. ARJOVSKY, V. DUMOULIN, AND A. C. COURVILLE, *Improved training of Wasserstein GANs*, in Advances in Neural Information Processing Systems 30, 2017, pp. 5767–5777.

[77] D. GUO, Y. WU, S. S. SHITZ, AND S. VERDU, *Estimation in Gaussian noise: Properties of the minimum mean-square error*, IEEE Transactions on Information Theory, 57 (2011), pp. 2371–2385.

[78] S. GUO AND H. XU, *Distributionally robust shortfall risk optimization model and its approximation*, Mathematical Programming, 174 (2019), pp. 473–498.

[79] L. R. HAFF, *The variational form of certain Bayes estimators*, The Annals of Statistics, 19 (1991), pp. 1163–1190.

[80] J. HAMILTON, *Time Series Analysis*, Princeton University Press, 1994.

[81] G. HANASUSANTO, D. KUHN, S. W. WALLACE, AND S. ZYMLER, *Distributionally robust multi-item newsvendor problems with multimodal demand distributions*, Mathematical Programming, 152 (2015), pp. 1–32.

[82] T. HASTIE, R. TIBSHIRANI, AND J. FRIEDMAN, *The Elements of Statistical Learning*, Springer, 2001.

[83] J. P. HESPANHA, *Linear Systems Theory*, Princeton Press, 2009.

[84] C.-J. HSIEH, M. A. SUSTIK, I. S. DHILLON, AND P. RAVIKUMAR, *QUIC: Quadratic approximation for sparse inverse covariance estimation*, Journal of Machine Learning Research, 15 (2014), pp. 2911–2947.

[85] H. HULT AND F. LINDSKOG, *Multivariate extremes, aggregation and dependence in elliptical distributions*, Advances in Applied Probability, 34 (2002), pp. 587–608.

[86] K. ISII, *The extrema of probability determined by generalized moments (i) bounded random variables*, Annals of the Institute of Statistical Mathematics, 12 (1960), pp. 119–134.

[87] R. JAGANNATHAN AND T. MA, *Risk reduction in large portfolios: Why imposing the wrong constraints helps*, The Journal of Finance, 58 (2003), pp. 1651–1683.

[88] M. JAGGI, *Sparse Convex Optimization Methods for Machine Learning*, PhD thesis, ETH Zurich, 2011.

[89] M. JAGGI, *Revisiting Frank-Wolfe: Projection-free sparse convex optimization*, in Proceedings of the 30th International Conference on Machine Learning, 2013, pp. 427–435.

[90] W. James and C. Stein, *Estimation with quadratic loss*, in Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, Berkeley, CA, 1961, University of California Press, pp. 361–379.

[91] S. R. Jaschke, *The Cornish-Fisher-Expansion in the context of Delta - Gamma - Normal approximations*, Journal of Risk, 4 (2002), pp. 33–52.

[92] E. Jondeau, S. Poon, and M. Rockinger, *Financial Modeling Under Non-Gaussian Distributions*, Springer, 2007.

[93] P. Jorion, *Value at Risk: The New Benchmark for Managing Financial Risk*, McGraw-Hill, 2006.

[94] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, *Generalized power method for sparse principal component analysis*, Journal of Machine Learning Research, 11 (2010), pp. 517–553.

[95] A. Juditsky and A. Nemirovski, *Lectures on statistical inferences via convex optimization*, 2018.

[96] A. Juditsky and A. Nemirovski, *Near-optimality of linear recovery in gaussian observation scheme under $\|\cdot\|_2^2$-loss*, Annals of Statistics, 46 (2018), pp. 1603–1629.

[97] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, 1993.

[98] D. Kelker, *Distribution theory of spherical distributions and a location-scale parameter generalization*, Sankhyā: The Indian Journal of Statistics, Series A, (1970), pp. 419–430.

[99] D. Kuhn and D. G. Luenberger, *Analysis of the rebalancing frequency in log-optimal portfolio selection*, Quantitative Finance, 10 (2010), pp. 221–234.

[100] D. Kuhn, P. Mohajerin Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, *Wasserstein distributionally robust optimization: Theory and applications in machine learning*, To appear in INFORMS TutORials in Operations Research, (2019).

[101] S. Kusuoka, *On law invariant coherent risk measures*, in Advances in Mathematical Economics, S. Kusuoka and T. Maruyama, eds., Springer, 2001, pp. 83–95.

[102] S. L. Lauritzen, *Graphical Models*, Oxford University Press, 1996.

[103] O. Ledoit and M. Wolf, *Improved estimation of the covariance matrix of stock returns with an application to portfolio selection*, Journal of Empirical Finance, 10 (2003), pp. 603–621.

[104] ——, *Honey, I shrunk the sample covariance matrix*, The Journal of Portfolio Management, 30 (2004), pp. 110–119.

[105] ——, *A well-conditioned estimator for large-dimensional covariance matrices*, Journal of Multivariate Analysis, 88 (2004), pp. 365–411.

[106] ——, *Nonlinear shrinkage estimation of large-dimensional covariance matrices*, The Annals of Statistics, 40 (2012), pp. 1024–1060.

[107] E. S. LEVITIN AND B. T. POLYAK, *Constrained minimization methods*, USSR Computational Mathematics and Mathematical Physics, 6 (1966), pp. 1–50.

[108] B. C. LEVY AND R. NIKOUKHAH, *Robust least-squares estimation with a relative entropy constraint*, IEEE Transactions on Information Theory, 50 (2004), pp. 89–104.

[109] ——, *Robust state space filtering under incremental model perturbations subject to a relative entropy tolerance*, IEEE Transactions on Automatic Control, 58 (2012), pp. 682–695.

[110] J. Y.-M. LI, *Technical note - Closed-form solutions for worst-case law invariant risk measures with application to robust portfolio optimization*, Operations Research, 66 (2018), pp. 1533–1541.

[111] M. S. LOBO, L. VANDENBERGHE, S. BOYD, AND H. LEBRET, *Applications of second-order cone programming*, Linear Algebra and its Applications, 284 (1998), pp. 193 – 228.

[112] J. LÖFBERG, *YALMIP: A toolbox for modeling and optimization in MATLAB*, in 2004 IEEE International Conference on Robotics and Automation, 2004, pp. 284–289.

[113] D. MACKAY, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.

[114] S. MALLAT, *A Wavelet Tour of Signal Processing*, Academic Press, 1999.

[115] H. MARKOWITZ, *Portfolio selection*, The Journal of Finance, 7 (1952), pp. 77–91.

[116] R. MICHAUD, *The Markowitz optimization enigma: Is 'optimized' optimal?*, Financial Analysts Journal, 45 (1989), pp. 31–42.

[117] M. K. MIHCAK, I. KOZINTSEV, K. RAMCHANDRAN, AND P. MOULIN, *Low-complexity image denoising based on statistical modeling of wavelet coefficients*, IEEE Signal Processing Letters, 6 (1999), pp. 300–303.

[118] P. MOHAJERIN ESFAHANI AND D. KUHN, *Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations*, Mathematical Programming, 171 (2018), pp. 115–166.

[119] P. MOHAJERIN ESFAHANI, S. SHAFIEEZADEH-ABADEH, G. A. HANASUSANTO, AND D. KUHN, *Data-driven inverse optimization with imperfect information*, Mathematical Programming, (2017).

[120] K. MURPHY, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.

[121] K. Natarajan, M. Sim, and J. Uichanco, *Tractable robust expected utility and risk models for portfolio optimisation*, Mathematical Finance, 20 (2010), pp. 695–731.

[122] Y. Nesterov, *Complexity bounds for primal-dual methods minimizing the model of objective function*, Mathematical Programming, 171 (2018), pp. 311–330.

[123] V. Nguyen, D. Kuhn, and P. Mohajerin Esfahani, *Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator*, arXiv preprint arXiv:1805.07194, (2018).

[124] V. Nguyen, S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani, *Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization*, Working paper, (2019).

[125] L. Ning, T. T. Georgiou, A. Tannenbaum, and S. P. Boyd, *Linear models based on noisy data and the Frisch scheme*, SIAM Review, 57 (2015), pp. 167–197.

[126] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, 2006.

[127] K. Ogata, *Modern Control Engineering*, Pearson, 2009.

[128] I. Olkin and F. Pukelsheim, *The distance between two random vectors with given dispersion matrices*, Linear Algebra and its Applications, 48 (1982), pp. 257–263.

[129] A. Oppenheim and G. Verghese, *Signals, Systems and Inference*, Pearson, 2015.

[130] F. Oztoprak, J. Nocedal, S. Rennie, and P. A. Olsen, *Newton-like methods for sparse inverse covariance estimation*, in Advances in Neural Information Processing Systems, 2012, pp. 755–763.

[131] V. Y. Pan and Z. Q. Chen, *The complexity of the matrix eigenproblem*, in Proceedings of the Thirty-first Annual ACM Symposium on Theory of Computing, 1999, pp. 507–516.

[132] B. V. Parys, P. Mohajerin Esfahani, and D. Kuhn, *From data to decisions: Distributionally robust optimization is optimal*, arXiv preprint arXiv:1704.04118, (2017).

[133] F. Pedregosa, A. Askari, G. Negiar, and M. Jaggi, *Step-size adaptivity in projection-free optimization*, arXiv preprint arXiv:1806.05123, (2018).

[134] M. D. Perlman, *STAT 542: Multivariate statistical analysis.* University of Washington, 2007. Lecture Notes.

[135] G. Peyré and M. Cuturi, *Computational optimal transport*, Foundations and Trends® in Machine Learning, 11 (2019), pp. 355–607.

[136] G. Pflug and A. Pichler, *Multistage Stochastic Optimization*, Springer, 2014.

[137] G. Pflug and D. Wozabal, *Ambiguity in portfolio selection*, Quantitative Finance, 7 (2007), pp. 435–442.

[138] G. C. Pflug, A. Pichler, and D. Wozabal, *The $1/N$ investment strategy is optimal under high model ambiguity*, Journal of Banking & Finance, 36 (2012), pp. 410–417.

[139] I. Pólik and T. Terlaky, *A survey of the S-lemma*, SIAM Review, 49 (2007), pp. 371–418.

[140] I. Popescu, *Robust mean-covariance solutions for stochastic optimization*, Operations Research, 55 (2007), pp. 98–112.

[141] A. Posekany, K. Felsenstein, and P. Sykacek, *Biological assessment of robust noise models in microarray data analysis*, Bioinformatics, 27 (2011), pp. 807–814.

[142] A. Ribes, J.-M. Azaïs, and S. Planton, *Adaptation of the optimal fingerprint method for climate change detection using a well-conditioned covariance matrix estimate*, Climate Dynamics, 33 (2009), pp. 707–722.

[143] T. Rippl, A. Munk, and A. Sturm, *Limit laws of the empirical Wasserstein distance: Gaussian distributions*, Journal of Multivariate Analysis, 151 (2016), pp. 90–109.

[144] R. T. Rockafellar and S. Uryasev, *Optimization of Conditional Value-at-Risk*, Journal of Risk, 2 (2000), pp. 21–41.

[145] N. Rujeerapaiboon, D. Kuhn, and W. Wiesemann, *Robust growth-optimal portfolios*, Management Science, 62 (2015), pp. 2090–2109.

[146] N. Rujeerapaiboon, D. Kuhn, and W. Wiesemann, *Chebyshev inequalities for products of random variables*, Mathematics of Operations Research, 43 (2018), pp. 887–918.

[147] U. E. Ruttimann, M. Unser, R. R. Rawlings, D. Rio, N. F. Ramsey, V. S. Mattay, D. W. Hommer, J. A. Frank, and D. R. Weinberger, *Statistical analysis of functional mri data in the wavelet domain*, IEEE Transactions on Medical Imaging, 17 (1998), pp. 142–154.

[148] J. Schäfer and K. Strimmer, *A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics*, Statistical Applications in Genetics and Molecular Biology, 4 (2005). Article 32.

[149] L. Sendur and I. W. Selesnick, *Bivariate shrinkage with local variance estimation*, IEEE Signal Processing Letters, 9 (2002), pp. 438–441.

[150] S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani, *Regularization via mass transportation*, To appear in Journal of Machine Learning Research, (2019).

[151] S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani, and D. Kuhn, *Distributionally robust logistic regression*, in Advances in Neural Information Processing Systems 28, 2015, pp. 1576–1584.

[152] S. Shafieezadeh-Abadeh, V. Nguyen, D. Kuhn, and P. Mohajerin Esfahani, *Wasserstein distributionally robust Kalman filtering*, in Advances in Neural Information Processing Systems 31, 2018, pp. 8483–8492.

[153] A. Shapiro, *Monte Carlo sampling methods*, in Stochastic Programming, vol. 10 of Handbooks in Operations Research and Management Science, Elsevier, 2003, pp. 353–425.

[154] A. Shapiro, *On Kusuoka representation of law invariant risk measures*, Mathematics of Operations Research, 38 (2013), pp. 142–152.

[155] A. Shapiro and A. Kleywegt, *Minimax analysis of stochastic problems*, Optimization Methods and Software, 17 (2002), pp. 523–542.

[156] M. Sion, *On general minimax theorems*, Pacific Journal of Mathematics, 8 (1958), pp. 171–176.

[157] J. Smith and R. Winkler, *The optimizer's curse: Skepticism and postdecision surprise in decision analysis*, Management Science, 52 (2006), pp. 311–322.

[158] J. Solomon, F. D. Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas, *Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains*, ACM Transactions on Graphics, 34 (2015), p. 66.

[159] C. Stein, *Estimation of a covariance matrix*, Rietz Lecture, 39th Annual Meeting IMS, Atlanta, GA, (1975).

[160] ——, *Lectures on the theory of estimation of many parameters*, Journal of Soviet Mathematics, 34 (1986), pp. 1373–1403.

[161] G. V. G. Stevens, *On the inverse of the covariance matrix in portfolio analysis*, The Journal of Finance, 53 (1998), pp. 1821–1827.

[162] J. Stock and M. Watson, *Introduction to Econometrics*, Prentice Hall, 2015.

[163] B. Taşkesen, S. Shafieezadeh-Abadeh, and D. Kuhn, *On the complexity of computing Wasserstein distances*, Working paper, (2019).

[164] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, *Wasserstein auto-encoders*, in International Conference on Learning Representations, 2018.

[165] G. Torri, R. Giacometti, and S. Paterlini, *Sparse precision matrices for minimum variance portfolios*, Computational Management Science, (2019). Forthcoming.

[166] A. Touloumis, *Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings*, Computational Statistics & Data Analysis, 83 (2015), pp. 251–261.

[167] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Mathematical Programming, 117 (2009), pp. 387–423.

[168] R. H. Tütüncü, K. C. Toh, and M. J. Todd, *Solving semidefinite-quadratic-linear programs using SDPT3*, Mathematical Programming, 95 (2003), pp. 189–217.

[169] H. R. van der Vaart, *On certain characteristics of the distribution of the latent roots of a symmetric random matrix under general conditions*, The Annals of Mathematical Statistics, 32 (1961), pp. 864–873.

[170] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Review, 38 (1996), pp. 49–95.

[171] A. Wiesel, Y. Eldar, and A. Hero, *Covariance estimation in decomposable Gaussian graphical models*, IEEE Transactions on Signal Processing, 58 (2010), pp. 1482–1492.

[172] W. Wiesemann, D. Kuhn, and M. Sim, *Distributionally robust convex optimization*, Operations Research, 62 (2014), pp. 1358–1376.

[173] J.-H. Won, J. Lim, S.-J. Kim, and B. Rajaratnam, *Condition number regularized covariance estimation*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75 (2013), pp. 427–450.

[174] J. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, MIT Press, 2010.

[175] D. Wozabal, *A framework for optimization under ambiguity*, Annals of Operations Research, 193 (2012), pp. 21–47.

[176] R. Yang and J. O. Berger, *Estimation of a covariance matrix using the reference prior*, The Annals of Statistics, 22 (1994), pp. 1195–1211.

[177] Y.-l. Yu, Y. Li, D. Schuurmans, and C. Szepesvári, *A general projection property for distribution families*, in Advances in Neural Information Processing Systems 22, 2009, pp. 2232–2240.

[178] C. Zhao and Y. Guan, *Data-driven risk-averse stochastic optimization with Wasserstein metric*, Operations Research Letters, 46 (2018), pp. 262 – 267.

[179] J. Zhen, D. Kuhn, and W. Wiesemann, *Distributionally robust nonlinear optimization*, Working paper, (2019).

[180] S. Zhu and M. Fukushima, *Worst-case conditional value-at-risk with application to robust portfolio management*, Operations Research, 57 (2009), pp. 1155–1168.

[181] M. Zorzi, *Robust Kalman filtering under model perturbations*, IEEE Transactions on Automatic Control, 62 (2016), pp. 2902–2907.

[182] ———, *On the robustness of the Bayes and Wiener estimators under model uncertainty*, Automatica, 83 (2017), pp. 133–140.

[183] S. Zymler, D. Kuhn, and B. Rustem, *Distributionally robust joint chance constraints with second-order moment information*, Mathematical Programming, 137 (2013), pp. 167–198.

[184] S. Zymler, D. Kuhn, and B. Rustem, *Worst-case value at risk of nonlinear portfolios*, Management Science, 59 (2013), pp. 172–188.

Curriculum Vitae

Viet Anh Nguyen
vietanh.nguyen226@gmail.com

Education

Risk Analytics and Optimization Chair, École Polytechnique Fédérale de Lausanne
Doctor of Philosophy in Risk Analytics and Optimization (2013 – 2019, Switzerland)
Thesis: "Adversarial Analytics"

Study Center Gerzensee
Swiss Program for Beginning Doctoral Students in Economics (2013 – 2014, Switzerland)

Department of Industrial and Systems Engineering, National University of Singapore
Master of Engineering (2011 – 2013, Singapore)
Thesis: "Routing and Planning for the Last Mile Mobility System"

École Centrale de Paris
Diplôme d'Ingénieur (2008 – 2011, France)

Department of Industrial and Systems Engineering, National University of Singapore
Bachelor of Engineering (2006 – 2011, Singapore)

Selected Honors

First place, George Nicholson Student Paper Competition
INFORMS 2018

Best Teaching Assistant Award
École Polytechnique Fédérale de Lausanne (2018)

Teaching Assistantship
National University of Singapore (2010-2013)

Eiffel Excellence Scholarship
French Ministry of Education (2008-2010)

ASEAN Undergraduate Scholarship
Singaporean Ministry of Education (2006-2010).

Publications

Distributionally Robust Risk Measures with Structured Ambiguity Sets (with Damir Filipović, Soroosh Shafieezadeh Abadeh and Daniel Kuhn). Working paper.

Bridging Bayesian and Minimax Mean Square Error Estimation via Wasserstein Distributionally Robust Optimization (with Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani and Daniel Kuhn). Working paper.

Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning (with Daniel Kuhn, Peyman Mohajerin Esfahani and Soroosh Shafieezadeh-Abadeh). INFORMS TutORials in Operations Research, 2019.

Wasserstein Distributionally Robust Kalman Filtering (with Soroosh Shafieezadeh-Abadeh, Daniel Kuhn and Peyman Mohajerin Esfahani). NeurIPS, 2018.

Distributionally Robust Inverse Covariance Estimation: The Wasserstein Shrinkage Estimator (with Daniel Kuhn and Peyman Mohajerin Esfahani). Minor revision at Operations Research - Resubmitted.

Energy and Reserve Dispatch with Distributionally Robust Joint Chance Constraints (with Christos Ordoudis, Daniel Kuhn and Pierre Pinson). Working paper.

A Linear-Quadratic Gaussian Approach to Dynamic Information Acquisition (with Thomas Weber). European Journal of Operational Research, 2018.

Satisficing Measure Approach for Vehicle Routing Problem with Time Windows under Uncertainty (with Jun Jiang, Kien Ming Ng and Kwong Meng Teo). European Journal of Operational Research, 2016.

Commuter Cycling Policy in Singapore: a Farecard Data Analytics based Approach (with Ashwani Kumar and Kwong Meng Teo). Annals of Operations Research, 2016.

Invited Presentations

INFORMS Annual Meeting, Seattle, United States (2019).

International Conference on Continuous Optimization (ICCOPT), Berlin, Germany (2019).

International Conference on Stochastic Programming (ICSP), Trondheim, Norway (2019).

INFORMS Annual Meeting, Arizona, United States (2018).

International Symposium on Mathematical Programming (ISMP), Bordeaux, France (2018).

Conference on Computational Management Science (CMS), Trondheim, Norway (2018).

European Conference on Operational Research (EURO), Glasgow, UK (2015).

Teaching Experience

Risk Analytics and Optimization Chair, École Polytechnique Fédérale de Lausanne
Doctoral Assistant (2013 – 2019, Switzerland)
List of courses: Optimal Decision Making (Master), Applied Probability and Stochastic Processes (Master), Microeconomics (Master), Operations: Economics and Strategy (Master), Network Analytics (Master).

Department of Industrial and Systems Engineering, National University of Singapore
Graduate Teaching Assistant (2011, Singapore)
List of courses: Operations Research (Bachelor), Probability Models and Applications (Bachelor).

Subject to change are all conditioned things.
Strive on with heedfulness.
— Gautama Buddha

Ce document a été imprimé au Centre d’impression EPFL,
imprimerie climatiquement neutre, certifiée myClimate.