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

Nguyen, Viet Anh; Shafieezadeh-Abadeh, Soroosh; Kuhn, Daniel; Esfahani, Peyman Mohajerin

DOI
10.1287/moor.2021.1176

Publication date
2023

Document Version
Final published version

Published in
Mathematics of Operations Research

Citation (APA)
Nguyen, V. A., Shafieezadeh-Abadeh, S., Kuhn, D., & Esfahani, P. M. (2023). Bridging Bayesian and Minimax Mean Square Error Estimation via Wasserstein Distributionally Robust Optimization. *Mathematics of Operations Research*, *48*(1), 1-37. https://doi.org/10.1287/moor.2021.1176

Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

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

Viet Anh Nguyen, Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, Peyman Mohajerin Esfahani

To cite this article:
Viet Anh Nguyen, Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, Peyman Mohajerin Esfahani (2023) Bridging Bayesian and Minimax Mean Square Error Estimation via Wasserstein Distributionally Robust Optimization. Mathematics of Operations Research 48(1):1-37. https://doi.org/10.1287/moor.2021.1176

https://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 48, No. 1, February 2023, pp. 1–37
ISSN 0364-765X (print), ISSN 1526-5471 (online)

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

Viet Anh Nguyen,[a] Soroosh Shafieezadeh-Abadeh,[b] Daniel Kuhn,[c] Peyman Mohajerin Esfahani[d]

[a] Department of Management Science and Engineering, Stanford University, Stanford, California 94305; [b] Tepper School of Business, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213; [c] Risk Analytics and Optimization Chair, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; [d] Delft Center for Systems and Control, Delft University of Technology, 2628 CD Delft, Netherlands

Contact: viet-anh.nguyen@stanford.edu, https://orcid.org/0000-0002-9607-7891 (VAN); soroosh.shafiee@gmail.com, https://orcid.org/0000-0001-9095-2686 (SS-A); daniel.kuhn@epfl.ch (DK); p.mohajerinesfahani@tudelft.nl (PME)

Received: November 8, 2019
Revised: September 21, 2020
Accepted: January 26, 2021
Published Online in Articles in Advance: December 17, 2021

MSC2000 Subject Classification: Primary: 90-10, 90B99

https://doi.org/10.1287/moor.2021.1176

Abstract. We introduce a distributionally robust minimium mean square error estimation model with a Wasserstein ambiguity set to recover an unknown signal from a noisy observation. The proposed model can be viewed as a zero-sum game between a statistician choosing an estimator—that is, a measurable function of the observation—and a fictitious adversary choosing a prior—that is, a pair of signal and noise distributions ranging over independent Wasserstein balls—with the goal to minimize and maximize the expected squared estimation error, respectively. We show that, if the Wasserstein balls are centered at normal distributions, then the zero-sum game admits a Nash equilibrium, by which the players' optimal strategies are given by an affine estimator and a normal prior, respectively. We further prove that this Nash equilibrium can be computed by solving a tractable convex program. Finally, we develop a Frank–Wolfe algorithm that can solve this convex program orders of magnitude faster than state-of-the-art general-purpose solvers. We show that this algorithm enjoys a linear convergence rate and that its direction-finding subproblems can be solved in quasi-closed form.

Funding: This research was supported by the Swiss National Science Foundation [Grants BSCGI0_157733 and 51NF40_180545], an Early Postdoc.Mobility Fellowship [Grant P2ELP2_195149], and the European Research Council [Grant TRUST-949796].

Keywords: distributionally robust optimization • minimum mean square error estimation • Wasserstein distance • affine estimator

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

$$y = Hx + w, \tag{1.1}$$

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 that w 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. We emphasize that none of the subsequent results rely on a particular ordering of the dimension n of the parameter x and the dimension m of the measurement y. The linear measurement Model (1.1) is fundamental for numerous applications in engineering (e.g., linear systems theory, Golnaraghi and Kuo [34], Ogata [56]), econometrics (e.g., linear regression, Stock and Watson [71], Wooldridge [74]; time series analysis, Chatfield [13], Hamilton [36]), machine learning and signal processing (e.g., Kalman filtering, Kay [44], Murphy [53], Oppenheim and Verghese [58]), or information theory (e.g., multiple-input, multiple-output systems, Cover and Thomas [15], MacKay [51]), etc. In addition, Model (1.1) also emerges naturally in many applications in operations research, such as traffic management and control (van Lint and Djukic [73]), inventory control (Aviv [2]), advertising and promotion budgeting (Sriram and Kalwani [70]), or resource management (Rubel and Naik [63]).

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 $\mathcal{F}$. The quality of an estimator is measured by a risk function $R : \mathcal{F} \times \mathbb{R}^n \to \mathbb{R}$, which quantifies the mismatch between the parameter x and its prediction $\psi(y)$. A popular risk function is the mean square error (MSE):

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

which defines the estimation error as the expected squared Euclidean distance between $\psi(y)$ and x. If x is known, then $R(\psi, x)$ can be minimized directly, and the constant estimator $\psi^\star(y) \equiv x$ is 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 the statistician's beliefs about x before seeing y (Kay [44, section 11.4]) and solves the minimum MSE (MMSE) estimation problem

$$\underset{\psi \in \mathcal{F}}{\text{minimize}} \;\; \mathbb{E}_{\mathbb{P}_x}\big[R(\psi, x)\big]. \tag{1.2}$$

If the distribution $\mathbb{P}_x$ of x has finite second moments, then (1.2) is solvable. In this case, the optimal estimator, which is usually termed the Bayesian MMSE estimator, is of the form $\psi^\star_{\mathcal{B}}(y) = \mathbb{E}_{\mathbb{P}_{x|y}}[x]$, 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, $\psi^\star_{\mathcal{B}}$ is highly sensitive to the prior distribution $\mathbb{P}_x$, which is troubling if the statistician has little confidence in the statistician's beliefs. Second, computing $\psi^\star_{\mathcal{B}}$ requires precise knowledge of the noise distribution $\mathbb{P}_w$, which is typically unobservable and, thus, uncertain at least to some extent. Moreover, $\psi^\star_{\mathcal{B}}$ may generically have a complicated functional form, and evaluating $\psi^\star_{\mathcal{B}}(y)$ 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 $\mathcal{F}$ of all measurable estimators in (1.2) to the space

$$\mathcal{A} = \{\psi \in \mathcal{F} : \exists A \in \mathbb{R}^{n\times m},\, b \in \mathbb{R}^n \text{ with } \psi(y) = Ay + b \;\; \forall y \in \mathbb{R}^m\} \tag{1.3}$$

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 $\psi^\star_{\mathcal{A}}(y) = A^\star y + b^\star$, 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 (1.1) together with the independence of x and w, one can show that

$$A^\star = \Sigma_x H^\top (H\Sigma_x H^\top + \Sigma_w)^{-1} \quad \text{and} \quad b^\star = \mu_x - A^\star(H\mu_x + \mu_w). \tag{1.4}$$

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 (1.4) 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 $\mathbb{X} \subseteq \mathbb{R}^n$ 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 $\psi \in \mathcal{F}$ with the goal to minimize the MSE, and nature, who chooses the parameter value $x \in \mathbb{X}$ with the goal to maximize the MSE:

$$\underset{\psi \in \mathcal{F}}{\text{minimize}} \; \sup_{x \in \mathbb{X}} \; R(\psi, x). \tag{1.5a}$$

By construction, any minimizer $\psi^\star_{\mathcal{M}}$ of (1.5a) incurs the smallest possible estimation error under the worst parameter realization within the uncertainty set $\mathbb{X}$. For this reason $\psi^\star_{\mathcal{M}}$ is called a minimax estimator. Note that the MSE $R(\psi, x)$ generically displays a complicated nonconcave dependence on x for any fixed ψ, which implies that nature's inner maximization problem in (1.5a) is usually nonconvex. Thus, we should not expect the zero-sum game (1.5a) 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 (1.5a) as the (equivalent) convex–concave saddle point problem

$$\underset{\psi \in \mathcal{F}}{\text{minimize}} \; \sup_{\mathbb{Q}_x \in \mathcal{M}(\mathbb{X})} \; \mathbb{E}_{\mathbb{Q}_x}\big[R(\psi, x)\big], \tag{1.5b}$$

where $\mathcal{M}(\mathbb{X})$ stands for the family of all distributions supported on $\mathbb{X}$ 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, and $\mathcal{F}$ and $\mathcal{M}(\mathbb{X})$ are both convex sets, the zero-sum game (1.5b) 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 herein, this means that $\psi^\star_{\mathcal{M}}$ is the Bayesian MMSE estimator

corresponding to the prior $\mathbb{Q}_x^\star$. For this reason, $\mathbb{Q}_x^\star$ is usually referred to as the *least favorable prior*. Even though the minimax approach exonerates the statistician from narrowing down the statistician's 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 $\mathbb{X}$, the minimax approach is often regarded as overly pessimistic. Moreover, as in the case of the Bayesian MMSE estimation problem, $\psi_{\mathcal{M}}^\star$ may generically have a complicated functional form, and evaluating $\psi_{\mathcal{M}}^\star(y)$ 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 $\mathbb{X}$ is discussed in Juditsky and Nemirovski [42, section 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

$$\underset{\psi\in\mathcal{F}}{\text{minimize}} \ \max_{\mathbb{Q}_x\in\mathcal{Q}_x} \ \mathbb{E}_{\mathbb{Q}_x}\big[R(\psi,x)\big], \tag{1.6}$$

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 (1.6) reduces to the Bayesian MMSE estimation problem (1.2). Similarly, under the ambiguity set $\mathcal{Q}_x = \mathcal{M}(\mathbb{X})$ for some convex closed uncertainty set $\mathbb{X} \subseteq \mathbb{R}^n$, Problem (1.6) reduces to the minimax mean square error estimation problem (1.5b). 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 (1.6) 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 (1.6) 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* $\mathcal{R}: \mathcal{F} \times \mathcal{M}(\mathbb{R}^{n+m}) \to \mathbb{R}$ through

$$\mathcal{R}(\psi,\mathbb{P}) = \ E_{\mathbb{P}}[\|x-\psi(Hx+w)\|^2].$$

If $\mathbb{P} = \mathbb{P}_x \times \mathbb{P}_w$ for some marginal distributions $\mathbb{P}_x \in \mathcal{M}(\mathbb{R}^n)$ and $\mathbb{P}_w \in \mathcal{M}(\mathbb{R}^m)$, 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 $\mathcal{R}(\psi,\mathbb{P}) = \ \mathbb{E}_{\mathbb{P}_x}[R(\psi,x)]$. Thus, the average risk $\mathcal{R}(\psi,\mathbb{P})$ corresponds indeed to the risk $R(\psi,x)$ averaged under the marginal distribution $\mathbb{P}_x$. In the remainder of this paper, we study generalized distributionally robust estimation problems of the form

$$\underset{\psi\in\mathcal{F}}{\text{minimize}} \ \sup_{\mathbb{Q}\in\mathbb{B}(\widehat{\mathbb{P}})} \ \mathcal{R}(\psi,\mathbb{Q}), \tag{1.7}$$

where the ambiguity set $\mathbb{B}(\widehat{\mathbb{P}}) \subseteq \mathcal{M}(\mathbb{R}^{n+m})$ captures distributional uncertainty in both $\mathbb{P}_x$ and $\mathbb{P}_w$. Specifically, we 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 1.1 (Wasserstein Distance). For any $d \in \mathbb{N}$, the type 2 Wasserstein distance between two distributions $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{M}(\mathbb{R}^d)$ is defined as

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

where $\Pi(\mathbb{Q}_1,\mathbb{Q}_2)$ denotes the set of all joint distributions or couplings $\pi \in \mathcal{M}(\mathbb{R}^d \times \mathbb{R}^d)$ 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

$$\mathbb{B}(\widehat{\mathbb{P}}) = \left\{ \mathbb{Q}_x \times \mathbb{Q}_w : \begin{array}{lll} \mathbb{Q}_x & \in \mathcal{M}(\mathbb{R}^n), & \mathbb{W}(\mathbb{Q}_x,\widehat{\mathbb{P}}_x) \le \rho_x \\ \mathbb{Q}_w & \in \mathcal{M}(\mathbb{R}^m), & \mathbb{W}(\mathbb{Q}_w,\widehat{\mathbb{P}}_w) \le \rho_w \end{array} \right\}, \tag{1.8}$$

where $\widehat{\mathbb{P}}_x$ and $\widehat{\mathbb{P}}_w$ represent prescribed nominal distributions that can be constructed via statistical analysis or

expert judgment, and the Wasserstein radii $\rho_x \geq 0$ and $\rho_w \geq 0$ constitute hyperparameters that quantify the statistician's uncertainty about the nominal distributions of x and w. We emphasize that the distributionally robust estimation model (1.7) generalizes all preceding models. Indeed, if $\rho_w = 0$, then (1.7) reduces to the first distributionally robust model (1.6), which, in turn, encompasses both the MMSE estimation problem (1.2) (for $\rho_x = 0$) and the minimax estimation problem (1.5b) (for $\rho_x = \infty$) as special cases.

The distributionally robust estimation model (1.7) is conceptually attractive because the hyperparameters ρ_x and ρ_w allow the statistician to specify the statistician's 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 show that (1.7) 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 nonconvex.

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

$$\mathbb{B}'(\widehat{\mathbb{P}}) = \left\{\mathbb{Q}_x \times \mathbb{Q}_w : \mathbb{Q}_x \in \mathcal{M}(\mathbb{R}^n),\ \mathbb{Q}_w \in \mathcal{M}(\mathbb{R}^m),\ \mathbb{W}(\mathbb{Q}_x \times \mathbb{Q}_w, \widehat{\mathbb{P}}_x \times \widehat{\mathbb{P}}_w) \leq \rho\right\}, \tag{1.9}$$

which involves only a single hyperparameter $\rho \geq 0$ 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 $\mathbb{B}'(\widehat{\mathbb{P}})$. The proof of this result is relegated to the appendix.

Lemma 1.1 (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 $\mathbb{W}(\mathbb{Q}_x^1 \times \mathbb{Q}_w^1, \mathbb{Q}_x^2 \times \mathbb{Q}_w^2)^2 = \mathbb{W}(\mathbb{Q}_x^1, \mathbb{Q}_x^2)^2 + \mathbb{W}(\mathbb{Q}_w^1, \mathbb{Q}_w^2)^2$.*

If we denote the ambiguity sets (1.8) and (1.9) 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 1.1 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.

In the following, we summarize the key contributions of this paper.

1. We construct a safe approximation for the distributionally robust MMSE estimation problem (1.7) by restricting attention to affine estimators and by maximizing the average risk over an *outer* approximation of the Wasserstein ambiguity set, which is described through first- and second-order moment conditions. We then prove that this safe approximation is equivalent to a tractable convex program.

2. We also study a *dual* estimation problem, which is obtained by interchanging the minimization and maximization operations in the *primal* problem (1.7) and, thus, lower bounds the optimal value of (1.7). We then construct a safe approximation for this dual problem by restricting the Wasserstein ambiguity set to contain only *normal* distributions. Assuming that the nominal distribution is normal, we prove that this safe approximation is again equivalent to a tractable convex program.

3. By construction, the primal and dual estimation problems are upper and lower bounded by their respective safe approximations. We prove, however, that the optimal values of the safe approximations collapse if the nominal distribution is normal. This result has three important implications.

   a. The primal and dual estimation problems and their safe approximations are all equivalent. This implies via contributions (1) and (2) that both original estimation problems are tractable.

   b. The primal estimation problem is solved by an affine estimator, and the dual estimation problem is solved by a normal distribution. In other words, we have discovered a new class of adaptive, distributionally robust optimization problems for which affine decision rules are optimal.

   c. The affine estimator and the normal distribution that solve the primal and dual estimation problems, respectively, form a *Nash equilibrium* for the zero-sum game between the statistician and nature. Thus, the optimal normal distribution constitutes a least favorable prior, and the optimal affine estimator represents the corresponding *Bayesian MMSE estimator*.

4. We leverage these insights to prove that the optimal affine estimator can be constructed easily from the least favorable prior without the need to solve another optimization problem.

5. We argue that our main results remain valid if the nominal distribution is any elliptical distribution.

6. We develop a tailor-made Frank–Wolfe (FW) algorithm that can solve the dual estimation problem orders of magnitude faster than state-of-the-art general-purpose solvers. We show that this algorithm enjoys a linear convergence rate. Moreover, we prove that the direction-finding subproblems can be solved in quasi-closed form, which means that the algorithm offers a favorable iteration complexity.

We highlight that the Wasserstein ambiguity set (1.8) is nonconvex as it contains only distributions under which the signal and the noise are independent. To our best knowledge, we describe the first distributionally robust optimization model with independence conditions that admits a tractable reformulation. We also emphasize that some of our results hold only if the nominal distribution $\widehat{\mathbb{P}}$ is normal or elliptical. Although this is restrictive, there is strong evidence 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 (Cover and Thomas [15, section 12]). Therefore, it has appeal as the least prejudiced baseline model. Similarly, if the parameter x in (1.1) is normally distributed, then a normal distribution minimizes the mutual information between x and the observation y among all noise distributions with bounded variance (Diggavi and Cover [18, lemma II.2]). In this sense, normally distributed noise renders the observations least informative. Conversely, if the noise in (1.1) is normally distributed, then a normal distribution maximizes the MMSE across all distributions of x with bounded variance (Guo et al. [35, 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.

In the following, we briefly survey the landscape of existing MMSE estimation models that have a robustness flavor. Several authors have addressed the *minimax* MMSE estimation problem (1.5a) from the perspective of classical robust optimization (Beck and Eldar [3], Beck et al. [5], Eldar et al. [25, 26, 27], Juditsky and Nemirovski [43]). To guarantee computational tractability, in all of these papers, the estimators are restricted to be affine functions of the measurements. In this case, the minimax MMSE estimation problem can be reformulated as a tractable semidefinite program (SDP) if the uncertainty set $\mathbb{X}$ is an ellipsoid (Eldar et al. [26, 27]) or an intersection of two ellipsoids (Beck and Eldar [3]). Similar SDP reformulations are available if the observation matrix H is also subject to uncertainty and ranges over a spectral norm ball (Eldar et al. [27]) or displays a block circulant structure, with each block ranging over a Frobenius norm ball (Beck et al. [5]). If the uncertainty set is described by an intersection of several ellipsoids, then the minimax MMSE estimation problem admits an (inexact) SDP relaxation (Eldar et al. [25]). Even though the restriction to affine estimators may incur a loss of optimality, affine estimators are known to be near-optimal in all of these minimax estimation models (Juditsky and Nemirovski [43]).

Another stream of literature investigates the *distributionally robust* estimation model (1.6) with an ambiguous signal distribution and a crisp noise distribution. When focusing on affine estimators only, this model can be reformulated as a tractable SDP if the uncertainty in the signal distribution is characterized through spectral constraints on its covariance matrix (Eldar and Merhav [24]). This tractability result also extends to uncertain observation matrices. Similar SDP reformulations are available for the distributionally robust estimation model (1.7) when both the signal and the noise distribution are ambiguous and their covariance matrices are subject to spectral constraints (Eldar [23]). Extensions to uncertain block circulant observation matrices are discussed in Beck et al. [4].

Some authors study less structured distributionally robust estimation problems in which the signal x and the measurement y are governed by a distribution that may *not* obey the linear measurement model (1.1). In this case, the zero-sum game between the statistician and nature admits a Nash equilibrium if nature may choose any distribution that has a bounded Kullback–Leibler divergence with respect to a normal nominal distribution (Levy and Nikoukhah [48]). Intriguingly, the (affine) Bayesian MMSE estimator for the nominal distribution is optimal in this model and, thus, enjoys strong robustness properties. On the downside, there is no hope to improve this estimator's performance by tuning the size of the Kullback–Leibler ambiguity set. The underlying distributionally robust estimation model also serves as a fundamental building block for a robust Kalman filter (Levy and Nikoukhah [49]). Extensions to general τ-divergence ambiguity sets that contain only normal distributions are reported in Zorzi [77]. We emphasize that all papers surveyed so far merely derive SDP reformulations or SDP relaxations that can be addressed with general-purpose solvers, but none of them develops a customized solution algorithm.

The present paper extends the distributionally robust MMSE estimation model introduced in Shafieezadeh-Abadeh et al. [67], which accommodates a simple Wasserstein ambiguity set for the distribution of the signal-measurement pairs and makes no structural assumptions about the measurement noise. Note, however, that the linear measurement model (1.1) abounds in the literature on control theory, signal processing, and information theory, implying that there are numerous applications in which the measurement noise is *known* to be additive and independent of the signal. As we see in Section 7, ignoring this structural information may result in weak estimators that sacrifice predictive performance. In Sections 2–4, we further see that constructing an explicit Nash equilibrium is considerably more difficult in the presence of structural information. Finally, we describe here an accelerated Frank–Wolfe algorithm that improves the sublinear convergence rate established in Shafieezadeh-Abadeh et al. [67] to a linear rate. We emphasize that, in contrast to the robust MMSE estimators derived in Levy

and Nikoukhah [48] and Zorzi [77] that are insensitive to the radii of the underlying divergence-based ambiguity sets, the estimators constructed here change with the Wasserstein radii ρ_x and ρ_w. Thus, using a Wasserstein ambiguity set to robustify the nominal MMSE estimation problem has a regularizing effect and leads to a parametric family of estimators that can be tuned to attain maximum prediction accuracy. Similar connections between robustification and regularization have previously been discovered in the context of statistical learning (Shafieezadeh-Abadeh et al. [66]) and covariance estimation (Nguyen et al. [55]).

The paper is structured as follows. Sections 2 and 3 develop conservative approximations for the primal and dual distributionally robust MMSE estimation problems, respectively, both of which are equivalent to tractable convex programs. Section 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 nonnormal nominal distributions are discussed in Section 5. Section 6 develops an efficient Frank–Wolfe algorithm for the dual MMSE estimation problem, and Section 7 reports on numerical results.

1.1. Notation

For any $A \in \mathbb{R}^{d\times d}$, we use $\mathrm{Tr}[A]$ to denote the trace and $\|A\| = \sqrt{\mathrm{Tr}[A^\top A]}$ 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 $\langle A, B\rangle = \mathrm{Tr}[A^\top B]$ to denote the inner product and $A \otimes B$ to denote the Kronecker product of A and B. 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. The Gelbrich MMSE Estimation Problem

The distributionally robust estimation problem (1.7) poses two fundamental challenges. First, checking feasibility of the inner maximization problem in (1.7) 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 and $\mathbb{Q}_x$ and $\mathbb{Q}_w$ are uniform distributions on hypercubes (Taşkesen et al. [72]). Efficient algorithms for computing Wasserstein distances are available only if both involved distributions are discrete (Cuturi [16], Peyré and Cuturi [60], Solomon et al. [69]), and analytical formulas are only known in exceptional cases (e.g., if both distributions are Gaussian (Givens and Shortt [33]) or belong to the same family of elliptical distributions (Gelbrich [32])). The second challenge is that the outer minimization problem in (1.7) constitutes an infinite-dimensional functional optimization problem. In order to bypass these computational challenges, we first seek a conservative approximation for (1.7) by relaxing the ambiguity set $\mathbb{B}(\widehat{\mathbb{P}})$ and restricting the feasible set $\mathcal{F}$. 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 2.1 (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

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

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, nonnegative, vanishes if and only if $(\mu_1, \Sigma_1) = (\mu_2, \Sigma_2)$, and satisfies the triangle inequality (Givens and Shortt [33]).

Proposition 2.1 (Commuting Covariance Matrices (Givens and Shortt [33])). *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 $\mathbb{G}((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)) = \|\sqrt{\Sigma_1} - \sqrt{\Sigma_2}\|$.*

Although the Gelbrich distance is nonconvex, the squared Gelbrich distance is convex in all of its arguments.

Proposition 2.2 (Convexity and Continuity of the Squared Gelbrich Distance). *The squared Gelbrich distance $\mathbb{G}((\mu_1,\Sigma_1),(\mu_2,\Sigma_2))^2$ is jointly convex and continuous in $\mu_1, \mu_2 \in \mathbb{R}^d$ and $\Sigma_1, \Sigma_2 \in \mathbb{S}^d_+$.*

Proof of Proposition 2.2. By Malagò et al. [52, proposition 2], the squared Gelbrich distance $\mathbb{G}((\mu_1,\Sigma_1),(\mu_2,\Sigma_2))^2$ coincides with the optimal value of the semidefinite program

$$\begin{array}{ll}\min & \|\mu_1-\mu_2\|_2^2+\mathrm{Tr}[\Sigma_1+\Sigma_2-2C]\\ \text{s.t.} & C\in\mathbb{R}^{d\times d}\\ & \begin{bmatrix}\Sigma_1 & C\\ C^\top & \Sigma_2\end{bmatrix}\succeq 0,\end{array}$$

and see also Dowson and Landau [19, section 3]). Less general results that hold when one of the matrices Σ_1 or Σ_2 is positive definite are reported in Givens and Shortt [33], Knott and Smith [46], and Olkin and Pukelsheim [57]. The convexity of the squared Gelbrich distance then follows from Bertsekas [8, proposition 3.3.1], which guarantees that convexity is preserved under partial minimization. Moreover, the continuity of the squared Gelbrich distance follows from the continuity of the matrix square root established in Lemma A.2. □

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 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 2.3 (Moment Bound on the Wasserstein Distance (Gelbrich [32, Theorem 2.1])). *For any distributions* $\mathbb{Q}_1$, $\mathbb{Q}_2\in\mathcal{M}(\mathbb{R}^d)$ *with mean vectors* μ_1, $\mu_2\in\mathbb{R}^d$ *and covariance matrices* Σ_1, $\Sigma_2\in\mathbb{S}_+^d$, *respectively, we have*

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

Proposition 2.3 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

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

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 2.1 (Relation Between Gelbrich and Wasserstein Ambiguity Sets). *For any* $\widehat{\mathbb{P}}=\widehat{\mathbb{P}}_x\times\widehat{\mathbb{P}}_w$ *with* $\widehat{\mathbb{P}}_x\in\mathcal{M}(\mathbb{R}^n)$ *and* $\widehat{\mathbb{P}}_w\in\mathcal{M}(\mathbb{R}^m)$, *we have* $\mathbb{B}(\widehat{\mathbb{P}})\subseteq\mathbb{G}(\widehat{\mathbb{P}})$.

Proof of Corollary 2.1. 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 2.3, 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 $\mathcal{F}$ to the set $\mathcal{A}$ 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 (1.7).

$$\underset{\psi\in\mathcal{A}}{\text{minimize}}\ \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})}\ \mathcal{R}(\psi,\mathbb{Q})\tag{2.1}$$

From now on we call (1.7) and (2.1) the Wasserstein and Gelbrich MMSE estimation problems, and we refer to their minimizers as Wasserstein and Gelbrich MMSE estimators, respectively. As the average risk $\mathcal{R}(\psi,\mathbb{Q})$ of a fixed affine estimator $\psi\in\mathcal{A}$ 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 (2.1) is nonconvex. Thus, one might suspect that the Gelbrich MMSE estimation problem is intractable. We show, however, that (2.1) 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 (2.1) is stable with respect to changes of its input parameters.

Proposition 2.4 (Regularity of the Gelbrich MMSE Estimation Problem). *The Gelbrich MMSE estimation problem* (2.1) *enjoys the following regularity properties:*

i. *Conservativeness: Problem* (2.1) *upper bounds the Wasserstein MMSE estimation problem* (1.7).
ii. *Solvability: If* $\rho_x>0$ *and* $\rho_w>0$, *then the minimum of* (2.1) *is attained.*
iii. *Stability: If* $\rho_x>0$ *and* $\rho_w>0$, *then the minimum of* (2.1) *is continuous in* $(\widehat{\mu}_x,\widehat{\mu}_w,\widehat{\Sigma}_x,\widehat{\Sigma}_w)$.

The proof of Proposition 2.4 is lengthy and technical and is, therefore, relegated to the appendix. We are now ready to prove the main result of this section.

Theorem 2.1 (Gelbrich MMSE Estimation Problem). *The Gelbrich MMSE estimation problem* (2.1) *is equivalent to the finite convex optimization problem*

$$\begin{array}{ll} \inf & \gamma_x(\rho_x^2 - \mathrm{Tr}[\widehat{\Sigma}_x]) + \gamma_x^2\langle[\gamma_x I_n - (I_n - AH)^\top(I_n - AH)]^{-1}, \widehat{\Sigma}_x\rangle \\ & \quad + \gamma_w(\rho_w^2 - \mathrm{Tr}[\widehat{\Sigma}_w]) + \gamma_w^2\langle(\gamma_w I_m - A^\top A)^{-1}, \widehat{\Sigma}_w\rangle \\ \text{s.t.} & A \in \mathbb{R}^{n\times m}, \quad \gamma_x, \gamma_w \in \mathbb{R}_+ \\ & \gamma_x I_n - (I_n - AH)^\top(I_n - AH) \succ 0, \quad \gamma_w I_m - A^\top A \succ 0. \end{array} \tag{2.2}$$

Moreover, if $\rho_x > 0$ *and* $\rho_w > 0$, *then* (2.2) *admits an optimal solution*[1] $A^\star$, *and the infimum of* (2.1) *is attained by the affine estimator* $\psi^\star(y) = A^\star y + b^\star$, *where* $b^\star = \widehat{\mu}_x - A^\star(H\widehat{\mu}_x + \widehat{\mu}_w)$.

The *strict* semidefinite inequalities in (2.2) ensure that the inverse matrices in the objective function exist.

Proof of Theorem 2.1. Throughout this proof, we denote by $\psi_{A,b} \in \mathcal{A}$ 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 $\mathcal{R}(\psi, \mathbb{Q})$ and the Gelbrich ambiguity set $\mathbb{G}(\widehat{\mathbb{P}})$, we then have

$$\inf_b \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathcal{R}(\psi_{A,b}, \mathbb{Q}) = \left\{ \begin{array}{lll} \inf\limits_b & \sup\limits_{\substack{\mu_x,\mu_w \\ \Sigma_x,\Sigma_w\succeq 0}} & \langle K^\top K, \Sigma_x + \mu_x\mu_x^\top\rangle + \langle A^\top A, \Sigma_w + \mu_w\mu_w^\top\rangle + b^\top b \\ & & \quad - 2\mu_x^\top K^\top A\mu_w - 2b^\top(K\mu_x - A\mu_w) \\ & \text{s.t.} & \mathbb{G}((\mu_x, \Sigma_x), (\widehat{\mu}_x, \widehat{\Sigma}_x))^2 \le \rho_x^2 \\ & & \mathbb{G}((\mu_w, \Sigma_w), (\widehat{\mu}_w, \widehat{\Sigma}_w))^2 \le \rho_w^2. \end{array} \right. \tag{2.3}$$

The outer minimization problem in (2.3) 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 (2.3) is nonconvex because its objective function is convex in (μ_x, μ_w). This observation prompts us to maximize over (μ_x, μ_w) and (Σ_x, Σ_w) sequentially and to reformulate (2.3) as

$$\begin{array}{lll} \inf\limits_b & \sup\limits_{\substack{\mu_x,\mu_w \\ \|\mu_x - \widehat{\mu}_x\|\le\rho_x \\ \|\mu_w - \widehat{\mu}_w\|\le\rho_w}} & \sup\limits_{\Sigma_x,\Sigma_w\succeq 0} \langle K^\top K, \Sigma_x + \mu_x\mu_x^\top\rangle + \langle A^\top A, \Sigma_w + \mu_w\mu_w^\top\rangle + b^\top b \\ & & \quad - 2\mu_x^\top K^\top A\mu_w - 2b^\top(K\mu_x - A\mu_w) \\ & \text{s.t.} & \mathbb{G}((\mu_x, \Sigma_x), (\widehat{\mu}_x, \widehat{\Sigma}_x))^2 \le \rho_x^2 \\ & & \mathbb{G}((\mu_w, \Sigma_w), (\widehat{\mu}_w, \widehat{\Sigma}_w))^2 \le \rho_w^2. \end{array} \tag{2.4}$$

As $\|\mu_x - \widehat{\mu}_x\| \le \mathbb{G}((\mu_x, \Sigma_x), (\widehat{\mu}_x, \widehat{\Sigma}_x))$ and as this inequality is tight for $\Sigma_x = \widehat{\Sigma}_x$, the extra constraint $\|\mu_x - \widehat{\mu}_x\| \le \rho_x$ 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 (2.4) admits the Lagrangian dual

$$\begin{array}{ll} \inf\limits_{\gamma_x,\gamma_w\ge 0} \sup\limits_{\Sigma_x,\Sigma_w\succeq 0} & \langle K^\top K, \Sigma_x + \mu_x\mu_x^\top\rangle + \langle A^\top A, \Sigma_w + \mu_w\mu_w^\top\rangle + b^\top b - 2\mu_x^\top K^\top A\mu_w - 2b^\top(K\mu_x - A\mu_w) \\ & + \gamma_x(\rho_x^2 - \|\mu_x - \widehat{\mu}_x\|^2 - \mathrm{Tr}[\Sigma_x + \widehat{\Sigma}_x - 2(\widehat{\Sigma}_x^{\frac{1}{2}}\Sigma_x\widehat{\Sigma}_x^{\frac{1}{2}})^{\frac{1}{2}}]) \\ & + \gamma_w(\rho_w^2 - \|\mu_w - \widehat{\mu}_w\|^2 - \mathrm{Tr}[\Sigma_w + \widehat{\Sigma}_w - 2(\widehat{\Sigma}_w^{\frac{1}{2}}\Sigma_w\widehat{\Sigma}_w^{\frac{1}{2}})^{\frac{1}{2}}]). \end{array} \tag{2.5}$$

Strong duality holds by Bertsekas [8, proposition 5.5.4], which applies because the primal problem has a nonempty compact feasible set. Next, we observe that the inner maximization problem in (2.5) can be solved analytically by using Proposition A.1, and thus, the dual problem (2.5) is equivalent to

$$\begin{array}{ll} \inf\limits_{\substack{\gamma_x,\gamma_w \\ \gamma_x I_n \succ K^\top K \\ \gamma_w I_m \succ A^\top A}} & \langle K^\top K, \mu_x\mu_x^\top\rangle + \langle A^\top A, \mu_w\mu_w^\top\rangle + b^\top b - 2\mu_x^\top K^\top A\mu_w - 2b^\top(K\mu_x - A\mu_w) \\ & \quad + \gamma_x(\rho_x^2 - \|\mu_x - \widehat{\mu}_x\|^2 - \mathrm{Tr}[\widehat{\Sigma}_x] + \gamma_x\langle(\gamma_x I_n - K^\top K)^{-1}, \widehat{\Sigma}_x\rangle) \\ & \quad + \gamma_w(\rho_w^2 - \|\mu_w - \widehat{\mu}_w\|^2 - \mathrm{Tr}[\widehat{\Sigma}_w] + \gamma_w\langle(\gamma_w I_m - A^\top A)^{-1}, \widehat{\Sigma}_w\rangle). \end{array} \tag{2.6}$$

Substituting (2.6) back into (2.4) then allows us to reformulate the Gelbrich MMSE estimation problem (2.6) as

$$\begin{array}{lll} \inf\limits_{b} & \sup\limits_{\substack{\mu_x,\mu_w \\ \|\mu_x-\widehat{\mu}_x\|\le\rho_x \\ \|\mu_w-\widehat{\mu}_w\|\le\rho_w}} & \inf\limits_{\substack{\gamma_x,\gamma_w \\ \gamma_x I_n \succ K^\top K \\ \gamma_w I_m \succ A^\top A}} \\ & & \begin{array}{l} \langle K^\top K, \mu_x\mu_x^\top\rangle + \langle A^\top A, \mu_w\mu_w^\top\rangle + b^\top b - 2\mu_x^\top K^\top A\mu_w - 2b^\top(K\mu_x - A\mu_w) \\ \quad + \gamma_x(\rho_x^2 - \|\mu_x-\widehat{\mu}_x\|^2 - \mathrm{Tr}[\widehat{\Sigma}_x] + \gamma_x\langle(\gamma_x I_n - K^\top K)^{-1}, \widehat{\Sigma}_x\rangle) \\ \quad + \gamma_w(\rho_w^2 - \|\mu_w-\widehat{\mu}_w\|^2 - \mathrm{Tr}[\widehat{\Sigma}_w] + \gamma_w\langle(\gamma_w I_m - A^\top A)^{-1}, \widehat{\Sigma}_w\rangle). \end{array} \end{array} \tag{2.7}$$

The infimum of the inner minimization problem over (γ_x, γ_w) in (2.7) 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 (2.7) then yields

$$\begin{array}{lll} \sup\limits_{\substack{\mu_x,\mu_w \\ \|\mu_x-\widehat{\mu}_x\|\le\rho_x \\ \|\mu_w-\widehat{\mu}_w\|\le\rho_w}} & \inf\limits_{\substack{\gamma_x,\gamma_w \\ \gamma_x I_n \succ K^\top K \\ \gamma_w I_m \succ A^\top A}} & \begin{array}{l} \gamma_x(\rho_x^2 - \|\mu_x-\widehat{\mu}_x\|^2 - \mathrm{Tr}[\widehat{\Sigma}_x]) + \gamma_x^2\langle(\gamma_x I_n - K^\top K)^{-1}, \widehat{\Sigma}_x\rangle \\ \quad + \gamma_w(\rho_w^2 - \|\mu_w-\widehat{\mu}_w\|^2 - \mathrm{Tr}[\widehat{\Sigma}_w]) + \gamma_w^2\langle(\gamma_w I_m - A^\top A)^{-1}, \widehat{\Sigma}_w\rangle. \end{array} \end{array} \tag{2.8}$$

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 (2.3) finally yields

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

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

$$\inf_{\psi\in\mathcal{A}} \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathcal{R}(\psi, \mathbb{Q}) = \inf_{A,b} \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathcal{R}(\psi_{A,b}, \mathbb{Q}) \tag{2.9}$$

is indeed equivalent to the finite convex optimization problem (2.2).

Assume now that $\rho_x > 0$ and $\rho_w > 0$. In this case, we know from Proposition 2.4(ii) that the Gelbrich MMSE estimation problem (2.9) 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 (2.2). Moreover, it implies that $b^\star$ is optimal in (2.3) when we fix $A = A^\star$. As (2.3) is equivalent to (2.7) and as the unique optimal solution of (2.7) 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 (2.2) and $b^\star$ is defined as before, then the affine estimator $\psi^\star(y) = A^\star y + b^\star$ is optimal in (2.9). This observation completes the proof. □

Using Schur complement arguments, the convex program (2.2) can be further simplified to a standard SDP, which can be addressed with off-the-shelf solvers.

Corollary 2.2 (SDP Reformulation). *The Gelbrich MMSE estimation problem* (2.1) *is equivalent to the SDP*

$$\begin{array}{ll} \inf & \gamma_x(\rho_x^2 - \mathrm{Tr}[\widehat{\Sigma}_x]) + \mathrm{Tr}[U_x] + \gamma_w(\rho_w^2 - \mathrm{Tr}[\widehat{\Sigma}_w]) + \mathrm{Tr}[U_w] \\ \text{s.t.} & A \in \mathbb{R}^{n\times m}, \quad \gamma_x, \gamma_w \in \mathbb{R}_+ \\ & U_x \in \mathbb{S}_+^n, \quad V_x \in \mathbb{S}_+^n, \quad U_w \in \mathbb{S}_+^m, \quad V_w \in \mathbb{S}_+^m \\ & \begin{bmatrix} U_x & \gamma_x\widehat{\Sigma}_x^{\frac{1}{2}} \\ \gamma_x\widehat{\Sigma}_x^{\frac{1}{2}} & V_x \end{bmatrix} \succeq 0, \quad \begin{bmatrix} \gamma_x I_n - V_x & I_n - H^\top A^\top \\ I_n - AH & I_n \end{bmatrix} \succeq 0 \\ & \begin{bmatrix} U_w & \gamma_w\widehat{\Sigma}_w^{\frac{1}{2}} \\ \gamma_w\widehat{\Sigma}_w^{\frac{1}{2}} & V_w \end{bmatrix} \succeq 0, \quad \begin{bmatrix} \gamma_w I_m - V_w & A^\top \\ A & I_n \end{bmatrix} \succeq 0. \end{array} \tag{2.10}$$

Proof of Corollary 2.2. Define the extended real-valued function $h_w : \mathbb{R}^{n\times m} \times \mathbb{R}_+ \to (-\infty, \infty]$ through

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

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

$$\begin{aligned} h_w(A, \gamma_w) &= \inf_{U_w \succeq 0} \left\{ \mathrm{Tr}[U_w] \;:\; U_w \succeq \gamma_w^2 \widehat{\Sigma}_w^{\frac{1}{2}} (\gamma_w I_m - A^\top A)^{-1} \widehat{\Sigma}_w^{\frac{1}{2}} \right\} \\ &= \inf_{U_w \succeq 0, V_w \succ 0} \left\{ \mathrm{Tr}[U_w] \;:\; U_w \succeq \gamma_w^2 \widehat{\Sigma}_w^{\frac{1}{2}} V_w^{-1} \widehat{\Sigma}_w^{\frac{1}{2}},\ \gamma_w I_m - A^\top A \succeq V_w \right\} \\ &= \inf_{U_w \succeq 0, V_w \succ 0} \left\{ \mathrm{Tr}[U_w] \;:\; \begin{bmatrix} \gamma_w I_m - V_w & A^\top \\ A & I_n \end{bmatrix} \succeq 0,\ \begin{bmatrix} U_w & \gamma_w \widehat{\Sigma}_w^{\frac{1}{2}} \\ \gamma_w \widehat{\Sigma}_w^{\frac{1}{2}} & V_w \end{bmatrix} \succeq 0 \right\}, \end{aligned} \tag{2.11}$$

where the first equality holds because of the cyclicity of the trace operator and because $U_w \succeq \bar{U}_w$ implies $\mathrm{Tr}[U_w] \geq \mathrm{Tr}[\bar{U}_w]$ for all $U_w, \bar{U}_w \succeq 0$, the second equality holds because $V_w \succeq \bar{V}_w$ is equivalent to $V_w^{-1} \preccurlyeq \bar{V}_w^{-1}$ for all $V_w, \bar{V}_w \succ 0$, and the last equality follows from standard Schur complement arguments; see, for example, Boyd and Vandenberghe [10, section A.5.5]. If $\gamma_w I_m - A^\top A \not\succ 0$, on the other hand, then the first matrix inequality in (2.11) implies that V_w must have at least one nonpositive eigenvalue, which contradicts the constraint $V_w \succ 0$. The SDP (2.11) is, therefore, infeasible, and its infimum evaluates to ∞. Thus, $h_w(A, \gamma_w)$ coincides with the optimal value of the SDP (2.11) 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

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

The claim now follows by substituting the SDP reformulations for $h_w(A, \gamma_w)$ and $h_x(A, \gamma_x)$ into (2.2). 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 (nonempty) feasible set and does not change the infimum of Problem (2.2). This observation completes the proof. □

Remark 2.1 (Numerical Stability). The SDP (2.10) 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 (2.10) that involve $\widehat{\Sigma}_x^{\frac{1}{2}}$ and $\widehat{\Sigma}_w^{\frac{1}{2}}$ with

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

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.

3. The Dual Wasserstein MMSE Estimation Problem over Normal Priors

We now examine the dual Wasserstein MMSE estimation problem

$$\sup_{\mathbb{Q} \in \mathbb{B}(\widehat{\mathbb{P}})} \inf_{\psi \in \mathcal{F}} \mathcal{R}(\psi, \mathbb{Q}), \tag{3.1}$$

which is obtained from (1.7) by interchanging the order of minimization and maximization. Any maximizer $\mathbb{Q}^\star$ of this dual estimation problem, if it exists, is henceforth called a least favorable prior. Unfortunately, Problem (3.1) is generically intractable. We demonstrate, however, that (3.1) becomes tractable if the nominal distribution $\widehat{\mathbb{P}}$ is normal.

Definition 3.1 (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 $\mathrm{supp}(\mathbb{P}) = \{\mu + \mathrm{Ev} : \mathrm{v} \in \mathbb{R}^{\mathrm{k}}\}$ and if the density function of $\mathbb{P}$ with respect to the Lebesgue measure on $\mathrm{supp}(\mathbb{P})$ is given by

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

where $k = \mathrm{rank}(\Sigma)$, $\mathrm{D} \in \mathbb{S}_{++}^{\mathrm{k}}$ 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 3.1 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 3.1 (Affine Transformations (Fang et al. [28, Theorem 2.16])). *If $\xi \in \mathbb{R}^d$ follows the normal distribution $\mathcal{N}(\mu, \Sigma)$ and $A \in \mathbb{R}^{k\times d}$ 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 3.2 (Affine Conditional Expectations (Cambanis et al. [11, Corollary 5])). *Assume that $\xi \in \mathbb{R}^d$ follows the normal distribution $\mathbb{P}\mathcal{N}(\mu, \Sigma)$ and that*

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

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 $\mathbb{E}_{\mathbb{P}}[\xi_1 \mid \xi_2] = A\xi_2 + b$ $\mathbb{P}$-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 3.3 (Wasserstein Distance Between Normal Distributions (Givens and Shortt [33, 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,

$$\widehat{\mathbb{P}} = \mathcal{N}(\widehat{\mu}, \widehat{\Sigma}) \quad \text{where} \quad \widehat{\mu} = \begin{bmatrix} \widehat{\mu}_x \\ \widehat{\mu}_w \end{bmatrix} \quad \text{and} \quad \widehat{\Sigma} = \begin{bmatrix} \widehat{\Sigma}_x & 0 \\ 0 & \widehat{\Sigma}_w \end{bmatrix}. \tag{3.2}$$

Armed with the fundamental results on normal distributions summarized before, we are now ready to address the dual Wasserstein MMSE estimation problem (3.1) with a normal nominal distribution. Analogous to Section 2, in which we proposed the Gelbrich MMSE estimation problem as an easier conservative approximation for the original primal estimation problem (1.7), we now construct an easier conservative approximation for the original dual estimation problem (3.1). To this end, we define the restricted ambiguity set

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

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 for which 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 nonconvex 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 (3.1):

$$\operatorname*{maximize}_{\mathbb{Q}\in\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})} \inf_{\psi\in\mathcal{F}} \mathcal{R}(\psi, \mathbb{Q}). \tag{3.3}$$

We henceforth refer to (3.3) as the dual Wasserstein MMSE estimation problem *over normal priors*. The following main theorem shows that (3.3) is equivalent to a finite convex optimization problem.

Theorem 3.1 (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 (3.2). Then, the dual Wasserstein MMSE estimation problem over normal priors (3.3) is equivalent to the finite convex optimization problem*

$$\begin{array}{ll} \sup & \operatorname{Tr}[\Sigma_x - \Sigma_x H^\top (H\Sigma_x H^\top + \Sigma_w)^{-1} H\Sigma_x] \\ \text{s.t.} & \Sigma_x \in \mathbb{S}_+^n, \quad \Sigma_w \in \mathbb{S}_{++}^m \\ & \operatorname{Tr}[\Sigma_x + \widehat{\Sigma}_x - 2(\widehat{\Sigma}_x^{\frac{1}{2}} \Sigma_x \widehat{\Sigma}_x^{\frac{1}{2}})^{\frac{1}{2}}] \le \rho_x^2 \\ & \operatorname{Tr}[\Sigma_w + \widehat{\Sigma}_w - 2(\widehat{\Sigma}_w^{\frac{1}{2}} \Sigma_w \widehat{\Sigma}_w^{\frac{1}{2}})^{\frac{1}{2}}] \le \rho_w^2. \end{array} \tag{3.4}$$

If $\widehat{\Sigma}_w \succ 0$, then (3.4) 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 (3.3) 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 of Theorem 3.1. 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 3.1, and the average risk $\mathcal{R}(\psi, \mathbb{Q})$ is minimized by the Bayesian MMSE estimator $\psi_{\mathcal{B}}^\star(y) = \mathbb{E}_{\mathbb{P}_{x|y}}[x]$, which is affine because of Proposition 3.2. 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,

$$\sup_{\mathbb{Q} \in \mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})} \inf_{\psi \in \mathcal{F}} \mathcal{R}(\psi, \mathbb{Q}) = \sup_{\mathbb{Q} \in \mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})} \inf_{\psi \in \mathcal{A}} \mathcal{R}(\psi, \mathbb{Q}). \tag{3.5}$$

As the average risk $\mathcal{R}(\psi, \mathbb{Q})$ of an affine estimator $\psi \in \mathcal{A}$ 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 3.3. Thus, the maximization problem over $\mathbb{Q}$ on the right-hand side of (3.5) can be recast as an equivalent maximization problem over the first and second moments of $\mathbb{Q}$. Specifically, by the definitions of $\mathcal{R}(\psi, \mathbb{Q})$ and $\mathbb{B}_\phi(\widehat{\mathbb{P}})$, we find

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

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 preceding problem and recalling the definition of the Gelbrich distance then yields

$$\sup_{\mathbb{Q} \in \mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})} \inf_{\psi \in \mathcal{A}} \mathcal{R}(\psi, \mathbb{Q}) = \begin{cases} \sup\limits_{\Sigma_x, \Sigma_w} \inf\limits_{\substack{A,K \\ K = I_n - AH}} \ \langle K^\top K, \Sigma_x \rangle + \langle A^\top A, \Sigma_w \rangle \\ \text{s.t.} \quad \mathrm{Tr}[\Sigma_x + \widehat{\Sigma}_x - 2(\widehat{\Sigma}_x^{\frac{1}{2}} \Sigma_x \widehat{\Sigma}_x^{\frac{1}{2}})^{\frac{1}{2}}] \le \rho_x^2 \\ \qquad \mathrm{Tr}[\Sigma_w + \widehat{\Sigma}_w - 2(\widehat{\Sigma}_w^{\frac{1}{2}} \Sigma_w \widehat{\Sigma}_w^{\frac{1}{2}})^{\frac{1}{2}}] \le \rho_w^2 \\ \qquad \Sigma_x \succeq 0, \ \Sigma_w \succ 0. \end{cases} \tag{3.6}$$

By using the equality $K = I_n - AH$ to eliminate K, the inner minimization problem in (3.6) 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

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

Substituting $A^\star$ as well as the corresponding auxiliary decision variable $K^\star = I_n - A^\star H$ into the objective function of (3.6) finally yields the postulated convex program (3.4).

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

$$\mathcal{S}_x = \left\{ \Sigma_x \in \mathbb{S}_+^n : \ \mathbb{G}((\widehat{\mu}_x, \Sigma_x), (\widehat{\mu}_x, \widehat{\Sigma}_x)) \le \rho_x \right\} \quad \text{and} \quad \mathcal{S}_w = \left\{ \Sigma_w \in \mathbb{S}_+^m : \ \mathbb{G}((\widehat{\mu}_w, \Sigma_w), (\widehat{\mu}_w, \widehat{\Sigma}_w)) \le \rho_w \right\}.$$

Equations (3.5) and (3.6) imply that

$$\begin{aligned} \sup_{\mathbb{Q} \in \mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})} \inf_{\psi \in \mathcal{F}} \mathcal{R}(\psi, \mathbb{Q}) &\le \sup_{\Sigma_x \in \mathcal{S}_x} \sup_{\Sigma_w \in \mathcal{S}_w} \inf_{\substack{A,K \\ K = I_n - AH}} \ \langle K^\top K, \Sigma_x \rangle + \langle A^\top A, \Sigma_w \rangle \\ &= \sup_{\substack{\Sigma_x \in \mathcal{S}_x \\ \Sigma_x \succeq \lambda_{\min}(\widehat{\Sigma}_x) I_n}} \sup_{\substack{\Sigma_w \in \mathcal{S}_w \\ \Sigma_w \succeq \lambda_{\min}(\widehat{\Sigma}_w) I_m}} \inf_{\substack{A,K \\ K = I_n - AH}} \ \langle K^\top K, \Sigma_x \rangle + \langle A^\top A, \Sigma_w \rangle, \end{aligned} \tag{3.7}$$

where the inequality holds because we relax the requirement that Σ_w be strictly positive definite, and the equality follows from applying Lemma A.3 consecutively to each of the two maximization problems. If $\widehat{\Sigma}_w \succ 0$, then

Problem (3.7) constitutes a restriction of (3.6) and, therefore, provides also a lower bound on the dual Wasserstein MMSE estimation problem. In summary, we, thus, have

$$\sup_{\mathbb{Q}\in\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})}\inf_{\psi\in\mathcal{F}}\mathcal{R}(\psi,\mathbb{Q})=\left\{\begin{array}{ll}\displaystyle\sup_{\Sigma_x,\Sigma_w}\ \inf_{\substack{A,K\\ K=I_n-AH}} & \langle K^\top K,\Sigma_x\rangle+\langle A^\top A,\Sigma_w\rangle\\ \text{s.t.} & \mathrm{Tr}[\Sigma_x+\widehat{\Sigma}_x-2(\widehat{\Sigma}_x^{\frac12}\Sigma_x\widehat{\Sigma}_x^{\frac12})^{\frac12}]\le\rho_x^2\\ & \mathrm{Tr}[\Sigma_w+\widehat{\Sigma}_w-2(\widehat{\Sigma}_w^{\frac12}\Sigma_w\widehat{\Sigma}_w^{\frac12})^{\frac12}]\le\rho_w^2\\ & \Sigma_x\succeq\lambda_{\min}(\widehat{\Sigma}_x)I_n,\ \Sigma_w\succeq\lambda_{\min}(\widehat{\Sigma}_w)I_m.\end{array}\right.\tag{3.8}$$

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 (3.6) and, consequently, to Problem (3.4) without altering their common optimal value. Problem (3.4) 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 (3.4), which has no redundant constraints. Thus, Problem (3.4) is solvable.

It remains to show that $\mathbb{Q}^\star$ as constructed in the theorem statement is optimal in (3.3). The feasibility of $(\Sigma_x^\star,\Sigma_w^\star)$ in (3.4) implies that $\mathbb{Q}^\star\in\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})$, and thus, $\mathbb{Q}^\star$ is feasible in (3.3). Moreover, we have

$$\sup_{\mathbb{Q}\in\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})}\inf_{\psi\in\mathcal{F}}\mathcal{R}(\psi,\mathbb{Q})\ge\inf_{\psi\in\mathcal{F}}\mathcal{R}(\psi,\mathbb{Q}^\star)=\mathrm{Tr}[\Sigma_x^\star-\Sigma_x^\star H^\top(H\Sigma_x^\star H^\top+\Sigma_w^\star)^{-1}H\Sigma_x^\star],\tag{3.9}$$

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

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

is the Bayesian MMSE estimator for the normal distribution $\mathbb{Q}^\star$. As the right-hand side of (3.9) coincides with the maximum of (3.4) and as Problem (3.4) is equivalent to the dual Wasserstein MMSE estimation problem (3.3) over normal priors, we may, thus, conclude that the inequality in (3.9) 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 (3.3). This observation completes the proof.

Remark 3.1 (Singular Covariance Matrices). A nonlinear SDP akin to (3.4) is derived in Shafieezadeh-Abadeh et al. [67] under the stronger assumption that the covariance matrix of the nominal distribution $\widehat{\mathbb{P}}$ is nondegenerate, 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 (3.4) is well defined. Therefore, Theorem 3.1 remains valid if the nominal covariance matrix $\widehat{\Sigma}_x$ is singular, which occurs in many applications. On the other hand, it is common to require that $\widehat{\Sigma}_w=\sigma^2 I_m$ for some $\sigma>0$; see, for example, Chang et al. [12].

Corollary 3.1 asserts that the convex program (3.4) admits a canonical linear SDP reformulation. The proof is omitted as it relies on standard Schur complement arguments familiar from the proof of Corollary 2.2.

Corollary 3.1 (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 (3.2) with noise covariance matrix $\widehat{\Sigma}_w\succ 0$. Then, the dual Wasserstein MMSE estimation problem (3.3) over normal priors is equivalent to the SDP*

$$\begin{array}{ll}\max & \mathrm{Tr}[\Sigma_x]-\mathrm{Tr}[U]\\ \text{s.t.} & \Sigma_x\in\mathbb{S}_+^n,\ \Sigma_w\in\mathbb{S}_+^m,\ V_x\in\mathbb{S}_+^n,\ V_w\in\mathbb{S}_+^m,\ U\in\mathbb{S}_+^n\\ & \begin{bmatrix}\widehat{\Sigma}_x^{\frac12}\Sigma_x\widehat{\Sigma}_x^{\frac12} & V_x\\ V_x & I_n\end{bmatrix}\succeq 0,\quad\begin{bmatrix}\widehat{\Sigma}_w^{\frac12}\Sigma_w\widehat{\Sigma}_w^{\frac12} & V_w\\ V_w & I_m\end{bmatrix}\succeq 0\\ & \mathrm{Tr}[\Sigma_x+\widehat{\Sigma}_x-2V_x]\le\rho_x^2,\quad\mathrm{Tr}[\Sigma_w+\widehat{\Sigma}_w-2V_w]\le\rho_w^2\\ & \begin{bmatrix}U & \Sigma_x H^\top\\ H\Sigma_x & H\Sigma_x H^\top+\Sigma_w\end{bmatrix}\succeq 0,\quad\Sigma_x\succeq\lambda_{\min}(\widehat{\Sigma}_x)I_n,\quad\Sigma_w\succeq\lambda_{\min}(\widehat{\Sigma}_w)I_m.\end{array}\tag{3.10}$$

We emphasize that the lower bounds on Σ_x and Σ_w are redundant but have been made explicit in (3.10).

4. Nash Equilibrium and Optimality of Affine Estimators

If $\widehat{\mathbb{P}}$ is a normal distribution of the form (3.2), then we have

$$\inf_{\psi\in\mathcal{A}} \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathcal{R}(\psi,\mathbb{Q}) \geq \inf_{\psi\in\mathcal{F}} \sup_{\mathbb{Q}\in\mathbb{B}(\widehat{\mathbb{P}})} \mathcal{R}(\psi,\mathbb{Q}) \geq \sup_{\mathbb{Q}\in\mathbb{B}(\widehat{\mathbb{P}})} \inf_{\psi\in\mathcal{F}} \mathcal{R}(\psi,\mathbb{Q}) \geq \sup_{\mathbb{Q}\in\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})} \inf_{\psi\in\mathcal{F}} \mathcal{R}(\psi,\mathbb{Q}), \tag{4.1}$$

where the first inequality follows from the inclusions $\mathcal{A}\subseteq\mathcal{F}$ and $\mathbb{B}(\widehat{\mathbb{P}})\subseteq\mathbb{G}(\widehat{\mathbb{P}})$, the second inequality exploits weak duality, and the last inequality holds because of the inclusion $\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})\subseteq\mathbb{B}(\widehat{\mathbb{P}})$. Note that the leftmost minimax problem is the Gelbrich MMSE estimation problem (2.1) studied in Section 2, and the rightmost maximin problem is the dual Wasserstein MMSE estimation problem (3.3) over normal priors studied in Section 3. We also highlight that these restricted primal and dual estimation problems sandwich the original Wasserstein estimation problems (1.7) and (3.1), which coincide with the second and third problems in (4.1), respectively. The following theorem asserts that all inequalities in (4.1) actually collapse to equalities.

Theorem 4.1 (Sandwich Theorem). *If $\widehat{\mathbb{P}}$ is a normal distribution of the form (3.2), then the optimal values of the restricted primal and dual estimation problems (2.1) and (3.3) coincide, that is,*

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

Proof of Theorem 4.1. By Theorem 2.1, the Gelbrich MMSE estimation problem (2.1) can be expressed as

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

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

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

and

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

both of which are convex and compact by virtue of Lemma A.4. Using Proposition A.2(i) to reformulate the inner minimization problem over γ_x and γ_w, we then obtain

$$\begin{aligned} \inf_{\psi\in\mathcal{A}} \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathcal{R}(\psi,\mathbb{Q}) &= \inf_{\substack{A,K \\ K=I_n-AH}} \sup_{\substack{\Sigma_x\in\mathcal{S}_x \\ \Sigma_w\in\mathcal{S}_w}} \langle K^\top K,\Sigma_x\rangle + \langle A^\top A,\Sigma_w\rangle \\ &= \sup_{\Sigma_x\in\mathcal{S}_x} \inf_{\substack{A,K \\ K=I_n-AH}} \sup_{\Sigma_w\in\mathcal{S}_w} \langle K^\top K,\Sigma_x\rangle + \langle A^\top A,\Sigma_w\rangle, \end{aligned}$$

where the second equality holds because of Sion's [68] minimax theorem. Define now the auxiliary function

$$f(A) = \sup_{\Sigma_w\in\mathcal{S}_w} \langle A^\top A,\Sigma_w\rangle.$$

As $\Sigma_w \succeq 0$ for any $\Sigma_w \in \mathcal{S}_w$, f constitutes a pointwise maximum of convex functions and is, therefore, itself convex. In addition, as the set $\mathcal{S}_w$ is compact by Lemma A.4, f is everywhere finite and, thus, continuous thanks to Rockafellar and Wets [62, theorem 2.35]. This allows us to conclude that

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

where the second equality exploits Lemma A.3, and the last equality follows from Sion's [68] minimax theorem. Another (trivial) application of Lemma A.3 then allows us to append the constraint $\Sigma_w \succeq \lambda_{\min}(\widehat{\Sigma}_w) I_m$ to the maximization problem over Σ_w. Sion's [68] minimax theorem finally implies that

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

where the last equality has already been established in the proof of Theorem 3.1; see Equation (3.8). Thus, the claim follows.

Theorem 4.1 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 4.1 (Nash Equilibrium). *If $\widehat{\mathbb{P}}$ is a normal distribution of the form (3.2) with $\widehat{\Sigma}_w \succ 0$, then the affine estimator $\psi^\star$ that solves (2.1) is optimal in the primal Wasserstein MMSE estimation problem (1.7), and the normal distribution $\mathbb{Q}^\star$ that solves (3.3) is optimal in the dual Wasserstein MMSE estimation problem (3.1). Moreover, $\psi^\star$ and $\mathbb{Q}^\star$ form a Nash equilibrium for the game between the statistician and nature, that is,*

$$\mathcal{R}(\psi^\star,\mathbb{Q}) \le \mathcal{R}(\psi^\star,\mathbb{Q}^\star) \le \mathcal{R}(\psi,\mathbb{Q}^\star) \quad \forall \psi\in\mathcal{F},\ \mathbb{Q}\in\mathbb{B}(\widehat{\mathbb{P}}). \tag{4.2}$$

Proof of Corollary 4.1. As $\widehat{\Sigma}_w \succ 0$, the Gelbrich MMSE estimation problem (2.1) is solved by the affine estimator $\psi^\star$ defined in Theorem 2.1, and the dual Wasserstein MMSE estimation problem (3.1) over normal priors is solved by the normal distribution $\mathbb{Q}^\star$ defined in Theorem 3.1. Thus, we have

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

where the three equalities follow from the definition of $\mathbb{Q}^\star$, Theorem 4.1, and the definition of $\psi^\star$, respectively. As the left- and right-hand sides of the preceding expression coincide, we may then conclude that

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

Moreover, as $\mathbb{B}(\widehat{\mathbb{P}}) \subseteq \mathbb{G}(\widehat{\mathbb{P}})$, the preceding relation implies (4.2).

It remains to be shown that $\psi^\star$ and $\mathbb{Q}^\star$ solve the primal and dual Wasserstein MMSE estimation problems (1.7) and (3.1), respectively. As for $\psi^\star$, we have

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

where the inequality holds because $\mathbb{B}(\widehat{\mathbb{P}}) \subseteq \mathbb{G}(\widehat{\mathbb{P}})$. The first equality follows from the definition of $\psi^\star$, and the second equality exploits Theorem 4.1, which implies that all inequalities in (4.1) are, in fact, equalities. This reasoning shows that $\psi^\star$ is optimal in (1.7). The optimality of $\mathbb{Q}^\star$ in (3.1) can be proved similarly. □

Corollary 4.1 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 argue that $\psi^\star$ cannot only be constructed from the solution of the convex program (2.2), which is equivalent to the Gelbrich MMSE estimation problem (2.1), but also from the solution of the convex program (3.4), which is equivalent to the dual MMSE estimation problem (3.3) over normal priors. This alternative construction is useful because Problem (3.4) is amenable to highly efficient first-order methods derived in Section 6.

Corollary 4.2 (Dual Construction of the Optimal Estimator). *If $\widehat{\mathbb{P}}$ is a normal distribution of the form (3.2) with $\widehat{\Sigma}_w \succ 0$, and $(\Sigma_x^\star, \Sigma_w^\star)$ is a maximizer of (3.4), then the affine estimator $\psi^\star(y) = A^\star y + b^\star$ with*

$$A^\star = \Sigma_x^\star H^\top \big(H\Sigma_x^\star H^\top + \Sigma_w^\star\big)^{-1} \quad \text{and} \quad b^\star = \widehat{\mu}_x - A^\star(H\widehat{\mu}_x + \widehat{\mu}_w) \tag{4.3}$$

solves the Wasserstein MMSE estimation problem (1.7).

Proof of Corollary 4.2. Define $\psi^\star$ as the affine estimator that solves (2.1) and $\mathbb{Q}^\star$ as the normal distribution that solves (3.3). By Corollary 4.1, the second inequality in (4.2) 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 (4.3). □

5. Nonnormal Nominal Distributions

We first show that the results of Sections 2–4 remain valid if $\widehat{\mathbb{P}}$ is an arbitrary elliptical (but maybe nonnormal) distribution. To this end, we first review some basic results on elliptical distributions.

Definition 5.1 (Elliptical Distributions). The distribution $\mathbb{P}$ of $\xi \in \mathbb{R}^d$ is called elliptical if the characteristic function $\Phi_{\mathbb{P}}(t) = \mathbb{E}_{\mathbb{P}}[\exp(it^\top \xi)]$ of $\mathbb{P}$ is given by $\Phi_{\mathbb{P}}(t) = \exp(it^\top \mu)\phi(t^\top S t)$ 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 $\mathbb{P} = \mathcal{E}_\phi^d(\mu, S)$. The class of all d-dimensional elliptical distributions with characteristic generator ϕ is denoted by $\mathcal{E}_\phi^d$.

The class of elliptical distributions is introduced in Kelker [45] 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 $M_{\mathbb{P}}(t) = \mathbb{E}_{\mathbb{P}}[\exp(t^\top \xi)]$, 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 5.1 is general enough to cover also heavy-tailed distributions with nonzero tail dependence coefficients (Hult and Lindskog [38]). Examples of elliptical distributions include the Laplace, logistic, and t-distribution, etc. Useful theoretical properties of elliptical distributions are discussed in Cambanis et al. [11] and Fang et al. [28]. We also highlight that elliptical distributions are central to a wide spectrum of diverse applications ranging from genomics (Posekany et al. [61]) and medical imaging (Ruttimann et al. [64]) to finance (Jondeau et al. [40, section 6.2.1]), to name a few.

If the dispersion matrix $S \in \mathbb{S}_+^d$ has rank r, then there exists $\Lambda \in \mathbb{R}^{d\times r}$ with $S = \Lambda\Lambda^\top$, and there exists a generalized inverse $\Lambda^{-1} \in \mathbb{R}^{r\times d}$ with $\Lambda^{-1}\Lambda = I_r = \Lambda^\top(\Lambda^{-1})^\top$. One easily verifies that, if $\xi \in \mathbb{R}^d$ follows an elliptical distribution $\mathbb{P} = \mathcal{E}_\phi^d(\mu, S)$, then $\tilde{\xi} = \Lambda^{-1}(\xi - \mu) \in \mathbb{R}^r$ follows the spherically symmetric distribution $\tilde{\mathbb{P}} = \mathcal{E}_\phi^d(0, I_r)$ 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 $\phi(0) = 1$, and 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

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

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 (Cambanis et al. [11, theorem 4]). We focus on elliptical distributions with finite first- and second-order moments (i.e., we only consider characteristic generators with $|\phi'(0)| < \infty$), and we 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; see, for example, Cambanis et al. [11, theorem 1 and corollary 5]. Moreover, the Wasserstein distance between two elliptical distributions with the same characteristic generator equals the Gelbrich distance between their mean vectors and covariance matrices (Gelbrich [32, theorem 2.4]). Thus, Propositions 3.1–3.3 extend verbatim from the class of normal distributions to *any* class of elliptical distributions that share the same characteristic generator. For the sake of brevity, we do not restate these results for elliptical distributions.

This discussion suggests that the results of Sections 2–4 carry over almost verbatim to MMSE estimation problems involving elliptical nominal distributions. In the following we, therefore, assume that

$$\widehat{\mathbb{P}} = \mathcal{E}_\phi^{n+m}(\widehat{\mu}, \widehat{\Sigma}) \qquad \text{with} \qquad \widehat{\mu} = \begin{bmatrix} \widehat{\mu}_x \\ \widehat{\mu}_w \end{bmatrix} \qquad \text{and} \qquad \widehat{\Sigma} = \begin{bmatrix} \widehat{\Sigma}_x & 0 \\ 0 & \widehat{\Sigma}_w \end{bmatrix}, \tag{5.1}$$

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, although 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

$$\mathbb{B}(\widehat{\mathbb{P}}) = \left\{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^{n+m}) : \ \mathbb{E}_{\mathbb{Q}}[xw^\top] = \ \mathbb{E}_{\mathbb{Q}}[x] \cdot \mathbb{E}_{\mathbb{Q}}[w]^\top, \ \mathbb{W}(\mathbb{Q}_x, \widehat{\mathbb{P}}_x) \le \rho_x, \ \ \mathbb{W}(\mathbb{Q}_w, \widehat{\mathbb{P}}_w) \le \rho_w \right\}, \tag{5.2}$$

which relaxes the independence condition in (1.8) and merely requires x and w to be uncorrelated. When using the new ambiguity set (5.2) 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 5.1 (Elliptical Distributions). *Assume that $\widehat{\mathbb{P}}$ is an elliptical distribution of the form (5.1) with characteristic generator ϕ and noise covariance matrix $\widehat{\Sigma}_w \succ 0$, and define the ambiguity set $\mathbb{B}(\widehat{\mathbb{P}})$ as in (5.2). If $(\Sigma_x^\star, \Sigma_w^\star)$ solves the finite convex program (3.3), then the affine estimator $\psi^\star(y) = A^\star y + b^\star$ with*

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

solves the Wasserstein MMSE estimation problem (1.7), and the elliptical distribution

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

solves the dual Wasserstein MMSE estimation problem (3.1). Moreover, $\psi^\star$ and $\mathbb{Q}^\star$ form a Nash equilibrium for the game between the statistician and nature, that is,

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

Proof of Theorem 5.1. The proof replicates the arguments used to establish Theorems 2.1, 3.1, and 4.1 as well as Corollary 4.1 with obvious minor modifications. Details are omitted for brevity. □

Theorem 5.1 asserts that the optimal estimator depends only on the first and second moments of the nominal elliptical 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.

If the nominal distribution fails to be elliptical, the minimum of the Gelbrich MMSE estimation problem (2.1) may strictly exceed the maximum of the dual Wasserstein MMSE estimation problem (3.3) over normal priors. Note that, in this case, the ambiguity set $\mathbb{B}_{\mathcal{N}}(\widehat{\mathbb{P}})$ may even be empty. Moreover, although typically suboptimal for the original Wasserstein MMSE estimation problem (1.7), the usual affine estimator constructed from a solution of the nonlinear SDP (3.4) still solves the Gelbrich MMSE estimation problem (2.1).

Proposition 5.1 (Nonelliptical Nominal Distributions). *Suppose 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 mean vectors $\widehat{\mu}_x$ and $\widehat{\mu}_w$ and covariance matrices $\widehat{\Sigma}_x \succeq 0$ and $\widehat{\Sigma}_w \succ 0$, respectively. Then, the nonlinear SDP (3.4) is solvable, and for any optimal solution $(\Sigma_x^\star, \Sigma_w^\star)$ of (3.4), the affine estimator $\psi^\star(y) = A^\star y + b^\star$ with*

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

solves the Gelbrich MMSE estimation problem (2.1).

Proof of Proposition 5.1. 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 $\widehat{\mathbb{P}}$. As $\widehat{\Sigma}_w \succ 0$, the nonlinear SDP (3.4) is then solvable by virtue of Theorem 3.1. Theorem 4.1 further implies that the first inequality in (4.1) with $\widehat{\mathbb{P}}'$ instead of $\widehat{\mathbb{P}}$ collapses to the equality

$$\inf_{\psi \in \mathcal{A}} \sup_{\mathbb{Q} \in \mathbb{G}(\widehat{\mathbb{P}}')} \mathcal{R}(\psi, \mathbb{Q}) = \inf_{\psi \in \mathcal{F}} \sup_{\mathbb{Q} \in \mathbb{B}(\widehat{\mathbb{P}}')} \mathcal{R}(\psi, \mathbb{Q}). \tag{5.3}$$

In addition, Corollary 4.2 ensures that the affine estimator $\psi^\star$ defined in the proposition statement solves the modified Wasserstein MMSE estimation problem with normal nominal distribution $\widehat{\mathbb{P}}'$ on the right-hand side of (5.3). Because $\psi^\star$ is affine, it is also feasible in the modified Gelbrich MMSE estimation problem on the left-hand side. In addition, the average risk of any affine estimator depends only on the mean vectors and covariance

matrices of x and w. If we denote by T the mean-covariance projection that maps any distribution $\mathbb{Q} = \mathbb{Q}_x \times \mathbb{Q}_w$ of (x, w) to the mean vectors and covariance matrices of x and w under $\mathbb{Q}$, then the images of the ambiguity sets $\mathbb{G}(\widehat{\mathbb{P}}')$ and $\mathbb{B}(\widehat{\mathbb{P}}')$ under T coincide by Proposition 3.3. These observations imply that the affine estimator $\psi^\star$ also solves the estimation problem on the left-hand side of (5.3). As $\widehat{\mathbb{P}}$ and $\widehat{\mathbb{P}}'$ share the same first and second moments, the Gebrich ball $\mathbb{G}(\widehat{\mathbb{P}})$ around the generic distribution $\widehat{\mathbb{P}}$ coincides with the Gelbrich ambiguity set $\mathbb{G}(\widehat{\mathbb{P}}')$ around the normal distribution $\widehat{\mathbb{P}}'$. Thus, we find

$$\sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathcal{R}(\psi^\star, \mathbb{Q}) = \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}}')} \mathcal{R}(\psi^\star, \mathbb{Q}) = \inf_{\psi\in\mathcal{A}} \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}}')} \mathcal{R}(\psi, \mathbb{Q}) = \inf_{\psi\in\mathcal{A}} \sup_{\mathbb{Q}\in\mathbb{G}(\widehat{\mathbb{P}})} \mathcal{R}(\psi, \mathbb{Q}),$$

where the second equality holds because $\psi^\star$ solves the Gelbrich MMSE estimation problem with the normal nominal distribution $\widehat{\mathbb{P}}'$ on the left-hand side of (5.3). Hence, $\psi^\star$ solves the Gelbrich MMSE estimation problem (2.1) with the generic nominal distribution $\widehat{\mathbb{P}}$.

6. Numerical Solution of Wasserstein MMSE Estimation Problems

By Corollaries 2.2 and 3.1, the primal and dual Wasserstein MMSE estimation problems (1.7) and (3.1) can be addressed with off-the-shelf SDP solvers. Unfortunately, however, general-purpose interior-point methods quickly run out of memory when the signal dimension n and the noise dimension m grow. It is, therefore, expedient to look for customized first-order algorithms that can handle larger problem instances.

In this section, we develop a Frank–Wolfe method for the nonlinear SDP (3.4), which is equivalent to the dual Wasserstein MMSE estimation problem (3.1). This approach is meaningful because any solution to (3.4) allows us to construct both an optimal estimator as well as a least favorable prior that form a Nash equilibrium in the sense of Corollary 4.1; see also Corollary 4.2. Addressing the nonlinear SDP (3.4) directly with a Frank–Wolfe method has great promise because the subproblems that identify the local search directions can be shown to admit quasi-closed form solutions and can, therefore, be solved very quickly.

In Section 6.1, we first review three variants of the Frank–Wolfe algorithm corresponding to a static, an adaptive, and a more flexible *fully* adaptive step-size rule, and we prove that the fully adaptive rule offers a linear convergence guarantee under standard regularity conditions. In Section 6.2, we then show that the nonlinear SDP (3.4) is amenable to the fully adaptive Frank–Wolfe algorithm and can, thus, be solved efficiently.

6.1. Frank–Wolfe Algorithm for Generic Convex Optimization Problems

Consider a generic convex minimization problem of the form

$$f^\star = \min_{s\in\mathcal{S}} \; f(s) \tag{6.1}$$

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

$$(F(s) - s)^\top \nabla f(s) \leq \delta \min_{z\in\mathcal{S}} \; (z - s)^\top \nabla f(s). \tag{6.2}$$

By the standard optimality condition for convex optimization problems, the minimum on the right-hand side of (6.2) vanishes if and only if s solves the original problem (6.1). 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 (6.2), one can design a Frank–Wolfe algorithm whose iterates obey the recursion

$$s_{t+1} = s_t + \eta_t(F(s_t) - s_t) \quad \forall t \in \mathbb{N} \bigcup \{0\}, \tag{6.3}$$

where $s_0 \in \mathcal{S}$ is an arbitrary initial feasible solution, δ is a prescribed precision, and $\eta_t \in [0, 1]$ is a step size that may depend on the current iterate s_t. The Frank–Wolfe algorithm was originally developed for quadratic programs (Frank and Wolfe [29]) and later extended to general convex programs with differentiable objective functions and compact convex feasible sets (Demyanov and Rubinov [17], Dunn [20, 21], Dunn and Harshbarger [22], Levitin and Polyak [47]). Convergence guarantees for the Frank–Wolfe algorithm typically rely on the assumption that the gradient of f is Lipschitz continuous (Dunn [20, 21], Freund and Grigas [30], Garber and Hazan [31], Levitin and Polyak [47]), that f has a bounded curvature constant (Clarkson [14], Jaggi [39]), or that the gradient of f is Hölder continuous (Nesterov [54]).

Throughout this section, we assume that the decision variable can be represented as $s = (s^{[1]}, \ldots, s^{[K]})$, where $s^{[k]} \in \mathbb{R}^{d_k}$ and $\sum_{k=1}^K d_k = d$. Moreover, we assume that the feasible set $\mathcal{S} = \times_{k=1}^K \mathcal{S}^{[k]}$ constitutes a K-fold Cartesian

product, for which the marginal feasible set $\mathcal{S}^{[k]} \subseteq \mathbb{R}^{d_k}$ is convex and compact for each $k = 1, \ldots, K$. This assumption is unrestrictive because we are free to set $K = 1$ and $\mathcal{S}^{[1]} = \mathcal{S}$. For ease of notation, we use from now on $\nabla_{[k]}$ to denote the partial gradient with respect to the subvector $s^{[k]} \in \mathcal{S}^{[k]}$, $k = 1, \ldots, K$.

The subsequent convergence analysis relies on the following regularity conditions.

Assumption 6.1 (Regularity Conditions).

i. *The objective function is β-smooth for some $\beta > 0$, that is,*

$$\|\nabla f(s) - \nabla f(\bar{s})\| \leq \beta \|s - \bar{s}\| \quad \forall s, \bar{s} \in \mathcal{S}.$$

ii. *The marginal feasible sets are α-strongly convex with respect to f for some $\alpha > 0$, that is,*

$$\theta s^{[k]} + (1-\theta)\bar{s}^{[k]} - \theta(1-\theta)\frac{\alpha}{2}\left\| s^{[k]} - \bar{s}^{[k]} \right\|^2 \frac{\nabla_{[k]} f(s)}{\|\nabla_{[k]} f(s)\|} \in \mathcal{S}^{[k]} \quad \forall s, \bar{s} \in \mathcal{S},\ \theta \in [0,1],\ k = 1, \ldots, K.$$

iii. *The objective function is ε-steep for some $\varepsilon > 0$, that is,*

$$\|\nabla_{[k]} f(s)\| \geq \varepsilon \quad \forall s \in \mathcal{S},\ k = 1, \ldots, K.$$

Assumption 6.1(ii) relaxes the standard strong convexity condition prevailing in the literature, which is obtained by setting $K = 1$ and requiring that the condition stated here remains valid when the normalized gradient $\nabla_{[1]} f(s)/\|\nabla_{[1]} f(s)\|$ is replaced with any other vector in the Euclidean unit ball; see, for example, Journée et al. [41, equation (25)]. We emphasize that our weaker condition is sufficient for the standard convergence proofs of the Frank–Wolfe algorithm but is necessary for our purposes because the feasible set of Problem (3.4) fails to be strongly convex in the traditional sense. Similarly, Assumption 6.1(iii) generalizes the usual ε-steepness condition from the literature, which is recovered by setting $K = 1$; see, for example, Journée et al. [41, assumption 1]. Under this assumption, the gradient never vanishes on $\mathcal{S}$, and the minimum of (6.1) is attained on the boundary of $\mathcal{S}$.

In the following, we distinguish three variants of the Frank–Wolfe algorithm with different step-size rules. The *vanilla Frank-Wolfe* algorithm employs the harmonically decaying static step size

$$\eta_t = \frac{2}{2+t},$$

which results in a sublinear $\mathcal{O}(1/t)$ convergence whenever Assumption 6.1(i) holds (Dunn and Harshbarger [22], Frank and Wolfe [29]). The *adaptive Frank-Wolfe* algorithm uses the step size

$$\eta_t = \min\left\{1, \frac{(s_t - F(s_t))^\top \nabla f(s_t)}{\beta \|s_t - F(s_t)\|^2}\right\}, \tag{6.4}$$

which adapts to the iterate s_t. If all of Assumption 6.1(i–iii) holds, then the adaptive Frank–Wolfe algorithm enjoys a linear $\mathcal{O}(c^t)$ convergence guarantee, with which $c \in (0,1)$ is an explicit function of the oracle precision δ, the smoothness parameter β, the strong convexity parameter α, and the steepness parameter ε (Garber and Hazan [31], Levitin and Polyak [47]). Note that the step size (6.4) is constructed as the unique solution of the univariate quadratic program

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

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

The adaptive step-size rule (6.4) undergoes further scrutiny in Pedregosa et al. [59], where it is discovered that one may improve the algorithm's convergence behavior by replacing the global smoothness parameter β in (6.4) with an adaptive smoothness parameter β_t that captures the smoothness of f along the line segment from s_t to $F(s_t)$. This extra flexibility is useful because β_t can be chosen smaller than the unnecessarily conservative global smoothness parameter β and because β_t is easier to estimate than β, which may not even be accessible.

Following Pedregosa et al. [59], we henceforth only require that $\beta_t > 0$ satisfies the inequality

$$f(s_t - \eta_t(\beta_t)(s_t - F(s_t))) \leq f(s_t) - \eta_t(\beta_t)(s_t - F(s_t))^\top \nabla f(s_t) + \frac{1}{2}\beta_t \eta_t(\beta_t)^2 \|s_t - F(s_t)\|^2, \tag{6.5}$$

where $\eta_t(\beta_t)$ is defined as the adaptive step size (6.4) with β replaced by β_t. As it adapts to both s_t and β_t, we from now on refer to $\eta_t = \eta_t(\beta_t)$ as the *fully* adaptive step size. This discussion implies that (6.5) is always satisfiable if Assumption 6.1(i) holds, in which case one may simply set β_t to the global smoothness parameter β. In practice, however, Inequality (6.5) is often satisfiable for much smaller values $\beta_t \ll \beta$ that may not even be related to the

smoothness properties of the objective function. A close upper bound on the smallest $\beta_t > 0$ that satisfies (6.5) can be found efficiently via backtracking line search. Specifically, the *fully adaptive Frank-Wolfe* algorithm sets β_t to the smallest element of the discrete search space $\frac{\beta_{t-1}}{\zeta} \cdot \{1, \tau, \tau^2, \tau^3, \dots\}$ that satisfies (6.5), 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 1.

Algorithm 1 (Fully Adaptive Frank-Wolfe Algorithm)

Input: initial feasible point $s_0 \in \mathcal{S}$, initial smoothness parameter $\beta_{-1} > 0$, line search parameters $\tau > 1$, $\zeta > 1$, initial iteration counter $t = 0$

while stopping criterion is not met **do**
  solve the oracle subproblem to find $\tilde{s}_t = F(s_t)$
  set $d_t \leftarrow \tilde{s}_t - s_t$ and $g_t \leftarrow -d_t^\top \nabla f(s_t)$
  set $\beta_t \leftarrow \beta_{t-1}/\zeta$ and $\eta \leftarrow \min\{1, g_t/(\beta_t \|d_t\|^2)\}$
  while $f(s_t + \eta d_t) > f(s_t) - \eta g_t + \frac{\eta^2 \beta_t}{2}\|d_t\|^2$ **do**
    $\beta_t \leftarrow \tau\beta_t$ and $\eta \leftarrow \min\{1, g_t/(\beta_t\|d_t\|^2)\}$
  end while
  set $\eta_t \leftarrow \eta$ and $s_{t+1} \leftarrow s_t + \eta_t d_t$
  set $t \leftarrow t + 1$
end while
Output: s_t

It is shown in Pedregosa et al. [59] that Algorithm 1 enjoys the same sublinear $\mathcal{O}(1/t)$ convergence guarantee as the vanilla Frank–Wolfe algorithm when Assumption 6.1(i) holds. We leverage techniques from Garber and Hazan [31] and Levitin and Polyak [47] to show that Algorithm 1 offers indeed a linear convergence rate if all of Assumption 6.1(i)–(iii) holds.

Theorem 6.1 (Linear Convergence of the Fully Adaptive Frank-Wolfe Algorithm). *If Assumption 6.1 holds and* $\bar{\beta} = \max\{\tau\beta, \beta_{-1}\}$, *then Algorithm 1 enjoys the linear convergence guarantee*

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

The proof of Theorem 6.1 relies on the following preparatory lemma.

Lemma 6.1 (Bounds on the Surrogate Duality Gap). *The surrogate duality gap* $g_t = -d_t^\top \nabla f(s_t)$ *corresponding to the search direction* $d_t = F(s_t) - s_t$ *admits the following lower bounds.*

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

ii. *If the marginal feasible sets are* α*-strongly convex with respect to f for some* $\alpha > 0$ *in the sense of Assumption 6.1*(ii), *then*

$$g_t \ge \min_{k \in \{1,\dots,K\}} \frac{(1-\sqrt{1-\delta})\alpha}{2\delta}\|d_t\|^2 \|\nabla_{[k]} f(s_t)\|.$$

Proof of Lemma 6.1. By the definition of g_t, we have

$$g_t = (s_t - F(s_t))^\top \nabla f(s_t) \ge \delta(s_t - s)^\top \nabla f(s_t) \qquad \forall s \in \mathcal{S}, \tag{6.6}$$

where the inequality follows from the defining property (6.2) of the inexact oracle with precision δ. Setting s in (6.6) to a global minimizer $s^\star$ of (6.1) then implies via the first-order convexity condition for f that

$$g_t \ge \delta(s_t - s^\star)^\top \nabla f(s_t) \ge \delta(f(s_t) - f^\star).$$

This observation establishes assertion (i). To prove assertion (ii), we first rewrite the estimate (6.6) as

$$g_t \ge \delta(s_t - s)^\top \nabla f(s) = \delta \sum_{t=1}^{T} (s_t^{[k]} - s^{[k]})^\top \nabla_{[k]} f(s_t) \quad \forall s \in \mathcal{S}. \tag{6.7}$$

In the following, we denote by $F^{[k]} : \mathcal{S} \to \mathcal{S}^{[k]}$ the kth suboracle for $k = 1, \dots, K$, which is defined through the identity $F = (F^{[1]}, \dots, F^{[K]})$. Similarly, for any $\theta \in [0,1]$, we define $s(\theta) = (s^{[1]}(\theta), \dots, s^{[K]}(\theta))$ through

$$s^{[k]}(\theta) = \theta F^{[k]}(s_t) + (1-\theta)s_t^{[k]} - \frac{\alpha}{2}\theta(1-\theta)\left\|F^{[k]}(s_t) - s_t^{[k]}\right\|^2 \frac{\nabla_{[k]} f(s_t)}{\|\nabla_{[k]} f(s_t)\|}.$$

By Assumption 6.1(ii), we have $s^{[k]}(\theta) \in \mathcal{S}^{[k]}$ for every $k = 1, \ldots, K$. Thanks to the rectangularity of the feasible set, this implies that $s(\theta) \in \mathcal{S}$. Setting s in (6.7) to $s(\theta)$, we, thus, find

$$\begin{aligned} g_t &\geq \delta \sum_{t=1}^{T} \left(\theta(s_t^{[k]} - F^{[k]}(s_t)) + \frac{\alpha}{2}\theta(1-\theta)\left\|F^{[k]}(s_t) - s_t^{[k]}\right\|^2 \frac{\nabla_{[k]} f(s_t)}{\|\nabla_{[k]} f(s_t)\|} \right)^\top \nabla_{[k]} f(s_t) \\ &= \delta\left(\theta g_t + \frac{\alpha}{2}\theta(1-\theta)\left[\sum_{k=1}^{K} \left\|F^{[k]}(s_t) - s_t^{[k]}\right\|^2 \|\nabla_{[k]} f(s_t)\| \right] \right) \\ &\geq \delta\left(\theta g_t + \frac{\alpha}{2}\theta(1-\theta)\|F(s_t) - s_t\|^2 \left(\min_{k\in\{1,\ldots,K\}} \|\nabla_{[k]} f(s_t)\| \right) \right) \quad \forall \theta \in [0,1], \end{aligned}$$

where the equality follows from the definition of g_t and the last inequality exploits the Pythagorean theorem. Reordering this inequality to bring g_t to the left-hand side yields

$$g_t \geq \min_{k\in\{1,\ldots,K\}} \frac{\alpha}{2}\|F(s_t) - s_t\|^2 \|\nabla_{[k]} f(s_t)\| \frac{\delta\theta(1-\theta)}{1-\delta\theta} \qquad \forall \theta \in [0,1]. \tag{6.8}$$

A tedious but straightforward calculation shows that the lower bound on the right-hand side of (6.8) is maximized by $\theta^\star = (1 - \sqrt{1-\delta})/\delta$. Assertion (ii) then follows by substituting $\theta^\star$ into (6.8). □

Proof of Theorem 6.1. By Assumption 6.1(i), the function f is β-smooth, and thus, one can show that

$$f(s_t + \eta d_t) \leq f(s_t) - \eta g_t + \frac{\eta^2\beta}{2}\|d_t\|^2 \qquad \forall \eta \in [0,1], \tag{6.9}$$

where the surrogate duality gap $g_t \geq 0$ and the search direction $d_t \in \mathbb{R}^d$ are defined as in Lemma 6.1. We emphasize that (6.9) holds, in fact, for all $\eta \in \mathbb{R}$. However, the next iterate $s_{t+1} = s_t + \eta d_t$ may be infeasible unless $\eta \in [0,1]$. Inequality (6.9) implies that any $\beta_t \geq \beta$ satisfies the condition of the inner while loop of Algorithm 1, 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_t and a step size η_t that satisfy Inequality (6.5). We henceforth denote by $h_t = f(s_t) - f^\star$ the suboptimality of the kth iterate and note that

$$h_{t+1} = f(s_t + \eta_t d_t) - f(s_t) + h_t \leq -g_t + \frac{1}{2}\beta_t \eta_t^2 \|d_t\|^2 + h_t, \tag{6.10}$$

where the inequality exploits (6.5) and the definitions of g_t and d_t. In order to show that h_t decays geometrically, we distinguish cases (i) $g_t/(\beta_t\|d_t\|^2) \geq 1$ and (ii) $g_t/(\beta_t\|d_t\|^2) < 1$. In case (i), the step size η_t defined in (6.5) satisfies $\eta_t = \min\{1, g_t/(\beta_t\|d_t\|^2)\} = 1$, and thus, we have

$$h_{t+1} \leq \left(\frac{\beta_t\|d_t\|^2}{2g_t} - 1 \right) g_t + h_t \leq -\frac{g_t}{2} + h_t \leq \left(1 - \frac{\delta}{2}\right) h_t, \tag{6.11}$$

where the first inequality follows from (6.10), and the third inequality holds because of Lemma 6.1(i).

In case (ii), the step size satisfies $\eta_t = g_t/\beta_t\|d_t\|^2 < 1$, and thus, we find

$$\begin{aligned} h_{t+1} &\leq -g_t + \frac{g_t^2}{2\beta_t\|d_t\|^2} + h_t \leq -\frac{g_t^2}{2\beta_t\|d_t\|^2} + h_t \leq \left(1 - \frac{\delta g_t}{2\beta_t\|d_t\|^2}\right) h_t \\ &\leq \left(1 - \min_{k\in\{1,\ldots,K\}} \frac{(1-\sqrt{1-\delta})\alpha}{4\beta_t} \|\nabla_{[k]} f(s_t)\|\right) h_t \leq \left(1 - \frac{(1-\sqrt{1-\delta})\alpha\varepsilon}{4\bar{\beta}}\right) h_t, \end{aligned} \tag{6.12}$$

where the first and second inequalities follow from (6.10) and from multiplying $-g_t$ with $\eta_t < 1$, respectively, and the third and the fourth inequalities exploit Lemma 6.1(i) and (ii), respectively. The last inequality in (6.12) holds because of Assumption 6.1(iii) and because $\beta_t \leq \bar{\beta}$ for all $t \in \mathbb{N}$; see Pedregosa et al. [59, proposition 2]. By Estimates (6.11) and (6.12), the suboptimality of the current iterate decays at least by

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

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

6.2. Frank-Wolfe Algorithm for Wasserstein MMSE Estimation Problems

We now use the fully adaptive Frank–Wolfe algorithm of Section 6.1 to solve the nonlinear SDP (3.4), which is equivalent to the dual Wasserstein MMSE estimation problem over normal priors. Recall from Corollary 4.2 that any solution of (3.4) can be used to construct a least favorable prior and an optimal estimator that form a Nash equilibrium. Unlike the generic convex program (6.1), the nonlinear SDP (3.4) is a convex *maximization* problem. This prompts us to apply Algorithm 1 to the convex minimization problem obtained from Problem (3.4) by turning the objective function upside down.

Throughout this section, we assume that $\widehat{\Sigma}_x \succ 0$, $\widehat{\Sigma}_w \succ 0$, $\rho_x > 0$ and $\rho_w > 0$, which implies via Theorem 3.1 that the nonlinear SDP (3.4) is solvable and can be reformulated more concisely as

$$\max_{\Sigma_x \in \mathcal{S}_x^+,\, \Sigma_w \in \mathcal{S}_w^+} f(\Sigma_x, \Sigma_w)\,, \tag{6.13}$$

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

$$f(\Sigma_x, \Sigma_w) = \mathrm{Tr}[\Sigma_x - \Sigma_x H^\top (H\Sigma_x H^\top + \Sigma_w)^{-1} H \Sigma_x]\,,$$

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

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

and

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

respectively. One readily verifies that f is concave and differentiable. Moreover, in the terminology of Section 6.1, the feasible set of the nonlinear SDP (6.13) constitutes a Cartesian product of $K = 2$ marginal feasible sets $\mathcal{S}_x^+$ and $\mathcal{S}_w^+$, both of which are convex and compact thanks to Lemma A.4. Note that $\mathcal{S}_x^+$ and $\mathcal{S}_w^+$ constitute restrictions of the feasible sets $\mathcal{S}_x$ and $\mathcal{S}_w$, respectively, which appear in the proofs of Theorems 3.1 and 4.1. The oracle problem that linearizes the objective function of the nonlinear SDP (6.13) around a fixed feasible solution $\Sigma_x \in \mathcal{S}_x^+$ and $\Sigma_w^+ \in \mathcal{S}_w$ can now be expressed concisely as

$$\max_{L_x \in \mathcal{S}_x^+,\, L_w \in \mathcal{S}_w^+} \langle L_x - \Sigma_x, D_x \rangle + \langle L_w - \Sigma_w, D_w \rangle\,, \tag{6.14}$$

where $D_x = \nabla_{\Sigma_x} f(\Sigma_x, \Sigma_w)$ and $D_w = \nabla_{\Sigma_w} f(\Sigma_x, \Sigma_w)$ represent the gradients of f with respect to Σ_x and Σ_w. Lemma A.5 offers analytical formulas for D_x and D_w and shows that they are both positive semidefinite.

The oracle problem (6.14) is manifestly separable in L_x and L_w and can, therefore, be decomposed into a sum of two structurally identical marginal subproblems. The Frank–Wolfe algorithm is an ideal method to address the nonlinear SDP (6.13) because these two marginal oracle subproblems admit quasi-closed form solutions. Specifically, Proposition A.2(iii) implies that Problem (6.14) is uniquely solved by

$$L_x^\star = (\gamma_x^\star)^2 (\gamma_x^\star I_n - D_x)^{-1} \widehat{\Sigma}_x (\gamma_x^\star I_n - D_x)^{-1} \quad \text{and} \quad L_w^\star = (\gamma_w^\star)^2 (\gamma_w^\star I_m - D_w)^{-1} \widehat{\Sigma}_w (\gamma_w^\star I_m - D_w)^{-1},$$

where $\gamma_x^\star \in (\lambda_{\max}(D_x), \infty)$ and $\gamma_w^\star \in (\lambda_{\max}(D_w), \infty)$ are the unique solutions of the algebraic equations

$$\rho_x^2 - \langle \widehat{\Sigma}_x, (I_n - \gamma_x^\star (\gamma_x^\star I_n - D_x)^{-1})^2 \rangle = 0 \quad \text{and} \quad \rho_w^2 - \langle \widehat{\Sigma}_w, (I_m - \gamma_w^\star (\gamma_w^\star I_m - D_w)^{-1})^2 \rangle = 0, \tag{6.15}$$

respectively. In practice, these algebraic equations need to be solved numerically. However, the numerical errors in $\gamma_x^\star$ and $\gamma_w^\star$ must be contained to ensure that $L_x^\star$ and $L_w^\star$ give rise to a δ-approximate solution for (6.14) in the sense of (6.2). In the following, we show that δ-approximate solutions for each of the two oracle subproblems in (6.14) and for each $\delta \in (0, 1)$ can be computed with an efficient bisection algorithm.

Theorem 6.2 (Approximate Oracle). *For any fixed $\rho \in \mathbb{R}_{++}$, $\widehat{\Sigma} \in \mathbb{S}_{++}^d$ and $D \in \mathbb{S}_+^d$, $D \neq 0$, consider the generic oracle subproblem*

$$\begin{array}{ll} \max\limits_{L \in \mathbb{S}_+^d} & \langle L - \Sigma, D \rangle \\ \text{s.t.} & \mathrm{Tr}[L + \widehat{\Sigma} - 2(\widehat{\Sigma}^{\frac{1}{2}} L \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}] \le \rho^2,\ L \succeq \lambda_{\min}(\widehat{\Sigma}) I_d\,, \end{array} \tag{6.16}$$

where $\Sigma \in \mathbb{S}_+^d$ represents a feasible reference solution. Moreover, denote the feasible set of Problem (6.16) by $\mathcal{S}^+$, let $\delta \in (0, 1)$ be the desired oracle precision, and define $\varphi(\gamma) = \gamma(\rho^2 + \langle \gamma(\gamma I_d - D)^{-1} - I_d, \widehat{\Sigma} \rangle) - \langle \Sigma, D \rangle$ for any $\gamma > \lambda_{\max}(D)$. Then, Algorithm 2 returns in finite time a matrix $\tilde{L} \in \mathbb{S}_+^d$ with the following properties.

i. *Feasibility:* $\tilde{L} \in \mathcal{S}^+$
ii. δ*-Suboptimality:* $\langle \tilde{L} - \Sigma, D\rangle \geq \delta \max_{L\in\mathcal{S}^+} \langle L - \Sigma, D\rangle$

Algorithm 2 (Bisection Algorithm for the Oracle Subproblem)
Input: nominal covariance matrix $\widehat{\Sigma} \in \mathbb{S}_{++}^d$, radius $\rho \in \mathbb{R}_{++}$, reference covariance matrix $\Sigma \in \mathbb{S}_+^d$ feasible in (6.14), gradient matrix $D \in \mathbb{S}_+^d$, $D \neq 0$, precision $\delta \in (0,1)$, dual objective function $\varphi(\gamma)$ defined in Theorem 6.2
set $\lambda_1 \leftarrow \lambda_{\max}(D)$, and let $v_1 \in \mathbb{R}^d$ be an eigenvector for λ_1
set $\underline{\gamma} \leftarrow \lambda_1(1 + (v_1^\top \widehat{\Sigma} v_1)^{\frac{1}{2}}/\rho)$ and $\bar{\gamma} \leftarrow \lambda_1(1 + \mathrm{Tr}[\widehat{\Sigma}]^{\frac{1}{2}}/\rho)$
repeat
Set $\tilde{\gamma} \leftarrow (\bar{\gamma} + \underline{\gamma})/2$ and $\tilde{L} \leftarrow (\tilde{\gamma})^2(\tilde{\gamma} I_d - D)^{-1}\widehat{\Sigma}(\tilde{\gamma} I_d - D)^{-1}$
if $\frac{\mathrm{d}\varphi}{\mathrm{d}\gamma}(\tilde{\gamma}) < 0$, **then** set $\underline{\gamma} \leftarrow \tilde{\gamma}$ **else** Set $\bar{\gamma} \leftarrow \tilde{\gamma}$ **end if**
until $\frac{\mathrm{d}\varphi}{\mathrm{d}\gamma}(\tilde{\gamma}) > 0$ and $\langle \tilde{L} - \Sigma, D\rangle \geq \delta\, \varphi(\tilde{\gamma})$
Output: $\tilde{L}$

Proof of Theorem 6.2. Proposition A.2(iii) guarantees that the lower bound on L in (6.16) is redundant and can be omitted without affecting the problem's optimal value. By Proposition A.2(i), the oracle subproblem (6.16), thus, admits the strong Lagrangian dual

$$\min_{\gamma > \lambda_{\max}(D)} \varphi(\gamma),$$

where the convex and differentiable function $\varphi(\gamma)$ is defined as in the theorem statement. In the following, we denote by $\lambda_1 > 0$ the largest eigenvalue of D and let $v_1 \in \mathbb{R}^d$ be a corresponding eigenvalue. By Proposition A.2(iii), the dual oracle subproblem admits a minimizer $\gamma^\star \in [\underline{\gamma}, \bar{\gamma}]$, that is, uniquely determined by the first-order optimality condition $\frac{\mathrm{d}\varphi}{\mathrm{d}\gamma}(\gamma^\star) = 0$, where

$$\underline{\gamma} = \lambda_1\left(1 + \sqrt{v_1^\top \widehat{\Sigma} v_1}/\rho\right) \quad \text{and} \quad \bar{\gamma} = \lambda_1\left(1 + \sqrt{\mathrm{Tr}[\widehat{\Sigma}]}/\rho\right),$$

and the primal problem (6.16) admits a unique maximizer $L^\star = L(\gamma^\star)$, where

$$L(\gamma) = \gamma^2(\gamma I_d - D)^{-1}\widehat{\Sigma}(\gamma I_d - D)^{-1}.$$

From the proof of Proposition A.2(iii), it is evident that $L(\gamma) \succ \lambda_{\min}(\widehat{\Sigma}) I_d$ for every $\gamma > 0$. A direct calculation further shows that

$$\frac{\mathrm{d}\varphi}{\mathrm{d}\gamma}(\gamma) = \rho^2 - \langle \widehat{\Sigma}, (I_d - \gamma(\gamma I_d - D)^{-1})^2\rangle = \rho^2 - \mathrm{Tr}[L(\gamma) + \widehat{\Sigma} - 2(\widehat{\Sigma}^{\frac{1}{2}} L(\gamma) \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}].$$

Recalling that $\varphi(\gamma)$ is convex and that its derivative is nonnegative for all $\gamma \geq \gamma^\star$, the reasoning implies that $L(\gamma) \in \mathcal{S}^+$ for all $\gamma \geq \gamma^\star$. Note also that the optimal value of the primal problem (6.16) is nonnegative because $\Sigma \in \mathcal{S}^+$. The continuity of $\langle L(\gamma) - \Sigma, D\rangle$ at $\gamma = \gamma^\star$, thus, ensures that there exists $\delta' > 0$ with

$$\langle L(\gamma) - \Sigma, D\rangle \leq \delta \langle L(\gamma^\star) - \Sigma, D\rangle = \delta \max_{L\in\mathcal{S}^+} \langle L - \Sigma, D\rangle \quad \forall \gamma \in [\gamma^\star, \gamma^\star + \delta'].$$

In summary, computing a feasible and δ-suboptimal matrix $\tilde{L} \in \mathbb{S}_+^d$ is tantamount to finding $\tilde{\gamma} \in [\gamma^\star, \gamma^\star + \delta']$. Algorithm 2 uses bisection over the interval $[\underline{\gamma}, \bar{\gamma}]$ to find a $\tilde{\gamma}$ with these properties. □

Theorem 6.2 complements Shafieezadeh-Abadeh et al. [67, theorem 3.2], which constructs an approximate oracle for a nonlinear SDP similar to (6.13) that offers an *additive* error guarantee. The *multiplicative* error guarantee of the oracle constructed here is needed to ensure the linear convergence of the fully adaptive Frank–Wolfe algorithm. Next, we prove that the nonlinear SDP (6.13) satisfies all regularity conditions listed in Assumption 6.1.

Proposition 6.1 (Regularity Conditions of the Nonlinear SDP (6.13)). *If* $\rho_x \in \mathbb{R}_{++}$, $\rho_w \in \mathbb{R}_{++}$, $\widehat{\Sigma}_x \in \mathbb{S}_+^n$ *and* $\widehat{\Sigma}_w \in \mathbb{S}_+^m$, *then the nonlinear SDP* (6.13) *obeys the following regularity conditions.*

i. *The objective function of Problem* (6.13) *is* β*-smooth in the sense of Assumption* 6.1(i), *in which*

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

which depends on the auxiliary constant $C = \lambda_{\max}(H^\top H) \cdot \lambda_{\min}^{-2}(\widehat{\Sigma}_w) \cdot (\rho_x + \mathrm{Tr}[\widehat{\Sigma}_x]^{\frac{1}{2}})^4$.

ii. *The marginal feasible sets $\mathcal{S}_x^+$ and $\mathcal{S}_w^+$ of Problem (6.13) are α-strongly convex with respect to $-f$ in the sense of Assumption 6.1(ii), in which $\alpha = \min\{\alpha_x, \alpha_w\}$, which depends on the auxiliary constants*

$$\alpha_x = \frac{\lambda_{\min}^{\frac{5}{4}}(\widehat{\Sigma}_x)}{2\rho_x(\rho_x + \mathrm{Tr}[\widehat{\Sigma}_x]^{\frac{1}{2}})^{\frac{7}{2}}} \quad \text{and} \quad \alpha_w = \frac{\lambda_{\min}^{\frac{5}{4}}(\widehat{\Sigma}_w)}{2\rho_w(\rho_w + \mathrm{Tr}[\widehat{\Sigma}_w]^{\frac{1}{2}})^{\frac{7}{2}}}.$$

iii. *The objective function of Problem (6.13) is ε-steep in the sense of Assumption 6.1(iii), in which $\varepsilon = \min\{\varepsilon_x, \varepsilon_w\}$, which depends on the auxiliary constants*

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

and

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

Proof of Proposition 6.1. The proof repeatedly uses the fact that, for any $A \in \mathbb{R}^{d_1 \times d_2}$ and $B \in \mathbb{S}_+^{d_2}$, we have

$$\lambda_{\max}(ABA^\top) = \lambda_{\max}(A^\top AB) \le \lambda_{\max}(A^\top A)\,\lambda_{\max}(B). \tag{6.17}$$

The equality in (6.17) holds because all eigenvalues of $ABA^\top$ are nonnegative and because the nonzero spectrum of $ABA^\top$ is identical to that of $A^\top AB$ because of Bernstein [7, proposition 4.4.10]. The inequality follows from the observation that $\lambda_{\max}(A^\top A)$ and $\lambda_{\max}(B)$ coincide with the operator norms of the positive semidefinite matrices $A^\top A$ and B, respectively.

As for assertion (i), recall first that the objective function f of the nonlinear SDP (3.4) is concave. In order to show that f is β-smooth for some $\beta > 0$, it, thus, suffices to prove that the largest eigenvalue of the positive semidefinite Hessian matrix of $-f$ admits an upper bound uniformly across $\mathcal{S}_x^+ \times \mathcal{S}_w^+$. By Lemma A.5, the partial gradients of f evaluated at $\Sigma_x \succ 0$ and $\Sigma_w \succ 0$ are given by

$$\begin{aligned} D_x &= \nabla_{\Sigma_x} f(\Sigma_x, \Sigma_w) = (I_n - \Sigma_x H^\top G^{-1} H)^\top (I_n - \Sigma_x H^\top G^{-1} H) \\ D_w &= \nabla_{\Sigma_w} f(\Sigma_x, \Sigma_w) = G^{-1} H \Sigma_x^2 H^\top G^{-1}, \end{aligned}$$

where $G = H\Sigma_x H^\top + \Sigma_w$. Moreover, the Hessian matrix

$$\mathcal{H} = \begin{bmatrix} \mathcal{H}_{xx} & \mathcal{H}_{xw} \\ \mathcal{H}_{xw}^\top & \mathcal{H}_{ww} \end{bmatrix} \succeq 0$$

of the convex function $-f$ evaluated at $\Sigma_x \succ 0$ and $\Sigma_w \succ 0$ consists of the submatrices

$$\begin{aligned} \mathcal{H}_{xx} &= -\nabla_{xx}^2 f(\Sigma_x, \Sigma_w) &&= 2D_x \otimes H^\top G^{-1} H \\ \mathcal{H}_{xw} &= -\nabla_{xw}^2 f(\Sigma_x, \Sigma_w) &&= H^\top G^{-1} \otimes (H^\top D_w - \Sigma_x H^\top G^{-1}) + (H^\top D_w - \Sigma_x H^\top G^{-1}) \otimes H^\top G^{-1} \\ \mathcal{H}_{ww} &= -\nabla_{ww}^2 f(\Sigma_x, \Sigma_w) &&= 2D_w \otimes G^{-1}, \end{aligned}$$

where ∇_x and ∇_w are used as shorthand for the nabla operators with respect to $\mathrm{vec}(\Sigma_x)$ and $\mathrm{vec}(\Sigma_w)$, respectively. To construct an upper bound on $\lambda_{\max}(\mathcal{H})$ uniformly across $\mathcal{S}_x^+ \times \mathcal{S}_w^+$, we note first that

$$\lambda_{\max}(\mathcal{H}) \le \lambda_{\max}(\mathcal{H}_{xx}) + \lambda_{\max}(\mathcal{H}_{ww}) = 2\left(\lambda_{\max}(D_x)\lambda_{\max}\left(H^\top G^{-1} H\right) + \lambda_{\max}(D_w)\lambda_{\max}\left(G^{-1}\right)\right), \tag{6.18}$$

where the inequality follows from Bernstein [7, fact 5.12.20] and the subadditivity of the maximum eigenvalue, whereas the equality exploits the trace rule of the Kronecker product Bernstein [7, proposition 7.1.10]. In the remainder of the proof, we derive an upper bound for each term on the right-hand side of the preceding expression.

By the definition of G and because $\Sigma_w \in \mathcal{S}_w^+$, we have $G \succeq \lambda_{\min}(\Sigma_w) I_m \succeq \lambda_{\min}(\widehat{\Sigma}_w) I_m$. As $\widehat{\Sigma}_w \succ 0$ by assumption, we may, thus, conclude that $\lambda_{\max}(G^{-1}) \le \lambda_{\min}^{-1}(\widehat{\Sigma}_w)$, which, in turn, implies via (6.17) that

$$\lambda_{\max}(H^\top G^{-1} H) \le \lambda_{\max}(G^{-1})\,\lambda_{\max}(HH^\top) \le \lambda_{\min}^{-1}(\widehat{\Sigma}_w)\,\lambda_{\max}(H^\top H).$$

Similarly, by the definition of D_w, we find

$$\begin{aligned} \lambda_{\max}(D_w) &= \lambda_{\max}(G^{-1} H \Sigma_x^2 H^\top G^{-1}) \\ &\le \lambda_{\max}^2(\Sigma_x)\,\lambda_{\max}^2(G^{-1})\,\lambda_{\max}(H^\top H) \le (\rho_x + \mathrm{Tr}[\widehat{\Sigma}_x]^{\frac{1}{2}})^4\,\lambda_{\min}^{-2}(\widehat{\Sigma}_w)\,\lambda_{\max}(H^\top H), \end{aligned}$$

where the first inequality follows from applying Estimate (6.17) twice, and the last inequality reuses the bound on $\lambda_{\max}(G^{-1})$ derived earlier and exploits Lemma A.4. Finally, by the definition of D_w, we have

$$\begin{aligned}\lambda_{\max}(D_x) &\le 1 + \lambda_{\max}(-H^\top G^{-1} H \Sigma_x) + \lambda_{\max}(-\Sigma_x H^\top G^{-1} H) + \lambda_{\max}(H^\top G^{-1} H \Sigma_x^2 H^\top G^{-1} H) \\ &\le 1 + \lambda_{\max}(H^\top G^{-1} H \Sigma_x^2 H^\top G^{-1} H) \\ &\le 1 + \lambda_{\max}^2(\Sigma_x)\, \lambda_{\max}^2(G^{-1})\, \lambda_{\max}^2(H^\top H) \\ &\le 1 + (\rho_x + \mathrm{Tr}[\widehat{\Sigma}_x]^{\frac{1}{2}})^4\, \lambda_{\min}^{-2}(\widehat{\Sigma}_w)\, \lambda_{\max}^2(H^\top H).\end{aligned}$$

where the first inequality holds because of the subadditivity of the maximum eigenvalue and Bernstein [7, proposition 4.4.10], which implies that the nonzero spectra of $-\Sigma_x H^\top G^{-1} H$ and $-H^\top G^{-1} H \Sigma_x$ are both real and coincide with the nonzero spectrum of the negative semidefinite matrix $-\Sigma_x^{\frac{1}{2}} H^\top G^{-1} H \Sigma_x^{\frac{1}{2}}$. The third inequality follows from applying Estimate (6.17) three times, and the fourth inequality reuses the bound on $\lambda_{\max}(G^{-1})$ and exploits Lemma A.4. Substituting all these bounds into (6.18) completes the proof of assertion (i).

As for assertion (ii), we first show that the feasible set $\mathcal{S}_x^+$ is α_x-strongly convex with respect to $-f$ in the sense of Assumption 6.1(ii). To see this, fix any $\Sigma_x, \Sigma_x' \in \mathcal{S}_x^+$ and $\theta \in [0,1]$, and set

$$\Sigma_\theta = \theta \Sigma_x + (1-\theta)\Sigma_x' + \theta(1-\theta)\frac{\alpha_x}{2}\|\Sigma_x - \Sigma_x'\|^2 \frac{D_x}{\|D_x\|}, \tag{6.19}$$

where $\alpha_x > 0$ is defined as in the proposition statement and D_x denotes again the partial gradient of f with respect to Σ_x. To prove strong convexity of $\mathcal{S}_x^+$ with respect to $-f$, we show that $\Sigma_\theta \in \mathcal{S}_x^+$. Note first that $\Sigma_\theta \succeq \lambda_{\min}(\widehat{\Sigma}_x) I_n$ because $\Sigma_x, \Sigma_x' \in \mathcal{S}_x^+$ and because D_x is positive semidefinite. Next, define

$$\bar{\mathcal{S}}_x = \left\{ \Sigma_x \in \mathbb{S}_+^n : \lambda_{\min}(\widehat{\Sigma}_x) I_n \preceq \Sigma_x \preceq (\rho_x + \mathrm{Tr}[\widehat{\Sigma}_x]^{\frac{1}{2}})^2 I_n \right\},$$

and note that $\mathcal{S}_x^+ \subseteq \bar{\mathcal{S}}_x$ by Lemma A.4. Moreover, Bhatia et al. [9, theorem 1] implies that the function $g_x : \mathbb{S}_+^n \to \mathbb{R}$ defined through $g_x(\Sigma_x) = \mathrm{Tr}[\Sigma_x + \widehat{\Sigma}_x - 2(\widehat{\Sigma}_x^{\frac{1}{2}} \Sigma_x \widehat{\Sigma}_x^{\frac{1}{2}})^{\frac{1}{2}}]$ is κ_1-strongly convex and κ_2-smooth over $\bar{\mathcal{S}}_x$, where

$$\kappa_1 = \frac{\lambda_{\min}^{\frac{1}{2}}(\widehat{\Sigma}_x)}{2(\rho_x + \mathrm{Tr}[\widehat{\Sigma}_x]^{\frac{1}{2}})^3} \quad \text{and} \quad \kappa_2 = \frac{\rho_x + \mathrm{Tr}[\widehat{\Sigma}_x]^{\frac{1}{2}}}{2\lambda_{\min}^{\frac{3}{2}}(\widehat{\Sigma}_x)}.$$

By Journée et al. [41, theorem 12], the sublevel set $\mathcal{S}_x^+ = \{\Sigma_x \in \bar{\mathcal{S}}_x : g_x(\Sigma_x) \le \rho_x^2\}$ is, thus, strongly convex in the canonical sense—relative to $\bar{\mathcal{S}}_x$—with convexity parameter $\alpha_x = \kappa_1/(\sqrt{2\kappa_2}\rho_x)$. This insight implies that $g_x(\bar{\Sigma}_x) \le \rho_x^2$, which, in turn, ensures that $\bar{\Sigma}_x \in \mathcal{S}_x^+$. As $\theta \in [0,1]$ was chosen arbitrarily, we may conclude that $\mathcal{S}_x^+$ is α_x-strongly convex with respect to f. Using an analogous argument, one can show that $\mathcal{S}_w^+$ is α_w-strongly convex with respect to f, and $\alpha_w > 0$ is defined as in the proposition statement. In summary, $\mathcal{S}_x^+ \times \mathcal{S}_w^+$ is, therefore, α-strongly convex with respect to f in the sense of Assumption 6.1(ii), in which $\alpha = \min\{\alpha_x, \alpha_w\}$.

In order to prove assertion (iii), we establish lower bounds on $\lambda_{\max}(D_x)$ and $\lambda_{\max}(D_w)$ uniformly across $\mathcal{S}_x^+ \times \mathcal{S}_w^+$. The claim then follows from the observation that $\lambda_{\max}(D_x) \le \|D_x\|$ and $\lambda_{\max}(D_w) \le \|D_w\|$. We first derive a lower bound on $\lambda_{\max}(D_x)$. To this end, set $T_1 = I_n - \Sigma_x H^\top G^{-1} H$ and $T_2 = H \Sigma_x H^\top G^{-1}$, and note that both T_1 and T_2 have real spectra thanks to Bernstein [7, proposition 4.4.10]. As $D_x = T_1^\top T_1$, we find

$$\lambda_{\max}(D_x) = \lambda_{\max}(T_1 T_1^\top) \ge \max_{\lambda \in \mathrm{spec}(T_1)} |\lambda|^2 = \max_{\lambda \in \mathrm{spec}(T_2)} |1-\lambda|^2 = \lambda_{\max^2}(I - T_2) = \lambda_{\max^2}(\Sigma_w G^{-1}), \tag{6.20}$$

where $\mathrm{spec}(T)$ denotes the eigenvalue spectrum of any square matrix T. The inequality in (6.20) follows from Browne's theorem Bernstein [7, fact 5.11.21], the second equality holds because the nonzero spectrum of $\Sigma_x H^\top G^{-1} H$ matches that of $H \Sigma_x H^\top G^{-1}$ thanks to Bernstein [7, proposition 4.4.10], and the last equality follows from the identity $T_2 = I_m - \Sigma_w G^{-1}$. Notice that all eigenvalues of $\Sigma_w G^{-1}$ are real because T_2 has a real spectrum.

Estimate (6.20) implies that a uniform lower bound on the largest eigenvalue of D_x can be obtained by maximizing the largest eigenvalue of $\Sigma_w G^{-1}$ over $\mathcal{S}_x^+ \times \mathcal{S}_w^+$. By the definition of G, we have

$$\begin{aligned}\inf_{\substack{\Sigma_x \in \mathcal{S}_x^+ \\ \Sigma_w \in \mathcal{S}_w^+}} \lambda_{\max}(\Sigma_w G^{-1}) &\ge \inf_{\substack{\Sigma_x \in \mathcal{S}_x^+ \\ \Sigma_w \in \mathcal{S}_w^+}} \lambda_{\min}(\Sigma_w (H \Sigma_x H^\top + \Sigma_w)^{-1}) \\ &\ge \inf_{\substack{\Sigma_x \in \mathcal{S}_x^+ \\ \Sigma_w \in \mathcal{S}_w^+}} \lambda_{\min}(\Sigma_w)\, \lambda_{\min}\left((H \Sigma_x H^\top + \Sigma_w)^{-1}\right) \\ &\ge \inf_{\substack{\Sigma_x \in \mathcal{S}_x^+ \\ \Sigma_w \in \mathcal{S}_w^+}} \frac{\lambda_{\min}(\Sigma_w)}{\lambda_{\max}(\Sigma_x)\lambda_{\max}(H^\top H) + \lambda_{\max}(\Sigma_w)} \\ &\ge \frac{\lambda_{\min}(\widehat{\Sigma}_w)}{(\rho_x + \mathrm{Tr}[\widehat{\Sigma}_x]^{\frac{1}{2}})^2 \lambda_{\max}(H^\top H) + (\rho_w + \mathrm{Tr}[\widehat{\Sigma}_w]^{\frac{1}{2}})^2},\end{aligned} \tag{6.21}$$

where the second inequality holds because $\lambda_{\min}(T) = \lambda_{\max}^{-1}(T^{-1})$ for any $T \succ 0$ and because the maximum eigenvalue of a positive definite matrix coincides with its operator norm. The third inequality exploits (6.17) and the subadditivity of the maximum eigenvalue, and the last inequality follows from Lemma A.4. Combining (6.20) and (6.21) shows that $\|D_x\| \geq \lambda_{\max}(D_x) \geq \varepsilon_x$, where ε_x is defined as in the proposition statement.

Using similar arguments, we can also derive a uniform lower bound on $\lambda_{\max}(D_w)$. Specifically, we have

$$\begin{aligned}\lambda_{\max}(D_w)\lambda_{\max}(H^\top H) &\geq \lambda_{\max}(H^\top D_w H) = \lambda_{\max}(H^\top G^{-1} H\Sigma_x (H^\top G^{-1} H\Sigma_x)^\top)\\ &\geq \lambda_{\max}^2(H\Sigma_x H^\top G^{-1}) = \left(1 - \lambda_{\min}(\Sigma_w G^{-1})\right)^2\\ &= \left(1 - \frac{1}{1+\lambda_{\max}(H\Sigma_x H^\top \Sigma_w^{-1})}\right)^2\\ &= \left(1 - \frac{1}{1+\lambda_{\max}\left(\Sigma_w^{-\frac{1}{2}} H\Sigma_x H^\top \Sigma_w^{-\frac{1}{2}}\right)}\right)^2,\end{aligned} \tag{6.22}$$

where the first inequality follows from (6.17), and the first equality holds because of the definition of D_w. Moreover, the second inequality exploits Brown's theorem Bernstein [7, fact 5.11.21], and the second equality uses the definition of G. Finally, the third equality follows from the relation $\lambda_{\min}(\Sigma_w G^{-1}) = \lambda_{\max}^{-1}(G\Sigma_w)$, and the fourth equality holds because of Bernstein [7, proposition 4.4.10]. A uniform lower bound on D_w can, thus, be obtained from the estimate

$$\Sigma_w^{-\frac{1}{2}} H\Sigma_x H^\top \Sigma_w^{-\frac{1}{2}} \geq \lambda_{\min}(\Sigma_x)\Sigma_w^{-\frac{1}{2}} HH^\top \Sigma_w^{-\frac{1}{2}} \geq \frac{\lambda_{\min}(\Sigma_x)}{\lambda_{\max}(\Sigma_w)} HH^\top,$$

which implies via Lemma A.4 that

$$\inf_{\substack{\Sigma_x \in \mathcal{S}_x^+ \\ \Sigma_w \in \mathcal{S}_w^+}} \lambda_{\max}(\Sigma_w^{-\frac{1}{2}} H\Sigma_x H^\top \Sigma_w^{-\frac{1}{2}}) \geq \inf_{\substack{\Sigma_x \in \mathcal{S}_x^+ \\ \Sigma_w \in \mathcal{S}_w^+}} \frac{\lambda_{\min}(\Sigma_x)}{\lambda_{\max}(\Sigma_w)} \lambda_{\max}(H^\top H) \geq \frac{\lambda_{\min}(\widehat{\Sigma}_x)}{(\rho_w + \operatorname{Tr}[\widehat{\Sigma}_w]^{\frac{1}{2}})^2} \lambda_{\max}(H^\top H). \tag{6.23}$$

Combining (6.22) and (6.23) shows that $\|D_w\| \geq \lambda_{\max}(D_w) \geq \varepsilon_w$, where ε_w is defined as in the proposition statement. We, thus, conclude that f is ε-steep in the sense of Assumption 6.1(iii) with $\varepsilon = \min\{\varepsilon_x, \varepsilon_w\}$. □

By Theorem 6.1, which is applicable because of Proposition 6.1, the fully adaptive Frank–Wolfe algorithm (see Algorithm 1) solves the (minimization problem equivalent to the) nonlinear SDP (6.13) at a linear convergence rate. Moreover, Theorem 6.2 ensures that the oracle problem (6.14), which needs to be solved in each iteration of Algorithm 1, can be solved highly efficiently via bisection (see Algorithm 2).

We emphasize that, if $\widehat{\Sigma}_x$ is singular, then the strong convexity parameter α of Assumption 6.1(ii) vanishes, and therefore, Theorem 6.1 is no longer applicable. In that case, however, Algorithm 1 is still guaranteed to converge, albeit at a sublinear rate; see Pedregosa et al. [59] for further details.

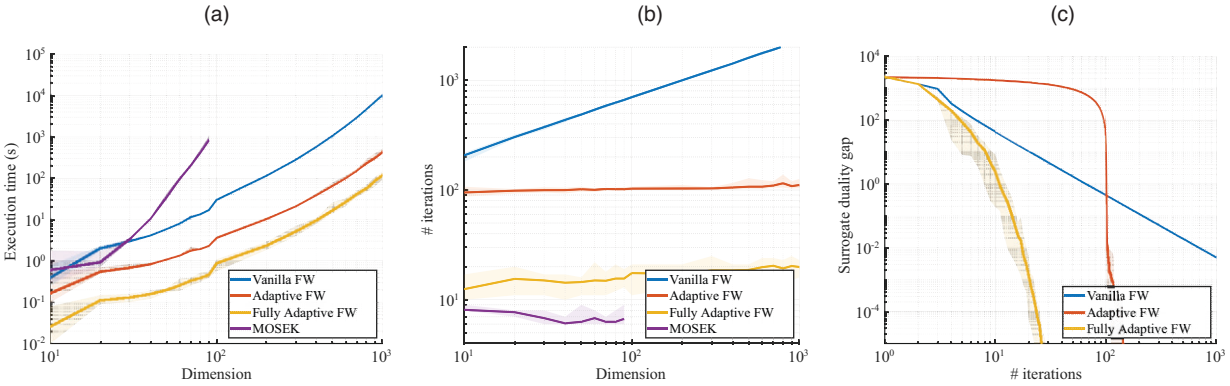
7. Numerical Experiments

All experiments are run on an Intel XEON CPU with 3.40 GHz clock speed and 16 GB of RAM. All (linear) SDPs are solved with MOSEK 8 using the YALMIP interface Löfberg [50]. In order to ensure the reproducibility of our experiments, we make all source codes available at https://github.com/sorooshafiee/WMMSE.

7.1. Scalability of the Frank–Wolfe Algorithm

We first compare the convergence behavior of the Frank–Wolfe algorithm developed in Section 6 against that of MOSEK. Each experiment consists of 10 independent simulation runs, in all of which we fix the signal and noise dimensions to $n = m = d$ and the Wasserstein radii to $\rho_x = \rho_w = \sqrt{d}$ for some $d \in \mathbb{N}$. In each simulation run, we randomly generate two nominal covariance matrices $\widehat{\Sigma}_x$ and $\widehat{\Sigma}_w$ as follows. First, we sample 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 set $\widehat{\Sigma}_x = R_x\Lambda_x(R_x)^\top$ and $\widehat{\Sigma}_w = R_w\Lambda_w R_w^\top$, where Λ_x and Λ_w are diagonal matrices whose main diagonals are sampled uniformly from $[1,5]^d$ and $[1,2]^d$, respectively. Finally, we set $\widehat{\mu}_x = 0$ and $\widehat{\mu}_w = 0$. The Wasserstein MMSE estimator can then be computed by solving either the nonlinear SDP (3.4) with a Frank–Wolfe algorithm or the linear SDP (3.10) with MOSEK. Figure 1, (a) and (b), shows the execution time and the number of iterations needed by the vanilla, adaptive, and

Figure 1. (Color online) Scalability properties of different methods for computing the Wasserstein MMSE estimator (shown are the means (solid lines) and the ranges (shaded areas) of the respective performance measures across 10 simulation runs). (a) Scaling of execution time. (b) Scaling of iteration count. (c) Convergence for $d = 1{,}000$.



fully adaptive versions of the FW algorithm as well as by MOSEK to drive the (surrogate) duality gap below 10^{-3}. MOSEK runs out of memory for all dimensions $d > 100$. Figure 1(c) visualizes the empirical convergence behavior of the three different Frank–Wolfe algorithms. We observe that the fully adaptive Frank–Wolfe algorithm finds highly accurate solutions already after 20 iterations for problem instances of dimension $d = 1{,}000$.

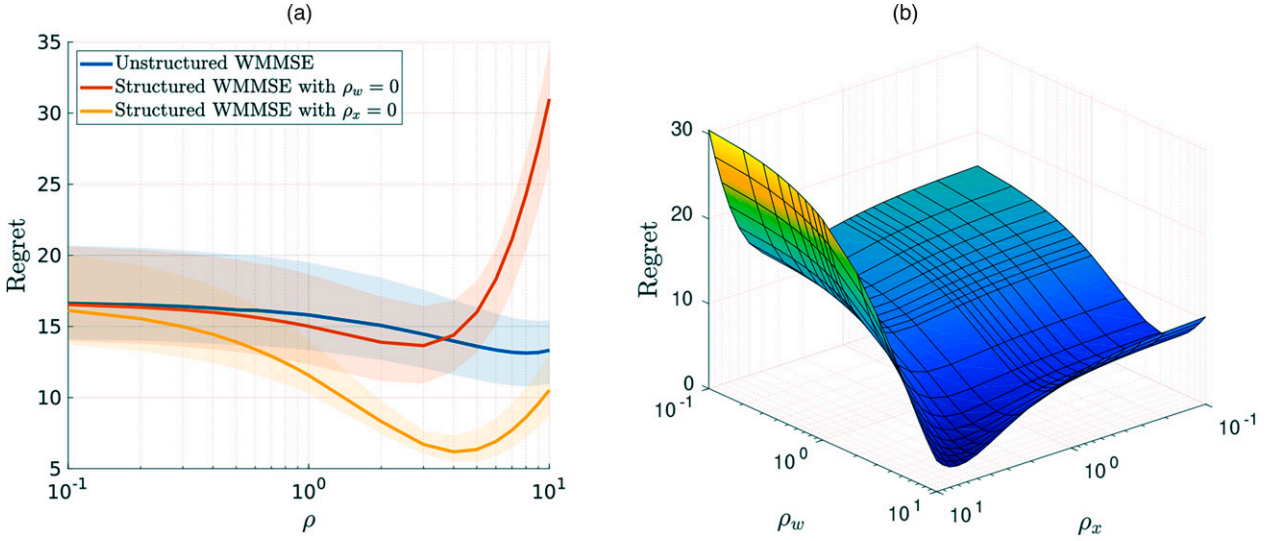
7.2. The Value of Structural Information

In the second experiment, we study the predictive power of different MMSE estimators. The experiment consists of 100 independent simulation runs. In each run, we use the same procedure as in Section 7.1 to generate two random covariance matrices Σ_x and Σ_w of dimensions $n = m = 50$, and we set the true signal and noise distributions to $\mathbb{P}_x = \mathcal{N}(0, \Sigma_x)$ and $\mathbb{P}_w = \mathcal{N}(0, \Sigma_w)$, respectively. Next, we define $\widehat{\Sigma}_x$ and $\widehat{\Sigma}_w$ as the sample covariance matrices corresponding to 100 independent samples from $\mathbb{P}_x$ and $\mathbb{P}_w$, respectively. Moreover, we set $H = I_n$. To assess the value of structural information, we compare the Wasserstein MMSE estimator proposed in this paper against the Bayesian MMSE estimator associated with the nominal signal and noise distributions and the unstructured Wasserstein MMSE estimator proposed in Shafieezadeh-Abadeh et al. [67]. The latter uses a single Wasserstein ball to model the ambiguity of the joint distribution of x and y, thereby ignoring the structural information that $w = Hy - x$ is independent of x. Both the structured and unstructured Wasserstein MMSE estimators collapse to the nominal Bayesian MMSE estimator when the underlying Wasserstein radii are set to zero. Recall also that the nominal Bayesian MMSE estimator is optimal in distributionally robust estimation problems whose ambiguity sets are defined via information divergences (Levy and Nikoukhah [48, 49], Zorzi [76, 77]). This robustness property makes the nominal Bayesian MMSE estimator an interesting benchmark.

We quantify the performance of a given estimator by its *regret*, which is defined as the difference between the estimator's average risk and the least possible average risk of *any* estimator under the unknown true distributions $\mathbb{P}_x$ and $\mathbb{P}_w$. Note that the minimum average risk is attained by the (affine) Bayesian MMSE estimator corresponding to the (normal) distributions $\mathbb{P}_x$ and $\mathbb{P}_w$. Figure 2(a) shows the regret of the Wasserstein MMSE estimator with $\rho_x = \rho$ and $\rho_w = 0$, the Wasserstein MMSE estimator with $\rho_x = 0$ and $\rho_w = \rho$, and the unstructured Wasserstein MMSE estimator from Shafieezadeh-Abadeh et al. [67] with Wasserstein radius ρ for all $\rho \in [0.1, 10]$. The solid lines represent the averages and the shaded areas the ranges of the regret across all 100 simulation runs. The regret of the structured Wasserstein MMSE estimator with Wasserstein radii $\rho_x, \rho_w \in [0.1, 10]$ averaged across all 100 simulation runs is visualized by the surface plot in Figure 2(b).

We observe that the average regret of the nominal Bayesian MMSE estimator amounts to 16.7 (the leftmost value of all curves in Figure 2(a)), and the best unstructured Wasserstein MMSE estimator attains a significantly lower average regret of 13.1 (the minimum of the blue curve in Figure 2(a)). The best structured Wasserstein MMSE estimator without noise ambiguity ($\rho_w = 0$) displays a similar performance, attaining an average regret of 13.2 (the minimum of the red curve in Figure 2(a)), and the one without signal ambiguity ($\rho_x = 0$) further reduces the average regret by more than 50% to 6.0 (the minimum of the yellow curve in Figure 2(a)). Finally, the best among *all* structured Wasserstein MMSE estimators, which is obtained by tuning both ρ_x and ρ_w, attains an even lower average regret of 2.2 (the minimum of the surface plot in Figure 2(b)). This experiment confirms our

Figure 2. (Color online) Regret of different estimators averaged across 100 simulation runs. (a) Regret of different Wasserstein MMSE estimators involving a single hyperparameter ρ. (b) Regret of the Wasserstein MMSE estimator involving two hyperparameters ρ_x and ρ_w.



hypothesis that structural (independence) information as well as information about distributional ambiguity can improve the predictive power of MMSE estimators.

Unlike in data-driven optimization, in which the nominal distribution and the radii of the ambiguity sets can be tuned from data, we assume here that the nominal distribution and the radii of the ambiguity sets reflect the modeler's prior distributional information. Thus, they are reminiscent of prior distributions in Bayesian statistics. The radii could be tuned using standard cross-validation techniques, for example, if we had access to samples $(\hat{x}_i, \hat{y}_i)$, $i = 1, \ldots, N$ drawn independently from the true joint distribution of (x, y) under $\mathbb{P}$. This conflicts, however, with our central assumption that only y is observable. In this case, it is fundamentally impossible to assess the empirical performance of an estimator and to tune the radii via cross validation.

Acknowledgments

The authors are grateful to Erick Delage for valuable comments on an earlier version of this paper.

Appendix. Auxiliary Results

We first prove Pythagoras' theorem for the Wasserstein distance.

Proof of Lemma 1.1. By the definition of the Wasserstein distance and by Pythagoras' theorem for the Euclidean distance, we have

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

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

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

where the first inequality exploits the superadditivity of the infimum operator, and 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.

In order to prove Proposition 2.4, we establish first general conditions for the solvability and stability of parametric minimax problems and prove that the matrix square root is Hölder continuous.

Lemma A.1 (Parametric Minimax Problems). *Consider the parametric minimax problem*

$$J(\theta) = \inf_{u \in \mathbb{U}} \sup_{v \in V(\theta)} f(u, v), \tag{A.1}$$

where $\mathbb{U}$, $\mathbb{V}$ and Θ are nonempty convex subsets of Euclidean spaces equipped with the respective subspace topologies, $f : \mathbb{U} \times \mathbb{V} \to \mathbb{R}$ and $g : \mathbb{V} \times \Theta \to \mathbb{R}$ are continuous functions, and $V : \Theta \rightrightarrows \mathbb{V}$ and $V^\circ : \Theta \rightrightarrows \mathbb{V}$ are set-valued mappings defined through $V(\theta) = \{v \in \mathbb{V} : g(v, \theta) \leq 0\}$ and $V^\circ(\theta) = \{v \in \mathbb{V} : g(v, \theta) < 0\}$, respectively. Assume that, for each $\theta' \in \Theta$, there exists a compact neighborhood $\Theta' \subseteq \Theta$ of θ' such that

i. *$V(\theta)$ is convex and bounded uniformly across all $\theta \in \Theta'$.*

ii. *$V^\circ(\theta)$ is nonempty and coincides with the interior of $V(\theta)$ for all $\theta \in \Theta'$.*

iii. *There exist $v' \in \mathbb{V}$ and $J' \in \mathbb{R}$ such that $v' \in V(\theta)$ and $J(\theta) \leq J'$ for all $\theta \in \Theta'$ and such that the set $\mathbb{U}' = \{u \in \mathbb{U} : f(u, v') \leq J'\}$ is nonempty and compact.*

Then, the minimax problem (A.1) is solvable for all $\theta \in \Theta$, and J is continuous on Θ.

Proof of Lemma A.1. Define $F(u, \theta) = \sup_{v \in V(\theta)} f(u, v)$ and note that $F(u, \theta)$ is finite because it is the maximum of a continuous function over a compact feasible set. As V is locally bounded thanks to assumption (i) and as the graph of V is closed thanks to the continuity of g, the closed graph theorem Aubin and Frankowska [1, proposition 1.4.8] ensures that the set-valued mapping V is upper semicontinuous in the sense of Berge [6, chapter VI]. By Berge [6, theorem 2], the optimal value function F is, thus, upper semicontinuous on $\mathbb{U} \times \Theta$. Next, note that $F(u, \theta) = \sup_{v \in V^\circ(\theta)} f(u, v)$ because f is continuous and because $V^\circ(\theta)$ coincides with the nonempty interior of $V(\theta)$ thanks to assumption (ii). As V° is convex-valued thanks to assumption (i) and as the graph of V is open thanks to the continuity of g, the open graph theorem Zhou [75, proposition 2] implies that the set-valued mapping V is lower semicontinuous in the sense of Berge [6, chapter VI]. By Berge [6, theorem 1], the function F is, thus, lower semicontinuous on $\mathbb{U} \times \Theta$. In summary, F is, therefore, continuous on $\mathbb{U} \times \Theta$.

To prove that J is lower semicontinuous, we select an arbitrary $\theta' \in \Theta$ and a compact neighborhood $\Theta' \subseteq \Theta$ of θ' that satisfies assumption (iii). For any $\theta \in \Theta'$, we may then restrict the feasible set of the outer minimization problem in (A.1) to $\mathbb{U}'$ without affecting $J(\theta)$. Indeed, any $u \notin \mathbb{U}'$ and $\theta \in \Theta'$ satisfies

$$F(u, \theta) = \sup_{v \in V(\theta)} f(u, v) \geq f(u, v') > J' \geq J(\theta),$$

where the three inequalities hold because $v' \in V(\theta)$ for all $\theta \in \Theta'$, $u \notin \mathbb{U}'$ and $J(\theta) \leq J'$ for all $\theta \in \Theta'$, respectively. Thus, $u \notin \mathbb{U}'$ is strictly suboptimal, and we have $J(\theta) = \inf_{u \in \mathbb{U}'} F(u, \theta)$ for all $\theta \in \Theta'$. As $\mathbb{U}'$ is compact and $F(u, \theta)$ is continuous, the restricted minimax problem with $\mathbb{U}'$ replacing $\mathbb{U}$ is solvable, and any minimizer also solves (A.1). Moreover, the constant feasible set mapping that assigns each $\theta \in \Theta'$ the same set $\mathbb{U}'$ is continuous in the sense of Berge [6, chapter VI]. By Berge's [6, theorem 1] maximum theorem, the function J is, thus, continuous on Θ' and, as $\theta' \in \Theta$ was chosen arbitrarily, on all of Θ. □

Lemma A.2 (Hölder Continuity of the Matrix Square Root). *The square root of a positive semidefinite matrix is Hölder-continuous with exponent 1/2. More precisely, we have*

$$\|\sqrt{A_1} - \sqrt{A_2}\| \le 2\sqrt{d}\,\|A_1 - A_2\|^{1/2} \quad \forall A_1, A_2 \in \mathbb{S}_+^d .$$

Proof of Lemma A.2. The proof exploits the following two facts.

i. By Schmitt [65, lemma 2.2], we have

$$\|\sqrt{A_1} - \sqrt{A_2}\| \le \frac{1}{\lambda_{\min}(A_1) + \lambda_{\min}(A_2)}\|A_1 - A_2\| \quad \forall A_1, A_2 \in \mathbb{S}_{++}^d .$$

ii. One can show that

$$\|\sqrt{A + \varepsilon I_d} - \sqrt{A}\| \le d\sqrt{\varepsilon} \quad \forall A \in \mathbb{S}_+^d,\ \varepsilon \ge 0 .$$

To prove assertion (ii), denote by $\lambda_i \ge 0$, $i = 1, \ldots, d$, the eigenvalues of $A \in \mathbb{S}_+^d$ and by $a_i \in \mathbb{R}^d$, $i = 1, \ldots, d$, the corresponding orthonormal eigenvectors, which implies that $A = \sum_{i=1}^d \lambda_i a_i a_i^\top$. Thus, we find

$$\|\sqrt{A + \varepsilon I_d} - \sqrt{A}\| = \left\|\sum_{i=1}^d (\sqrt{(\lambda_i + \varepsilon)} - \sqrt{\lambda_i}) a_i a_i^\top\right\| = \left\|\sum_{i=1}^d \varepsilon(\sqrt{(\lambda_i + \varepsilon)} + \sqrt{\lambda_i})^{-1} a_i a_i^\top\right\|$$
$$\le \sum_{i=1}^d \frac{\varepsilon}{\sqrt{(\lambda_i + \varepsilon)} + \sqrt{\lambda_i}}\|a_i a_i^\top\| \le d\sqrt{\varepsilon},$$

where the first inequality follows from the triangle inequality and the second inequality holds because the denominator of the ith fraction grows with λ_i and is, therefore, minimal for $\lambda_i = 0$.

To prove the lemma, select now arbitrary $A_1, A_2 \in \mathbb{S}_+^d$ and $\varepsilon > 0$. The triangle inequality implies that

$$\|\sqrt{A_1} - \sqrt{A_2}\| \le \|\sqrt{A_1} - \sqrt{A_1 + \varepsilon I_d}\| + \|\sqrt{A_1 + \varepsilon I_d} - \sqrt{A_2 + \varepsilon I_d}\| + \|\sqrt{A_2 + \varepsilon I_d} - \sqrt{A_2}\|$$
$$\le d\sqrt{\varepsilon} + \frac{1}{\lambda_{\min}(\sqrt{A_1 + \varepsilon I_d}) + \lambda_{\min}(\sqrt{A_2 + \varepsilon I_d})}\|A_1 - A_2\| + d\sqrt{\varepsilon} \le 2d\sqrt{\varepsilon} + \frac{\|A_1 - A_2\|}{2\sqrt{\varepsilon}},$$

where the second exploits assertions (i) and (ii). The claim now follows by setting $\varepsilon = \|A_1 - A_2\|/4d$, which actually minimizes the right-hand side of the last expression. □

Armed with Lemmas A.1 and A.2, we are now ready to prove Proposition 2.4.

Proof of Proposition 2.4. The Gelbrich MMSE estimation problem (2.1) upper bounds the Wasserstein MMSE estimation problem (1.7) because $\mathcal{A} \subseteq \mathcal{F}$ and $\mathbb{G}(\widehat{\mathbb{P}}) \supseteq \mathbb{B}(\widehat{\mathbb{P}})$; see Corollary 2.1. 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 (2.1) depend on ψ and $\mathbb{Q}$ only through $u = (A, b)$, which ranges over $\mathbb{U} = \mathbb{R}^{n\times m} \times \mathbb{R}^n$, and $v = (\mu_x, \mu_w, \Sigma_x, \Sigma_w)$, which ranges over $\mathbb{V} = \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{S}_+^n \times \mathbb{S}_+^m$. Indeed, the average risk of ψ under $\mathbb{Q}$ satisfies

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

Similarly, the constraint $\mathbb{Q} \in \mathbb{G}(\widehat{\mathbb{P}})$ can be reformulated as $g(v, \theta) \le 0$, where $\theta = (\widehat{\mu}_x, \widehat{\mu}_w, \widehat{\Sigma}_x, \widehat{\Sigma}_w)$ is shorthand for the problem's input parameters ranging over the set $\Theta = \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{S}_+^n \times \mathbb{S}_+^m$, and where

$$g(v, \theta) = \max\left\{\mathbb{G}((\mu_x, \Sigma_x), (\widehat{\mu}_x, \widehat{\Sigma}_x))^2 - \rho_x^2,\ \mathbb{G}((\mu_w, \Sigma_w), (\widehat{\mu}_w, \widehat{\Sigma}_w))^2 - \rho_w^2\right\}.$$

Thus, the Gelbrich MMSE estimation problem (2.1) can be re-expressed more concisely as

$$J(\theta) = \inf_{u\in\mathbb{U}} \sup_{v\in V(\theta)} f(u, v), \tag{A.2}$$

where $V(\theta) = \{v \in \mathbb{V} : g(v, \theta) \le 0\}$ for all $\theta \in \Theta$. Assume now that $\rho_x > 0$ and $\rho_w > 0$. In the remainder, we prove that the optimal value $J(\theta)$ of the minimax problem (A.2) is attained and continuous in θ by showing that all assumptions of Lemma A.1 are satisfied.

To this end, note first that f is continuous by construction and that g is convex and continuous thanks to Proposition 2.2. Next, select an arbitrary reference point $\theta' = (\widehat{\mu}_x', \widehat{\mu}_w', \widehat{\Sigma}_x', \widehat{\Sigma}_w') \in \Theta$, and define

$$\Theta' = \left\{ (\widehat{\mu}_x, \widehat{\mu}_w, \widehat{\Sigma}_x, \widehat{\Sigma}_w) \in \Theta : \begin{array}{ll} \mathbb{G}((\widehat{\mu}_x, \widehat{\Sigma}_x), (\widehat{\mu}_x', \widehat{\Sigma}_x')) & \leq \frac{\rho_x}{2} \\ \mathbb{G}((\widehat{\mu}_w, \widehat{\Sigma}_w), (\widehat{\mu}_w', \widehat{\Sigma}_w')) & \leq \frac{\rho_w}{2} \end{array} \right\}.$$

Note that Θ' is a neighborhood of θ' because ρ_x and ρ_w are strictly positive and because Proposition 2.2 ensures that the (squared) Gelbrich distance is continuous. Moreover, Θ' is compact because of Lemma A.4.

In order to verify assumption (i) of Lemma A.1, we note first that $V(\theta)$ is a convex set for every $\theta \in \Theta$ because g is a convex function. Moreover, we have

$$\begin{aligned} V(\theta) &= \left\{ (\mu_x, \mu_w, \Sigma_x, \Sigma_w) \in \mathbb{V} : \begin{array}{ll} \mathbb{G}((\mu_x, \Sigma_x), (\widehat{\mu}_x, \widehat{\Sigma}_x)) & \leq \rho_x \\ \mathbb{G}((\mu_w, \Sigma_w), (\widehat{\mu}_w, \widehat{\Sigma}_w)) & \leq \rho_w \end{array} \right\} \\ &\subseteq \left\{ (\mu_x, \mu_w, \Sigma_x, \Sigma_w) \in \mathbb{V} : \begin{array}{ll} \mathbb{G}((\mu_x, \Sigma_x), (\widehat{\mu}_x', \widehat{\Sigma}_x')) & \leq \frac{3\rho_x}{2} \\ \mathbb{G}((\mu_w, \Sigma_w), (\widehat{\mu}_w', \widehat{\Sigma}_w')) & \leq \frac{3\rho_w}{2} \end{array} \right\} = V' \quad \forall \theta \in \Theta', \end{aligned}$$

where the first equality holds because of the definition of $V(\theta)$; the inclusion holds because of the definition of Θ' and because the Gelbrich distance satisfies the triangle inequality; and the last equality defines the set V', which is compact thanks to Lemma A.4. This reasoning shows that $V(\theta)$ is uniformly bounded on Θ'.

In order to verify assumption (ii) of Lemma A.1, define

$$V^\circ(\theta) = \{ v \in \mathbb{V} : g(v, \theta) < 0 \} = \left\{ (\mu_x, \mu_w, \Sigma_x, \Sigma_w) \in \mathbb{V} : \begin{array}{ll} \mathbb{G}((\mu_x, \Sigma_x), (\widehat{\mu}_x, \widehat{\Sigma}_x)) & < \rho_x \\ \mathbb{G}((\mu_w, \Sigma_w), (\widehat{\mu}_w, \widehat{\Sigma}_w)) & < \rho_w \end{array} \right\}$$

for any $\theta \in \Theta$. As the Gelbrich distance satisfies the identity of indiscernibles and as ρ_x and ρ_w are strictly positive, we have $\theta \in V^\circ(\theta)$, which implies that $V^\circ(\theta)$ is nonempty. It remains to be shown that $V^\circ(\theta)$ coincides with the interior of $V(\theta)$. As the Gelbrich distance is continuous by virtue of Proposition 2.2, it is clear that $V^\circ(\theta)$ is an open subset of $V(\theta)$ and, thus, contained in $\mathrm{int}(V(\theta))$. To prove the converse inclusion, assume for the sake of argument that $V^\circ(\theta)$ is a strict subset of $\mathrm{int}(V(\theta))$. Thus, there must exist an open set $O \subseteq \mathrm{int}(V(\theta)) \backslash V^\circ(\theta)$. Otherwise, each point in $\mathrm{int}(V(\theta)) \backslash V^\circ(\theta)$ would belong to the boundary of $\mathrm{int}(V(\theta))$, which is impossible because $\mathrm{int}(V(\theta))$ is open. As $O \subseteq V(\theta) \backslash V^\circ(\theta)$, it is clear that at least one of the equalities $\mathbb{G}((\mu_x, \Sigma_x), (\widehat{\mu}_x, \widehat{\Sigma}_x)) = \rho_x$ or $\mathbb{G}((\mu_w, \Sigma_w), (\widehat{\mu}_w, \widehat{\Sigma}_w)) = \rho_w$ is satisfied at any point in O. In fact, as the Gelbrich distance is continuous, one of these equalities holds throughout an open set $O' \subseteq O$. Without loss of generality, we may, thus, assume that $\mathbb{G}((\mu_x, \Sigma_x), (\widehat{\mu}_x, \widehat{\Sigma}_x))^2 = \rho_x^2$ on an open set $O' \subseteq O$, which implies that any point in O' is a local minimizer of the squared Gelbrich distance. As the squared Gelbrich distance is convex because of Proposition 2.2, this means that all points in O' are, in fact, global minimizers. This is not possible, however, because the (squared) Gelbrich distance adopts its global minimum only at points at which $\mu_x = \widehat{\mu}_x$ and $\Sigma_x = \widehat{\Sigma}_x$. No such point belongs to O' because $O' \cap V^\circ(\theta) = \emptyset$. Thus, we have $\mathrm{int}(V(\theta)) = V^\circ(\theta)$.

In order to verify assumption (iii) of Lemma A.1, select a point $v' = (\mu_x', \mu_w', \Sigma_x', \Sigma_w') \in \Theta'$ with $\Sigma_w' > 0$. Such a point exists because $\rho_w > 0$ and because the (squared) Gelbrich distance is continuous by virtue of Proposition 2.2, which implies that

$$(\mu_x', \mu_w', \Sigma_x', \Sigma_w') = (\widehat{\mu}_x, \widehat{\mu}_w, \widehat{\Sigma}_x, \widehat{\Sigma}_w + \lambda \cdot I_m) \in \Theta'$$

for all sufficiently small $\lambda > 0$. The triangle inequality for the Gelbrich distance then ensures that $v' \in V(\theta)$ for all $\theta \in \Theta'$. Next, set $J' = \sup_{v \in V'} f(0, v)$, which is finite because V' is compact and f is continuous, and note that

$$J(\theta) = \inf_{u \in \mathbb{U}} \sup_{v \in V(\theta)} f(u, v) \leq \sup_{v \in V'} f(0, v) = J' \quad \forall \theta \in \Theta',$$

where the inequality holds because $V(\theta) \subseteq V'$ whenever $\theta \in \Theta'$. Finally, define $\mathbb{U}' = \{ u \in \mathbb{U} : f(u, v') \leq J' \}$, which is nonempty. Indeed, $\mathbb{U}'$ contains at least the point $u = 0$ because $v' \in \Theta' \subseteq V'$. As $\Sigma_w' > 0$, it is easy to verify that $f(u, v')$ is strictly convex and quadratic in u, which implies that $\mathbb{U}'$ is a compact ellipsoid.

In summary, Problem (A.2), which is equivalent to the Gelbrich MMSE estimation problem (2.1), satisfies all assumptions of Lemma A.1. Therefore, its optimal value $J(\theta)$ is attained and continuous in θ. □

In order to derive a tractable reformulation for the Gelbrich MMSE estimation problem studied in Section 2, we need to be able to solve nonlinear SDPs of the form

$$J^\star = \sup_{\Sigma \geq 0} \ \langle D, \Sigma \rangle - \gamma \mathrm{Tr}[\Sigma - 2(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}] \tag{A.3}$$

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 (A.3) admits a unique optimal solution that is available in closed form Nguyen et al. [55]. In the following, we review the construction of this optimal solution under slightly more general conditions.

Proposition A.1 (Closed-Form Solution of (A.3)). *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 (A.3) is given by*

$$J^\star = \begin{cases} \gamma^2 \langle (\gamma I_d - D)^{-1}, \widehat{\Sigma} \rangle & \text{if } \gamma > \lambda_{\max}(D), \\ \liminf\limits_{\bar{\gamma} \downarrow \gamma} \bar{\gamma}^2 \langle (\bar{\gamma} I_d - D)^{-1}, \widehat{\Sigma} \rangle & \text{if } \gamma = \lambda_{\max}(D), \\ +\infty & \text{if } \gamma < \lambda_{\max}(D). \end{cases}$$

Moreover, Problem (A.3) 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 of Proposition A.1. 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 (A.3) as

$$\begin{aligned} J^\star &= \sup_{B \succeq 0} \ \langle D, \widehat{\Sigma}^{-\frac{1}{2}} B^2 \widehat{\Sigma}^{-\frac{1}{2}} \rangle - \gamma \mathrm{Tr}[\widehat{\Sigma}^{-\frac{1}{2}} B^2 \widehat{\Sigma}^{-\frac{1}{2}} - 2B] \\ &= \sup_{B \succeq 0} \ \langle B^2, \widehat{\Sigma}^{-\frac{1}{2}} (D - \gamma I_d) \widehat{\Sigma}^{-\frac{1}{2}} \rangle + 2\gamma \langle B, I_d \rangle, \end{aligned}$$

where the second equality exploits the cyclicity of the trace operator. Introducing the auxiliary parameter $\Delta = \widehat{\Sigma}^{-\frac{1}{2}} (D - \gamma I_d) \widehat{\Sigma}^{-\frac{1}{2}}$, we can then rewrite the last maximization problem over B more concisely as

$$J^\star = \sup_{B \succeq 0} \ \langle B^2, \Delta \rangle + 2\gamma \langle B, I_d \rangle. \tag{A.4}$$

Note that (A.4) 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 (A.4) 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 Hespanha [37, theorem 12.5]. As it is strictly positive definite, $B^\star$, thus, uniquely solves (A.4), 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 (A.3). Substituting $\Sigma^\star$ back into the objective function of (A.3) further shows that $J^\star = \gamma^2 \langle (\gamma I_d - D)^{-1}, \widehat{\Sigma} \rangle$.

Next, we 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

$$J^\star(\widehat{\Sigma}) = \gamma^2 \langle (\gamma I_d - D)^{-1}, \widehat{\Sigma} \rangle \quad \text{and} \quad \Sigma^\star(\widehat{\Sigma}) = \gamma^2 (\gamma I_d - D)^{-1} \widehat{\Sigma} (\gamma I_d - D)^{-1}$$

as explicit continuous functions of the parameter $\widehat{\Sigma} \in \mathbb{S}_+^d$. Similarly, define the function

$$F(\Sigma, \widehat{\Sigma}) = \langle D, \Sigma \rangle - \gamma \mathrm{Tr}[\Sigma - 2(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}],$$

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 $J^\star(\widehat{\Sigma})$, and 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 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. These arguments imply that $J^\star(\widehat{\Sigma})$ and $\Sigma^\star(\widehat{\Sigma})$ represent the optimal value and an optimal solution of (A.3), respectively, even if $\widehat{\Sigma} \in \mathbb{S}_+^d$ is rank-deficient.

Assume next that $\gamma < \lambda_{\max}(D)$, and denote by $\bar{v} \in \mathbb{R}^d$ a normalized eigenvector of D corresponding to the eigenvalue $\lambda_{\max}(D)$. By optimizing only over matrices of the form $\Sigma = t \bar{v} \bar{v}^\top$ for some $t \geq 0$, we find

$$\begin{aligned} J^\star &\geq \sup_{t \geq 0} \ t \langle D - \gamma I_d, \bar{v} \bar{v}^\top \rangle + 2\sqrt{t} \gamma \mathrm{Tr}[(\widehat{\Sigma}^{\frac{1}{2}} \bar{v} \bar{v}^\top \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}] \\ &= \sup_{t \geq 0} \ t (\lambda_{\max}(D) - \gamma) + 2\sqrt{t} \gamma \mathrm{Tr}[(\widehat{\Sigma}^{\frac{1}{2}} \bar{v} \bar{v}^\top \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}] = \infty. \end{aligned}$$

Assume finally that $\gamma = \lambda_{\max}(D)$. To investigate this limiting case, note that the objective function of (A.3) is linear in γ, which implies that the optimal value of (A.3) 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 (A.3) must be given by $J^\star = \liminf_{\bar{\gamma} \downarrow \gamma} \bar{\gamma}^2 \langle (\bar{\gamma} I_d - D)^{-1}, \widehat{\Sigma} \rangle$. This observation completes the proof.

In order to derive search directions for the Frank–Wolfe algorithm developed in Section 6, we need to be able to solve constrained nonlinear SDPs of the form

$$\begin{array}{ll} \sup\limits_{\Sigma\succeq 0} & \langle D,\Sigma\rangle \\ \text{s.t.} & \mathrm{Tr}[\Sigma+\widehat{\Sigma}-2(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}]\le\rho^2 \end{array} \tag{A.5}$$

parameterized by $D\in\mathbb{S}^d$, $\widehat{\Sigma}\in\mathbb{S}_+^d$ and $\rho\in\mathbb{R}_+$. It is known that Problem (A.5) admits a unique optimal solution that is available in quasi-closed form (Nguyen et al. [55]). We review the construction of this optimal solution under more general conditions and uncover several previously unknown properties of this solution.

Proposition A.2 (Quasi-Closed Form Solution of (A.5)). *The following statements hold:*

i. *If $D\in\mathbb{S}^d$, $\widehat{\Sigma}\in\mathbb{S}_+^d$ and $\rho\in\mathbb{R}_+$, then Problem (A.5) is solvable, and its maximum matches that of the univariate convex minimization problem*

$$\inf_{\substack{\gamma\ge 0\\ \gamma>\lambda_{\max}(D)}} \gamma\Big(\rho^2+\langle\gamma(\gamma I_d-D)^{-1}-I_d,\widehat{\Sigma}\rangle\Big). \tag{A.6}$$

ii. *If $D\neq 0$, $\widehat{\Sigma}\succ 0$ and $\rho>0$, then Problem (A.6) has a unique minimizer $\gamma^\star\in(\lambda_{\max}(D),\infty)$, and Problem (A.5) 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*

$$\rho^2-\langle\widehat{\Sigma},(I_d-\gamma^\star(\gamma^\star I_d-D)^{-1})^2\rangle=0 \tag{A.7}$$

in the interval $(\lambda_{\max}(D),\infty)$, the matrix $\Sigma^\star$ defined in assertion (iii) is the unique maximizer of (A.5), the Gelbrich distance constraint in (A.5) is binding at $\Sigma^\star$, and $\Sigma^\star\succ\lambda_{\min}(\widehat{\Sigma})I_d$. Moreover, we have

$$\underline{\gamma}=\lambda_1\Big(1+\sqrt{v_1^\top\widehat{\Sigma}v_1}/\rho\Big)\le\gamma^\star\le\lambda_1\Big(1+\sqrt{\mathrm{Tr}[\widehat{\Sigma}]}/\rho\Big)=\bar{\gamma},$$

where λ_1 is the largest eigenvalue of D, and v_1 is an eigenvector corresponding to λ_1.

Proof of Proposition A.2. As for assertion (i), note that the Lagrangian dual of (A.5) can be represented as

$$\inf_{\gamma\ge 0}\ \sup_{\Sigma\succeq 0}\ \langle D,\Sigma\rangle-\gamma\mathrm{Tr}[\Sigma+\widehat{\Sigma}-2(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}]+\gamma\rho^2. \tag{A.8}$$

Strong duality as well as primal solvability follow from Bertsekas [8, proposition 5.5.4], which applies because the primal problem (A.5) has a continuous objective function and—by virtue of Lemma A.4—a nonempty, compact, and convex feasible set. The postulated reformulation (A.6) then follows immediately from replacing the supremum of the inner maximization problem in (A.8) with the analytical formula derived in Proposition A.1. We emphasize that, for $\gamma=\lambda_{\max}(D)$, depending on the problem data, the inner supremum in (A.8) 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 (A.6) without affecting the problem's infimum.

As for assertion (ii), note that $\Sigma=\widehat{\Sigma}$ represents a Slater point for the primal problem (A.5) because $\rho>0$. Thus, the dual problem (A.8) is solvable by Bertsekas [8, proposition 5.3.1]. To prove that (A.6) is also solvable, it remains to be shown that (A.8) does not attain its maximum at the boundary point $\gamma=\lambda_{\max}(D)$, which has been excluded from the feasible set of (A.6). 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 (A.8) amounts to $+\infty$. We may, thus, conclude that (A.6) admits a minimizer $\gamma^\star\in(\lambda_{\max}(D),\infty)$. This minimizer is unique because the objective function of (A.6) is strictly convex when $\widehat{\Sigma}\succ 0$. Finally, the Karush–Kuhn–Tucker optimality conditions Bertsekas [8, proposition 5.3.2] imply that any solution of the primal problem (A.5) also solves the inner maximization problem in (A.8) at $\gamma=\gamma^\star$. The formula for $\Sigma^\star$, thus, follows from Proposition A.1.

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 (A.6) are binding at optimality. As the objective function of (A.6) is smooth and strictly convex, $\gamma^\star$ is, thus, uniquely determined by the first-order optimality condition (A.7), 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 (A.8); see Proposition A.1. Moreover, as $\gamma^\star>0$, the Gelbrich distance constraint in (A.6) is binding at $\Sigma_\rho^\star$ because of complementary slackness. Next, we have

$$\begin{aligned}\frac{1}{\lambda_{\min}(\Sigma^\star)}&=\lambda_{\max}\Big((\Sigma^\star)^{-1}\Big)=\lambda_{\max}\Big(\Big(\gamma^{\star 2}(\gamma^\star I_n-D)^{-1}\widehat{\Sigma}(\gamma^\star I_d-D)^{-1}\Big)^{-1}\Big)\\&=\lambda_{\max}\Big((I_d-D/\gamma^\star)\widehat{\Sigma}^{-1}(I_d-D/\gamma^\star)\Big)\\&\le\lambda_{\max}(I_d-D/\gamma^\star)^2\,\lambda_{\max}\Big(\widehat{\Sigma}^{-1}\Big)<\lambda_{\max}\Big(\widehat{\Sigma}^{-1}\Big)=\frac{1}{\lambda_{\min}(\widehat{\Sigma})},\end{aligned}$$

where the strict inequality holds because $\gamma^\star>\lambda_{\max}(D)>0$. Thus, we conclude that $\lambda_{\min}(\Sigma^\star)>\lambda_{\min}(\widehat{\Sigma})$.

To in order derive upper and lower bounds on $\gamma^\star$, we let $D = \sum_{i=1}^d \lambda_i v_i v_i^\top$ be the eigendecomposition of D, where $\lambda_1, \ldots, \lambda_d$ denote the eigenvalues of D indexed in descending order, and $v_1, \ldots, v_d$ represent the corresponding normalized eigenvectors. The left-hand side of (A.7) can, thus, be re-expressed as

$$\rho^2 - \sum_{i=1}^d \left(\frac{\lambda_i}{\gamma - \lambda_i}\right)^2 v_i^\top \widehat{\Sigma} v_i.$$

This expression is manifestly nondecreasing in $\gamma \in (\lambda_1, \infty)$. Moreover, the sum admits the simple bounds

$$\left(\frac{\lambda_1}{\gamma - \lambda_1}\right)^2 v_1^\top \widehat{\Sigma} v_1 \le \sum_{i=1}^d \left(\frac{\lambda_i}{\gamma - \lambda_i}\right)^2 v_i^\top \widehat{\Sigma} v_i \le \left(\frac{\lambda_1}{\gamma - \lambda_1}\right)^2 \mathrm{Tr}[\widehat{\Sigma}].$$

Equating these lower and upper bounds to ρ^2 and solving the resulting equation for γ yields $\underline{\gamma}$ and $\bar{\gamma}$, respectively. This observation concludes the proof. □

In Sections 3 and 4, we repeatedly encounter nonlinear SDPs of the form

$$\begin{array}{ll} \sup\limits_{\Sigma \succeq 0} & \inf\limits_{L \in \mathcal{C}} \ \langle L^\top L, \Sigma \rangle + f(L) \\ \text{s.t.} & \mathrm{Tr}\left[\Sigma + \widehat{\Sigma} - 2\left(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \le \rho^2 \end{array} \tag{A.9}$$

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 (A.9) is reminiscent of (A.5) but accommodates a nonlinear convex objective function. We do not attempt to characterize the maximizers of (A.9) 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 zero.

Lemma A.3 (Structural Properties of the Maximizers of (A.9)). *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 (A.9) admits a maximizer $\Sigma^\star \succeq \lambda_{\min}(\widehat{\Sigma}) I_d$.*

Proof of Lemma A.3. 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 (A.9) by

$$\mathcal{S} = \left\{\Sigma \in \mathbb{S}_+^d : \mathrm{Tr}[\Sigma + \widehat{\Sigma} - 2(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}] \le \rho^2\right\},$$

we then find

$$\begin{aligned} \sup_{\Sigma \in \mathcal{S}} \inf_{L \in \mathcal{C}} \langle L^\top L, \Sigma \rangle + f(L) &= \inf_{L \in \mathcal{C}} \sup_{\Sigma \in \mathcal{S}} \langle L^\top L, \Sigma \rangle + f(L) \\ &= \inf_{L \in \mathcal{C}} \sup_{\substack{\Sigma \in \mathcal{S} \\ \Sigma \succeq \lambda_{\min}(\widehat{\Sigma}) I_d}} \langle L^\top L, \Sigma \rangle + f(L) = \sup_{\substack{\Sigma \in \mathcal{S} \\ \Sigma \succeq \lambda_{\min}(\widehat{\Sigma}) I_d}} \inf_{L \in \mathcal{C}} \langle L^\top L, \Sigma \rangle + f(L), \end{aligned}$$

where the first and the third equality follow from Sion's [68] minimax theorem, which applies because $\langle L^\top L, \Sigma \rangle$ is convex and continuous in L for every fixed $\Sigma \succeq 0$ and because $\mathcal{S}$ is convex and compact by virtue of Lemma A.4. The second equality follows readily from Proposition A.2(iii). The last maximization problem in the preceding 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 (A.9), and thus, the claim follows. □

The proofs of Proposition A.2 and Lemma A.3 rely on the following auxiliary lemma, which extends Shafieezadeh-Abadeh et al. [67, lemma A.6] to situations in which $\widehat{\Sigma}$ may be an arbitrary positive semidefinite covariance matrix.

Lemma A.4 (Compactness of the Feasible Set). *For any $\widehat{\Sigma} \in \mathbb{S}_+^d$ and $\rho \in \mathbb{R}_+$, the set*

$$\mathcal{S} = \left\{\Sigma \in \mathbb{S}_+^d : \mathrm{Tr}[\Sigma + \widehat{\Sigma} - 2(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}] \le \rho^2\right\}$$

is convex and compact. Moreover, for any $\Sigma \in \mathcal{S}$, we have $\mathrm{Tr}[\Sigma] \le (\rho + \mathrm{Tr}[\widehat{\Sigma}]^{\frac{1}{2}})^2$.

Proof of Lemma A.4. The convexity of the feasible set $\mathcal{S}$ follows from the convexity of the squared Gelbrich distance proven in Proposition 2.2. To prove that $\mathcal{S}$ is compact, we recall from Malagò et al. [52, proposition 2] that

$$\begin{aligned} \mathrm{Tr}[(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}] &= \max_{C \in \mathbb{R}^{d \times d}} \left\{\mathrm{Tr}[C] : \begin{bmatrix} \Sigma & C \\ C^\top & \widehat{\Sigma} \end{bmatrix} \succeq 0\right\} \\ &\le \max_{C \in \mathbb{R}^{d \times d}} \left\{\mathrm{Tr}[C] : C_{ij}^2 \le \Sigma_{ii} \widehat{\Sigma}_{jj} \ \ \forall i, j = 1, \ldots, d\right\} = \sum_{i=1}^d \sqrt{\Sigma_{ii} \widehat{\Sigma}_{ii}} \le \sqrt{\mathrm{Tr}[\Sigma] \mathrm{Tr}[\widehat{\Sigma}]}, \end{aligned}$$

where the first inequality holds because all second principal minors of a positive semidefinite matrix are nonnegative, and the second inequality follows from the Cauchy–Schwarz inequality. Thus, any $\Sigma \in \mathcal{S}$ satisfies

$$\rho^2 \geq \mathrm{Tr}[\Sigma + \widehat{\Sigma} - 2(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}})] \geq \left(\sqrt{\mathrm{Tr}[\Sigma]} - \sqrt{\mathrm{Tr}[\widehat{\Sigma}]}\right)^2,$$

which implies that $\mathrm{Tr}[\Sigma] \leq (\rho + \mathrm{Tr}[\widehat{\Sigma}]^{\frac{1}{2}})^2$. This allows us to conclude that $\mathcal{S}$ is bounded. Moreover, $\mathcal{S}$ is closed because of the continuity of the matrix square root established in Lemma A.2. □

Finally, we derive the second-order Taylor expansion of the objective function

$$f(\Sigma_x, \Sigma_w) = \mathrm{Tr}[\Sigma_x - \Sigma_x H^\top (H\Sigma_x H^\top + \Sigma_w)^{-1} H\Sigma_x]$$

of the nonlinear SDP (6.13), which is needed for the proof of Proposition 6.1.

Lemma A.5 (Gradient and Hessian of *f*). *If* $(\Sigma_x, \Sigma_w) \in \mathbb{S}_{++}^n \times \mathbb{S}_{++}^m$ *and* $G = H\Sigma_x H^\top + \Sigma_w \in \mathbb{S}_{++}^m$, *then*

$$\begin{aligned}
D_x &= \nabla_{\Sigma_x} f(\Sigma_x, \Sigma_w) = (I_n - \Sigma_x H^\top G^{-1} H)^\top (I_n - \Sigma_x H^\top G^{-1} H) \\
D_w &= \nabla_{\Sigma_w} f(\Sigma_x, \Sigma_w) = G^{-1} H \Sigma_x^2 H^\top G^{-1} \\
\mathcal{H}_{xx} &= -\nabla_{xx}^2 f(\Sigma_x, \Sigma_w) = 2D_x \otimes H^\top G^{-1} H \\
\mathcal{H}_{xw} &= -\nabla_{xw}^2 f(\Sigma_x, \Sigma_w) = H^\top G^{-1} \otimes (H^\top D_w - \Sigma_x H^\top G^{-1}) + (H^\top D_w - \Sigma_x H^\top G^{-1}) \otimes H^\top G^{-1} \\
\mathcal{H}_{ww} &= -\nabla_{ww}^2 f(\Sigma_x, \Sigma_w) = 2D_w \otimes G^{-1},
\end{aligned}$$

where ∇_x *and* ∇_w *stand for the nabla operators with respect to* $\mathrm{vec}(\Sigma_x)$ *and* $\mathrm{vec}(\Sigma_w)$, *respectively.*

Proof of Lemma A.5. We first derive the second-order Taylor expansion of G^{-1}. Specifically, if $\Delta_x \in \mathbb{S}^n$ and $\Delta_w \in \mathbb{S}^m$ represent symmetric perturbation directions of Σ_x and Σ_w, respectively, then we have

$$\begin{aligned}
&(H[\Sigma_x + t\Delta_x]H^\top + [\Sigma_w + t\Delta_w])^{-1} \\
&\quad = (G + t[H\Delta_x H^\top + \Delta_w])^{-1} \\
&\quad = G^{-\frac{1}{2}}(I_m + tG^{-\frac{1}{2}}[H\Delta_x H^\top + \Delta_w]G^{-\frac{1}{2}})^{-1}G^{-\frac{1}{2}} \\
&\quad = G^{-\frac{1}{2}}(I_m - tG^{-\frac{1}{2}}[H\Delta_x H^\top + \Delta_w]G^{-\frac{1}{2}} + t^2(G^{-\frac{1}{2}}[H\Delta_x H^\top + \Delta_w]G^{-\frac{1}{2}})^2 + \mathcal{O}(|t|^3))G^{-\frac{1}{2}} \\
&\quad = G^{-1} - tG^{-1}[H\Delta_x H^\top + \Delta_w]G^{-1} + t^2 G^{-1}[H\Delta_x H^\top + \Delta_w]G^{-1}[H\Delta_x H^\top + \Delta_w]G^{-1} + \mathcal{O}(|t|^3),
\end{aligned}$$

where the third equality follows from a Neumann series expansion Bernstein [7, proposition 9.4.13]. Thanks to Bernstein [7, fact 7.4.9], the second-order Taylor expansion of f can, thus, be represented as

$$\begin{aligned}
&f(\Sigma_x + t\Delta_x, \Sigma_w + t\Delta_w) \\
&\quad = \mathrm{Tr}[[\Sigma_x + t\Delta_x] - [\Sigma_x + t\Delta_x]H^\top (H[\Sigma_x + t\Delta_x]H^\top + [\Sigma_w + t\Delta_w])^{-1} H[\Sigma_x + t\Delta_x]] \\
&\quad = f(\Sigma_x, \Sigma_w) + t\langle D_x, \Delta_x\rangle + t\langle D_w, \Delta_w\rangle - \frac{t^2}{2}\begin{pmatrix}\mathrm{vec}(\Delta_x)\\ \mathrm{vec}(\Delta_w)\end{pmatrix}^\top \begin{pmatrix}\mathcal{H}_{xx} & \mathcal{H}_{xw}\\ \mathcal{H}_{xw}^\top & \mathcal{H}_{ww}\end{pmatrix}\begin{pmatrix}\mathrm{vec}(\Delta_x)\\ \mathrm{vec}(\Delta_w)\end{pmatrix} + \mathcal{O}(|t|^3),
\end{aligned}$$

where the matrices D_x, D_w, $\mathcal{H}_{xx}$, $\mathcal{H}_{xw}$ and $\mathcal{H}_{ww}$ are defined as in the statement of the lemma.

Endnote

[1] We say that $A^\star$ solves (2.2) if adding the constraint $A = A^\star$ does not change the infimum of (2.2). Note that the infimum of the resulting problem over (γ_x, γ_w) may not be attained, that is, the existence of a solution $A^\star$ does not imply that (2.2) is solvable.

References

[1] Aubin J-P, Frankowska H (1990) *Set-Valued Analysis* (Birkhäuser).
[2] Aviv Y (2003) A time-series framework for supply chain inventory management. *Oper. Res.* 51(2):210–227.
[3] Beck A, Eldar YC (2007) Regularization in regression with bounded noise: A Chebyshev center approach. *SIAM J. Matrix Anal. Appl.* 29 (2):606–625.
[4] Beck A, Ben-Tal A, Eldar YC (2006) Robust mean-squared error estimation of multiple signals in linear systems affected by model and noise uncertainties. *Math. Programming* 107:155–187.
[5] Beck A, Eldar YC, Ben-Tal A (2007) Mean-squared error estimation for linear systems with block circulant uncertainty. *SIAM J. Matrix Anal. Appl.* 29(3):712–730.
[6] Berge C (1963) *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity* (Courier Corporation).
[7] Bernstein DS (2009) *Matrix Mathematics: Theory, Facts, and Formulas* (Princeton University Press).
[8] Bertsekas D (2009) *Convex Optimization Theory* (Athena Scientific).

[9] Bhatia R, Jain T, Lim Y (2018) Strong convexity of sandwiched entropies and related optimization problems. *Rev. Math. Phys.* 30(9).
[10] Boyd S, Vandenberghe L (2004) *Convex Optimization* (Cambridge University Press).
[11] Cambanis S, Huang S, Simons G (1981) On the theory of elliptically contoured distributions. *J. Multivariate Anal.* 11(3):368–385.
[12] Chang SG, Yu B, Vetterli M (2000) Adaptive wavelet thresholding for image denoising and compression. *IEEE Trans. Image Processing* 9(9):1532–1546.
[13] Chatfield C (2016) *The Analysis of Time Series: An Introduction* (CRC Press).
[14] Clarkson KL (2010) Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Trans. Algorithms* 6(4):1–30.
[15] Cover TM, Thomas JA (2006) *Elements of Information Theory* (Wiley-Interscience).
[16] Cuturi M (2013) Sinkhorn distances: Lightspeed computation of optimal transport. *Adv. Neural Inform. Processing Systems* 26:2292–2300.
[17] Demyanov VF, Rubinov AM (1970) *Approximate Methods in Optimization Problems* (American Elsevier Publishing).
[18] Diggavi SN, Cover TM (2001) The worst additive noise under a covariance constraint. *IEEE Trans. Inform. Theory* 47(7):3072–3081.
[19] Dowson DC, Landau BV (1982) The Fréchet distance between multivariate normal distributions. *J. Multivariate Anal.* 12(3):450–455.
[20] Dunn JC (1979) Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. *SIAM J. Control Optim.* 17(2):187–211.
[21] Dunn JC (1980) Convergence rates for conditional gradient sequences generated by implicit step length rules. *SIAM J. Control Optim.* 18 (5):473–487.
[22] Dunn JC, Harshbarger S (1978) Conditional gradient algorithms with open loop step size rules. *J. Math. Anal. Appl.* 62(2):432–444.
[23] Eldar YC (2006) Robust competitive estimation with signal and noise covariance uncertainties. *IEEE Trans. Inform. Theory* 52(10):4532–4547.
[24] Eldar YC, Merhav N (2004) A competitive minimax approach to robust estimation of random parameters. *IEEE Trans. Signal Processing* 52(7):1931–1946.
[25] Eldar YC, Beck A, Teboulle M (2008) A minimax Chebyshev estimator for bounded error estimation. *IEEE Trans. Signal Processing* 56(4): 1388–1397.
[26] Eldar YC, Ben-Tal A, Nemirovski A (2004) Linear minimax regret estimation of deterministic parameters with bounded data uncertainties. *IEEE Trans. Signal Processing* 52(8):2177–2188.
[27] Eldar YC, Ben-Tal A, Nemirovski A (2004) Robust mean-squared error estimation in the presence of model uncertainties. *IEEE Trans. Signal Processing* 53(1):168–181.
[28] Fang K-T, Kotz S, Ng KW (1990) *Symmetric Multivariate and Related Distributions* (Chapman & Hall).
[29] Frank M, Wolfe P (1956) An algorithm for quadratic programming. *Naval Res. Logist.* 3(1–2):95–110.
[30] Freund RM, Grigas P (2016) New analysis and results for the Frank–Wolfe method. *Math. Programming* 155:199–230.
[31] Garber D, Hazan E (2015) Faster rates for the Frank–Wolfe method over strongly-convex sets. *Internat. Conf. Machine Learn.*, 541–549.
[32] Gelbrich M (1990) On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten* 147(1):185–203.
[33] Givens CR, Shortt RM (1984) A class of Wasserstein metrics for probability distributions. *Michigan Math. J.* 31(2):231–240.
[34] Golnaraghi F, Kuo B (2017) *Automatic Control Systems* (McGraw-Hill Education).
[35] Guo D, Wu Y, Shitz SS, Verdu S (2011) Estimation in Gaussian noise: Properties of the minimum mean-square error. *IEEE Trans. Inform. Theory* 57(4):2371–2385.
[36] Hamilton J (1994) *Time Series Analysis* (Princeton University Press).
[37] Hespanha JP (2009) *Linear Systems Theory* (Princeton University Press).
[38] Hult H, Lindskog F (2002) Multivariate extremes, aggregation and dependence in elliptical distributions. *Adv. Appl. Probab.* 34(3):587–608.
[39] Jaggi M (2013) *Revisiting Frank–Wolfe: Projection-free sparse convex optimization*. Proc. 30th Internat. Conf. Machine Learn., 427–435.
[40] Jondeau E, Poon S, Rockinger M (2007) *Financial Modeling under Non-Gaussian Distributions* (Springer).
[41] Journée M, Nesterov Y, Richtárik P, Sepulchre R (2010) Generalized power method for sparse principal component analysis. *J. Machine Learn. Res.* 11:517–553.
[42] Juditsky A, Nemirovski A (2018) *Lectures on Statistical Inferences via Convex Optimization*.
[43] Juditsky A, Nemirovski A (2018) Near-optimality of linear recovery in Gaussian observation scheme under $|\cdot|_2^2$-loss. *Ann. Statist.* 46(4): 1603–1629.
[44] Kay SM (1993) *Fundamentals of Statistical Signal Processing: Estimation Theory* (Prentice Hall).
[45] Kelker D (1970) Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhya Ser. A.* 32(4): 419–430.
[46] Knott M, Smith CS (1984) On the optimal mapping of distributions. *J. Optim. Theory Appl.* 43:39–49.
[47] Levitin ES, Polyak BT (1966) Constrained minimization methods. *USSR Comput. Math. Math. Phys.* 6(5):1–50.
[48] Levy BC, Nikoukhah R (2004) Robust least-squares estimation with a relative entropy constraint. *IEEE Trans. Inform. Theory* 50(1): 89–104.
[49] Levy BC, Nikoukhah R (2012) Robust state space filtering under incremental model perturbations subject to a relative entropy tolerance. *IEEE Trans. Automatic Control* 58(3):682–695.
[50] Löfberg J (2004) YALMIP: A toolbox for modeling and optimization in MATLAB. *IEEE Internat. Conf. Robotics Automation*, 284–289.
[51] MacKay D (2003) *Information Theory, Inference and Learning Algorithms* (Cambridge University Press).
[52] Malagò L, Montrucchio L, Pistone G (2018) Wasserstein Riemannian geometry of Gaussian densities. *Inform. Geometry* 1:137–179.
[53] Murphy K (2012) *Machine Learning: A Probabilistic Perspective* (MIT Press).
[54] Nesterov Y (2018) Complexity bounds for primal-dual methods minimizing the model of objective function. *Math. Programming* 171: 311–330.
[55] Nguyen VA, Kuhn D, Mohajerin Esfahani P (2021) Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *Oper. Res.*, ePub ahead of print July 23, https://doi.org/10.1287/opre.2020.2076.
[56] Ogata K (2009) *Modern Control Engineering* (Pearson).
[57] Olkin I, Pukelsheim F (1982) The distance between two random vectors with given dispersion matrices. *Linear Algebra Appl.* 48:257–263.
[58] Oppenheim AV, Verghese GC (2015) *Signals, Systems and Inference* (Pearson).

[59] Pedregosa F, Negiar G, Askari A, Jaggi M (2020) Linearly convergent Frank-Wolfe with backtracking line-search. *Internat. Conf. Artificial Intelligence Statist.*, 1–10.
[60] Peyré G, Cuturi M (2019) Computational optimal transport. *Foundations and Trends® in Machine Learning*, vol. 11, 355–607. https://www.nowpublishers.com/article/Details/MAL-073.
[61] Posekany A, Felsenstein K, Sykacek P (2011) Biological assessment of robust noise models in microarray data analysis. *Bioinformatics* 27 (6):807–814.
[62] Rockafellar RT, Wets RJ-B (2010) *Variational Analysis* (Springer).
[63] Rubel O, Naik PA (2017) Robust dynamic estimation. *Marketing Sci.* 36(3):453–467.
[64] Ruttimann UE, Unser M, Rawlings RR, Rio D, Ramsey NF, Mattay VS, Hommer DW, Frank JA, Weinberger DR (1998) Statistical analysis of functional MRI data in the wavelet domain. *IEEE Trans. Medical Imaging* 17(2):142–154.
[65] Schmitt BA (1992) Perturbation bounds for matrix square roots and Pythagorean sums. *Linear Algebra Appl.* 174:215–227.
[66] Shafieezadeh-Abadeh S, Mohajerin Esfahani P, Kuhn D (2019) Regularization via mass transportation. *J. Machine Learn. Res.* 20(103):1–68.
[67] Shafieezadeh-Abadeh S, Nguyen V, Kuhn D, Mohajerin Esfahani P (2018) Wasserstein distributionally robust Kalman filtering. *Adv. Neural Inform. Processing Systems* 31:8483–8492.
[68] Sion M (1958) On general minimax theorems. *Pacific J. Math.* 8(1):171–176.
[69] Solomon J, Goes FD, Peyré G, Cuturi M, Butscher A, Nguyen A, Du T, Guibas L (2015) Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graphics* 34(4):1–11.
[70] Sriram S, Kalwani MU (2007) Optimal advertising and promotion budgets in dynamic markets with brand equity as a mediating variable. *Management Sci.* 53(1):46–60.
[71] Stock J, Watson M (2015) *Introduction to Econometrics* (Prentice Hall).
[72] Taşkesen B, Shafieezadeh Abadeh S, Kuhn D (2021) Semi-discrete optimal transport: Hardness, regularization and numerical solution. Preprint, submitted March 10, https://arxiv.org/abs/2103.06263.
[73] van Lint H, Djukic T (2012) Applications of Kalman filtering in traffic management and control. Mirchandani P, Cole Smith J, eds. *New Directions in Informatics, Optimization, Logistics, and Production* (INFORMS, Hanover, PA), 59–91.
[74] Wooldridge J (2010) *Econometric Analysis of Cross Section and Panel Data* (MIT Press).
[75] Zhou J (1995) On the existence of equilibrium for abstract economies. *J. Math. Anal. Appl.* 193(3):839–858.
[76] Zorzi M (2016) Robust Kalman filtering under model perturbations. *IEEE Trans. Automatic Control* 62(6):2902–2907.
[77] Zorzi M (2017) On the robustness of the Bayes and Wiener estimators under model uncertainty. *Automatica* 83:133–140.