Wasserstein Distributionally Robust Kalman Filtering

Soroosh Shafieezadeh-Abadeh **Viet Anh Nguyen** **Daniel Kuhn**
École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland
{soroosh.shafiee,viet-anh.nguyen,daniel.kuhn} @epfl.ch

Peyman Mohajerin Esfahani
Delft Center for Systems and Control, TU Delft, The Netherlands
P.MohajerinEsfahani@tudelft.nl

Abstract

We study a distributionally robust mean square error estimation problem over a nonconvex Wasserstein ambiguity set containing only normal distributions. We show that the optimal estimator and the least favorable distribution form a Nash equilibrium. Despite the non-convex nature of the ambiguity set, we prove that the estimation problem is equivalent to a tractable convex program. We further devise a Frank-Wolfe algorithm for this convex program whose direction-searching subproblem can be solved in a quasi-closed form. Using these ingredients, we introduce a distributionally robust Kalman filter that hedges against model risk.

1 Introduction

The Kalman filter is the workhorse for the online tracking and estimation of a dynamical system's internal state based on indirect observations [1]. It has been applied with remarkable success in areas as diverse as automatic control, brain-computer interaction, macroeconomics, robotics, signal processing, weather forecasting and many more. The classical Kalman filter critically relies on the availability of an accurate state-space model and is therefore susceptible to model risk. This observation has led to several attempts to robustify the Kalman filter against modeling errors.

The $\mathcal{H}_\infty$-filter targets situations in which the statistics of the noise process is uncertain and where one aims to minimize the worst case instead of the variance of the estimation error [3, 29]. This filter bounds the $\mathcal{H}_\infty$-norm of the transfer function that maps the disturbances to the estimation errors. However, in transient operation, the desired $\mathcal{H}_\infty$-performance is lost, and the filter may diverge unless some (typically restrictive) positivity condition holds in each iteration. In set-valued estimation the disturbance vectors are modeled through bounded sets such as ellipsoids [6, 25]. In this framework, one attempts to construct the smallest ellipsoids around the state estimates that are consistent with the observations and the exogenous disturbance ellipsoids. However, the resulting robust filters ignore any distributional information and thus have a tendency to be over-conservative. A filter that is robust against more general forms of (set-based) model uncertainty was first studied in [22]. This filter iteratively minimizes the worst-case mean square error across all models in the vicinity of a nominal state space model. While performing well in the face of large uncertainties, this filter may be too conservative under small uncertainties. A generalized Kalman filter that addresses this shortcoming and strikes the balance between nominal and worst-case performance has been proposed in [28]. A risk-sensitive Kalman filter is obtained by minimizing the moment-generating function instead of the mean of the squared estimation error [27]. This risk-sensitive Kalman filter is equivalent to a distributionally robust filter proposed in [15], which minimizes the worst-case mean square error across all joint state-output distributions in a Kullback-Leibler (KL) ball around a nominal distribution. Extensions to more general τ-divergence balls are investigated in [30].

Preprint. Work in progress.

In this paper we use ideas from distributionally robust optimization to design a Kalman-type filter that is immunized against model risk. Specifically, we assume that the joint distribution of the states and outputs is uncertain but known to reside in a given *ambiguity set* that contains all distributions in the proximity of the nominal distribution generated by a nominal state-space model. The ambiguity set thus reflects our level of (dis)trust in the nominal model. We then construct the most accurate filter under the least favorable distribution in this set. The hope is that hedging against the worst-case distribution has a regularizing effect and will lead to a filter that performs well under the unknown true distribution. Distributionally robust filters of this type have been studied in [10, 19] using uncertainty sets for the covariance matrix of the state vector and in [15, 30] using ambiguity sets defined via information divergences. Inspired by recent progress in data-driven distributionally robust optimization [17], we construct here the ambiguity set as a ball around the nominal distribution with respect to the type-2 Wasserstein distance. The Wasserstein distance has seen widespread application in machine learning [2, 9, 21], and an intimate relation between regularization and Wasserstein distributional robustness has been discovered in [24, 23, 26, 18]. Also, the Wasserstein distance is known to be more statistically robust than other information divergences [8].

We summarize our main contributions as follows:

- We introduce a distributionally robust mean square estimation problem over a nonconvex Wasserstein ambiguity set containing *normal* distributions only, and we demonstrate that the optimal estimator and the least favorable distribution form a Nash equilibrium.
- Leveraging modern reformulation techniques from [18], we prove that this problem is equivalent to a tractable convex program—despite the nonconvex nature of the underlying ambiguity set—and that the optimal estimator is an affine function of the observations.
- We devise an efficient Frank-Wolfe-type first-order method inspired by [13] to solve the resulting convex program. We show that the direction-finding subproblem can be solved in quasi-closed form, and we derive the algorithm's convergence rate.
- We introduce a Wasserstein distributionally robust Kalman filter that hedges against model risk. The filter can be computed efficiently by solving a sequence of robust estimation problems via the proposed Frank-Wolfe algorithm. Its performance is validated on standard test instances.

All proofs are relegated to Appendix A, and additional numerical results are reported in Appendix B.

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

2 Robust Estimation with Wasserstein Ambiguity Sets

Consider the problem of estimating a signal $x \in \mathbb{R}^n$ from a potentially noisy observation $y \in \mathbb{R}^m$. In practice, the joint distribution of x and y is never directly observable and thus fundamentally uncertain. This distributional uncertainty should be taken into account in the estimation procedure. In this paper, we model distributional uncertainty through an *ambiguity set* $\mathcal{P}$, that is, a family of distributions on $\mathbb{R}^d$, $d = n + m$, that are sufficiently likely to govern x and y in view of the available data or that are sufficiently close to a prescribed nominal distribution. We then seek a robust estimator that minimizes the worst-case mean square error across all distributions in the ambiguity set. In the following, we propose to use the Wasserstein distance in order to construct ambiguity sets.

Definition 2.1 (Wasserstein distance)**.** The type-2 Wasserstein distance between two distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$ on $\mathbb{R}^d$ is defined as

$$W_2(\mathbb{Q}_1, \mathbb{Q}_2) \triangleq \inf_{\pi\in\Pi(\mathbb{Q}_1,\mathbb{Q}_2)} \left\{ \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} \|z_1 - z_2\|^2 \, \pi(\mathrm{d}\, z_1, \mathrm{d}\, z_2) \right)^{\frac{1}{2}} \right\}, \tag{1}$$

where $\Pi(\mathbb{Q}_1, \mathbb{Q}_2)$ is the set of all probability distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals $\mathbb{Q}_1$ and $\mathbb{Q}_2$.

Proposition 2.2 ([12, Proposition 7]). The type-2 Wasserstein distance between two normal distributions $\mathbb{Q}_1 = \mathcal{N}_d(\mu_1, \Sigma_1)$ and $\mathbb{Q}_2 = \mathcal{N}_d(\mu_2, \Sigma_2)$ with $\mu_1, \mu_2 \in \mathbb{R}^d$ and $\Sigma_1, \Sigma_2 \in \mathbb{S}_+^d$ equals

$$W_2(\mathbb{Q}_1, \mathbb{Q}_2) = \sqrt{\|\mu_1 - \mu_2\|^2 + \operatorname{Tr}\left[\Sigma_1 + \Sigma_2 - 2\left(\Sigma_2^{\frac{1}{2}}\Sigma_1\Sigma_2^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]}.$$

Consider now a d-dimensional random vector $z = [x^\top, y^\top]^\top$ comprising the signal $x \in \mathbb{R}^n$ and the observation $y \in \mathbb{R}^m$, where $d = n + m$. For a given ambiguity set $\mathcal{P}$, the *distributionally robust minimum mean square error estimator* of x given y is a solution of the outer minimization problem in

$$\inf_{\psi \in \mathcal{L}} \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}^{\mathbb{Q}} \left[\|x - \psi(y)\|^2\right], \tag{2}$$

where $\mathcal{L}$ denotes the family of all measurable functions from $\mathbb{R}^m$ to $\mathbb{R}^n$. Problem (2) can be viewed as a zero-sum game between a statistician choosing the estimator ψ and a fictitious adversary (or nature) choosing the distribution $\mathbb{Q}$. By construction, the minimax estimator performs best under the worst possible distribution $\mathbb{Q} \in \mathcal{P}$. From now on we assume that $\mathcal{P}$ is the Wasserstein ambiguity set

$$\mathcal{P} = \Big\{\mathbb{Q} \in \mathcal{N}_d \,:\, W_2(\mathbb{Q}, \mathbb{P}) \leq \rho\Big\}, \tag{3}$$

which can be interpreted as a ball of radius $\rho \geq 0$ in the space of normal distributions. We will further assume that $\mathcal{P}$ is centered at a normal distribution $\mathbb{P} = \mathcal{N}_d(\mu, \Sigma)$ with covariance matrix $\Sigma \succ 0$.

Even though the Wasserstein ambiguity set $\mathcal{P}$ is nonconvex (as mixtures of normal distributions are generically not normal), we can prove a minimax theorem, which ensures that one may interchange the infimum and the supremum in (2) without affecting the problem's optimal value.

Theorem 2.3 (Minimax theorem). If $\mathcal{P}$ is a Wasserstein ambiguity set of the form (3), then

$$\inf_{\psi \in \mathcal{L}} \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}^{\mathbb{Q}} \left[\|x - \psi(y)\|^2\right] = \sup_{\mathbb{Q} \in \mathcal{P}} \inf_{\psi \in \mathcal{L}} \mathbb{E}^{\mathbb{Q}} \left[\|x - \psi(y)\|^2\right]. \tag{4}$$

Remark 2.4 (Connection to Bayesian estimation). The optimal solutions $\psi^\star$ and $\mathbb{Q}^\star$ of the two dual problems in (4) represent the minimax strategies of the statistician and nature, respectively. Theorem 2.3 implies that $(\psi^\star, \mathbb{Q}^\star)$ forms a saddle point (and thus a Nash equilibrium) of the underlying zero-sum game. Hence, the robust estimator $\psi^\star$ is also the optimal Bayesian estimator for the prior $\mathbb{Q}^\star$. For this reason, $\mathbb{Q}^\star$ is often referred to as the *least favorable prior* [14].

We now demonstrate that the minimax problem (2) is equivalent to a tractable convex program, whose solution allows us to recover both the optimal estimator $\psi^\star$ as well as the least favorable prior $\mathbb{Q}^\star$.

Theorem 2.5 (Tractable reformulation). The minimax problem (2) with the Wasserstein ambiguity set (3) centered at $\mathbb{P} = \mathcal{N}_d(\mu, \Sigma)$, $\underline{\sigma} \triangleq \lambda_{\min}(\Sigma) > 0$, is equivalent to the finite convex program

$$\begin{array}{ll} \sup & \operatorname{Tr}\left[S_{xx} - S_{xy}S_{yy}^{-1}S_{yx}\right] \\ \text{s.t.} & S = \begin{bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{bmatrix} \in \mathbb{S}_+^d, \quad S_{xx} \in \mathbb{S}_+^n, \quad S_{yy} \in \mathbb{S}_+^m, \quad S_{xy} = S_{yx}^\top \in \mathbb{R}^{n \times m} \\ & \operatorname{Tr}\left[S + \Sigma - 2\left(\Sigma^{\frac{1}{2}} S \Sigma^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \leq \rho^2, \quad S \succeq \underline{\sigma} I_d. \end{array} \tag{5}$$

If $S^\star$, $S_{xx}^\star$, $S_{yy}^\star$ and $S_{xy}^\star$ is optimal in (5) and $\mu = [\mu_x^\top, \mu_y^\top]^\top$ for some $\mu_x \in \mathbb{R}^n$ and $\mu_y \in \mathbb{R}^m$, then the affine function $\psi^\star(y) = S_{xy}^\star (S_{yy}^\star)^{-1}(y - \mu_y) + \mu_x$ is the distributionally robust minimum mean square error estimator, and the normal distribution $\mathbb{Q}^\star = \mathcal{N}_d(\mu, S^\star)$ is the least favorable prior.

Theorem 2.5 provides a tractable procedure for constructing a Nash equilibrium $(\psi^\star, S^\star)$ for the statistician's game against nature. Note that if $\rho = 0$, then $S^\star = \Sigma$ is the unique solution to (5). In this case the estimator $\psi^\star$ reduces to the Bayesian estimator corresponding to the nominal distribution $\mathbb{P} = \mathcal{N}_d(\mu, \Sigma)$. We emphasize that the choice of the Wasserstein radius ρ may have a significant impact on the resulting estimator. In fact, this is a key distinguishing feature of the Wasserstein ambiguity set (3) with respect to other popular divergence-based ambiguity sets.

Remark 2.6 (Divergence-based ambiguity sets)**.** As a natural alternative, one could replace the Wasserstein distance in (3) with an information divergence. For example, ambiguity sets defined via τ-divergences, which encapsulate the popular KL divergence as a special case, have been studied in [15, 30]. As shown in [15, Theorem 1] and [30, Theorem 2.1], the optimal estimator corresponding to any τ-divergence ambiguity set always coincides with the Bayesian estimator for the nominal distribution $\mathbb{P} = \mathcal{N}_d(\mu, \Sigma)$ irrespective of ρ. Thus, in stark contrast to the setting considered here, the size of a τ-divergence ambiguity set has no impact on the corresponding optimal estimator. Moreover, the least favorable prior $\mathbb{Q} = \mathcal{N}_d(\mu, S^\star)$ for a τ-divergence ambiguity set always satisfies

$$S^\star = \begin{bmatrix} S^\star_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}. \tag{6}$$

Thus, in order to harm the statistician, nature only perturbs the second moments of the signal but sets all second moments of the observation as well as all cross moments to their nominal values.

Example 2.7 (Impact of ρ on the Nash equilibrium)**.** We illustrate the dependence of the saddle point $(\psi^\star, \mathbb{Q}^\star)$ on the size ρ of the ambiguity set in a 2-dimensional example. Suppose that the nominal distribution $\mathbb{P}$ of $[x, y] \in \mathbb{R}^2$ satisfies $\mu_x = \mu_y = 0$, $\Sigma_{xx} = \Sigma_{xy} = 1$ and $\Sigma_{yy} = 1.1$, implying that the noise $w \triangleq y - x$ and the signal x are independent ($\mathbb{E}^{\mathbb{P}}[xw] = \Sigma_{xy} - \Sigma_{xx} = 0$). Figure 1 visualizes the canonical 90% confidence ellipsoids of the the least favorable priors as well as the graphs of the optimal estimators for different sizes of the Wasserstein and KL ambiguity sets. As ρ increases, the least favorable prior for the Wasserstein ambiguity set displays the following interesting properties: (i) the signal variance $S^\star_{xx}$ increases, (ii) the measurement variance $S^\star_{yy}$ decreases, (iii) the signal-measurement covariance $S^\star_{xy}$ decreases towards 0, and (iv) the noise variance $\mathbb{E}^{\mathbb{Q}^\star}[w^2] = S^\star_{yy} - 2S^\star_{xy} + S^\star_{xx}$ increases. Hence, (v) the signal-noise covariance $\mathbb{E}^{\mathbb{Q}^\star}[xw] = S^\star_{xy} - S^\star_{xx}$ decreases and is *negative* for all $\rho > 0$, and (vi) the optimal estimator $\psi^\star$ tends to the zero function. Note that the optimal estimator and the measurement variance remain constant in ρ when working with a KL ambiguity set.

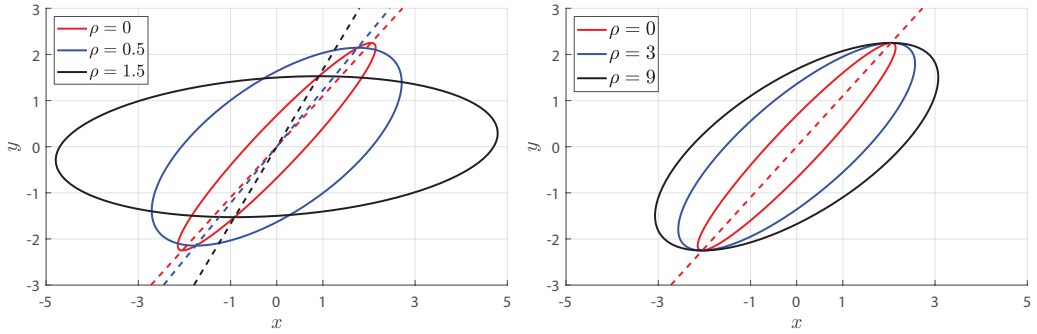


Figure 1: Least favorable priors (solid ellipsoids) and optimal estimators (dashed lines) for Wasserstein (left) and KL (right) ambiguity sets with different radii ρ. The Wasserstein estimators vary with ρ, while the KL estimators remain unaffected by ρ.

Remark 2.8 (Ambiguity sets with non-normal distributions)**.** Theorem 2.3 can be generalized to Wasserstein ambiguity set of the form $\mathcal{Q} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^d) \,:\, W_2(\mathbb{Q}, \mathbb{P}) \leq \rho\}$, where $\mathcal{M}(\mathbb{R}^d)$ denotes the set of all (possibly non-normal) probability distributions on $\mathbb{R}^d$ with finite second moments, and $\mathbb{P} = \mathcal{N}_d(\mu, \Sigma)$. In this case, the minimax result (4) remains valid provided that the set $\mathcal{L}$ of all measurable estimators is restricted to the set $\mathcal{A}$ of all affine estimators. Theorem 2.5 also remains valid under this alternative setting.

3 Efficient Frank-Wolfe Algorithm

The finite convex optimization problem (5) is numerically challenging as it constitutes a *nonlinear* semi-definite program (SDP). In principle, it would be possible to eliminate all nonlinearities by using Schur complements and to reformulate (5) as a *linear* SDP, which is formally tractable. However, it is folklore knowledge that general-purpose SDP solvers are yet to be developed that can reliably solve large-scale problem instances. We thus propose a tailored first-order method to solve the nonlinear SDP (5) directly, which exploits a covert structural property of the problem's objective function

$$f(S) \triangleq \mathrm{Tr}\left[S_{xx} - S_{xy} S_{yy}^{-1} S_{yx}\right].$$

Definition 3.1 (Unit total elasticity[1])**.** We say that a function $\varphi : \mathbb{S}_+^d \to \mathbb{R}_+$ has unit total elasticity if

$$\varphi(S) = \big\langle S, \nabla\varphi(S) \big\rangle \quad \forall S \in \mathbb{S}_+^d.$$

It is clear that every linear function has unit total elasticity. Maybe surprisingly, however, the objective function $f(S)$ of problem (5) also enjoys unit total elasticity because

$$\big\langle S, \nabla f(S) \big\rangle = \left\langle \begin{bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{bmatrix}, \begin{bmatrix} I_n & -S_{xy}S_{yy}^{-1} \\ -S_{yy}^{-1}S_{yx} & S_{yy}^{-1}S_{yx}S_{xy}S_{yy}^{-1} \end{bmatrix} \right\rangle = f(S).$$

Moreover, as will be explained below, it turns out problem (5) can be solved highly efficiently if its objective function is replaced with a linear approximation. These observations motivate us to solve (5) with a Frank-Wolfe algorithm [11], which starts at $S^{(0)} = \Sigma$ and constructs iterates

$$S^{(k+1)} = \alpha_k F\big(S^{(k)}\big) + (1 - \alpha_k) S^{(k)} \quad \forall k \in \mathbb{N} \cup \{0\}, \tag{7a}$$

where α_k represents a judiciously chosen step-size, while the oracle mapping $F : \mathbb{S}_+ \to \mathbb{S}_+$ returns the unique solution of the direction-finding subproblem

$$F(S) \triangleq \left\{ \begin{array}{ll} \underset{L \succeq \underline{\sigma} I_d}{\arg\max} & \big\langle L, \nabla f(S) \big\rangle \\ \text{s.t.} & \mathrm{Tr}\left[L + \Sigma - 2\left(\Sigma^{\frac{1}{2}} L \Sigma^{\frac{1}{2}} \right)^{\frac{1}{2}} \right] \leq \rho^2 \,. \end{array} \right. \tag{7b}$$

In each iteration, the Frank-Wolfe algorithm thus minimizes a linearized objective function over the original feasible set. In contrast to other commonly used first-order methods, the Frank-Wolfe algorithm thus obviates the need for a potentially expensive projection step to recover feasibility. It is easy to convince oneself that any solution of the nonlinear SDP (5) is indeed a fixed point of the operator F. To make the Frank-Wolfe algorithm (7) work in practice, however, one needs

(i) an efficient routine for solving the direction-finding subproblem (7b);

(ii) a step-size rule that offers rigorous guarantees on the algorithm's convergence rate.

In the following, we propose an efficient bisection algorithm to address (i). As for (ii), we show that the convergence analysis portrayed in [13] applies to the problem at hand. The procedure for solving (7b) is outlined in Algorithm 1, which involves an auxiliary function $h : \mathbb{R}_+ \to \mathbb{R}$ defined via

$$h(\gamma) \triangleq \rho^2 - \big\langle \Sigma, \big(I_d - \gamma(\gamma I_d - \nabla f(S))^{-1} \big)^2 \big\rangle. \tag{8}$$

Theorem 3.2 (Direction-finding subproblem)**.** For any fixed inputs $\rho, \varepsilon \in \mathbb{R}_{++}$, $\Sigma \in \mathbb{S}_{++}^d$ and $S \in \mathbb{S}_+^d$, Algorithm 1 outputs a feasible and ε-suboptimal solution to (7b).

We emphasize that the most expensive operation in Algorithm 1 is the matrix inversion $(\gamma I_d - D)^{-1}$, which needs to be evaluated repeatedly for different values of γ. These computations can be accelerated by diagonalizing D only once at the beginning. The repeat loop in Algorithm 1 carries out the actual bisection algorithm, and a suitable initial bisection interval is determined by a pair of a priori bounds LB and UB, which are available in closed form (see Appendix A).

The overall structure of the proposed Frank-Wolfe method is summarized in Algorithm 2. We borrow the step-size rule suggested in [13] to establish rigorous convergence guarantees. This is accomplished by showing that the objective function f has a bounded *curvature constant*. Our convergence result is formalized in the next theorem.

Theorem 3.3 (Convergence analysis)**.** If $\Sigma \succ 0$, $\rho > 0$, $\delta > 0$ and $\alpha_k = 2/(2+k)$ for any $k \in \mathbb{N}$, then the k-th iterate $S^{(k)}$ computed by Algorithm 2 is feasible in (5) and satisfies

$$f(S^\star) - f(S^{(k)}) \leq \frac{4\bar{\sigma}^4}{\underline{\sigma}^3(k+2)}(1+\delta),$$

where $S^\star$ is an optimal solution of (5), $\underline{\sigma}$ is the smallest eigenvalue of Σ, and $\bar{\sigma} \triangleq (\rho + \sqrt{\mathrm{Tr}\,[\Sigma]})^2$.

[1]Our terminology is inspired by the definition of the elasticity of a univariate function $\varphi(s)$ as $\frac{\mathrm{d}\varphi(s)}{\mathrm{d}s}\frac{s}{\varphi(s)}$.

Algorithm 1 Bisection algorithm to solve (7b)

Input: Covariance matrix $\Sigma \succ 0$
Gradient matrix $D \triangleq \nabla f(S) \succeq 0$
Wasserstein radius $\rho > 0$
Tolerance $\varepsilon > 0$

Denote the largest eigenvalue of D by λ_1
Let v_1 be an eigenvector of λ_1
Set $LB \leftarrow \lambda_1(1 + \sqrt{v_1^\top \Sigma v_1}/\rho)$
Set $UB \leftarrow \lambda_1(1 + \sqrt{\mathrm{Tr}\,[\Sigma]}/\rho)$
repeat
  Set $\gamma \leftarrow (UB + LB)/2$
  Set $L \leftarrow \gamma^2(\gamma I_d - D)^{-1}\Sigma(\gamma I_d - D)^{-1}$
  if $h(\gamma) < 0$ **then**
    Set $LB \leftarrow \gamma$
  else
    Set $UB \leftarrow \gamma$
  end if
  Set $\Delta \leftarrow \gamma(\rho^2 - \mathrm{Tr}\,[\Sigma]) - \langle L, D\rangle + \gamma^2\langle(\gamma I_d - D)^{-1}, \Sigma\rangle$
until $h(\gamma) > 0$ and $\Delta < \varepsilon$

Output: L

Algorithm 2 Frank-Wolfe algorithm to solve (5)

Input: Covariance matrix $\Sigma \succ 0$
Wasserstein radius $\rho > 0$
Tolerance $\delta > 0$

Set $\underline{\sigma} \leftarrow \lambda_{\min}(\Sigma), \bar{\sigma} \leftarrow (\rho + \sqrt{\mathrm{Tr}\,[\Sigma]})^2$
Set $\overline{C} \leftarrow 2\bar{\sigma}^4/\underline{\sigma}^3$
Set $S^{(0)} \leftarrow \Sigma, k \leftarrow 0$
while Stopping criterion is not met **do**
  Set $\alpha_k \leftarrow \frac{2}{k+2}$
  Set $G \leftarrow S_{xy}^{(k)}(S_{yy}^{(k)})^{-1}$
  Compute gradient $D \leftarrow \nabla f(S^{(k)})$ by $D \leftarrow [I_n, \ -G]^\top [I_n, \ -G]$
  Set $\varepsilon \leftarrow \alpha_k \delta \overline{C}$
  Solve the subproblem (7b) by Algorithm 1 $L \leftarrow \text{Bisection}(\Sigma, D, \rho, \varepsilon)$
  Set $S^{(k+1)} \leftarrow S^{(k)} + \alpha_k(L - S^{(k)})$
  Set $k \leftarrow k + 1$
end while

Output: $S^{(k)}$

4 The Wasserstein Distributionally Robust Kalman Filter

Consider a discrete-time dynamical system whose (unobservable) state $x_t \in \mathbb{R}^n$ and (observable) output $y_t \in \mathbb{R}^m$ evolve randomly over time. At any time $t \in \mathbb{N}$, we aim to estimate the current state x_t based on the output history $Y_t \triangleq (y_1, \ldots, y_t)$. We assume that the joint state-output process $z_t = [x_t^\top, \ y_t^\top]^\top, t \in \mathbb{N}$, is governed by an unknown Gaussian distribution $\mathbb{Q}$ in the neighborhood of a known nominal distribution $\mathbb{P}^\star$. The distribution $\mathbb{P}^\star$ is determined through the linear state-space model

$$\left.\begin{array}{l} x_t = A_t x_{t-1} + B_t v_t \\ y_t = C_t x_t + D_t v_t \end{array}\right\} \quad \forall t \in \mathbb{N}, \tag{9}$$

where A_t, B_t, C_t, and D_t are given matrices of appropriate dimensions, while $v_t \in \mathbb{R}^d$, $t \in \mathbb{N}$, denotes a Gaussian white noise process independent of $x_0 \sim \mathcal{N}_n(\hat{x}_0, V_0)$. Thus, $v_t \sim \mathcal{N}_d(0, I_d)$ for all t, while v_t and $v_{t'}$ are independent for all $t \neq t'$. Note that we may restrict the dimension of v_t to the dimension $d = n + m$ of z_t without loss of generality. Otherwise, all linearly dependent columns of $[B_t^\top, \ D_t^\top]^\top$ and the corresponding components of v_t can be eliminated systematically.

By the law of total probability and the Markovian nature of the state-space model (9), the nominal distribution $\mathbb{P}^\star$ is uniquely determined by the marginal distribution $\mathbb{P}_{x_0}^\star = \mathcal{N}_n(\hat{x}_0, V_0)$ of the initial state x_0 and the conditional distributions

$$\mathbb{P}_{z_t|x_{t-1}}^\star = \mathcal{N}_d\left(\begin{bmatrix} A_t \\ C_t A_t \end{bmatrix} x_{t-1}, \begin{bmatrix} B_t \\ C_t B_t + D_t \end{bmatrix} \begin{bmatrix} B_t \\ C_t B_t + D_t \end{bmatrix}^\top\right)$$

of z_t given x_{t-1} for all $t \in \mathbb{N}$.

Unlike $\mathbb{P}^\star$, the true distribution $\mathbb{Q}$ governing z_t, $t \in \mathbb{N}$, is unknown, and thus the estimation problem at hand is not well-defined. We will therefore estimate the conditional mean $\hat{x}_t$ and covariance matrix V_t of x_t given Y_t under some worst-case distribution $\mathbb{Q}^\star$ to be constructed recursively. First, we assume that the marginal distribution $\mathbb{Q}_{x_0}^\star$ of x_0 under $\mathbb{Q}^\star$ equals $\mathbb{P}_{x_0}$, that is, $\mathbb{Q}_{x_0}^\star = \mathcal{N}_n(\hat{x}_0, V_0)$. Next, fix any $t \in \mathbb{N}$ and assume that the conditional distribution $\mathbb{Q}_{x_{t-1}|Y_{t-1}}^\star$ of x_{t-1} given Y_{t-1} under $\mathbb{Q}^\star$ has already been computed as $\mathbb{Q}_{x_{t-1}|Y_{t-1}}^\star = \mathcal{N}_n(\hat{x}_{t-1}, V_{t-1})$. The construction of $\mathbb{Q}_{x_t|Y_t}^\star$ is then split into a prediction step and an update step. The prediction step combines the previous state estimate $\mathbb{Q}_{x_{t-1}|Y_{t-1}}^\star$ with the nominal transition kernel $\mathbb{P}_{z_t|x_{t-1}}^\star$ to generate a pseudo-nominal

Algorithm 3 Robust Kalman filter at time t

Input: Covariance matrix $V_{t-1} \succeq 0$
State estimate $\hat{x}_{t-1}$
Wasserstein radius $\rho_t > 0$
Tolerance $\delta > 0$

Prediction:
Form the pseudo-nominal distribution $\mathbb{P}_{z_t|Y_{t-1}} = \mathcal{N}_d(\mu_t, \Sigma_t)$ using (10)

Observation:
Observe the output y_t

Update:
Use Algorithm 2 to solve (11)
$S_t^\star \leftarrow$ Frank-Wolfe$(\Sigma_t, \mu_t, \rho_t, \delta)$

Output: $V_t = S_{t,xx} - S_{t,xy}(S_{t,yy})^{-1}S_{t,yx}$
$\hat{x}_t = S^\star_{t,xy}(S^\star_{t,yy})^{-1}(y_t - \mu_{t,y}) + \mu_{t,x}$

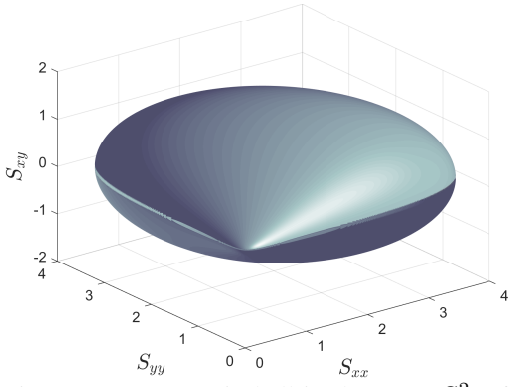


Figure 2: Wasserstein ball in the space $\mathbb{S}_+^2$ of covariance matrices centered at I_2 with radius 1.

distribution $\mathbb{P}_{z_t|Y_{t-1}}$ of z_t conditioned on Y_{t-1}, which is defined through

$$\mathbb{P}_{z_t|Y_{t-1}}(B|Y_{t-1}) = \int_{\mathbb{R}^n} \mathbb{P}^\star_{z_t|x_{t-1}}(B|x_{t-1})\mathbb{Q}^\star_{x_{t-1}|Y_{t-1}}(\mathrm{d}x_{t-1}|Y_{t-1})$$

for every Borel set $B \subseteq \mathbb{R}^d$ and observation history $Y_{t-1} \in \mathbb{R}^{m\times(t-1)}$. The well-known formula for the convolution of two multivariate Gaussians reveals that $\mathbb{P}_{z_t|Y_{t-1}} = \mathcal{N}_d(\mu_t, \Sigma_t)$, where

$$\mu_t = \begin{bmatrix} A_t \\ C_tA_t \end{bmatrix} \hat{x}_{t-1} \quad \text{and} \quad \Sigma_t = \begin{bmatrix} A_t \\ C_tA_t \end{bmatrix} V_{t-1} \begin{bmatrix} A_t \\ C_tA_t \end{bmatrix}^\top + \begin{bmatrix} B_t \\ C_tB_t + D_t \end{bmatrix} \begin{bmatrix} B_t \\ C_tB_t + D_t \end{bmatrix}^\top . \quad (10)$$

Note that the construction of $\mathbb{P}_{z_t|Y_{t-1}}$ resembles the prediction step of the classical Kalman filter but uses the least favorable distribution $\mathbb{Q}^\star_{x_{t-1}|Y_{t-1}}$ instead of the nominal distribution $\mathbb{P}^\star_{x_{t-1}|Y_{t-1}}$.

In the update step, the pseudo-nominal a priori estimate $\mathbb{P}_{z_t|Y_{t-1}}$ is updated by the measurement y_t and robustified against model uncertainty to yield a refined a posteriori estimate $\mathbb{Q}^\star_{x_t|Y_t}$. This a posteriori estimate is found by solving the minimax problem

$$\inf_{\psi_t \in \mathcal{L}} \sup_{\mathbb{Q} \in \mathcal{P}_{z_t|Y_{t-1}}} \mathbb{E}^{\mathbb{Q}} \left[\|x_t - \psi_t(y_t)\|^2 \right] \quad (11)$$

equipped with the Wasserstein ambiguity set

$$\mathcal{P}_{z_t|Y_{t-1}} = \left\{ \mathbb{Q} \in \mathcal{N}_d : W_2(\mathbb{Q}, \mathbb{P}_{z_t|Y_{t-1}}) \leq \rho_t \right\}.$$

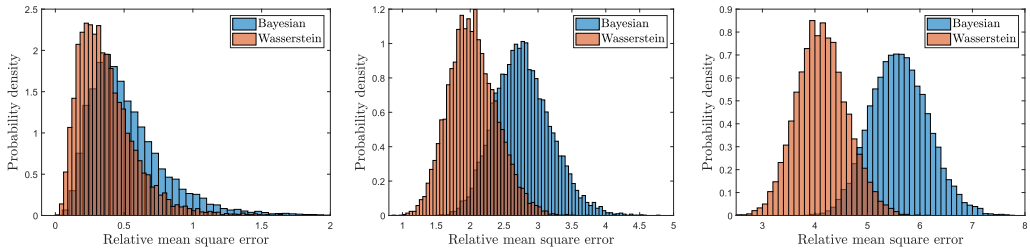
Note that the Wasserstein radius ρ_t quantifies our distrust in the pseudo-nominal a priori estimate and can therefore be interpreted as a measure of model uncertainty. Practically, we reformulate (11) as an equivalent finite convex program of the form (5), which is amenable to efficient computational solution via the Frank-Wolfe algorithm detailed in Section 3. By Theorem 2.5, the optimal solution $S_t^\star$ of problem (5) yields the least favorable conditional distribution $\mathbb{Q}^\star_{z_t|Y_{t-1}} = \mathcal{N}_d(\mu_t, S_t^\star)$ of z_t given Y_{t-1}. By using the well-known formulas for conditional normal distributions (see, *e.g.*, [20, page 522]), we then obtain the least favorable conditional distribution $\mathbb{Q}^\star_{x_t|Y_t} = \mathcal{N}_d(\hat{x}_t, V_t)$ of x_t given Y_t, where

$$\hat{x}_t = S^\star_{t,xy}(S^\star_{t,yy})^{-1}(y_t - \mu_{t,y}) + \mu_{t,x} \quad \text{and} \quad V_t = S^\star_{t,\,xx} - S^\star_{t,\,xy}(S^\star_{t,\,yy})^{-1}S^\star_{t,\,yx}.$$

The distributionally robust Kalman filtering approach is summarized in Algorithm 3. Note that the robust update step outlined above reduces to the usual update step of the classical Kalman filter for $\rho \downarrow 0$.

5 Numerical Results

We showcase the performance of the proposed Frank-Wolfe algorithm and the distributionally robust Kalman filter in a suite of synthetic experiments. All optimization problems are implemented in MATLAB and run on an Intel XEON CPU with 3.40GHz clock speed and 16GB of RAM, and the corresponding codes are made publicly available at `https://github.com/sorooshafiee/WKF`.



(a) $d = 10$ (b) $d = 50$ (c) $d = 100$

Figure 3: Distribution of the difference between the errors of the robust MMSE (Bayesian MMSE) and the ideal MMSE* estimator.

5.1 Distributionally Robust Minimum Mean Square Error Estimation

We first assess the distributionally robust minimum mean square error (robust MMSE) estimator, which is obtained by solving (2), against the classical Bayesian MMSE estimator, which can be viewed as the solution of problem (2) over a singleton ambiguity set that contains only the nominal distribution. Recall from Remark 2.6 that the optimal estimator corresponding to a KL or τ-divergence ambiguity set of the type studied in [15, 30] coincides with the Bayesian MMSE estimator irrespective of ρ. Thus, we may restrict attention to Wasserstein ambiguity sets. In order to develop a geometric intuition, Figure 2 visualizes the set of all bivariate normal distributions with zero mean that have a Wasserstein distance of at most 1 from the standard normal distribution—projected to the space of covariance matrices.

In the first experiment we aim to predict a signal $x \in \mathbb{R}^{4d/5}$ from an observation $y \in \mathbb{R}^{d/5}$, where the random vector $z = [x^\top,\ y^\top]^\top$ follows a d-variate Gaussian distribution with $d \in \{10, 50, 100\}$. The experiment comprises 10^4 simulation runs. In each run we randomly generate two covariance matrices $\Sigma^\star$ and Σ as follows. First, we draw two matrices $A^\star$ and A from the standard normal distribution on $\mathbb{R}^{d\times d}$, and we denote by $R^\star$ and R the orthogonal matrices whose columns correspond to the orthonormal eigenvectors of $A^\star + (A^\star)^\top$ and $A + A^\top$, respectively. Then, we define $\Delta^\star = R^\star\Lambda^\star(R^\star)^\top$ and $\Sigma = R\Lambda R^\top$, where $\Lambda^\star$ and Λ are diagonal matrices whose main diagonals are sampled uniformly from $[0, 1]^d$ and $[0.1, 10]^d$, respectively. Finally, we set $\Sigma^\star = (\Sigma^{\frac{1}{2}} + (\Delta^\star)^{\frac{1}{2}})^2$ and define the normal distributions $\mathbb{P}^\star = \mathcal{N}_d(0, \Sigma^\star)$ and $\mathbb{P} = \mathcal{N}_d(0, \Sigma)$. By construction, we have

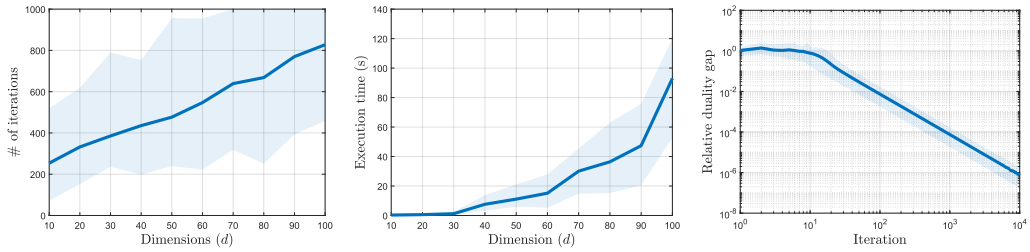
$$W_2(\mathbb{P}^\star, \mathbb{P}) \leq \|(\Sigma^\star)^{\frac{1}{2}} - \Sigma^{\frac{1}{2}}\|_F \leq \sqrt{d},$$

where $\|\cdot\|_F$ stands for the Frobenius norm, and the first inequality follows from [16, Proposition 3]. We assume that $\mathbb{P}^\star$ is the true distribution and $\mathbb{P}$ our nominal prior. The robust MMSE estimator is obtained by solving (5) for $\rho = \sqrt{d}$ via the Frank-Wolfe algorithm from Section 3, while the Bayesian MMSE estimator under $\mathbb{P}$ is calculated analytically. In order to provide a meaningful comparison between these two approaches, we also compute the Bayesian MMSE estimator under the true distribution $\mathbb{P}^\star$ (denoted by MMSE*), which is indeed the best possible estimator. Figure 3 visualizes the distribution of the difference between the mean square errors under $\mathbb{P}^\star$ of the robust MMSE (Bayesian MMSE) and MMSE* estimators. We observe that the robust MMSE estimator produces better results consistently across all experiments, and the effect is more pronounced for larger dimensions d. Figures 4(a) and 4(b) report the execution time and the iteration complexity of the Frank-Wolfe algorithm for $d \in \{10, \ldots, 100\}$ when the algorithm is stopped as soon as the relative duality gap $\big\langle F(S^k) - S^k, \nabla f(S^k)\big\rangle / f(S^k)$ drops below 0.01%. Note that the execution time grows polynomially due to the matrix inversion in the bisection algorithm. Figure 4(c) shows the relative duality gap of the current solution as a function of the iteration count.

5.2 Wasserstein Distributionally Robust Kalman Filtering

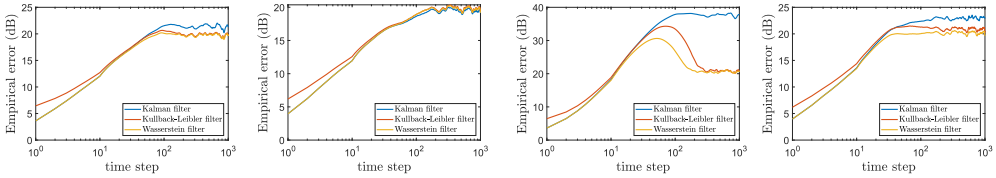
We assess the performance of the proposed Wasserstein distributionally robust Kalman filter against that of the classical Kalman filter and the Kalman filter with the KL ambiguity set from [15]. To this end, we borrow the standard test instance from [22, 28, 15] with $n = 2$ and $m = 1$. The system matrices satisfy

$$A_t = \begin{bmatrix} 0.9802 & 0.0196 + 0.099\Delta_t \\ 0 & 0.9802 \end{bmatrix},\ B_tB_t^\top = \begin{bmatrix} 1.9608 & 0.0195 \\ 0.0195 & 1.9605 \end{bmatrix},\ C_t = [1,\ -1],\ D_tD_t^\top = 1,$$



(a) Scaling of iteration count (b) Scaling of execution time (c) Convergence for $d = 100$

Figure 4: Convergence behavior of the Frank-Wolfe algorithm (shown are the average (solid line) and the range (shaded area) of the respective performance measures across 100 simulation runs)



(a) Small time-invariant uncertainty (b) Small time varying uncertainty (c) Large time-invariant uncertainty (d) Large time-varying uncertainty

Figure 5: Empirical means square estimation error of different filters

and $B_t D_t^\top = 0$, where Δ_t represents a scalar uncertainty, and the initial state satisfies $x_0 \sim \mathcal{N}_2(0, I_2)$. In all numerical experiments we simulate the different filters over 1000 periods starting from $\hat{x}_0 = 0$ and $V_0 = I_2$. Figure 5 shows the empirical mean square error $\frac{1}{500}\sum_{j=1}^{500} \|x_t^j - \hat{x}_t^j\|^2$ across 500 independent simulation runs, where $\hat{x}_t^j$ denotes the state estimate at time t in the j^{th} run. We distinguish four different scenarios: time-invariant uncertainty ($\Delta_t^j = \Delta^j$ sampled uniformly from $[-\bar{\Delta}, \bar{\Delta}]$ for each j) versus time-varying uncertainty (Δ_t^j sampled uniformly from $[-\bar{\Delta}, \bar{\Delta}]$ for each t and j), and small uncertainty ($\bar{\Delta} = 1$) versus large uncertainty ($\bar{\Delta} = 10$). All results are reported in decibel units ($10\log_{10}(\cdot)$). As for the filter design, the Wasserstein and KL radii are selected from the search grids $\{a \cdot 10^{-1} : a \in \{1, 1.1, \cdots, 2\}\}$ and $\{a \cdot 10^{-4} : a \in \{1, 1.1, \cdots, 2\}\}$, respectively. Figure 5 reports the results with minimum steady state error across all candidate radii.

Under small time-invariant uncertainty (Figure 5(a)), the Wasserstein and KL distributionally robust filters display a similar steady-state performance but outperform the classical Kalman filter. Note that the KL distributionally robust filter starts from a different initial point as we use the delayed implementation from [15]. Under small time-varying uncertainty (Figure 5(b)), both distributionally robust filters display a similar performance as the classical Kalman filter. Figures 5(c) and (d) corresponding to the case of large uncertainty are similar to Figures 5(a) and (b), respectively. However, the Wasserstein distributionally robust filter now significantly outperforms the classical Kalman filter and, to a lesser extent, the KL distributionally robust filter. Moreover, the Wasserstein distributionally robust filter exhibits the best transient behavior.

Appendix A Proofs

A.1 Proof of Theorem 2.3

The proof of Theorem 2.3 requires the following preparatory lemma, which we borrow from [18].

Lemma A.1 ([18, Proposition 2.8])**.** For any $\gamma \in \mathbb{R}_+$, $D \in \mathbb{S}_+^d \setminus \{0\}$ and $\Sigma \in \mathbb{S}_{++}^d$, we have

$$\sup_{S \succeq 0} \ \langle D, S\rangle - \gamma \operatorname{Tr}\left[S - 2\left(\Sigma^{\frac{1}{2}} S \Sigma^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] = \begin{cases} \gamma^2 \langle (\gamma I_d - D)^{-1}, \Sigma\rangle & \text{if } \gamma I_d \succ D, \\ +\infty & \text{otherwise.} \end{cases}$$

Moreover, if $\gamma I_d \succ D$, the unique optimal solution of the above maximization problem is given by

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

Proof of Theorem 2.3. The optimal value of the minimax problem (2) satisfies

$$\inf_{\psi\in\mathcal{L}} \sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}^{\mathbb{Q}}\left[\|x-\psi(y)\|^2\right] \geq \sup_{\mathbb{Q}\in\mathcal{P}} \inf_{\psi\in\mathcal{L}} \mathbb{E}^{\mathbb{Q}}\left[\|x-\psi(y)\|^2\right] \tag{A.1a}$$

$$= \sup_{\mathbb{Q}\in\mathcal{P}} \inf_{G,g} \mathbb{E}^{\mathbb{Q}}\left[\|x-Gy-g\|^2\right], \tag{A.1b}$$

where (A.1a) follows from the max-min inequality, and (A.1b) holds because the inner minimization problem over ψ is solved by the conditional expectation function $\psi^\star(y)=\mathbb{E}^{\mathbb{Q}}[x|y]$, which is affine in y for every fixed Gaussian distribution $\mathbb{Q}\in\mathcal{P}$, see, *e.g.*, [20, page 522]. Without loss of generality, one can thus restrict the set of measurable functions $\mathcal{L}$ to the set of affine functions parametrized by a sensitivity matrix $G\in\mathbb{R}^{n\times m}$ and an intercept vector $g\in\mathbb{R}^n$. Recalling the definition of the Wasserstein ambiguity set $\mathcal{P}$ in (3) and encoding each normal distribution $\mathbb{Q}\in\mathcal{P}$ by its mean vector $c\in\mathbb{R}^d$ and covariance matrix $S\in\mathbb{S}_+^d$, we can use Proposition 2.2 to reformulate (A.1b) as

$$\begin{array}{ll} \sup & \inf\limits_{G,g} \left\langle I_n, S_{xx}+c_xc_x^\top\right\rangle + \left\langle G^\top G, S_{yy}+c_yc_y^\top\right\rangle - \left\langle G, S_{xy}+c_xc_y^\top\right\rangle \\ & \qquad -\left\langle G^\top, S_{yx}+c_yc_x^\top\right\rangle + 2\left\langle g, Gc_y-c_x\right\rangle + g^\top g \\ \text{s.t.} & c\in\mathbb{R}^d, \quad c_x\in\mathbb{R}^n, \quad c_y\in\mathbb{R}^m \\ & S\in\mathbb{S}_+^d, \quad S_{xx}\in\mathbb{S}_+^n, \quad S_{yy}\in\mathbb{S}_+^m, \quad S_{xy}=S_{yx}^\top\in\mathbb{R}^{n\times m} \\ & c=\begin{bmatrix} c_x \\ c_y\end{bmatrix},\ S=\begin{bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy}\end{bmatrix}\succeq 0 \\ & \|c-\mu\|^2 + \operatorname{Tr}\left[S+\Sigma-2\left(\Sigma^{\frac12}S\Sigma^{\frac12}\right)^{\frac12}\right]\leq\rho^2. \end{array} \tag{A.2a}$$

Solving the inner minimization problem over g analytically and substituting the optimal solution $g^\star=c_x-Gc_y$ back into the objective function shows that (A.2a) is equivalent to

$$\begin{array}{ll} \sup & \inf\limits_{G} \left\langle \begin{bmatrix} I_n & -G \\ -G^\top & G^\top G\end{bmatrix}, S\right\rangle \\ \text{s.t.} & c\in\mathbb{R}^d, \quad S\in\mathbb{S}_+^d \\ & \|c-\mu\|^2 + \operatorname{Tr}\left[S+\Sigma-2\left(\Sigma^{\frac12}S\Sigma^{\frac12}\right)^{\frac12}\right]\leq\rho^2. \end{array} \tag{A.2b}$$

The minimization over G may now be interchanged with the maximization over c and S by using the classical minimax theorem [5, Proposition 5.5.4], which applies because c and S range over a compact feasible set. The inner maximization problem over c is then solved by $c^\star=\mu$, which maximizes the slack of the Wasserstein constraint. Thus, the minimax problem (A.2b) simplifies to

$$\begin{array}{lll} \inf\limits_{G} & \sup\limits_{S\succeq 0} & \left\langle \begin{bmatrix} I_n & -G \\ -G^\top & G^\top G\end{bmatrix}, S\right\rangle \\ & \text{s.t.} & \operatorname{Tr}\left[S+\Sigma-2\left(\Sigma^{\frac12}S\Sigma^{\frac12}\right)^{\frac12}\right]\leq\rho^2. \end{array} \tag{A.2c}$$

Assigning a Lagrange multiplier $\gamma\geq 0$ to the Wasserstein constraint and dualizing the inner maximization problem yields

$$\inf_{G}\ \inf_{\gamma\geq 0}\ \sup_{S\succeq 0}\ \left\langle \begin{bmatrix} I_n & -G \\ -G^\top & G^\top G\end{bmatrix}, S\right\rangle + \gamma\left(\rho^2-\operatorname{Tr}\left[S+\Sigma-2\left(\Sigma^{\frac12}S\Sigma^{\frac12}\right)^{\frac12}\right]\right). \tag{A.2d}$$

Strong duality holds because $S=\Sigma\succ 0$ represents a Slater point for the primal maximization problem. Finally, by using Lemma A.1, problem (A.2d) can be reformulated as

$$\begin{array}{ll} \inf & \gamma\left(\rho^2-\operatorname{Tr}\left[\Sigma\right]\right)+\gamma^2\left\langle(\gamma I_d-[I_n,\ -G]^\top[I_n,\ -G])^{-1},\Sigma\right\rangle \\ \text{s.t.} & G\in\mathbb{R}^{n\times m}, \quad \gamma\in\mathbb{R}_+ \\ & \gamma I_d\succ[I_n,\ -G]^\top[I_n,\ -G]. \end{array} \tag{A.3}$$

By construction, the optimal value of (A.3) provides a lower bound on that of the minimax problem (2). Next, we construct an upper bound by restricting $\mathcal{L}$ to the class of affine estimators.

$$\inf_{\psi\in\mathcal{L}} \sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}^{\mathbb{Q}}\left[\|x-\psi(y)\|^2\right] \leq \inf_{G,g} \sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}^{\mathbb{Q}}\left[\|x-Gy-g\|^2\right] \tag{A.4}$$

As $\mathcal{P}$ is non-convex, we cannot simply use Sion's minimax theorem to show that the right-hand side of (A.4) equals (A.1b). Instead, we need a more involved argument. Recalling the definition of $\mathcal{P}$ in (3) and encoding each normal distribution $\mathbb{Q} \in \mathcal{P}$ by its mean vector $c \in \mathbb{R}^d$ and covariance matrix $S \in \mathbb{S}_+^d$, we can use Proposition 2.2 to reformulate the right-hand side of (A.4) as

$$\begin{array}{lll} \inf\limits_{G,g} & \sup & \big\langle I_n, S_{xx} + c_x c_x^\top \big\rangle + \big\langle G^\top G, S_{yy} + c_y c_y^\top \big\rangle - \big\langle G, S_{xy} + c_x c_y^\top \big\rangle \\ & & \qquad -\big\langle G^\top, S_{yx} + c_y c_x^\top \big\rangle + 2\big\langle g, Gc_y - c_x \big\rangle + g^\top g \\ & \text{s.t.} & c \in \mathbb{R}^d, \quad c_x \in \mathbb{R}^n, \quad c_y \in \mathbb{R}^m \\ & & S \in \mathbb{S}_+^d, \quad S_{xx} \in \mathbb{S}_+^n, \quad S_{yy} \in \mathbb{S}_+^m, \quad S_{xy} = S_{yx}^\top \in \mathbb{R}^{n \times m} \\ & & c = \begin{bmatrix} c_x \\ c_y \end{bmatrix}, \, S = \begin{bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{bmatrix} \succeq 0 \\ & & \|c - \mu\|^2 + \operatorname{Tr}\left[S + \Sigma - 2\left(\Sigma^{\frac{1}{2}} S \Sigma^{\frac{1}{2}} \right)^{\frac{1}{2}} \right] \leq \rho^2. \end{array} \tag{A.5a}$$

Next, we introduce the set $\mathcal{C} \triangleq \{c \in \mathbb{R}^d : \|c - \mu\| \leq \rho\}$ as well as the auxiliary functions

$$D(G) \triangleq \begin{bmatrix} I_n & -G \\ -G^\top & G^\top G \end{bmatrix} \quad \text{and} \quad b(G, g) \triangleq \begin{bmatrix} -g \\ G^\top g \end{bmatrix}$$

to reformulate problem (A.5a) as

$$\begin{array}{lll} \inf\limits_{G,g} & \sup\limits_{\substack{c \in \mathcal{C} \\ S \succeq 0}} & \big\langle D(G), S + c\, c^\top \big\rangle + 2\big\langle b(G,g), c \big\rangle + g^\top g \\ & \text{s.t.} & \|c - \mu\|^2 + \operatorname{Tr}\left[S + \Sigma - 2\left(\Sigma^{\frac{1}{2}} S \Sigma^{\frac{1}{2}} \right)^{\frac{1}{2}} \right] \leq \rho^2. \end{array} \tag{A.5b}$$

We emphasize that the constraint $c \in \mathcal{C}$ is redundant in (A.5b) but will facilitate further simplifications below. Note also that $D(G) \succeq 0$, and thus the minimax problem (A.5b) involves a cumbersome convex maximization problem over c. By employing a penalty formulation of the Wasserstein constraint, the inner maximization problem over c and S in (A.5b) can be re-expressed as

$$\begin{array}{ll} \sup\limits_{\substack{c \in \mathcal{C} \\ S \succeq 0}} & \inf\limits_{\gamma \geq 0} \big\langle D(G), S + c\, c^\top \big\rangle + 2\big\langle b(G,g), c \big\rangle + g^\top g \\ & \qquad + \gamma \left(\rho^2 - \|c - \mu\|^2 + \operatorname{Tr}\left[S + \Sigma - 2\left(\Sigma^{\frac{1}{2}} S \Sigma^{\frac{1}{2}} \right)^{\frac{1}{2}} \right] \right). \end{array}$$

Here, the minimization over γ and the maximization over S may be interchanged by strong duality, which holds because $S = \Sigma \succ 0$ constitutes a Slater point for the primal problem, see, *e.g.*, [5, Proposition 5.3.1]. We note that when $\|c - \mu\| = \rho$, the feasible set of S reduces to a singleton, and thus strong duality holds trivially. The emerging inner maximization problem over S can then be solved analytically by using Lemma A.1. In summary, the minimax problem (A.5b) is equivalent to

$$\begin{array}{llll} \inf\limits_{G,g} & \sup\limits_{c \in \mathcal{C}} & \inf\limits_{\gamma \geq 0} & \big\langle D(G), c\, c^\top \big\rangle + 2\big\langle b(G,g), c \big\rangle + g^\top g + \gamma \left(\rho^2 - \|c - \mu\|^2 - \operatorname{Tr}\left[\Sigma \right] \right) \\ & & & \qquad + \gamma^2 \big\langle (\gamma I_d - D(G))^{-1}, \Sigma \big\rangle \\ & & \text{s.t.} & \gamma I_d \succ D(G). \end{array} \tag{A.5c}$$

Observe now that the optimal value function of the innermost minimization problem over γ in (A.5c) is convex in g and, thanks to the constraint $\gamma I_d - D(G) \succ 0$, concave in c for every fixed G. By the classical minimax theorem [5, Proposition 5.5.4], which applies because c ranges over the compact set $\mathcal{C}$, we may thus interchange the infimum over g with the supremum over c. After replacing $D(G)$ and $b(G, g)$ with their definitions, it becomes clear that the innermost minimization problem over g admits the analytical solution $g^\star = \mu_x - G\mu_y$. Thus, problem (A.5c) is equivalent to

$$\begin{array}{llll} \inf\limits_{G} & \sup\limits_{c \in \mathcal{C}} & \inf\limits_{\gamma \geq 0} & \gamma \left(\rho^2 - \|c - \mu\|^2 - \operatorname{Tr}\left[\Sigma \right] \right) + \gamma^2 \big\langle (\gamma I_d - [I_n, \ -G]^\top [I_n, \ -G])^{-1}, \Sigma \big\rangle \\ & & \text{s.t.} & \gamma I_d \succ [I_n, \ -G]^\top [I_n, \ -G]. \end{array} \tag{A.5d}$$

By invoking the minimax theorem [5, Proposition 5.5.4] once again, the inner infimum over γ can be interchanged with the supremum over c. As the resulting inner maximization problem over c is solved by $c^\star = \mu$, problem (A.5d) is thus equivalent to (A.3). In summary, we have shown that (A.3) provides both an upper bound on the left-hand side of (A.1) as well as a lower bound on the right-hand side of (A.1). Thus, the inequality in (A.1) is in fact an equality. □

A.2 Proof of Theorem 2.5

The proof of Theorem 2.5 relies on the following lemma, which extends a similar result from [18].

Lemma A.2 (Analytical solution of direction-finding subproblem)**.** For any fixed $\Sigma \in \mathbb{S}_{++}^d$ and $D \in \mathbb{S}_+^d \setminus \{0\}$, the optimization problem

$$\begin{array}{ll} \sup\limits_{S \in \mathbb{S}_+^d} & \big\langle S, D \big\rangle \\ \text{s.t.} & \operatorname{Tr}\left[S + \Sigma - 2\left(\Sigma^{\frac{1}{2}} S \Sigma^{\frac{1}{2}} \right)^{\frac{1}{2}} \right] \leq \rho^2 \end{array}$$

is solved by

$$S^\star = (\gamma^\star)^2 (\gamma^\star I_d - D)^{-1} \Sigma (\gamma^\star I_d - D)^{-1},$$

where $\gamma^\star$ is the unique solution with $\gamma^\star I_d \succ D$ of the algebraic equation

$$\rho^2 - \left\langle \Sigma, \left(I_d - \gamma^\star (\gamma^\star I_d - D)^{-1} \right)^2 \right\rangle = 0.$$

Moreover, we have $S^\star \succeq \underline{\sigma} I_d$, where $\underline{\sigma} \triangleq \lambda_{\min}(\Sigma)$.

Proof of Lemma A.2. The optimality of $S^\star$ follows immediately from [18, Theorem 5.1]. Moreover, the spectral norm of $(S^\star)^{-1}$ obeys the following estimate.

$$\|(S^\star)^{-1}\| \leq \|I_d - \frac{1}{\gamma^\star} D\| \cdot \|\Sigma^{-1}\| \cdot \|I_d - \frac{1}{\gamma^\star} D\| \leq \|\Sigma^{-1}\| = \underline{\sigma}^{-1}$$

As the largest eigenvalue of $(S^\star)^{-1}$ is bounded by $\underline{\sigma}^{-1}$, we may conclude that $S^\star \succeq \underline{\sigma} I_d$. □

Proof of Theorem 2.5. The proof of Theorem 2.3 has shown that the original infinite-dimensional minimax problem (2) is equivalent to the finite-dimensional minimax problem (A.2c). By Lemma A.2, the solution of the inner maximization problem in (A.2c) satisfies $S^\star \succeq \underline{\sigma} I_d$. Thus, one may append the redundant constraint $S \succeq \underline{\sigma} I_d$ to this inner problem without sacrificing optimality. By interchanging the minimization over G with the maximization over S, which is allowed by [5, Proposition 5.5.4], problem (A.2c) can thus be reformulated as

$$\begin{array}{lll} \sup\limits_{S \succeq 0} & \inf\limits_{G} & \left\langle \begin{bmatrix} I_n & -G \\ -G^\top & G^\top G \end{bmatrix}, S \right\rangle \\ \text{s.t.} & & \operatorname{Tr}\left[S + \Sigma - 2\left(\Sigma^{1/2} S \Sigma^{1/2} \right)^{1/2} \right] \leq \rho^2 \\ & & S \succeq \underline{\sigma} I_d. \end{array} \tag{A.6}$$

Recall that $\underline{\sigma} > 0$, which implies that $S \succ 0$. Hence, the unconstrained quadratic minimization problem over G in (A.6) has a unique solution $G^\star$, which can be obtained analytically by solving the problem's first-order optimality condition. Specifically, we have

$$2G^\star S_{yy} - 2S_{xy} = 0 \quad \Longleftrightarrow \quad G^\star = S_{xy} S_{yy}^{-1}.$$

Substituting $G^\star$ into (A.6) yields the desired maximization problem (5). By construction, this convex program is equivalent to nature's decision problem on the right-hand side of (4), and thus it is easy to see that the least favorable prior is given by $\mathbb{Q}^\star = \mathcal{N}_d(\mu, S^\star)$. Next, we solve the Bayesian estimation problem

$$\inf_{\psi \in \mathcal{L}} \mathbb{E}^{\mathbb{Q}^\star} \left[\|x - \psi(y)\|^2 \right].$$

An elementary analytical calculation reveals that this problem is solved by $\psi^\star(y) = S_{xy}^\star (S_{yy}^\star)^{-1}(y - \mu_y) + \mu_x$. Moreover, this solution is unique because $S^\star \succeq \underline{\sigma} I_d$, which implies that the objective function is strictly convex. By Theorem 2.3 and [7, Section 5.5.5], we may then conclude that $\psi^\star$ is also optimal in (2). This observation completes the proof. □

A.3 Proof of Theorem 3.2

The following lemma suggests upper and lower bounds on the (unique) root $\gamma^\star$ of the function $h(\gamma)$ defined in (8). Note that this root is computed approximately using bisection in Algorithm 1.

Lemma A.3 (Bisection interval)**.** For any $\rho > 0$, the solution of the algebraic equation $h(\gamma^\star) = 0$ resides in the interval $[\gamma_{\min}, \gamma_{\max}]$, where

$$\gamma_{\min} \triangleq \lambda_1 \left(1 + \sqrt{v_1^\top \Sigma v_1}/\rho\right), \quad \gamma_{\max} \triangleq \lambda_1 \left(1 + \sqrt{\operatorname{Tr}[\Sigma]}/\rho\right), \tag{A.7}$$

the scalar λ_1 is the largest eigenvalue of $D \triangleq \nabla f(S)$, and v_1 is a corresponding eigenvector.

Proof of Lemma A.3. Let $D = \sum_{i=1}^d \lambda_i v_i v_i^\top$ be the spectral decomposition of D. The function h can be equivalently rewritten as

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

where the summation admits the following bounds:

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

Equating the two bounds to ρ^2 yields $\gamma_{\min}$ and $\gamma_{\max}$, respectively. □

Proof of Theorem 3.2. The proof of Lemma A.2 implies that $L(\gamma) \triangleq \gamma^2(\gamma I_d - D)^{-1}\Sigma(\gamma I_d - D)^{-1}$ is feasible in (7b) for every γ with $\gamma I_d \succ D$ and $h(\gamma) > 0$. Moreover, $L(\gamma^\star)$ is optimal in (7b) if $\gamma^\star I_d \succ D$ and $h(\gamma^\star) = 0$. Algorithm 1 uses a bisection procedure to compute an approximation γ of $\gamma^\star$ such that $L(\gamma)$ is feasible and ε-suboptimal in (7b). The degree of suboptimality of $L(\gamma)$ equals $\big\langle L(\gamma^\star) - L(\gamma), D\big\rangle$. The true optimal value $\big\langle L(\gamma^\star), D\big\rangle$ is inaccessible but can be estimated above by the objective value of γ in the Lagrangian dual of (7b), which can be expressed as

$$\min_{\gamma:\ \gamma I_d \succ D} \gamma(\rho^2 - \operatorname{Tr}[\Sigma]) + \gamma^2 \big\langle (\gamma I_d - D)^{-1}, \Sigma \big\rangle,$$

see also [18, Proposition 2.8]. Thus, the suboptimality of $L(\gamma)$ is bounded above by

$$\big\langle L(\gamma^\star) - L(\gamma), D\big\rangle \leq \gamma(\rho^2 - \operatorname{Tr}[\Sigma]) + \gamma^2 \big\langle (\gamma I_d - D)^{-1}, \Sigma \big\rangle - \big\langle L(\gamma), D\big\rangle.$$

Lemma A.3 ensures that $\gamma^\star \in [\gamma_{\min}, \gamma_{\max}]$, and therefore it suffices to search over this interval. □

A.4 Proof of Theorem 3.3

The proof of Theorem 3.3 widely parallels that of [13, Theorem 1]. The key ingredient is to prove that the *curvature constant* of the problem's (negative) objective function $-f$ is bounded.

Definition A.4 (Curvature constant)**.** The curvature constant C_g of the convex function g with respect to a compact domain $\mathcal{S}$ is defined as

$$C_g \triangleq \left\{ \begin{array}{ll} \sup\limits_{X,Y,Z,\alpha} & \frac{2}{\alpha^2}\left(g(Z) - g(X) - \big\langle Z - X, \nabla g(X)\big\rangle\right) \\ \text{s.t.} & Z = (1-\alpha)X + \alpha Y \\ & X, Y \in \mathcal{S}, \quad \alpha \in [0,1]. \end{array} \right.$$

In order to bound the curvature constant of $-f$, we need several preparatory lemmas.

Lemma A.5 ([4, Fact 7.4.9])**.** For any $A \in \mathbb{R}^{n\times m}, B \in \mathbb{R}^{m\times l}, C \in \mathbb{R}^{l\times k}$, and $D \in \mathbb{R}^{k\times n}$, we have

$$\operatorname{Tr}[ABCD] = \operatorname{vec}(A)^\top (B \otimes D^\top) \operatorname{vec}(C^\top),$$

where '$\otimes$' stands for the Kronecker product, while '$\operatorname{vec}(\cdot)$' denotes the vectorization of a matrix.

Lemma A.6 (Bounded feasible set)**.** If S is feasible in (5), then $S \preceq \bar{\sigma} I_d$, where $\bar{\sigma} \triangleq (\rho + \sqrt{\operatorname{Tr}[\Sigma]})^2$.

Proof of Lemma A.6. We seek an upper bound on the maximum eigenvalue of S uniformly across all covariance matrices S feasible in (5), that is, we seek an upper bound on the optimal value of

$$\begin{array}{ll} \sup\limits_{S\succeq 0} & \|S\| \\ \text{s. t.} & \mathrm{Tr}\left[S+\Sigma-2\left(\Sigma^{\frac{1}{2}}S\Sigma^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]\leq\rho^2. \end{array} \tag{A.8}$$

Problem (A.8) is a non-convex optimization problem because we maximize a convex function (the spectral norm of S) over a convex set. An easily computable upper bound is obtained by solving

$$\begin{array}{ll} \sup\limits_{S\succeq 0} & \langle S, I_d\rangle \\ \text{s. t.} & \mathrm{Tr}\left[S+\Sigma-2\left(\Sigma^{\frac{1}{2}}S\Sigma^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]\leq\rho^2. \end{array} \tag{A.9}$$

Indeed, note that $\mathrm{Tr}\,[S]=\langle S, I_d\rangle \geq \|S\|$, where the inequality holds because $S\succeq 0$. By Lemma A.2, which studies a more general problem with an arbitrary linear objective function $\langle S, D\rangle$, problem (A.9) has an analytical solution that is found by solving the following algebraic equation in γ.

$$\rho^2-\left\langle \Sigma, \left(I_d-\gamma^\star(\gamma^\star I_d-I_d)^{-1}\right)^2\right\rangle=0 \iff \rho^2-\left(\frac{1}{\gamma^\star-1}\right)^2\mathrm{Tr}\,[\Sigma]=0$$

In the special case considered here, this equation can be solved in closed form, and there is no need for a bisection algorithm. Specifically, we have $\gamma^\star=1+\sqrt{\mathrm{Tr}\,[\Sigma]}/\rho$, and thus (A.9) is solved by

$$S^\star=(\gamma^\star)^2(\gamma^\star I_d-I_d)^{-1}\Sigma(\gamma^\star I_d-I_d)^{-1}=\left(\frac{\gamma^\star}{\gamma^\star-1}\right)^2\Sigma=\frac{(\rho+\sqrt{\mathrm{Tr}\,[\Sigma]})^2}{\mathrm{Tr}\,[\Sigma]}\Sigma.$$

Therefore, problem (A.8) is upper bounded by $\mathrm{Tr}\,[S^\star]=(\rho+\sqrt{\mathrm{Tr}\,[\Sigma]})^2$. □

For ease of exposition, we now define the (compact) feasible set of problem (5) as

$$\mathcal{S}\triangleq\left\{S\in\mathbb{S}_+^d : \mathrm{Tr}[S+\Sigma-2(\Sigma^{\frac{1}{2}}S\Sigma^{\frac{1}{2}})^{\frac{1}{2}}]\leq\rho^2,\quad S\succeq\underline{\sigma}I_d\right\} \tag{A.10}$$

Lemma A.7 (Curvature bound)**.** The curvature constant C_{-f} of the (negative) objective function $-f$ over the feasible set $\mathcal{S}$ satisfies $C_{-f}\leq\overline{C}\triangleq 2\bar{\sigma}^4/\underline{\sigma}^3$.

Proof of Lemma A.7. We first expand the negative objective function $-f$ at $S\in\mathbb{S}_+^d$. By Lemma A.5, for any symmetric perturbation matrix Δ with a characteristic block structure of the form

$$\Delta=\begin{bmatrix}\Delta_{xx} & \Delta_{xy}\\ \Delta_{xy}^\top & \Delta_{yy}\end{bmatrix}\in\mathbb{S}^d,$$

the negative objective function $-f(S+\Delta)$ can be expressed as

$$\begin{aligned}
&\mathrm{Tr}\left[-S_{xx}-\Delta_{xx}+(S_{xy}+\Delta_{xy})(S_{yy}+\Delta_{yy})^{-1}(S_{yx}+\Delta_{xy}^\top)\right]\\
&=\mathrm{Tr}\left[-S_{xx}-\Delta_{xx}\right]+\\
&\qquad\mathrm{Tr}\left[(S_{xy}+\Delta_{xy})S_{yy}^{-1}(I_m-\Delta_{yy}S_{yy}^{-1}+(\Delta_{yy}S_{yy}^{-1})^2+\mathcal{O}(\|\Delta_{yy}\|^3))(S_{yx}+\Delta_{xy}^\top)\right]\\
&=\mathrm{Tr}\left[-S_{xx}+S_{xy}S_{yy}^{-1}S_{yx}\right]-\mathrm{Tr}\left[\Delta_{xx}-\Delta_{xy}S_{yy}^{-1}S_{yx}+S_{xy}S_{yy}^{-1}\Delta_{yy}S_{yy}^{-1}S_{yx}-S_{xy}S_{yy}^{-1}\Delta_{xy}^\top\right]\\
&\quad-\mathrm{Tr}\left[\Delta_{xy}S_{yy}^{-1}\Delta_{yy}S_{yy}^{-1}S_{yx}-\Delta_{xy}S_{yy}^{-1}\Delta_{xy}^\top+S_{xy}S_{yy}^{-1}\Delta_{yy}S_{yy}^{-1}\Delta_{xy}^\top-S_{xy}S_{yy}^{-1}(\Delta_{yy}S_{yy}^{-1})^2S_{yx}\right]\\
&\quad+\mathcal{O}(\|\Delta\|^3)\\
&=\mathrm{Tr}\left[-S_{xx}+S_{xy}S_{yy}^{-1}S_{yx}\right]-\langle D,\Delta\rangle+\frac{1}{2}\begin{bmatrix}\mathrm{vec}\,\Delta_{xx}\\ \mathrm{vec}\,\Delta_{xy}\\ \mathrm{vec}\,\Delta_{yy}\end{bmatrix}^\top H\begin{bmatrix}\mathrm{vec}\,\Delta_{xx}\\ \mathrm{vec}\,\Delta_{xy}\\ \mathrm{vec}\,\Delta_{yy}\end{bmatrix}+\mathcal{O}(\|\Delta\|^3),
\end{aligned}$$

where

$$D=\begin{bmatrix}I_n & -S_{xy}S_{yy}^{-1}\\ -S_{yy}^{-1}S_{yx} & S_{yy}^{-1}S_{yx}S_{xy}S_{yy}^{-1}\end{bmatrix}\in\mathbb{S}_+^d \tag{A.11}$$

and

$$H = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2S_{yy}^{-1} \otimes I_n & -2S_{yy}^{-1} \otimes S_{xy}S_{yy}^{-1} \\ 0 & -2S_{yy}^{-1} \otimes S_{yy}^{-1}S_{yx} & 2S_{yy}^{-1} \otimes S_{yy}^{-1}S_{yx}S_{xy}S_{yy}^{-1} \end{bmatrix} \in \mathbb{S}_+^{(n^2+nm+m^2)}.$$

Note that the matrix D represents the gradient of f, which plays a crucial role in the Frank-Wolfe algorithm. Similarly, H can be viewed as a compressed version of the Hessian matrix of $-f$, where the redundant rows and columns corresponding to S_{yx} have been eliminated. Thus, the Lipschitz constant of the gradient ∇f can be upper bounded by the largest eigenvalue of H, which is given by

$$\|H\| = 2\|S_{yy}^{-1} \otimes D\| = 2\|S_{yy}^{-1}\| \cdot \|D\|. \tag{A.12}$$

By a standard Schur complement argument, we then have

$$S = \begin{bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{bmatrix} = \begin{bmatrix} I_n & S_{xy}S_{yy}^{-1} \\ 0 & I_m \end{bmatrix} \begin{bmatrix} S_{xx} - S_{xy}S_{yy}^{-1}S_{yx} & 0 \\ 0 & S_{yy} \end{bmatrix} \begin{bmatrix} I_n & 0 \\ S_{yy}^{-1}S_{yx} & I_m \end{bmatrix}.$$

Next, define the set

$$\mathcal{V} \triangleq \left\{ z = [x^\top, y^\top]^\top \in \mathbb{R}^d : S_{yy}^{-1}S_{yx}\, x + y = 0 \right\},$$

and note that any $z \in \mathcal{V}$ satisfies

$$\begin{bmatrix} I_n & 0 \\ S_{yy}^{-1}S_{yx} & I_m \end{bmatrix} z = \begin{bmatrix} x \\ 0 \end{bmatrix}.$$

Thus, by the definition of the smallest eigenvalue, we have

$$\lambda_{\min}(S) = \min_{z \neq 0} \frac{z^\top S z}{z^\top z} \leq \min_{\substack{z \neq 0 \\ z \in \mathcal{V}}} \frac{z^\top S z}{z^\top z} \leq \lambda_{\min}(S_{xx} - S_{xy}S_{yy}^{-1}S_{yx}) \implies S_{xx} - S_{xy}S_{yy}^{-1}S_{yx} \succeq \underline{\sigma} I_n.$$

Moreover, by the Cauchy interlacing theorem [4, Theorem 8.4.5], Lemma A.6, and basic properties of the spectral norm, we have

$$\|S_{yy}^{-1}\| \leq \|S^{-1}\| \leq \frac{1}{\underline{\sigma}}, \quad \|S_{xx}\| \leq \bar{\sigma} \quad \text{and} \quad \|S_{yy}\| \leq \bar{\sigma}.$$

Using the above inequalities, one can show that

$$\frac{1}{\bar{\sigma}} I_m \preceq S_{yy}^{-1} \implies S_{xy}S_{yx} \preceq \bar{\sigma} S_{xy}S_{yy}^{-1}S_{yx}$$

and

$$\bar{\sigma} I_n \succeq S_{xx} \succeq S_{xx} - S_{xy}S_{yy}^{-1}S_{yx} \succeq \underline{\sigma} I_n \implies S_{xy}S_{yy}^{-1}S_{yx} \preceq (\bar{\sigma} - \underline{\sigma})\, I_n.$$

Setting $B = [I_n, -S_{xy}S_{yy}^{-1}]$, the above inequalities imply that

$$\begin{aligned} \|D\| = \|B^\top B\| = \|BB^\top\| = \|I_n + S_{xy}S_{yy}^{-2}S_{yx}\| &= 1 + \|S_{xy}S_{yy}^{-2}S_{yx}\| \\ &= 1 + \|S_{yy}^{-1}S_{yx}S_{xy}S_{yy}^{-1}\| \\ &\leq 1 + \|S_{yy}^{-1}\|^2 \cdot \|S_{yx}S_{xy}\| \\ &\leq 1 + \frac{\bar{\sigma}(\bar{\sigma} - \underline{\sigma})}{\underline{\sigma}^2} \leq \frac{\bar{\sigma}^2}{\underline{\sigma}^2}. \end{aligned}$$

By combining the last estimate with (A.12), we then find that the Lipschitz constant of ∇f satisfies

$$\mathrm{Lip}(\nabla f) = \|H\| \leq \frac{2\bar{\sigma}^2}{\underline{\sigma}^3}.$$

The diameter of the feasible set $\mathcal{S}$ with respect to the Frobenius norm satisfies

$$\mathrm{diam}(\mathcal{S}) = \sup_{S_1, S_2 \in \mathcal{S}} \|S_1 - S_2\|_F \leq \sup_{S_1, S_2 \in \mathcal{S}} \mathrm{Tr}\,[S_1 - S_2] \leq \sup_{S \in \mathcal{S}} \mathrm{Tr}\,[S] \leq \bar{\sigma},$$

where the first inequality holds due to [4, Equation (9.2.16)], and the last inequality follows from the proof of Lemma A.6. Therefore, by [13, Lemma 7], the curvature constant C_{-f} admits the estimate

$$C_{-f} \leq (\mathrm{diam}(\mathcal{S}))^2 \mathrm{Lip}(\nabla f) \leq \frac{2\bar{\sigma}^4}{\underline{\sigma}^3}.$$

This observation completes the proof. □

Proof of Theorem 3.3. By Lemma A.7, the curvature constant C_{-f} is bounded, and thus one can directly apply [13, Theorem 1] to find the convergence rate. □

Appendix B Sequential versus Static Estimation

We have resolved the filtering problem underlying Figure 4(c) as a single (static) estimation problem in the spirit of Section 2, where the entire observation history $Y_t \triangleq (y_1, \ldots, y_1)$ is interpreted as a single observation used to predict x_t. To our surprise, we found that the sequential filtering approach advocated in Section 4 outperforms this alternative static approach even if an oracle reveals the optimal radius of the ambiguity set (for $t = 100$, *e.g.*, the static estimation error is 37.5 dB, while the sequential estimation error is only 24.5 dB). In fact, for the static estimation problem the optimal radius of the Wasserstein ball is $\rho = 0$ whenever $t \geq 5$, that is, robustification does not improve performance. In contrast, in the sequential filtering approach robustification always helps. A possible explanation for this observation is that in the static approach our lack of information about the system uncertainty propagates through the dynamics. As such, it renders robust estimation ineffective when applied globally to the entire observation history at once. In contrast, in the sequential approach the robustification at each stage appears to limit such an uncertainty propagation.

Acknowledgments We gratefully acknowledge financial support from the Swiss National Science Foundation under grant BSCGI0_157733.

References

[1] B. D. Anderson and J. B. Moore. *Optimal Filtering*. Prentice Hall, 1979.

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[3] T. Başar and P. Bernhard. *$\mathcal{H}_\infty$-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Springer, 2008.

[4] D. S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, 2009.

[5] D. Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.

[6] D. Bertsekas and I. Rhodes. Recursive state estimation for a set-membership description of uncertainty. *IEEE Transactions on Automatic Control*, 16(2):117–128, 1971.

[7] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[8] Y. Chen, J. Ye, and J. Li. A distance for HMMs based on aggregated Wasserstein metric and state registration. In *European Conference on Computer Vision*, pages 451–466, 2016.

[9] M. Cuturi and D. Avis. Ground metric learning. *The Journal of Machine Learning Research*, 15(1):533–564, 2014.

[10] Y. C. Eldar and N. Merhav. A competitive minimax approach to robust estimation of random parameters. *IEEE Transactions on Signal Processing*, 52(7):1931–1946, 2004.

[11] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics*, 3(1-2):95–110, 1956.

[12] C. R. Givens and R. M. Shortt. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

[13] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435, 2013.

[14] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, 2006.

[15] B. C. Levy and R. Nikoukhah. Robust state space filtering under incremental model perturbations subject to a relative entropy tolerance. *IEEE Transactions on Automatic Control*, 58(3):682–695, 2013.

[16] V. Masarotto, V. M. Panaretos, and Y. Zemel. Procrustes metrics on covariance operators and optimal transportation of Gaussian processes. *preprint at arXiv:1801.01990*, 2018.

[17] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

[18] V. A. Nguyen, D. Kuhn, and P. Mohajerin Esfahani. Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *Optimization Online*, 2018.

[19] L. Ning, T. Georgiou, A. Tannenbaum, and S. Boyd. Linear models based on noisy data and the Frisch scheme. *SIAM Review*, 57(2):167–197, 2015.

[20] C. R. Rao. *Linear Statistical Inference and its Applications*. Wiley, 1973.

[21] A. Rolet, M. Cuturi, and G. Peyré. Fast dictionary learning with a smoothed Wasserstein loss. In *Artificial Intelligence and Statistics*, pages 630–638, 2016.

[22] A. H. Sayed. A framework for state-space estimation with uncertain models. *IEEE Transactions on Automatic Control*, 46(7):998–1013, 2001.

[23] S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani. Regularization via mass transportation. *preprint at arXiv:1710.10016*, 2017.

[24] S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.

[25] S. Shtern and A. Ben-Tal. A semi-definite programming approach for robust tracking. *Mathematical Programming*, 156(1-2):615–656, 2016.

[26] A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[27] J. L. Speyer, C. Fan, and R. N. Banavar. Optimal stochastic estimation with exponential cost criteria. In *IEEE Conference on Decision and Control*, pages 2293–2299, 1992.

[28] H. Xu and S. Mannor. A Kalman filter design based on the performance/robustness tradeoff. *IEEE Transactions on Automatic Control*, 54(5):1171–1175, 2009.

[29] K. Zhou, J. C. Doyle, and K. Glover. *Robust and Optimal Control*. Prentice Hall, 1996.

[30] M. Zorzi. Robust Kalman filtering under model perturbations. *IEEE Transactions on Automatic Control*, 62(6):2902–2907, 2017.