Wasserstein Distributionally Robust Optimization with Wasserstein Barycenters

Tim Tsz-Kit Lau[*] Han Liu[†]

May 30, 2022

Abstract

In many applications in statistics and machine learning, the availability of data samples from multiple possibly heterogeneous sources has become increasingly prevalent. On the other hand, in distributionally robust optimization, we seek data-driven decisions which perform well under the most adverse distribution from a nominal distribution constructed from data samples within a certain discrepancy of probability distributions. However, it remains unclear how to achieve such distributional robustness in model learning and estimation when data samples from multiple sources are available. In this work, we propose constructing the nominal distribution in optimal transport-based distributionally robust optimization problems through the notion of Wasserstein barycenter as an aggregation of data samples from multiple sources. Under specific choices of the loss function, the proposed formulation admits a tractable reformulation as a finite convex program, with powerful finite-sample and asymptotic guarantees. As an illustrative example, we demonstrate with the problem of distributionally robust sparse inverse covariance matrix estimation for zero-mean Gaussian random vectors that our proposed scheme outperforms other widely used estimators in both the low- and high-dimensional regimes.

1 Introduction

In various statistical and machine learning applications, data samples are collected from multiple sources, which can be viewed as samples drawn from multiple data distributions. A notable example is federated learning (Kairouz et al., 2021; McMahan et al., 2017), in which many users collaboratively learn a common model but the samples collected by the clients might have highly heterogeneous distributions. This distribution heterogeneity leads to difficulty in building a robust model in two aspects: (i) how to aggregate estimations of the distributions with data samples from these sources; (ii) how to perform robust estimation with this data aggregation given distributional uncertainty.

In practice, the first issue is usually dealt with by simply taking a simple (weighted) average of the distribution estimates, whereas the second one is tackled by minimizing the weighted aggregate loss with possibly different weights. However, the mixture distribution constructed from the weighted average of distributions does not take into account the geometric structure of data samples, thus failing to well summarize the characteristics from all sources. In this work, we consider the notion of barycenter (a.k.a. Fréchet mean) in the space of probability distributions endowed with the Wasserstein distance, called *Wasserstein barycenter* (Agueh and Carlier, 2011), which is a nonlinear interpolation between distributions.

To perform robust estimation of models against distributional uncertainty, *distributionally robust optimization* (DRO; Delage and Ye, 2010; Goh and Sim, 2010; Wiesemann et al., 2014) has been shown to be a powerful modeling framework, which has aroused much attention in the machine learning community lately attributed to its connections to generalization, regularization and robustness.

[*]Department of Statistics and Data Science, Northwestern University, Evanston, IL 60208, USA; Email: timlautk@u.northwestern.edu.

[†]Department of Computer Science and Department of Statistics and Data Science, Northwestern University, Evanston, IL 60208, USA; Email: hanliu@northwestern.edu.

Contributions. We thus propose a unified approach to overcome these two aspects of difficulty. We first construct an aggregate distribution of multiple data distributions through Wasserstein barycenter, followed by using it to define an ambiguity set in a DRO problem, which is a family of distributions lying within a certain Wasserstein distance from this Wasserstein barycenter, coined the *Wasserstein barycentric ambiguity set*. We hence introduce *Wasserstein Barycentric DRO* (WBDRO) as a general aggregate data-driven decision making framework with distributional robustness against uncertainty arising in multiple possibly heterogeneous unknown true distributions.

We establish finite-sample guarantees and asymptotic consistency results for WBDRO. We also consider an approximation of the Wasserstein ambiguity set by characterizing it using only the first two moments of the family of distributions and those of the nominal distribution, called the *Gelbrich ambiguity set*. We further extend this construction in the case of multiple nominal distributions using the 2-Wasserstein barycenter. We also exemplify WBDRO through distributionally robust maximum likelihood estimation for sparse inverse covariance matrices of zero-mean Gaussian random vectors, which numerically outperforms other widely-used estimators.

1.1 Related Work

Distributionally Robust Optimization. As a powerful modeling framework, DRO has recently found a wide range of applications in statistics and machine learning (Bertsimas and Van Parys, 2022; Blanchet et al., 2019a; Duchi and Namkoong, 2021; Duchi et al., 2021; Li et al., 2021; Nguyen et al., 2021b, 2022; Shafieezadeh-Abadeh et al., 2015, 2019; Taskesen et al., 2021b), signal processing (Shafieezadeh-Abadeh et al., 2018), portfolio selection and maximization (Blanchet et al., 2021a; Nguyen et al., 2021a,c; Obłój and Wiesel, 2021), etc. One key component of DRO is the choice of data-driven ambiguity sets, which can be defined through f-divergence (Ben-Tal et al., 2013; Duchi and Namkoong, 2021; Duchi et al., 2021), Wasserstein distance (Blanchet et al., 2019b; Gao, 2020; Gao and Kleywegt, 2016; Gao et al., 2017; Pflug and Wozabal, 2007), generalized moment constraints (Bertsimas et al., 2018; Delage and Ye, 2010; Goh and Sim, 2010; Wiesemann et al., 2014), maximum mean discrepancy (MMD; Staib and Jegelka, 2019), etc. Tractable reformulation as finite convex programs are available for DRO problems with these different ambiguity sets. We refer to Carmon and Hausler (2022); Haddadpour et al. (2022); Jin et al. (2021); Levy et al. (2020); Li et al. (2021); Yu et al. (2022) for recent advances in the computational perspectives of DRO, and Zhen et al. (2021) for a review of the mathematical foundations of DRO.

Notion of Mean Distributions. Fréchet mean or barycenter in different metric spaces, as a notion of mean distributions, has long been a central object in statistical analysis. Notably, a (weighted) average of distributions on $\mathbb{R}^d$ is a barycenter in Euclidean space. The Wasserstein barycenter (Agueh and Carlier, 2011; Kroshnin, 2018) is a more appropriate notion of mean distributions since the geometric structure of the distributions can be considered (see e.g., Backhoff-Veraguas et al., 2018). This notion has already appeared in various applications in statistics and machine learning (Backhoff-Veraguas et al., 2018; Bigot et al., 2019b; Bishop, 2014; Bishop and Doucet, 2021; Schmitz et al., 2018; Srivastava et al., 2018; Yang and Tabak, 2021). In particular, the work Álvarez-Esteban et al. (2018) shares similar motivation to ours, which is to perform consensus-based estimation combining several estimations of probability distributions.

Learning with Data from Multiple Sources. Modern machine learning applications involve the use of data collected from multiple sources. Such examples include federated learning (Kairouz et al., 2021; McMahan et al., 2017; Wang et al., 2021a), (multiple-source) domain adaptation (Mansour et al., 2021; Zhang et al., 2021), information fusion and network consensus (Bishop, 2014; Bishop and Doucet, 2021). However, the consensus problem indeed has a much longer history (see e.g., DeGroot, 1974).

A more detailed discussion on other related prior work can be found in Appendix A.

2 Preliminaries

Notation. We denote by $I_d \in \mathbb{R}^{d\times d}$ the $d \times d$ identity matrix and $\mathbf{1}_d \in \mathbb{R}^d$ the d-dimensional all-one vector. The subscripts for dimensions are suppressed if they are clear from context. We define $[\![n]\!] := \{1, \ldots, n\}$ for

$n \in \mathbb{N}^*$. Let $\mathbb{S}_{++}^d$ (resp. $\mathbb{S}_+^d$) denote the set of symmetric positive (resp. semi-)definite matrices. $\mathscr{P}(\mathcal{X})$ is the set of Borel probability measures over the Polish space $\mathcal{X}$, $\mathscr{P}_k(\mathcal{X})$ is the set of probability measures over $\mathcal{X}$ with finite k-order moments, and $\mathscr{P}_k^{\mathrm{ac}}(\mathcal{X})$ is the set of absolutely continuous probability measures over $\mathcal{X}$ (w.r.t. the Lebesgue measure) with finite k-order moments. The set $\triangle^d \coloneqq \{\boldsymbol{p} \in [0, +\infty)^d : \langle \boldsymbol{p}, \mathbf{1}_d \rangle = 1\}$ is the $(d-1)$-dimensional probability simplex, where $\langle \cdot, \cdot \rangle$ is the usual inner product. We denote by $\mathcal{N}(\mu, \Sigma)$ a Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{S}_+^d$, and δ_x a Dirac measure at point $x \in \mathcal{X}$.

Optimal Transport. We introduce several notions from optimal transport (OT) used throughout the whole paper, which can be found in various monographs on the subject (Ambrosio et al., 2021; Figalli and Glaudo, 2021; Peyré and Cuturi, 2019; Santambrogio, 2015; Villani, 2003, 2009). We also refer to Panaretos and Zemel (2019, 2020); Peyré and Cuturi (2019) for comprehensive reviews of recent advances of optimal transport in machine learning and statistics. Let $\Omega \subseteq \mathbb{R}^m$ be a closed convex set. For $p \in [1, +\infty)$, the p-Wasserstein distance between two probability measures $\rho, \nu \in \mathscr{P}_p(\Omega)$ is defined by

$$\mathsf{W}_p(\rho, \nu) \coloneqq \left(\inf_{\pi \in \Pi(\rho, \nu)} \int_{\Omega \times \Omega} \|x - y\|^p \, \mathrm{d}\pi(x, y) \right)^{1/p}, \tag{2.1}$$

where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^m$, and $\Pi(\rho, \nu)$ denotes the set of joint distributions on $\mathbb{R}^m \times \mathbb{R}^m$ with ρ and ν as marginals. The p-Wasserstein distance is a distance on the space $\mathscr{P}_p(\Omega)$ (see e.g., Figalli and Glaudo, 2021, Theorem 3.1.5). We call the metric space $\mathscr{W}_p(\Omega) \coloneqq (\mathscr{P}_p(\Omega), \mathsf{W}_p)$ the *p-Wasserstein space* (Ambrosio et al., 2005).

The notion of Wasserstein barycenter (Agueh and Carlier, 2011) can be viewed as the mean of probability distributions in the Wasserstein space. For $p \in [1, +\infty)$, the *p-Wasserstein barycenter* of $\mathbb{P} \in \mathscr{W}_p(\mathscr{P}_p(\Omega))$ is defined by

$$\mathsf{b}_p(\mathbb{P}) \coloneqq \operatorname*{argmin}_{\nu \in \mathscr{P}_p(\mathbb{R}^m)} \mathbb{E}_{\rho \sim \mathbb{P}}\big[\mathsf{W}_p^p(\nu, \rho)\big], \tag{2.2}$$

where $\rho \in \mathscr{P}_p(\Omega)$ is a random measure with distribution $\mathbb{P}$. If we take $\mathbb{P} = \sum_{k=1}^K \lambda_k \delta_{\rho_k}$ in (2.2), where $\boldsymbol{\lambda} = (\lambda_k)_{k \in [\![K]\!]} \in \triangle^K$, we recover the *$\boldsymbol{\lambda}$-weighted empirical p-Wasserstein barycenter*, defined by

$$\widehat{\mathsf{b}}_{\boldsymbol{\lambda}, p}(\rho_1, \ldots, \rho_K) \coloneqq \operatorname*{argmin}_{\nu \in \mathscr{P}_p(\mathbb{R}^m)} \sum_{k=1}^K \lambda_k \mathsf{W}_p^p(\nu, \rho_k).$$

Thus, we use the term p-Wasserstein barycenter to refer to both empirical and population p-Wasserstein barycenters whenever it is clear from context. Note that Wasserstein barycenters do not always exist, and might not be unique if exist. Technical conditions for their existence and uniqueness are studied in e.g., Agueh and Carlier (2011); Le Gouic and Loubes (2017).

Distributionally Robust Optimization. In DRO, we investigate a learning problem under distributional uncertainty which is casted as a generic expected loss minimization framework. The loss function $\ell \colon \mathbb{R}^m \to \overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{\pm\infty\}$ is a function of the the uncertainty vector $\xi \in \mathbb{R}^m$ whose distribution $\mathbb{P}$ is supported on $\Xi \subseteq \mathbb{R}^m$. The *risk* (or *expected loss*) of a decision $\ell \in \mathcal{L}$ is defined as

$$\mathcal{R}_{\mathbb{P}}(\ell) \coloneqq \mathbb{E}_{\xi \sim \mathbb{P}}[\ell(\xi)],$$

where $\mathcal{L}$ is the set of all admissible loss functions. The *optimal risk* is then defined as the infimum of the risk over $\mathcal{L}$. However, $\mathbb{P}$ is often unknown in practice except for some limited statistical and structural information about it. We thus assume that $\mathbb{P}$ is known to lie in an *ambiguity set* $\mathcal{U}_\varepsilon(\widehat{\mathbb{P}})$, which is a ball of radius $\varepsilon \geqslant 0$ in $\mathscr{P}(\Xi)$ centered at the nominal distribution $\widehat{\mathbb{P}}$ in some discrepancy between probability distributions. We can then define the *worst-case risk* of $\ell \in \mathcal{L}$ by

$$\mathcal{R}_{\mathcal{U}_\varepsilon(\widehat{\mathbb{P}})}(\ell) \coloneqq \sup_{\mathbb{P} \in \mathcal{U}_\varepsilon(\widehat{\mathbb{P}})} \mathcal{R}_{\mathbb{P}}(\ell). \tag{2.3}$$

Such a nominal distribution $\widehat{\mathbb{P}}$ is usually constructed from a set of observed data $\mathcal{D} := \{z_i\}_{i=1}^n \subset \mathbb{R}^m$, e.g., its empirical measure $\frac{1}{n}\sum_{i=1}^n \updelta_{z_i}$. The distributionally robust optimization (DRO) problem seeks decisions achieving the *optimal worst-case risk*

$$\mathcal{R}_{\mathcal{U}_\varepsilon(\widehat{\mathbb{P}})}(\mathcal{L}) := \inf_{\ell\in\mathcal{L}} \mathcal{R}_{\mathcal{U}_\varepsilon(\widehat{\mathbb{P}})}(\ell). \tag{2.4}$$

Remark 2.1. Note that if the loss function is parameterized by the decision $x \in \mathcal{X} \subseteq \mathbb{R}^d$, i.e., $\ell\colon \mathcal{X}\times\mathbb{R}^m \to \overline{\mathbb{R}}$, then the risk can be defined in terms of x as $\mathcal{R}_{\mathbb{P}}(x) := \mathbb{E}_{\xi\sim\mathbb{P}}[\ell(x,\xi)]$. The worst-case risk and the worst-case optimal risk can be defined similarly.

In this paper, we consider the ambiguity set defined via the p-Wasserstein distance. Then, the *p-Wasserstein ambiguity set* is defined by

$$\mathcal{W}_{\varepsilon,p}(\widehat{\mathbb{P}}) := \{\mathbb{Q} \in \mathscr{P}_p(\Xi) : \mathsf{W}_p(\mathbb{Q}, \widehat{\mathbb{P}}) \leqslant \varepsilon\},$$

where $\Xi \subseteq \mathbb{R}^m$ is a closed set which is known to contain the support of the unknown true distribution $\mathbb{P}^\star$ and $\varepsilon \geqslant 0$. Such a DRO formulation is called the *Wasserstein DRO* (WDRO).

3 Learning with Aggregation of Multiple Distributions

Problem Formulation. In different centralized model learning scenarios with data from multiple sources, such as federated learning, the learning objective can usually be casted as a *stochastic composition optimization* problem (see e.g., Wang et al., 2021a; Yuan et al., 2022):

$$\operatorname*{minimize}_{x\in\mathcal{X}}\ F(x) := \mathbb{E}_{k\sim\mathbb{D}}[f_k(x)], \quad \text{where} \quad f_k(x) := \mathbb{E}_{\xi\sim\mathbb{P}_k}[\ell(x,\xi)], \tag{3.1}$$

where $x \in \mathcal{X}$ is the parameter of the global model for some closed convex set $\mathcal{X} \subseteq \mathbb{R}^d$, $\xi \in \Xi$ is a random vector representing an input-output pair for some sample space $\subseteq \mathbb{R}^m$, $f_k\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ is the local objective function of the kth source, $\mathbb{P}_k$ is the distribution associated to the kth source, and $\mathbb{D}$ is a distribution supported on the set of sources $\mathcal{K}$. Assuming there is only a finite number of K sources, i.e., $\mathcal{K} = [\![K]\!]$, then the objective in (3.1) can be written as $F_{\boldsymbol{\lambda}}(x) := \sum_{k=1}^K \lambda_k f_k(x)$, where $\mathbb{D}$ is taken to be a categorical distribution with probabilities $\boldsymbol{\lambda} = (\lambda_k)_{k\in[\![K]\!]} \in \triangle^K$.

Usually, each data source k has a finite number of local samples, denoted by $\mathcal{D}_k = (z_{k,1}, \ldots, z_{k,n_k})$, where n_k is the sample size of the kth source and $N := \sum_{k=1}^K n_k$. Using their empirical distributions $\widehat{\mathbb{P}}_k := \frac{1}{n_k}\sum_{i=1}^{n_k} \updelta_{z_{k,i}}$, with $\widehat{f}_k(x) := \mathbb{E}_{\xi\sim\widehat{\mathbb{P}}_k}[\ell(x,\xi)]$, we usually solve the following *empirical risk minimization* (ERM) problem in practice:

$$\operatorname*{minimize}_{x\in\mathcal{X}}\ \widehat{F}_{\boldsymbol{\lambda}}(x) := \sum_{k=1}^K \lambda_k \widehat{f}_k(x) = \sum_{k=1}^K \frac{\lambda_k}{n_k}\sum_{i=1}^{n_k} \ell(x, z_{k,i}), \tag{3.2}$$

Note that $\boldsymbol{\lambda}$ is usually taken as the uniform distribution over the numbers of samples from the sources, i.e., $\lambda_k = n_k/N$, so that the ERM objective (3.2) is amount to an ERM objective with the union of all the local data samples.

However, as argued by Mohri et al. (2019); Ro et al. (2021), this choice of the uniform distribution is questionable since there is often a mismatch between the target distribution (for which the centralized model is learned) and the mixture distribution $\sum_{k=1}^K n_k\mathbb{P}_k/N$. Instead, the target distribution is better expressed as a $\boldsymbol{\lambda}$-mixture of $\mathbb{P}_1, \ldots, \mathbb{P}_K$, i.e., $\mathbb{P}_{\boldsymbol{\lambda}} := \sum_{k=1}^K \lambda_k\mathbb{P}_k$ for some $\boldsymbol{\lambda} \in \triangle^K$. Then, with the $\boldsymbol{\lambda}$-mixture of the empirical distributions $\widehat{\mathbb{P}}_{\boldsymbol{\lambda}} := \sum_{k=1}^K \lambda_k\widehat{\mathbb{P}}_k = \sum_{k=1}^K \frac{\lambda_k}{n_k}\sum_{i=1}^{n_k} \updelta_{z_{k,i}}$, it is not hard to see that the objective in (3.2) is equivalent to $\mathbb{E}_{\xi\sim\widehat{\mathbb{P}}_{\boldsymbol{\lambda}}}[\ell(x,\xi)]$.

Stochastic Barycentric Optimization. Let us recall that $\mathbb{P}_{\boldsymbol{\lambda}}$ is the $\boldsymbol{\lambda}$-weighted Euclidean barycenter of the distributions $\mathbb{P}_1, \ldots, \mathbb{P}_K$. Leveraging this important fact, we consider more generally a $\boldsymbol{\lambda}$-weighted barycenter $\widehat{\mathsf{b}}_{\boldsymbol{\lambda}}(\mathbb{P}_1, \ldots, \mathbb{P}_K)$ of $\mathbb{P}_1, \ldots, \mathbb{P}_K$ defined via some discrepancy between distributions such as the Wasserstein distance, so that (3.1) can be formulated with the objective

$$F_{\boldsymbol{\lambda}}^{\mathsf{b}}(x) := \mathbb{E}_{\xi\sim\widehat{\mathsf{b}}_{\boldsymbol{\lambda}}(\mathbb{P}_1,\ldots,\mathbb{P}_K)}[\ell(x,\xi)],$$

which we refer to as a *stochastic barycentric optimization (SBO)* problem. With the data samples $\mathcal{D}_1, \ldots, \mathcal{D}_K$, we also have its surrogate objective defined with the empirical distributions $\widehat{\mathbb{P}}_1, \ldots, \widehat{\mathbb{P}}_K$, given by

$$\widehat{F}^{\mathsf{b}}_{\boldsymbol{\lambda}}(x) \coloneqq \mathbb{E}_{\xi \sim \widehat{\mathsf{b}}_{\boldsymbol{\lambda}}(\widehat{\mathbb{P}}_1, \ldots, \widehat{\mathbb{P}}_K)}[\ell(x, \xi)].$$

Unfortunately, except for the case of the Euclidean barycenter, $\widehat{F}^{\mathsf{b}}_{\boldsymbol{\lambda}}$ usually cannot be expressed as a finite sum. This appears to be unfavorable computationally compared to (3.2). To solve it computationally, one can resort to ERM by drawing samples from the barycenter of empirical distributions. However, solving such an ERM problem requires (i) the computation of a barycenter; and (ii) sampling from such a barycenter. In the case of the Wasserstein barycenter, these tasks could be computationally intensive (Altschuler and Boix-Adserà, 2022) or not well addressed until recently (Daaloul et al., 2021). However, the choice of the Wasserstein barycenter over the Euclidean barycenter in SBO is justified in the sense that the Euclidean barycenter usually fails to take into account the underlying geometry of these distributions (see e.g., Backhoff-Veraguas et al., 2018). Thus, despite the potential computational obstacles, we specifically consider the Wasserstein barycenter (and possibly its entropic-regularized variants). We also provide further discussion on the connections of this formulation to other related machine learning paradigms in Appendix A.

4 Wasserstein Barycentric Distributionally Robust Optimization

Although the use of the Wasserstein barycenter might give a better consensus representation of samples from different sources, discrepancy between the target distribution and the Wasserstein barycenter of the empirical distributions might still arise, due to e.g., sampling errors and data heterogeneity across the sources. In a similar spirit to the single-source case, we propose to hedge against the impact of such model misspecification through the lens of WDRO.

Wasserstein Ambiguity Sets with Wasserstein Barycenters. Suppose that we have a finite number of K data sources with probability distributions $\mathbb{Q}_1, \ldots, \mathbb{Q}_K$ respectively. Aside from directly considering the notion of Wasserstein barycenters, one way to construct an ambiguity set out of these K probability distributions is to consider the intersection of the individual Wasserstein ambiguity sets $\bigcap_{k=1}^K \mathcal{W}_{\varepsilon,p}(\mathbb{Q}_k)$, but if the data sources are very heterogeneous then this could lead to an overly conservative ambiguity set using the same large ε. Alternatively, a less conservative ambiguity set based on these K distributions can be defined by

$$\widetilde{\mathcal{W}}_{\varepsilon,p}(\mathbb{Q}_1, \ldots, \mathbb{Q}_K; \boldsymbol{\lambda}) \coloneqq \left\{ \mathbb{P} \in \mathscr{P}_p(\Xi) : \sum_{k=1}^K \lambda_k \mathsf{W}_p^p(\mathbb{P}, \mathbb{Q}_k) \leqslant \varepsilon^p \right\}. \tag{4.1}$$

It is straightforward to observe that $\bigcap_{k=1}^K \mathcal{W}_{\varepsilon,p}(\mathbb{Q}_k) \subseteq \widetilde{\mathcal{W}}_{\varepsilon,p}(\mathbb{Q}_1, \ldots, \mathbb{Q}_K; \boldsymbol{\lambda})$ for any $\boldsymbol{\lambda} \in \triangle^K$. In the following, we illustrate that how the ambiguity set (4.1) is related to another ambiguity set defined with the $\boldsymbol{\lambda}$-weighted p-Wasserstein barycenter of $\mathbb{Q}_1, \ldots, \mathbb{Q}_K$.

Definition 4.1 (Wasserstein barycentric ambiguity set)**.** For $\boldsymbol{\lambda} \in \triangle^K$, the *$p$-Wasserstein barycentric ambiguity set* with radius $\varepsilon \geqslant 0$ centered at a $\boldsymbol{\lambda}$-weighted p-Wasserstein barycenter of $\mathbb{Q}_1, \ldots, \mathbb{Q}_K$ (if exists), denoted by $\overline{\mathbb{Q}}_{\boldsymbol{\lambda},p} \coloneqq \widehat{\mathsf{b}}_{\boldsymbol{\lambda},p}(\mathbb{Q}_1, \ldots, \mathbb{Q}_K)$, is defined by

$$\overline{\mathcal{W}}_{\varepsilon,p}(\mathbb{Q}_1, \ldots, \mathbb{Q}_K; \boldsymbol{\lambda}) \coloneqq \left\{ \mathbb{P} \in \mathscr{P}_p(\Xi) : \mathsf{W}_p(\mathbb{P}, \overline{\mathbb{Q}}_{\boldsymbol{\lambda},p}) \leqslant \varepsilon \right\}. \tag{4.2}$$

Note that $\overline{\mathcal{W}}_{\varepsilon,p}(\mathbb{Q}_1, \ldots, \mathbb{Q}_K; \boldsymbol{\lambda}) = \mathcal{W}_{\varepsilon,p}(\overline{\mathbb{Q}}_{\boldsymbol{\lambda},p})$. The ambiguity sets (4.1) and (4.2) of different radii (differed by a factor of 2^p) can be related by the following inclusion.

Theorem 4.2. *For $\boldsymbol{\lambda} \in \triangle^K$, suppose that a $\boldsymbol{\lambda}$-weighted p-Wasserstein barycenter of $\mathbb{Q}_1, \ldots, \mathbb{Q}_K$ exists. Then, for any $\varepsilon \geqslant 0$, the following inclusion of the ambiguity sets (4.1) and (4.2) of different radii holds:*

$$\widetilde{\mathcal{W}}_{\varepsilon,p}(\mathbb{Q}_1, \ldots, \mathbb{Q}_K; \boldsymbol{\lambda}) \subseteq \overline{\mathcal{W}}_{2^p \cdot \varepsilon,p}(\mathbb{Q}_1, \ldots, \mathbb{Q}_K; \boldsymbol{\lambda}). \tag{4.3}$$

All proofs of this paper are deferred to Appendix C. Consequently, it is more feasible to use the Wasserstein barycentric ambiguity set (4.2) since ε is often tuned in practice. We can therefore apply existing results of WDRO if the Wasserstein barycenter $\overline{\mathbb{Q}}_{\boldsymbol{\lambda}}$ exists and is available, without the need of redeveloping tools to handle the ambiguity set (4.1).

Remark 4.3. The overall rationale of the construction of (4.2) is that the most adverse distribution should be close to a population p-Wasserstein barycenter $\mathsf{b}^\star := \mathsf{b}_p(\mathbb{P}^\star) \in \mathscr{P}_p(\Xi)$ of the unknown true distribution (of distributions) $\mathbb{P}^\star \in \mathscr{W}_p(\mathscr{P}_p(\Xi))$ within a radius ε in W_p distance. This population barycenter can be approximated by its empirical counterpart (see Section 4.1 for details). This is as opposed to the usual WDRO in which the most adverse distribution is close to the unknown true distribution $\mathbb{P}^\star \in \mathscr{P}_p(\Xi)$ approximated by a nominal (empirical) distribution $\widehat{\mathbb{P}}$.

We now revisit the WDRO problem (2.4) with K available nominal distributions $\widehat{\mathbb{P}} = (\widehat{\mathbb{P}}_1, \dots, \widehat{\mathbb{P}}_K)$ and the Wasserstein barycentric ambiguity set $\mathcal{U}_\varepsilon(\widehat{\mathbb{P}}) = \overline{\mathcal{W}}_{\varepsilon,p}(\widehat{\mathbb{P}}; \boldsymbol{\lambda})$, which we refer to as the *Wasserstein barycentric DRO* (WBDRO). Consequently, due to the definition of the Wasserstein barycentric ambiguity set (4.2), solving a WBDRO involves two consecutive steps: (i) computing a Wasserstein barycenter $\widehat{\mathsf{b}}_{\boldsymbol{\lambda},p}(\widehat{\mathbb{P}}_1, \dots, \widehat{\mathbb{P}}_K)$; (ii) solving a WDRO problem. Such a problem decomposition enables us to leverage existing theoretical results and computational tools for Wasserstein barycenters (e.g., Álvarez-Esteban et al., 2016; Carlier et al., 2015; Chewi et al., 2020; Heinemann et al., 2022) and WDRO (e.g., Blanchet et al., 2021d) respectively.

In the original WDRO formulation (2.4), we usually observe some i.i.d. realizations $\mathcal{D} := \{z_i\}_{i=1}^n$ of the unknown true distribution $\mathbb{P}^\star$, from which we can construct a nominal distribution $\widehat{\mathbb{P}}^n$, e.g., the empirical distribution $\widehat{\mathbb{P}}^n = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$. However, in WBDRO, the way of constructing the K nominal distributions $\widehat{\mathbb{P}}_1, \dots, \widehat{\mathbb{P}}_K$ as approximations of their corresponding unknown true distributions $\mathbb{P}_1^\star, \dots, \mathbb{P}_K^\star$ becomes more subtle. For instance, they can be constructed from the same set of observed data $\mathcal{D}_n$ via resampling techniques such as bootstrap (requiring $K \leqslant n$), or from K different sources where the observed data at the kth source $\mathcal{D}_k = \{z_{i,k}\}_{i=1}^n$ are i.i.d. realizations of its unknown true distribution $\mathbb{P}_k^\star$ (assuming the same sample size n for simplicity). The former scenario is often used to avoid overfitting, while the latter is often encountered in the setting of federated learning with possibly heterogeneous data sources. Again, for each $k \in [\![K]\!]$, the nominal distributions $\widehat{\mathbb{P}}_k^n$ can be taken as the empirical distributions. Note that $\widehat{\mathbb{P}}_k^n$ converges to $\mathbb{P}_k^\star$ in W_p distance as $n \to \infty$ (see e.g., Bolley et al., 2007; Fournier and Guillin, 2015).

Under this construction, we are interested in statistical properties of WBDRO in the following two case: (i) $n \to \infty$ with fixed and finite K; (ii) $K \to \infty$ with fixed and finite n. We argue that both cases are well motivated by the two statistical frameworks with applications discussed in Boissard et al. (2015); Le Gouic and Loubes (2017).

4.1 Two Statistical Paradigms

Asymptotic Sample Size. Under this paradigm, there are K unknown true distributions $\mathbb{P}_1^\star, \dots, \mathbb{P}_K^\star \in \mathscr{P}_p(\Xi)$, where $K \in \mathbb{N}^*$ is finite and fixed. These are approximated by a sequence of distributions constructed from data samples, e.g., the empirical measures $\widehat{\mathbb{P}}_k^n := \frac{1}{n}\sum_{i=1}^n \delta_{z_{i,k}}$, $k \in [\![K]\!]$. A crucial result in this case is that, for $\boldsymbol{\lambda} \in \triangle^K$, a p-Wasserstein barycenter of $\widehat{\rho}_{\boldsymbol{\lambda}}^n := \sum_{k=1}^K \lambda_k \delta_{\widehat{\mathbb{P}}_k^n}$ (if exists) converges to a p-Wasserstein barycenter of the limit $\rho_{\boldsymbol{\lambda}} := \sum_{k=1}^K \lambda_k \delta_{\widehat{\mathbb{P}}_k^\star}$ in W_p distance as $n \to \infty$, by Le Gouic and Loubes (2017, Theorem 3).

Asymptotic Number of Data Sources. For now, we consider the case where the unknown true distribution $\mathbb{P}^\star \in \mathscr{W}_p(\mathscr{P}_p(\Xi))$ is approximated by a growing discrete distribution ρ_K supported on K elements, with $K \to \infty$. Consider a sequence of K distributions $\mathbb{P}_k \in \mathscr{P}_p(\Xi)$ with weights $\lambda_k^K \geqslant 0$ for each $k \in [\![K]\!]$, from which we define the sequence of distributions by $\rho_K := \sum_{k=1}^K \lambda_k^K \delta_{\mathbb{P}_k}$, where $K \in \mathbb{N}^*$. Assume that ρ_K converges to some distribution $\mathbb{P}^\star$ in W_p distance. Then the p-Wasserstein barycenter of ρ_K converges to the p-Wasserstein barycenter of $\mathbb{P}^\star$ in W_p distance as $K \to \infty$, by Le Gouic and Loubes (2017, Theorem 3).

Remark 4.4. The case of n and K both growing to infinity is even more of interest to our case. Indeed, since $\widehat{\mathbb{P}}_k^n \to \mathbb{P}_k^\star$ in W_p distance as $n \to \infty$, for each $k \in [\![K]\!]$, using the above argument with $\mathbb{P}_k$ replaced by $\mathbb{P}_k^\star$ for each $k \in [\![K]\!]$, the p-Wasserstein barycenter of $\sum_{k=1}^K \lambda_k^K \delta_{\mathbb{P}_k^\star}$ converges to the p-Wasserstein barycenter of $\mathbb{P}^\star$ in W_p distance as $K \to \infty$.

Remark 4.5. Under some more specific frameworks, e.g., in deformation models (Allassonnière et al., 2007, 2013), the 2-Wasserstein barycenter of $\widehat{\mathbb{P}}_1, \ldots, \widehat{\mathbb{P}}_K$ is a consistent estimate of $\mathbb{P}^\star$, in the sense that $\widehat{\mathsf{b}}_{\mathbf{1}/K,2}(\widehat{\mathbb{P}}_1, \ldots, \widehat{\mathbb{P}}_K) \to \mathbb{P}^\star$ as $K \to \infty$ in W_2 distance. In the case with empirical observations available, as both $n \to \infty$ and $K \to \infty$, $\widehat{\rho}_{n,K} \coloneqq \frac{1}{K}\sum_{k=1}^{K} \delta_{\widehat{\mathbb{P}}_k^n} \to \mathbb{P}^\star$ in W_2 distance. We refer to Boissard et al. (2015, Theorem 4.2 and Proposition 5.1) and Bigot and Klein (2018); Zemel and Panaretos (2019) for details.

4.2 Performance Guarantees

We now study the implications of the two statistical paradigms in Section 4.1 on performance guarantees of WBDRO. To simplify discussion, we only consider the case of $p = 2$ and equal weights, i.e., $\lambda_k = 1/K$ for each $k \in \llbracket K \rrbracket$. To simplify discussion, we also assume that all 2-Wasserstein barycenters exist (subject to some technical regularity conditions). In this subsection, with slight abuse of notation, we write $\widehat{\mathbb{P}} = (\widehat{\mathbb{P}}_k)_{k \in \llbracket K \rrbracket}$, $\widehat{\mathbb{P}}^n = (\widehat{\mathbb{P}}_k^n)_{k \in \llbracket K \rrbracket}$, and $\widehat{\mathbb{P}}^\star = (\widehat{\mathbb{P}}_k^\star)_{k \in \llbracket K \rrbracket}$.

The finite-sample guarantees of WBDRO are derived by the rates of convergence of Wasserstein barycenters (Ahidar-Coutrix et al., 2020; Le Gouic et al., 2021; Schötz, 2019), characterized by measure concentration of the 2-Wasserstein barycenter of the nominal distributions $\widehat{\mathbb{P}}_1^n, \ldots, \widehat{\mathbb{P}}_K^n$.

With data samples $\mathcal{D}_{n,k} = \{z_{i,k}\}_{i=1}^n$ for each $k \in \llbracket K \rrbracket$, let $\widehat{\rho}_{n,K} \coloneqq \frac{1}{K}\sum_{k=1}^{K} \delta_{\widehat{\mathbb{P}}_k^n}$ where $\widehat{\mathbb{P}}_k^n \coloneqq \frac{1}{n}\sum_{i=1}^{n} \delta_{z_{i,k}}$. For now, we let $K \in \mathbb{N}^*$ be finite and fixed. The following measure concentration result simplified from Le Gouic et al. (2021, Theorem 12) states that the 2-Wasserstein barycenter of sub-Gaussian $\widehat{\rho}_K^\star \coloneqq \frac{1}{K}\sum_{k=1}^{K} \delta_{\widehat{\mathbb{P}}_k^\star}$ should be contained in the 2-Wasserstein barycentric ambiguity set centered at the 2-Wasserstein barycenter of $\widehat{\rho}_{n,K}$ in WBDRO with high probability.

Theorem 4.6 (Concentration inequality). *Let $K \in \mathbb{N}^*$ be finite and fixed. Suppose that $\widehat{\rho}_K^\star \in \mathscr{W}_2(\mathscr{P}_2(\Xi))$ is sub-Gaussian with 2-Wasserstein barycenter $\widehat{\mathsf{b}}_K^\star \coloneqq \mathsf{b}_2(\widehat{\rho}_K^\star) = \widehat{\mathsf{b}}_{\mathbf{1}/K,2}(\widehat{\mathbb{P}}^\star) \in \mathscr{P}_2(\Xi)$. Then $\widehat{\mathsf{b}}_K^\star$ is unique and there exist constants $(c_1, c_2) \in (0, +\infty)^2$ independent of n such that for any $\beta \in (0,1)$, the concentration inequality*

$$\mathbb{P}^n\Big\{\widehat{\mathsf{b}}_K^\star \in \overline{\mathbb{W}}_{\varepsilon,2}(\widehat{\mathbb{P}}^n; \mathbf{1}/K)\Big\} \geqslant 1 - \beta - \mathrm{e}^{-c_2 n}$$

holds whenever ε exceeds

$$\varepsilon_n(\beta) = \sqrt{\frac{c_1}{n}\log\left(\frac{2}{\beta}\right)}. \tag{4.4}$$

Theorem 4.6 indicates that any 2-Wasserstein barycentric ambiguity set $\overline{\mathbb{W}}_\varepsilon(\widehat{\mathbb{P}}^n; \mathbf{1}/K)$ with radius $\varepsilon \geqslant \varepsilon_n(\beta)$ represents an approximate $(1-\beta)$-confidence region for $\widehat{\mathsf{b}}_K^\star$, which is a K-sample approximation of the 2-Wasserstein barycenter of the unknown true distribution $\mathbb{P}^\star$.

Note that the distributional uncertainty radius $\varepsilon_n(\beta)$ decays as $\mathscr{O}(n^{-1/2})$. Therefore, there is no curse of dimensionality in the uncertainty dimension m when choosing the distributional uncertainty radius $\varepsilon_n(\beta)$, as opposed to the case of WDRO (see Kuhn et al., 2019, §3 and Shafieezadeh-Abadeh et al., 2019, Remark 37).

From Theorem 4.6, we can immediately derive the following finite-sample guarantee.

Theorem 4.7 (Finite-sample guarantee). *Suppose that all conditions of Theorem 4.6 hold with $\varepsilon_n(\beta)$ defined in (4.4). Then for all $\beta \in (0,1)$ and $\varepsilon \geqslant \varepsilon_n(\beta)$, we have*

$$\mathbb{P}^n\Big\{(\forall \ell \in \mathcal{L})\ \mathcal{R}_{\widehat{\mathsf{b}}_K^\star}(\ell) \leqslant \mathcal{R}_{\overline{\mathbb{W}}_{\varepsilon,2}(\widehat{\mathbb{P}}^n;\mathbf{1}/K)}(\ell)\Big\} \geqslant 1 - \beta - \mathrm{e}^{-c_2 n}.$$

Theorem 4.7 asserts that the worst-case risk provides an upper confidence bound on the approximate true risk under the 2-Wasserstein barycenter of the unknown true distribution $\mathbb{P}^\star$ uniformly across all loss functions $\ell \in \mathcal{L}$. In particular, if we take ℓ to be an optimizer of $\mathcal{R}_{\overline{\mathbb{W}}_{\varepsilon,2}(\widehat{\mathbb{P}}^n;\mathbf{1}/K)}$, then this result implies that the optimal value of WBDRO provides an upper confidence bound on the out-of-sample performance of its optimizers.

When K is fixed and finite, we recall that $\varepsilon_n \to 0$ as $n \to \infty$. We can then derive asymptotic consistency of WBDRO in the sample size n, which asserts that the solution of WBDRO converges to the worst-case optimal risk under $\widehat{\mathsf{b}}_K^\star$ with a suitably chosen $\beta = \beta_n \to 0$ decaying to 0 as $n \to \infty$. On the other hand, let us also recall the 2-Wasserstein barycenter $\widehat{\mathsf{b}}_{\mathbf{1}/K,2}(\widehat{\mathbb{P}}^\star)$ converges to a 2-Wasserstein barycenter $\mathsf{b}_2(\mathbb{P}^\star)$ of the unknown true distribution $\mathbb{P}^\star \in \mathscr{W}_2(\mathscr{P}_2(\Xi))$ in W_2 distance as $K \to \infty$. Using this fact, we can also derive the asymptotic consistency of WBDRO in the number of data sources K.

Theorem 4.8 (Asymptotic consistency)**.** *Suppose that all conditions of Theorem 4.6 hold. If $K \in \mathbb{N}^*$ is finite and fixed, we can choose $\beta_n \in (0,1)$ and $\varepsilon_n = \varepsilon_n(\beta_n)$ given in (4.4), $n \in \mathbb{N}^*$, satisfying $\sum_{n=1}^{\infty} \beta_n < \infty$ and $\lim_{n\to\infty} \varepsilon_n(\beta_n) = 0$. If ℓ is upper semicontinuous and there exists $C > 0$ such that $|\ell(\xi)| \leqslant C(1 + \|\xi\|^2)$ for all $\ell \in \mathcal{L}$ and $\xi \in \Xi$, then we have for $\mathbb{P}^\infty$-almost surely, as $n \to \infty$,*

$$\mathcal{R}_{\overline{\mathbb{W}}_{\varepsilon_n(\beta_n),2}(\widehat{\mathbb{P}}^n;\mathbf{1}/K)}(\mathcal{L}) \to \mathcal{R}_{\widehat{\mathsf{b}}^\star_K}(\mathcal{L}).$$

Furthermore, suppose that $\mathbb{P}^\star$ is sub-Gaussian with 2-Wasserstein barycenter $\mathsf{b}^\star := \mathsf{b}_2(\mathbb{P}^\star)$. If we choose $\gamma_K \in (0,1)$ and $\omega_K = \sqrt{c_3 \log(2/\gamma_K)/K}$ with $c_3 > 0$ independent of K satisfying $\sum_{K=1}^{\infty} \gamma_K < \infty$ and $\lim_{K\to\infty} \omega_K(\gamma_K) = 0$, then we have for $\mathbb{P}^\infty$-almost surely, $\mathcal{R}_{\widehat{\mathsf{b}}^\star_K}(\mathcal{L}) \to \mathcal{R}_{\mathsf{b}^\star}(\mathcal{L})$ as $K \to \infty$.

The second part of Theorem 4.8 implies that, if $\widehat{\mathsf{b}}^\star_K$ (or $\widehat{\mathbb{P}}^\star$) is available, the solution of WBDRO converges to the worst-case optimal risk under $\mathsf{b}^\star$ with a suitably chosen γ_K decaying to 0 as $K \to \infty$.

5 Gelbrich Ambiguity Set with 2-Wasserstein Barycenter

In WDRO, more precise structural assumptions about the nominal distribution $\widehat{\mathbb{P}}$ can be made, e.g., it belongs to some family of distributions. In this section, we consider the case of $\widehat{\mathbb{P}} \in \mathscr{P}_2(\mathbb{R}^m)$ with mean $\widehat{\mu} \in \mathbb{R}^m$ and covariance matrix $\widehat{\Sigma} \in \mathbb{S}^m_+$. While computing the Wasserstein distance between any two distributions is NP-hard in general (Taskesen et al., 2021a), an analytical lower bound of their W_2 distance is indeed given by the Gelbrich distance (Gelbrich, 1990), which only involves their first two moments.

Definition 5.1 (Gelbrich distance)**.** For any $\mathbb{P}_1, \mathbb{P}_2 \in \mathscr{P}_2(\mathbb{R}^m)$ with means $\mu_1, \mu_2 \in \mathbb{R}^m$ and covariance matrices $\Sigma_1, \Sigma_2 \in \mathbb{S}^m_{++}$, their *Gelbrich distance* is defined through

$$\mathsf{G}((\mu_1,\Sigma_1),(\mu_2,\Sigma_2)) := \sqrt{\|\mu_1 - \mu_2\|_2^2 + \mathsf{B}^2(\Sigma_1,\Sigma_2)},$$
$$\text{where } \mathsf{B}^2(\Sigma_1,\Sigma_2) := \operatorname{tr}(\Sigma_1) + \operatorname{tr}(\Sigma_2) - 2\operatorname{tr}\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2} \quad (5.1)$$

is the squared *Bures–Wasserstein distance* (Bhatia et al., 2019).

For any $\mathbb{P}_1, \mathbb{P}_2 \in \mathscr{P}_2(\mathbb{R}^m)$, the *Gelbrich bound* (Gelbrich, 1990):

$$\mathsf{W}_2(\mathbb{P}_1,\mathbb{P}_2) \geqslant \mathsf{G}((\mu_1,\Sigma_1),(\mu_2,\Sigma_2))$$

holds, where equality holds if $\mathbb{P}_1, \mathbb{P}_2$ belongs to the same location-scatter family $\mathcal{F}(\mathbb{P}_0)$ for some $\mathbb{P}_0 \in \mathscr{P}_2^{\mathrm{ac}}(\mathbb{R}^m)$, which is defined as follows.

Definition 5.2 (Location-scatter family)**.** Let $X_0 \in \mathbb{R}^m$ be a random vector with $\mathrm{Law}(X_0) = \mathbb{P}_0 \in \mathscr{P}_2^{\mathrm{ac}}(\mathbb{R}^m)$. The set $\mathcal{F}(\mathbb{P}_0) := \{\mathrm{Law}(AX_0 + b) : A \in \mathbb{S}^m_{++}, b \in \mathbb{R}^m\}$ of probability distributions induced by positive definite affine transformations from $\mathbb{P}_0$ is called a *location-scatter family*. Notably, location-scatter families encompass the Gaussian and elliptical distributions (see e.g., Muzellec and Cuturi, 2018).

We then define the *mean-covariance ambiguity set* (Kuhn et al., 2019; Nguyen et al., 2021a) as a ball centered at $(\widehat{\mu}, \widehat{\Sigma})$ with radius $\varepsilon \geqslant 0$ in terms of the Gelbrich distance by $\mathcal{V}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}) := \{(\mu,\Sigma) \in \mathbb{R}^m \times \mathbb{S}^m_+ : \mathsf{G}((\mu,\Sigma),(\widehat{\mu},\widehat{\Sigma})) \leqslant \varepsilon\}$. Next, we define the *Gelbrich ambiguity set* (Kuhn et al., 2019; Nguyen et al., 2021a,b), which is the preimage of $\mathcal{V}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$ under the mean-covariance projection, through

$$\mathcal{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}) := \left\{\mathbb{Q} \in \mathscr{P}_2(\Xi) : (\mathbb{E}_{\mathbb{Q}}[\xi], \mathrm{Var}_{\mathbb{Q}}(\xi)) \in \mathcal{V}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})\right\}. \quad (5.2)$$

Now we consider the setting of Section 4 with $p = 2$, and that all $\mathbb{Q}_k$'s belong to the same location-scatter family. The 2-Wasserstein barycenter $\overline{\mathbb{Q}}_{\boldsymbol{\lambda},2}$ also belongs to the same location-scatter family (Álvarez-Esteban et al., 2018, see Proposition B.7 for details).

Note that we can recover the Gelbrich ambiguity set centered at $\overline{\mathbb{Q}}_{\boldsymbol{\lambda},2}$ if we impose further distributional restrictions on the 2-Wasserstein barycentric ambiguity set.

Proposition 5.3. *Let $\mathbb{P}_0 \in \mathscr{P}_2^{\mathrm{ac}}(\mathbb{R}^m)$ and $\mathbb{Q}_1, \ldots, \mathbb{Q}_K \in \mathcal{F}(\mathbb{P}_0)$ with means $\mu_1, \ldots, \mu_K \in \mathbb{R}^m$ and covariance matrices $\Sigma_1, \ldots, \Sigma_K \in \mathbb{S}_{++}^m$ respectively. For $\boldsymbol{\lambda} \in \triangle^K$, the $\boldsymbol{\lambda}$-weighted 2-Wasserstein barycenter of $\mathbb{Q}_1, \ldots, \mathbb{Q}_K$ is given by $\overline{\mathbb{Q}}_{\boldsymbol{\lambda},2} \in \mathcal{F}(\mathbb{P}_0)$ with mean $\overline{\mu}_{\boldsymbol{\lambda}} \in \mathbb{R}^m$ and covariance matrix $\overline{\Sigma}_{\boldsymbol{\lambda}} \in \mathbb{S}_{++}^m$. If the family of distributions in the 2-Wasserstein barycentric ambiguity set also belongs to $\mathcal{F}(\mathbb{P}_0)$, i.e, $\overline{\mathbb{W}}^{\mathsf{ls}}_{\varepsilon,2}(\mathbb{Q}_1, \ldots, \mathbb{Q}_K; \boldsymbol{\lambda}) \coloneqq \{\mathbb{P} \in \mathcal{F}(\mathbb{P}_0) : \mathsf{W}_2(\mathbb{P}, \overline{\mathbb{Q}}_{\boldsymbol{\lambda},2}) \leqslant \varepsilon\}$, then this ambiguity set equals the Gelbrich ambiguity set (5.2) centered at $(\overline{\mu}_{\boldsymbol{\lambda}}, \overline{\Sigma}_{\boldsymbol{\lambda}})$ restricted to $\mathcal{F}(\mathbb{P}_0)$, i.e.,*

$$\overline{\mathbb{W}}^{\mathsf{ls}}_{\varepsilon,2}(\mathbb{Q}_1, \ldots, \mathbb{Q}_K; \boldsymbol{\lambda}) = \mathcal{G}_\varepsilon(\overline{\mu}_{\boldsymbol{\lambda}}, \overline{\Sigma}_{\boldsymbol{\lambda}}) \cap \mathcal{F}(\mathbb{P}_0).$$

Similar to WDRO (cf. Corollary B.6), we can derive the following (optimal) worst-case risk upper bounds for 2-WBDRO with the *Gelbrich risk* (i.e., risk under the Gelbrich ambiguity set).

Theorem 5.4. *Assume that $\widehat{\mathbb{P}}_k \in \mathscr{P}_2^{\mathrm{ac}}(\Xi)$ has mean $\mu_k \in \mathbb{R}^m$ and covariance matrix $\Sigma_k \in \mathbb{S}_{++}^m$ for each $k \in [\![K]\!]$. Then, we have*

$$(\forall \ell \in \mathcal{L}) \quad \mathcal{R}_{\overline{\mathbb{W}}_{\varepsilon,2}(\widehat{\mathbb{P}}_1,\ldots,\widehat{\mathbb{P}}_K;\boldsymbol{\lambda})}(\ell) \leqslant \mathcal{R}_{\mathcal{G}_\varepsilon(\overline{\mu}_{\boldsymbol{\lambda}},\overline{\Sigma}_{\boldsymbol{\lambda}})}(\ell) \quad \text{and} \quad \mathcal{R}_{\overline{\mathbb{W}}_{\varepsilon,2}(\widehat{\mathbb{P}}_1,\ldots,\widehat{\mathbb{P}}_K;\boldsymbol{\lambda})}(\mathcal{L}) \leqslant \mathcal{R}_{\mathcal{G}_\varepsilon(\overline{\mu}_{\boldsymbol{\lambda}},\overline{\Sigma}_{\boldsymbol{\lambda}})}(\mathcal{L}),$$

where $\overline{\mu}_{\boldsymbol{\lambda}}$ and $\overline{\Sigma}_{\boldsymbol{\lambda}}$ are the mean and covariance matrix of $\mathsf{b}_{\boldsymbol{\lambda},2}(\widehat{\mathbb{P}}_1, \ldots, \widehat{\mathbb{P}}_K)$ respectively.

The above risk bounds indeed reveal a trade-off between tractability and the use of available information. While the worst-case Gelbrich risk minimization problem is more tractable, it uses merely information of the nominal distributions up to their first two moments and discards higher-order moment information.

6 Distributionally Robust Inverse Covariance Matrix Estimation

We demonstrate the proposed WBDRO via an example of sparse inverse covariance (precision) matrix estimation with a Wasserstein barycentric ambiguity set for a Gaussian random vector $\xi \in \mathbb{R}^m$ with covariance matrix $\Sigma \in \mathbb{S}_{++}^m$, where n independent samples are obeserved for each of the K possibly heterogeneous data sources. Estimation of precision matrices is of more interest than that of covariance matrices since it finds various applications to, e.g., mean-variance portfolio optimization and linear discriminant analysis. However, the sample covariance matrix $\widehat{\Sigma}$ is usually rank-deficient when $m > n$ even if Σ has full rank, so naïvely inverting $\widehat{\Sigma}$ to obtain a meaningful precision matrix estimator is not viable.

For simplicity, we assume that the unknown true distribution $\mathbb{P}^\star$ has zero mean. In this case the precision matrix is usually estimated via maximum likelihood estimation (MLE) by minimizing

$$f(X) \coloneqq -\log\det X + \frac{1}{n}\sum_{i=1}^{n} \langle z_i, X z_i \rangle$$

over $\mathbb{S}_{++}^m$ with n independent samples $\{z_i\}_{i=1}^n$. However, this MLE problem is unbounded for $n \leqslant m$. Nguyen et al. (2022) alleviate this issue by incorporating distributional robustness using a 2-Wasserstein ambiguity set centered at the nominal distribution $\mathcal{N}(0, \widehat{\Sigma})$, leading to the *Wasserstein Shrinkage Estimator* (WSE), which can be solved in a quasi-closed form.

With observed data from K data sources $\{z_{i,k}\}_{i \in [\![n]\!], k \in [\![K]\!]}$, it is unclear how to construct a common estimator using aggregate information from them even in the low-dimensional regime using the MLE approach other than a simple weighted average. In view of this, we propose the use of WBDRO to construct a distributionally robust aggregate estimator. We consider the Bures–Wasserstein ambiguity set centered at the $\boldsymbol{\lambda}$-weighted 2-Wasserstein barycenter $\overline{\Sigma}_{\boldsymbol{\lambda}}$ of $\widehat{\mathbb{P}}_1^n, \ldots, \widehat{\mathbb{P}}_K^n$, where $\widehat{\mathbb{P}}_k^n = \mathcal{N}(0, \widehat{\Sigma}_k^n)$ with empirical covariance $\widehat{\Sigma}_k^n = \frac{1}{n}\sum_{i=1}^n z_{i,k} z_{i,k}^\top$ for each $k \in [\![K]\!]$. The *Bures–Wasserstein ambiguity set* is defined by

$$\mathcal{B}_\varepsilon(\widehat{\Sigma}) \coloneqq \{\mathbb{Q} \sim \mathcal{N}(0, \Sigma) : \mathsf{B}(\Sigma, \widehat{\Sigma}) \leqslant \varepsilon\}.$$

Note that $\overline{\Sigma}_{\boldsymbol{\lambda}}$ also coincides the $\boldsymbol{\lambda}$-weighted Bures–Wasserstein barycenter (Kroshnin et al., 2021b) of $\widehat{\Sigma}_1, \dots, \widehat{\Sigma}_K$ (see Appendix B.5). The distributionally robust maximum likelihood estimation (DRMLE) problem can hence be formulated as

$$\underset{X \in \mathbb{S}^m_+}{\text{minimize}} \left\{ -\log\det X + \sup_{\mathbb{P} \in \mathcal{B}_\varepsilon(\overline{\Sigma}_{\boldsymbol{\lambda}})} \mathbb{E}_{\xi \sim \mathbb{P}}[\langle \xi, X\xi \rangle] \right\}. \tag{6.1}$$

Note that the population Wasserstein barycenter of $\mathbb{P}^\star \in \mathscr{W}_2(\mathscr{P}_2(\mathbb{R}^m))$ is also Gaussian since any location-scatter family (which includes Gaussian) is closed for Wasserstein barycenters (Álvarez-Esteban et al., 2018). The problem (6.1) is indeed equivalent to a WDRO problem with a Wasserstein ambiguity set centered at the Wasserstein barycenter $\mathcal{N}(0, \overline{\Sigma}_{\boldsymbol{\lambda}})$, which also admits an analytical solution and is referred to as the *Wasserstein Barycentric Shrinkage Estimator* (WBSE). Further details are given in Appendix D.

Simulations. We compare WBSE with two other estimators constructed from widely used precision matrix estimators for single data source, namely linear shrinkage (LS) and L_1-regularized maximum likelihood estimators (L_1). We choose $m = 20$, $\boldsymbol{\lambda} = \mathbf{1}/K$, $n \in \{50, 100, 200\}$, $K \in \{25, 50, 100\}$, in order to observe the effects of both the sample size n and the number of data sources K on each estimator. We generate K sparse matrices in $\mathbb{S}^m_{++}$ with sparsity level $s = 50\%$ as the true precision matrices Σ_k^{-1}. The true covariance matrix $\Sigma^\star$ is approximated by the Bures–Wasserstein barycenter of another 1000 samples of $\Sigma_{k'}$. Then the true precision matrix is $X^\star = (\Sigma^\star)^{-1}$. We then generate samples $\{z_{i,k}\}_{i=1}^n$ from each of $\mathcal{N}(0, \Sigma_k)$ to construct the empirical covariance matrices $\widehat{\Sigma}_k$, which are used to compute $\overline{\Sigma}_{\boldsymbol{\lambda}}$ and construct the three estimators. See Appendix D for additional details.

We measure performance of estimators using the Stein loss $L(\widehat{X}, \Sigma^\star) := -\log\det(\widehat{X}\Sigma^\star) + \operatorname{tr}(\widehat{X}^\top \Sigma^\star) - m$, which vanishes if $\widehat{X} = (\Sigma^\star)^{-1}$. The losses of the estimators are given in Table 1, averaged over 20 independent trials. We observe that WBSE outperforms the other two estimators by a large margin. The performance of WBSE also improves as n and K increase.

Table 1: Stein losses of LS, L_1 and WBSE.

n	K	LS	L_1	WBSE
50	25	6.77 ± 0.58	7.66 ± 0.63	1.77 ± 0.30
	50	6.81 ± 0.43	7.72 ± 0.46	1.27 ± 0.20
	100	6.91 ± 0.29	7.84 ± 0.31	0.99 ± 0.14
100	25	6.72 ± 0.49	7.61 ± 0.53	1.74 ± 0.27
	50	6.76 ± 0.33	7.67 ± 0.35	1.32 ± 0.19
	100	6.90 ± 0.27	7.83 ± 0.29	1.12 ± 0.16
200	25	6.68 ± 0.56	7.57 ± 0.60	1.76 ± 0.28
	50	6.72 ± 0.36	7.63 ± 0.38	1.34 ± 0.21
	100	6.87 ± 0.29	7.79 ± 0.31	0.62 ± 0.18

Despite being motivated by the high-dimensional setting, the proposed WBS estimator is not feasible when $n < m$ since all $\widehat{\Sigma}_k$'s are singular and their Bures–Wasserstein barycenter $\overline{\Sigma}_{\boldsymbol{\lambda}}$ does not exist, as opposed to the applicability of WSE. A possible remedy is to consider the entropic-regularized variants of the barycenter (Bigot et al., 2019c; Carlier et al., 2021; Janati et al., 2020b; Mallasto et al., 2021; Minh, 2022). We use the Sinkhorn barycenter (see Appendix B) and give additional simulation results under the high-dimensional setting in Appendix D.

7 Concluding Remarks

In this paper, we propose the use of Wasserstein barycenter in the construction of ambiguity sets in WDRO to aggregate data samples from multiple sources. In addition to the performance guarantees established in this paper, extending the statistical analysis (Bartl et al., 2021; Blanchet and Kang, 2021; Blanchet et al., 2021b,c) and generalization bounds (An and Gao, 2021) for WDRO to WBDRO are also important research directions. Motivated by computational tractability and different use cases, alternative barycenters based on other optimal transport distances can also be considered (Bigot et al., 2019c; Bonneel et al., 2015; Carlier et al., 2021; Cazelles et al., 2021; Friesecke et al., 2021; Janati et al., 2020a; Kim and Pass, 2018; Li et al., 2020; Peyré et al., 2016). The same applies to the choice of discrepancy between probability distributions in the ambiguity set in WBDRO (Azizian et al., 2022; Wang et al., 2021b). Finally, it is also interesting to build more general machine learning applications upon the general framework of WBDRO.

References

Anshul Adve and Alpár Mészáros. On nonexpansiveness of metric projection operators on Wasserstein spaces. *arXiv preprint arXiv:2009.01370*, 2020.

Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

Adil Ahidar-Coutrix, Thibaut Le Gouic, and Quentin Paris. Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics. *Probability Theory and Related Fields*, 177 (1):323–368, 2020.

Stéphanie Allassonnière, Yali Amit, and Alain Trouvé. Towards a coherent statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):3–29, 2007.

Stéphanie Allassonnière, Jérémie Bigot, Joan Alexis Glaunès, Florian Maire, and Frédéric J.P. Richard. Statistical models for deformable templates in image and shape analysis. *Annales mathématiques Blaise Pascal*, 20(1):1–35, 2013.

Jason M. Altschuler and Enric Boix-Adserà. Wasserstein barycenters are NP-hard to compute. *SIAM Journal on Mathematics of Data Science*, 4(1):179–203, 2022.

Pedro C. Álvarez-Esteban, E. Del Barrio, J.A. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

Pedro C. Álvarez-Esteban, Eustasio del Barrio, Juan A. Cuesta-Albertos, and Carlos Matrán. Wide consensus aggregation in the Wasserstein space. Application to location-scatter families. *Bernoulli*, 24(4A):3147–3179, 2018.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Basel, 2005.

Luigi Ambrosio, Elia Brué, and Daniele Semola. *Lectures on Optimal Transport*, volume 130 of *UNITEXT*. Springer, 2021.

Yang An and Rui Gao. Generalization bounds for (Wasserstein) robust optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Waïss Azizian, Franck Iutzeler, and Jérôme Malick. Regularization for Wasserstein distributionally robust optimization. *arXiv preprint arXiv:2205.08826*, 2022.

Julio Backhoff-Veraguas, Joaquin Fontbona, Gonzalo Rios, and Felipe Tobar. Bayesian learning with Wasserstein barycenters. *arXiv preprint arXiv:1805.10833*, 2018.

Daniel Bartl, Samuel Drapeau, Jan Obłój, and Johannes Wiesel. Sensitivity analysis of Wasserstein distributionally robust optimization problems. *Proceedings of the Royal Society A*, 477(2256):20210176, 2021.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Espen Bernton, Promit Ghosal, and Marcel Nutz. Entropic optimal transport: Geometry and large deviations. *arXiv preprint arXiv:2102.04397*, 2021.

Dimitris Bertsimas and Bart Van Parys. Bootstrap robust prescriptive analytics. *Mathematical Programming*, 2022.

Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, 2018.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

Jérémie Bigot and Thierry Klein. Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *ESAIM: Probability and Statistics*, 22:35–57, 2018.

Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electronic Journal of Statistics*, 13(2):5120–5150, 2019a.

Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. Data-driven regularization of Wasserstein barycenters with an application to multivariate density registration. *Information and Inference: A Journal of the IMA*, 8(4):719–755, 2019b.

Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. Penalization of barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 51(3):2261–2285, 2019c.

Adrian N. Bishop. Information fusion via the Wasserstein barycenter in the space of probability measures: Direct fusion of empirical measures and Gaussian fusion with unknown correlation. In *17th International Conference on Information Fusion (FUSION)*. IEEE, 2014.

Adrian N. Bishop and Arnaud Doucet. Network consensus in the Wasserstein metric space of probability measures. *SIAM Journal on Control and Optimization*, 59(5):3261–3277, 2021.

Jose Blanchet and Yang Kang. Sample out-of-sample inference based on Wasserstein distance. *Operations Research*, 69(3):985–1013, 2021.

Jose Blanchet and Nian Si. Optimal uncertainty size in distributionally robust inverse covariance estimation. *Operations Research Letters*, 47(6):618–621, 2019.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019a.

Jose Blanchet, Yang Kang, Karthyek Murthy, and Fan Zhang. Data-driven optimal transport cost selection for distributionally robust optimization. In *Proceedings of the Winter Simulation Conference (WSC)*. IEEE, 2019b.

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

Jose Blanchet, Karthyek Murthy, and Viet Anh Nguyen. Statistical analysis of Wasserstein distributionally robust estimators. In *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*, pages 227–254. INFORMS, 2021b.

Jose Blanchet, Karthyek Murthy, and Nian Si. Confidence regions in Wasserstein distributionally robust estimation. *Biometrika*, 2021c.

Jose Blanchet, Karthyek Murthy, and Fan Zhang. Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes. *Mathematics of Operations Research*, 2021d.

Emmanuel Boissard, Thibaut Le Gouic, and Jean-Michel Loubes. Distribution's template estimate with Wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.

François Bolley, Arnaud Guillin, and Cédric Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3-4):541–593, 2007.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Guillaume Carlier, Adam Oberman, and Edouard Oudet. Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1621–1642, 2015.

Guillaume Carlier, Katharina Eichinger, and Alexey Kroshnin. Entropic-Wasserstein barycenters: PDE characterization, regularity, and CLT. *SIAM Journal on Mathematical Analysis*, 53(5):5880–5914, 2021.

Yair Carmon and Danielle Hausler. Distributionally robust optimization via ball oracle acceleration. *arXiv preprint arXiv:2203.13225*, 2022.

Elsa Cazelles, Felipe Tobar, and Joaquin Fontbona. A novel notion of barycenter for probability distributions based on optimal weak mass transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J. Stromme. Gradient descent algorithms for Bures-Wasserstein barycenters. In *Proceedings of the Conference on Learning Theory (COLT)*, 2020.

Lénaïc Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster Wasserstein distance estimation with the Sinkhorn divergence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Kai Lai Chung. *A Course in Probability Theory.* Academic Press, 3rd edition, 2001.

Corinna Cortes, Mehryar Mohri, Dmitry Storcheus, and Ananda Theertha Suresh. Boosting with multiple sources. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

Marco Cuturi and Gabriel Peyré. Semidual regularized optimal transport. *SIAM Review*, 60(4):941–965, 2018.

Chiheb Daaloul, Thibaut Le Gouic, Jacques Liandrat, and Magali Tournus. Sampling from the Wasserstein barycenter. *arXiv preprint arXiv:2105.01706*, 2021.

Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

John C. Duchi, Peter W. Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.

Pavel Dvurechenskii, Darina Dvinskikh, Alexander Gasnikov, Cesar Uribe, and Angelia Nedich. Decentralize and randomize: Faster algorithm for Wasserstein barycenters. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171 (1-2):115–166, 2018.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Alessio Figalli and Federico Glaudo. *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows*. EMS Textbooks in Mathematics. EMS Press, Zürich, 2021.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

Gero Friesecke, Daniel Matthes, and Bernhard Schmitzer. Barycenters for the Hellinger–Kantorovich distance over $\mathbb{R}^d$. *SIAM Journal on Mathematical Analysis*, 53(1):62–110, 2021.

Rui Gao. Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *arXiv preprint arXiv:2009.04382*, 2020.

Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

Rui Gao, Xi Chen, and Anton J. Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *arXiv preprint arXiv:1712.06050*, 2017.

Matthias Gelbrich. On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Promit Ghosal, Marcel Nutz, and Espen Bernton. Stability of entropic optimal transport and Schrödinger bridges. *arXiv preprint arXiv:2106.03670*, 2021.

Joel Goh and Melvyn Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58(4):902–917, 2010.

Ziv Goldfeld, Kengo Kato, Gabriel Rioux, and Ritwik Sadhu. Statistical inference with regularized optimal transport. *arXiv preprint arXiv:2205.04283*, 2022.

Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Amin Karbasi. Learning distributionally robust models at scale via composite optimization. In *International Conference on Learning Representations (ICLR)*, 2022.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825): 357–362, 2020.

Florian Heinemann, Axel Munk, and Yoav Zemel. Randomized Wasserstein barycenter computation: Resampling with statistical guarantees. *SIAM Journal on Mathematics of Data Science*, 4(1):229–259, 2022.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Debiased Sinkhorn barycenters. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020a.

Hicham Janati, Boris Muzellec, Gabriel Peyré, and Marco Cuturi. Entropic optimal transport between unbalanced Gaussian measures has a closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020b.

Jikai Jin, Bohang Zhang, Haiyang Wang, and Liwei Wang. Non-convex distributionally robust optimization: Non-asymptotic analysis. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Young-Heon Kim and Brendan Pass. A canonical barycenter via Wasserstein regularization. *SIAM Journal on Mathematical Analysis*, 50(2):1817–1828, 2018.

Martin Knott and Cyril S. Smith. On a generalization of cyclic monotonicity and distances among random vectors. *Linear Algebra and Its Applications*, 199:363–371, 1994.

Alexey Kroshnin. Fréchet barycenters in the Monge-Kantorovich spaces. *Journal of Convex Analysis*, 25(4): 1371–1395, 2018.

Alexey Kroshnin, Vladimir Spokoiny, and Alexandra Suvorikova. Multiplier bootstrap for Bures-Wasserstein barycenters. *arXiv preprint arXiv:2111.12612*, 2021a.

Alexey Kroshnin, Vladimir Spokoiny, and Alexandra Suvorikova. Statistical inference for Bures–Wasserstein barycenters. *The Annals of Applied Probability*, 31(3):1264–1298, 2021b.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.

Long Tan Le, Josh Nguyen, Canh T. Dinh, and Nguyen Hoang Tran. On the generalization of Wasserstein robust federated learning, 2022. URL https://openreview.net/forum?id=nWprF5r2spe.

Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, 168(3):901–917, 2017.

Thibaut Le Gouic, Quentin Paris, Philippe Rigollet, and Austin J. Stromme. Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space. *arXiv preprint arXiv:1908.00828v4*, 2021.

Daniel Levy, Yair Carmon, John C. Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Lingxiao Li, Aude Genevay, Mikhail Yurochkin, and Justin M. Solomon. Continuous regularized Wasserstein barycenters. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Mengmeng Li, Tobias Sutter, and Daniel Kuhn. Distributionally robust optimization with Markovian data. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Giulia Luise, Saverio Salzo, Massimiliano Pontil, and Carlo Ciliberto. Sinkhorn barycenters with free support via Frank-Wolfe algorithm. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Anton Mallasto, Augusto Gerolin, and Hà Quang Minh. Entropy-regularized 2-Wasserstein distance between Gaussian measures. *Information Geometry*, 2021.

Yishay Mansour, Mehryar Mohri, Jae Ro, Ananda Theertha Suresh, and Ke Wu. A theory of multiple-source adaptation with limited target labeled data. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Hà Quang Minh. Entropic regularization of Wasserstein distance between infinite-dimensional Gaussian measures and Gaussian processes. *Journal of Theoretical Probability*, pages 1–96, 2022.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

Boris Muzellec and Marco Cuturi. Generalizing point embeddings using the Wasserstein space of elliptical distributions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

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

Viet Anh Nguyen, Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization. *Mathematics of Operations Research*, 2021b.

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

Viet Anh Nguyen, Daniel Kuhn, and Peyman Mohajerin Esfahani. Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *Operations Research*, 70(1):490–515, 2022.

Marcel Nutz. *Introduction to Entropic Optimal Transport*. 2021. URL https://www.math.columbia.edu/~mnutz/docs/EOT_lecture_notes.pdf. Lecture Notes, Columbia University.

Marcel Nutz and Johannes Wiesel. Entropic optimal transport: Convergence of potentials. *Probability Theory and Related Fields*, pages 1–24, 2021.

Jan Obłój and Johannes Wiesel. Distributionally robust portfolio maximization and marginal utility pricing in one period financial markets. *Mathematical Finance*, 31(4):1454–1493, 2021.

Victor M. Panaretos and Yoav Zemel. Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6:405–431, 2019.

Victor M. Panaretos and Yoav Zemel. *An Invitation to Statistics in Wasserstein Space*. SpringerBriefs in Probability and Mathematical Statistics. Springer Nature, 2020.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

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

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

Amirhossein Reisizadeh, Farzan Farnia, Ramtin Pedarsani, and Ali Jadbabaie. Robust federated learning: The case of affine distribution shifts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jae Ro, Mingqing Chen, Rajiv Mathews, Mehryar Mohri, and Ananda Theertha Suresh. Communication-efficient agnostic federated averaging. *arXiv preprint arXiv:2104.02748*, 2021.

Ludger Rüschendorf and Ludger Uckelmann. On the n-coupling problem. *Journal of Multivariate Analysis*, 81(2):242–258, 2002.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2019.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser, 2015.

Morgan A. Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngole, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

Christof Schötz. Convergence rates for the generalized Fréchet mean via the quadruple inequality. *Electronic Journal of Statistics*, 13(2):4280–4345, 2019.

Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Soroosh Shafieezadeh-Abadeh, Viet Anh Nguyen, Daniel Kuhn, and Peyman Mohajerin Esfahani. Wasserstein distributionally robust Kalman filtering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

Agnieszka Słowik and Léon Bottou. On distributionally robust optimization and data rebalancing. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Sanvesh Srivastava, Cheng Li, and David B. Dunson. Scalable Bayes via barycenter in Wasserstein space. *Journal of Machine Learning Research*, 19(1):312–346, 2018.

Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Bahar Taskesen, Soroosh Shafieezadeh-Abadeh, and Daniel Kuhn. Semi-discrete optimal transport: Hardness, regularization and numerical solution. *arXiv preprint arXiv:2103.06263*, 2021a.

Bahar Taskesen, Man-Chung Yue, Jose Blanchet, Daniel Kuhn, and Viet Anh Nguyen. Sequential domain adaptation by synthesizing distributionally robust experts. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021b.

Cédric Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2009.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H. Brendan McMahan, Blaise Aguera y Arcas, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, Suhas Diggavi, Hubert Eichner, Advait Gadhikar, Zachary Garrett, Antonious M. Girgis, Filip Hanzely, Andrew Hard, Chaoyang He, Samuel Horvath, Zhouyuan Huo, Alex Ingerman, Martin Jaggi, Tara Javidi, Peter Kairouz, Satyen Kale, Sai Praneeth Karimireddy, Jakub Konecny, Sanmi Koyejo, Tian Li, Luyang Liu, Mehryar Mohri, Hang Qi, Sashank J. Reddi, Peter Richtarik, Karan Singhal, Virginia Smith, Mahdi Soltanolkotabi, Weikang Song, Ananda Theertha Suresh, Sebastian U. Stich, Ameet Talwalkar, Hongyi Wang, Blake Woodworth, Shanshan Wu, Felix X. Yu, Honglin Yuan, Manzil Zaheer, Mi Zhang, Tong Zhang, Chunxiang Zheng, Chen Zhu, and Wennan Zhu. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021a.

Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *arXiv preprint arXiv:2109.11926*, 2021b.

Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

Hongkang Yang and Esteban G. Tabak. Clustering, factor discovery and optimal transport. *Information and Inference: A Journal of the IMA*, 10(4):1353–1387, 2021.

Yaodong Yu, Tianyi Lin, Eric Mazumdar, and Michael I. Jordan. Fast distributionally robust learning with variance reduced min-max optimization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Honglin Yuan, Warren Richard Morningstar, Lin Ning, and Karan Singhal. What do we mean by generalization in federated learning? In *International Conference on Learning Representations (ICLR)*, 2022.

Yoav Zemel and Victor M. Panaretos. Fréchet means and procrustes analysis in Wasserstein space. *Bernoulli*, 25(2):932–976, 2019.

Ningshan Zhang, Mehryar Mohri, and Judy Hoffman. Multiple-source adaptation theory and algorithms. *Annals of Mathematics and Artificial Intelligence*, 89(3):237–270, 2021.

Jianzhe Zhen, Daniel Kuhn, and Wolfram Wiesemann. Mathematical foundations of robust and distributionally robust optimization. *arXiv preprint arXiv:2105.00760*, 2021.

Appendix

A Other Related Work

In this section, we provide a more detailed discussion on the connections of the proposed framework to other existing machine learning paradigms.

We now introduce some additional notation. In the following, we consider a supervised learning setting with the input space $\mathcal{X}$ and the output space $\mathcal{Y}$. With n samples $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$, the most notable framework for building machine learning models is the ERM formulation, which solves

$$\underset{\theta\in\Theta}{\text{minimize}} \ \mathbb{E}_{(x,y)\sim\widehat{\mathbb{P}}}[\ell(h_\theta(x), y)] = \frac{1}{n}\sum_{i=1}^n \ell(h_\theta(x_i), y_i),$$

where $h_\theta\colon \mathcal{X} \to \mathcal{Y}$ represents the model with parameter $\theta \in \Theta \subseteq \mathbb{R}^d$, $\ell(\cdot,\cdot)$ is the loss function, and $\widehat{\mathbb{P}} \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$ is the empirical distribution of $\mathcal{D}_n$.

Federated Learning. Following the discussion in Section 3 of the main text, federated learning (FL) and its optimization formulation is a crucial motivation of the problem considered in this paper. The agnostic federated learning (AFL) framework (Mohri et al., 2019; Ro et al., 2021) considers the worst-case setting in which the learner seeks a solution that is favorable for any $\boldsymbol{\lambda} \in \Lambda \subseteq \triangle^K$, where Λ is a closed convex set. Again, if we define the $\boldsymbol{\lambda}$-mixture of distributions $\mathbb{P}_{\boldsymbol{\lambda}} \coloneqq \sum_{k=1}^K \lambda_k \mathbb{P}_k$, then the *agnostic risk* is given by

$$\mathcal{L}_{\mathbb{P}_\Lambda}(\theta) \coloneqq \sup_{\boldsymbol{\lambda}\in\Lambda} \ \mathbb{E}_{(x,y)\sim\mathbb{P}_{\boldsymbol{\lambda}}}[\ell(h_\theta(x), y)].$$

In practice, only the empirical distributions $\widehat{\mathbb{P}}_k$'s are accessible (constructed from finite samples), so we can define the $\boldsymbol{\lambda}$-mixture of empirical distributions $\overline{\mathbb{P}}_{\boldsymbol{\lambda}} \coloneqq \sum_{k=1}^K \lambda_k \widehat{\mathbb{P}}_k$. Then, the *agnostic empirical risk* is given by

$$\mathcal{L}_{\overline{\mathbb{P}}_\Lambda}(\theta) \coloneqq \sup_{\boldsymbol{\lambda}\in\Lambda} \ \mathbb{E}_{(x,y)\sim\overline{\mathbb{P}}_{\boldsymbol{\lambda}}}[\ell(h_\theta(x), y)]. \tag{A.1}$$

Two notable differences between AFL and our proposed framework are that in AFL the choice of $\boldsymbol{\lambda}$ is also optimized, and the use of mixture distributions instead of Wasserstein barycenters. Following this line of work, Deng et al. (2020) develop communication-efficient distributed algorithms for minimizing the agnostic empirical risk (A.1), whereas Reisizadeh et al. (2020) study the notion of robustness against affine distribution drifts in clients' data in federated learning.

On the other hand, Wasserstein distributionally robust federated learning (WAFL; Le et al., 2022) shares a very similar spirit to our work which considers a WDRO formulation under the federated learning setting, but again with the mixture distribution (Euclidean barycenter) usually considered in FL instead of the notion of (entropic-regularized) Wasserstein barycenters used in this work.

Remark A.1. In this work, we however do not aim at solving the FL problem, which requires distributed and decentralized computations. Yet, it is very interesting to make our proposed paradigm amenable to the full federated learning setting, which might require decentralized distributed computation of Wasserstein barycenters (Dvurechenskii et al., 2018).

(Multiple-Source) Domain Adaptation. In the multiple-source domain adaptation problem (see e.g., Mansour et al., 2021; Zhang et al., 2021, and references therein), each domain is defined by the corresponding distribution $\mathbb{P}_k$. The target distribution could be assumed to be close to some convex combination of source distributions, $\sum_{k=1}^K \lambda_k \mathbb{P}_k$. The learner wants to learn a model on the target domain. Similar to the AFL formulation, the learner can do this by solving

$$\underset{\theta\in\Theta}{\text{minimize}} \ \mathbb{E}_{(x,y)\sim\overline{\mathbb{P}}_{\boldsymbol{\lambda}^\star}}[\ell(h_\theta(x), y)],$$

where $\boldsymbol{\lambda}^\star \coloneqq \operatorname{argmin}_{\boldsymbol{\lambda}\in\triangle^K} \mathcal{E}(\boldsymbol{\lambda})$ for some discrepancy measure $\mathcal{E}$ between the empirical target distribution $\widehat{\mathbb{P}}_0$ and $\overline{\mathbb{P}}_{\boldsymbol{\lambda}}$, e.g., a Bregman divergence $B(\widehat{\mathbb{P}}_0 \,\|\, \overline{\mathbb{P}}_{\boldsymbol{\lambda}})$. Another related work Taskesen et al. (2021b) study a distributionally robust formulation for supervised domain adaptation with scarce labeled target data.

Boosting. Also inspired by the agnostic loss in the AFL framework (Mohri et al., 2019), Cortes et al. (2021) study boosting in the presence of multiple source domains. They put forward the so-called Q-ensembles, which are convex combinations weighted by a domain classifier Q. The typical assumption that the target distribution is a mixture of the source distributions is also made. They also provide an algorithmic extension to the federated learning scenario. Further related work can be found in Cortes et al. (2021).

Group DRO. Machine learning models might rely on *spurious correlations*—misleading heuristics which hold for most training examples but are wrongly linked to the target. Thus, these models could suffer high risk on minority groups where these correlations do not hold. The Group DRO framework (Hu et al., 2018; Sagawa et al., 2019), which aims to obtain high performance across all groups, minimizes the *worst-group risk*:

$$\operatorname*{minimize}_{\theta\in\Theta} \sup_{\mathbb{P}\in\mathcal{Q}(\mathbb{P}_1,\ldots,\mathbb{P}_K;\boldsymbol{\lambda})} \mathbb{E}_{\xi\sim\mathbb{P}}[\ell(h_\theta(x),y)], \tag{A.2}$$

where the ambiguity set is defined as $\mathcal{Q}(\mathbb{P}_1,\ldots,\mathbb{P}_K;\boldsymbol{\lambda}) \coloneqq \left\{\sum_{k=1}^K \lambda_k\mathbb{P}_k : \boldsymbol{\lambda}\in\triangle^K\right\}$.

Note that (A.2) is equivalent to

$$\operatorname*{minimize}_{\theta\in\Theta} \sup_{k\in[\![K]\!]} \mathbb{E}_{\xi\sim\mathbb{P}_k}[\ell(h_\theta(x),y)].$$

Since in practice we only observe the empirical distributions $\widehat{\mathbb{P}}_k$'s, we instead minimize the *empirical worst-group risk*:

$$\operatorname*{minimize}_{\theta\in\Theta} \sup_{k\in[\![K]\!]} \mathbb{E}_{\xi\sim\widehat{\mathbb{P}}_k}[\ell(h_\theta(x),y)],$$

which can be rewritten as

$$\operatorname*{minimize}_{\theta\in\Theta} \sup_{\boldsymbol{\lambda}\in\triangle^K} \left\{\sum_{k=1}^K \lambda_k\mathbb{E}_{\xi\sim\widehat{\mathbb{P}}_k}[\ell(h_\theta(x),y)] = \mathbb{E}_{\xi\sim\overline{\mathbb{P}}_{\boldsymbol{\lambda}}}[\ell(h_\theta(x),y)]\right\}.$$

This has an almost identical formulation to the AFL framework (A.1) above (when $\Lambda = \triangle^K$). A recent work Carmon and Hausler (2022) study an accelerated optimization method for Group DRO.

As a work close to Group DRO, Słowik and Bottou (2022) study the relation between solving a DRO problem and optimizing the expected error for a single distribution constructed by the mixture distribution $\sum_{k=1}^K \lambda_k\mathbb{P}_k$ for some $\boldsymbol{\lambda}\in\triangle^K$, particularly with nonconvex loss functions.

B Additional Technical Details and Results

B.1 Additional Notation and Definitions

For any two matrices $A \coloneqq (a_{i,j})_{i\in[\![d_1]\!],j\in[\![d_2]\!]}$ and $B \coloneqq (b_{i,j})_{i\in[\![d_1]\!],j\in[\![d_2]\!]}$ in $\mathbb{R}^{d_1\times d_2}$, we denote the Frobenius inner product of A and B by $\langle\!\langle A, B\rangle\!\rangle_{\mathrm{F}} \coloneqq \operatorname{tr}(A^\top B) = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} a_{i,j}b_{i,j}$ and the Frobenius norm of A by $|\!|\!|A|\!|\!|_{\mathrm{F}} \coloneqq \sqrt{\langle\!\langle A, A\rangle\!\rangle_{\mathrm{F}}}$. The elementwise L_1-norm of A is denoted by $\|A\|_1 \coloneqq \sum_{i=1}^{d_1}\sum_{j=1}^{d_2}|a_{i,j}|$. The Lebesgue measure over $\mathcal{X}$ is denoted by $\mathcal{L}_{\mathcal{X}}$. For two probability measures μ and ν on $\mathcal{B}(\mathcal{X})$, the relative entropy or the Kullback–Leibler (KL) divergence from μ to ν is $\mathsf{D}_{\mathrm{KL}}(\mu\,\|\,\nu) \coloneqq \int_{\mathcal{X}} \log(\mathrm{d}\mu/\mathrm{d}\nu)\,\mathrm{d}\mu$ if μ is absolutely continuous w.r.t. ν (denoted by $\mu \ll \nu$) with the Radon–Nikodym derivative $\mathrm{d}\mu/\mathrm{d}\nu$ and $+\infty$ otherwise. The product measure $\mu\otimes\nu \in \mathscr{P}(\mathbb{R}^d\times\mathbb{R}^d)$ of μ and ν is characterized by $(\mu\otimes\nu)(\mathcal{X}\times\mathcal{Y}) = \mu(\mathcal{X})\nu(\mathcal{Y})$ for any pair of Borel sets $\mathcal{X},\mathcal{Y}\subset\mathbb{R}^d$.

Definition B.1 (Sub-Gaussian measure)**.** Let $\Xi \subseteq \mathbb{R}^m$ be a closed convex set. A probability measure $\mathbb{P}\in\mathscr{W}_p(\mathscr{P}_p(\Xi))$ is sub-Gaussian with variance proxy $\sigma^2 > 0$ if

$$\mathbb{E}_{\rho\sim\mathbb{P}}\left[\exp\left\{\frac{1}{2\sigma^2}\mathsf{W}_p^2(\mathsf{b}_p(\mathbb{P}),\rho)\right\}\right] \leqslant 2,$$

where $\rho\in\mathscr{P}_p(\Xi)$ is a random measure with distribution $\mathbb{P}$.

B.2 Entropic Optimal Transport

Based on computational consideration, entropic regularization has been introduced to approximate Wasserstein distances. The so-called *entropic(-regularized) optimal transport* has aroused much theoretical and computational interests across the fields of machine learning, statistics, economics, image processing, and theoretical and applied probability. We refer to Bernton et al. (2021); Bigot et al. (2019a); Ghosal et al. (2021); Goldfeld et al. (2022); Nutz (2021); Nutz and Wiesel (2021) for recent theoretical advances in probability and statistics.

In the machine learning community, Cuturi (2013) proposes the so-called *Sinkhorn distance* (which is indeed not a metric), which is referred to as the *entropic-p-Wasserstein distance* in this paper and is the central object in entropic optimal transport. Let us recall that $\Omega \subseteq \mathbb{R}^m$ is a closed convex set.

Definition B.2 (Entropic-Wasserstein distance)**.** For $\sigma > 0$, the *entropic-p-Wasserstein distance* between $\rho, \nu \in \mathscr{P}(\Omega)$ is defined by

$$\mathsf{OT}_{p,\sigma}^{\omega_1,\omega_2}(\rho,\nu) \coloneqq \inf_{\pi\in\Pi(\rho,\nu)} \left\{ \int_{\Omega\times\Omega} \|x-y\|^p \,\mathrm{d}\pi(x,y) + \sigma \mathsf{D}_{\mathrm{KL}}(\pi \,\|\, \omega_1 \otimes \omega_2) \right\}, \tag{B.1}$$

where ω_1 and ω_2 are two reference measures such that $\rho \ll \omega_1$ and $\nu \ll \omega_2$. Note that the entropic-p-Wasserstein distance equals the p-Wasserstein distance when $\sigma \to 0$.

The choice of the reference measures in (B.1) is known to induce different types of entropy bias (Janati et al., 2020a), since in general $\mathsf{OT}_{p,\sigma}^{\omega_1,\omega_2}(\rho,\rho) \neq 0$ due to the regularization term. For example, the Lebesgue measure ($\omega_1 = \omega_2 = \mathcal{L}_{\mathbb{R}^m}$) induces a blurring bias, whereas simply taking the product measure with $\omega_1 = \rho$ and $\omega_2 = \nu$ induces a shrinking bias. To circumvent the entropy bias, the *p-Sinkhorn divergence* (Chizat et al., 2020; Feydy et al., 2019; Genevay et al., 2018; Luise et al., 2019; Ramdas et al., 2017) can be defined without specifying any reference measures (Feydy et al., 2019):

$$\mathsf{S}_{p,\sigma}(\rho,\nu) \coloneqq \mathsf{OT}_{p,\sigma}(\rho,\nu) - \frac{\mathsf{OT}_{p,\sigma}(\rho,\rho) + \mathsf{OT}_{p,\sigma}(\nu,\nu)}{2}. \tag{B.2}$$

The Sinkhorn divergence can be viewed as an interpolation between the (unregularized) Wasserstein distance (when $\sigma \to 0$) and *maximum mean discrepancy* (MMD; when $\sigma \to \infty$) (Feydy et al., 2019; Ramdas et al., 2017). Note that there have also been lots of recent results regarding the computational efficiency guarantees of entropic optimal transport, see e.g., Chizat et al. (2020); Genevay et al. (2019).

To avoid confusion, we reserve *Sinkhorn* to solely refer to notions defined via the Sinkhorn divergence (B.2), whereas *entropic-Wasserstein* to solely refer to notions defined via the entropic-Wasserstein distance (B.1).

The Gaussian Case. Similar to the (unregularized) Wasserstein distance, entropic-Wasserstein distances and Sinkhorn divergences usually do not admit closed forms, with the notable exception for the one between two multivariate Gaussians. The following closed form expressions of entropic-2-Wasserstein distance and Sinkhorn divergence between two multivariate Gaussians are directly stated from Janati et al. (2020b); Mallasto et al. (2021); Minh (2022) without proofs.

Proposition B.3. *The entropic-2-Wasserstein distance between two Gaussians $\rho_k = \mathcal{N}(\mu_k, \Sigma_k)$ with $\mu_k \in \mathbb{R}^m$ and $\Sigma_k \in \mathbb{S}_+^m$ for $k \in [\![2]\!]$ is given by*

$$\mathsf{OT}_{2,\gamma}^{\otimes}(\rho_1,\rho_2) = \|\mu_1-\mu_2\|_2^2 + \operatorname{tr}\bigl(\Sigma_1+\Sigma_2 - 2D_\gamma^{\Sigma_1,\Sigma_2}\bigr) + \frac{\gamma}{2}\bigl[m(1-\log\gamma) + \log\det(2D_\gamma^{\Sigma_1,\Sigma_2} + \gamma I/2)\bigr],$$

where $D_\gamma^{\Sigma_1,\Sigma_2} \coloneqq (\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2} + \gamma^2 I/16)^{1/2}$. Then using (B.2), the Sinkhorn divergence between two Gaussians $\rho_k = \mathcal{N}(\mu_k,\Sigma_k)$ for $k \in [\![2]\!]$ is given by

$$\begin{aligned}\mathsf{S}_{2,\gamma}(\rho_1,\rho_2) = \|\mu_1-\mu_2\|_2^2 + \operatorname{tr}\bigl(D_\gamma^{\Sigma_1,\Sigma_1} + D_\gamma^{\Sigma_2,\Sigma_2} - 2D_\gamma^{\Sigma_1,\Sigma_2}\bigr) + \frac{\gamma}{2}\log\det(2D_\gamma^{\Sigma_1,\Sigma_2} + \gamma I/2)\\ - \frac{\gamma}{4}\bigl[\log\det(2D_\gamma^{\Sigma_1,\Sigma_1} + \gamma I/2) + \log\det(2D_\gamma^{\Sigma_2,\Sigma_2} + \gamma I/2)\bigr].\end{aligned}$$

Note that, unlike the (squared) 2-Wasserstein distance between two Gaussians (i.e., Gelbrich distance), the entropic-2-Wasserstein distance and the Sinkhorn divergence are both defined for degenerate Gaussians, i.e., when (any of or both) Σ_1 and Σ_2 are singular.

B.2.1 Entropic-Wasserstein and Sinkhorn Barycenters

Similar to Wasserstein distances, Wasserstein barycenters are also NP-hard to compute in general (Altschuler and Boix-Adserà, 2022). A potential remedy is to introduce entropic regularization, which we refer to as the entropic-Wasserstein barycenters (Bigot et al., 2019b,c; Carlier et al., 2021; Cuturi and Doucet, 2014; Cuturi and Peyré, 2018; Janati et al., 2020a; Kim and Pass, 2018). Recent theoretical results on entropic-Wasserstein barycenters such as their existence and uniqueness can be found in Carlier et al. (2021). A variant of the entropic-Wasserstein barycenter is the Sinkhorn barycenter (Janati et al., 2020a; Luise et al., 2019), defined via the Sinkhorn divergence (B.2), in order to debias the entropic-Wasserstein barycenter due to the entropic regularization. We give the definitions of both the empirical entropic-Wasserstein and Sinkhorn barycenters below.

Definition B.4 (Empirical entropic-Wasserstein and Sinkhorn barycenters)**.** For $p \in [1, +\infty)$ and $\boldsymbol{\lambda} = (\lambda_k)_{k\in[\![K]\!]} \in \triangle^K$, the *$\boldsymbol{\lambda}$-weighted entropic-$p$-Wasserstein barycenter* is

$$\widehat{\mathsf{b}}^{\mathsf{OT},\sigma}_{\boldsymbol{\lambda},p}(\rho_1, \ldots, \rho_K) \coloneqq \operatorname*{argmin}_{\nu\in\mathscr{P}(\mathbb{R}^m)} \sum_{k=1}^{K} \lambda_k \mathsf{OT}^{\otimes}_{p,\sigma}(\nu, \rho_k),$$

where $\mathsf{OT}^{\otimes}_{p,\sigma}(\rho,\nu) \equiv \mathsf{OT}^{\rho,\nu}_{p,\sigma}(\rho,\nu)$. Likewise, for $p \in [1, +\infty)$ and $\boldsymbol{\lambda} = (\lambda_k)_{k\in[\![K]\!]} \in \triangle^K$, the *$\boldsymbol{\lambda}$-weighted p-Sinkhorn barycenter* is

$$\widehat{\mathsf{b}}^{\mathsf{S},\sigma}_{\boldsymbol{\lambda},p}(\rho_1, \ldots, \rho_K) \coloneqq \operatorname*{argmin}_{\nu\in\mathscr{P}(\mathbb{R}^m)} \sum_{k=1}^{K} \lambda_k \mathsf{S}_{p,\sigma}(\nu, \rho_k).$$

The Gaussian Case. Unlike the unregularized case, the entropic-2-Wasserstein barycenter and Sinkhorn barycenter of Gaussians are no longer guaranteed to be Gaussian, so we have to restrict them to the manifold of Gaussians. Then, under this assumption, similar to the unregularized case, both the entropic-2-Wasserstein barycenter and the Sinkhorn barycenter can be computed by solving fixed-point equations (Janati et al., 2020b; Mallasto et al., 2021; Minh, 2022). For completeness, we state (without proof) the most general results from Minh (2022, Theorems 11–12) below, and refer the readers to Minh (2022) for details.

Proposition B.5. *Let $\rho_1, \ldots, \rho_K$ be K possibly degenerate Gaussian distributions $\rho_k = \mathcal{N}(\mu_k, \Sigma_k)$ with $\mu_k \in \mathbb{R}^m$ and $\Sigma_k \in \mathbb{S}^m_+$ for $k \in [\![K]\!]$. Their entropic-2-Wasserstein barycenter restricted to the manifold of Gaussians is $\widehat{\mathsf{b}}^{\mathsf{OT},\sigma}_{\boldsymbol{\lambda},p}(\rho_1, \ldots, \rho_K) = \mathcal{N}(\overline{\mu}_{\boldsymbol{\lambda}}, \overline{\Sigma}_{\boldsymbol{\lambda},\sigma})$, where $\overline{\mu}_{\boldsymbol{\lambda}} = \sum_{k=1}^K \lambda_k\mu_k$ and $\overline{\Sigma}_{\boldsymbol{\lambda},\sigma}$ satisfies the equation*

$$\overline{\Sigma}_{\boldsymbol{\lambda},\sigma} = \frac{\sigma}{4}\sum_{k=1}^{K} \lambda_k \left(-I + \left(I + \frac{16}{\sigma^2}\overline{\Sigma}^{1/2}_{\boldsymbol{\lambda},\sigma}\Sigma_k\overline{\Sigma}^{1/2}_{\boldsymbol{\lambda},\sigma}\right)^{1/2}\right).$$

Furthermore, their Sinkhorn barycenter restricted to the manifold of Gaussians is also $\widehat{\mathsf{b}}^{\mathsf{S},\sigma}_{\boldsymbol{\lambda},p}(\rho_1, \ldots, \rho_K) = \mathcal{N}(\overline{\mu}_{\boldsymbol{\lambda}}, \widetilde{\Sigma}_{\boldsymbol{\lambda},\sigma}) \in \mathbb{S}^m_+$, where $\widetilde{\Sigma}_{\boldsymbol{\lambda},\sigma}$ is the unique solution of the following equation

$$\widetilde{\Sigma}_{\boldsymbol{\lambda},\sigma} = \varphi_\sigma(\widetilde{\Sigma}_{\boldsymbol{\lambda},\sigma})\sum_{k=1}^{K} \lambda_k \left[\Sigma_k^{1/2}\left(I + \left(I + \frac{16}{\gamma^2}\Sigma_k^{1/2}\widetilde{\Sigma}_{\boldsymbol{\lambda},\sigma}\Sigma_k^{1/2}\right)^{1/2}\right)^{-1}\Sigma_k^{1/2}\right]\varphi_\sigma(\widetilde{\Sigma}_{\boldsymbol{\lambda},\sigma}),$$

where $\varphi_\sigma(M) \coloneqq \left(I + \left(I + 16M^2/\sigma^2\right)^{1/2}\right)^{1/2}$. Furthermore, if $\widetilde{\Sigma}_{\boldsymbol{\lambda},\sigma} \in \mathbb{S}^m_{++}$, then $\widetilde{\Sigma}_{\boldsymbol{\lambda},\sigma}$ is the unique solution of the equation

$$\widetilde{\Sigma}_{\boldsymbol{\lambda},\sigma} = \frac{\sigma}{4}\left(-I + \left[\sum_{k=1}^{K} \lambda_k \left(I + \frac{16}{\sigma^2}\widetilde{\Sigma}^{1/2}_{\boldsymbol{\lambda},\sigma}\Sigma_k\widetilde{\Sigma}^{1/2}_{\boldsymbol{\lambda},\sigma}\right)^{1/2}\right]^2\right)^{1/2}.$$

B.3 An Example for Section 3

We now illustrate the formulation of stochastic barycentric optimization (SBO) through an example with Gaussian distribution which admits closed form for their 2-Wasserstein barycenters. Suppose that the source distributions are nondegenerate Gaussians, i.e., $\mathbb{P}_k = \mathcal{N}(\mu_k, \Sigma_k)$, where $\mu_k \in \mathbb{R}^m$ and $\Sigma_k \in \mathbb{S}_+^m$ for $k \in [\![K]\!]$. Given the assumption that the source distributions are Gaussian, the empirical distributions are instead taken as $\widehat{\mathbb{P}}_k = \mathcal{N}(\widehat{\mu}_k, \widehat{\Sigma}_k)$ for $k \in [\![K]\!]$, where $\widehat{\mu}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} z_{k,i}$ and $\widehat{\Sigma}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}(z_{k,i} - \widehat{\mu}_k)(z_{k,i} - \widehat{\mu}_k)^\top$ are their respective empirical means and empirical covariance matrices (which we assume to be full-rank).

Then, by Proposition B.7, the $\boldsymbol{\lambda}$-weighted 2-Wasserstein barycenter of $\widehat{\mathbb{P}}_1, \ldots, \widehat{\mathbb{P}}_K$ is $\widehat{\mathsf{b}}_{\boldsymbol{\lambda}}(\widehat{\mathbb{P}}_1, \ldots, \widehat{\mathbb{P}}_K) = \mathcal{N}(\overline{\mu}_{\boldsymbol{\lambda}}, \overline{\Sigma}_{\boldsymbol{\lambda}})$, where

$$\overline{\mu}_{\boldsymbol{\lambda}} = \sum_{k=1}^K \lambda_k \mu_k \qquad \text{and} \qquad \overline{\Sigma}_{\boldsymbol{\lambda}} = \operatorname*{argmin}_{\Sigma \in \mathbb{S}_{++}^m} \sum_{k=1}^K \lambda_k \mathsf{B}^2(\Sigma, \widehat{\Sigma}_k).$$

Instead of directly minimizing the SBO objective $\widehat{F}^{\mathsf{b}}_{\boldsymbol{\lambda}}(x) \coloneqq \mathbb{E}_{\xi \sim \widehat{\mathsf{b}}_{\boldsymbol{\lambda}}(\widehat{\mathbb{P}}_1, \ldots, \widehat{\mathbb{P}}_K)}[\ell(x, \xi)]$, we first draw N independent samples $\{\widehat{\xi}_i\}_{i=1}^N$ from $\mathcal{N}(\overline{\mu}_{\boldsymbol{\lambda}}, \overline{\Sigma}_{\boldsymbol{\lambda}})$ and then minimize the ERM objective

$$\widehat{F}^{\mathsf{b}}_{\mathsf{ERM}}(x) \coloneqq \frac{1}{N}\sum_{i=1}^N \ell(x, \widehat{\xi}_i).$$

B.4 Additional Results for Section 4

Following the development of the Gelbrich ambiguity set in Section 5, we define similar notions in the case of Wasserstein barycentric ambiguity set.

Let $\mu_1, \ldots, \mu_K \in \mathbb{R}^m$ and $\Sigma_1, \ldots, \Sigma_K \in \mathbb{S}_{++}^m$. The *mean-covariance barycentric ambiguity set* is defined as

$$\widetilde{\mathcal{U}}_\varepsilon\big((\mu_k, \Sigma_k)_{k \in [\![K]\!]}; \boldsymbol{\lambda}\big) \coloneqq \left\{(\mu, \Sigma) \in \mathbb{R}^m \times \mathbb{S}_+^m : \sum_{k=1}^K \lambda_k \mathsf{G}^2((\mu, \Sigma), (\mu_k, \Sigma_k)) \leqslant \varepsilon^2\right\}.$$

Then, the *Gelbrich barycentric ambiguity set* is defined as

$$\widetilde{\mathcal{G}}_\varepsilon\big((\mu_k, \Sigma_k)_{k \in [\![K]\!]}; \boldsymbol{\lambda}\big) \coloneqq \big\{\mathbb{Q} \in \mathscr{P}_2(\Xi) : (\mathbb{E}_{\mathbb{Q}}[\xi], \mathrm{Cov}_{\mathbb{Q}}(\xi)) \in \mathcal{U}_\varepsilon\big((\mu_k, \Sigma_k)_{k \in [\![K]\!]}\big)\big\}.$$

Now let $\mathbb{P}, \mathbb{Q}_1, \ldots, \mathbb{Q}_K \in \mathscr{P}_2(\Xi)$ with means $\mu, \mu_1, \ldots, \mu_K \in \mathbb{R}^m$ and covariance matrices $\Sigma, \Sigma_1, \ldots, \Sigma_K \in \mathbb{S}_{++}^m$ respectively. Then, by the Gelbrich bound, we have

$$\sum_{k=1}^K \lambda_k \mathsf{W}_2^2(\mathbb{P}, \mathbb{Q}_k) \geqslant \sum_{k=1}^K \lambda_k \mathsf{G}^2((\mu, \Sigma), (\mu_k, \Sigma_k)).$$

The above inequality becomes an equality if $\mathbb{P}, \mathbb{Q}_1, \ldots, \mathbb{Q}_K$ belong to the same location-scatter family. Furthermore, this inequality implies

$$\widetilde{\mathcal{W}}_{\varepsilon,2}(\mathbb{Q}_1, \ldots, \mathbb{Q}_K; \boldsymbol{\lambda}) \subseteq \widetilde{\mathcal{G}}_\varepsilon\big((\mu_k, \Sigma_k)_{k \in [\![K]\!]}; \boldsymbol{\lambda}\big).$$

B.5 Additional Results for Section 5

The Gelbrich bound implies the Gelbrich ambiguity set is an outer approximation of the p-Wasserstein ambiguity set with $p \geqslant 2$, i.e., $\mathcal{W}_{\varepsilon,p}(\widehat{\mathbb{P}}) \subseteq \mathcal{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$ for every $p \geqslant 2$. This result immediately leads to an upper bound on the (optimal) worst-case risk by the (optimal) *Gelbrich risk* (i.e., risk under the Gelbrich ambiguity set; Kuhn et al., 2019; Nguyen et al., 2021a, Corollary 1).

Corollary B.6 (Worst-case risk bounds). *If the nominal distribution $\widehat{\mathbb{P}} \in \mathscr{P}_2(\Xi)$ has mean $\widehat{\mu} \in \mathbb{R}^m$ and covariance matrix $\widehat{\Sigma} \in \mathbb{S}_+^m$, then, for every $p \geqslant 2$, we have*

$$(\forall \ell \in \mathcal{L}) \quad \mathcal{R}_{\mathcal{W}_{\varepsilon,p}(\widehat{\mathbb{P}})}(\ell) \leqslant \mathcal{R}_{\mathcal{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})}(\ell) \qquad \textit{and} \qquad \mathcal{R}_{\mathcal{W}_{\varepsilon,p}(\widehat{\mathbb{P}})}(\mathcal{L}) \leqslant \mathcal{R}_{\mathcal{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})}(\mathcal{L}).$$

We now state the following proposition from Peyré and Cuturi (2019, Remark 9.5), which is proved in Agueh and Carlier (2011, Theorem 6.1) and previously known from Knott and Smith (1994); Rüschendorf and Uckelmann (2002).

Proposition B.7. *Let $\mathbb{P}_0 \in \mathscr{P}_2^{\mathrm{ac}}(\mathbb{R}^m)$ and $\mathbb{Q}_1, \ldots, \mathbb{Q}_K \in \mathcal{F}(\mathbb{P}_0)$ with means $\mu_1, \ldots, \mu_K \in \mathbb{R}^m$ and covariance matrices $\Sigma_1, \ldots, \Sigma_K \in \mathbb{S}_{++}^m$ respectively. Then, for $\boldsymbol{\lambda} \in \triangle^K$, the (unique) $\boldsymbol{\lambda}$-weighted 2-Wasserstein barycenter of $\mathbb{Q}_1, \ldots, \mathbb{Q}_K$ is $\overline{\mathbb{Q}}_{\boldsymbol{\lambda},2} \in \mathcal{F}(\mathbb{P}_0)$ with mean $\overline{\mu}_{\boldsymbol{\lambda}} \in \mathbb{R}^m$ and covariance matrix $\overline{\Sigma}_{\boldsymbol{\lambda}} \in \mathbb{S}_{++}^m$ given by, for each $k \in \llbracket K \rrbracket$,*

$$\overline{\mu}_{\boldsymbol{\lambda}} = \sum_{k=1}^{K} \lambda_k \mu_k \qquad \text{and} \qquad \overline{\Sigma}_{\boldsymbol{\lambda}} = \operatorname*{argmin}_{\Sigma \in \mathbb{S}_{++}^m} \sum_{k=1}^{K} \lambda_k \mathsf{B}^2(\Sigma, \Sigma_k), \tag{B.3}$$

where $\mathsf{B}(\cdot,\cdot)$ is the Bures–Wasserstein distance (5.1). The covariance matrix $\overline{\Sigma}_{\boldsymbol{\lambda}}$ can be obtained by finding the unique positive definite fixed point of the equation

$$\Sigma = \sum_{k=1}^{K} \lambda_k \left(\Sigma^{1/2} \Sigma_k \Sigma^{1/2}\right)^{1/2}. \tag{B.4}$$

Bures–Wasserstein Barycenter. Another important example is the Bures–Wasserstein barycenter of $\Sigma_1, \ldots, \Sigma_K \in \mathbb{S}_{++}^m$, defined via the Bures–Wasserstein distance, which coincides with the Wasserstein barycenter of $\mathbb{Q}_1, \ldots, \mathbb{Q}_K$ in the same zero-mean location-scatter family with covariances $\Sigma_1, \ldots, \Sigma_K$ respectively, i.e., $\overline{\Sigma}_{\boldsymbol{\lambda}}$ in (B.3). The issues of approximation, statistical inference and computational algorithms of the Bures–Wasserstein barycenter can be found in Chewi et al. (2020); Kroshnin et al. (2021a,b).

B.6 An Example of Tractability Results

We now derive a tractability result from Propositions 5.3 and B.7 for nominal distributions of location-scatter families and a quadratic loss function, similar to the one for WDRO (Kuhn et al., 2019, Theorem 16). Assume that $\Xi = \mathbb{R}^m$ and consider the quadratic loss function $\ell(\xi) = \langle \xi, Q\xi \rangle + 2\langle q, \xi \rangle$ with $Q \in \mathbb{S}^m$ and $q \in \mathbb{R}^m$. If $\widehat{\mu}_k \in \mathbb{R}^m$ and $\widehat{\Sigma}_k \in \mathbb{S}_{+}^m$ for each $k \in \llbracket K \rrbracket$, then the Gelbrich risk is equal to the optimal values of the following tractable SDP:

$$\mathcal{R}_{\mathcal{G}_\varepsilon(\overline{\mu}_{\boldsymbol{\lambda}}, \overline{\Sigma}_{\boldsymbol{\lambda}})}(\ell) = \inf \left\{ \alpha\left(\varepsilon^2 - \|\overline{\mu}_{\boldsymbol{\lambda}}\|_2^2 - \operatorname{tr}(\overline{\Sigma}_{\boldsymbol{\lambda}})\right) + x + \operatorname{tr}(X) \right\},$$

subject to $\alpha \geqslant 0$, $x \geqslant 0$, $X \in \mathbb{S}_{+}^m$,

$$\begin{pmatrix} \alpha I - Q & q + \alpha \overline{\mu}_{\boldsymbol{\lambda}} \\ q^\top + \alpha \overline{\mu}_{\boldsymbol{\lambda}}^\top & x \end{pmatrix} \succeq 0, \quad \begin{pmatrix} \alpha I - Q & \alpha \overline{\Sigma}_{\boldsymbol{\lambda}}^{1/2} \\ \alpha \overline{\Sigma}_{\boldsymbol{\lambda}}^{1/2} & X \end{pmatrix} \succeq 0.$$

If for each $k \in \llbracket K \rrbracket$, $\widehat{\mathbb{P}}_k \in \mathcal{F}(\mathbb{P}_0)$ for some $\mathbb{P}_0 \in \mathscr{P}_2^{\mathrm{ac}}(\mathbb{R}^m)$ with mean $\widehat{\mu}_k \in \mathbb{R}^m$ and covariance matrix $\widehat{\Sigma}_k \in \mathbb{S}_{++}^m$, and $p = 2$, then the optimal value of the above SDP, the worst-case risk (2.3), and the Gelbrich risk all coincide.

To verify the above claim, we assume that $\mathbb{Q}^\star \in \mathcal{F}(\mathbb{P}_0)$ with mean $\widehat{\mu}^\star$ and covariance $\widehat{\Sigma}^\star$ is the extremal distribution, i.e.,

$$\mathbb{Q}^\star = \operatorname*{argmax}_{\mathbb{Q} \in \mathcal{G}_\varepsilon(\overline{\mu}_{\boldsymbol{\lambda}}, \overline{\Sigma}_{\boldsymbol{\lambda}})} \mathcal{R}_{\mathbb{Q}}(\ell) \quad \text{and} \quad \mathcal{R}_{\mathbb{Q}^\star}(\ell) = \mathcal{R}_{\mathcal{G}_\varepsilon(\overline{\mu}_{\boldsymbol{\lambda}}, \overline{\Sigma}_{\boldsymbol{\lambda}})}(\ell).$$

Note that $\mathbb{Q}^\star \in \overline{\mathcal{W}}_{\varepsilon,p}(\widehat{\mathbb{P}}_1, \ldots, \widehat{\mathbb{P}}_K; \boldsymbol{\lambda})$. Then the worst-case risk satisfies

$$\mathcal{R}_{\mathbb{Q}^\star}(\ell) \leqslant \mathcal{R}_{\overline{\mathcal{W}}_{\varepsilon,p}(\widehat{\mathbb{P}}_1, \ldots, \widehat{\mathbb{P}}_K; \boldsymbol{\lambda})}(\ell) \leqslant \mathcal{R}_{\mathcal{G}_\varepsilon(\overline{\mu}_{\boldsymbol{\lambda}}, \overline{\Sigma}_{\boldsymbol{\lambda}})}(\ell),$$

which follows from the Gelbrich bound and Theorem 5.4. Details of how to solve this SDP can be found in Kuhn et al. (2019).

C Proofs

C.1 Preparatory Lemmas

We first state some useful preparatory lemmas.

Lemma C.1. *For $1 \leqslant p < q < \infty$, we have*

$$\mathsf{W}_p(\rho, \nu) \leqslant \mathsf{W}_q(\rho, \nu)$$

for any $\rho, \nu \in \mathscr{P}_q(\Omega)$ with $\Omega \subseteq \mathbb{R}^m$.

Proof of Lemma C.1. Since $\|x-y\|^p$ is convex in $\|x-y\|$ for all $p \in [1, +\infty)$, by Jensen's inequality, for $p \leqslant q$ we have

$$\left(\int_{\Omega\times\Omega} \|x-y\|^p \,\mathrm{d}\pi(x,y)\right)^{1/p} \leqslant \left(\int_{\Omega\times\Omega} \|x-y\|^q \,\mathrm{d}\pi(x,y)\right)^{1/q},$$

which implies that $\mathsf{W}_p(\rho, \nu) \leqslant \mathsf{W}_q(\rho, \nu)$. □

Lemma C.2. *For $a, b \in \mathbb{R}_+$ and $p \in [1, +\infty)$, we have $(a+b)^p \leqslant 2^{p-1}(a^p + b^p)$.*

Proof of Lemma C.2. by the convexity of $\mathbb{R}_+ \ni s \mapsto s^p$, we get

$$\left(\frac{a+b}{2}\right)^p \leqslant \frac{a^p + b^p}{2},$$

which is equivalent to $(a+b)^p \leqslant 2^{p-1}(a^p + b^p)$. □

C.2 Proof of Theorem 4.2

Proof of Theorem 4.2. Assume that $\mathbb{Q}_1, \ldots, \mathbb{Q}_K \in \mathscr{P}_p(\Xi)$ have a $\boldsymbol{\lambda}$-weighted p-Wasserstein barycenter $\overline{\mathbb{Q}}_{\boldsymbol{\lambda},p}$, where $\boldsymbol{\lambda} \in \triangle^K$. Using the argument of Figalli and Glaudo (2021, Remark 3.1.2), by the triangle inequality and Lemma C.2, for any $\mathbb{P} \in \mathscr{P}_p(\Xi)$ and $k \in [\![K]\!]$, we have

$$\mathsf{W}_p^p(\mathbb{P}, \overline{\mathbb{Q}}_{\boldsymbol{\lambda},p}) \leqslant \big(\mathsf{W}_p(\mathbb{P}, \mathbb{Q}_k) + \mathsf{W}_p(\overline{\mathbb{Q}}_{\boldsymbol{\lambda},p}, \mathbb{Q}_k)\big)^p \leqslant 2^{p-1}\big(\mathsf{W}_p^p(\mathbb{P}, \mathbb{Q}_k) + \mathsf{W}_p^p(\overline{\mathbb{Q}}_{\boldsymbol{\lambda},p}, \mathbb{Q}_k)\big).$$

This implies that

$$\begin{aligned}(\forall \mathbb{P} \in \mathscr{P}_p(\Xi)) \quad \mathsf{W}_p^p(\mathbb{P}, \overline{\mathbb{Q}}_{\boldsymbol{\lambda},p}) &= \sum_{k=1}^K \lambda_k \mathsf{W}_p^p(\mathbb{P}, \overline{\mathbb{Q}}_{\boldsymbol{\lambda},p}) \\ &\leqslant 2^{p-1}\sum_{k=1}^K \lambda_k \mathsf{W}_p^p(\mathbb{P}, \mathbb{Q}_k) + 2^{p-1}\sum_{k=1}^K \lambda_k \mathsf{W}_p^p(\overline{\mathbb{Q}}_{\boldsymbol{\lambda},p}, \mathbb{Q}_k) \\ &\leqslant 2^{p-1}\sum_{k=1}^K \lambda_k \mathsf{W}_p^p(\mathbb{P}, \mathbb{Q}_k) + 2^{p-1}\sum_{k=1}^K \lambda_k \mathsf{W}_p^p(\mathbb{P}, \mathbb{Q}_k) \\ &= 2^p \sum_{k=1}^K \lambda_k \mathsf{W}_p^p(\mathbb{P}, \mathbb{Q}_k),\end{aligned}$$

where the second inequality follows from the definition of the $\boldsymbol{\lambda}$-weighted p-Wasserstein barycenter.

Consequently, for any $\varepsilon \geqslant 0$,

$$\sum_{k=1}^K \lambda_k \mathsf{W}_p^p(\mathbb{P}, \mathbb{Q}_k) \leqslant \varepsilon \implies \mathsf{W}_p^p(\mathbb{P}, \overline{\mathbb{Q}}_{\boldsymbol{\lambda},p}) \leqslant 2^p \cdot \varepsilon,$$

which implies the desired result. □

C.3 Proof of Theorem 4.6

Proof of Theorem 4.6. First note that the 2-Wasserstein space $\mathscr{W}_2(\mathbb{R}^m) = (\mathscr{P}_2(\mathbb{R}^m), \mathsf{W}_2)$ is positively curved in the sense of Alexandrov (Ambrosio et al., 2005, §7.3). Then, by Le Gouic et al. (2021, Theorem 12), we have, for any $\beta \in (0,1)$,

$$\mathbb{P}^n\left\{\mathsf{W}_2^2\left(\widehat{\mathsf{b}}_{\mathbf{1}/K,2}(\widehat{\mathbb{P}}^n), \widehat{\mathsf{b}}_K^\star\right) \leqslant \frac{c_1}{n}\log\left(\frac{2}{\beta}\right)\right\} \geqslant 1 - \beta - \mathrm{e}^{-c_2 n},$$

where $\widehat{\mathsf{b}}_{\mathbf{1}/K,2}(\widehat{\mathbb{P}}^n) = \mathsf{b}_2(\widehat{\rho}_{n,K})$, and $c_1, c_2 \in \mathbb{R}_{++}$ are independent of n. By the definition of $\overline{\mathbb{W}}_{\varepsilon,2}(\widehat{\mathbb{P}}^n; \mathbf{1}/K)$, we also have

$$\mathbb{P}^n\left\{\widehat{\mathsf{b}}_K^\star \in \overline{\mathbb{W}}_{\varepsilon,2}(\widehat{\mathbb{P}}^n; \mathbf{1}/K)\right\} = \mathbb{P}^n\left\{\mathsf{W}_2\left(\widehat{\mathsf{b}}_{\mathbf{1}/K,2}(\widehat{\mathbb{P}}^n), \widehat{\mathsf{b}}_K^\star\right) \leqslant \varepsilon\right\},$$

hence the desired result. □

Remark C.3. Note that Adve and Mészáros (2020) provides a heuristic argument for the non-negative curvature of $\mathscr{W}_p(\mathbb{R}^m)$ when $p \in (1, p(m))$, where $p(m) > 1$ is close to 1 (e.g., $p(m) = 1 + 1/\mathscr{O}(m^2 \log m)$). Consequently, Theorem 4.6 should also hold for $p \in (1, p(m))$. For $p \notin (1, p(m)) \cup \{2\}$, it remains unclear whether $\mathscr{W}_p(\mathbb{R}^m)$ is positively curved even though it is expected. In this case, the general concentration inequality in Le Gouic et al. (2021, Theorem 12) involving various abstract constants cannot be easily simplified.

C.4 Proof of Theorem 4.7

Proof of Theorem 4.7. By Theorem 4.6, the inequality $\mathbb{P}^n\left\{\widehat{\mathsf{b}}_K^\star \in \overline{\mathbb{W}}_{\varepsilon,2}(\widehat{\mathbb{P}}^n; \mathbf{1}/K)\right\} \geqslant 1-\beta-\mathrm{e}^{-c_2 n}$ immediately implies that

$$\mathcal{R}_{\widehat{\mathsf{b}}_K^\star}(\ell) \leqslant \sup_{\mathbb{P}\in\overline{\mathbb{W}}_{\varepsilon,2}(\widehat{\mathbb{P}}^n;\mathbf{1}/K)} \mathcal{R}_{\mathbb{P}}(\ell) =: \mathcal{R}_{\overline{\mathbb{W}}_{\varepsilon_n,2}(\widehat{\mathbb{P}}^n;\mathbf{1}/K)}(\ell)$$

with probability at least $1 - \beta - \mathrm{e}^{-c_2 n}$, where $c_2 > 0$ is independent of n. □

C.5 Proof of Theorem 4.8

To prove Theorem 4.8, we need the following lemma which is a slight modification of Esfahani and Kuhn (2018, Lemma 3.7).

Lemma C.4. *Let $K \in \mathbb{N}^*$ be finite and fixed. Suppose that $\mathbb{P} \in \mathscr{W}_2(\mathscr{P}_2(\Xi))$ is sub-Gaussian, $\beta_n \in (0,1)$ and $\varepsilon_n = \varepsilon_n(\beta_n) \in \mathbb{R}_{++}$ for $n \in \mathbb{N}^*$, satisfies $\sum_{n=1}^\infty \beta_n < \infty$ and $\lim_{n\to\infty} \varepsilon_n(\beta_n) = 0$, then any sequence $\widehat{\mathbb{Q}}_n \in \overline{\mathbb{W}}_{\varepsilon_n(\beta_n),2}(\widehat{\mathbb{P}}^n; \mathbf{1}/K)$, $n \in \mathbb{N}^*$, where $\widehat{\mathbb{Q}}_n$ may depend on the observations, converges in the W_2 distance to $\widehat{\mathsf{b}}_K^\star$ almost surely with respect to $\mathbb{P}^\infty$, i.e,*

$$\mathbb{P}^\infty\left\{\lim_{n\to\infty} \mathsf{W}_2\left(\widehat{\mathsf{b}}_K^\star, \widehat{\mathbb{Q}}_n\right)\right\} = 1.$$

Proof of Lemma C.4. For any $\widehat{\mathbb{Q}}_n \in \overline{\mathbb{W}}_{\varepsilon_n(\beta_n),2}(\widehat{\mathbb{P}}^n; \mathbf{1}/K)$, the triangle inequality implies

$$\mathsf{W}_2\left(\widehat{\mathsf{b}}_K^\star, \widehat{\mathbb{Q}}_n\right) \leqslant \mathsf{W}_2\left(\widehat{\mathsf{b}}_K^\star, \widehat{\mathsf{b}}_{\mathbf{1}/K,2}(\widehat{\mathbb{P}}^n)\right) + \mathsf{W}_2\left(\widehat{\mathsf{b}}_{\mathbf{1}/K,2}(\widehat{\mathbb{P}}^n), \widehat{\mathbb{Q}}_n\right) \leqslant \mathsf{W}_2\left(\widehat{\mathsf{b}}_K^\star, \widehat{\mathsf{b}}_{\mathbf{1}/K,2}(\widehat{\mathbb{P}}^n)\right) + \varepsilon_n(\beta_n).$$

In addition, by Theorem 4.7, we have $\mathbb{P}^n\left\{\mathsf{W}_2\left(\widehat{\mathsf{b}}_K^\star, \widehat{\mathsf{b}}_{\mathbf{1}/K,2}(\widehat{\mathbb{P}}^n)\right) \leqslant \varepsilon_n(\beta_n)\right\} \geqslant 1 - \beta_n - \mathrm{e}^{-c_2 n}$, which implies

$$\mathbb{P}^n\left\{\mathsf{W}_2\left(\widehat{\mathsf{b}}_K^\star, \widehat{\mathbb{Q}}_n\right) \leqslant 2\varepsilon_n(\beta_n)\right\} \geqslant 1 - \beta_n - \mathrm{e}^{-c_2 n}.$$

Since $\sum_{n=1}^\infty \beta_n < \infty$ and $\sum_{n=1}^\infty \mathrm{e}^{-c_2 n} < \infty$ as $c_2 > 0$, invoking the (second) Borel–Cantelli lemma (Chung, 2001, Theorem 4.2.4) yields

$$\mathbb{P}^\infty\left\{\mathsf{W}_2\left(\widehat{\mathsf{b}}_K^\star, \widehat{\mathbb{Q}}_n\right) \leqslant \varepsilon_n(\beta_n) \text{ for sufficiently large } n\right\} = 1.$$

Since $\varepsilon_n(\beta) \to 0$ as $n \to \infty$, we conclude that $\lim_{n\to\infty} \mathsf{W}_2\left(\widehat{\mathsf{b}}_K^\star, \widehat{\mathbb{Q}}_n\right)$ $\mathbb{P}^\infty$-almost surely. □

Now we are ready to prove Theorem 4.8.

Proof of Theorem 4.8. Let $K \in \mathbb{N}^*$ be finite and fixed. Let us define

$$\ell^\star_{n,K} \coloneqq \operatorname*{argmin}_{\ell \in \mathcal{L}} \mathcal{R}_{\overline{\mathcal{W}}_{\varepsilon,2}(\widehat{\mathbb{P}}^n; \mathbf{1}/K)} \quad \text{and} \quad \mathcal{R}_{\widehat{\mathsf{b}}^\star_K}(\mathcal{L}) \coloneqq \inf_{\ell \in \mathcal{L}} \mathcal{R}_{\widehat{\mathsf{b}}^\star_K}(\ell).$$

Since $\ell^\star_{n,K} \in \mathcal{L}$, we have $\mathcal{R}_{\widehat{\mathsf{b}}^\star_K}(\ell^\star_{n,K}) \leqslant \mathcal{R}_{\widehat{\mathsf{b}}^\star_K}(\mathcal{L})$. Theorem 4.6 implies

$$\begin{aligned} \mathbb{P}^n\Big\{\mathcal{R}_{\widehat{\mathsf{b}}^\star_K}(\ell^\star_{n,K}) \leqslant \mathcal{R}_{\widehat{\mathsf{b}}^\star_K}(\mathcal{L}) \leqslant \mathcal{R}_{\overline{\mathcal{W}}_{\varepsilon_n(\beta_n),2}(\widehat{\mathbb{P}}^n; \mathbf{1}/K)}(\ell^\star_{n,K})\Big\} &\geqslant \mathbb{P}^n\Big\{\widehat{\mathsf{b}}^\star_K \in \overline{\mathcal{W}}_{\varepsilon_n(\beta_n),2}(\widehat{\mathbb{P}}^n; \mathbf{1}/K)\Big\} \\ &\geqslant 1 - \beta - \mathrm{e}^{-c_2 n} \end{aligned}$$

for all $n \in \mathbb{N}^*$. Since $\sum_{n=1}^\infty \beta_n < \infty$ and $\sum_{n=1}^\infty \mathrm{e}^{-c_2 n} < \infty$ as $c_2 > 0$, invoking the (second) Borel–Cantelli lemma again yields

$$\mathbb{P}^\infty\Big\{\mathcal{R}_{\widehat{\mathsf{b}}^\star_K}(\ell^\star_{n,K}) \leqslant \mathcal{R}_{\widehat{\mathsf{b}}^\star_K}(\mathcal{L}) \leqslant \mathcal{R}_{\overline{\mathcal{W}}_{\varepsilon_n(\beta_n),2}(\widehat{\mathbb{P}}^n; \mathbf{1}/K)}(\ell^\star_{n,K}) \text{ for sufficiently large } n\Big\} = 1.$$

Thus, it remains to show that $\limsup_{n\to\infty} \mathcal{R}_{\overline{\mathcal{W}}_{\varepsilon_n(\beta_n),2}(\widehat{\mathbb{P}}^n; \mathbf{1}/K)}(\ell^\star_{n,K}) \leqslant \mathcal{R}_{\widehat{\mathsf{b}}^\star_K}(\ell^\star_{n,K})$ with probability one. This part is more subtle and involves overwhelming technicalities—we refer to the proof of Esfahani and Kuhn (2018, Theorem 3.6) as our proof follows exactly the same steps.

The second part of Theorem 4.8 follows exactly the same procedures as that of the first part. By considering that n has already been taken to ∞, results similar to Theorems 4.6 and 4.7 also hold, by replacing n with K, $\widehat{\mathbb{P}}^n$ with $\widehat{\mathbb{P}}^\star$, $\widehat{\mathsf{b}}^\star_K$ with $\mathsf{b}^\star$, etc. □

C.6 Proof of Proposition 5.3

Proof of Proposition 5.3. Since $\mathbb{Q}_1, \ldots, \mathbb{Q}_K$ belong to the same location-scatter family $\mathcal{F}(\mathbb{P}_0)$ with $\mathbb{P}_0 \in \mathscr{P}_2^{\mathrm{ac}}(\mathbb{R}^m)$, so is their $\boldsymbol{\lambda}$-weighted 2-Wasserstein barycenter $\overline{\mathbb{Q}}_{\boldsymbol{\lambda},2}$, since location-scatter families are closed for barycenters (Álvarez-Esteban et al., 2018, Theorem 3.8).

Let us recall the definition of the 2-Wasserstein ambiguity set

$$\overline{\mathcal{W}}^{\mathsf{ls}}_{\varepsilon,2}(\mathbb{Q}_1, \ldots, \mathbb{Q}_K; \boldsymbol{\lambda}) \coloneqq \big\{\mathbb{P} \in \mathcal{F}(\mathbb{P}_0) : \mathsf{W}_2(\mathbb{P}, \overline{\mathbb{Q}}_{\boldsymbol{\lambda},2}) \leqslant \varepsilon\big\}.$$

Recall that the W_2 distance between two distributions of the same location-scatter family is the Gelbrich distance, i.e., $\mathsf{W}_2(\mathbb{P}, \overline{\mathbb{Q}}_{\boldsymbol{\lambda},2}) = \mathsf{G}((\mu, \Sigma), (\overline{\mu}_{\boldsymbol{\lambda}}, \overline{\Sigma}_{\boldsymbol{\lambda}}))$, where μ and Σ are the mean and the covariance matrix of $\mathbb{P} \in \mathcal{F}(\mathbb{P}_0)$ respectively. Therefore, we have

$$\begin{aligned} \overline{\mathcal{W}}^{\mathsf{ls}}_{\varepsilon,2}(\mathbb{Q}_1, \ldots, \mathbb{Q}_K; \boldsymbol{\lambda}) &= \big\{\mathbb{P} \in \mathcal{F}(\mathbb{P}_0) : \mathsf{G}((\mu, \Sigma), (\overline{\mu}_{\boldsymbol{\lambda}}, \overline{\Sigma}_{\boldsymbol{\lambda}})) \leqslant \varepsilon\big\} \\ &= \big\{\mathbb{P} \in \mathscr{P}_2(\Xi) : \mathsf{G}((\mu, \Sigma), (\overline{\mu}_{\boldsymbol{\lambda}}, \overline{\Sigma}_{\boldsymbol{\lambda}})) \leqslant \varepsilon\big\} \cap \mathcal{F}(\mathbb{P}_0) \\ &= \mathcal{G}_\varepsilon(\overline{\mu}_{\boldsymbol{\lambda}}, \overline{\Sigma}_{\boldsymbol{\lambda}}) \cap \mathcal{F}(\mathbb{P}_0). \end{aligned}$$

□

C.7 Proof of Theorem 5.4

Proof of Theorem 5.4. Let $\mathbb{P} \in \mathscr{P}_2(\Xi)$ with mean $\mu \in \mathbb{R}^m$ and covariance matrix $\Sigma \in \mathbb{S}^m_+$, and let $\widehat{\mathbb{P}}_1, \ldots, \widehat{\mathbb{P}}_K \in \mathscr{P}_2^{\mathrm{ac}}(\Xi)$ with means $\mu_1, \ldots, \mu_K \in \mathbb{R}^m$ and covariance matrices $\Sigma_1, \ldots, \Sigma_K \in \mathbb{S}^m_{++}$ respectively. By the Gelbrich bound, we have

$$\mathsf{W}_2(\mathbb{P}, \widehat{\mathsf{b}}_{\boldsymbol{\lambda},2}(\widehat{\mathbb{P}}_1, \ldots, \widehat{\mathbb{P}}_K)) \geqslant \mathsf{G}((\mu, \Sigma), (\overline{\mu}_{\boldsymbol{\lambda}}, \overline{\Sigma}_{\boldsymbol{\lambda}})),$$

where $\overline{\mu}_{\boldsymbol{\lambda}}$ and $\overline{\Sigma}_{\boldsymbol{\lambda}}$ are the mean and the covariance matrix of $\widehat{\mathsf{b}}_{\boldsymbol{\lambda},2}(\widehat{\mathbb{P}}_1, \ldots, \widehat{\mathbb{P}}_K)$. The two worst-case risk bounds immediately follow from this inequality due to the definitions of the worst-case risk and the optimal worst-case risk. □

C.8 Proof of Corollary B.6

Proof of Corollary B.6. The Gelbrich bound and Lemma C.1 together imply that, for any $\mathbb{P}, \widehat{\mathbb{P}} \in \mathscr{P}_2(\Xi)$ with means $\mu, \widehat{\mu} \in \mathbb{R}^m$ and covariance matrices $\Sigma, \widehat{\Sigma} \in \mathbb{S}_{++}^m$, we have

$$\mathsf{G}((\mu, \Sigma), (\widehat{\mu}, \widehat{\Sigma})) = \mathsf{W}_2(\rho, \nu) \leqslant \mathsf{W}_p(\rho, \nu).$$

Consequently, we have the inclusion property

$$\mathcal{W}_{\varepsilon,p}(\widehat{\mathbb{P}}) \subseteq \mathcal{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$$

for every $p \geqslant 2$. Thus, according to the definitions of the worst-case risk (2.3) and the optimal worst-case risk (2.4), the desired results follow. □

C.9 Proof of Proposition B.7

Proof of Proposition B.7. See e.g., Agueh and Carlier (2011, Theorem 6.1) and Knott and Smith (1994); Rüschendorf and Uckelmann (2002). □

D Experimental Details

In this section, we give further details about Section 6.

D.1 The Wasserstein Barycentric Shrinkage Estimator

The following theorem indicates the way to compute the solution $X^\star$ of the DRMLE problem (6.1).

Theorem D.1. *Assume that $\varepsilon > 0$ and $\widehat{\Sigma}_k \in \mathbb{S}_+^m$ for each $k \in [\![K]\!]$ with all least one of the $\widehat{\Sigma}_k$'s in $\mathbb{S}_{++}^m$, and that the $\boldsymbol{\lambda}$-weighted Bures–Wasserstein barycenter $\overline{\Sigma}_{\boldsymbol{\lambda}}$ of $\widehat{\Sigma}_1, \dots, \widehat{\Sigma}_K$ admits the spectral decomposition $\overline{\Sigma}_{\boldsymbol{\lambda}} = \sum_{j=1}^m \zeta_j v_j v_j^\top$ with eigenvalues $\zeta_j \in \mathbb{R}_+$ and corresponding orthonormal eigenvectors $v_j \in \mathbb{R}^m$, $j \in [\![m]\!]$. Then the unique minimizer of the DRMLE problem (6.1) is given by $X^\star = \sum_{j=1}^m x_j^\star v_j v_j^\top$, where, for each $j \in [\![m]\!]$,*

$$x_j^\star = \chi^\star \left[1 - \frac{1}{2}\left(\sqrt{\zeta_j^2 (\chi^\star)^2 + 4\zeta_j^2 \chi^\star} - \zeta_j \chi^\star\right)\right],$$

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

$$\left(\varepsilon^2 - \frac{1}{2}\sum_{j=1}^m \zeta_j\right)\chi - m + \frac{1}{2}\sum_{j=1}^m \sqrt{\zeta_j^2 \chi^2 + 4\zeta_j \chi} = 0. \tag{D.1}$$

Proof of Theorem D.1. According to the formulation of the DRMLE problem (6.1). The above theorem is simply Nguyen et al. (2022, Theorem 3.1) with $\widehat{\Sigma}$ replaced by $\overline{\Sigma}_{\boldsymbol{\lambda}}$. □

Recall that $\overline{\Sigma}_{\boldsymbol{\lambda}}$ can be obtained by finding the unique positive definite fixed point of the equation (B.4). Thus, to approximate the barycenter $\overline{\Sigma}_{\boldsymbol{\lambda}}$, one iterates

$$(\forall t \in \mathbb{N}^*) \quad S_{t+1} = S_t^{-1/2}\left(\sum_{k=1}^K \lambda_k (S_t^{1/2} \Sigma_k S_t^{1/2})^{1/2}\right)^2 S_t^{-1/2},$$

which gives $\lim_{t\to\infty} S_t = \overline{\Sigma}_{\boldsymbol{\lambda}}$. Details of this iterative scheme can be found in Álvarez-Esteban et al. (2016).

D.2 The Averaged Linear Shrinkage Estimator

A naïve estimator is constructed by simply replacing the sample covariance matrix in the widely-used linear shrinkage estimator by the $\boldsymbol{\lambda}$-weighted average of the sample covariance matrices $\widehat{\Sigma}_1, \ldots, \widehat{\Sigma}_K$, defined as follows.

Definition D.2 (Averaged linear shrinkage estimator)**.** Given K empirical covariance matrices $\widehat{\Sigma}_1, \ldots, \widehat{\Sigma}_K \in \mathbb{S}_{++}^m$, the $\boldsymbol{\lambda}$-weighted linear shrinkage estimator for the precision matrix $X \in \mathbb{S}_{++}^m$ is defined by

$$X^\star = \left[(1-\alpha)\sum_{k=1}^{K}\lambda_k\widehat{\Sigma}_k + \alpha\sum_{k=1}^{K}\lambda_k \operatorname{Diag}(\widehat{\Sigma}_k)\right]^{-1},$$

where $\alpha \in [0,1]$, $\boldsymbol{\lambda} \in \triangle^K$ and $\operatorname{Diag}(A)$ is the diagonal matrix with the same diagonal of the square matrix $A \in \mathbb{R}^{m\times m}$.

Let us recall that the $\boldsymbol{\lambda}$-weighted average of the sample covariance matrices $\widehat{\Sigma}_1, \ldots, \widehat{\Sigma}_K$ is indeed the Frobenius barycenter of $\widehat{\Sigma}_1, \ldots, \widehat{\Sigma}_K$:

$$\sum_{k=1}^{K}\lambda_k\widehat{\Sigma}_k = \operatorname*{argmin}_{\Sigma\in\mathbb{S}_{++}^m}\sum_{k=1}^{K}\lambda_k\|\Sigma - \widehat{\Sigma}_k\|_{\mathrm{F}}^2.$$

D.3 The Averaged L_1-Regularized Maximum Likelihood Estimator

An averaged L_1-regularized maximum likelihood estimator of the precision matrix can be obtained from simply minimizing a $\boldsymbol{\lambda}$-weighted loss function of the original L_1-regularized maximum likelihood estimation problem.

Definition D.3 (Averaged L_1-regularized maximum likelihood estimator)**.** Let us recall that the original L_1-regularized maximum likelihood estimation problem takes the following objective:

$$\operatorname*{minimize}_{X\in\mathbb{S}_{++}^m}\ g(X,\widehat{\Sigma}) \coloneqq -\log\det X + \left\langle\!\!\left\langle \widehat{\Sigma}, X\right\rangle\!\!\right\rangle_{\mathrm{F}} + \tau\|X\|_1,$$

where $\tau \geqslant 0$. Then, given K empirical covariance matrices $\widehat{\Sigma}_1, \ldots, \widehat{\Sigma}_K \in \mathbb{S}_{++}^m$, the $\boldsymbol{\lambda}$-weighted L_1-regularized maximum likelihood estimator for the precision matrix $X \in \mathbb{S}_{++}^m$ is defined by

$$X^\star = \operatorname*{argmin}_{X\in\mathbb{S}_{++}^m}\sum_{k=1}^{K}\lambda_k g(X,\widehat{\Sigma}_k) = -\log\det X + \sum_{k=1}^{K}\lambda_k\left\langle\!\!\left\langle \widehat{\Sigma}_k, X\right\rangle\!\!\right\rangle_{\mathrm{F}} + \tau\|X\|_1 = g\left(X, \sum_{k=1}^{K}\lambda_k\widehat{\Sigma}_k\right),$$

where $\tau \geqslant 0$ and $\boldsymbol{\lambda} \in \triangle^K$.

D.4 The Sinkhorn Barycentric Shrinkage Estimator

The Sinkhorn Barycentric Shrinkage Estimator (SBSE) is simply an estimator similar to WBSE with the 2-Wasserstein barycenter replaced by the Sinkhorn barycenter, which is used because of the non-existence of the 2-Wasserstein barycenter under the high-dimensional setting and computational consideration used in simulations. Theoretical treatment of this estimator is left for future work. We do not use the entropic-2-Wasserstein barycenter in simulations since it does *not* make much sense (see Minh, 2022, Remark 4). Also recall from Proposition B.5 for the Sinkhorn barycenter of Gaussians (restricted to the manifold of Gaussians).

D.5 Simulation Settings

All experiments were run with a laptop with Intel Core i7-7700HQ CPU (2.80 GHz) and 32GB RAM, using Python 3.9 with libraries `numpy` (Harris et al., 2020), `scipy` (Virtanen et al., 2020) and `scikit-learn` (Pedregosa et al., 2011).

The equation (D.1) is solved via the Netwon–Raphson method in `scipy`. We give the choice of tuning parameters in Table 2, which are obtained based on grid search.

Table 2: Tuning parameters of LS, L_1 and WBSE.

n	K	α	τ	ε
50	25	0.1	0.1	0.3
	50	0.1	0.1	0.3
	100	0.1	0.1	0.3
100	25	0.1	0.1	0.03
	50	0.1	0.1	0.03
	100	0.1	0.1	0.03
200	25	0.1	0.1	0.03
	50	0.1	0.1	0.03
	100	0.1	0.1	0.005

Remark D.4. While we treat the choice of ε as a tuning parameter in simulations, in Blanchet and Si (2019), the optimal distributional uncertainty size $\varepsilon = \varepsilon_n$ is studied as a function of the sample size n for the Wasserstein Shrinkage Estimator (Nguyen et al., 2022). Blanchet and Si (2019) prove that ε_n should scale at rate $\varepsilon_n = \varepsilon^\star n^{-1}(1 + o(1)) = \mathscr{O}(n^{-1})$, which aligns with the empirical findings of Nguyen et al. (2022). This is as opposed to the theoretical rate of $\mathscr{O}(n^{-1/2})$. It is interesting to find the optimal scaling of ε in our proposed Wasserstein Barycentric Shrinkage Estimator in terms of both n and K, and is left for future work.

E Simulations under High-Dimensional Setting

In this section, we study the performance of the proposed Sinkhorn barycentric shrinkage estimator (see Appendix D.4) under the high-dimensional setting ($m > n$). Let us recall that under the high-dimensional setting, the sample covariance matrices $\widehat{\Sigma}_k$'s are all rank-deficient, so the empirical (unregularized) 2-Wasserstein barycenter of K Gaussians does not exist. We thus resort to the Sinkhorn barycenter with entropic regularization strength $\sigma > 0$. In particular, we choose $m = 20$, $n \in \{5, 10, 15\}$, $K \in \{25, 50, 100\}$ and $\sigma = 0.1$. The Stein losses of the estimators are given in Table 3, averaged over 20 independent trials.

Table 3: Stein losses of LS, L_1 and SBSE.

n	K	LS	L_1	SBSE
5	25	7.61 ± 0.73	8.43 ± 0.79	2.72 ± 0.23
	50	7.75 ± 0.59	8.67 ± 0.63	2.54 ± 0.21
	100	7.70 ± 0.47	8.67 ± 0.51	2.43 ± 0.15
10	25	7.12 ± 0.69	7.99 ± 0.74	1.94 ± 0.29
	50	7.12 ± 0.44	8.04 ± 0.47	1.37 ± 0.17
	100	7.20 ± 0.35	8.15 ± 0.38	1.07 ± 0.10
15	25	6.96 ± 0.55	7.84 ± 0.60	1.91 ± 0.27
	50	6.99 ± 0.40	7.90 ± 0.43	1.12 ± 0.16
	100	6.97 ± 0.27	7.90 ± 0.29	0.71 ± 0.10

Similar to WBSE under the low-dimensional setting, we observe that SBSE also outperforms the other two estimators by a large margin. The performance of SBSE also improves as n and K increase.

We also give the choice of tuning parameters in Table 4, which are obtained based on grid search.

The Effect of Entropic Regularization Strength σ in SBSE. We also numerically study how different entropic regularization strengths in the Sinkhorn barycenter affect the performance of the proposed Sinkhorn

Table 4: Tuning parameters of LS, L_1 and SBSE.

n	K	α	τ	ε
5	25	0.1	0.1	1
	50	0.1	0.1	1
	100	0.1	0.1	1
10	25	0.1	0.1	0.5
	50	0.1	0.1	0.5
	100	0.1	0.1	0.5
15	25	0.1	0.1	0.3
	50	0.1	0.1	0.3
	100	0.1	0.1	0.3

barycentric shrinkage estimator. We choose $m = 20$, $n = 5$, $K = 25$, $\alpha = 0.1$, $\tau = 0.1$, $\varepsilon = 0.8$ and $\sigma \in \{0.01, 0.1, 1, 10, 100\}$.

Table 5: Stein losses of LS, L_1 and SBSE with different σ's.

σ	LS	L_1	SBSE
0.01	7.61 ± 0.73	8.43 ± 0.79	4.47 ± 0.35
0.1			2.03 ± 0.22
1			5.58 ± 0.36
10			9.82 ± 0.71
100			11.70 ± 0.88

We observe that, given a fixed set of (m, n, K, ε), smaller σ (i.e., closer approximation to the unregularized 2-Wasserstein barycenter) does not necessarily yield lower Stein loss. However, as the strength of entropic regularization grows, the performance of SBSE decays sharply, which could be even worse than the averaged linear shrinkage and averaged L_1-regularized maximum likelihood estimators.