Received: 8 November 2018 | Accepted: 21 May 2020

DOI: 10.1111/mafi.12280

MATHEMATICAL FINANCE An International Journal of Mathematics, Statistics and Financial Economics WILEY

ORIGINAL ARTICLE

Robust risk aggregation with neural networks

Stephan Eckstein[1] | **Michael Kupper**[1] | **Mathias Pohl**[2]

[1] Department of Mathematics, University of Konstanz, Konstanz, Germany

[2] Faculty of Business, Economics & Statistics, University of Vienna, Vienna, Austria

Correspondence
Mathias Pohl, Faculty of Business, Economics, and Statistics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria.
Email: mathias.pohl@univie.ac.at

Funding information
Austrian Science Fund, Grant/Award Number: P28661

Abstract

We consider settings in which the distribution of a multivariate random variable is partly ambiguous. We assume the ambiguity lies on the level of the dependence structure, and that the marginal distributions are known. Furthermore, a current best guess for the distribution, called reference measure, is available. We work with the set of distributions that are both close to the given reference measure in a transportation distance (e.g., the Wasserstein distance), and additionally have the correct marginal structure. The goal is to find upper and lower bounds for integrals of interest with respect to distributions in this set. The described problem appears naturally in the context of risk aggregation. When aggregating different risks, the marginal distributions of these risks are known and the task is to quantify their joint effect on a given system. This is typically done by applying a meaningful risk measure to the sum of the individual risks. For this purpose, the stochastic interdependencies between the risks need to be specified. In practice, the models of this dependence structure are however subject to relatively high model ambiguity. The contribution of this paper is twofold: First, we derive a dual representation of the considered problem and prove that strong duality holds. Second, we propose a generally applicable and computationally feasible method, which relies on neural networks, in order to numerically solve the derived dual problem. The latter method is tested on

a number of toy examples, before it is finally applied to perform robust risk aggregation in a real-world instance.

KEYWORDS
average value at risk, dependence uncertainty, model uncertainty, neural networks, optimal transport, penalization, risk bounds, Wasserstein distance

1 | INTRODUCTION

1.1 | Motivation

Risk aggregation is the process of combining multiple types of risk within a firm. The aim is to obtain meaningful measures for the overall risk the firm is exposed to. The stochastic interdependencies between the different risk types are crucial in this respect. There is a variety of different approaches to model these interdependencies. One generally observes that these models for the dependence structure between the risk types are significantly less accurate than the models for the individual types of risk.

We take the following approach to address this issue: We assume that the distributions of the marginal risks are known and fixed. This assumption is justified in many cases of practical interest. Moreover, risk aggregation is per definition not concerned with the computation of the marginal risks' distributions. Additionally, we take a probabilistic model for the dependence structure linking the marginal risks as given. Note that there are at least two different approaches in the literature to specify this *reference dependence structure*: The construction of copulas and factor models. The particular form of this reference model is not relevant for our approach as long as it allows us to generate random samples. Independently of the employed method, the choice of a reference dependence structure is typically subject to high uncertainty. Our contribution is to model the ambiguity with respect to the specified reference model, while fixing the marginal distributions. We address the following question in this paper:

> How can we account for model ambiguity with respect to a specific dependence structure when aggregating different risks?

We propose an intuitive approach to this problem: We compute the aggregated risk with respect to the worst-case dependence structure in a neighborhood around the specified reference dependence structure. For the construction of this neighborhood, we use transportation distances. These distance measures between probability distributions are flexible enough to capture different kinds of *model ambiguity*. At the same time, they allow us to generally derive numerical methods, which solve the corresponding problem of robust risk aggregation in reasonable time. To highlight some of the further merits of our approach, we are able to determine the worst-case dependence structure for a problem at hand. Hence, our method for robust risk *measurement* is arguably a useful tool also for risk *management* as it provides insights about which scenarios stress a given system the most. Moreover, it should be emphasized that our approach is restricted neither to a particular risk measure nor a particular aggregation function.[1]

In summary, the approach presented provides a flexible way to include model ambiguity in situations where a reference dependence structure is given and the marginals are fixed. It is generally applicable and computationally feasible. In the subsequent subsection, we outline our approach in some more details before discussing the related literature.

1.2 | Overview

We aim to evaluate

$$\int_{\mathbb{R}^d} f d\bar{\mu},$$

for some $f : \mathbb{R}^d \to \mathbb{R}$ in the presence of ambiguity with respect to the probability measure $\bar{\mu} \in \mathcal{P}(\mathbb{R}^d)$, where $\mathcal{P}(\mathbb{R}^d)$ denotes the set of all Borel probability measures on $\mathbb{R}^d$. In particular, we assume that the marginals $\bar{\mu}_1, \dots, \bar{\mu}_d$ of $\bar{\mu}$ are known and the ambiguity lies solely on the level of the dependence structure. Moreover, we assume a reference dependence structure, namely, the one implied by the reference measure $\bar{\mu}$, is given and that the degree of ambiguity with respect to the reference measure $\bar{\mu}$ can be modeled by the transportation distance d_c, which is defined in (2). Hence, we consider the following problem

$$\phi(f) := \max_{\substack{\mu \in \Pi(\bar{\mu}_1, \dots, \bar{\mu}_d) \\ d_c(\bar{\mu}, \mu) \leq \rho}} \int_{\mathbb{R}^d} f \, d\mu, \tag{1}$$

where the set $\Pi(\bar{\mu}_1, \dots, \bar{\mu}_d)$ consists of all $\mu \in \mathcal{P}(\mathbb{R}^d)$ satisfying $\mu_i = \bar{\mu}_i$ for all $i = 1, \dots, d$, where $\mu_i \in \mathcal{P}(\mathbb{R})$ and $\bar{\mu}_i \in \mathcal{P}(\mathbb{R})$ denote the ith marginal distributions of μ and $\bar{\mu}$. We fix a continuous function $c : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ such that $c(x, x) = 0$ for all $x \in \mathbb{R}$. The cost of transportation between $\bar{\mu}$ and μ in $\mathcal{P}(\mathbb{R}^d)$ with respect to the cost function c is defined as

$$d_c(\bar{\mu}, \mu) := \inf_{\pi \in \Pi(\bar{\mu}, \mu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) \pi(dx, dy), \tag{2}$$

where $\Pi(\bar{\mu}, \mu)$ denotes the set of all couplings of the marginals $\bar{\mu}$ and μ. For the cost function $c(x, y) = ||x - y||^p$ with $p \geq 1$, the mapping $d_c^{1/p}$ corresponds to the Wasserstein distance of order p.

The numerical methods to solve problem (1), which are developed in this paper, build on the following dual formulation of problem (1):

$$\inf_{\lambda \geq 0, \, h_i \in C_b(\mathbb{R})} \left\{ \rho\lambda + \sum_{i=1}^{d} \int_{\mathbb{R}} h_i \, d\bar{\mu}_i + \int_{\mathbb{R}^d} \sup_{y \in \mathbb{R}^d} \left[f(y) - \sum_{i=1}^{d} h_i(y_i) - \lambda c(x, y) \right] \bar{\mu}(dx) \right\}, \tag{3}$$

where $C_b(\mathbb{R})$ is the set of all continuous and bounded functions $h : \mathbb{R} \to \mathbb{R}$. This dual formulation was initially derived by Gao and Kleywegt (2017a). These authors show that strong duality holds, that is, problems (1) and (3) coincide, for upper semicontinuous functions $f : X \to \mathbb{R}$ satisfying

the growth condition $\sup_{x\in X} \frac{f(x)}{c(x,y_0)} < \infty$ for some $y_0 \in X$, where $X = X_1 \times \cdots \times X_d$ for possibly noncompact subsets $X_1, \dots, X_d$ of $\mathbb{R}$.

Theorem 2.1 extends the duality in the following aspects: First, the functions $f : X \to \mathbb{R}$ need not satisfy a growth condition that depends on the cost c. Our results allow for upper semicontinuous functions of bounded growth. Second, we can consider a space $X = X_1 \times \cdots \times X_d$, where X_i can be arbitrary polish spaces. We emphasize that the problem setting can therefore include an information structure where multivariate marginals are known and fixed. Finally, Theorem 2.1 extends the constraint $d_c(\bar{\mu}, \mu) \le \rho$ to a more general way of penalizing with respect to $d_c(\bar{\mu}, \mu)$.

We now turn to the question how the dual formulation (3) can be used to solve problem (1). One approach is to assume that the reference distribution $\bar{\mu}$ is a discrete distribution. In this context, Gao and Kleywegt (2017a) show that the dual problem (3) can be reformulated as a linear program under the following assumptions: First, the function f can be written as the maximum of affine functions. Second, the reference distribution $\bar{\mu}$ is given by an empirical distribution on n points $x^1, \dots, x^n$ in $\mathbb{R}^d$. Third, the cost function c has to be additively separable, that is, $c(x, y) = \sum_{i=1}^d c_i(x_i, y_i)$. For further details we refer to Corollary 2.5.

This linear programming approach is especially useful when only few observations are available to construct the reference distribution $\bar{\mu}$—a case where accounting for ambiguity with respect to the dependence structure is often required. Nevertheless, the assumptions under which problem (3) can be solved by means of linear programming exclude many cases of practical interest. Even in cases that linear programming is applicable, the resulting size of the linear program quickly becomes intractable in higher dimensions. Hence, this paper presents a generally applicable and computationally feasible method to numerically solve problem (3), which uses neural networks.

The basic idea is to use neural networks to parameterize the functions $h_i \in C_b(\mathbb{R})$ and then solve the resulting finite dimensional problem. Theoretically, such an approach is justified by the universal approximation properties of neural networks, see, for example, Hornik (1991).

To use neural networks, we first dualize the pointwise supremum inside the integral of (3). Under mild assumptions, this leads to

$$\inf_{\substack{\lambda\ge 0,\\ h_i\in C_b(\mathbb{R}),\, g\in C_b(\mathbb{R}^d):\\ g(x)\ge f(y)-\sum_{i=1}^d h_i(y_i)-\lambda c(x,y)}} \left\{ \lambda\rho + \sum_{i=1}^d \int_{\mathbb{R}} h_i \, d\bar{\mu}_i + \int_{\mathbb{R}^d} g \, d\bar{\mu} \right\}.$$

As the pointwise inequality constraint prevents a direct implementation with neural networks, the constraint is penalized. This is done by introducing a measure $\theta \in \mathcal{P}(\mathbb{R}^{2d})$, which we refer to as the *sampling measure*. Further, we are given a family of *penalty functions* $(\beta_\gamma)_{\gamma>0}$, which increase the accuracy of the penalization for increasing γ, for example, $\beta_\gamma(x) = \gamma \max\{0, x\}^2$. The resulting optimization problem reads

$$\begin{aligned} \phi_{\theta,\gamma}(f) := \inf_{\substack{\lambda\ge 0,\\ h_i\in C_b(\mathbb{R}),\, g\in C_b(\mathbb{R}^d)}} \Bigg\{ & \lambda\rho + \sum_{i=1}^d \int_{\mathbb{R}} h_i \, d\bar{\mu}_i + \int_{\mathbb{R}^d} g \, d\bar{\mu} \qquad (4) \\ & + \int_{\mathbb{R}^{2d}} \beta_\gamma\left(f(y) - \sum_{i=1}^d h_i(y_i) - \lambda c(x,y) - g(x) \right) \theta(dx, dy) \Bigg\}. \qquad (4) \end{aligned}$$

Before we develop numerical methods to evaluate $\phi_{\theta,\gamma}(f)$ and thereby approximate $\phi(f)$, we need to study the convergence

$$\phi_{\theta,\gamma}(f) \to \phi(f) \quad \text{for } \gamma \to \infty. \tag{5}$$

A sufficient condition for this convergence is derived in Proposition 2.7. We additionally give a general instance where this derived condition is satisfied. It states that (5) holds whenever the cost function c satisfies a mild growth condition and the sampling measure θ is the product measure between the reference measure and the respective marginals, that is, $\theta = \bar{\mu} \otimes (\bar{\mu}_1 \otimes \cdots \otimes \bar{\mu}_d)$.

Besides the optimal value of problem (1) also the corresponding optimizer is of interest. To this end, we develop duality for problem (4). This duality leads to a simple formula to obtain an approximate optimizer of the initial problem (1) once the dual formulation (4) is solved. It shows that any optimizer $(\lambda^\star, (h_i^\star)_{i=1,\dots,d}, g^\star)$ of (4) gives an approximate optimizer $\mu^\star$ of (1) by setting $\mu^\star$ equal to the second marginal of $\pi^\star$, where $\pi^\star$ is defined by the Radon–Nikodym derivative

$$\frac{d\pi^\star}{d\theta}(x,y) := \beta_\gamma'\left(f(y) - g^\star(x) - \sum_{i=1}^{d} h_i^\star(y_i) - \lambda^\star c(x,y) \right). \tag{6}$$

Problem $\phi_{\theta,\gamma}(f)$ fits into the standard framework in which neural networks can be applied to parameterize the functions $h_i \in C_b(\mathbb{R})$ and $g \in C_b(\mathbb{R}^d)$. We justify this parameterization theoretically by giving conditions under which the approximation error vanishes for an infinite-size neural network. In Section 3, we give details concerning the numerical solution of $\phi_{\theta,\gamma}(f)$ using neural networks, which encompasses the choice of the neural network structure, hyperparameters, and optimization method.

This approach based on neural networks is the main reason to derive and study the penalized problem (4). Nonetheless, problem (4) is interesting in its own right and by no means limited to the application of neural networks: it may be efficiently solved using advanced first-order methods (see, e.g., Nesterov, 2012). We thank an anonymous referee for pointing this out to us.

The remainder of the paper is structured as follows. In Subsection 1.3, we provide an overview of the relevant literature. Our main results can be found in Section 2, which consists of three parts: First, we state and prove the general form of the duality between (1) and (3) and derive some implications thereof. In the second part of Section 2, we study the penalization introduced in Equation (4). Third, we give conditions under which $\phi_{\theta,\gamma}(f)$ can be approximated with neural networks. Section 3 gives implementation details. Section 4 is devoted to three toy examples, which aim to shed some light on the developed concepts. Finally in Section 5, the acquired techniques are applied to a real-world example. We thereby demonstrate how to implement robust risk aggregation with neural networks in practice.

1.3 | Related literature

There are three different strings of literature, which are relevant in the present context: First, literature on risk aggregation; second, literature on model ambiguity and particularly on ambiguity sets constructed using the Wasserstein distance; third, recent application of neural networks in finance and related optimization problems.

1.3.1 | Risk aggregation

In Section 5, we motivate from an applied point of view why there is interest in risk bounds for the sum of losses of which the marginal distributions are known. The theoretical interest in this topic started with the following questions: How can one compute bounds for the distribution function of a sum of two random variables when the marginal distributions are fixed? This problem was solved in 1982 by Makarov (1982) and Rüschendorf (1982). Starting with the work of Embrechts and Puccetti (2006) more than 20 years later, the higher dimensional version of this problem was studied extensively due to its relevance for risk management. We refer to Embrechts, Wang, and Wang (2015) and Puccetti and Wang (2015) for an overview of the developments concerning *risk aggregation under dependence uncertainty*, as this problem was coined. Let us mention that Puccetti and Rüschendorf (2012b) introduced the so-called *rearrangement algorithm*, which is a fast procedure to numerically compute the bounds of interest. Applying this algorithm to real-world examples demonstrates a conceptual drawback of the assumption that no information concerning the dependence of the marginal risk is available: The implied lower and upper bound for the aggregated risk are impractically far apart.

Hence, some authors recently tried to overcome this drawback and to come up with more realistic bounds by including partial information about the dependence structure. For instance, Puccetti and Rüschendorf (2012a) discuss how positive, negative, or independence information influence the above risk bounds; Bernard, Rüschendorf, and Vanduffel (2017) derive risk bounds with constraints on the variance of the aggregated risk; Bernard, Rüschendorf, Vanduffel, and Wang (2017) consider partially specified factor models for the dependence structure. The interested reader is referred to Rüschendorf (2017) for a recent review of these and related approaches. Finally, we want to point out the intriguing contribution by Lux and Papapantoleon (2016). These authors provide a framework that allows them to derive value at risk (VaR)-bounds if (a) extreme value information is available, (b) the copula linking the marginals is known on a subset of its domain and (c) the latter copula lies in the neighborhood of a reference copula as measured by a statistical distance.

As our paper aims to contribute to this string of literature, let us point out that the latter mentioned type of partial information about the dependence structure used in Lux and Papapantoleon (2016) is similar in spirit to our approach. We emphasize that Lux and Papapantoleon use statistical distances that are different to the transportation distance d_c defined in the previous subsection.

1.3.2 | Model ambiguity

There is an obvious connection of problem (1), which is studied in this paper, with the following minimax stochastic optimization problem:

$$\min_{x\in\mathbb{X}} \max_{Q\in\mathcal{Q}} \mathbb{E}^Q[f(x,\xi)], \tag{7}$$

where $\mathbb{X}\subset\mathbb{R}^m$, $f:\mathbb{R}^m\times\Xi\to\mathbb{R}$, ξ is a random vector whose distribution Q is supported on $\Xi\subset\mathbb{R}^d$ and $\mathcal{Q}$ is a nonempty set of probability distributions, referred to as *ambiguity set*. Problems of this form recently became known as distributionally robust stochastic optimization problems. As pointed out by Shapiro (2017), there are two natural and somewhat different approaches to constructing the ambiguity set $\mathcal{Q}$. On the one hand, ambiguity sets have been defined by moment

constraints, see Delage and Ye (2010) and references therein. An alternative approach is to assume a reference probability distribution $\bar{Q}$ is given and define the ambiguity set by all distributions that are in the neighborhood of $\bar{Q}$ as measured by a statistical distance. To the best of our knowledge two distinct choices of this statistical distance have been established in the literature: The ϕ-divergence and the Wasserstein distance. Concerning ambiguity sets constructed using the ϕ-divergence we refer to Bayraksan and Love (2015), and references therein. In the following, we focus on approaches that rely on the Wasserstein distance to account for model ambiguity. Pflug and Wozabal (2007) were the first to study these particular ambiguity sets. Esfahani and Kuhn (2018) showed that distributionally robust stochastic optimization problems over Wasserstein balls centered around a discrete reference distribution possess a tractable reformulation: under mild assumptions these problems belong to the same complexity class as their nonrobust counterparts. The duality result driving this insight was also proven by Blanchet and Murthy (2019), Gao and Kleywegt (2016), and Bartl, Drapeau, and Tangpi (2019) based on different techniques and assumptions. These contributions indicate that distributionally robust stochastic optimization using the Wasserstein distance developed into an active field of research in recent years. For instance, Zhao and Guan (2018) and Hanasusanto and Kuhn (2018) adapted similar ideas in the context of two-stage stochastic programming and Chen, Yu, and Haskell (2019) and Yang (2017) study distributionally robust Markov decision processes using the Wasserstein distance. Obloj and Wiesel (2018) analyze a robust estimation method for superhedging prices relying on a Wasserstein ball around the empirical measure.

Most relevant in the context of our paper are the following two references: Gao and Kleywegt (2017b) put two Wasserstein-type constraints on the probability distribution Q in (7): Q has to be close in Wasserstein distance to a reference distribution $\bar{Q}$, while the dependence structure implied by Q has to be close, again in Wasserstein distance, to a specified reference dependence structure. In their follow-up paper, Gao and Kleywegt (2017a) consider problem (1) in the context of stochastic optimization, that is, in the framework (7). The contribution of this paper, that is, their duality result and the linear programming (LP) formulation, is already reviewed in the above overview. In addition, the authors provide numerical experiments in portfolio selection and nonparametric density estimation.

1.3.3 | Neural networks in finance and optimization

Applications of neural networks have vastly increased in recent years. Most of the popularity arose from successes of neural networks related to data representation tasks, for example, related to pattern recognition, image classification, or task-specific artificial intelligence. In contrast to such an utilization, neural networks have also been applied strictly as a tool to solve certain optimization problems. This is the way we use neural networks in this paper, and they have found similar uses in various areas related to finance. Among others, they were applied to solve high-dimensional partial differential equations and stochastic differential equations (see, e.g., Beck, Becker, Grohs, Jaafari, & Jentzen, 2018; Berner, Grohs, & Jentzen, 2018; Weinan, Han, & Jentzen, 2017) as well as backward stochastic differential equation (Henry-Labordere, 2017), in optimal stopping (Becker, Cheridito, & Jentzen, 2019), optimal hedging with respect to a risk measure (Buehler, Gonon, Teichmann, & Wood, 2019), and superhedging (Eckstein & Kupper, 2019).

For more classical learning tasks where neural networks are applied, ideas from optimal transport and distributional robustness are also used. Although the settings are often quite different in nature to the one in this paper, the optimization problems that are eventually implemented are

nevertheless similar. Most related to the current paper are settings in which optimal transport type of constraints are solved via a penalization or regularization method. Examples include generative models for images (see, e.g., Gulrajani, Ahmed, Arjovsky, Dumoulin, & Courville, 2017; Roth, Lucchi, Nowozin, & Hofmann, 2017), optimal transport and calculation of barycenters for images (see, e.g., Seguy et al., 2017), martingale optimal transport (see, e.g., Henry-Labordere, 2019), or distributional robustness methods applied to learning tasks (see, e.g., Blanchet, Kang, & Murthy, 2016; Gao, Chen, & Kleywegt, 2017).

2 | RESULTS

2.1 | Duality

Let $X = X_1 \times X_2 \times \cdots \times X_d$ be a Polish space, and denote by $\mathcal{P}(X)$ the set of all Borel probability measures on X. Throughout, we fix a reference probability measure $\bar{\mu} \in \mathcal{P}(X)$. For $i = 1, \dots, d$, we denote by $\mu_i := \mu \circ \mathrm{pr}_i^{-1}$ the ith marginal of $\mu \in \mathcal{P}(X)$, where $\mathrm{pr}_i : X \to X_i$ is the projection $\mathrm{pr}_i(x) := x_i$. Further, let $\kappa : X \to [1, \infty)$ be a growth function of the form $\kappa(x_1, \dots, x_d) = \sum_{i=1}^d \kappa_i(x_i)$, where each $\kappa_i : X_i \to [1, \infty)$ is continuous and satisfies $\int_{X_i} \kappa_i \, d\bar{\mu}_i < \infty$. We further assume one of the following: Either κ has compact sublevel sets,[2] or $X_i = \mathbb{R}^{d_i}$ for all $i = 1, \dots, d$. Denote by $C_\kappa(X)$ and $U_\kappa(X)$ the spaces of all continuous, respectively, upper semicontinuous functions $f : X \to \mathbb{R}$ such that f/κ is bounded. Recall that $C_b(X)$ denotes the set of all continuous and bounded functions on X.

In the following, we fix a continuous function $c : X \times X \to [0, \infty)$ such that $c(x, x) = 0$ for all $x \in X$. The cost of transportation between $\bar{\mu}$ and μ in $\mathcal{P}(X)$ with respect to the cost function c is defined as

$$d_c(\bar{\mu}, \mu) := \inf_{\pi \in \Pi(\bar{\mu}, \mu)} \int_{X \times X} c(x, y) \pi(dx, dy), \tag{8}$$

where $\Pi(\bar{\mu}_1, \dots, \bar{\mu}_d)$ denotes the set of all $\mu \in \mathcal{P}(X)$ such that $\mu_i = \bar{\mu}_i$ for all $i = 1, \dots, d$. The elements in $\Pi(\bar{\mu}_1, \dots, \bar{\mu}_d)$ are referred to as couplings of the marginals $\bar{\mu}_1, \dots, \bar{\mu}_d$. Although the computation of the convex conjugate in the following result relies on Bartl, Drapeau, et al. (2019), we do not need their growth condition on the cost function c. The main reason we do not require this condition is that continuity from above of the functional (9)—which corresponds to tightness of the considered set of measures—is already obtained by the imposed marginal constraints.

Theorem 2.1. *For every convex and lower semicontinuous function $\varphi : [0, \infty] \to [0, \infty]$ such that $\varphi(0) = 0$ and $\varphi(\infty) = \infty$, and all $f \in U_\kappa(X)$, it holds that*

$$\begin{aligned} &\max_{\mu \in \Pi(\bar{\mu}_1, \dots, \bar{\mu}_d)} \left\{ \int_X f \, d\mu - \varphi(d_c(\bar{\mu}, \mu)) \right\} \\ &= \inf_{\lambda \geq 0,\, h_i \in C_{\kappa_i}(X_i)} \left\{ \varphi^*(\lambda) + \sum_{i=1}^d \int_{X_i} h_i d\bar{\mu}_i + \int_X \sup_{y \in X} \left[f(y) - \sum_{i=1}^d h_i(y_i) - \lambda c(x, y) \right] \bar{\mu}(dx) \right\}, \end{aligned} \tag{9}$$

where φ^* denotes the convex conjugate of φ, that is, $\varphi^*(\lambda) = \sup_{x \geq 0} \{\lambda x - \varphi(x)\}$.

Proof. 1. Define the optimal transport functional $\psi_1 : C_\kappa(X) \to \mathbb{R}$ by

$$\psi_1(f) := \inf \left\{ \sum_{i=1}^d \int_{X_i} h_i \, d\bar{\mu}_i \, : \, h_i \in C_{\kappa_i}(X_i) \text{ such that } \oplus_{i=1}^d h_i \geq f \right\},$$

where $\oplus_{i=1}^d h_i : X \to \mathbb{R}$ is defined as $\oplus_{i=1}^d h_i(x) := \sum_{i=1}^d h_i(x_i)$. We show that ψ_1 is continuous from above on $C_\kappa(X)$, that is, for every sequence (f^n) in $C_\kappa(X)$ such that $f^n \downarrow f \in C_\kappa(X)$ one has $\psi_1(f^n) \downarrow \psi_1(f)$. Fix $\varepsilon > 0$ and $h_i \in C_{\kappa_i}(X_i)$ such that $\oplus_{i=1}^d h_i \geq f$ and $\psi_1(f) + \frac{\varepsilon}{3} \geq \sum_{i=1}^d \int_{X_i} h_i \, d\bar{\mu}_i$. As $f^1 \in C_\kappa(X)$ and $h_i \in C_{\kappa_i}(X_i)$, there exists a constant $M > 0$ such that $|f^1| \leq M\kappa$ and $|h_i| \leq M\kappa_i$. By assumption, $\int_{X_i} \kappa_i \, d\bar{\mu}_i < +\infty$ for all $i = 1, \dots, d$. We now show that $\psi_1(f^{n_0}) \leq \psi_1(f) + \varepsilon$, which we do separately depending on whether we assume κ has compact sublevel sets, or $X_i = \mathbb{R}^{d_i}$.

- Let κ have compact sublevel sets. Choose $z > 0$ such that $\sum_{i=1}^d \int_{X_i} 4M(\kappa_i - \frac{z}{d})^+ \, d\bar{\mu}_i \leq \frac{\varepsilon}{3}$. By Dini's lemma, there exists $n_0 \in \mathbb{N}$ such that $f^{n_0} \leq \oplus_{i=1}^d h_i + \frac{\varepsilon}{3}$ on the compact $\{\kappa \leq 2z\}$. As it further holds $\kappa \mathbb{1}_{\{\kappa > 2z\}} \leq 2(\kappa - z)^+ \leq 2 \oplus_{i=1}^d (\kappa_i - \frac{z}{d})^+$, one obtains

$$\begin{aligned} f^{n_0} &= \mathbb{1}_{\{\kappa \leq 2z\}} f^{n_0} + \mathbb{1}_{\{\kappa > 2z\}} f^{n_0} \\ &\leq \mathbb{1}_{\{\kappa \leq 2z\}} \oplus_{i=1}^d h_i + \mathbb{1}_{\{\kappa > 2z\}} f^{n_0} + \frac{\varepsilon}{3} \\ &= \oplus_{i=1}^d h_i + \mathbb{1}_{\{\kappa > 2z\}} \big(f^{n_0} - \oplus_{i=1}^d h_i \big) + \frac{\varepsilon}{3} \\ &\leq \oplus_{i=1}^d h_i + \mathbb{1}_{\{\kappa > 2z\}} 2M\kappa + \frac{\varepsilon}{3} \\ &\leq \oplus_{i=1}^d \left(h_i + 4M \left(\kappa_i - \frac{z}{d} \right)^+ \right) + \frac{\varepsilon}{3} \end{aligned}$$

  and hence $\psi_1(f^{n_0}) \leq \sum_{i=1}^d \int_{X_i} h_i + 4M(\kappa_i - \frac{z}{d})^+ \, d\bar{\mu}_i + \frac{\varepsilon}{3} \leq \psi_1(f) + \varepsilon$.

- Let $X_i = \mathbb{R}^{d_i}$. Choose $z > 0$ such that $\sum_{i=1}^d \int_{X_i} 4M(\kappa_i - \frac{z}{d})^+ \, d\bar{\mu}_i \leq \frac{\varepsilon}{6}$. Choose $R_i > 0$ such that $\sum_{i=1}^d \bar{\mu}_i(\overline{B(0, R_i)}^c) \cdot 4Mz < \frac{\varepsilon}{6}$, where $B(0, r)$ is the open Euclidean ball around 0 of radius r. By Dini's lemma, there exists $n_0 \in \mathbb{N}$ such that $f^{n_0} \leq \oplus_{i=1}^d h_i + \frac{\varepsilon}{3}$ on the compact $K := K_1 \times \cdots \times K_d := \overline{B(0, R_1 + 2)} \times \cdots \times \overline{B(0, R_d + 2)}$. As $\mathbb{1}_{B(0,R_i+1)^c}$ is upper semicontinuous, we can find continuous and bounded functions g_i such that $\mathbb{1}_{B(0,R_i+1)^c} \leq g_i$ and $\sum_{i=1}^d \int_{X_i} g_i \, d\bar{\mu}_i \cdot 4Mz < \frac{\varepsilon}{6}$ (as g_i approximates $\mathbb{1}_{B(0,R_i+1)^c}$ and $\mathbb{1}_{B(0,R_i+1)^c} \leq \mathbb{1}_{\overline{B(0,R_i)}^c}$). With some of the same steps as in the case where κ has compact sublevel sets, one obtains

$$\begin{aligned} f^{n_0} &= \mathbb{1}_K f^{n_0} + \mathbb{1}_{K^c} f^{n_0} \\ &\leq \oplus_{i=1}^d h_i + \mathbb{1}_{K^c} \big(f^{n_0} - \oplus_{i=1}^d h_i \big) + \frac{\varepsilon}{3} \end{aligned}$$

$$\leq \oplus_{i=1}^{d} h_i + \mathbb{1}_{K^c}\mathbb{1}_{\{\kappa>2z\}} 2M\kappa + \mathbb{1}_{K^c}\mathbb{1}_{\{\kappa\leq 2z\}} 2M\kappa + \frac{\varepsilon}{3}$$

$$\leq \oplus_{i=1}^{d} \left(h_i + 4M\left(\kappa_i - \frac{z}{d}\right)^+ \right) + \mathbb{1}_{K^c} 4Mz + \frac{\varepsilon}{3}$$

$$\leq \oplus_{i=1}^{d} \left(h_i + 4M\left(\kappa_i - \frac{z}{d}\right)^+ \right) + (\oplus_{i=1}^{d} \mathbb{1}_{K_i^c}) 4Mz + \frac{\varepsilon}{3}$$

$$\leq \oplus_{i=1}^{d} \left(h_i + 4M\left(\kappa_i - \frac{z}{d}\right)^+ + 4Mz \cdot g_i \right) + \frac{\varepsilon}{3}$$

and hence $\psi_1(f^{n_0}) \leq \sum_{i=1}^{d} \int_{X_i} h_i + 4M(\kappa_i - \frac{z}{d})^+ + 4Mz \cdot g_i \, d\bar{\mu}_i + \frac{\varepsilon}{3} \leq \psi_1(f) + \varepsilon$.

This shows that ψ_1 is continuous from above on $C_\kappa(X)$. Moreover, its convex conjugate is given by

$$\psi^*_{1,C_\kappa}(\mu) = \sup_{f \in C_\kappa(X)} \left(\int_X f \, d\mu - \inf_{\substack{h_i \in C_{\kappa_i}(X_i) \\ \oplus_{i=1}^{d} h_i \geq f}} \sum_{i=1}^{d} \int_{X_i} h_i \, d\bar{\mu}_i \right)$$

$$= \sup_{h_i \in C_{\kappa_i}(X_i)} \sup_{\substack{f \in C_\kappa(X) \\ \oplus_{i=1}^{d} h_i \geq f}} \left(\int_X f \, d\mu - \sum_{i=1}^{d} \int_{X_i} h_i \, d\bar{\mu}_i \right)$$

$$= \sup_{h_i \in C_{\kappa_i}(X_i)} \sum_{i=1}^{d} \left(\int_X h_i \, d\mu - \int_{X_i} h_i \, d\bar{\mu}_i \right) = \begin{cases} 0 & \text{if } \mu \in \Pi(\bar{\mu}_1, \cdots, \bar{\mu}_d) \\ +\infty & \text{else} \end{cases} \tag{10}$$

for all $\mu \in \mathcal{P}_\kappa(X)$, where $\mathcal{P}_\kappa(X)$ denotes the set of all $\mu \in \mathcal{P}(X)$ such that $\kappa \in L^1(\mu)$. Note that $\Pi(\bar{\mu}_1, \ldots, \bar{\mu}_d) \subset \mathcal{P}_\kappa(X)$.2. Define $\psi_2 : C_\kappa(X) \to \mathbb{R} \cup \{+\infty\}$ by

$$\psi_2(f) := \inf_{\lambda \geq 0} \left(\varphi^*(\lambda) + \int_X \sup_{y \in X} [f(y) - \lambda c(x, y)] \, \bar{\mu}(dx) \right).$$

By definition ψ_2 is convex and increasing. Further, as $\inf_{\lambda \geq 0} \varphi^*(\lambda) = \varphi^*(0) = 0$ and $f^{\lambda c}(x) := \sup_{y \in X}\{f(y) - \lambda c(x, y)\} \geq f(x)$ for all $\lambda \geq 0$, it follows that

$$\psi_2(f) \geq \inf_{\lambda \geq 0} \left(\varphi^*(\lambda) + \int_X f \, d\bar{\mu} \right) > -\infty$$

for all $f \in C_\kappa(X)$, where we use that $f \in L^1(\bar{\mu})$. For the convex conjugates one has

$$\psi^*_{2,C_\kappa}(\mu) := \sup_{f \in C_\kappa(X)} \left(\int_X f \, d\mu - \psi_2(f) \right)$$

$$= \sup_{f \in U_\kappa(X)} \left(\int_X f \, d\mu - \psi_2(f) \right) =: \psi^*_{2,U_\kappa}(\mu) = \varphi(d_c(\bar{\mu}, \mu)) \tag{11}$$

for all $\mu \in \mathcal{P}_\kappa(X)$. Indeed, for every $\mu \in \mathcal{P}_\kappa(X)$ one has

$$\psi^*_{2,U_\kappa}(\mu) \geq \psi^*_{2,C_\kappa}(\mu) \geq \psi^*_{2,C_b}(\mu) = \varphi(d_c(\bar{\mu}, \mu)),$$

where the last equality is shown in Bartl, Drapeau, et al. (2019, Proof of Theorem 2.4, Step 4), notably without using the growth condition for c imposed in Bartl, Drapeau, et al. (2019). It remains to show that $\psi^*_{2,U_\kappa}(\mu) \leq \varphi(d_c(\bar{\mu}, \mu))$. As $\varphi(\infty) = \infty$, the case $d_c(\bar{\mu}, \mu) = \infty$ is obvious. Suppose $d_c(\bar{\mu}, \mu) < +\infty$. Note that $\int_X f^{\lambda c} \, d\bar{\mu}$ is well defined as $f^{\lambda c} \geq f \in L^1(\bar{\mu})$, so that the negative part of the integral is finite. Further, by eliminating redundant choices in supremum and infimum of the convex conjugate, one obtains

$$\psi^*_{2,U_\kappa}(\mu) = \sup_{\substack{f \in U_\kappa(X) \\ \psi_2(f) < \infty}} \left\{ \int_X f d\mu - \inf_{\substack{\lambda \geq 0, \varphi^*(\lambda) < \infty, \\ \int_X f^{\lambda c} d\bar{\mu} < \infty}} \left(\varphi^*(\lambda) + \int_X f^{\lambda c} d\bar{\mu} \right) \right\}.$$

For every $\varepsilon > 0$, $f \in U_\kappa(X)$ and $\lambda \geq 0$ such that $\psi_2(f) < +\infty$, $\varphi^*(\lambda) < +\infty$, $\int_X f^{\lambda c} d\bar{\mu} < +\infty$, it follows that $\int_X f \, d\mu$, $\varphi^*(\lambda)$ and $\int_X f^{\lambda c} \, d\bar{\mu}$ are real numbers, so that

$$\begin{aligned}
&\int_X f \, d\mu - \varphi^*(\lambda) - \int_X f^{\lambda c} \, d\bar{\mu} - \varepsilon \leq \int_X f \, d\mu - \lambda d_c(\bar{\mu}, \mu) + \varphi(d_c(\bar{\mu}, \mu)) - \int_X f^{\lambda c} d\bar{\mu} - \varepsilon \\
&\quad \leq \int_{X \times X} f(y) \, \pi(dx, dy) - \int_{X \times X} \lambda c(x, y) \, \pi(dx, dy) - \int_{X \times X} f^{\lambda c}(x) \, \pi(dx, dy) + \varphi(d_c(\bar{\mu}, \mu)) \\
&\quad \leq \int_{X \times X} \left[\lambda c(x, y) + f^{\lambda c}(x) - \lambda c(x, y) - f^{\lambda c}(x) \right] \pi(dx, dy) + \varphi(d_c(\bar{\mu}, \mu)) \\
&\quad = \varphi(d_c(\bar{\mu}, \mu)),
\end{aligned}$$

where $\pi \in \Pi(\bar{\mu}, \mu)$ is such that $\lambda d_c(\bar{\mu}, \mu) + \varepsilon \geq \int_{X \times X} \lambda c \, d\pi$, and where we used that $\varphi^*(\lambda) \geq \lambda d_c(\bar{\mu}, \mu) - \varphi(d_c(\bar{\mu}, \mu))$ and $f(y) \leq \lambda c(x, y) + f^{\lambda c}(x)$. Taking the supremum over all such f and λ implies $\psi^*_{2,U_\kappa}(\mu) \leq \varphi(d_c(\bar{\mu}, \mu))$.3. For $f \in U_\kappa(X)$ define the convolution

$$\begin{aligned}
\psi(f) &:= \inf_{g \in C_\kappa(X)} \{ \psi_1(g) + \psi_2(f - g) \} \\
&= \inf_{\lambda \geq 0, \, h_i \in C_b(X_i)} \left\{ \varphi^*(\lambda) + \sum_{i=1}^d \int_{X_i} h_i d\bar{\mu}_i + \int_X \sup_{y \in X} \left[f(y) - \sum_{i=1}^d h_i(y_i) - \lambda c(x, y) \right] \bar{\mu}(dx) \right\}.
\end{aligned}$$

For the associated convex conjugates, it follows from (10) and (11) that

$$\begin{aligned}
\psi^*_{C_\kappa}(\mu) &= \sup_{f\in C_\kappa(X)}\sup_{g\in C_\kappa(X)}\left(\int_X f\,d\mu - \psi_1(g) - \psi_2(f-g)\right)\\
&= \sup_{g\in C_\kappa(X)}\left(\int_X g\,d\mu - \psi_1(g)\right) + \sup_{f\in C_\kappa(X)}\left(\int_X f\,d\mu - \psi_2(f)\right) = \psi^*_{1,C_\kappa}(\mu) + \psi^*_{2,C_\kappa}(\mu)\\
&= \psi^*_{1,C_\kappa}(\mu) + \psi^*_{2,U_\kappa}(\mu) = \sup_{g\in C_\kappa(X)}\left(\int_X g\,d\mu - \psi_1(g)\right) + \sup_{f\in U_\kappa(X)}\left(\int_X f\,d\mu - \psi_2(f)\right)\\
&= \sup_{f\in U_\kappa(X)}\sup_{g\in C_\kappa(X)}\left(\int_X f\,d\mu - \psi_1(g) - \psi_2(f-g)\right)\\
&= \psi^*_{U_\kappa}(\mu) = \begin{cases}\phi\left(d_c(\bar\mu,\mu)\right) & \text{if } \mu\in\Pi(\bar\mu_1,\cdots,\bar\mu_d)\\ +\infty & \text{else}\end{cases}
\end{aligned}$$

for all $\mu\in\mathcal{P}_\kappa(X)$.4. For every $f\in U_\kappa(X)$ one has

$$\psi(f) \geq \int_X f\,d\bar\mu - \psi^*_{U_\kappa}(\bar\mu) = \int_X f\,d\bar\mu > -\infty$$

as $\psi^*_{U_\kappa}(\bar\mu) = \varphi(d_c(\bar\mu,\bar\mu)) = \varphi(0) = 0$ and $f\in L^1(\mu)$. This shows that $\psi : U_\kappa(X)\to\mathbb{R}$. By definition, ψ is convex and increasing. Moreover, ψ is continuous from above on $C_\kappa(X)$, as for every sequence (f^n) in $C_\kappa(X)$ such that $f^n\downarrow 0$ one has

$$\begin{aligned}
\inf_{n\in\mathbb{N}}\psi(f^n) &= \inf_{n\in\mathbb{N}}\inf_{g\in C_\kappa(X)}\left(\psi_1(g) + \psi_2(f^n - g)\right)\\
&= \inf_{g\in C_\kappa(X)}\inf_{n\in\mathbb{N}}\left(\psi_1(f^n - g) + \psi_2(g)\right)\\
&= \inf_{g\in C_\kappa(X)}\left(\psi_1(-g) + \psi_2(g)\right) = \psi(0),
\end{aligned}$$

where we use that ψ_1 is continuous from above on $C_\kappa(X)$ by the first step. As also $\psi^*_{C_\kappa} = \psi^*_{U_\kappa}$ on $\mathcal{P}_\kappa(X)$ by the third step, it follows from Bartl, Cheridito, and Kupper (2019, Theorem 2.2, Proposition 2.3) that ψ has the dual representation

$$\psi(f) = \max_{\mu\in\mathcal{P}_\kappa(X)}\left\{\int_X f\,d\mu - \psi^*_{C_\kappa}(\mu)\right\} = \max_{\mu\in\Pi(\bar\mu_1,\ldots,\bar\mu_d)}\left\{\int_X f\,d\mu - \varphi(d_c(\bar\mu,\mu))\right\}$$

for all $f\in U_\kappa(X)$. □

Corollary 2.2. *For every $f \in U_\kappa(X)$, one has*

$$\max_{\substack{\mu\in\Pi(\bar\mu_1,\dots,\bar\mu_d)\\ d_c(\bar\mu,\mu)\le\rho}} \int_X f\, d\mu \tag{12}$$

$$= \inf_{\lambda\ge 0,\, h_i\in C_{\kappa_i}(X_i)} \left\{ \rho\lambda + \sum_{i=1}^d \int_{X_i} h_i\, d\bar\mu_i + \int_X \sup_{y\in X}\left[f(y) - \sum_{i=1}^d h_i(y_i) - \lambda c(x,y)\right]\bar\mu(dx)\right\} \tag{13}$$

for each radius $\rho \ge 0$.

Proof. This follows directly from Theorem 2.1 for φ given by $\varphi(x) = 0$ if $x \le \rho$ and $\varphi(x) = +\infty$ if $x > \rho$. In that case, the conjugate is given by $\varphi^*(\lambda) = \rho\lambda$. □

Remark 2.3. Let us comment on the interpretation of the dual problem (13): Roughly speaking, in case $\rho = \infty$, the above result collapses to the duality of multi-marginal optimal transport. On the other hand, if $\rho = 0$, both the primal problem (12) and the dual problem (13) reduce to $\int f\, d\bar\mu$. Finally, if one drops the constraint $\mu \in \Pi(\bar\mu_1, \dots, \bar\mu_d)$ in the primal formulation (12), the functions $h_1 = h_2 = \cdots = 0$.

From a computational point of view, the penalty function $\varphi(x) = x$ is of particular interest as the optimization in Theorem 2.1 over the Lagrange multiplier λ disappears.

Corollary 2.4. *For every $f \in U_\kappa(X)$, one has*

$$\max_{\mu\in\Pi(\bar\mu_1,\dots,\bar\mu_d)} \left\{ \int_X f\, d\mu - d_c(\bar\mu,\mu)\right\}$$

$$= \inf_{h_i\in C_{\kappa_i}(X_i)} \left\{ \sum_{i=1}^d \int_{X_i} h_i d\bar\mu_i + \int_X \sup_{y\in X}\left[f(y) - \sum_{i=1}^d h_i(y_i) - c(x,y)\right]\bar\mu(dx)\right\}.$$

Proof. This follows from Theorem 2.1 for $\varphi(y) = y$. Indeed, as the convex conjugate is given by $\varphi^*(\lambda) = 0$ for $0 \le \lambda \le 1$ and $\varphi^*(\lambda) = +\infty$ for $\lambda > 1$, the infimum in Theorem 2.1 is attained at $\lambda = 1$. □

Corollary 2.5 (Gao & Kleywegt, 2017a). *Let $f(x) = \max_{1\le m\le M}(a^m)^\top x + b^m$ for $x \in \mathbb{R}^d$, $a^m \in \mathbb{R}^d$, and $b^m \in \mathbb{R}$. Let $\bar\mu = \frac{1}{n}\sum_{j=1}^n \delta_{x^j}$ for given points $x^1, \dots, x^n$ in $\mathbb{R}^d$.*[3] *Let the same points $x^1, \dots, x^n$ define the sets X_i, that is, $X_i = \{x_i^1, \dots, x_i^n\}$ and $X = X_1 \times \cdots \times X_d$. Let the cost function c be additively separable, that is, $c(x,y) = \sum_{i=1}^d c_i(x_i, y_i)$. Then, the dual problem (13) is equivalent to the linear program*

$$\min_{\lambda, h_i(j), g(j), u_i(j,m)} \left\{ \lambda\rho + \frac{1}{n}\sum_{i=1}^d\sum_{j=1}^n h_i(j) + \frac{1}{n}\sum_{j=1}^n g(j)\right\} \tag{14}$$

$$s.t.: g(j) \geq b^m + \sum_{i=1}^{d} u_i(j,m) \quad j = 1,\dots,n;\ m = 1,\dots,M \tag{15}$$

$$u_i(j,m) \geq a_i^m x_i^k - h_i(k) - \lambda c_i(x_i^j, x_i^k) \quad i = 1,\dots,d;\ m = 1,\dots,M;\ j,k = 1,\dots,n \tag{16}$$

$$\lambda \geq 0. \tag{17}$$

The proof can be found in Gao and Kleywegt (2017a). For the convenience of the reader, we also present a direct proof of Corollary 2.5.

Proof. Due to the assumptions that $X_i = \{x_i^1,\dots,x_i^n\}$ and $\bar{\mu} = \frac{1}{n}\sum_{j=1}^n \delta_{x^j}$, the term $\int_{X_i} h_i \, d\bar{\mu}_i$ in (13) can be written as $\frac{1}{n}\sum_{j=1}^n h_i(x^j)$ and we shall use that $h_i(x^j) = h_i(x_i^j)$. Combing these facts with the assumption $c(x,y) = \sum_{i=1}^d c_i(x_i,y_i)$, the dual problem (13) can be reformulated as

$$\min_{\lambda\geq 0,\, h_i} \left\{ \lambda\rho + \frac{1}{n}\sum_{i=1}^{d}\sum_{j=1}^{n} h_i(x^j) + \frac{1}{n}\sum_{j=1}^{n} \max_{y\in X} \left\{ \max_{1\leq m\leq M} \left(\sum_{i=1}^{d} a_i^m y_i + b^m \right) - \sum_{i=1}^{d} h_i(y) - \lambda c(x^j, y) \right\} \right\}$$

$$= \min_{\lambda\geq 0,\, h_i} \left\{ \lambda\rho + \frac{1}{n}\sum_{i=1}^{d}\sum_{j=1}^{n} h_i(x^j) + \frac{1}{n}\sum_{j=1}^{n} \max_{1\leq m\leq M} \left\{ \max_{y\in X} b^m + \sum_{i=1}^{d} \left(a_i^m y_i - h_i(y_i) - \lambda c_i(x_i^j, y_i) \right) \right\} \right\}.$$

The assumption $X = X_1 \times \cdots \times X_d$ implies that for any $y \in X$ we can find indices $k_1,\dots,k_d$ with $1 \leq k_i \leq n$ for $i = 1,\dots,d$ such that $y = (x_1^{k_1},\dots,x_d^{k_d})$. We introduce the auxiliary variables $g(j) \in \mathbb{R}$ for $j = 1,\dots,n$ and write the above problem as

$$\min_{\lambda\geq 0,\, h_i,\, g(j)} \left\{ \lambda\rho + \frac{1}{n}\sum_{i=1}^{d}\sum_{j=1}^{n} h_i(x^j) + \frac{1}{n}\sum_{j=1}^{n} g(j) \,:\, g(j) \geq \max_{k_1,\cdots,k_d} b^m \right.$$

$$\left. + \sum_{i=1}^{d} \left(a_i^m x_i^{k_i} - h_i\left(x_i^{k_i}\right) - \lambda c_i\left(x_i^j, x_i^{k_i}\right) \right), 1 \leq j \leq n, 1 \leq m \leq M \right\}$$

$$= \min_{\lambda\geq 0,\, h_i,\, g(j)} \left\{ \lambda\rho + \frac{1}{n}\sum_{i=1}^{d}\sum_{j=1}^{n} h_i(x^j) + \frac{1}{n}\sum_{j=1}^{n} g(j) \,:\, g(j) \geq b^m \right.$$

$$\left. + \sum_{i=1}^{d} \max_{1\leq k\leq n} \left(a_i^m x_i^{k_i} - h_i\left(x_i^{k_i}\right) - \lambda c_i\left(x_i^j, x_i^{k_i}\right) \right), 1 \leq j \leq n, 1 \leq m \leq M \right\},$$

where we use that

$$\max_{k_1,\dots,k_d} \sum_{i=1}^{d} a_i^m x_i^{k_i} - h_i(x^{k_i}) - \lambda c_i(x_i^j, x_i^{k_i}) = \sum_{i=1}^{d} \max_{1\le k\le n} \left(a_i^m x_i^k - h_i(x_i^k) - \lambda c_i(x_i^j, x_i^k) \right).$$

Introducing the auxiliary variables $u_i(j,m) \in \mathbb{R}$, where $i = 1,\dots,d$, $j = 1,\dots,n$ and $m = 1,\dots,M$, in order to remove the remaining max function, together with the notation $h_i(j) := h_i(x^j) \in \mathbb{R}$ yields the assertion. □

2.2 | Penalization

The aim of this section is to modify the functional (9), so that it allows for a numerical solution by neural networks.

To focus on the main ideas, we assume that κ is bounded, that is, we restrict to continuous bounded functions, as well as $\varphi = \infty \mathbb{1}_{(\rho,\infty)}$ as in the overview in Subsection 1.2. Hence, in line with Corollary 2.2, we consider the functional

$$\begin{aligned} \phi(f) &:= \max_{\substack{\mu \in \Pi(\bar\mu_1,\dots,\bar\mu_d) \\ d_c(\bar\mu,\mu)\le\rho}} \int_X f \, d\mu \qquad (18) \\ &= \inf_{\lambda\ge 0,\, h_i \in C_{\kappa_i}(X_i)} \left\{ \rho\lambda + \sum_{i=1}^{d} \int_{X_i} h_i \, d\bar\mu_i + \int_X \sup_{y\in X} \left[f(y) - \sum_{i=1}^{d} h_i(y_i) - \lambda c(x,y) \right] \bar\mu(dx) \right\} \end{aligned}$$

for all $f \in C_b(X)$ and a fixed radius $\rho > 0$. For simplicity, we assume that the function $f^{\lambda c}(x) = \sup_{y\in X}\{f(y) - \lambda c(x,y)\}$ is continuous for all $\lambda \ge 0$ and $f \in C_b(X)$.[4] In that case, the functional $\phi_1 : C_b(X^2) \to \mathbb{R}$ defined as

$$\phi_1(f) := \inf_{\substack{\lambda\ge 0,\, h_i\in C_b(X_i),\, g\in C_b(X): \\ g(x)\ge f(x,y) - \sum_{i=1}^d h_i(y_i) - \lambda c(x,y)}} \left\{ \lambda\rho + \sum_{i=1}^{d} \int_{X_i} h_i \, d\bar\mu_i + \int_X g \, d\bar\mu \right\} \qquad (19)$$

satisfies $\phi(\tilde f) = \phi_1(\tilde f \circ \mathrm{pr}_2)$ for all $\tilde f \in C_b(X)$, that is, ϕ_1 is an extension of ϕ from $C_b(X)$ to $C_b(X^2)$. The functional ϕ_1 can be regularized by penalizing the inequality constraint. To do so, we consider the functional

$$\begin{aligned} \varphi_{\theta,\gamma}(f) := \inf_{\substack{\lambda\ge 0,\, h_i\in C_b(X_i), \\ g\in C_b(X)}} &\left\{ \lambda\rho + \sum_{i=1}^{d} \int_{X_i} h_i \, d\bar\mu_i + \int_X g \, d\bar\mu \right. \\ &\left. + \int_{X^2} \beta_\gamma \left(f(x,y) - g(x) - \sum_{i=1}^{d} h_i(y_i) - \lambda c(x,y) \right) \theta(dx,dy) \right\} \qquad (20) \end{aligned}$$

for a sampling measure $\theta \in \mathcal{P}(X^2)$, and a penalty function $\beta_\gamma(x) := \frac{1}{\gamma}\beta(\gamma x)$, $\gamma > 0$, where $\beta : \mathbb{R} \to [0, \infty)$ is convex, nondecreasing, differentiable, and satisfies $\frac{\beta(x)}{x} \to \infty$ for $x \to \infty$. Let $\beta_\gamma^*(y) := \sup_{x\in\mathbb{R}}\{xy - \beta_\gamma(x)\}$ for $y \in \mathbb{R}_+$, and note that $\beta_\gamma^*(y) = \frac{1}{\gamma}\beta^*(y)$.

Note that the introduced penalization method is in no way specific to the penalized constraint and hence, rather general. It includes as a special case the well-studied entropic penalization related to the Sinkhorn algorithm, which is often applied to optimal transport problems. The penalization can also be seen as a regularization because it introduces a slight smoothness bias for the probability measures in the optimization problem. On the one hand, this leads to an approximation error, which can be made arbitrarily small theoretically, see Propositions 2.7 and 2.8. On the other hand, the resulting smoothness is also seen as a feature that produces good empirical results (see, e.g., Cuturi, 2013; Genevay, Peyré, & Cuturi, 2017).

The following lemma sets the stage for Proposition 2.7, in which we provide a duality result for $\phi_{\theta,\gamma}(f)$, study the respective relation of primal and dual optimizers, and outline convergence $\phi_{\theta,\gamma}(f) \to \phi(f)$ for $\gamma \to \infty$.

Lemma 2.6. *For every $f \in C_b(X^2)$, one has*

$$\phi_{\theta,\gamma}(f) = \inf_{\tilde{f}\in C_b(X^2)} \left\{\phi_1(\tilde{f}) + \phi_2(f - \tilde{f})\right\}, \tag{21}$$

where $\phi_2(f) := \int_{X^2} \beta_\gamma(f)\, d\theta$. Moreover, the convex conjugate of $\phi_{\theta,\gamma}$ is given by

$$\varphi_{\theta,\gamma}^*(\pi) = \begin{cases} \int_{X^2} \beta_\gamma^*\left(\frac{d\pi}{d\theta}\right) d\theta & \text{if } \pi_1 = \bar{\mu},\ \pi_2 \in \Pi(\bar{\mu}_1, \dots, \bar{\mu}_d) \text{ and } \int_{X^2} c\, d\pi \le \rho \\ \infty & \text{else} \end{cases}$$

for all $\pi \in \mathcal{P}(X^2)$ with the convention $\frac{d\pi}{d\theta} = +\infty$ if π is not absolutely continuous with respect to θ.

Proof. Observe that for every $f \in C_b(X^2)$, one has

$$\begin{aligned} &\inf_{\tilde{f}\in C_b(X^2)} \left\{\phi_1(\tilde{f}) + \phi_2(f - \tilde{f})\right\} \\ &= \inf_{\substack{\lambda\ge 0,\, h_i\in C_b(X_i),\, g\in C_b(X),\, \tilde{f}\in C_b(X^2): \\ \tilde{f}(x,y)\le g(x)+\sum_{i=1}^d h_i(y_i)+\lambda c(x,y)}} \left\{\lambda\rho + \sum_{i=1}^d \int_{X_i} h_i\, d\bar{\mu}_i + \int_X g\, d\bar{\mu} + \int_{X^2} \beta_\gamma(f - \tilde{f})\, d\theta\right\} \end{aligned}$$

where the right-hand side is equal to $\phi_{\theta,\gamma}(f)$. This follows from the dominated convergence theorem applied on the sequence $\tilde{f}_n(x, y) = \min\{n,\ g(x) + \sum_{i=1}^d h_i(y_i) + \lambda c(x, y)\}$.

As for the calculation of the convex conjugate, we first show that $\phi_{\theta,\gamma}^*(\pi) = \infty$ whenever $\pi_1 \neq \bar{\mu}$ or $\pi_2 \notin \Pi(\bar{\mu}_1, \dots, \bar{\mu}_d)$. Indeed, as

$$\phi_{\theta,\gamma}(f) \leq \inf_{h_i \in C_b(X_i),\, g \in C_b(X)} \left\{ \sum_{i=1}^{d} \int_{X_i} h_i \, d\bar{\mu}_i + \int_X g \, d\bar{\mu} + \int_{X^2} \beta_\gamma \left(f(x,y) - g(x) - \sum_{i=1}^{d} h_i(y_i) \right) \theta(dx,dy) \right\}$$

$$\leq \inf_{\substack{h_i \in C_b(X_i),\, g \in C_b(X): \\ g(x) + \sum_i h_i(y_i) \geq f(x,y)}} \left\{ \sum_{i=1}^{d} \int_{X_i} h_i \, d\bar{\mu}_i + \int_X g \, d\bar{\mu} \right\} + \beta_\gamma(0),$$

it follows that $\phi_{\theta,\gamma}$ is bounded above by a multi-marginal transport problem. As the respective convex conjugate is $+\infty$, it follows that $\phi^*_{\theta,\gamma}(\pi) = \infty$ for all $\pi \in \mathcal{P}(X^2)$ such that $\pi_1 \neq \bar{\mu}$ or $\pi_2 \notin \Pi(\bar{\mu}_1, \dots, \bar{\mu}_d)$. Conversely, if $\pi_1 = \bar{\mu}$ and $\pi_2 \in \Pi(\bar{\mu}_1, \dots, \bar{\mu}_d)$ one has

$$\begin{aligned}
\phi^*_{\theta,\gamma}(\pi) &= \sup_{f \in C_b(X^2)} \left\{ \int_{X^2} f \, d\pi - \phi_{\theta,\gamma}(f) \right\} \\
&= \sup_{\lambda \geq 0} \sup_{\tilde{f} \in C_b(X^2)} \left\{ -\lambda\rho + \int_{X^2} \tilde{f} \, d\pi - \int_{X^2} \beta_\gamma(\tilde{f} - \lambda c) \, d\theta \right\} \\
&= \sup_{\lambda \geq 0} \sup_{\bar{f} \in C_b(X^2)} \left\{ -\lambda\rho + \lambda \int_{X^2} c \, d\pi + \int_{X^2} \bar{f} \, d\pi - \int_{X^2} \beta_\gamma(\bar{f}) \, d\theta \right\} \\
&= \sup_{\lambda \geq 0} \lambda \left(\int_{X^2} c \, d\pi - \rho \right) + \int_{X^2} \beta^*_\gamma \left(\frac{d\pi}{d\theta} \right) d\theta. \\
&= \begin{cases} \int_{X^2} \beta^*_\gamma \left(\frac{d\pi}{d\theta} \right) d\theta & \text{if } \int_{X^2} c \, d\pi \leq \rho \\ +\infty & \text{else} \end{cases}.
\end{aligned}$$

Here, the second equality follows by substituting $\tilde{f}(x,y) = f(x,y) - \sum_{i=1}^{d} h_i(y_i) - g(x)$ and using the structure of the marginals of π. The third equality follows by setting $\bar{f}^n = \tilde{f} + \min\{n, \lambda c\}$ and using the dominated convergence theorem. Finally, the fourth equality follows by a standard selection argument, see, for example, the proof of Bartl, Cheridito, et al. (2019, Lemma 3.5). □

Proposition 2.7. *Suppose there exists $\pi \in \mathcal{P}(X^2)$ such that $\phi^*_{\theta,\gamma}(\pi) < \infty$. Then it holds:*

(i) For every $f \in C_b(X^2)$, one has

$$\phi_{\theta,\gamma}(f) = \max_{\substack{\pi \in \Pi(\bar{\mu}, \bar{\mu}_1, \dots, \bar{\mu}_d): \\ \int c \, d\pi \leq \rho}} \int_{X^2} f \, d\pi - \int_{X^2} \beta^*_\gamma \left(\frac{d\pi}{d\theta} \right) d\theta. \tag{22}$$

(ii) Let $f \in C_b(X^2)$. If $g^\star \in C_b(X)$, $h_i^\star \in C_b(X_i)$, $i = 1, \dots, d$, and $\lambda^\star \geq 0$ are optimizers of (20), then the probability measure $\pi^\star$ defined by

$$\frac{d\pi^\star}{d\theta}(x,y) := \beta'_\gamma \left(f(x,y) - g^\star(x) - \sum_{i=1}^{d} h_i^\star(y_i) - \lambda^\star c(x,y) \right)$$

is a maximizer of (22). Hence, $\mu^\star := \pi^\star \circ pr_2^{-1}$ is a feasible solution to (18).

(iii) Fix $f \in C_b(X)$ and $\varepsilon > 0$. Suppose that $\mu_\varepsilon \in \mathcal{P}(X)$ is an ε-optimizer of (18), and $\pi_\varepsilon \in \Pi(\bar{\mu}, \mu_\varepsilon)$ satisfies $\alpha := \int_{X^2} \beta^*(\frac{d\pi_\varepsilon}{d\theta})\, d\theta < \infty$, and $\int_{X^2} c\, d\pi_\varepsilon \le \rho$. Then one has

$$\phi_{\theta,\gamma}(f \circ \mathrm{pr}_2) - \frac{\beta(0)}{\gamma} \le \phi(f) \le \phi_{\theta,\gamma}(f \circ pr_2) + \varepsilon + \frac{\alpha}{\gamma}.$$

Proof. (a) To show duality, we check condition (R1) from Bartl, Cheridito, et al. (2019, Theorem 2.2), that is, we have to show that $\phi_{\theta,\gamma}$ is real-valued and continuous from above. That $\phi_{\theta,\gamma}$ is real-valued follows from the assumption that there exists $\pi \in \mathcal{P}(X^2)$ such that $\phi^*_{\theta,\gamma}(\pi) < \infty$. Indeed, it holds $\infty > \phi^*_{\theta,\gamma}(\pi) \ge \int f d\pi - \phi_{\theta,\gamma}(f)$ and hence, $\phi_{\theta,\gamma}(f) > -\infty$ (while $\phi_{\theta,\gamma}(f) < \infty$ is clear) for all $f \in C_b(X^2)$.

To show continuity from above, let (f_n) be a sequence in $C_b(X^2)$ such that $f_n \downarrow 0$. In view of (21), one has

$$\inf_{n\in\mathbb{N}} \phi_{\theta,\gamma}(f_n) = \inf_{\tilde{f}\in C_b(X^2)} \inf_{n\in\mathbb{N}} \left\{\phi_1(\tilde{f}) + \phi_2(f_n - \tilde{f})\right\} = \inf_{\tilde{f}\in C_b(X^2)} \left\{\phi_1(\tilde{f}) + \phi_2(-\tilde{f})\right\} = \phi_{\theta,\gamma}(0),$$

as $\inf_{n\in\mathbb{N}} \phi_2(f_n - \tilde{f}) = \phi_2(-\tilde{f})$ by dominated convergence. By Lemma 2.6, the claim follows.

(b) That $\pi^\star$ is a feasible solution in the sense that $\pi^\star_1 = \bar{\mu}$, $\pi^\star_2 \in \Pi(\bar{\mu}_1, \dots, \bar{\mu}_d)$, and $\int_{X^2} c\, d\pi^\star = \rho$ whenever $\lambda^\star > 0$, follows from the first-order conditions. For instance, as the derivative of (20) in direction $g^\star + tg$ vanishes at $t = 0$, it follows $\int_X g\, d\bar{\mu} - \int_{X^2} g \circ \mathrm{pr}_1\, d\pi^\star = 0$ for all $g \in C_b(X)$, which shows that $\pi^\star_1 = \bar{\mu}$. This also implies that $\pi^\star$ is a probability measure. Similarly, $\pi^\star_2 \in \Pi(\bar{\mu}_1, \dots, \bar{\mu}_d)$ follows by considering the derivative in direction $h^\star_i + th_i$, and $\int_{X^2} \lambda^\star c\, d\pi^\star = \lambda^\star \rho$ from the first-order condition for λ. Hence, as $\pi^\star$ is feasible it follows from Lemma 2.6 that

$$\begin{aligned}
\phi_{\theta,\gamma}(f) &\ge \int_{X^2} f\, d\pi^\star - \phi^*_{\theta,\gamma}(\pi^\star) \\
&= \int_{X^2} f\beta'_\gamma\left(f - g^\star - \sum_i h^\star_i - \lambda^\star c\right) - \beta^*_\gamma\left(\beta'_\gamma\left(f - g^\star - \sum_i h^\star_i - \lambda^\star c\right)\right) d\theta \\
&= \int_{X^2} g^\star + \sum_i h^\star_i + \lambda^\star c\, d\pi^\star + \int_{X^2} \beta_\gamma\left(f - \hat{g} - \sum_i h^\star_i - \lambda^\star c\right) d\theta \\
&= \lambda^\star \rho + \sum_i \int_{X_i} h^\star_i\, d\bar{\mu}_i + \int_X g^\star\, d\bar{\mu} + \int_{X^2} \beta_\gamma\left(f - \hat{g} - \sum_i h^\star_i - \lambda^\star c\right) d\theta \\
&= \phi_{\theta,\gamma}(f),
\end{aligned}$$

where we use that $\beta^*_\gamma(\beta'_\gamma(x)) = \beta'_\gamma(x)x - \beta_\gamma(x)$ for all $x \in \mathbb{R}$. This shows that $\pi^\star$ is an optimizer.

(c) By restricting the infimum in (20) to those $\lambda \ge 0$, $h_i \in C_b(X_i)$, $g \in C_b(X)$ such that $g(x) \ge f(y) - \sum_i h_i(y_i) - \lambda c(x, y)$, it follows that

$$\phi_{\theta,\gamma}(f \circ \mathrm{pr}_2) \le \inf_{\substack{\lambda\ge 0,\, h_i\in C_b(X_i),\, g\in C_b(X):\\ g(x)\ge f(y) - \sum_{i=1}^d h_i(y_i) - \lambda c(x,y)}} \left\{\lambda\rho + \sum_{i=1}^d \int_{X_i} h_i\, d\bar{\mu}_i + \int_X g\, d\bar{\mu}\right\} + \beta_\gamma(0)$$

$$= \phi(f) + \frac{\beta(0)}{\gamma},$$

where the last equality follows from (19). As for the second inequality, as $\mu_\varepsilon \in \mathcal{P}(X)$ is an ε-optimizer of (18), and $\pi_\varepsilon \in \Pi(\bar{\mu}, \mu_\varepsilon)$ one has

$$\phi(f) \leq \int_X f \, d\mu_\varepsilon + \varepsilon = \int_{X^2} f \circ \mathrm{pr}_2 \, d\pi_\varepsilon - \phi^*_{\theta,\gamma}(\pi_\varepsilon) + \phi^*_{\theta,\gamma}(\pi_\varepsilon) + \varepsilon \leq \phi_{\theta,\gamma}(f \circ \mathrm{pr}_2) + \frac{\alpha}{\gamma} + \varepsilon.$$

The proof is complete. □

The following Proposition shows that the convergence result from Proposition 2.7(c) can be applied whenever the sampling measure is chosen as $\theta = \bar{\mu} \otimes \bar{\mu}_1 \otimes \cdots \otimes \bar{\mu}_d$, and a minimal growth condition on the cost function c is imposed, see Proposition 2.8(b)(ii). In this case, existence of $\pi \in \mathcal{P}(X^2)$ such that $\phi^*_{\theta,\gamma}(\pi) < \infty$ holds as well, so that Proposition 2.7 applies in full. It is worth pointing out that this result below trivially transfers to all reference measures θ for which the Radon–Nikodym derivative $\frac{d\bar{\mu}\otimes\bar{\mu}_1\otimes\cdots\bar{\mu}_d}{d\theta}$ is bounded. As pointed out by a referee, it is especially desirable to have the values $\alpha_\varepsilon := \int_{X^2} \beta^*(\frac{d\pi_\varepsilon}{d\theta})\, d\theta$ uniformly bounded in ε (respectively, growing in a certain order depending on ε), so that a linear convergence $\phi_{\theta,\gamma}(f) \to \phi(f)$ for $\gamma \to \infty$ (respectively, a slower order of convergence) is implied. The result below does not achieve this, and we believe this to be a nontrivial task left open for future work.

Proposition 2.8.

(a) Let $\mu_i \in \mathcal{P}(X_i)$ for $i = 1, \dots, d$. Let $\nu \in \Pi(\mu_1, \dots, \mu_d)$ and let $\mu := \mu_1 \otimes \mu_2 \otimes \cdots \otimes \mu_d$. Then there exist $\nu^n \in \Pi(\mu_1, \dots, \mu_d)$ for $n \in \mathbb{N}$ such that $\nu^n \xrightarrow{w} \nu$ for $n \to \infty$, $\nu^n \ll \mu$ and there exist constants $0 < C_n < \infty$ such that $\frac{d\nu^n}{d\mu} \leq C_n$ μ-a.s..

(b) Let $\eta_i : X_i \to [0, \infty)$ be Borel measurable with $\int_{X_i} \eta_i \, d\bar{\mu}_i < \infty$. Let $\eta(x) = \sum_{i=1}^d \eta_i(x_i)$. Assume there is a constant $C > 0$ such that for all $x, y \in X$ it holds $c(x, y) \leq C(\eta(x) + \eta(y))$. Let $\theta = \bar{\mu} \otimes \bar{\mu}_1 \otimes \cdots \otimes \bar{\mu}_d$. Then it holds:

(i) For $\pi^* \in \Pi(\bar{\mu}, \bar{\mu}_1, \dots, \bar{\mu}_d)$ with $\int c \, d\pi^* \leq \rho$, there exist $\pi_\varepsilon \in \Pi(\bar{\mu}, \bar{\mu}_1, \dots, \bar{\mu}_d)$ for $\varepsilon > 0$ such that $\pi_\varepsilon \xrightarrow{w} \pi^*$ for $\varepsilon \to 0$, $\pi_\varepsilon \ll \theta$, $\frac{d\pi_\varepsilon}{d\theta}$ is θ-a.s. bounded and $\int c \, d\pi_\varepsilon \leq \rho$.

(ii) The condition for Proposition 2.7(c) is satisfied, that is, for every $\varepsilon > 0$ there exists $\mu_\varepsilon \in \mathcal{P}(X)$ that is an ε-optimizer of (18), and $\pi_\varepsilon \in \Pi(\bar{\mu}, \mu_\varepsilon)$ satisfying $\alpha_\varepsilon := \int_{X^2} \beta^*(\frac{d\pi_\varepsilon}{d\theta})\, d\theta < \infty$, and $\int_{X^2} c \, d\pi_\varepsilon \leq \rho$.

Proof. Proof of (a): We endow each X_i by a compatible metric m_i and without loss of generality we specify the metric on $X = X_1 \times X_2 \times \cdots \times X_d$ as $m(x, y) = \sum_{i=1}^d m_i(x_i, y_i)$.

Step 1: Construction of ν^n: Let $K_i^n \subseteq X_i$ be compact for $i = 1, \dots, d$ such that $\mu_i(K_i^n) \to 1$ for $n \to \infty$, and $K^n = K_1^n \times \cdots \times K_d^n$. Notably $\nu(K^n) \to 1$ follows.

Further, as each K_i^n is compact we can choose a Borel partition $\mathcal{A}_n$ of K^n, that is,

$$\dot{\cup}_{A \in \mathcal{A}_n} A = K^n,$$

where each $A \in \mathcal{A}_n$ is Borel measurable and satisfies $A = A_1 \times \cdots \times A_d$ as well as

$$\max_{A \in \mathcal{A}_n} \sup_{x,y \in A} m(x,y) \leq \frac{1}{n}.$$

A simple way of obtaining such a partition is to first cover each K_i^n by countably many open balls of radius $1/(2dn)$, choosing a finite subcover, and building a partition of K_i^n out of that subcover. For the partition of K^n simply choose all product sets that can be formed from the partitions of the K_i^n.

Note that $(K^n)^c$ is the disjoint union of $2^d - 1$ many product sets, namely $K_{n,1}^c \times K_{n,2} \times \cdots \times K_{n,d}, \ldots, K_{n,1}^c \times \cdots \times K_{n,d}^c$. We denote the union of $\mathcal{A}_n$ with the family of these $2^d - 1$ many product sets by $\overline{\mathcal{A}}_n$, which is a partition of X. Define

$$\nu^n := \sum_{A \in \overline{\mathcal{A}}_n} \nu(A) \cdot (\nu_{|A})_1 \otimes \cdots \otimes (\nu_{|A})_d$$

where implicitly the sum is understood to only include those terms where $\nu(A) > 0$ and $\nu_{|A}$ is then defined as $\nu_{|A}(B) = \nu(A \cap B)/\nu(A)$. We do not make this explicit, but every time we divide by $\nu(A)$ or $\mu_i(A_i)$ we will assume it is one of the relevant terms with $\nu(A) > 0$, where of course $\nu(A) > 0$ implies $\mu_i(A_i) > 0$ for all $i = 1, \ldots, d$ and $A \in \overline{\mathcal{A}}_n$.

Step 2: Verifying marginals of ν^n: We only show that $\nu_1^n = \mu_1$, while the other marginals follow in the same way by symmetry. Let $B_1 \subseteq X_1$ be Borel. It holds

$$\begin{aligned}
\nu^n(B_1 \times X_2 \times \cdots \times X_d) &= \sum_{A \in \overline{\mathcal{A}}_n} \nu(A) \cdot (\nu_{|A})_1(B_1) \\
&= \sum_{A \in \overline{\mathcal{A}}_n} \nu(A) \cdot \nu_{|A}(B_1 \times X_2 \times \cdots \times X_d) \\
&= \sum_{A \in \overline{\mathcal{A}}_n} \nu(A \cap (B_1 \times X_2 \times \cdots \times X_d)) \\
&= \nu(B_1 \times X_2 \times \cdots \times X_d) = \nu_1(B_1).
\end{aligned}$$

Step 3: Convergence $\nu^n \overset{w}{\to} \nu$ for $n \to \infty$: Let $f : X \to \mathbb{R}$ be bounded and Lipschitz continuous with constant L. We have to show $\int f \, d\nu^n \to \int f \, d\nu$ for $n \to \infty$. As $\nu(A) = \nu^n(A)$ for all $A \in \overline{\mathcal{A}}_n$ (and in particular $\nu^n((K^n)^c) = \nu((K^n)^c)$), it holds

$$\begin{aligned}
\left| \int f \, d\nu^n - \int f \, d\nu \right| &\leq \|f\|_\infty 2\nu((K^n)^c) + \sum_{A \in \mathcal{A}_n} \left| \int f \mathbb{1}_A \, d\nu^n - \int f \mathbb{1}_A \, d\nu \right| \\
&\leq \|f\|_\infty 2\nu((K^n)^c) + \sum_{A \in \mathcal{A}_n} \sup_{x,y \in A} |f(x) - f(y)| \nu(A) \\
&\leq \|f\|_\infty 2\nu((K^n)^c) + \sum_{A \in \mathcal{A}_n} \nu(A) L \frac{1}{n} \\
&\overset{n \to \infty}{\longrightarrow} 0.
\end{aligned}$$

Step 4: Absolute continuity and boundedness of $\frac{d\nu^n}{d\mu}$: Let

$$C_n = \max_{\substack{A\in\overline{\mathcal{A}}_n:\\ \nu(A)>0}} \frac{1}{\nu(A)^{d-1}}.$$

Given arbitrary Borel sets $B_i \subseteq X_i$ for $i = 1, \dots, d$, we show that $\nu^n(B_1 \times \cdots \times B_d) \leq C_n\, \mu(B_1 \times \cdots \times B_d)$. Once this is shown, $\nu^n(S) \leq C_n\, \mu(S)$ follows for all Borel sets $S \subseteq X$ by the monotone class theorem, which will immediately yield both absolute continuity $\nu^n \ll \mu$ and $\frac{d\nu^n}{d\mu} \leq C_n$.

For $A \in \overline{\mathcal{A}}_n$ it holds

$$\begin{aligned}(\nu_{|A})_i(B_i) &= \nu(A)^{-1}\nu(A_1 \times \cdots \times (A_i \cap B_i) \times \cdots \times A_d)\\ &\leq \nu(A)^{-1} \cdot \nu_i(A_i \cap B_i) = \nu(A)^{-1} \cdot \mu_i(A_i \cap B_i).\end{aligned}$$

It follows

$$\begin{aligned}\nu^n(B_1 \times \cdots \times B_d) &= \sum_{A\in\overline{\mathcal{A}}_n} \nu(A) \cdot (\nu_{|A})_1(B_1) \cdot \cdots \cdot (\nu_{|A})_d(B_d)\\ &\leq \sum_{A\in\overline{\mathcal{A}}_n} \frac{1}{\nu(A)^{d-1}} \mu_1(B_1 \cap A_1) \cdot \cdots \cdot \mu_d(B_d \cap A_d)\\ &\leq C_n \sum_{A\in\overline{\mathcal{A}}_n} \mu((B_1 \times \cdots \times B_d) \cap A) = C_n\, \mu(B_1 \times \cdots \times B_d),\end{aligned}$$

where we note that for the last equality to hold, the second to last sum over $A \in \overline{\mathcal{A}}_n$ includes all terms, not just the ones where $\nu(A) > 0$ (this only makes the sum larger). The proof of part (a) is complete.

Proof of (b)(i): We first show the following: If $\Pi(\bar{\mu}, \bar{\mu}_1, \dots, \bar{\mu}_d) \ni \pi_\varepsilon \overset{w}{\to} \pi \in \Pi(\bar{\mu}, \bar{\mu}_1, \dots, \bar{\mu}_d)$ for $\varepsilon \to 0$, then $\int c\, d\pi_\varepsilon \to \int c\, d\pi$ for $\varepsilon \to 0$.

To prove it, note that the growth condition implies that for every $\delta > 0$, we can choose a compact set $K \in X \times X$ such that $\sup_{\varepsilon>0} \int_{K^c} c\, d\pi_\varepsilon \leq \delta$ and $\int_{K^c} c\, d\pi \leq \delta$. Restricted to K, c is bounded from above, say by a constant $M > 0$ (note c is nonnegative). Hence, for all $\varepsilon > 0$, it holds $|\int c\, d\pi_\varepsilon - \int \min\{c, M\}\, d\pi_\varepsilon| \leq 2\delta$, and the same for π instead of π_ε. As $\min\{c, M\}$ is continuous and bounded, we get $|\int c\, d\pi_\varepsilon - \int c\, d\pi| \leq 4\delta + |\int \min\{c, M\}\, d\pi_\varepsilon - \int \min\{c, M\}\, d\pi| \to 4\delta$ for $\varepsilon \to 0$. Letting δ go to zero yields the claim.

For the statement of the part (b)(i), consider for $\lambda \in (0, 1)$ the coupling $\pi_\lambda := \lambda\pi^* + (1 - \lambda)(\bar{\mu} \otimes (x \mapsto \delta_x))$. Then it holds $\int c\, d\pi_\lambda < \rho$ as $c(x, x) = 0$. Further $\pi_\lambda \overset{w}{\to} \pi^*$ for $\lambda \to 1$. By approximating every π_λ by a $\pi_{\lambda,\varepsilon}$ via part (a), $\int c\, d\pi_{\lambda,\varepsilon} \leq \rho$ follows automatically for ε small enough, which follows by $\int c\, d\pi_{\lambda,\varepsilon} \to \int c\, d\pi_\lambda$ for $\varepsilon \to 0$ as shown above. The claim hence follows by a diagonal argument.

Proof of (b)(ii): Any optimizer $\mu^\star \in \Pi(\bar{\mu}_1, \dots, \bar{\mu}_d)$ of (18) and a corresponding coupling $\pi^* \in \Pi(\bar{\mu}, \bar{\mu}_1, \dots, \bar{\mu}_d)$ with $\int c\, d\pi^* \leq \rho$ can be approximated via part (b) by $(\pi_\varepsilon)_{\varepsilon>0}$ that satisfies all required properties. Taking μ_ε as the projection of π_ε onto the second component of $X \times X$, that

is, $\mu_\varepsilon = \pi_\varepsilon \circ ((x,y) \mapsto y)^{-1}$, we get that $\mu_\varepsilon \overset{w}{\to} \mu^\star$ and hence $\int f \, d\mu_\varepsilon \to \int f \, d\mu^\star$ that means after a possible change of indices, μ_ε is an ε-optimizer of (18). □

2.3 | Approximation with neural networks

Let us shortly recap. In Subsection 2.1, we show that our original problem,

$$\phi(f) := \max_{\substack{\mu \in \Pi(\bar{\mu}_1, \dots, \bar{\mu}_d) \\ d_c(\bar{\mu}, \mu) \le \rho}} \int_{\mathbb{R}^d} f \, d\mu,$$

can be written as

$$\inf_{\lambda \ge 0,\, h_i \in C_b(\mathbb{R})} \left\{ \rho\lambda + \sum_{i=1}^d \int_{\mathbb{R}} h_i \, d\bar{\mu}_i + \int_{\mathbb{R}^d} \sup_{y \in \mathbb{R}^d} \left[f(y) - \sum_{i=1}^d h_i(y_i) - \lambda c(x,y) \right] \bar{\mu}(dx) \right\}$$

for all continuous and bounded functions $f \in C_b(X)$. We then proceed in Subsection 2.2 with considering the penalized version of the latter problem

$$\begin{aligned} \varphi_{\theta,\gamma}(f) := \inf_{\substack{\lambda \ge 0, \\ h_i \in C_b(\mathbb{R}),\, g \in C_b(\mathbb{R}^d)}} & \left\{ \lambda\rho + \sum_{i=1}^d \int_{\mathbb{R}} h_i \, d\bar{\mu}_i + \int_{\mathbb{R}^d} g \, d\bar{\mu} \right) \\ & + \left(\int_{\mathbb{R}^{2d}} \beta_\gamma \left(f(y) - \sum_{i=1}^d h_i(y_i) - \lambda c(x,y) - g(x) \right) \theta(dx, dy) \right\}. \end{aligned}$$

We provide sufficient conditions for the convergence $\phi_{\theta,\gamma}(f) \to \phi(f)$ for $\gamma \to \infty$. The subsequent and final step is to theoretically justify that neural networks can indeed be used to approximate $\phi_{\theta,\gamma}(f)$ and thereby $\phi(f)$.

To do so, let us introduce the following notation: We denote by $A_0, \dots, A_l$ affine transformations with A_0 mapping form $\mathbb{R}^{d_0}$ to $\mathbb{R}^m$, $A_1, \dots, A_{l-1}$ mapping form $\mathbb{R}^m$ to $\mathbb{R}^m$ and A_l mapping form $\mathbb{R}^m$ to $\mathbb{R}$. We further fix a nonconstant, continuous and bounded *activation function* $\varphi : \mathbb{R} \to \mathbb{R}$. The evaluation of φ at a vector $y \in \mathbb{R}^m$ is understood pointwise, that is, $\varphi(y) = (\varphi(y_1), \dots, \varphi(y_m))$. Then,

$$\mathfrak{N}(m, d_0) := \left\{ g : \mathbb{R}^{d_0} \to \mathbb{R} : x \mapsto A_l \circ \varphi \circ A_{l-1} \circ \dots \circ \varphi \circ A_0(x) \right\}$$

defines the set of neural network functions mapping to the real numbers $\mathbb{R}$ with a fixed number of layers $l \ge 2$ (at least one hidden layer), input dimension d_0 and hidden dimension m. The following is a classical *universal approximation theorem* for neural networks.

Theorem 2.9 (Hornik, 1991). *Let $h \in C_b(\mathbb{R}^N)$. For any finite measure $\nu \in \mathcal{P}(\mathbb{R}^N)$ and $\varepsilon > 0$ there exists $m \in \mathbb{N}$ and $h^m \in \mathcal{N}(m, N)$ such that $\|h - h^m\|_{L^p(\nu)} \le \varepsilon$.*

For Proposition 2.10, let $X_i := \mathbb{R}^{N_i}$ for $i = 1, \dots, d$ and thus $X = \mathbb{R}^N$ with $N = \sum_{i=1}^d N_i$. Define the function $F(\lambda, h_1, \dots, h_d, g)$ by

$$\phi_{\theta,\gamma}(f) = \inf_{\substack{\lambda \geq 0, \\ h_i \in C_b(\mathbb{R}^{N_i}),\, g \in C_b(\mathbb{R}^N)}} F(\lambda, h_1, \dots, h_d, g).$$

We define the neural network approximation of $\phi_{\theta,\gamma}(f)$ by

$$\phi^m_{\theta,\gamma}(f) = \inf_{\substack{\lambda \geq 0, \\ h_i^m \in \mathfrak{N}(m,N_i),\, g^m \in \mathfrak{N}(m,N)}} F(\lambda, h_1^m, \dots, h_d^m, g^m).$$

The following result showcases a simple, yet general setting in which the neural network approximation is asymptotically precise.

Proposition 2.10. *Fix $f \in C_b(\mathbb{R}^N)$. Let $p > 1$, $\beta_\gamma(x) := \frac{1}{\gamma}(\gamma x)_+^p$ and assume $c \in L_p(\theta)$. Then*

$$\phi^m_{\theta,\gamma}(f) \to \phi_{\theta,\gamma}(f) \quad \text{for } m \to \infty.$$

Proof. By the choice of the activation function, all network functions are continuous and bounded, and hence $\phi^m_{\theta,\gamma}(f) \geq \phi_{\theta,\gamma}(f)$.

It therefore suffices to show that for any $\varepsilon > 0$, there exists an $m \in \mathbb{N}$ such that

$$\phi_{\theta,\gamma}(f) \geq \phi^m_{\theta,\gamma}(f) - \varepsilon.$$

Choose any feasible $(\lambda, h_1, \dots, h_d, g)$ for $\phi_{\theta,\gamma}(f)$. By Theorem 2, we can find a sequence $(\lambda^m, h_1^m, \dots, h_d^m, g^m)$ with $h_i^m \in \mathfrak{N}(m, N_i)$ for $i = 1, \dots, d$ and $g^m \in \mathfrak{N}(m, N)$ such that for $m \to \infty$ it holds

$$\begin{aligned}
\lambda^m &\to \lambda, \\
h_i^m &\to h_i && \text{in } L_p(\bar{\mu}_i) \text{ for } i = 1, \dots, d, \\
((x,y) \mapsto h_i^m(y_i)) &\to ((x,y) \mapsto h_i(y_i)) && \text{in } L_p(\theta) \text{ for } i = 1, \dots, d, \\
g^m &\to g && \text{in } L_p(\bar{\mu}), \\
((x,y) \mapsto g^m(x)) &\to ((x,y) \mapsto g(x)) && \text{in } L_p(\theta).
\end{aligned}$$

As $c \in L_p(\theta)$, it also holds $\lambda^m c \to \lambda c$ in $L_p(\theta)$ and hence

$$\begin{aligned}
&\left((x,y) \mapsto f(x,y) - g^m(x) - \sum_{i=1}^d h_i^m(y_i) - \lambda^m c(x,y) \right) \\
&\to \left((x,y) \mapsto f(x,y) - g(x) - \sum_{i=1}^d h_i(y_i) - \lambda c(x,y) \right)
\end{aligned}$$

in $L_p(\theta)$ as $m \to \infty$. As the mapping $x \mapsto x^+$ is Lipschitz-1, taking only the positive parts lets the above convergence remain valid. As convergence in $L_p(\theta)$ implies convergence of the pth moment, we obtain $F(\lambda^m, h_1^m, \dots, h_d^m, g^m) \to F(\lambda, h_1, \dots, h_d, g)$ as $m \to \infty$.

For a given $\varepsilon > 0$, choose a feasible $(\lambda, h_1, \dots, h_d, g)$ for $\phi_{\theta,\gamma}(f)$, such that $\phi_{\theta,\gamma}(f) \geq F(\lambda, h_1, \dots, h_d, g) - \frac{\varepsilon}{2}$. Due to the above proven convergence, we can find $(\lambda^m, h_1^m, \dots, h_d^m, g^m)$ with $h_i^m \in \mathfrak{N}(m, N_i)$ for $i = 1, \dots, d$ and $g^m \in \mathfrak{N}(m, N)$ such that

$$\phi_{\theta,\gamma}(f) \geq F(\lambda, h_1, \dots, h_d, g) - \frac{\varepsilon}{2} \geq \left(F(\lambda^m, h_1^m, \dots, h_d^m, g^m) - \frac{\varepsilon}{2}\right) - \frac{\varepsilon}{2} \geq \phi_{\theta,\gamma}^m(f) - \varepsilon.$$

□

Remark 2.11. Although the previous result obtains, for a fixed $f \in C_b(\mathbb{R}^N)$, the convergence $\phi_{\theta,\gamma}^m(f) \to \phi_{\theta,\gamma}(f)$ for $m \to \infty$, approximation errors for finite values of m are also of interest. In general, there is hope to achieve this. In the setting of Proposition 2.10 with $p \in \mathbb{N}_{\geq 2}$: Assume $\lambda, h_1, \dots, h_d, g$ are optimizers of $\phi_{\theta,\gamma}(f)$ that have sufficient moments and $h_1^m, \dots, h_d^m, g^m$ are network functions that approximate $h_1, \dots, h_d, g$ up to ε accuracy for the respective L^p-norms. Then it holds

$$\left|\phi_{\theta,\gamma}(f) - \phi_{\theta,\gamma}^m(f)\right| \leq C \cdot \varepsilon,$$

where C is a constant only depending on $f, \lambda, c, h_1, \dots, h_d, g$.

Proof. First, define $T(x,y) := f(y) - \sum_{i=1}^d h_i(y_i) - \lambda c(x,y) - g(x)$ and $T^m(x,y) := f(y) - \sum_{i=1}^d h_i^m(y_i) - \lambda c(x,y) - g^m(x)$. Using the function F introduced above, we have that $\phi_{\theta,\gamma}^m(f) \leq F(\lambda, h_1^m, \dots, h_d^m, g^m)$ and hence

$$\begin{aligned}\phi_{\theta,\gamma}^m(f) - \phi_{\theta,\gamma}(f) &\leq |F(\lambda, h_1^m, \dots, h_d^m, g^m) - F(\lambda, h_1^m, \dots, h_d^m, g)| \\ &\leq \sum_{i=1}^d \|h_i - h_i^m\|_{L^1(\bar{\mu}_i)} + \|g - g^m\|_{L^1(\bar{\mu})} + \|T_+^p - (T^m)_+^p\|_{L^1(\theta)}\end{aligned}$$

and by the inequality quoted in Lemma A.1 (see Appendix A.1) it holds

$$\|T_+^p - (T^m)_+^p\|_{L^1(\theta)} \leq \tilde{C}\|T_+ - T_+^m\|_{L_p(\theta)} \leq \tilde{C}\|T - T^m\|_{L^p(\theta)} \leq \tilde{C}(d+1)\varepsilon$$

with $\tilde{C} = \sum_{k=0}^{p-1} \|T_+\|_{L^p(\theta)}^k \|T_+^m\|_{L^p(\theta)}^{p-1-k}$. As $\phi_{\theta,\gamma}(f) \leq \phi_{\theta,\gamma}^m(f)$, we obtain

$$\left|\phi_{\theta,\gamma}(f) - \phi_{\theta,\gamma}^m(f)\right| = \phi_{\theta,\gamma}^m(f) - \phi_{\theta,\gamma}(f) \leq (\tilde{C}(d+1) + (d+1)) \cdot \varepsilon.$$

Although the constant $\tilde{C}$ formally depends on the pth moments of both T and T^m, to eliminate the dependence on T^m one can use $\|T^m\|_{L^p(\theta)} \leq \|T\|_{L^p(\theta)} + \|T - T^m\|_{L^p(\theta)} \leq \|T\|_{L^p(\theta)} + 1$ for ε small enough. □

3 | IMPLEMENTATION

This section aims to give specifics regarding the implementation of problem (4) as an approximation of problem (1). In particular, the following points are discussed:

1. The choice of θ, β_γ, and neural network structure.
2. The optimization method for the parameters of the neural network.
3. How to evaluate the quality of the obtained solution.
4. The typical runtime.

3.1 | Choice of θ, β_γ, and neural network parameters

The neural network structure to approximate the space $C_b(\mathbb{R}^d)$ is chosen as a feedforward neural network with five layers (input, output, three hidden layers) with hidden dimension $64 \cdot d$. The basic idea behind this was to increase the size of the neural networks until a further increase no longer changes the outcome of the optimization. As an activation function, we use the ReLu function.

To be precise, the neural network functions we work with are of the form

$$x \mapsto \underbrace{A_4}_{\substack{\text{output}\\ \text{layer}}} \circ \underbrace{\varphi \circ A_3}_{\text{4th layer}} \circ \underbrace{\varphi \circ A_2}_{\text{3rd layer}} \circ \underbrace{\varphi \circ A_1}_{\text{2nd layer}} \circ \underbrace{\varphi \circ A_0}_{\substack{\text{input}\\ \text{layer}}}(x),$$

where the activation function φ is chosen as $\varphi(x) = \max\{0, x\}$. The mappings A_i are affine transformations, that is, $A_i(x) = M_i\, x + b_i$ for a matrix $M_i \in \mathbb{R}^{d_{i,2} \times d_{i,1}}$ and a vector $b_i \in \mathbb{R}^{d_{i,2}}$. The matrices $M_0, \dots, M_4$ and vectors $b_0, \dots, b_4$ are the parameters of the network that one optimizes for. As described above, the dimensions of these parameters are chosen as follows: The input dimension $d_{0,1} = d$ is given by the dimension of the input vector x, and $d_{i,2} = d_{i+1,1}$ has to hold for compatibility. The hidden dimension $d_{i,1}$ for $i = 1, 2, 3, 4$ is set to $64 \cdot d$, while the output dimension $d_{4,2}$ is always 1.

The penalization function β_γ is set to $\beta_\gamma(x) = \gamma \max\{0, x\}^2$. On the one hand, this choice has shown to be stable across all examples. On the other hand, the theory in Proposition 2.10 applies precisely to penalization functions of this kind. Regarding the parameter γ, we usually first solve the problem with a low choice, like $\gamma = 50$, which leads to stable performance. Then, we gradually increase γ until a further increment no longer leads to a significant change in the objective value of (4). Regarding instabilities when γ is set too large, see Subsection 3.3.

Concerning the sampling measure θ, the basic choice is to use $\theta^{\text{prod}} = \bar{\mu} \otimes \bar{\mu}_1 \otimes \cdots \otimes \bar{\mu}_d$. Particularly for low values of ρ, this is suboptimal: Indeed, for $\rho = 0$ in problem (22), we know that the optimizer is always of the form $\pi^{\text{diag}} = \bar{\mu} \otimes K$, where K is the stochastic kernel $K : \mathbb{R}^d \to \mathcal{P}(\mathbb{R}^d)$ given by $K(x) = \delta_x$. As π^{diag} is singular with respect to θ^{prod}, using only θ^{prod} as sampling measure, one can expect high errors arising from penalization for small values of ρ. It hence makes sense to use (among other possibilities) $\theta^{\text{half}} := \frac{1}{2}\theta^{\text{prod}} + \frac{1}{2}\pi^{\text{diag}}$. This is very specific, however, and most solutions will not put mass precisely where π^{diag} puts mass. Hence, we add some noise to π^{diag}, for example, via a Gaussian measure with covariance matrix ε^2: $\theta^{\text{third}} := \frac{1}{2}\theta^{\text{prod}} + \frac{1}{4}\pi^{\text{diag}} + \frac{1}{4}(\pi^{\text{diag}} * \mathcal{N}(0, \varepsilon^2))$, where $*$ denotes convolution of measures. The

sampling measure θ^{half} is used in all four toy examples in Section 4, whereas we rely on θ^{third} in the final case study in Section 5.

3.2 | Optimization method for the parameters of the neural network

This subsection may as well be called "Training." However, as we do not employ neural networks in a training-testing kind of environment, this might be misleading.

Regarding this topic, trial and error is especially useful, as the simple goal is to obtain a stable convergence. For the parameters of the neural network, we use the Adam Optimizer with parameters $\beta_1 = .99$ and $\beta_2 = .995$. For the learning rate, we start with $\alpha = .0001$ for the first N_0 iterations of training, and then decrease it by a factor of .98 each 50 iterations for a total of N_{fine} further iterations. We use a batch size (the number of samples generated in each iteration for the measures involved) of around 2^7 to 2^{16}, see Subsection 3.3 for more details. N_0 and N_{fine} are chosen problem specific: for simple problems in Section 4, $N_0 = 15,000$ and $N_{\text{fine}} = 5,000$, while for the DNB case study in Section 5 they are chosen as $N_0 = 60,000$ and $N_{\text{fine}} = 30,000$.

The parameter λ has to be optimized separately from the parameters of the neural network, as the value of λ is clearly more important than any single parameter of the network. To be precise after a fixed number N_λ of iterations, λ is updated by

$$\lambda \mapsto \lambda - \alpha_\lambda \frac{1}{N_\lambda} \sum_{i \in I} \nabla_\lambda^i,$$

where I are the previous N_λ many iterations, α_λ is the learning rate and ∇_λ^i is the sample derivative of the objective function with respect to λ in iteration i. Concerning the choice of α_λ and N_λ, we usually first set α_λ to around .1 (depending on the problem), and decrease it in the same fashion as α, while N_λ is set to 200. Before we update λ for the first time, we wait until the network parameters are in a sensible region, which typically takes around 1,000–10,000 iterations.

If another parameter is involved in the optimization (such as τ in the examples that calculate the Average Value at Risk [AVaR]), we employ the same method as for λ, but we update this parameter even more rarely (once every 1,000–2,500 iterations), and wait longer at the start to update it the first time (between 5,000 and 20,000 iterations).

3.3 | Evaluation of the solution quality

To evaluate the obtained solutions, we found that mostly three aspects have to be considered:

(a) Is the neural network structure rich enough?
(b) How large is the effect of penalization?
(c) Has the numerical optimization procedure converged to a (near) minimum?

Section A.2 shows how this is put into practice for an exemplary case.

Part (a) is the seemingly simplest, as we found the choice of network structure described in Subsection 3.1 to be sufficient for all problems, in the sense that further increasing the network size does not alter the obtained solution.

Regarding part (b), the most useful observation is the following: As described in Proposition 2.7, the numerical solution via neural networks can be used to obtain an approximate solution $\mu^\star$ of the primal problem. If we evaluate the integral $\int f d\mu^\star$ and compare it to $\phi_{\theta,\gamma}(f)$, the difference is $\phi^*_{\theta,\gamma}(\pi^\star)$, which can be seen as the effect of penalization. If $\phi^*_{\theta,\gamma}(\pi^\star)$ has a small value, it indicates a small effect of penalization. The second observation is that $\phi_{\theta,\gamma}(f)$ is increasing in γ, and under the conditions studied in Propositions 2.7 and 2.8 converges to $\phi(f)$. Hence, starting with a low value of γ and increasing it until no further change is observed is a good strategy. When doing so, values of γ that are too large can of course be detrimental regarding part (c), and hence when increasing γ a concurrent adaptation of training parameters (like learning rate or batch size) is often necessary.

Regarding part (c), we found that most instabilities could be solved by increasing the batch size. This increase naturally comes with longer run times. Especially if γ has to be increased a lot to allow for a small effect of penalization, very large batch sizes were required (e.g., in the DNB case study, we use a batch size of 2^{15}). To obtain structured criteria for convergence (compared to just evaluating convergence visually), we can again use the dual relation arising from Proposition 2.7. Indeed, we can exploit the fact the numerically obtained $\mu^\star$ (as the second marginal of $\pi^\star$ from Proposition 2.7(b)) is an approximately feasible solution to problem (1) if the algorithm has converged. Hence, as a necessary criteria for convergence, one can check whether $\mu^\star$ satisfies criteria for feasibility. To this end, one can compare the marginals of $\mu^\star$ to those of $\bar{\mu}$ (we did this mostly by visually evaluating empirical marginals of $\mu^\star$) as well as estimate $d_c(\bar{\mu}, \mu^\star)$.[5]

3.4 | Runtime

Generally speaking, calculations using neural networks can benefit greatly from parallelization, for example, by employing GPUs. For most of our examples, this was not necessary however, and the respective calculations could be performed quickly (i.e., in between 1 and 5 min) even with a regular CPU (intel i5-7200U; dual core with 2.5–3.1 GHz each). For the DNB case study, however, a single run with stable learning parameters takes around 20 hr on a CPU. By utilizing a single GPU (Nvidia GeForce RTX 2080 Ti) this is reduced to around 30 min. Notably, in the smaller examples there was less speed-up when using GPU compared to CPU, the reason being that the problems were too small to fully use parallel capabilities of a GPU.

4 | EXAMPLES

The aim of this section is to illustrate how the above introduced concepts can be used to numerically solve given problems. In particular we demonstrate that neural networks are able to (a) achieve a satisfactory empirical performance for all problems considered, (b) naturally determine the structure of the worst-case distribution via Proposition 2.7(b), and (c) deal with problems, that cannot be reformulated as linear programs. Concerning the latter point, we consider both a function f, which cannot be written as the maximum of affine functions, as well as a cost function c, which is not additively separable. Additionally, we make a case for the generality of our duality result: we replace the distance constraint by a distance penalty and fix the distribution of bivariate, rather than univariate, marginals. Furthermore by considering unbounded functions f, we shed some light on the necessity of the growth functions κ used in Theorem 2.1. To achieve all of these points, we consider three examples with increasing difficulty.

Concerning the notation in this section, c denotes the cost function

$$c(x,y) = ||x-y||_1 = \sum_i |x_i - y_i|.$$

This notation implies that

$$d_c(\bar{\mu},\mu) := \inf_{\pi\in\Pi(\bar{\mu},\mu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \sum_{i=1}^{d} |x_i - y_i|\, \pi(dx,dy)$$

is the first-order Wasserstein distance with respect to the L_1-metric. On the other hand, we consider the first-order Wasserstein distance with respect to the Euclidean metric

$$d_{c_2}(\bar{\mu},\mu) := \left\{ \inf_{\pi\in\Pi(\bar{\mu},\mu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \left(\sum_{i=1}^{d} (x_i - y_i)^2 \right)^{1/2} \pi(dx,dy) \right\}. \tag{23}$$

Note that the cost function $c_2(x,y) := ||x-y||_2$ is not additively separable.[6]

4.1 | Expected maximum of two comonotone standard uniforms

We start our exemplification with a toy example that is not connected to risk measurement. Consider the following problem

$$\phi(f_1) := \sup_{\substack{\binom{V}{U}\sim\mu\in\Pi(\bar{\mu}_1,\bar{\mu}_2),\\ d_c(\bar{\mu},\mu)\le\rho}} \mathbb{E}[\max(U,V)] = \sup_{\substack{\mu\in\Pi(\bar{\mu}_1,\bar{\mu}_2),\\ d_c(\bar{\mu},\mu)\le\rho}} \int_{[0,1]^2} \max(x_1,x_2)\,\mu(dx), \tag{24}$$

where $\bar{\mu}_1 = \bar{\mu}_2 = \mathcal{U}([0,1])$ are (univariate) standard uniformly distributed probability measures and $\bar{\mu}$ is the comonotone copula. In other words, $\bar{\mu}$ is a bivariate probability measure with standard uniformly distributed marginals that are perfectly dependent. In the notation of the Section 2, we choose the function f as $f_1(x) = \max(x_1,x_2)$ and $X = X_1 \times X_2 = [0,1]\times[0,1]$. Interpreting problem (24), we aim to compute the expected value of the maximum of two standard Uniforms under ambiguity with respect to the reference dependence structure, which is given by the comonotone coupling. Problem (24) possesses the following analytic solution

$$\phi(f_1) = \frac{1+\min(\rho,0.5)}{2}.$$

The derivation of this solution can be found in Appendix A.3 and is based on the duality result in Corollary 2.2. Hence, problem (24) is well suited to benchmark the solution method based on neural networks. In comparison, we also solve the problem with linear programming. To be precise, we consider the following two methods:

1. We discretize the reference copula $\bar{\mu}$ (and thereby the marginal distributions $\bar{\mu}_1$ and $\bar{\mu}_2$) and solve the resulting dual problem by means of linear programming (see Corollary 2.5). There are two distinct ways to discretize $\bar{\mu}$:

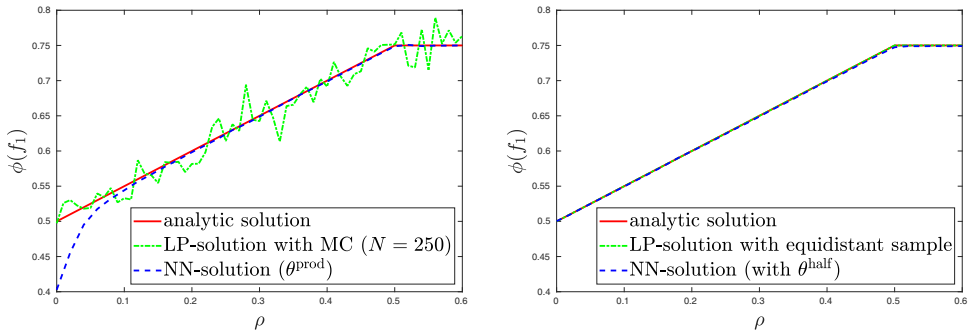

FIGURE 1 In the left panel, the analytic solution $\phi(f_1)$ of problem (24) is plotted as a function of ρ and compared to corresponding numerical solutions obtained by methods (1.a) and (2.a), which are described in Subsection 4.1. The right panel shows the same for the improved methods (1.b) and (2.b) [Color figure can be viewed at wileyonlinelibrary.com]

(a) We use Monte Carlo sampling. In the notation of Corollary 2.5, this means we sample n points $x_1^1, \dots, x_1^n$ in [0,1] from the standard Uniform distribution. Then, we set $x_2^j = x_1^j$ for $j = 1, \dots, n$.

(b) We set the points $x_1^j = x_2^j = \frac{2j-1}{2n}$ for $j = 1, \dots, n$. As the comonotonic copula lives only on the main diagonal of the unit square, this deterministic discretization of $\bar{\mu}$ in some sense minimizes the discretization error. The simple geometrical argument used to find this discretization can be applied only due to the special structure of the reference distribution at hand.

Let us emphasize that method (1.a) can be applied to any reference distribution $\bar{\mu}$. On the other hand, method (1.b) can only be used in this particular example as $\bar{\mu}$ is given by the comonotonic copula.

2. We solve the problem with the neural network approach described in the above Section 3. As discussed, some hyperparameters need to be chosen problem specific. In particular, we set: $N_0 = 15,000$, $N_{\text{fine}} = 5,000$, $\gamma = 1,280$, batch size $= 2^7$ and $\alpha_\lambda = .1$.[7] Concerning the sampling measure θ, for this example we compare
   (a) the basic choice $\theta = \theta^{\text{prod}}$ and
   (b) the improved choice $\theta = \theta^{\text{half}}$.

   To better understand these parameter choices and our neural network approach in general, we provide a detailed convergence analysis for this example in Appendix A.2.

Figure 1 compares the two above mentioned methods to solve problem (24) for different values of ρ. In the left panel of Figure 1, we observe that method (1.a) yields an unsatisfactory result even though $n = 250$ is chosen as large as possible for the resulting LP to be solvable by a commercial computer. This issue arises due to the poor quality of the discretization resulting from Monte Carlo simulation. If one chooses the discretization as done in method (1.b), we recover the analytic solution of problem (24) as can be seen in the right panel of Figure 1. Moreover, Figure 1 indicates that method (2), that is, the approach presented in this paper, yields quite good and stable results. The left panel, however, shows that for small ρ method (2.a) does not rediscover the true solution. The reason for this is that when drawing random samples from the chosen sampling measure θ^{prod}, it is unlikely that we sample from the relevant region, namely the main diagonal of the unit square. As discussed in Subsection 3.1, method (2.b) is designed to overcome precisely this weakness and the right panel of Figure 1 illustrates that it does.

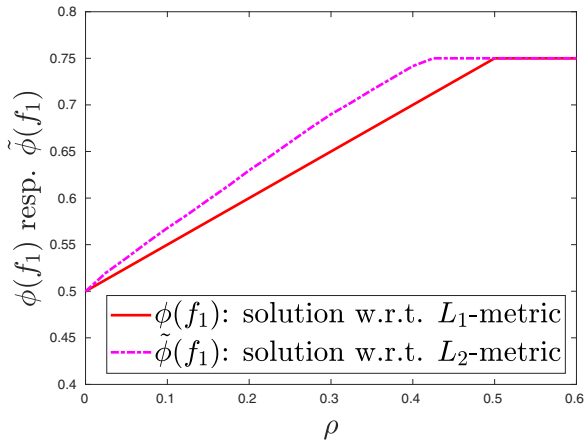

FIGURE 2 The analytic solution $\phi(f_1)$ of problem (24), which uses the first-order Wasserstein distance with respect to the L_1-metric, is compared to the numerical solution $\tilde{\phi}(f_1)$ of problem (25), which uses the first-order Wasserstein distance with respect to the Euclidean metric, that is, the L_2-metric [Color figure can be viewed at wileyonlinelibrary.com]

We finalize this example by considering the Wasserstein distance with respect to the Euclidean metric d_{c_2}, defined in Equation (23), rather than the Wasserstein distance with respect to the L_1 metric d_c. Thus, we compare problem (24) to

$$\tilde{\phi}(f_1) := \sup_{\substack{\mu\in\Pi(\bar{\mu}_1,\bar{\mu}_2),\\ d_{c_2}(\bar{\mu},\mu)\le\rho}} \int_{[0,1]^2} \max(x_1,x_2)\,\mu(dx). \tag{25}$$

As the cost function c_2 is not additively separable, $\tilde{\phi}(f_1)$—other than $\phi(f_1)$—cannot be approximated based on Corollary 2.5, that is, linear programming. Nevertheless, we can approximate $\tilde{\phi}(f_1)$ using neural networks, which demonstrates the flexibility of our approach.[8] Figure 2 compares $\phi(f_1)$ and $\tilde{\phi}(f_1)$ for different ρ. Note that as $c(x,y) \ge c_2(x,y)$ for all x, y, $d_c(\bar{\mu},\mu) \ge d_{c^2}(\bar{\mu},\mu)^{1/2}$ for all $\bar{\mu},\mu \in \mathcal{P}(X)$. Hence, $\phi(f_1) \le \tilde{\phi}(f_1)$ for fixed ρ. Figure 2 is in line with this observation.

4.2 | AVaR of two independent standard uniforms

We increase the level of complexity slightly compared to the previous example, as we now turn to robust risk aggregation. We aim to compute $\mathrm{AVaR}_\alpha(U+V)$, where U and V are *independent* standard Uniforms under ambiguity with respect to the independence assumption. Note that the Average Value at Risk (AVaR) is defined by

$$\mathrm{AVaR}_\alpha(Y) := \min_{\tau\in\mathbb{R}} \left\{ \tau + \frac{1}{1-\alpha}\mathbb{E}[\max(Y-\tau,0)] \right\},$$

see Rockafellar and Uryasev (2000). Using the first-order Wasserstein distance to construct an ambiguity set around the reference dependence structure, we are led to the following problem:

$$\Phi_2 := \sup_{\substack{\binom{V}{U}\sim\mu\in\Pi(\bar{\mu}_1,\bar{\mu}_2),\\ d_c(\bar{\mu},\mu)\le\rho}} \mathrm{AVaR}_\alpha(U+V) \tag{26}$$

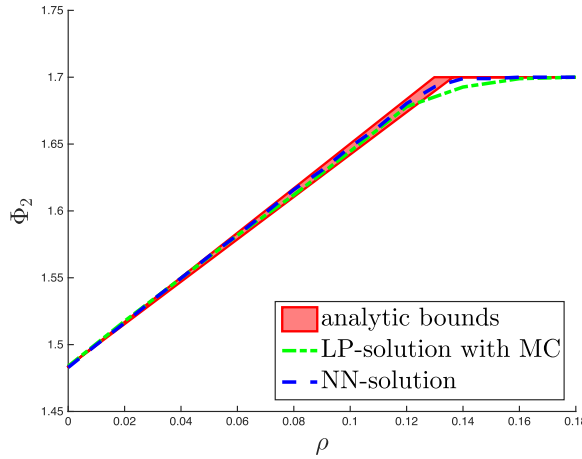

FIGURE 3 The analytic upper and lower bounds of problem (26) are compared to two distinct numerical solutions. The first numerical solution is obtained by Monte Carlo simulation with $n = 100$ sample points as well as linear programming and averaged over 100 simulations for each fixed ρ. The second numerical solution is obtained by penalization and neural networks. The confidence level of the AVaR considered in problem (26) is set to $\alpha = .7$ [Color figure can be viewed at wileyonlinelibrary.com]

$$= \sup_{\substack{\mu\in\Pi(\bar{\mu}_1,\bar{\mu}_2),\\ d_c(\bar{\mu},\mu)\le\rho}} \inf_{\tau\in\mathbb{R}} \left\{ \tau + \frac{1}{1-\alpha} \int_{[0,1]^2} \max(x_1 + x_2 - \tau, 0)\mu(dx) \right\} \tag{27}$$

$$= \inf_{\tau\in\mathbb{R}} \phi(f_2^\tau), \tag{28}$$

where $\bar{\mu}_1 = \bar{\mu}_2 = \mathcal{U}([0,1])$ are (univariate) standard uniformly distributed probability measures and $\bar{\mu}$ is the independence copula. In other words, $\bar{\mu} = \mathcal{U}([0,1]^2)$ is a bivariate probability measure with independent, standard uniformly distributed marginals. Moreover, we have that $f_2^\tau(x) = \tau + \frac{1}{1-\alpha}\max(x_1 + x_2 - \tau, 0)$ and $\phi(\cdot)$ is defined as in Equation (1).

Note that in the above formulation of the problem we can go from (27) to (28) as the problem is convex in τ and concave in μ and Wasserstein balls are weakly compact. Thus, we can apply Sion's Minimax Theorem to interchange the supremum and the infimum in (27).

In Appendix A.4, we derive an analytical upper and lower bound for Φ_2 in (26). These bounds are tight enough for the present purpose, which is to evaluate the performance of the two discussed numerical methods.

Figure 3 supports the latter claim: The analytic bounds for Φ_2 are rather tight when plotted as a function of ρ. The bounds are compared to the same two numerical methods as discussed in the previous example. With respect to the solution based on Monte Carlo simulation and linear programming, we now average over 100 simulations for each fixed ρ. Thus, the results in Figure 3 do not fluctuate as much as those we have seen in the left panel of Figure 1. Nevertheless, Figure 3 shows that the solution obtained via MC and LP does not stay within the analytic bounds—other than the solution based on our neural networks approach. Arguably this is due to the lack of symmetry when discretizing the reference distribution μ using Monte Carlo. Regarding runtime, both numerical methods take around the same time to calculate the values needed for Figure 3.

We now want to illustrate a further merit of the neural networks approach, namely that we can sample from the numerical optimizer $\mu^\star$ of problem (26). By doing so, we obtain information about the structure of the worst-case distribution. The samples are obtained by acceptance–rejection sampling from the density given by Proposition 2.7(b), where we replace true optimizers by numerical ones. Figure 4 plots samples of this worst-case distribution $\mu^\star$ for different values of ρ. To understand the intriguing nature of the results presented in Figure 4,

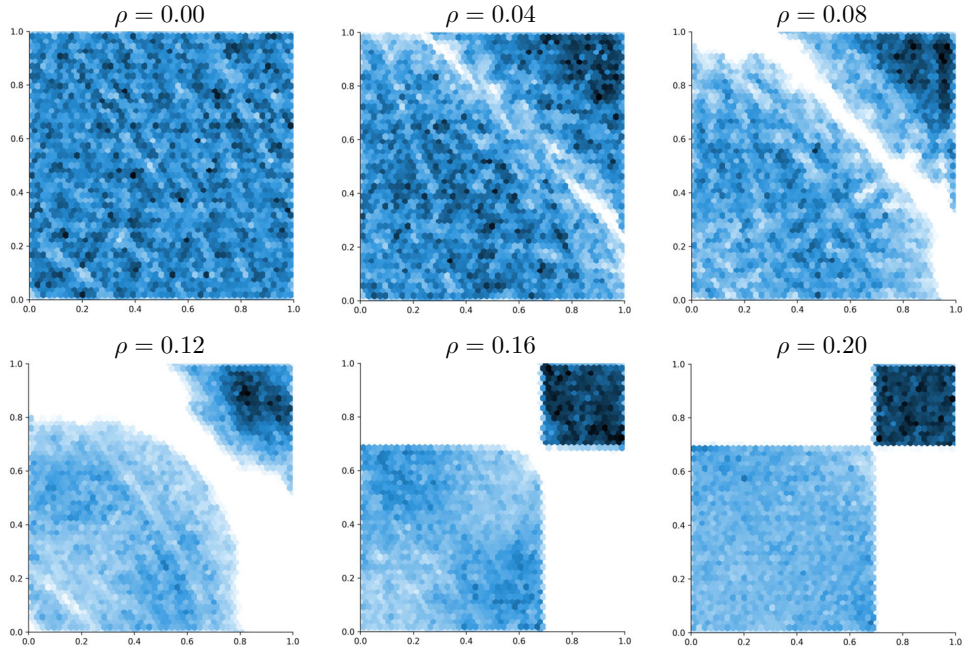

FIGURE 4 Samples from the optimizer $\mu^\star$ of problem (26) as obtained by the neural networks approach are shown in form of a heatplot for six different levels of ambiguity, that is, $\rho = 0, .04, .08, .12, .16, .2$ [Color figure can be viewed at wileyonlinelibrary.com]

we have to describe problem (26) in some more detail. It should be clear that the comonotone coupling of the Uniforms U and V is maximizing $\mathrm{AVaR}_\alpha(U+V)$ among all possible coupling of U and V. However, one can find many different maximizing couplings. Notably, the optimizer shown for $\rho = .2$ corresponds to the one which has the lowest relative entropy with respect to the independent coupling among the maximizers of $\mathrm{AVaR}_\alpha(U+V)$. On the other hand, the middle panel for $\rho = 0.16$ motivated us to derive a coupling that—among maximizers of $\mathrm{AVaR}_\alpha(U+V)$—we conjecture to have the lowest Wasserstein distance to the independent coupling. This coupling is used to derive the lower bound for problem (26) in Appendix A.4. Some features of the others couplings, for example, for $\rho = .08$ and $\rho = .12$ came as a surprise to us: For example, the curved lines as boundary for the support are unusual in an L_1-Wasserstein problem.

4.3 | Variance of three normally distributed random variables with distance penalization

We now leave the domain of uniformly distributed, univariate marginals and replace the distance constraint by a distance penalty. We analyze the following problem:

$$\begin{aligned}\chi(f_4) &:= \sup_{\binom{X}{Y}\sim\mu\in\Pi(\bar\mu_{12},\bar\mu_3)} \mathbb{V}\mathrm{ar}(X_1+X_2+Y) - \frac{1}{r}d_{\bar c}(\bar\mu,\mu)^r \\ &= \sup_{\mu\in\Pi(\bar\mu_{12},\bar\mu_3)} \int_{\mathbb{R}^3} \big((x_1+x_2+y)^2 - m^2\big)\mu(dx_1,dx_2,dy) - \frac{1}{r}d_{\bar c}(\bar\mu,\mu)^r, \end{aligned} \tag{29}$$

where the cost function $\tilde{c}(x, y) = 2||x - y||_1$.[9] We specify the reference distribution function as follows

$$\bar{\mu} = \mathcal{N}\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.8 & 0\\0.8 & 1 & 0\\0 & 0 & 1\end{pmatrix}\right).$$

In this examples, there are two novelties that are explained in the following.

First, the fact that we set $\mu \in \Pi(\bar{\mu}_{12}, \bar{\mu}_3)$, means we are fixing not only the univariate marginal distributions, which are standard normal, but also the dependence structure between the first and the second margin X_1 and X_2. In this case, we assume that X_1 and X_2 are jointly normal with correlation .8. We use $\bar{\mu}_{12}$ to denote the fixed, *bivariate* margin. As a consequence, the model ambiguity concerns solely the dependence structure between the third margin Y and the other two margins X_1 and X_2.

Second, rather than a distance constraint $d_{\tilde{c}}(\bar{\mu}, \mu) \leq \rho$, we now use a distance penalty to account for the described model ambiguity: we set $\varphi(x) = \frac{1}{r}x^r$ in Theorem 2.1. The parameter r accounts for the degree of penalization and hence is *not* comparable to the radius ρ of the Wasserstein balls described above. Instead, for $r \to \infty$ the penalization becomes closer and closer to the case where we impose the constraint $d_{\tilde{c}}(\bar{\mu}, \mu) \leq 1$.

These two specifications aim to demonstrate the value of the generality of Theorem 2.1 with respect to both the choice of polish spaces and the modeling of ambiguity.

Even though Subsection 2.2 focuses on the Wasserstein ball constraint, the solution method based on penalization and neural networks is trivially adapted to problems like (29). We state the resulting numerical solution of problem (29) for different values of r in Table 1. To make these results more concrete, we sampled 20,000 values from the respective worst-case distribution $\mu^\star$ and report the corresponding empirical covariance matrix $\hat{\Sigma}_{\mu^\star}$. Notably, the covariance matrix does not completely characterize $\mu^\star$, as $\mu^\star$ does not have to be a joint normal distribution.

5 | DNB CASE STUDY: AGGREGATION OF SIX GIVEN RISKS

Aas and Puccetti (2014) provide a very illustrative case study of the risk aggregation at the Den Norske Bank Bank (DNB), Norway's largest bank. We want to make use of this example to showcase the applicability of the novel framework presented in this paper.

The DNB is exposed to six different types of risks: credit, market, asset, operational, business, and insurance risk. Let the random variables $L_1, \dots, L_6$ represent the marginal risk exposures for these six risks. Per definition, risk aggregation is not concerned with the computation of the distribution of the marginal risks. Hence, we take the corresponding marginal distribution functions $F_1, \dots, F_6$ as given. In this particular case, F_1, F_2, and F_3 are empirical cdfs originating from given samples, while L_4, L_5, and L_6 are assumed to be log-normally distributed with given parameters, see Table 2.

For the purpose of risk management, the DNB needs to determine the capital to be reserved. According to the Basel Committee on Banking Supervision (2013), this capital requirement should be computed by the AVaR of the sum of these six losses.[10] The AVaR of the sum of these six losses

TABLE 1 Comparison of the numerical solutions $\chi(f_4)$ of problem (24), computed based on penalization and neural networks, for different values of r

	$\chi(f_4)$	$\int_{\mathbb{R}^3}(x_1+x_2+y)^2 d\mu^\star$	$d_{\tilde{c}}(\bar{\mu},\mu^\star)$	$\hat{\Sigma}_{\mu^\star}$
No penalization	4.6	4.6	0	$\begin{pmatrix} 1 & 0.8 & 0 \\ 0.8 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
$r=1$	6.16	8.08	1.92	$\begin{pmatrix} 0.998 & 0.801 & 0.847 \\ 0.801 & 1.008 & 0.853 \\ 0.847 & 0.853 & 1.011 \end{pmatrix}$
$r=2$	6.50	7.60	1.48	$\begin{pmatrix} 0.989 & 0.806 & 0.737 \\ 0.806 & 0.989 & 0.736 \\ 0.737 & 0.736 & 0.997 \end{pmatrix}$
$r=3$	6.57	7.29	1.29	$\begin{pmatrix} 0.975 & 0.795 & 0.675 \\ 0.795 & 0.991 & 0.682 \\ 0.675 & 0.682 & 0.980 \end{pmatrix}$
$r=4$	6.65	7.19	1.21	$\begin{pmatrix} 0.976 & 0.792 & 0.652 \\ 0.792 & 0.970 & 0.654 \\ 0.652 & 0.654 & 0.986 \end{pmatrix}$
$r=\infty$	6.76	6.76	1.00	$\begin{pmatrix} 0.991 & 0.803 & 0.554 \\ 0.803 & 0.998 & 0.551 \\ 0.554 & 0.551 & 0.993 \end{pmatrix}$

We define the worst-case distribution $\mu^\star \in \Pi(\bar{\mu}_{12},\bar{\mu}_3)$ such that $\chi(f_4)=\int_{\mathbb{R}^3}(x_1+x_2+y)^2\mu^\star(dx_1,dx_2,dy)-\frac{1}{r}d_{\tilde{c}}(\bar{\mu},\mu^\star)^r$ and report also the empirical covariance matrix $\hat{\Sigma}_{\mu^\star}$ computed from $N=20,000$ samples of $\mu^\star$. The case $r=\infty$ corresponds to the constraint $d_{\tilde{c}}(\bar{\mu},\mu)\le 1$.

TABLE 2 Overview of the information concerning the reference distribution in the DNB case study

	Description	**Type**	**Parameters/other details**
F_1	cdf of credit risk L_1	Empirical cdf	Given by 2.5 million samples; $SD\ \bar{\sigma}_1 = 644.602$
F_2	cdf of market risk L_2	Empirical cdf	Given by 2.5 million samples; $SD\ \bar{\sigma}_2 = 5,562.362$
F_3	cdf of asset risk L_3	Empirical cdf	Given by 2.5 million samples; $SD\ \bar{\sigma}_3 = 1,112.402$
F_4	cdf of operational risk L_4	Lognormal cdf	Mean $\bar{m}_4 = 840.735$; $SD\ \bar{\sigma}_4 = 694.613$
F_5	cdf of business risk L_5	Lognormal cdf	Mean $\bar{m}_5 = 743.345$; $SD\ \bar{\sigma}_5 = 465.064$
F_6	cdf of insurance risk L_6	Lognormal cdf	Mean $\bar{m}_6 = 438.978$; $SD\ \bar{\sigma}_6 = 111.011$
C_0	Reference copula linking $L_1,\dots,L_6$	Student-t copula	With 6 degrees of freedom and correlation matrix Σ_0

The correlation matrix Σ_0 is given in Appendix A.5. F_i denotes the cumulative distribution function of the marginal probability measure $\bar{\mu}_i$ for $i=1,\dots,6$.

TABLE 3 Comparison of AVaRs for different dependence structures

$\inf_{C\in\mathcal{C}} \mathrm{AVaR}_\alpha^C(L_6^+)$	$\mathrm{AVaR}_\alpha^\Pi(L_6^+)$	$\mathrm{AVaR}_\alpha^{C_0}(L_6^+)$	$\sup_{C\in\mathcal{C}} \mathrm{AVaR}_\alpha^C(L_6^+)$
24,165.52	26,980.64	30,498.94	36,410.12

Note that we set $\alpha = 0.95$. We use the rearrangement algorithm (see Aas and Puccetti, 2014) to approximate $\inf_{C\in\mathcal{C}} \mathrm{AVaR}_\alpha^C(L_6^+)$, while $\sup_{C\in\mathcal{C}} \mathrm{AVaR}_\alpha^C(L_6^+) = \sum_{i=1}^6 \mathrm{AVaR}_\alpha(L_i)$.
The two remaining entries are computed by averaging over 50 simulation runs where 10 million sample points are drawn in each run. Note that Π denotes the independence copula. Thus, $\mathrm{AVaR}_\alpha^\Pi(L_6^+)$ corresponds to the AVaR of the sum of the six losses given that they are independent.

at a specific confidence level α is defined as

$$\mathrm{AVaR}_\alpha\left(L_6^+\right) = \min_{\tau\in\mathbb{R}} \left\{ \tau + \frac{1}{1-\alpha}\mathbb{E}[\max(L_6^+ - \tau, 0)] \right\}, \tag{30}$$

where $L_6^+ := \sum_{i=1}^6 L_i$. To evaluate expression (30), the joint distribution of $L_1, \dots, L_6$ is needed. As the marginal distributions of $L_1, \dots, L_6$ are known, the DNB relies on the concept of copulas to model the dependence structure between these risks. From the above description, it is clear that joint observations of the $L_1, \dots, L_6$ are not available. Hence, standard techniques to determine the copula, for example, by fitting a copula family and the corresponding parameters to a multivariate data set, do not apply. A panel of experts at the DNB therefore chooses a specific *reference copula* C_0, in this case a student-t copula with six degrees of freedom and a particular correlation matrix. Such an approach is common in practice and referred to as *expert opinion*.

From an academic point of view, this method for risk aggregation is not very satisfying due to the fact that the experts' choice of a *reference dependence structure* between the different risk types might be very inaccurate. Hence, we say that there is *model ambiguity* with respect to the dependence structure. It should be emphasized that a misspecification of this *reference copula* chosen by expert opinion can have a significant impact on the aggregated risk and therefore on the required capital. Table 3 supports this statement by comparing the AVaR implied by the reference copula C_0 to the AVaR implied by other dependence structures: Without any information regarding the dependence structure between the six risk, the lower (respectively, upper) bound for the AVaR with confidence level $\alpha = 0.95$ is 24,165.52 (respectively, 36,410.12) million Norwegian kroner. Similar bounds are studied in Aas and Puccetti (2014). As we pointed out in the literature review in Subsection 1.3, these bounds have been criticized in the literature as they are too far apart for practical purposes. We therefore apply the results derived in this paper to compute bounds for the AVaR that depend on the level ρ of distrust concerning the reference copula C_0. Alternatively, the parameter ρ can be understood as the level of ambiguity with respect to the reference distribution $\bar{\mu}$.

We define the probability measure $\bar{\mu}$ of the reference distribution by the following joint cumulative distribution function

$$\bar{F}(x) = C_0(F_1(x_1), F_2(x_2), \dots, F_6(x_6),$$

for all $x \in \mathbb{R}^6$. Hence, the cdfs of the marginals $\bar{\mu}_i$ are given by $F_i(\cdot)$ for $i = 1, 2, \dots, 6$. The problem of interest can be formulated as follows:

$$\underline{\Phi}_4^{C_0}(\alpha, \rho) := \inf_{\substack{L_6^+\sim\mu\in\Pi(\bar{\mu}_1,\dots,\bar{\mu}_6),\\ d_c(\bar{\mu},\mu)\le\rho}} \mathrm{AVaR}_\alpha\left(L_6^+\right), \tag{31}$$

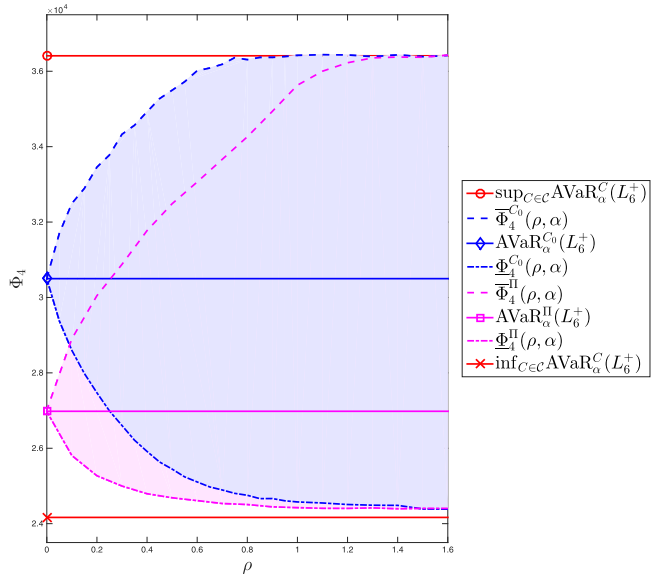

FIGURE 5 We consider two distinct reference dependence structures, the student-t copula C_0 defined in Table 2 and the independence copula Π. The corresponding robust solutions $\underline{\Phi}_4^{C_0}(\alpha,\rho)$ and $\overline{\Phi}_4^{C_0}(\alpha,\rho)$, defined in (31), and (32), respectively, $\underline{\Phi}_4^{\Pi}(\alpha,\rho)$ and $\overline{\Phi}_4^{\Pi}(\alpha,\rho)$, defined analogously, are plotted as a function of the level of ambiguity ρ. We compare these results, which were computed relying on the concept presented in this paper, to the known values of $\mathrm{AVaR}_\alpha(L_6^+)$ given in Table 2. Note that we fix $\alpha = .95$ [Color figure can be viewed at wileyonlinelibrary.com]

$$\overline{\Phi}_4^{C_0}(\alpha,\rho) := \sup_{\substack{L_6^+\sim\mu\in\Pi(\bar{\mu}_1,\ldots,\bar{\mu}_6),\\ d_c(\bar{\mu},\mu)\le\rho}} \mathrm{AVaR}_\alpha\left(L_6^+\right). \tag{32}$$

The cost function c defining the transportation distance d_c in problem (31) and (32) is set to

$$c(x,y) = \sum_{i=6}^{d} \frac{|x_i - y_i|}{\bar{\sigma}_i}, \tag{33}$$

where $\bar{\sigma}_i$ denotes the standard deviation of $\bar{\mu}_i$ and is given in Table 2. The rationale behind this definition of c is that we want to model the ambiguity such that it concerns solely the dependence structure of the reference distribution. Definition (33) is a simple way to achieve this.[11]

Figure 5 shows the numerical solutions of problems (31) and (32), which are computed relying on penalization and neural networks, as a function of ρ and for $\alpha = .95$. As a comparison, the same problem is also solved with respect to the independence coupling Π rather than the reference copula C_0 described in Table 2. The shaded regions outline the possible levels of risk for a given level of ambiguity ρ and the two reference structures. On one hand, the evolution of the risk levels in ρ, combined with the given optimizers of problems (31) and (32) can be used as an informative tool to better understand the risk the DNB is exposed to. On the other hand, if a certain level of ambiguity is justified in practice, the bank can assign their capital based on the corresponding worst-case value. If for example $\rho = .1$ is decided on, the bank would have to assign 32,490 capital compared to 30,499 as dictated by the reference structure C_0.

Analytically, one striking feature of the numerical solution with respect to C_0 is worth pointing out: The absolute upper bound is attained already for $\rho \approx .8$, while the distance from the reference measure to the comonotone joint distribution can be calculated to be around 1.7. This underlines the fact that even though the comonotone distribution is a maximizer of the worst-case AVaR, there are several more, and they may be significantly more plausible structurally than the comonotone one.

In conclusion, this paper introduces a flexible framework to aggregate different risks while accounting for ambiguity with respect to the chosen dependence structure between these risks.

The proposed numerical method allows us to perform this task without making restrictive assumptions about either the particular form of the aggregation functional, or the considered distributions, or the specific way to account for the model ambiguity.

ACKNOWLEDGMENTS

The authors confirm that the data supporting the findings of this study are available within the article by Aas and Puccetti (2014). They are grateful to Daniel Bartl, Ludovic Tangpi, Ruodu Wang, and the participants of the numerous conference and seminars, where they presented this paper, for helpful comments as well as interesting discussions. Moreover, Mathias Pohl thanks Philipp Schmocker for his help and acknowledges support by the Austrian Science Fund (FWF) under the grant P28661 and Stephan Eckstein sincerely thanks Jan Obłój for his hospitality. Finally, the authors thank the two referees for their helpful comments.

ORCID

Mathias Pohl https://orcid.org/0000-0001-5209-676X

ENDNOTES

[1] Note also that our methods can be applied to solve completely unrelated problems, such as the portfolio selection problem under dependence uncertainty introduced in Pflug and Pohl (2017).

[2] This is, for example, satisfied if all κ_i have compact sublevel sets, as the sublevel sets are by continuity closed and further by positivity of κ_i it holds $\{\kappa \leq c\} \subseteq \{\kappa_1 \leq c\} \times \cdots \times \{\kappa_d \leq c\}$.

[3] Note that $\delta_x(A) = 1$ if $x \in A$, and $\delta_x(A) = 0$ otherwise.

[4] By definition, $f^{\lambda c}$ is lower semicontinuous. Moreover, if $c(x, y) = \bar{c}(x - y)$ for a continuous function $\bar{c} : X \to [0, \infty)$ with compact sublevel sets, then $f^{\lambda c}$ is upper semicontinuous and therefore continuous. This, for instance, holds for $\bar{c}(x) = \sum_{i=1}^{d} |x_i|$ or $\bar{c}(x) = \sum_{i=1}^{d} |x_i|^2$ corresponding to the first and second-order Wasserstein distance on $\mathbb{R}^d$.

[5] Clearly, $d_c(\bar{\mu}, \mu^\star)$ should be bounded by ρ and in the optimum equal to ρ (if one is not already in the edge case where ρ is so large that it does not have an effect).

[6] In the literature, the Wasserstein distance with respect to the Euclidean metric is usually associated with order two, in which case the underlying cost function *is* additively separable.

[7] We wait for 2,500 iterations until updating λ for the first time where λ is initialized to $\lambda = .75$

[8] We use the same choices for the hyperparameters as given in point 2. above, but increase N_0 considerably for small ρ to guarantee convergence of λ, and use θ^{half} as a sampling measure.

[9] One can think of the factor 2 occurring in the cost function similarly to a particular choice of ρ for the radius of the Wasserstein ball.

[10] Aas and Puccetti (2014) focus on the VaR rather than the AVaR. As the Basel Committee on Banking Supervision recently shifted the quantitative risk metrics system from VaR to Expected Shortfall (see Chang, Jiménez-Martín, Maasoumi, McAleer, & Pérez-Amaral, 2019), which is equivalent to the AVaR, we consider the AVaR in our study.

[11] It should be mentioned that Gao and Kleywegt (2017a) promote the definition $c(x, y) = \sum_{i=1}^{6} |F_i(x_i) - F_i(y_i)|$, which implies that the transportation distance d_c is defined directly on the level of copulas. Even if this approach is arguably more intuitive, we stick to definition (33) mainly for the sake of computational efficiency.

[12] Note that $|u_1 - u_2|$ is the distance (and thereby the cost of transportation) of any point-mass $C(u_1, u_2)$ to the main diagonal $M(u_1, u_2)$.

REFERENCES

Aas, K., & Puccetti, G. (2014). Bounds for total economic capital: The DNB case study. *Extremes*, *17*(4), 693–715.

Bartl, D., Cheridito, P., & Kupper, M. (2019). Robust expected utility maximization with medial limits. *Journal of Mathematical Analysis and Applications*, *471*(1), 752–775.

Bartl, D., Drapeau, S., & Tangpi, L. (2019). Computational aspects of robust optimized certainty equivalents and option pricing. *Mathematical Finance*, *30*(1), 287–309.

Basel Committee on Banking Supervision. (2013). *Consultative document, fundamental review of the trading book: A revised market risk framework*. Retrieved from https://www.bis.org/publ/bcbs265.pdf.

Bayraksan, G., & Love, D. K. (2015). Data-driven stochastic programming using phi-divergences. *Tutorials in Operations Research*, 1–19. https://doi.org/10.1287/Feduc.2015.0134

Beck, C., Becker, S., Grohs, P., Jaafari, N., & Jentzen, A. (2018). Solving stochastic differential equations and Kolmogorov equations by means of deep learning. Preprint, arXiv:1806.00421.

Becker, S., Cheridito, P., & Jentzen, A. (2019). Deep optimal stopping. *Journal of Machine Learning Research*, *20*(74), 1–25.

Bernard, C., Rüschendorf, L., & Vanduffel, S. (2017). Value-at-risk bounds with variance constraints. *Journal of Risk and Insurance*, *84*(3), 923–959.

Bernard, C., Rüschendorf, L., Vanduffel, S., & Wang, R. (2017). Risk bounds for factor models. *Finance and Stochastics*, *21*(3), 631–659.

Berner, J., Grohs, P., & Jentzen, A. (2018). Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations. Preprint, arXiv:1809.03062.

Blanchet, J., Kang, Y., & Murthy, K. (2016). Robust Wasserstein profile inference and applications to machine learning. Preprint, arXiv:1610.05627.

Blanchet, J., & Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, *44*(2), 565–600.

Buehler, H., Gonon, L., Teichmann, J., & Wood, B. (2019). Deep hedging. *Quantitative Finance*, *19*(8), 1271–1291.

Chang, C.-L., Jiménez-Martín, J.-Á., Maasoumi, E., McAleer, M., & Pérez-Amaral, T. (2019). Choosing expected shortfall over VaR in Basel III using stochastic dominance. *International Review of Economics & Finance*, *60*, 95–113.

Chen, Z., Yu, P., & Haskell, W. B. (2019). Distributionally robust optimization for sequential decision-making. *Optimization*, *68*(12), 2397–2426.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26* (pp. 2292–2300). Curran Associates, Inc. http://papers.nips.cc/paper/4927-sinkhorn-distances-lightspeed-computation-of-optimal-transport.pdf

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

Eckstein, S., & Kupper, M. (2019). Computation of optimal transport and related hedging problems via penalization and neural networks. *Applied Mathematics & Optimization*. https://doi.org/10.1007/s00245-019-09558-1

Embrechts, P., & Puccetti, G. (2006). Bounds for functions of dependent risks. *Finance and Stochastics*, *10*(3), 341–352.

Embrechts, P., Wang, B., & Wang, R. (2015). Aggregation-robustness and model uncertainty of regulatory risk measures. *Finance and Stochastics*, *19*(4), 763–790.

Gao, R., Chen, X., & Kleywegt, A. J. (2017). Wasserstein distributional robustness and regularization in statistical learning. Preprint, arXiv:1712.06050.

Gao, R., & Kleywegt, A. J. (2016). Distributionally robust stochastic optimization with Wasserstein distance. Preprint, arXiv:1604.02199.

Gao, R., & Kleywegt, A. J. (2017a). *Data-driven robust optimization with known marginal distributions*. Working paper.

Gao, R., & Kleywegt, A. J. (2017b). Distributionally robust stochastic optimization with dependence structure. Preprint, arXiv:1701.04200.

Genevay, A., Peyré, G., & Cuturi, M. (2017). Learning generative models with sinkhorn divergences. Preprint, arXiv:1706.00292.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of Wasserstein gans. In *Advances in neural information processing systems* (pp. 5767–5777).

Hanasusanto, G. A., & Kuhn, D. (2018). Conic programming reformulations of two-stage distributionally robust linear programs over Wasserstein balls. *Operations Research*, *66*(3), 849–869.

Henry-Labordere, P. (2017). *Deep primal-dual algorithm for BSDEs: Applications of machine learning to CVA and IM*. Available at SSRN: 3071506.

Henry-Labordere, P. (2019). *(Martingale) optimal transport and anomaly detection with neural networks: A primal-dual algorithm*. Available at SSRN 3370910.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251–257.

Liebscher, E. (2014). Copula-based dependence measures. *Dependence Modeling*, *2*(1). https://doi.org/10.2478/demo-2014-0004

Lux, T., & Papapantoleon, A. (2016). Model-free bounds on value-at-risk using partial dependence information. Preprint, arXiv:1610.09734.

Makarov, G. (1982). Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory of Probability & its Applications*, *26*(4), 803–806.

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

Nelsen, R. B. (2007). *An introduction to copulas*. New York: Springer Science & Business Media.

Nesterov, Y. (2012). Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, *22*(2), 341–362.

Norbert (https://math.stackexchange.com/users/19538/norbert). (2013). *When does convergence in L^p imply convergence of the p-th moment*? Mathematics Stack Exchange. Retrieved from https://math.stackexchange.com/q/475146 (version: 2013-08-25).

Obloj, J., & Wiesel, J. (2018). Statistical estimation of superhedging prices. Preprint, arXiv:1807.04211.

Pflug, G., & Wozabal, D. (2007). Ambiguity in portfolio selection. *Quantitative Finance*, *7*(4), 435–442.

Pflug, G. C., & Pohl, M. (2017). A review on ambiguity in stochastic portfolio optimization. *Set-Valued and Variational Analysis*, 26(4), 733–757.

Puccetti, G., & Rüschendorf, L. (2012a). Bounds for joint portfolios of dependent risks. *Statistics & Risk Modeling with Applications in Finance and Insurance*, *29*(2), 107–132.

Puccetti, G., & Rüschendorf, L. (2012b). Computation of sharp bounds on the distribution of a function of dependent risks. *Journal of Computational and Applied Mathematics*, *236*(7), 1833–1840.

Puccetti, G., & Wang, R. (2015). Extremal dependence concepts. *Statistical Science*, *30*(4), 485–517.

Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, *2*, 21–42.

Roth, K., Lucchi, A., Nowozin, S., & Hofmann, T. (2017). Stabilizing training of generative adversarial networks through regularization. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 2018–2028). Curran Associates, Inc. http://papers.nips.cc/paper/6797-stabilizing-training-of-generative-adversarial-networks-through-regularization.pdf

Rüschendorf, L. (1982). Random variables with maximum sums. *Advances in Applied Probability*, *14*(3), 623–632.

Rüschendorf, L. (2017). Risk bounds and partial dependence information. In *From statistics to mathematical finance* (pp. 345–366). Berlin: Springer. https//doi.org/10.1007/978-3-319-50986-0_17

Seguy, V., Damodaran, B. B., Flamary, R., Courty, N., Rolet, A., & Blondel, M. (2017). Largescale optimal transport and mapping estimation. Preprint, arXiv:1711.02283.

Shapiro, A. (2017). Distributionally robust stochastic programming. *SIAM Journal on Optimization*, *27*(4), 2258–2275.

Weinan, E., Han, J., & Jentzen, A. (2017). Deep learning-based numerical methods for highdimensional parabolic partial differential equations and backward stochastic differential equations. *Communications in Mathematics and Statistics*, *5*(4), 349–380.

Yang, I. (2017). A convex optimization approach to distributionally robust Markov decision processes with Wasserstein distance. *IEEE Control Systems Letters*, *1*(1), 164–169.

Zhao, C., & Guan, Y. (2018). Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, *46*(2), 262–267.

How to cite this article: Eckstein S, Kupper M, Pohl M. Robust risk aggregation with neural networks. *Mathematical Finance*. 2020;30:1229–1272. https://doi.org/10.1111/mafi.12280

APPENDIX A

A.1 | A basic inequality between difference in *p*th moment and difference in *p*-norm

The following result is used in Remark 2.11. The statement and proof are taken from Norbert (2013).

Lemma A.1. *Let $p \in \mathbb{N}$ and $X, Y \in L^p$, then*

$$\|X^p - Y^p\|_1 \leq \|X - Y\|_p \sum_{k=0}^{p-1} \|X\|_p^k \|Y\|_p^{p-1-k}.$$

Proof. For $p = 1$, the inequality obviously holds. Let $p \geq 2$, then

$$\|X^p - Y^p\|_1 = \left\| (X - Y) \sum_{k=0}^{p-1} X^k Y^{p-1-k} \right\|_1 \leq \|X - Y\|_p \sum_{k=0}^{p-1} \left\| |X|^{kq} |Y|^{(p-1-k)q} \right\|_1^{1/q}. \tag{A.1}$$

It holds for all $k \in \{0, \dots, p-1\}$ that

$$\left\| |X|^{kq} |Y|^{(p-1-k)q} \right\|_1 \leq \|X\|_p^{kq} \|Y\|_p^{p-kq}. \tag{A.2}$$

For $k \in \{0, p-1\}$, (A.2) is immediate. For $0 < k < p-1$ (A.2) follows by Hölder's inequality applied with $r = p/kq$. Putting (A.1) and (A.2) together, we obtain

$$\|X^p - Y^p\|_1 \leq \|X - Y\|_p \sum_{k=0}^{p-1} \left(\|X\|_p^{kq} \|Y\|_p^{p-kq} \right)^{1/q} \leq \|X - Y\|_p \sum_{k=0}^{p-1} \|X\|_p^k \|Y\|_p^{p-1-k}$$

and thus the claim. □

A.2 | Convergence analysis for Example 4.1

This section is meant to demonstrate how to assess the quality of the obtained numerical solution. We consider the case $\rho = .25$ in Figure 1. We use the sampling measure θ^{prod} and compare the following three parameter settings:

(i) $\gamma = 100$, batch size = 1,024, $N_0 = 15,000$, and $N_{\text{fine}} = 5,000$.
(ii) $\gamma = 2,500$, batch size = 1,024, $N_0 = 15,000$, and $N_{\text{fine}} = 5,000$.
(iii) $\gamma = 2,500$, batch size = 16, $N_0 = 7,500$, and $N_{\text{fine}} = 2,500$.

Figure A.1 examines the contrast between setting (i) and (ii). As can be seen, in both settings the algorithm appears to converge in a stable way. In setting (i), there is, however, an apparent difference between the dual value $\phi_{\theta,\gamma}(f_1)$ and the primal value $\int f_1 d\mu^\star$, which is computed using the worst-case distribution $\mu^\star$. We can therefore conclude that the penalization in setting (i) is insufficient, that is, the penalization parameter γ is chosen too low. This is clearly not the case for setting (ii). Figure A.2 shows that both a small batch size and a small number of iterations lead to bad numerical behavior.

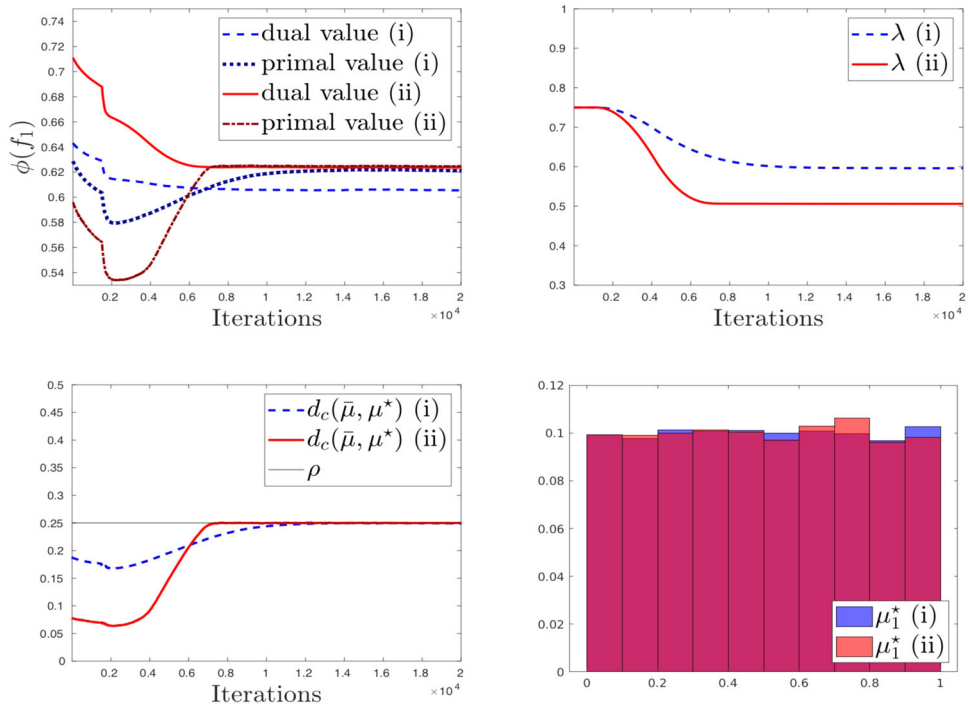

FIGURE A.1 Comparison of the parameter setting (i) with $\gamma = 100$ and (ii) with $\gamma = 2,500$, while batch size = 1,024, $N_0 = 15,000$, and $N_{\text{fine}} = 5,000$ in both settings. The upper left panel shows the dual value $\phi_{\theta,\gamma}(f_1)$ as well as the primal value $\int f_1 d\mu^\star$. The upper right, respectively, lower left panel illustrates the convergence of λ, respectively, $d_c(\bar{\mu}, \mu^\star)$. The lower right panel plots 5,000 samples from the first marginal $\mu_1^\star$ of the worst-case distribution $\mu^\star$. Note that this histogram is also representative for the second marginal $\mu_2^\star$. The computation time is 205 s in both cases [Color figure can be viewed at wileyonlinelibrary.com]

A.3 | Proof for Section 4.1

We want to derive the analytic solution of problem (24). To do so, the concept of copulas turns out to be rather useful. We refer to Nelsen (2007) for an introduction to this topic. Let $\mathcal{C}$ denote the set of all copulas and let the comonotonic copula be denoted by $M(u_1, u_2) = \min(u_1, u_2)$, for all $u_1, u_2 \in [0, 1]$. Using this notation, we can rewrite problem (24) and show the following:

$$\phi_1(f) = \sup_{\substack{C \in \mathcal{C}, \\ d_c(M,C) \le \rho}} \int_{[0,1]^2} \max(u_1, u_2) dC(u_1, u_2) = \frac{1 + \min(\rho, 0.5)}{2}.$$

Proof. First, we derive an upper bound for $d_c(M, C)$, where $C \in \mathcal{C}$. As M lives on the main diagonal of the unit square, the vertical (or horizontal) projection of the mass of an arbitrary copula $C \in \mathcal{C}$ onto M is feasible transportation plan with costs $\int_{[0,1]^2} |u_1 - u_2| dC(u)$.[12] The latter expression appears in the definition of a concordance measure called *Spearman's footrule* and it known to be maximized by the countermonotonic copula $W(u_1, u_2) := \max(u_1 + u_2 - 1, 0)$ for all $u_1, u_2 \in [0, 1]$ (see Liebscher, 2014). Hence, we obtain

$$d_c(M, C) \le \int_{[0,1]^2} |u_1 - u_2| dC(u) \le \int_{[0,1]^2} |u_1 - u_2| dW(u) = \int_0^1 |2u_1 - 1| du_1 = 0.5.$$

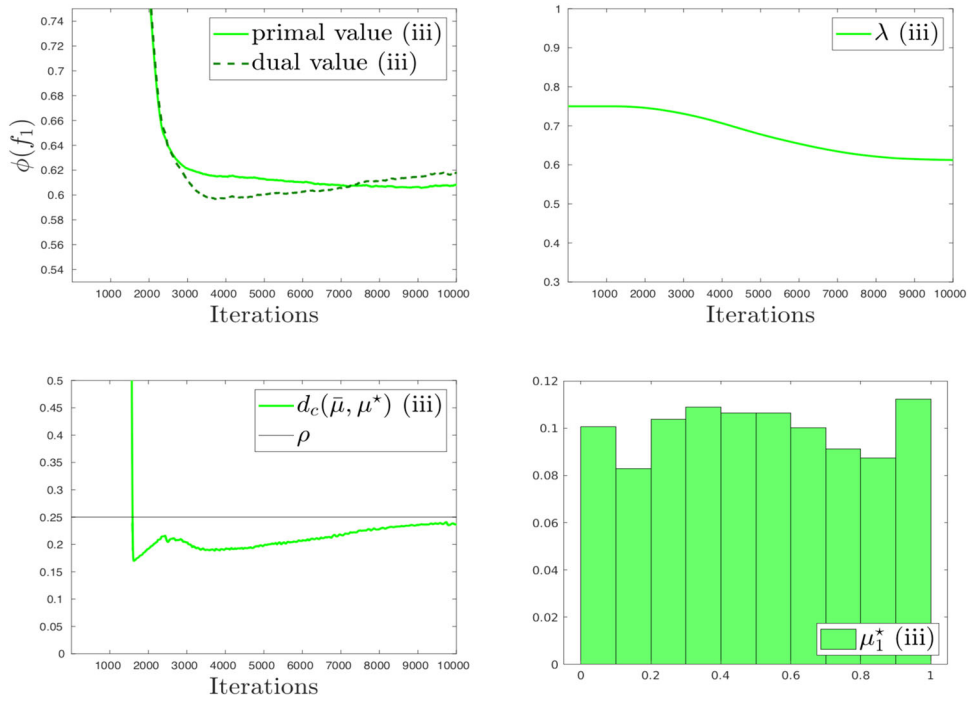

FIGURE A.2 Convergence analysis of the parameter setting (iii) where $\gamma = 2,500$, batch size $= 16$, $N_0 = 7,500$, and $N_{\text{fine}} = 2,500$. The upper left panel shows the dual value $\phi_{\partial,\gamma}(f_1)$ as well as the primal value $\int f_1 d\mu^\star$. The upper right, respectively, lower left panel illustrates the convergence of λ, respectively, $d_c(\bar{\mu}, \mu^\star)$. The lower right panel plots 5,000 samples from the first marginal $\mu_1^\star$ of the worst-case distribution $\mu^\star$. Note that this histogram is also representative for the second marginal $\mu_2^\star$. The computation time is 45 s [Color figure can be viewed at wileyonlinelibrary.com]

Second, we show that this upper bound is attend for $d_c(M, W)$. The Kantorovich Rubinstein duality yields that

$$d_c(M,W) = \sup_{|h(u)-h(v)|\le c(u,v)} \int_{[0,1]^2} h(u)dM(u) - \int_{[0,1]^2} h(v)dW(v)$$
$$\ge \int_{[0,1]^2} u_1 + u_2 dM(u) - \int_{[0,1]^2} v_1 + v_2 dW(v) = 1 - 0.5 = 0.5,$$

where we simply set $h(u) = u_1 + u_2$ to obtain the inequality. As $d_c(M, C) \le .5$ for all $C \in \mathcal{C}$, we have $d_c(M, W) = .5$.

Combining these two observations yields for $\rho > .5$ that

$$\phi(f_1) = \sup_{C\in\mathcal{C}} \int_{[0,1]^2} \max(u_1,u_2)dC(u_1,u_2) = \int_{[0,1]^2} \max(u_1,u_2)dW(u_1,u_2) = \frac{3}{4}.$$

It follows that we can assume $\rho \le .5$ for the remainder of the proof.

Let us define the copula R_α as follows:

$$R_\alpha(u_1,u_2)=\begin{cases} W(u_1,u_2) & \text{if } \frac{1-\alpha}{2}\le u_1,u_2\le\frac{1+\alpha}{2} \\ M(u_1,u_2) & \text{else}\end{cases},$$

for $\alpha\in[0,1]$. Using the same projection-argument as in the beginning of the proof, it follows that

$$d_c(M,R_\alpha)\le\int_{[0,1]^2}|u_1-u_2|dR_\alpha(u)=\int_{(1-\alpha)/2}^{(1+\alpha)/2}|2u_1-1|du_1=\alpha^2/2.$$

Thus, $d_c(M,R_{\sqrt{2\rho}})\le\rho$, which implies

$$\phi(f_1)\ge\int_{[0,1]^2}\max(u_1,u_2)dR_{\sqrt{2\rho}}(u_1,u_2)=\frac{1+\rho}{2}.$$

By Corollary 2.2, we have

$$\varphi(f_1)=\inf_{\lambda\ge0,h_i\in C([0,1])}\left\{\lambda\rho+\sum_{i=1}^2\int_0^1 h_i(u_i)du_i\right.$$
$$\left.+\int_{[0,1]^2}\sup_{v\in[0,1]^2}\left[\max(v_1,v_2)-\sum_{i=1}^2 h_i(v_i)-\lambda\sum_{i=1}^2|u_i-v_i|\right]dM(u)\right\}.$$

Plugging in the value $\lambda=.5$ and setting $h_1(u)=h_2(u)=u/2$, yields $\phi_1(f)\le\frac{\rho}{2}+\frac{1}{2}+0$. □

A.4 | Proof for Section 4.2

We now derive the analytic bounds for Φ_2, that is, the solution of problem (26), which are plotted in Figure 3.

Let us start by proving the following upper bound

$$\Phi_2\le\min\left(1+\alpha,2-\frac{2}{3}\sqrt{2-2\alpha}+\frac{\rho}{2(1-\alpha)}\right),\tag{A.3}$$

where Φ_2 is defined in (26).

Proof. Due to Corollary 2.2,

$$\Phi_2=\inf_{\tau,\lambda\ge0,h_i\in C([0,1])}\left\{\lambda\rho+\sum_{i=1}^2\int_0^1 h_i(u_i)du_i\right.\tag{A.4}$$
$$\left.+\int_{[0,1]^2}\sup_{v\in[0,1]^2}\left[\tau+\frac{1}{1-\alpha}\max(v_1+v_2-\tau,0)-\sum_{i=1}^2 h_i(v_i)-\lambda\sum_{i=1}^2|u_i-v_i|\right]d(u_1,u_2)\right\}.$$

The following choice of optimizers in Equation (A.4) yields the upper bound for Φ_2 given in (A.3):

$$\lambda = \frac{1}{2(1-\alpha)}, \qquad \tau = \tau^\star := 2 - \sqrt{2-2\alpha} \quad \text{and} \quad h_i(v) = \frac{1}{1-\alpha}\left(v - \frac{\alpha\tau^\star}{2}\right) \text{ for } i = 1, 2.$$

□

We now derive the following lower bound:

$$\Phi_2 \geq \min\left(1+\alpha, 2 - \frac{2}{3}\sqrt{2-2\alpha} + \frac{2(-3+2\sqrt{2-2\alpha}+3\alpha)\rho}{3(2-\alpha)(1-\alpha)\alpha}\right), \tag{A.5}$$

where Φ_2 is defined in (26).

Proof. It is straight forward to see that Φ_2 is concave in the radius ρ of the considered Wasserstein ball around $\bar{\mu}$. This is due to the fact that we defined the ground metric $c(\cdot,\cdot)$ of the transportation distance d_c by the L_1-metric, that is, $c(x,y) = ||x-y||_1$. Hence, to establish the lower bound (A.5), we only need to show that for $\rho^\star = \alpha(1-\alpha)(1-\alpha/2)$ it holds that $\Phi_2 \geq 1+\alpha$.

Therefore, we define the probability measure μ_α by the following bivariate copula

$$C_\alpha(u_1,u_2) = \begin{cases} u_1u_2 & \text{if } u \in [0,\alpha/2]^2 \cup [\alpha/2,\alpha]^2 \\ \frac{2-\alpha}{\alpha}u_1u_2 & \text{if } u \in ([0,\alpha/2]\times[\alpha/2,\alpha]) \cup ([\alpha/2,\alpha]\times[0,\alpha/2]) \\ \frac{1}{1-\alpha}u_1u_2 & \text{if } u \in [\alpha,1]^2 \\ \min(u_1,u_2) & \text{else} \end{cases}.$$

Tedious calculations show that $d_c(\bar{\mu},\mu_\alpha) \leq \alpha(1-\alpha)(1-\alpha/2) = \rho^\star$, where $\bar{\mu}$ is the bivariate probability measure with independent, standard uniformly distributed marginals defined in problem (26). Moreover, for $\binom{V}{U} \sim \mu_\alpha$ it holds that $\text{AVaR}_\alpha(U+V) = 1+\alpha$. □

A.5 | Correlation matrix

The purpose of this subsection is to give the correlation matrix Σ_0. Recall that Σ_0 defines the student-t copula C_0 with six degrees of freedom used as a reference dependence structure in the case study by Aas and Puccetti (2014), which we consider in Section 5. As this matrix is not given in the paper by Aas and Puccetti (2014), we simply choose the following arbitrary correlation matrix

$$\Sigma_0 = \begin{pmatrix} 1 & 0.36 & 0.35 & 0.44 & 0.45 & 0.30 \\ 0.36 & 1 & 0.37 & 0.36 & 0.41 & 0.43 \\ 0.35 & 0.37 & 1 & 0.44 & 0.32 & 0.42 \\ 0.44 & 0.36 & 0.44 & 1 & 0.41 & 0.29 \\ 0.45 & 0.41 & 0.32 & 0.41 & 1 & 0.28 \\ 0.30 & 0.43 & 0.42 & 0.29 & 0.28 & 1 \end{pmatrix}.$$