

Risk sharing with deep neural networks

M. Burzoni, A. Doldi & E. Monzio Compagnoni

To cite this article: M. Burzoni, A. Doldi & E. Monzio Compagnoni (06 Feb 2024): Risk sharing with deep neural networks, Quantitative Finance, DOI: [10.1080/14697688.2024.2307493](https://doi.org/10.1080/14697688.2024.2307493)

To link to this article: <https://doi.org/10.1080/14697688.2024.2307493>



Published online: 06 Feb 2024.



Submit your article to this journal [↗](#)






View related articles [↗](#)



View Crossmark data [↗](#)

Risk sharing with deep neural networks

M. BURZONI ^{*}†, A. DOLDI [‡]§ and E. MONZIO COMPAGNONI [§]¶

[†]Dipartimento di Matematica, Università degli Studi di Milano, Milano, Italy

[‡]Department of Economics and Management, Università degli Studi di Firenze, Firenze, Italy

[§]Department of Mathematics & Computer Science, University of Basel, Basel, Switzerland

(Received 22 December 2022; accepted 10 January 2024; published online 6 February 2024)

We consider the problem of optimally sharing a financial position among agents with potentially different reference risk measures. The problem is equivalent to computing the infimal convolution of the risk metrics and finding the so-called optimal allocations. We propose a neural network-based framework to solve the problem and we prove the convergence of the approximated inf-convolution, as well as the approximated optimal allocations, to the corresponding theoretical values. We support our findings with several numerical experiments.

Keywords: Risk sharing; Deep neural networks; Risk allocation; Inf-convolution; Universal approximation theorem

JEL Classification: C45, C63, G32

1. Introduction

We consider the problem faced by n economic agents, with reference risk measures ρ_1, \dots, ρ_n , who want to share the risk carried by a certain financial position, represented by a random variable X . The goal is to write X as the sum of n random variables X_1, \dots, X_n so that the sum of the risk of the single agents, $\rho_1(X_1) + \dots + \rho_n(X_n)$, is minimized. The problem is well known in the mathematical finance literature under the name of *risk sharing*, and it amounts to the calculation of the infimal convolution (*inf-convolution*) defined as follows:

$$\rho_1 \square \dots \square \rho_n(X) := \inf \left\{ \sum_{i=1}^n \rho_i(X_i) : \sum_{i=1}^n X_i = X \right\}. \quad (1)$$

The seminal paper Barriau and El Karoui (2005), which introduced inf-convolutions in the context of (convex) risk measures, originated a vast offspring of literature. Acciaio (2007) and Filipović and Svindland (2008) studied the case without monotonicity assumptions on the $(\rho_i)_{i=1, \dots, n}$, Mastrogioacomo and Rosazza Gianin (2015) considered the case of cash-subadditive and quasi-convex functionals, while multivariate risks are treated in Carlier and Dana (2013) and Carlier *et al.* (2012). We also mention Heath and Ku (2004), Tsanakas (2009), Dana and Le Van (2010), Weber (2018), Liebrich and Svindland (2019) and Embrechts *et al.* (2018, 2020) for fur-

ther extensions and we refer to Rüschendorf (2013) for a comprehensive overview of the topic.

The most relevant results for our analysis were established by Filipović and Svindland (2008) and Jouini *et al.* (2008) in the study of the so-called *optimal allocations* for $\rho_1 \square \dots \square \rho_n$, namely, the minimizers of the right-hand side of (1). For the case of law-invariant risk measures, it was demonstrated that comonotonicity plays a key role. In fact, optimal allocations can be found in the form $f_1(X), \dots, f_n(X)$ for some non-decreasing, real-valued, maps f_1, \dots, f_n , which sum up to the identity. This key aspect inspires the numerical framework that we propose in this paper. Indeed, it can be shown that the functions f_i , for $i = 1, \dots, n$, are Lipschitz continuous functions, thus, they can be very well approximated by neural networks. Despite the abundance of theoretical results on the risk-sharing topic, we are not aware of a general framework for the numerical computation of the solutions which works under very little assumptions, such as law-invariance and convexity. Indeed, the aforementioned literature usually focused on the explicit (or semi-explicit) computation of the optimal allocations in some special cases. This operation obviously requires an exact computation of the inf-convolution and their minimizers, which needs to be worked out case-by-case. Some risk measures of interest, for which explicit solutions can be provided, are the *entropic risk measure* and *expected shortfall* among the family of convex risk measures and, more recently, the (Range) Value at Risk among the non-convex ones, see Embrechts *et al.* (2018, 2020).

*Corresponding author. Email: alessandro.doldi@unifi.it

¶All authors contributed equally to this paper.

Using a suitable version of the Universal Approximation Theorem, we prove in section 2 that

$$\rho_1 \square \rho_2(X) = \inf \{ \rho_1(f(X)) + \rho_2(X - f(X)) : f \in \mathcal{NN} \},$$

where \mathcal{NN} is a suitable class of feed-forward neural networks. Deep neural networks (DNNs) have been proven to be very effective in solving a great variety of problems and in this paper, we show that this is the case also in a risk sharing context. The precise results are stated in theorems 2.8 and 2.11, which constitute the main results of the section. Note that the restriction to $n = 2$ agents is dictated by the convenience of exposition, but the case $n \geq 2$ can be covered similarly. Of course, we do not exclude that other methods could be successfully applied. In appendix 2, we discuss some possible alternative approaches.

In view of establishing a rigorous framework for our numerical experiments, we devote section 3 to the convergence analysis of the historical estimators of the inf-convolution, as well as of their corresponding optimal allocations. Such estimators are constructed simply by applying the risk measures of the agents to the empirical distribution of a large sample of X (see e.g. Cont *et al.* (2010) for an overview). The main convergence result of the section is theorem 3.3 which provides the theoretical justification of the experiments of section 4. We test our findings in a series of numerical experiments with different risk measures, different architectures, and different distributions for X (see section 4.1 for the details about the framework) obtaining consistent results. As for the risk measures, we use the following:

- (i) Entropic risk measure with parameter $\beta > 0$:

$$\text{Entr}_\beta(X) := \beta \log \mathbb{E} [e^{-X/\beta}]; \quad (2)$$

- (ii) Expected Shortfall (ES) at level $\alpha \in (0, 1)$:

$$\text{ES}_\alpha(X) = \frac{1}{\alpha} \int_0^\alpha V @ R_u(X) \, d\mu,$$

$$V @ R_u(X) := \inf \{ m \in \mathbb{R} : \mathbb{P}(X + m < 0) \leq u \};$$

- (iii) Distortion risk measure for $\mu \in \text{Prob}([0, 1])$, see (Föllmer and Schied 2016, section 4.6):

$$\rho_\mu(X) := \int_0^1 \text{ES}_\alpha(X) \mu(d\alpha). \quad (3)$$

The situation where all agents adopt an entropic risk measure, respectively ES, admits an explicit and simple solution for both the value of the inf-convolution and the optimal allocations. We test our numerical approximation in these cases only to confirm that the trained DNNs are converging to the known solutions. We then proceed in testing our algorithms in more complex situations. We cover the case of risk sharing between agents with distortion-type risk measures and between heterogeneous agents, that is, two agents using risk measures of different types—one has the entropic and the other adopts either the expected shortfall or a distortion risk measure. In all such cases, we confirm that the trained network is able to recover the expected form of the optimal allocations,

known from Embrechts *et al.* (2018) and Jouini *et al.* (2008). In the last experiment, where we consider the convolution of an entropic risk measure with a distortion risk measure, we do not have any information about the solution.

We also mention here that a neural-network based approach was successfully adopted in approximating optima for short-fall systemic risk measures with random allocations in Feng *et al.* (2022) and Doldi *et al.* (2023). Such a type of systemic risk measures was shown to be connected to sup-convolution problems in Doldi *et al.* (2024). We conclude this introduction with the frequently used notation. For a metric space \mathbb{X} , $\mathcal{B}(\mathbb{X})$ denotes the Borel σ -algebra and $m\mathcal{B}(\mathbb{X})$ denotes the class of real-valued, Borel-measurable functions on \mathbb{X} . We define the following sets:

$$\text{ca}(\mathbb{X}) := \{ \gamma : \mathcal{B}(\mathbb{X}) \rightarrow (-\infty, +\infty) \\ : \gamma \text{ is finite signed Borel measure on } \mathbb{X} \};$$

$$\text{Meas}(\mathbb{X}) := \{ \mu : \mathcal{B}(\mathbb{X}) \rightarrow [0, +\infty) : \mu \\ \text{is a non negative finite Borel measure on } \mathbb{X} \};$$

$$\text{Prob}(\mathbb{X}) := \{ Q : \mathcal{B}(\mathbb{X}) \rightarrow [0, 1] \\ : Q \text{ is a probability Borel measure on } \mathbb{X} \};$$

$$\mathcal{C}(\mathbb{X}) := \{ \varphi : \mathbb{X} \rightarrow \mathbb{R} : \varphi \text{ is continuous on } \mathbb{X} \};$$

$$\mathcal{C}_b(\mathbb{X}) := \{ \varphi : \mathbb{X} \rightarrow \mathbb{R} : \varphi \text{ is bounded and} \\ \text{continuous on } \mathbb{X} \};$$

$$\text{Prob}^p(\mathbb{R}) := \{ Q \in \text{Prob}(\mathbb{R}) :$$

$$\int_{\mathbb{R}} |x|^p \, dQ(x) < +\infty \}, \quad p \in [1, +\infty);$$

$$\text{Prob}_K^\infty(\mathbb{R}) := \{ Q \in \text{Prob}(\mathbb{R}) : Q([-K, K]) = 1 \}, \quad K > 0;$$

$$\text{Prob}^\infty(\mathbb{R}) := \bigcup_{K>0} \text{Prob}_K^\infty(\mathbb{R}).$$

2. The theoretical framework

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a standard non-atomic probability space (see e.g. Svidland (2010) for details about the possibility of dropping the standardness assumption). The Banach space $L^p(\Omega, \mathcal{F}, \mathbb{P})$ for $p \in [1, \infty)$ is the set of p -integrable random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, endowed with the norm $\| \cdot \|_p := (\mathbb{E}[|\cdot|^p])^{1/p}$. The Banach space L^∞ is the set of essentially bounded random variables, endowed with the supremum norm $\| \cdot \|_\infty$. The order relation \leq on such spaces is the one induced by the \mathbb{P} -a.s. ordering. We first recall the definition of monetary risk measures and some of their standard properties. We refer to the book Föllmer and Schied (2016) for a thorough presentation of the topic.

DEFINITION 1 Let $p \in [1, \infty]$ and $\rho : L^p(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (-\infty, \infty]$ a functional.

- ρ is normalized if $\rho(0) = 0$;
- ρ is finite if $\rho(X) < \infty$ for every $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$;
- ρ is monotone if $\rho(X) \leq \rho(Y)$, whenever $X \geq Y$ \mathbb{P} -a.s.;

- ρ is cash additive if $\rho(X + c) = \rho(X) - c$, for every $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ and $c \in \mathbb{R}$;
- ρ is convex if $\rho(\lambda X + (1 - \lambda)Y) \leq \lambda\rho(X) + (1 - \lambda)\rho(Y)$, for every $X, Y \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ and $\lambda \in [0, 1]$.

Any normalized, finite, monotone, and cash-additive ρ is called a monetary risk measure. If ρ is also convex, it is called a convex risk measure.

- ρ is law-invariant if $\rho(X) = \rho(Y)$ whenever $X \sim Y$;
- ρ satisfies the Lebesgue Property if $\rho(X) = \lim_{n \rightarrow +\infty} \rho(X_n)$ for any sequence $(X_n)_{n \in \mathbb{N}} \subseteq L^p(\Omega, \mathcal{F}, \mathbb{P})$ and $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ such that: there exists $Z \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ with $|X_n| \leq Z$ \mathbb{P} -a.s. for all $n \in \mathbb{N}$ and $\lim_{n \rightarrow +\infty} X_n = X$ \mathbb{P} -a.s. holds.

We next introduce the concept of infimal convolution (inf-convolution in short) of convex risk measures ρ_1, \dots, ρ_n . For ease of exposition, we restrict ourselves to the case of $n = 2$, however, all the results generalize to the case of an arbitrary $n \in \mathbb{N}$.

DEFINITION 2 Let $p \in [1, \infty]$. Given two functionals $\rho_1, \rho_2 : L^p(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (-\infty, \infty]$, their infimal convolution is defined as:

$$\rho_1 \square \rho_2(X) := \inf \{ \rho_1(X_1) + \rho_2(X_2) : X_1, X_2 \in L^p, X_1 + X_2 = X \}, \quad X \in L^p. \quad (4)$$

Every couple $(X_1, X_2) \subseteq L^p$ such that $X_1 + X_2 = X$ is called an allocation for X . Additionally, we say that an allocation is

- An optimal allocation if it is a minimizer of the right-hand side of (4);
- A comonotonic allocation if it is of the form $(f_1(X), f_2(X))$ for some increasing[†] functions $f_1, f_2 : \mathbb{R} \rightarrow \mathbb{R}$ such that $f_1 + f_2 = \text{Id}$, where $\text{Id} : \mathbb{R} \rightarrow \mathbb{R}$ denotes the identity function $\text{Id}(x) = x$.

The following well-known result, see Filipović and Svindland (2008, theorem 2.5), shows that for lower semi-continuous (l.s.c.) law-invariant convex risk measures, optimal allocations can be found among the class of comonotonic allocations.

THEOREM 2.1 Let $p \in [1, \infty]$ and $\rho_1, \rho_2 : L^p \rightarrow (-\infty, \infty]$ be l.s.c. law-invariant convex cash additive functions. Then $\rho_1 \square \rho_2 : L^p(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow [-\infty, \infty]$ is a l.s.c. law-invariant convex cash additive function. Moreover, there exist increasing functions $f_1, f_2 : \mathbb{R} \rightarrow \mathbb{R}$ such that $f_1 + f_2 = \text{Id}$ and

$$\rho_1 \square \rho_2(X) = \rho_1(f_1(X)) + \rho_2(f_2(X)).$$

It is not difficult to see that the functions f_1, f_2 are necessarily Lipschitz continuous. Indeed, for $x \geq y$, we can write $f_1(x) - f_1(y) + f_2(x) - f_2(y) = x - y$ and, using that both functions are increasing, we obtain the inequality $|f_i(x) - f_i(y)| \leq |x - y|$ for $i = 1, 2$. For the case $x \leq y$, the argument

is analogous. In particular, it holds $\|f_i\|_{\text{Lip}} \leq 1$ for $i = 1, 2$, where

$$\|f\|_{\text{Lip}} := \inf \{ L > 0 : |f(x) - f(y)| \leq L|x - y|, \forall x, y \in \mathbb{R} \}.$$

We also observe that, for monetary risk measures, if $(f_1(X), f_2(X))$ is an optimal allocation, the same is true for $(f_1(X) - c, f_2(X) + c)$ for an arbitrary $c \in \mathbb{R}$. This is also called *rebalancing of cash*. Without loss of generality, the function f_1 can be therefore chosen to satisfy $f_1(0) = 0$, while still preserving the Lipschitz property. Combining these two observations we obtain the following corollary to theorem 2.1.

COROLLARY 2.2 Under the assumptions of theorem 2.1, we have

$$\rho_1 \square \rho_2(X) = \min \left\{ \rho_1(f(X)) + \rho_2(X - f(X)) : f \in \mathcal{A}_{\text{Lip}}^0 \right\}, \quad (5)$$

where

$$\mathcal{A}_{\text{Lip}}^0 := \{ f : \mathbb{R} \rightarrow \mathbb{R} : f(0) = 0, \|f\|_{\text{Lip}} \leq 1, \|\text{Id} - f\|_{\text{Lip}} \leq 1 \} \quad (6)$$

is the set of normalized Lipschitz allocations.

Any function $f \in \mathcal{A}_{\text{Lip}}^0$ induces the allocation $(f(X), X - f(X))$. Indeed the sum equals X by construction and, using the Lipschitz property, it is clear that if $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ then $f(X) \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ as well. With a slight abuse of terminology, we call *allocation* also the pair of functions $(f, \text{Id} - f)$. By denoting \mathbb{P}_X the law of X under \mathbb{P} , this terminology becomes accurate when we work on the probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$, as it will often be the case below. The following is a sufficient criterion for guaranteeing the uniqueness of optimal allocations, see Filipović and Svindland (2008, theorem 2.5). For $X, Y \in L^p$ we use the notation $X - Y \notin \mathbb{R}$ for indicating that the difference $X - Y$ is not a constant random variable.

PROPOSITION 2.3 Under the assumptions of theorem 2.1, suppose additionally that ρ_1 is strictly convex in the sense that

$$\rho_1(\lambda Y + (1 - \lambda)Z) < \lambda\rho_1(Y) + (1 - \lambda)\rho_1(Z) \quad \forall \lambda \in (0, 1), Y, Z \in L^p \text{ s.t. } Y - Z \notin \mathbb{R}. \quad (7)$$

Then the optimal allocation for $\rho_1 \square \rho_2$ is unique up to rebalancing of cash, namely, for any pair of optima $(X_1, X_2), (\tilde{X}_1, \tilde{X}_2) \in L^p(\Omega, \mathcal{F}, \mathbb{P}) \times L^p(\Omega, \mathcal{F}, \mathbb{P})$ it holds $\tilde{X}_i = X_i + c_i$ for some $c_i \in \mathbb{R}$ with $c_1 = -c_2$ and for $i = 1, 2$.

This proposition generalizes to the case $n \geq 2$ when all but one of the initial risk measures are strictly convex (see e.g. the discussion after corollary 11.14 in Rüschendorf (2013)).

An example of strictly convex risk measure is the entropic risk measure of (2). We observe here that, for a given $X \in L^p$, uniqueness can be obtained via a small perturbation of ρ_1 by guaranteeing, at the same time, that the value of the infimal convolution is close.

[†] Increasing is understood in the non-strict sense.

LEMMA 2.4 *Let ρ_1, ρ_2 be law-invariant convex risk measures on $L^p(\Omega, \mathcal{F}, \mathbb{P})$, for $p \in [1, +\infty]$. Let $\tilde{\rho}$ be a strictly convex risk measure on $L^p(\Omega, \mathcal{F}, \mathbb{P})$. For every $\tilde{\varepsilon} > 0$, the risk measure $\rho_{1, \tilde{\varepsilon}} := (1 - \tilde{\varepsilon})\rho_1 + \tilde{\varepsilon}\tilde{\rho}$ is a strictly convex risk measure. Moreover, for every $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ and $\varepsilon > 0$, there exists $1 > \tilde{\varepsilon} = \tilde{\varepsilon}(X) > 0$ such that $|\rho_1 \square \rho_2(X) - \rho_{1, \tilde{\varepsilon}} \square \rho_2(X)| < \varepsilon$.*

Proof It is easy to see that all the properties of ρ_1 and $\tilde{\rho}$ are inherited by $\rho_{1, \tilde{\varepsilon}}$. As for the second statement, in view of corollary 2.2 it is enough to show that, defining $\Psi_{\tilde{\varepsilon}}(f) := \rho_{1, \tilde{\varepsilon}}(f(X)) + \rho_2(X - f(X))$, $f \in \mathcal{A}_{Lip}^0$, we have $\lim_{\tilde{\varepsilon} \downarrow 0} \sup_{f \in \mathcal{A}_{Lip}^0} |\Psi_{\tilde{\varepsilon}}(f) - \Psi_0(f)| = 0$. To see this, observe that

$$\begin{aligned} |\Psi_{\tilde{\varepsilon}}(f) - \Psi_0(f)| &= \tilde{\varepsilon} |\rho_1(f(X)) - \tilde{\rho}(f(X))| \\ &\leq \tilde{\varepsilon} (|\rho_1(f(X))| + |\tilde{\rho}(f(X))|) \end{aligned}$$

Since $f \in \mathcal{A}_{Lip}^0$, $|x| \geq |f(x)| \geq -|x|$ and by monotonicity and finiteness of ρ_1 we get $\rho_1(-|X|) \geq \rho_1(f(X)) \geq \rho_1(|X|)$, and $|\rho_1(f(X))| \leq |\rho_1(-|X|)| + |\rho_1(|X|)|$. The same argument applies to $\tilde{\rho}$, from which we deduce $|\Psi_{\tilde{\varepsilon}}(f) - \Psi_0(f)| \leq \tilde{\varepsilon}K$ for some constant $K > 0$ depending only on X . Since the right-hand side does not depend on f , the claim is proved. ■

Towards the aim of approximating the infimal convolutions using neural networks, we need some continuity of the risk functionals. For risk measures on L^∞ , the continuity is a consequence of the monotonicity and cash additivity properties. For the case $p \in [1, \infty)$, the Extended Namioka-Klee Theorem (see Biagini and Frittelli (2010)) guarantees that any proper convex and monotone functional on L^p is continuous with respect to the L^p -norm, on the interior of its domain. Thanks to the finiteness property, convex risk measures as in definition 2.1 are norm continuous for every $p \in [1, \infty]$ on the whole space. Throughout the paper, we will therefore make the following standing assumption.

ASSUMPTION 2.5 ρ_1 and ρ_2 are law-invariant convex risk measures.

2.1. Approximation of inf-convolutions via neural networks

In this section, we show that the inf-convolution of two risk measures in (5) can be approximated using neural networks in the construction of the allocations. This is achieved by means of appropriate versions of the universal approximation theorem (UAT). We first note that we can reduce our focus to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Consider indeed a functional $\rho : L^p(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (-\infty, +\infty]$ which is law-invariant. Since the underlying space is non-atomic, for every probability measure $\mathbb{Q} \in \text{Prob}^p(\mathbb{R})$ (or $\mathbb{Q} \in \text{Prob}^\infty(\mathbb{R})$ for $p = \infty$), there exists $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{Q} = \mathbb{P}_X$. Using the law invariance of ρ , the functional $\tilde{\rho}(\cdot | \mathbb{Q}) : L^p(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{Q}) \rightarrow (-\infty, +\infty]$ given by

$$\begin{aligned} \tilde{\rho}(\varphi | \mathbb{Q}) &:= \rho(\varphi \circ X) \text{ for any measurable} \\ X : \Omega &\rightarrow \mathbb{R} \text{ such that } \mathbb{P}_X = \mathbb{Q} \end{aligned} \quad (8)$$

is well defined and it inherits the properties listed in definition 2.1 from ρ . A similar procedure has been considered by Frittelli and Maggis (2018), although with a totally different aim. We stress some key consequences.

PROPOSITION 2.6 *Let $p \in [1, +\infty]$. Let $\rho_1, \rho_2 : L^p \rightarrow \mathbb{R}$ be law-invariant convex risk measures. Then,*

$$\begin{aligned} \rho_1 \square \rho_2(X) &= \inf \left\{ \rho_1(f(X)) + \rho_2(X - f(X)) : f \in \mathcal{A}_{Lip}^0 \right\} \\ &= \inf \left\{ \tilde{\rho}_1(f | \mathbb{P}_X) + \tilde{\rho}_2(\text{Id} - f | \mathbb{P}_X) \right. \\ &\quad \left. : f \in \mathcal{A}_{Lip}^0 \right\} = \tilde{\rho}_1(\cdot | \mathbb{P}_X) \square \tilde{\rho}_2(\cdot | \mathbb{P}_X)(\text{Id}) \end{aligned} \quad (9)$$

and $\hat{f} \in \mathcal{A}_{Lip}^0$ is a minimum in (9) if and only if it is a minimum in (10).

Proof The equality (9) is simply corollary 2.2. The first equality in (10) is given by definition of $\tilde{\rho}$ in (8) and the fact that any $f \in \mathcal{A}_{Lip}^0$ satisfies $f \in L^p(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$, thanks to the Lipschitz continuity. The last equality in (10) does not immediately follow from corollary 2.2, since we do not know if $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$ is non-atomic. The inequality \geq is clear. Using (8), we can rewrite

$$\begin{aligned} &\tilde{\rho}_1(\cdot | \mathbb{P}_X) \square \tilde{\rho}_2(\cdot | \mathbb{P}_X)(\text{Id}) \\ &= \inf \left\{ \tilde{\rho}_1(Y | \mathbb{P}_X) + \tilde{\rho}_2(\text{Id} - Y | \mathbb{P}_X) : Y \in L^p(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X) \right\} \\ &= \inf \left\{ \rho_1(Y \circ X) + \rho_2(X - Y \circ X) : Y \in L^p(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X) \right\} \\ &\geq \rho_1 \square \rho_2(X) \end{aligned}$$

which concludes the proof of the equality chain in (9), (10). The last statement follows from (8). ■

REMARK 1 The second equality in (10) holds, more generally, if we replace \mathbb{P}_X with an arbitrary $\mathbb{Q} \in \text{Prob}^p(\mathbb{R})$. Indeed, since the space is non atomic, \mathbb{Q} is the law of some $Y \in L^p(\Omega, \mathcal{F}, \mathbb{P})$. For such a Y it also holds that $\rho_1 \square \rho_2(Y) = \tilde{\rho}_1(\cdot | \mathbb{Q}) \square \tilde{\rho}_2(\cdot | \mathbb{Q})(\text{Id})$.

We next introduce the class of neural networks that we intend to use.

DEFINITION 3 *Let $L, N_0, \dots, N_L \in \mathbb{N}$ with $L \geq 2$, let $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ an activation function and for any $\ell = 1, \dots, L$, let $W_\ell : \mathbb{R}^{N_{\ell-1}} \rightarrow \mathbb{R}^{N_\ell}$ an affine function. A (feed-forward) neural network is a function $F : \mathbb{R}^{N_0} \rightarrow \mathbb{R}^{N_L}$ defined as*

$$F(x) := (W_L \circ \sigma \circ W_{L-1} \circ \dots \circ \sigma \circ W_1)(x),$$

where the activation function σ is applied componentwise.

We denote by \mathcal{NN} the vector space generated by the class of neural networks from $\mathbb{R}^K \rightarrow \mathbb{R}$ determined by a fixed activation function σ , continuous, nonconstant and bounded. Notice that \mathcal{NN} is a vector subspace of $\mathcal{C}_b(\mathbb{R}^K)$, and is, in particular, a convex cone. Moreover, imposing $L = 2$ in the above definition, \mathcal{NN} contains the functions generated by only one hidden layer and only one output unit as considered in Hornik (1991). A simple argument based on the classical

UAT of (Hornik 1991, theorem 1) yields the approximation result that we need, at least for the case $p < +\infty$. The case $p = \infty$ is not covered by this theorem and we need a slightly different approach.

THEOREM 2.7 *Let σ be continuous, bounded, and nonconstant. Then \mathcal{NN} is norm dense in $L^p(\mathbb{R}^K, \mathcal{B}(\mathbb{R}^K), \mu)$ for any finite measure $\mu \in \text{Meas}(\mathbb{R}^K)$ and $p \in [1, +\infty)$.*

The original theorem is stated for networks with two layers ($L = 2$). Since \mathcal{NN} contains this particular class, the density result also holds as stated in theorem 2.7. The following is our first approximation result.

THEOREM 2.8 *Let $p \in [1, \infty)$. Let $\rho_1, \rho_2 : L^p(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow \mathbb{R}$ be law-invariant convex risk measures. Then,*

$$\rho_1 \square \rho_2(X) = \inf \{ \rho_1(f(X)) + \rho_2(X - f(X)) : f \in \mathcal{NN} \}. \quad (11)$$

Proof Let $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$. Let \mathbb{P}_X be the law of X under \mathbb{P} . We claim that

$$\begin{aligned} \rho_1 \square \rho_2(X) &= \inf \{ \rho_1(f(X)) + \rho_2(X - f(X)) \\ &\quad : f : \mathbb{R} \rightarrow \mathbb{R} \text{ s.t. } f(X) \in L^p(\Omega, \mathcal{F}, \mathbb{P}) \} \\ &= \inf \{ \tilde{\rho}_1(f|\mathbb{P}_X) + \tilde{\rho}_2(\text{Id} - f|\mathbb{P}_X) \\ &\quad : f \in L^p(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X) \}. \end{aligned}$$

Indeed, from corollary 2.2, $\rho_1 \square \rho_2(X)$ attains the minimum over the set of allocations $(f(X), X - f(X))$ with $f \in \mathcal{A}_{\text{Lip}}^0$. Since $f(X) \in L^p(\Omega, \mathcal{F}, \mathbb{P})$, for every $f \in \mathcal{A}_{\text{Lip}}^0$, the first equality follows. The second equality is by (8). Notice now that $\tilde{\rho}_1(\cdot|\mathbb{P}_X)$ and $\tilde{\rho}_2(\cdot|\mathbb{P}_X)$ are law-invariant convex risk measures, which are $\|\cdot\|_p$ -continuous by the Extended Namioka-Klee Theorem in Biagini and Frittelli (2010). From theorem 2.7, applied with $\mu = \mathbb{P}_X$, we have

$$\begin{aligned} \rho_1 \square \rho_2(X) &= \inf \left\{ \tilde{\rho}_1(f|\mathbb{P}_X) + \tilde{\rho}_2(\text{Id} - f|\mathbb{P}_X) : f \in \overline{\mathcal{NN}}^{\|\cdot\|_p} \right\}, \end{aligned}$$

which in turns yields, by norm continuity, that (11) holds. \blacksquare

The UAT does not provide uniform approximations and, in particular, it does not cover the L^∞ case. We use here an approach based on weighted spaces in order to obtain theorem 2.10, which is inspired by the forthcoming paper Cuchiero *et al.* (2022). This theorem is only instrumental for our main results and it is certainly not the first time that the theory of weighted spaces has been exploited in Universal Approximation results (see for example Kratsios (2021), Cuchiero *et al.* (2022) and the references therein). We will also add a short proof for the sake of completeness. Let $r : \mathbb{X} \rightarrow [1, +\infty)$ be a continuous function with compact sublevels and define

$$C_r(\mathbb{X}) := \left\{ \phi \in \mathcal{C}(\mathbb{X}) : \|\phi\|_r := \sup_{x \in \mathbb{X}} \frac{|\phi(x)|}{r(x)} < +\infty \right\}.$$

The space $C_r(\mathbb{X})$ is a Banach lattice when endowed with the norm $\|\cdot\|_r$. This can be easily verified following verbatim the

argument for the standard case of bounded continuous functions with the supremum norm. We introduce the following sets:

$$\begin{aligned} \text{ca}_r(\mathbb{X}) &:= \{ \gamma : \mathcal{B}(\mathbb{X}) \rightarrow \mathbb{R} \\ &\quad : \mu \in \text{ca}(\mathbb{X}) \text{ and } \int_{\Omega} r(x) d|\gamma|(x) < +\infty \}; \end{aligned}$$

$$\begin{aligned} \text{Meas}_r(\mathbb{X}) &:= \{ \mu : \mathcal{B}(\mathbb{X}) \rightarrow [0, +\infty) : \\ &\quad : \mu \in \text{Meas}(\mathbb{X}) \text{ and } \int_{\Omega} r(x) d\mu(x) < +\infty \}; \end{aligned}$$

$$\begin{aligned} \text{Prob}_r(\mathbb{X}) &:= \{ \mathbb{Q} : \mathcal{B}(\mathbb{X}) \rightarrow [0, 1] \\ &\quad : \mathbb{Q} \in \text{Prob}(\mathbb{X}) \text{ and } \int_{\Omega} r(x) d\mathbb{P}(x) < +\infty \}; \end{aligned}$$

$$B_r(\mathbb{X}) := \overline{C_b(\mathbb{X})}^{\|\cdot\|_r}.$$

PROPOSITION 2.9 *$B_r(\mathbb{X})$ is a Banach space and for every continuous linear functional $\ell \in (B_r(\mathbb{X}))^*$ there exists a unique $\gamma \in \text{ca}_r(\mathbb{X})$ such that $\ell(\phi) = \int_{\mathbb{X}} \phi d\gamma$. Conversely, every $\gamma \in \text{ca}_r(\mathbb{X})$ defines a continuous linear functional in $(B_r(\mathbb{X}))^*$ in the same way.*

Proof This follows from Dörsek and Teichmann (2022), see theorems 2.4 and 2.7. \blacksquare

THEOREM 2.10 *Let $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ be continuous, bounded, and non-constant. Let $K \geq 1$ be a fixed integer and take $\mathbb{X} = \mathbb{R}^K$. Then the family \mathcal{NN} in definition 2.3 is $\|\cdot\|_r$ -dense in $B_r(\mathbb{R}^K)$.*

Proof As commented above, it is enough to prove the result for networks with two layers. By (Hornik 1991, theorem 5), σ is discriminatory, meaning that for any $\gamma \in \text{ca}_r(\mathbb{R}^K) = B_r^*$ we have

$$\begin{aligned} \int_{\mathbb{R}^K} \sigma \left(\sum_{j=1}^K a_j x_j + \theta \right) d\gamma = 0, \\ \forall a_1, \dots, a_K, \theta \in \mathbb{R} \Rightarrow \gamma \equiv 0. \end{aligned}$$

Let $\overline{\mathcal{NN}}$ be the weak closure of \mathcal{NN} in B_r with respect to the topology $\sigma(B_r, B_r^*) = \sigma(B_r, \text{ca}_r(\mathbb{R}^K))$, where the pairing is given by the integration and is well defined from proposition 2.9. Recall that, for a cone $C \subseteq B_r$, $C^\circ := \{ \gamma \in \text{ca}_r(\mathbb{R}^K) : \int_{\mathbb{R}^K} \phi d\gamma = 0 \forall \phi \in C \}$ is called the polar cone of C . Since σ is discriminatory, $\overline{\mathcal{NN}}^\circ \subseteq \mathcal{NN}^\circ = \{0\}$. By the bipolar theorem, we have $\overline{\mathcal{NN}} = \{0\}^\circ = B_r(\mathbb{R}^K)$. Since \mathcal{NN} in definition 2.3 is convex, we have $\overline{\mathcal{NN}} = \overline{\mathcal{NN}}^{\|\cdot\|_r}$, the latter being the $\|\cdot\|_r$ -closure of \mathcal{NN} . This proves that $\overline{\mathcal{NN}}^{\|\cdot\|_r} = B_r(\mathbb{R}^K)$, as desired. \blacksquare

The following is our second general approximation result.

THEOREM 2.11 *Let $\rho_1, \rho_2 : L^\infty(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow \mathbb{R}$ be law-invariant convex risk measures. Then,*

$$\begin{aligned} \rho_1 \square \rho_2(X) &= \inf \{ \rho_1(f(X)) + \rho_2(X - f(X)) \\ &\quad : f \in \mathcal{NN} \}, \quad X \in L^\infty(\Omega, \mathcal{F}, \mathbb{P}). \quad (12) \end{aligned}$$

Proof Let $X \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ and \mathbb{P}_X its law under \mathbb{P} . From proposition 2.6, we can rewrite

$$\rho_1 \square \rho_2(X) = \inf \left\{ \tilde{\rho}_1(f|\mathbb{P}_X) + \tilde{\rho}_2(\text{Id} - f|\mathbb{P}_X) : f \in \mathcal{A}_{\text{Lip}}^0 \right\}.$$

Note that $\mathcal{A}_{\text{Lip}}^0 \subseteq B^r(\mathbb{R})$ for $r(x) := 1 + |x|^{1+\varepsilon}$ with $\varepsilon > 0$. We thus deduce,

$$\begin{aligned} \rho_1 \square \rho_2(X) &= \inf \left\{ \tilde{\rho}_1(f|\mathbb{P}_X) + \tilde{\rho}_2(\text{Id} - f|\mathbb{P}_X) : f \in \mathcal{A}_{\text{Lip}}^0 \right\} \\ &\geq \inf \left\{ \tilde{\rho}_1(f|\mathbb{P}_X) + \tilde{\rho}_2(\text{Id} - f|\mathbb{P}_X) : f \in B^r(\mathbb{R}) \right\} \\ &\geq \inf \left\{ \tilde{\rho}_1(f|\mathbb{P}_X) + \tilde{\rho}_2(\text{Id} - f|\mathbb{P}_X) \right. \\ &\quad \left. : f \in L^\infty(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X) \right\} \\ &= \rho_1 \square \rho_2(X), \end{aligned}$$

where in the second inequality we have used that \mathbb{P}_X has compact support and r is continuous, whereas, in the third inequality we used again the law invariance. Finally, the last equality is a consequence of proposition 2.6. This shows that all the above inequalities are actually equalities. From theorem 2.10, $B^r(\mathbb{R}) = \overline{\mathcal{NN}^{\|\cdot\|_r}}$, so that

$$\rho_1 \square \rho_2(X) = \inf \left\{ \tilde{\rho}_1(f|\mathbb{P}_X) + \tilde{\rho}_2(\text{Id} - f|\mathbb{P}_X) : f \in \overline{\mathcal{NN}^{\|\cdot\|_r}} \right\}.$$

We conclude using the continuity of ρ_1, ρ_2 with respect to the uniform convergence. \blacksquare

REMARK 2 Note that if we replace $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ with $L^p(\Omega, \mathcal{F}, \mathbb{P})$ and $p \in [1, +\infty)$, the same arguments with the choice of $r(x) := 1 + |x|^{p+\varepsilon}$ yields that (12) holds for every $X \in L^{p+\varepsilon}(\Omega, \mathcal{F}, \mathbb{P})$. The proof presented for theorem 2.8 is however more direct and it does not require extra integrability on X . The advantage of the approach of theorem 2.11 is that the extra equality

$$\rho_1 \square \rho_2(X) = \inf \left\{ \tilde{\rho}_1(f|\mathbb{P}_X) + \tilde{\rho}_2(\text{Id} - f|\mathbb{P}_X) : f \in \overline{\mathcal{NN}^{\|\cdot\|_r}} \right\}$$

shows that the elements in the closure are approximated by the neural networks uniformly also for the case of $p < \infty$.

3. Convergence results

Let $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$, for some $p \in [1, +\infty]$. Consider an i.i.d. sequence $(X_n)_n \subseteq L^p(\Omega, \mathcal{F}, \mathbb{P})$ with common distribution \mathbb{P}_X and let $(\widehat{F}_N)_N$ denote the corresponding sequence of empirical cumulative distribution functions: for $x \in \mathbb{R}$, $\widehat{F}_N(x) : \Omega \rightarrow \mathbb{R}$ is defined as

$$\widehat{F}_N(x) := \frac{1}{N} \sum_{n=1}^N \mathbf{1}_{(-\infty, x]}(X_n), \quad x \in \mathbb{R}.$$

Denote by $\widehat{\mathbb{P}}_N$ the random measure associated to the empirical c.d.f. \widehat{F}_N , namely

$$\omega \mapsto \widehat{\mathbb{P}}_N(\omega)(\cdot) := \frac{1}{N} \sum_{n=1}^N \delta_{X_n(\omega)}(\cdot). \quad (13)$$

Finally, for $p \in [1, +\infty)$, let \mathcal{W}_p be the p -Wasserstein distance on $\text{Prob}^p(\mathbb{R})$ induced by the Euclidean norm, namely,

$$(\inf \{\mathbb{E}|X - Y|^p : X \sim \mu, Y \sim \nu\})^{\frac{1}{p}}$$

We refer to the book Villani (2009) for a thorough presentation of the topic.

LEMMA 3.1 *Let $p \in [1, +\infty]$ and $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$. Then $\widehat{\mathbb{P}}_N \Rightarrow_N \mathbb{P}_X$, \mathbb{P} -a.s., where \Rightarrow denotes the weak convergence of probability measures. If $p \in [1, +\infty)$, it holds additionally*

$$\lim_{N \rightarrow +\infty} W_p(\mathbb{P}, \widehat{\mathbb{P}}_N) = 0 \quad \mathbb{P}\text{-a.s.}$$

Proof The first statement follows from the Glivenko-Cantelli theorem. As for the second statement, theorem 6.9 in Villani (2009) shows that convergence in \mathcal{W}_p is equivalent to $\widehat{\mathbb{P}}_N \Rightarrow_N \mathbb{P}_X$ together with the convergence of the p th moments. The latter follows from the law of large numbers and the integrability of \mathbb{P}_X , so that, $\lim_{N \rightarrow +\infty} \int_{\mathbb{R}} |x|^p d\widehat{\mathbb{P}}_N(x) = \int_{\mathbb{R}} |x|^p d\mathbb{P}_X(x)$. \blacksquare

We aim at proving the convergence of the optimal values, namely,

$$\begin{aligned} &\rho_1 \square \rho_2(X) \\ &= \lim_{N \rightarrow +\infty} \tilde{\rho}_1(\cdot|\widehat{\mathbb{P}}_N(\omega)) \square \tilde{\rho}_2(\cdot|\widehat{\mathbb{P}}_N(\omega))(\text{Id}), \quad \mathbb{P}\text{-a.e. } \omega, \end{aligned} \quad (14)$$

and the convergence of the corresponding minimizers. We will need to establish some joint continuity results for the map $(f, \mathbb{Q}) \mapsto \tilde{\rho}(f|\mathbb{Q})$, defined in (8), on the spaces $\mathcal{A}_{\text{Lip}}^0 \times \text{Prob}^p(\mathbb{R})$ and $\mathcal{A}_{\text{Lip}}^0 \times \text{Prob}_K^\infty(\mathbb{R})$. Some preliminary topological considerations are useful.

REMARK 3 The space $\text{Prob}_K^\infty(\mathbb{R})$ is used for the $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ case. Indeed, for $X \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ and $(X_n)_n$ an i.i.d. sequence with common law \mathbb{P}_X , we obviously have $|X_n| \leq \|X\|_\infty$ \mathbb{P} -a.s.. Thus, the measure $\widehat{\mathbb{P}}_N := \frac{1}{N} \sum_{n=1}^N \delta_{X_n}$ satisfies $\widehat{\mathbb{P}}_N(\omega) \in \text{Prob}_K^\infty(\mathbb{R})$ for \mathbb{P} -a.e. ω and for $K \geq \|X\|_\infty$.

We endow $\text{Prob}^p(\mathbb{R})$ with the \mathcal{W}_p -topology, for $p \in [1, \infty)$ and Prob_K^∞ with the weak topology. We endow $\mathcal{A}_{\text{Lip}}^0$ with the topology induced by the metric:

$$\begin{aligned} d(\varphi, \psi) &:= \sum_{h=1}^{\infty} \frac{1}{2^h} \\ &\min \left(1, \sup_{x \in [-h, h]} |\varphi(x) - \psi(x)| \right), \quad \varphi, \psi \in \mathcal{A}_{\text{Lip}}^0. \end{aligned} \quad (15)$$

One can verify that $(\mathcal{C}(\mathbb{R}), d)$ is a complete metric space and d metrizes the uniform convergence on compact sets. As a

consequence of the Ascoli-Arzelà theorem, we have that $\mathcal{A}_{\text{Lip}}^0$ is compact with respect to the topology induced by d , as we prove next.

LEMMA 3.2 *Let $(\varphi_n)_n$ be a sequence in $\mathcal{A}_{\text{Lip}}^0$. There exists $\varphi \in \mathcal{A}_{\text{Lip}}^0$ and a subsequence $(\varphi_{n_k})_k$ such that $\lim_{n \rightarrow +\infty} d(\varphi_{n_k}, \varphi) = 0$.*

Proof Recall that for every $f \in \mathcal{A}_{\text{Lip}}^0$, $\|f\|_{\text{Lip}} \leq 1$. In particular, for every $h \in \mathbb{N}$, any family of functions in $\mathcal{A}_{\text{Lip}}^0$, restricted to $[-h, h]$, is equicontinuous and equibounded. We first construct a sequence of functions $(\varphi_h)_{h \in \mathbb{N}}$ in the following iterative way. For $h = 1$ we apply the Ascoli-Arzelà theorem to $(\varphi_n)_n$ restricted to $[-1, 1]$. This yields a subsequence, that we relabel again as $(\varphi_n)_n$, and a continuous function φ_1 on $[-1, 1]$ such that $(\varphi_n)_n$ converges uniformly to φ_1 on $[-1, 1]$. Note that φ_1 is continuous on $[-1, 1]$, being uniform limit of continuous functions, and satisfies $\varphi_1(0) = 0$, since $\|\varphi_n\|_{\text{Lip}} \leq 1$ and $\varphi_n(0) = 0$ for any $n \in \mathbb{N}$. At the step $h + 1$ we repeat the same argument to the sequence $(\varphi_n)_n$ obtained at step h . Note that the limiting function φ_{h+1} satisfies $\varphi_{h+1} = \varphi_h$ on $[-h, h]$, since $(\varphi_n)_n$ converges uniformly to φ_h on $[-h, h]$. Similarly as above, φ_{h+1} is continuous on $[-(h+1), h+1]$ and satisfies $\varphi_{h+1}(0) = 0$.

We are now able to construct the limiting $\varphi \in \mathcal{A}_{\text{Lip}}^0$. For every $h \in \mathbb{N}$, we extend φ_h to \mathbb{R} in an arbitrary way outside $[-h, h]$. We set $\varphi(x) := \lim_{h \rightarrow \infty} \varphi_h(x)$, for every $x \in \mathbb{R}$ and note that φ coincides with φ_h on every $[-h, h]$. In particular, we deduce that $\varphi \in \mathcal{A}_{\text{Lip}}^0$.

Finally, we construct the convergent subsequence $(\varphi_{n_k})_k$ of the original sequence. From the Ascoli-Arzelà argument above, for every, $k \in \mathbb{N}$, there exists $n_k \in \mathbb{N}$ such that

$$\sup_{x \in [-k, k]} |\varphi_{n_k}(x) - \varphi_k(x)| < \frac{1}{k}.$$

For every $h \in \mathbb{N}$ and $k \geq h$, using that $\varphi = \varphi_k$ on $[-k, k]$, we obtain

$$\begin{aligned} \sup_{x \in [-h, h]} |\varphi_{n_k}(x) - \varphi(x)| &\leq \sup_{x \in [-k, k]} |\varphi_{n_k}(x) - \varphi(x)| \\ &= \sup_{x \in [-k, k]} |\varphi_{n_k}(x) - \varphi_k(x)| < \frac{1}{k}, \end{aligned}$$

which implies the uniform convergence of the subsequence $(\varphi_{n_k})_k$ to φ on every interval $[-h, h]$. An application of Dominated Convergence Theorem then yields

$$\begin{aligned} \lim_{k \rightarrow +\infty} d(\varphi_{n_k}, \varphi) &= \lim_{k \rightarrow +\infty} \sum_{h=1}^{\infty} \frac{1}{2^h} \\ &\min \left(1, \sup_{x \in [-h, h]} |\varphi_{n_k}(x) - \varphi(x)| \right) = 0. \end{aligned}$$

■

We state now our main convergence result. We need the following functional, which is an almost sure version of the

distance d . For μ a measure on $\mathcal{B}(\mathbb{R})$ and $\varphi, \psi : \mathbb{R} \rightarrow \mathbb{R}$, $\mathcal{B}(\mathbb{R})$ -measurable functions,

$$d_\mu(\varphi, \psi) := \sum_{h=1}^{\infty} \frac{1}{2^h} \left\| \min(1, |\varphi - \psi| 1_{[-h, h]}) \right\|_{\infty, \mu}, \quad (16)$$

where in the above $L^\infty(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)$ -norm we made explicit the dependence on the reference measure μ in order to avoid ambiguity in what follows.

THEOREM 3.3 *Let $p \in [1, \infty]$. Let $\rho_1, \rho_2 : L^p(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow \mathbb{R}$ be law-invariant convex risk measures. Suppose that ρ_1 is strictly convex in the sense of (7). Only in the case of $p = \infty$ suppose that ρ_1 and ρ_2 satisfy the Lebesgue property. Then,*

$$\begin{aligned} \tilde{\rho}_1(\cdot | \mathbb{P}_X) \square \tilde{\rho}_2(\cdot | \mathbb{P}_X) (\text{Id}) \\ = \lim_{N \rightarrow +\infty} \tilde{\rho}_1(\cdot | \widehat{\mathbb{P}}_N(\omega)) \square \tilde{\rho}_2(\cdot | \widehat{\mathbb{P}}_N(\omega)) (\text{Id}) \quad \mathbb{P}\text{-a.e. } \omega. \end{aligned}$$

Furthermore, let $(\widehat{\varphi}, \text{Id} - \widehat{\varphi})$ and $(\widehat{\varphi}_N(\omega), \text{Id} - \widehat{\varphi}_N(\omega))$ be optimal allocations in $\mathcal{A}_{\text{Lip}}^0$, corresponding to \mathbb{P}_X and $\widehat{\mathbb{P}}_N(\omega)$ respectively. Then,

$$\lim_{N \rightarrow +\infty} d_{\mathbb{P}_X}(\widehat{\varphi}_N(\omega), \widehat{\varphi}) = 0 \quad \mathbb{P}\text{-a.e. } \omega.$$

The rest of the section is devoted to the proof of this theorem. We will need a number of auxiliary results, which are of independent interest. The first result is essentially (Delbaen 2021, proposition 1) or (Shapiro 2013, theorem 2.1), adapted to our context.

LEMMA 3.4 *Consider again a generic atomless probability space $(\Omega, \mathcal{F}, \mathbb{P})$.*

- (i) *Let $p \in [1, +\infty)$. Suppose $(\mathbb{Q}_n)_n, \mathbb{Q} \subseteq \text{Prob}^p(\mathbb{R})$ and $\lim_{n \rightarrow +\infty} W_p(\mathbb{Q}_n, \mathbb{Q}) = 0$. Then, there exists a sequence $(Y_n)_n$ in $L^p(\Omega, \mathcal{F}, \mathbb{P})$ and $Y \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{P}_{Y_n} = \mathbb{Q}_n$ for every $n \in \mathbb{N}$, $\mathbb{P}_Y = \mathbb{Q}$ and $\lim_{n \rightarrow +\infty} \|Y_n - Y\|_p = 0$.*
- (ii) *Let $p = \infty$. Suppose $(\mathbb{Q}_n)_n, \mathbb{Q} \subseteq \text{Prob}_K^\infty(\mathbb{R})$ and $\mathbb{Q}_n \Rightarrow_n \mathbb{Q}$. Then there exists a sequence $(Y_n)_n$ in $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ and $Y \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{P}_{Y_n} = \mathbb{Q}_n$ for every $n \in \mathbb{N}$, $\mathbb{P}_Y = \mathbb{Q}$, $\lim_{n \rightarrow +\infty} Y_n = Y$ \mathbb{P} -a.s. and $\sup_n \|Y_n\|_\infty < +\infty$.*

Proof For every $\mathbb{Q}_n \Rightarrow_n \mathbb{Q}$, by the Skorokhod Theorem (as in (Billingsley 1999, theorem 25.6)), there exist random variables $Y, (Y_n)_n$ such that $\mathbb{P}_Y = \mathbb{Q}$, $\mathbb{P}_{Y_n} = \mathbb{Q}_n$ for every $n \in \mathbb{N}$ and $\lim_{n \rightarrow +\infty} Y_n = Y$ \mathbb{P} -a.s. For item (ii), it is enough to additionally note that $\mathbb{P}(Y_n \in [-K, K]) = \mathbb{Q}_n([-K, K]) = 1$.

As for item (i), by the characterization of W_p -convergence in (Villani 2009, theorem 6.9), we have

$$\begin{aligned} \lim_{n \rightarrow +\infty} \mathbb{E}_{\mathbb{P}} [|Y_n|^p] &= \lim_{n \rightarrow +\infty} \int_{\mathbb{R}} |x|^p d\mathbb{Q}_n(x) \\ &= \int_{\mathbb{R}} |x|^p d\mathbb{Q}(x) = \mathbb{E}_{\mathbb{P}} [|Y|^p]. \end{aligned}$$

Now we proceed as in Delbaen (2021, proposition 1): by Scheffé's lemma we conclude that $(|Y_n|^p)_n$ converges in

$L^1(\Omega, \mathcal{F}, \mathbb{P})$ to $|Y|^p$. Hence $|Y_n - Y|^p \leq \frac{1}{2}|2Y_n|^p + \frac{1}{2}|2Y|^p = 2^{p-1}(|Y_n|^p + |Y|^p)$ is a uniformly integrable sequence. Indeed,

$$\begin{aligned} \mathbb{E}_{\mathbb{P}} \left[\mathbf{1}_{\{|Y_n - Y|^p \geq K\}} |Y_n - Y|^p \right] &\leq \mathbb{E}_{\mathbb{P}} \left[\mathbf{1}_{\{|Y_n|^p + |Y|^p \geq 2^{1-p}K\}} |q_n - q|^p \right] \\ &\leq \mathbb{E}_{\mathbb{P}} \left[\mathbf{1}_{\{|Y_n|^p + |Y|^p \geq 2^{1-p}K\}} (|Y_n|^p + |Y|^p) \right]. \end{aligned}$$

Since $|Y_n - Y|^p$ converges to zero \mathbb{P} -a.s., the proof is complete. \blacksquare

PROPOSITION 3.5 *Let $p \in [1, \infty]$ and $\rho : L^p(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow \mathbb{R}$ be a law-invariant convex risk measure. For $p \in [1, +\infty)$ the map $(f, \mathbb{Q}) \mapsto \tilde{\rho}(f|\mathbb{Q})$ of (8) is continuous on $\mathcal{A}_{\text{Lip}}^0 \times \text{Prob}^p(\mathbb{R})$. If, on the other hand $p = +\infty$ and ρ additionally satisfies the Lebesgue property on $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$, then $(f, \mathbb{Q}) \mapsto \tilde{\rho}(f|\mathbb{Q})$ is continuous on $\mathcal{A}_{\text{Lip}}^0 \times \text{Prob}_K^\infty(\mathbb{R})$, for every $K > 0$.*

Proof We start covering the case $p \in [1, +\infty)$. Since $\mathcal{A}_{\text{Lip}}^0 \times \text{Prob}^p(\mathbb{R})$ is a metric space, we check continuity along sequences. Take a convergent sequence $(f_n, \mathbb{Q}_n) \rightarrow_n (f, \mathbb{Q})$ in $\mathcal{A}_{\text{Lip}}^0 \times \text{Prob}^p(\mathbb{R})$. Take any subsequence. We prove that it admits a further subsequence for which $\lim_{k \rightarrow +\infty} \tilde{\rho}(f_{n_k}|\mathbb{Q}_{n_k}) = \tilde{\rho}(f|\mathbb{Q})$, which yields the convergence of the original sequence. The first extracted subsequence will be relabelled with the index $n \in \mathbb{N}$. Since $\lim_{n \rightarrow +\infty} W_p(\mathbb{Q}_n, \mathbb{Q}) = 0$, we can apply lemma 3.4. Since

$$\lim_{N \rightarrow +\infty} \|Y_n - Y\|_p = 0$$

up to taking a further subsequence (and relabeling again with n) we might suppose that there exists a $0 \leq Z \in L^p$ with $|Y_n| \leq Z \forall n \in \mathbb{N}$, \mathbb{P} -a.s. By Dominated Convergence Theorem, since $\lim_{n \rightarrow +\infty} f_n(Y_n) = f(Y)$ \mathbb{P} -a.s., we get

$$\lim_{n \rightarrow +\infty} \|f_n(Y_n) - f(Y)\|_p = 0.$$

Now, since ρ is real-valued, hence norm continuous by Extended Namioka-Klee Theorem, we deduce

$$\tilde{\rho}(f|\mathbb{Q}) := \rho(f(Y)) = \lim_{n \rightarrow +\infty} \rho(f_n(Y_n)) =: \lim_{n \rightarrow +\infty} \tilde{\rho}(f_n|\mathbb{Q}_n) \quad (17)$$

and the desired continuity follows.

For the second statement, the argument is very similar. Take a convergent sequence $(f_n, \mathbb{Q}_n) \rightarrow_n (f, \mathbb{Q})$ in $\mathcal{A}_{\text{Lip}}^0 \times \text{Prob}_K^\infty(\mathbb{R})$. Take any subsequence. We prove that it admits a further subsequence for which $\lim_{k \rightarrow +\infty} \tilde{\rho}(f_{n_k}|\mathbb{Q}_{n_k}) = \tilde{\rho}(f|\mathbb{Q})$. Use lemma 3.4 item (ii) to obtain the sequence $(Y_n)_n$. Since $f_n \in \mathcal{A}_{\text{Lip}}^0 \forall n \in \mathbb{N}$, we have $\sup_n \|f_n(Y_n)\|_\infty \leq \sup_n \|Y_n\|_\infty$ which is finite by lemma 3.4 item (ii). Since f_n converges to f uniformly on compact intervals by definition, we deduce $\lim_{n \rightarrow +\infty} f_n(Y_n) = f(Y)$ \mathbb{P} -a.s. Using the Lebesgue property we conclude that (17) holds true, providing continuity. \blacksquare

Proof of theorem 3.3 Consider first the case $p \in [1, \infty)$. We prove something stronger, namely, that the thesis holds for every $(\mathbb{Q}_n)_n, \mathbb{Q} \subseteq \text{Prob}^p(\mathbb{R})$ such that $\lim_{n \rightarrow +\infty} W_p(\mathbb{Q}_n, \mathbb{Q}) =$

0 instead of only for $(\widehat{\mathbb{P}}_N)_N$ and \mathbb{P}_X . First, observe that by proposition 3.5 the function

$$(\varphi, \mathbb{Q}) \mapsto \tilde{\rho}_1(\varphi|\mathbb{Q}) + \tilde{\rho}_2(\text{Id} - \varphi|\mathbb{Q})$$

is continuous on $\mathcal{A}_{\text{Lip}}^0 \times \text{Prob}^p(\mathbb{R})$, and $\mathcal{A}_{\text{Lip}}^0$ is compact by lemma 3.2. Berge's Theorem (Aliprantis and Border 2006, theorem 17.31) guarantees that

$$\tilde{\rho}_1(\cdot|\mathbb{Q}) \square \tilde{\rho}_2(\cdot|\mathbb{Q})(\text{Id}) = \lim_{N \rightarrow +\infty} \tilde{\rho}_1(\cdot|\mathbb{Q}_N) \square \tilde{\rho}_2(\cdot|\mathbb{Q}_N)(\text{Id})$$

and that the correspondence $\Gamma : \text{Prob}^p(\mathbb{R}) \rightrightarrows \mathcal{A}_{\text{Lip}}^0$ defined by

$$\Gamma(\mathbb{Q}) := \text{argmin} \left\{ \tilde{\rho}_1(\varphi|\mathbb{Q}) + \tilde{\rho}_2(\text{Id} - \varphi|\mathbb{Q}) : \varphi \in \mathcal{A}_{\text{Lip}}^0 \right\}$$

is upper hemicontinuous. Consider now the numerical sequence $(d_{\mathbb{Q}}(\widehat{\varphi}_n, \widehat{\varphi}))_n$. Take an arbitrary subsequence and relabeled it again by n . Using the upper hemicontinuity of Γ (Aliprantis and Border 2006, theorem 17.20) and the convergence of $(\mathbb{Q}_n)_n$ to \mathbb{Q} , the sequence $(\widehat{\varphi}_n \in \Gamma(\mathbb{Q}_n))_n \subseteq \mathcal{A}_{\text{Lip}}^0$ has a limit point in $\Gamma(\mathbb{Q})$, that we call $\widehat{\varphi}_\infty$. Up to passing to a further subsequence and relabeling, we may assume that $(\widehat{\varphi}_n)_n$ converges to $\widehat{\varphi}_\infty$ with respect to the distance d . By definition of Γ , $\widehat{\varphi}_\infty$ induces an optimal allocation under \mathbb{Q} and for $Y \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{P}_Y = \mathbb{Q}$ we get, and using proposition 2.6,

$$\rho_1 \square \rho_2(Y) = \rho_1(\widehat{\varphi}_\infty(Y)) + \rho_2(Y - \widehat{\varphi}_\infty(Y)).$$

Since ρ_1 is strictly convex, the minimizer is unique by proposition 2.3 (recall that we fixed $\widehat{\varphi}_\infty(0) = 0^\dagger$). Thus, $\mathbb{P}(\widehat{\varphi}(Y) \neq \widehat{\varphi}_\infty(Y)) = 0$, or equivalently $0 = \mathbb{P}_Y(\widehat{\varphi} \neq \widehat{\varphi}_\infty) = \mathbb{Q}(\widehat{\varphi} \neq \widehat{\varphi}_\infty)$. The latter \mathbb{Q} -a.s. equality property yields $d_{\mathbb{Q}}(\widehat{\varphi}_n, \widehat{\varphi}_\infty) = d_{\mathbb{Q}}(\widehat{\varphi}_n, \widehat{\varphi})$. Note now that, by construction, $d_{\mathbb{Q}} \leq d$. We conclude that

$$\begin{aligned} \limsup_{n \rightarrow +\infty} d_{\mathbb{Q}}(\widehat{\varphi}_n, \widehat{\varphi}) &= \limsup_{n \rightarrow +\infty} d_{\mathbb{Q}}(\widehat{\varphi}_n, \widehat{\varphi}_\infty) \\ &\leq \lim_{n \rightarrow +\infty} d(\widehat{\varphi}_n, \widehat{\varphi}_\infty) = 0. \end{aligned}$$

We have shown that starting from an arbitrary subsequence of $(d_{\mathbb{Q}}(\widehat{\varphi}_n, \widehat{\varphi}_\infty))_n$ there exists a further subsequence converging to 0. This shows the desired property. For the case $p = +\infty$, the argument is exactly the same. Note only that for applying proposition 3.5 we need to require the Lebesgue continuity.

The claims in the statement follow now from lemma 3.1. \blacksquare

The case of spectral risk measures: The convergence (14) can be established in the context of spectral risk measures by proving a stronger result.

\dagger As a consequence of translation invariance, we can assume, without loss of generality, that \mathbb{Q} gives positive mass to every neighborhood of 0. Then, given two optimal allocations $\varphi_1, \varphi_2 \in \mathcal{A}_{\text{Lip}}^0$, proposition 2.3 implies that $\varphi_1 - \varphi_2 = c$ \mathbb{Q} -a.s., for some $c \in \mathbb{R}$. However, the definition of $\mathcal{A}_{\text{Lip}}^0$ necessarily implies $c = 0$.

DEFINITION 4 Let $p \in [1, \infty]$. A functional $\rho : L^p \rightarrow (-\infty, \infty]$ is called a spectral risk measure if

$$\rho(X) = \int_0^1 V @ R_\alpha(X) h(\alpha) d\alpha, \quad (18)$$

for some non-increasing function $h : [0, 1] \rightarrow [0, \infty)$, called spectral density, satisfying $\int_0^1 h(p) dp = 1$.

We refer to Pichler (2013a, 2013b) for a thorough analysis of the topic. In particular, the properties of h ensure that ρ is convex, monotone, and cash additive. Moreover, due to the properties of $V @ R$, ρ is also law invariant and positive homogeneous. Whenever ρ_1 and ρ_2 are both finite spectral risk measures on $L^p(\Omega, \mathcal{F}, \mathbb{P})$, $p \in [1, +\infty]$ (which is the case for suitably integrable spectral densities as shown in (Pichler 2013b, proposition 5 and theorem 11), assumption 2.5 is thus satisfied. For $p = +\infty$, ρ_1 and ρ_2 also satisfy the Lebesgue property if the spectral densities h_1, h_2 are such that ρ_1 and ρ_2 are well defined and finite on $L^r(\Omega, \mathcal{F}, \mathbb{P})$ for some r big enough, by the Extended Namioka-Klee Theorem. This translates into an integrability requirement on h_1, h_2 , see (Pichler 2013b, proposition 5 and theorem 11), and holds true for example if h_1, h_2 are bounded themselves. In both cases, we are in the exact setup of theorem 3.3. However, we can provide an explicit estimate for the convergence (14), as detailed below.

PROPOSITION 3.6 Let $p \in (1, \infty)$ and $\rho_1, \rho_2 : L^p(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow \mathbb{R}$ be law-invariant spectral convex risk measures, with possibly different spectral densities $h_1, h_2 \in L^{\frac{p}{p-1}}([0, 1], \mathcal{B}([0, 1]), \text{Leb})$. Then,

$$\begin{aligned} & |\rho_1 \square \rho_2(X) - \tilde{\rho}_1(\cdot | \widehat{\mathbb{P}}_N(\omega)) \square \tilde{\rho}_2(\cdot | \widehat{\mathbb{P}}_N(\omega)) (\text{Id})| \\ & \leq \left[\|h_1\|_{\frac{p}{p-1}} + \|h_2\|_{\frac{p}{p-1}} \right] W_p(\widehat{\mathbb{P}}_N(\omega), \mathbb{P}_X) \quad \forall \omega \in \Omega. \end{aligned} \quad (19)$$

In particular, (14) holds.

Proof We see that, fixing $\omega \in \Omega$ and taking $\widehat{\mathbb{P}}_N(\omega)$ as a (deterministic) measure in $\text{Prob}^p(\mathbb{R})$, we also have $\mathbb{P}_X \in \text{Prob}^p(\mathbb{R})$ since $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$, and

$$\begin{aligned} \rho_1 \square \rho_2(X) & \stackrel{\text{Prop. 2.6}}{=} \inf \left\{ \tilde{\rho}_1(f | \mathbb{P}_X) \square \tilde{\rho}_2(\text{Id} - f | \mathbb{P}_X) : f \in \mathcal{A}_{\text{Lip}}^0 \right\} \\ & \leq \inf \left\{ \tilde{\rho}_1(f | \widehat{\mathbb{P}}_N(\omega)) + \tilde{\rho}_2(\text{Id} - f | \widehat{\mathbb{P}}_N(\omega)) : f \in \mathcal{A}_{\text{Lip}}^0 \right\} \\ & \quad + W_p(\mathbb{P}_X, \widehat{\mathbb{P}}_N) \|h_1\|_{\frac{p}{p-1}} + W_p(\mathbb{P}_X, \widehat{\mathbb{P}}_N) \|h_2\|_{\frac{p}{p-1}} \\ & \stackrel{\text{Prop. 2.6}}{=} \tilde{\rho}_1(\cdot | \widehat{\mathbb{P}}_N) \square \tilde{\rho}_2(\cdot | \widehat{\mathbb{P}}_N(\omega)) (\text{Id}) \\ & \quad + \left[\|h_1\|_{\frac{p}{p-1}} + \|h_2\|_{\frac{p}{p-1}} \right] W_p(\mathbb{P}_X, \widehat{\mathbb{P}}_N(\omega)) \end{aligned}$$

where the inequality follows from (Pichler 2013a, corollary 11). Interchanging the roles of $\mathbb{P}_X, \widehat{\mathbb{P}}_N(\omega)$, we get

$$\begin{aligned} & \tilde{\rho}_1(\cdot | \widehat{\mathbb{P}}_N(\omega)) \square \tilde{\rho}_2(\cdot | \widehat{\mathbb{P}}_N(\omega)) (\text{Id}) \leq \rho_1 \square \rho_2(X) \\ & \quad + \left[\|h_1\|_{\frac{p}{p-1}} + \|h_2\|_{\frac{p}{p-1}} \right] W_p(\widehat{\mathbb{P}}_N(\omega), \mathbb{P}_X) \end{aligned}$$

so that (19) holds. By lemma 3.1 there exists $E \in \mathcal{F}$ with $\mathbb{P}(E) = 0$ such that $\lim_{N \rightarrow +\infty} W_p(\mathbb{P}_X, \widehat{\mathbb{P}}_N(\omega)) = 0$ for all $\omega \in \Omega \setminus E$. Thus, by (19), we have

$$\rho_1 \square \rho_2(X) = \lim_{N \rightarrow +\infty} \tilde{\rho}_1(\cdot | \widehat{\mathbb{P}}_N) \square \tilde{\rho}_2(\cdot | \widehat{\mathbb{P}}_N) (\text{Id}) \quad \forall \omega \in \Omega \setminus E. \quad \blacksquare$$

4. Numerical experiments

In this section, we illustrate the results of a number of numerical experiments that showcase the usefulness of the approximation developed in section 2. We first test our findings in the case of entropic risk measures and expected shortfall, where simple explicit formulas for the optimal allocations and for the value of the inf-convolutions are known. We then consider the more complex case of distortion risk measures and, to conclude, we treat the case of heterogeneous agents adopting risk measures in two different classes, i.e. entropic and distortion-type.

4.1. Description of the framework

We model two agents with reference risk measures ρ_1 and ρ_2 as those in the introduction. For the sake of comparison, we consider a financial position $X \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$, as it belongs to the domain of each of those risk measures. The objective is to approximate optimal allocations for the inf-convolution of ρ_1 and ρ_2 , which takes the form

$$\rho_1 \square \rho_2(X) = \inf \left\{ \tilde{\rho}_1(f | \mathbb{P}_X) + \tilde{\rho}_2(\text{Id} - f | \mathbb{P}_X) : f \in \mathcal{A}_{\text{Lip}}^0 \right\}.$$

To model the functions f and $\text{Id} - f$, we use two Fully Connected Deep Neural Networks (DNNs) ϕ_1 and ϕ_2 , respectively. We observe that, while ϕ_1 and ϕ_2 explicitly parametrize f and $\text{Id} - f$ by design, the functions $\text{Id} - \phi_1$ and $\text{Id} - \phi_2$ are proxies for $\text{Id} - f$ and f respectively.

Let $\tilde{X} = (X_1, \dots, X_N)$ be a sample of \mathbb{P}_X of size N with $\widehat{\mathbb{P}}_N$ its empirical measure. For $i = 1, 2$, we denote by $\widehat{\rho}_i$ the historical risk measures associated to ρ_1 and ρ_2 , namely, $\widehat{\rho}_i(\cdot) := \tilde{\rho}_i(\cdot | \widehat{\mathbb{P}}_N)$. Since we aim at finding allocations that realize an inf-convolution value as close as possible to $\rho_1 \square \rho_2(X)$, we need to find an appropriate and robust estimate of such a quantity. While we could use the explicit parametrizations ϕ_1 of f and ϕ_2 of $\text{Id} - f$ and minimize $\widehat{\rho}_1(\phi_1(X)) + \widehat{\rho}_2(\phi_2(X))$, another valid alternative is to use the implicit parametrizations and minimize $\widehat{\rho}_1(\text{Id} - \phi_2(X)) + \widehat{\rho}_2(\text{Id} - \phi_1(X))$. To have a more robust estimate, we use their arithmetic average which leads to the following loss function

$$\begin{aligned} L_{\rho_1, \rho_2}(\tilde{X}) & = \frac{1}{2} \left(\widehat{\rho}_1(\phi_1(\tilde{X})) + \widehat{\rho}_2(\phi_2(\tilde{X})) \right. \\ & \quad \left. + \widehat{\rho}_1(\tilde{X} - \phi_2(\tilde{X})) + \widehat{\rho}_2(\tilde{X} - \phi_1(\tilde{X})) \right) \end{aligned} \quad (20)$$

and to the following parameterizations of f and $\text{Id} - f$

$$\begin{aligned} f_1(x) & := \frac{\phi_1(x) + x - \phi_2(x)}{2}, \\ f_2(x) & := \frac{\phi_2(x) + x - \phi_1(x)}{2} = \text{Id}(x) - f_1(x). \end{aligned}$$

Not only these are more robust estimates, but as they sum up to the identity, they provide acceptable allocations by construction. Theorem 3.3 guarantees the convergence of the induced optimal allocations for the risk-sharing problem.

To train the neural networks and obtain the estimators of the theoretical optimum $(\Phi_1, \Phi_2) := (\hat{\varphi}, \text{Id} - \hat{\varphi}) \in \mathcal{A}_{\text{Lip}}^0 \times \mathcal{A}_{\text{Lip}}^0$, we minimize (20) with the optimizer Adam Kingma and Ba (2014). The precise choices of the learning rate, batch size, the number of training epochs, and other implementation details are reported in the Appendix. To ensure a robust framework, we train the DNNs multiple times and use their average as the final estimate. To be more explicit, we train f_1 and f_2 for n times each starting from a different initialization. We obtain n couples of neural networks (f_1^k, f_2^k) and use their arithmetic averages

$$\hat{\Phi}_1(\cdot) := \frac{1}{n} \sum_{k=1}^n f_1^k(\cdot), \quad \hat{\Phi}_2(\cdot) := \frac{1}{n} \sum_{k=1}^n f_2^k(\cdot)$$

to estimate Φ_1 and Φ_2 , respectively. In all our experiments we chose $n = 3$.

To have a flexible framework, we allow the networks to have three types of different activation functions:

$$\sigma(x) = \text{Tanh}(x), \quad \sigma(x) = \text{ReLu}(x), \quad \sigma(x) = x.$$

While the non-linear activation functions Tanh and ReLu are standard choices in deep learning, the reason for including the linear one will be apparent below. We incidentally note that, since X is bounded, we have no issues in allowing for unbounded activation functions. We report a review of possible methodological and architectural enhancements in appendix A.1.

To verify the stability and reliability of our framework, we test our results with three different distributions

- (i) $X \sim \mathcal{U}[-1, 1]$, where \mathcal{U} is the uniform distribution;
- (ii) $X \sim \mathcal{N}(0, 1)$, where \mathcal{N} is the normal distribution;
- (iii) $X \sim -\text{Beta}(2, 5)$, where *Beta* is the *Beta* distribution.

The uniform distribution is the most basic example and it provides an easy setup to test our framework. A more interesting example is the normal distribution because of its well-known financial relevance. We note that in our experiments we restricted to $[-3, 3]$ in order to have a distribution with bounded support. Finally, the *Beta* distribution presents skewness and rare events, modeling the scarcity of data for extreme losses. We chose the opposite of a *Beta* distribution in order to represent the financially more relevant case of pure losses.

4.2. Initial tests: entropic and expected shortfall case

To begin with, we test our framework in the well-known cases of entropic risk measures and expected shortfalls, for which explicit formulas are known. We start by recalling

that, as in Examples 2.8 and 2.9 in Filipović and Svindland (2008),

$$\begin{aligned} \text{Entr}_\alpha \square \text{Entr}_\beta(X) &= \text{Entr}_{\alpha+\beta}(X), \\ \text{ES}_\alpha \square \text{ES}_\beta(X) &= \text{ES}_{\alpha \vee \beta}(X). \end{aligned} \quad (21)$$

This means that we can directly compute the theoretical value of the inf-convolution and compare it with the value $L_{\rho_1, \rho_2}(\tilde{X})$ obtained by the DNNs. Additionally, we can calculate the L^2 error (under \mathbb{P}_X) between the estimated $\hat{\Phi}_1$ and $\hat{\Phi}_2$ and the theoretical ones. As we show below, we found that the values of the inf-convolutions achieved by all our trained networks converge to the theoretical values and that $\hat{\Phi}_i$ approximates Φ_i up to a negligible error, for $i = 1, 2$.

- For the entropic case, we chose $\rho_1(X) = \text{Entr}_2(X)$ and $\rho_2(X) = \text{Entr}_3(X)$ which yield the optimal allocation $\Phi_1(x) = \frac{2}{2+3}x$ and $\Phi_2(x) = \frac{3}{2+3}x$;
- For the ES case, we chose $\rho_1(X) = \text{ES}_{0.8}(X)$ and $\rho_2(X) = \text{ES}_{0.7}(X)$. It is clear from (21) that $\Phi_1(x) = x$ and $\Phi_2(x) = 0$ is an optimal allocation.

We start by discussing the entropic case. Figure 1 shows the comparison between the theoretical optimal allocations and the average predicted $\hat{\Phi}_1$ and $\hat{\Phi}_2$ for the normal distribution case and for the three activation functions. Every trained DNN seems to match perfectly the theoretical allocations. Indeed, we point out that the average predicted allocations in figure 1 are plotted with their respective ± 3 standard deviation bands across the n networks. In particular, for this case, we notice that the uncertainty bands are invisible as they are almost null. In figure 2(a), we show the comparison between the average loss functions, along with the respective ± 3 standard deviation shaded band, and the theoretical infimum calculated using (21). All three types of NNs achieve a loss that is close to the theoretical value of the inf-convolution, up to a negligible error. In figure 2(b), for each model, we plot the standard deviation of the loss function (20). Since the variance of the three losses is decreasing, we are observing a stable convergence. Table 1 collects the data regarding the errors for the experiment with the standard normal distribution. We computed the average relative error with respect to the theoretical infimum, together with its standard deviation, and the L^2 error of $\hat{\Phi}_1$ with respect to the theoretical Φ_1 . Table A1 reports the same figures also for the cases of uniform and *Beta* distributions. We observe that the errors are all close to zero, meaning that all our NNs reached convergence and they exhibited low uncertainty, which is an indication of stable learning. Observe that for the entropic case (as well as for ES) the optimal allocation is a linear function, therefore, it belongs to the span of the linear-activated DNN. We thus expect the linear activation to achieve the best performance. As we can appreciate in table 1, this result is confirmed by our experiments. Additionally, we notice that also ReLu and Tanh are providing satisfactory performances.

Similar considerations apply to the case of ES and we obtain qualitatively and quantitatively the same results. As an example, in figure 3 we show the results for the uniform distribution case. Regarding the convergence analysis, we observe in figure 4 that all three types of NNs achieve a loss that is only marginally distant from the theoretical value

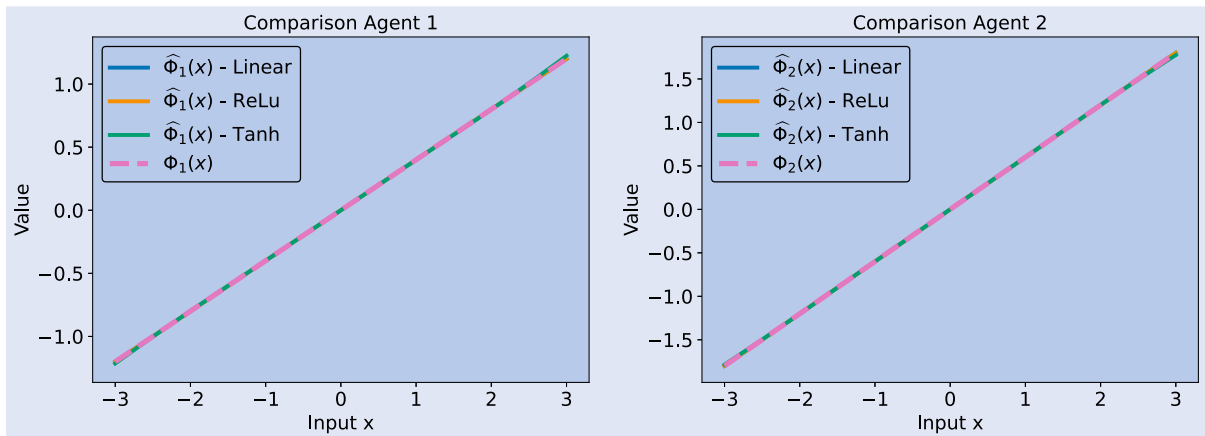


Figure 1. The entropic case – normal distribution – comparison of predicted vs. theoretical allocations. We train all models over 3 trials and plot the average predicted allocation along with the ± 3 standard deviations.

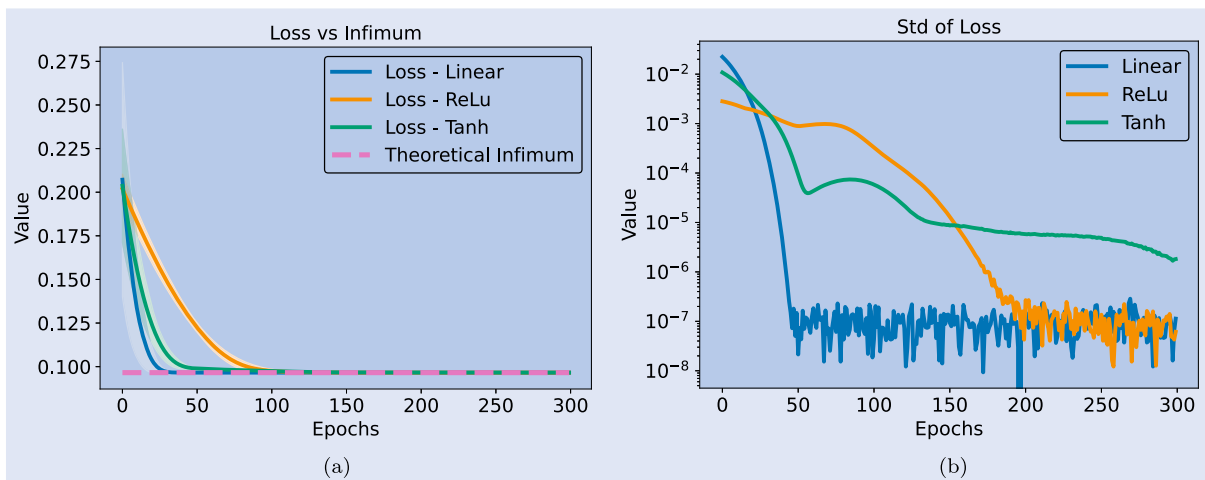


Figure 2. The entropic case – normal distribution – convergence analysis. (a) Average training loss along with ± 3 standard deviation and (b) Standard deviation of the loss.

Table 1. Average relative errors between the losses and the theoretical infimum, their standard deviation, and the average L^2 error between $\hat{\Phi}_1$ and Φ_1 .

Entropic case – $\mathcal{N}(0, 1)$ – Infimum = 0.09656			
	Avg. Rel. Error	Std. Rel. Error	Avg. L^2 Error $\hat{\Phi}_1$
Linear	$\%1.800 \cdot 10^{-5}$	$\%1.172 \cdot 10^{-5}$	$1.788 \cdot 10^{-7}$
ReLu	$\%5.143 \cdot 10^{-6}$	$\%6.341 \cdot 10^{-5}$	$4.096 \cdot 10^{-5}$
Tanh	$\%5.955 \cdot 10^{-6}$	$\%1.814 \cdot 10^{-6}$	$5.813 \cdot 10^{-5}$

of the inf-convolution. Additionally, we notice that the convergence to such a value takes place with decreasing variance of the losses, indicating a stable convergence. In figure 3 we present the comparison between the average predicted $\hat{\Phi}_1$ and $\hat{\Phi}_2$, along with their respective ± 3 standard deviation shaded band, and the theoretical optimal allocations for the uniform distribution case. A consideration, which is specific to the ES case, is now due. The optimal allocation is of the form $\Phi_1(x) = x$ and $\Phi_2(x) = 0$, and the DNN needs to learn the constant function in the latter case. Consistently with the known fact that using nonlinear functions for the (unsupervised) learning of constant functions is a challenging task, we find that ReLu and Tanh underperform with respect to the linear DNN.

In table 2, we finally present the average relative error with respect to the theoretical infimum, its standard deviation, and L^2 error of $\hat{\Phi}_1$ with respect to Φ_1 . Table A2 reports the same figures also for the case of normal and *Beta* distributions.

4.3. Convolution of distortion risk measures

We here consider the case where both ρ_1 and ρ_2 are distortion risk measures, as in (3), with respect to some discrete probabilities μ_1, μ_2 . Let N_1, N_2 be two given integers, and consider the risk measures

$$\rho_1(X) := \sum_{j=1}^{N_1} \mu_{1j} \text{ES}_{\alpha_{1j}}(X), \quad \rho_2(X) := \sum_{j=1}^{N_2} \mu_{2j} \text{ES}_{\alpha_{2j}}(X),$$

where $\mu_{ij} > 0$ with $\sum_{j=1}^{N_i} \mu_{ij} = 1$ and $0 < \alpha_{ij} < 1$ for $j = 1, \dots, N_i$ and for $i = 1, 2$. Some semi-explicit expressions of the optimal allocations are known for this case, in particular, an optimal allocation can be found as a linear combination of ReLu functions, possibly composed with translation maps—see Example 3.1 in Jouini *et al.* (2008) and also Appendix A of Embrechts *et al.* (2018) for a more general case[†]. Hence,

[†] We thank an anonymous referee for pointing out this fact.

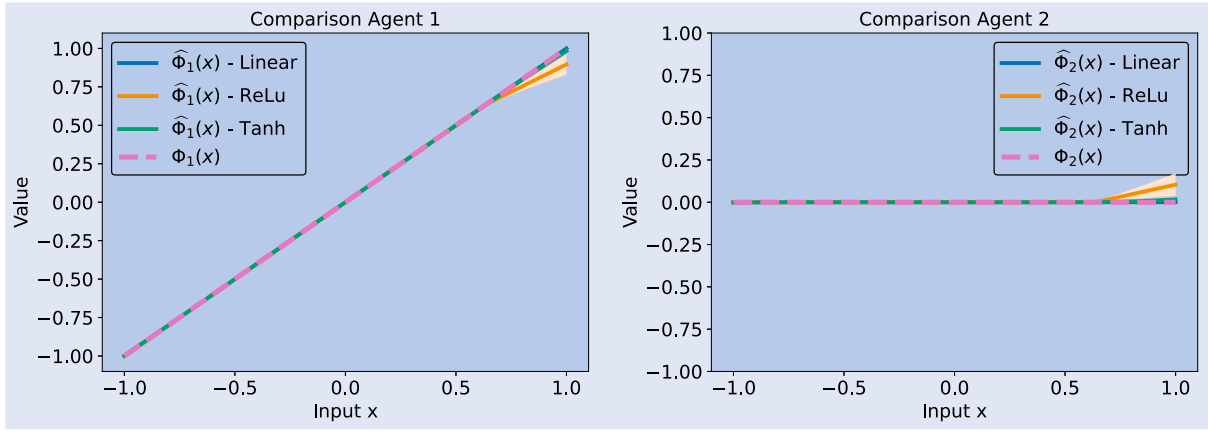


Figure 3. Expected shortfall – uniform distribution – comparison of predicted vs. theoretical allocations. We train all models over 3 trials and plot the average predicted allocation along with the ± 3 standard deviations.

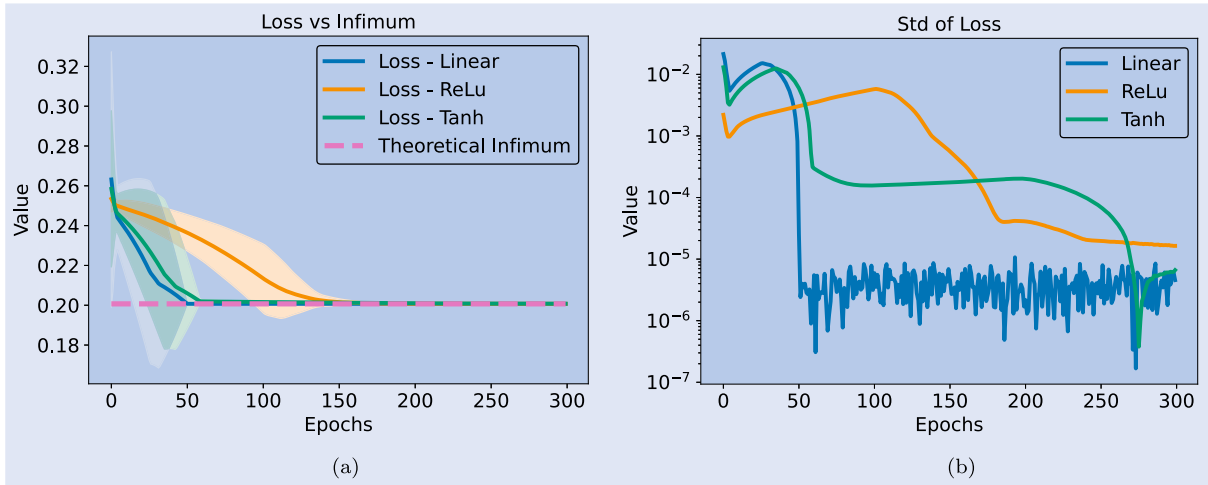


Figure 4. Expected shortfall – uniform distribution – convergence analysis. (a) Average training loss along with ± 3 standard deviation and (b) Standard deviation of the loss.

we expect the ReLu-activated DNN to achieve the best performance. Differently from the entropic and ES cases, the problem has a non-linear solution and we expect the linear-activated DNN to perform poorly. Nevertheless, for the sake of consistency in our tests, we included the linear activation in all experiments. As an example, we chose

$$\begin{aligned}\rho_1(X) &= 0.5ES_{0.8}(X) + 0.5ES_{0.7}(X), \\ \rho_2(X) &= 0.7ES_{0.9}(X) + 0.3ES_{0.5}(X).\end{aligned}$$

Figure 5 shows the average predicted $\hat{\Phi}_1$ and $\hat{\Phi}_2$ for the case of the *Beta* distribution and for the three activation functions. As we can observe, the DNNs trained with non-linear activation functions agree on the shape of the solution, whereas the linear-activated one is clearly different.

As anticipated before, we expect the optimal allocations to be linear combinations of ReLu activations. Consistently with the theory, if we look at the average predicted $\hat{\Phi}_1(x)$ for the case of ReLu, as in figure 6(b), we observe that such expected behavior is captured. Figure 6(a) shows the average loss functions as a function of the training epochs. First of all, we notice that the linear NN achieves a loss level that is sensibly larger than those achieved by the ReLu and the Tanh DNNs, confirming the expectations of its poor performance. Secondly, we

Table 2. Average relative errors between the losses and the theoretical infimum, their standard deviation, and the average L^2 error between $\hat{\Phi}_1$ and Φ_1 .

Expected Shortfall – $\mathcal{U}[-1, 1]$ – Infimum = 0.2006			
	Avg. Rel. Error	Std. Rel. Error	Avg. L^2 Error $\hat{\Phi}_1$
Linear	$\%2.846 \cdot 10^{-4}$	$\%2.270 \cdot 10^{-3}$	$9.701 \cdot 10^{-6}$
ReLu	$\%1.379 \cdot 10^{-2}$	$\%8.171 \cdot 10^{-3}$	$1.817 \cdot 10^{-2}$
Tanh	$\%2.742 \cdot 10^{-2}$	$\%3.301 \cdot 10^{-3}$	$2.521 \cdot 10^{-3}$

notice that the loss decreases with decreasing variance, indicating a stable convergence with low uncertainty in all three cases.

Finally, in table 3, we present the average achieved losses, together with the uncertainty of their estimates, for all activation functions and for all distributions. In line with the theoretical predictions, the DNN activated with a ReLu function is the one performing best in terms of average loss: all three loss values are, by construction, greater or equal to the theoretical infimum, and the best performance is understood in the sense of achieving the lowest value. The Tanh-activated DNN is comparably reliable. From table 3, we can see that in some cases the linear-activated DNN shows the most stable convergence, namely the lowest standard deviation of

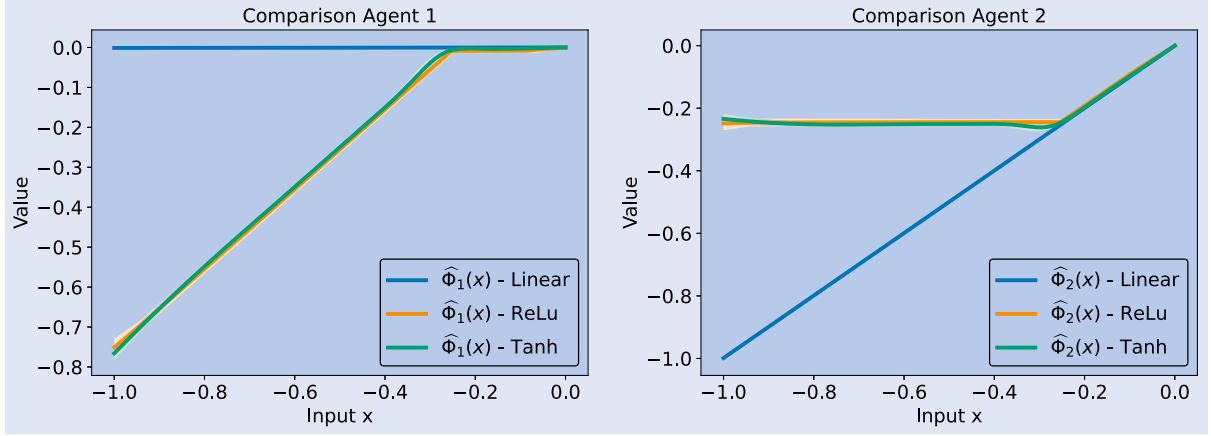


Figure 5. Distortion measures – beta distribution – predicted allocations. We train all models over 3 trials and plot the average predicted allocation along with the ± 3 standard deviations.

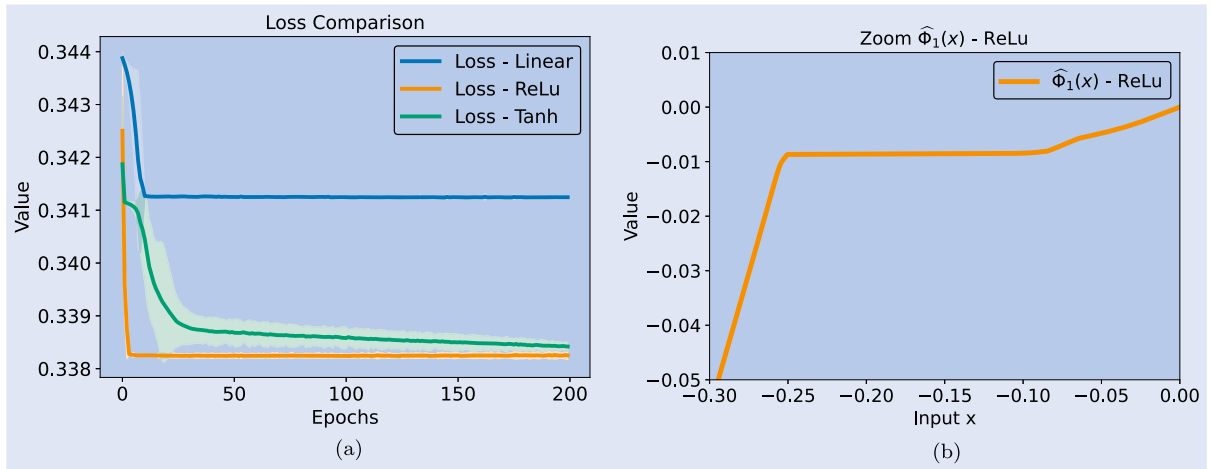


Figure 6. Distortion measures – beta distribution – convergence analysis and optimal ReLu. (a) Average training loss along with ± 3 standard deviation and (b) $\hat{\Phi}_1(x)$ for ReLu-activated DNN as a weighted sum of ReLu activations.

Table 3. Average loss of the achieved training losses together with its standard deviation.

Distortion Measure – $\mathcal{U}[-1, 1]$		
	Avg. Loss	Std. Loss
Linear	0.220786	$1.79364 \cdot 10^{-7}$
ReLu	0.210493	$1.43078 \cdot 10^{-6}$
Tanh	0.210722	$4.81990 \cdot 10^{-5}$
Distortion risk measures – $\mathcal{N}(0, 1)$		
Linear	0.371297	$1.12391 \cdot 10^{-7}$
ReLu	0.355218	$2.53271 \cdot 10^{-7}$
Tanh	0.355505	$3.12838 \cdot 10^{-5}$
Distortion Measure – $-Beta(2, 5)$		
Linear	0.341245	$9.27940 \cdot 10^{-6}$
ReLu	0.338251	$2.52160 \cdot 10^{-6}$
Tanh	0.338419	$2.66384 \cdot 10^{-6}$

losses. However, it converges to a loss value that is significantly higher than the other two. This is not unexpected since, by design, the linear-activated DNN is unable to represent a nonlinear function and, therefore, exhibits poorer performances.

4.4. Heterogeneous agents

In our last experiments, we consider two heterogeneous agents, in the sense that one adopts an entropic risk measure while the other one opts for a distortion-type risk measure. In the first of such experiments, the risk measures are

$$\rho_1(X) = \text{ES}_{0.9}(X), \quad \rho_2(X) = \text{Entr}_{0.3}(X). \quad (22)$$

From Jouini *et al.* (2008, proposition 3.2) or Rüschenendorf (2013, theorem 11.22), the optimal allocation is known to be induced by $(f, \text{Id} - f) = (-(x - k)^-, \max(x, k))$ for some (non-explicit) constant k . In line with the previous subsections, we show an example of the average predicted $\hat{\Phi}_1$ and $\hat{\Phi}_2$. In figure 7, we plot the predicted allocations for the *Beta* distribution, for the three different activation functions. Once again, we expect the solution to be non-linear and we can observe that the optimal allocations found by ReLu and Tanh-activated DNNs are comparable, whereas the one found by the linear-activated DNN differs significantly. In figure 8(b) we isolated the allocation $\hat{\Phi}_1$ found by the ReLu DNN which, as we will see below, is the one that performed best. We notice that the desired behavior of the optimal allocations is well-captured.

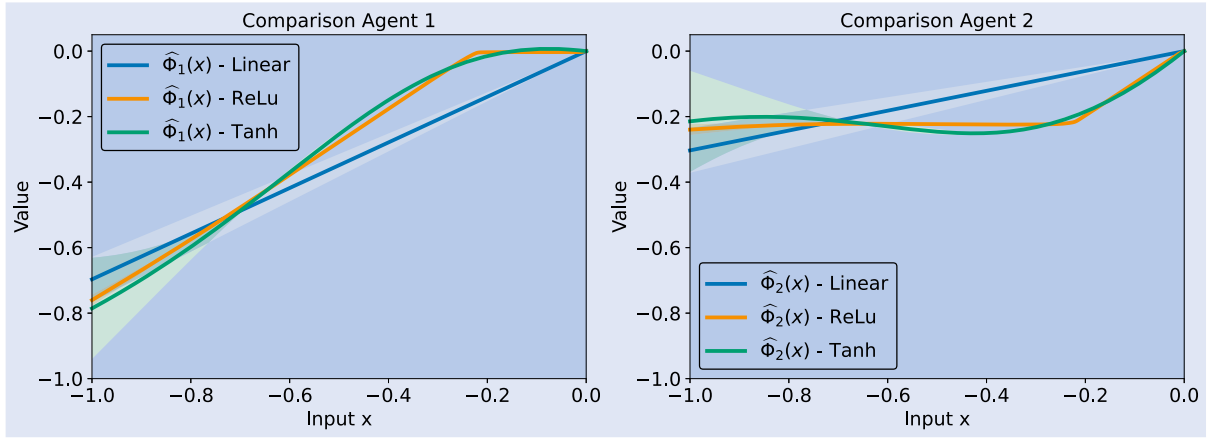


Figure 7. Heterogeneous agents – case equation (22) – beta distribution – predicted allocations. We train all models over 3 trials and plot the average Predicted Allocation along with the ± 3 standard deviations.

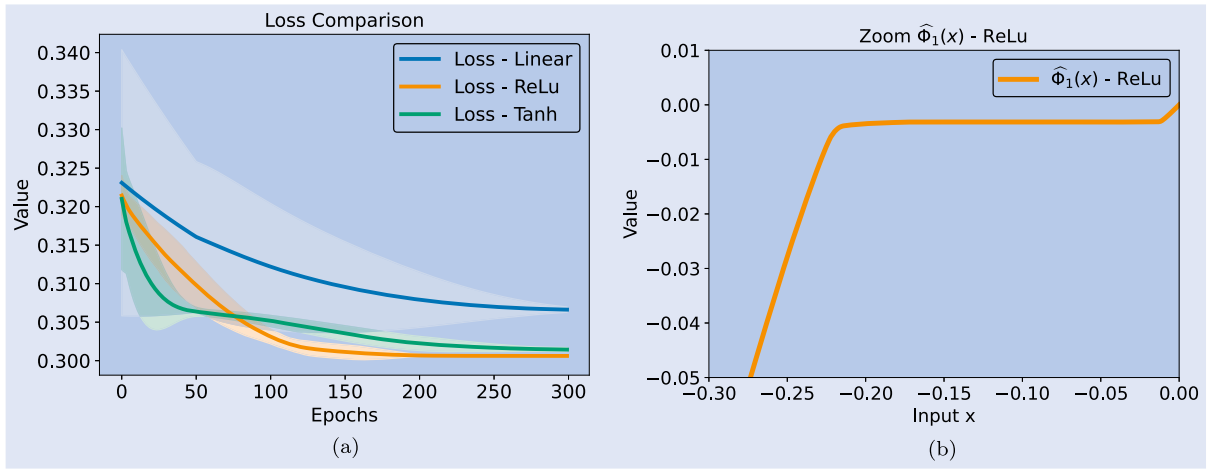


Figure 8. Heterogeneous measures – case equation (22) – beta distribution – convergence analysis and optimal ReLu. (a) Average training loss along with ± 3 standard deviation and (b) $\hat{\Phi}_1(x)$ for ReLu-activated DNN is (almost) as per theoretical prediction.

Figure 8(a) shows the average loss functions as a function of the training epochs, together with their uncertainty-shaded bands. We notice that all networks exhibit stable convergence. However, the linear-activated network achieves a loss level that is sensibly larger than those achieved by the ReLu and Tanh-activated ones. From the picture, it is already clear that ReLu is the one performing best in this case. This is confirmed by the data that we collect in table 4, namely, the average achieved loss together with the uncertainty of their estimates. Nevertheless, we note that while the Tanh NN underperforms with respect to the ReLu one, it provides comparable performances.

In our last experiment, we consider a case where, to the best of our knowledge, no theoretical information is available. Again, we consider two heterogeneous agents, the first one opting for a distortion risk measure, and the second one adopting an entropic risk measure. More precisely, the risk measures are

$$\rho_1(X) = 0.7ES_{0.8}(X) + 0.3ES_{0.7}(X), \quad \rho_2(X) = \text{Entr}_{0.3}(X). \quad (23)$$

In figure 9 we plot the average predicted $\hat{\Phi}_1$ and $\hat{\Phi}_2$ for the beta distribution, for the three different activation functions.

Table 4. Average loss of the achieved training losses together with its standard deviation.

Case equation (22) – $\mathcal{U}[-1, 1]$		
	Avg. Loss	Std. Loss
Linear	0.0962926	$1.62093 \cdot 10^{-6}$
ReLU	0.0837376	$7.15505 \cdot 10^{-6}$
Tanh	0.085397	$7.64700 \cdot 10^{-5}$
Case equation (22) – $\mathcal{N}(0, 1)$		
Linear	0.185575	$1.65563 \cdot 10^{-6}$
ReLU	0.166919	$2.50091 \cdot 10^{-4}$
Tanh	0.169095	$9.12596 \cdot 10^{-6}$
Case equation (22) – $\text{Beta}(2, 5)$		
Linear	0.306628	$1.16732 \cdot 10^{-4}$
ReLU	0.300616	$2.28480 \cdot 10^{-6}$
Tanh	0.301437	$8.87485 \cdot 10^{-5}$

As in the cases in section 4.3 and in the previous heterogeneous case, we anticipate a non-linear behavior, which translates into linear activated DNNs underperforming significantly. We can observe that the optimal allocations found by

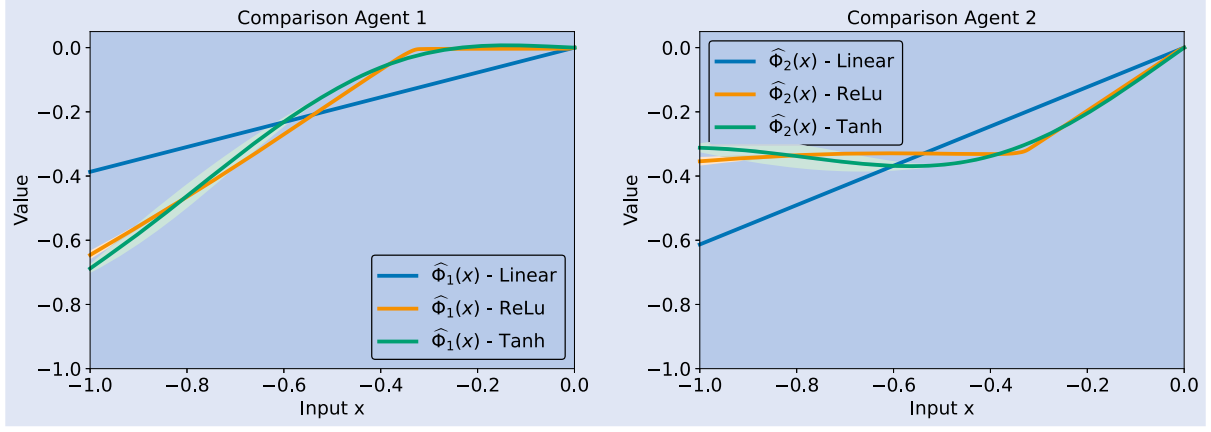


Figure 9. Heterogeneous agents – case equation (23) – beta distribution – predicted allocations. We train all models over 3 trials and plot the average predicted allocation along with the ± 3 standard deviations.

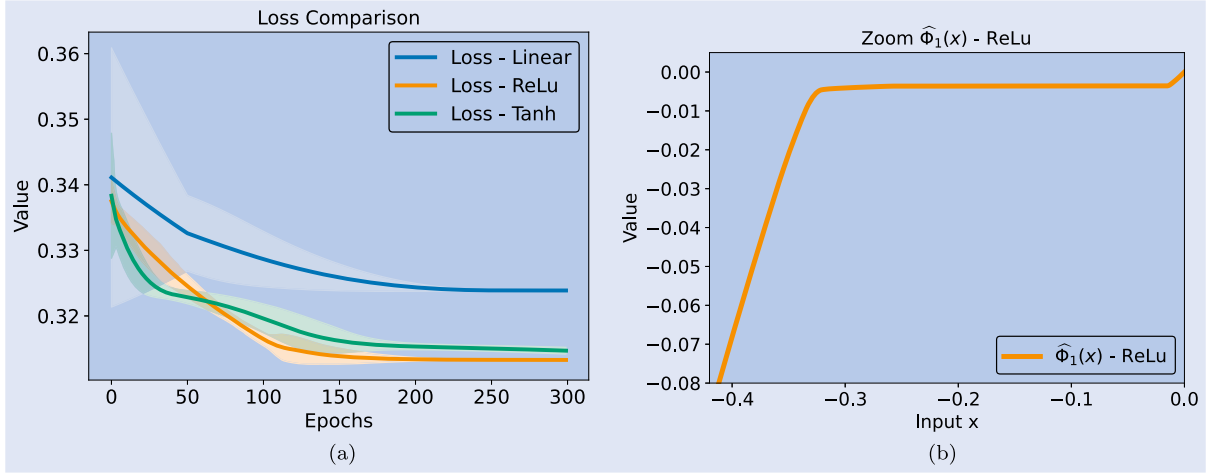


Figure 10. Heterogeneous measures – case equation (23) – beta distribution – convergence analysis and optimal ReLu. (a) Average training loss along with ± 3 standard deviation and (b) $\hat{\Phi}_1(x)$ for ReLu-activated DNN as a weighted sum of ReLu functions.

Table 5. Average loss of the achieved training losses together with its standard deviation.

Case equation (23) – $\mathcal{U}[-1, 1]$		
	Avg. Loss	Std. Loss
Linear	0.2064738	$8.26712 \cdot 10^{-7}$
ReLu	0.1786352	$1.22513 \cdot 10^{-6}$
Tanh	0.1807351	$4.32114 \cdot 10^{-5}$
Case equation (23) – $\mathcal{N}(0, 1)$		
	Avg. Loss	Std. Loss
Linear	0.3662076	$9.0044 \cdot 10^{-7}$
ReLu	0.32997002	$2.92006 \cdot 10^{-5}$
Tanh	0.3324212	$1.65775 \cdot 10^{-4}$
Case equation (23) – $-\text{Beta}(2, 5)$		
	Avg. Loss	Std. Loss
Linear	0.3238797	$1.00014 \cdot 10^{-6}$
ReLu	0.3132883	$1.25900 \cdot 10^{-6}$
Tanh	0.3147001	$1.70660 \cdot 10^{-4}$

ReLu and Tanh-activated DNNs are comparable. In figure 10 we isolated the allocation $\hat{\Phi}_1$ found by the ReLu DNN which, as in the previous heterogeneous case of section 4.4, is the one that performed best, which is confirmed by table 5.

All networks exhibit stable convergence. Still, as expected, the linear-activated DNN achieves a far larger loss level. The Tanh NN underperforms with respect to the ReLu one, yet still provides comparable performances.

Acknowledgments

The authors thank two anonymous referees for precious comments, and F.-B. Liebrich for addressing them to the reference (Shapiro 2013) and for pointing out the delicate point of the standardness requirements on the underlying probability space.

Disclosure statement

The authors report there are no competing interests to declare.

ORCID

M. Burzoni <http://orcid.org/0000-0002-3741-5790>

A. Doldi <http://orcid.org/0000-0001-6443-2409>

E. Monzio Compagnoni <http://orcid.org/0009-0004-7094-2586>

References

- Acciaio, B., Optimal risk sharing with non-monotone monetary functionals. *Finance Stoch.*, 2007, **11**(2), 267–289.
- Aliprantis, C.D. and Border, K.C., *Infinite Dimensional Analysis*, 3rd ed., 2006 (Springer: Berlin).
- Barrieu, P. and El Karoui, N., Inf-convolution of risk measures and optimal risk transfer. *Finance Stoch.*, 2005, **9**(2), 269–298.
- Biagini, S. and Frittelli, M., On the extension of the Namioka-Klee theorem and on the Fatou property for risk measures. In *Optimality and Risk-Modern Trends in Mathematical Finance: The Kabanov Festschrift*, pp. 1–28, 2010 (Springer: Berlin, Heidelberg).
- Billingsley, P., *Convergence of Probability Measures*, 2nd ed., Wiley Series in Probability and Statistics: Probability and Statistics, 1999 (John Wiley & Sons Inc.: New York).
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G. and Askell, A., et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.*, 2020, **33**, 1877–1901.
- Carlier, G. and Dana, R.A., Pareto optima and equilibria when preferences are incompletely known. *J. Econ. Theory*, 2013, **148**(4), 1606–1623.
- Carlier, G., Dana, R.A. and Galichon, A., Pareto efficiency for the concave order and multivariate comonotonicity. *J. Econ. Theory*, 2012, **147**(1), 207–229.
- Clevert, D.A., Unterthiner, T. and Hochreiter, S., Fast and accurate deep network learning by exponential linear units (elus), 2015, preprint, arXiv:1511.07289.
- Compagnoni, E.M., Orvieto, A., Biggio, L., Kersting, H., Proske, F.N. and Lucchi, A., An SDE for modeling SAM: Theory and insights. In *ICML 2023*, Honolulu, Hawaii, 2023.
- Compagnoni, E.M., Scapicchio, A., Biggio, L., Orvieto, A., Hofmann, T. and Teichmann, J., On the effectiveness of randomized signatures as reservoir for learning rough dynamics. In *IJCNN 2023*, Queensland, Australia, 2023.
- Cont, R., Deguest, R. and Scandolo, G., Robustness and sensitivity analysis of risk measurement procedures. *Quant. Finance*, 2010, **10**(6), 593–606.
- Cuchiero, C., Schmock, P. and Teichmann, J., Global universal approximation of functional input maps on weighted spaces. In preparation, 2022.
- Dana, R.A. and Le Van, C., Overlapping sets of priors and the existence of efficient allocations and equilibria for risk measures. *Math. Finance*, 2010, **20**(3), 327–339.
- Daniels, H. and Velikova, M., Monotone and partially monotone neural networks. *IEEE Trans. Neural Netw.*, 2010, **21**(6), 906–917.
- Delbaen, F., Law of large numbers for risk measures, 2021. arXiv preprint arXiv:2109.10612v1.
- Doldi, A., Feng, Y., Fouque, J.P. and Frittelli, M., Multivariate systemic risk measures and computation by deep learning algorithms. *Quant. Finance*, 2023, **23**(10), 1431–1444.
- Doldi, A., Frittelli, M. and Rosazza Gianin, E., Short communication: Are shortfall systemic risk measures one dimensional? *SIAM J. Financ. Math.*, 2024, **15**(1), SC1–SC14.
- Dörsek, P. and Teichmann, J., A semigroup point of view on splitting schemes for stochastic (partial) differential equations, 2022. arXiv preprint arXiv:1011.2651v1.
- Embrechts, P., Liu, H., Mao, T. and Wang, R., Quantile-based risk sharing with heterogeneous beliefs. *Math. Program.*, 2020, **181**(2), 319–347.
- Embrechts, P., Liu, H. and Wang, R., Quantile-based risk sharing. *Oper. Res.*, 2018, **66**(4), 936–949.
- Feng, Y., Min, M. and Fouque, J.P., Deep learning for systemic risk measures. In *Proceedings of the Third ACM International Conference on AI in Finance*, ICAIF '22, New York, NY, USA, pp. 62–69, 2022 (Association for Computing Machinery: New York, NY, USA).
- Filipović, D. and Svindland, G., Optimal capital and risk allocations for law- and cash-invariant convex functions. *Finance Stoch.*, 2008, **12**(3), 423–439.
- Föllmer, H. and Schied, A., *Stochastic Finance. An Introduction in Discrete Time*. Fourth revised and extended, 2016 (De Gruyter Graduate: De Gruyter, Berlin).
- Frittelli, M. and Maggis, M., Disentangling price, risk and model risk: V&R measures. *Math. Financial Econ.*, 2018, **12**(2), 219–247.
- Heath, D. and Ku, H., Pareto equilibria with coherent measures of risk. *Math. Finance*, 2004, **14**(2), 163–172.
- Hendrycks, D. and Gimpel, K., Gaussian error linear units (GELUS), 2016. arXiv preprint arXiv:1606.08415.
- Hornik, K., Approximation capabilities of multilayer feedforward networks. *Neural Netw.*, 1991, **4**(2), 251–257.
- Jouini, E., Schachermayer, W. and Touzi, N., Optimal risk sharing for law invariant monetary utility functions. *Math. Finance*, 2008, **18**(2), 269–292.
- Kailath, T. and Weinert, H., An RKHS approach to detection and estimation problems—II: Gaussian signal detection. *IEEE Trans. Inf. Theory*, 1975, **21**(1), 15–23.
- Kingma, D.P. and Ba, J., Adam: A method for stochastic optimization. Cite arxiv:1412.6980. Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015, 2014.
- Kratsios, A., The universal approximation property: Characterization, construction, representation, and existence. *Ann. Math. Artif. Intell.*, 2021, **89**(5–6), 435–469.
- Liebrich, F.B. and Svindland, G., Risk sharing for capital requirements with multidimensional security markets. *Finance Stoch.*, 2019, **23**(4), 925–973.
- Liu, X., Han, X., Zhang, N. and Liu, Q., Certified monotonic neural networks. *Adv. Neural Inf. Process. Syst.*, 2020, **33**, 15427–15438.
- Mastrogiovanni, E. and Rosazza Gianin, E., Pareto optimal allocations and optimal risk sharing for quasiconvex risk measures. *Math. Financ. Econ.*, 2015, **9**(2), 149–167.
- Pichler, A., Evaluations of risk measures for different probability measures. *SIAM J. Optim.*, 2013a, **23**(1), 530–551.
- Pichler, A., The natural Banach space for version independent risk measures. *Insur. Math. Econ.*, 2013b, **53**(2), 405–415.
- Rahimi, A. and Recht, B., Random features for large-scale kernel machines. In *Proceedings of the Advances in Neural Information Processing Systems*, edited by J. Platt, D. Koller, Y. Singer and S. Roweis, Vol. 20, 2007 (Curran Associates, Inc.: Vancouver, BC).
- Rahimi, A. and Recht, B., Uniform approximation of functions with random bases. In *Proceedings of the 2008 46th Annual Allerton Conference on Communication, Control, and Computing*, Vancouver, B.C., Canada, pp. 555–561, 2008.
- Rüschendorf, L., *Mathematical Risk Analysis: Dependence, Risk Bounds, Optimal Allocations and Portfolios*, Springer Series in Operations Research and Financial Engineering, 2013 (Springer: Heidelberg).
- Schölkopf, B., Muandet, K., Fukumizu, K., Harmeling, S. and Peters, J., Computing functions of random variables via reproducing kernel Hilbert space representations. *Stat. Comput.*, 2015, **25**(4), 755–766.
- Scholkopf, B. and Smola, A.J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, 2018 (MIT press).
- Shapiro, A., Consistency of sample estimates of risk averse stochastic programs. *J. Appl. Probab.*, 2013, **50**(2), 533–541.
- Svindland, G., Continuity properties of law-invariant (quasi-)convex risk functions on L^∞ . *Math. Financ. Econ.*, 2010, **3**(1), 39–43.
- Theodoridis, S. and Koutroumbas, K., *Pattern Recognition*, 2006 (Elsevier).
- Tsanakas, A., To split or not to split: Capital allocation with convex risk measures. *Insur. Math. Econ.*, 2009, **44**(2), 268–277.
- Villani, C., *Optimal Transport, Grundlehren der Mathematischen Wissenschaften Fundamental Principles of Mathematical Sciences*, Vol. 338, 2009 (Springer-Verlag: Berlin, Old and new).
- Weber, S., Solvency II, or how to sweep the downside risk under the carpet. *Insur. Math. Econ.*, 2018, **82**, 191–200.
- Wen, K., Ma, T. and Li, Z., How sharpness-aware minimization minimizes sharpness? In *Proceedings of the Eleventh International Conference on Learning Representations*, Kigali, Rwanda, 2023.

Appendices

Appendix 1. Implementation details and additional experimental results

All code is implemented in Python and the Deep Learning library used is PyTorch. In each experiment, the dataset \tilde{X} is of size $N = 100,000$, while the batch size is $b = 1000$. All the neural networks have 3 hidden layers of 100 neurons each and have been optimized with Adam. More precisely, the learning rate is 10^{-6} while all other settings of Adam are as per default setting. We remind that it is a very well-known result of convex optimization that the learning rate has to be smaller than twice the inverse of the largest eigenvalue of the loss function for Gradient Descent to converge. In practice, this is a valuable indication also in nonconvex optimization. Even if our choice for the learning rate might seem unusual, such a low value was necessary for our experiments, as we observed that higher ones would lead to instability in the optimization process. This is oftentimes an indication that the optimization problem at hand is rather nonlinear and the loss landscape is irregular, together with its derivatives. To make the convergence even more stable, we used the *ReduceLROnPlateau* scheduler for the learning rate, with *patience* equal to 1000 and *threshold* equal to 10^{-6} , while all other parameters are as per default specification. Finally, all experiments have been run for a number of *epochs* equal to 300, apart from those for the Distortion Measures where the number of *epochs* is 200. The optimal hyperparameters are the result of fine-tuning via extensive grid search.

We finally complete the exposition of the numerical results for the entropic risk measure and expected shortfall experiments. Tables A1 and A2 contain the average relative error with respect to the theoretical infimum, together with its standard deviation, and the L^2 error of $\hat{\Phi}_1$ with respect to the theoretical $\hat{\varphi}$ for all distributions and activation functions.

In our experiments, we observe that both ReLu and Tanh activation functions performed well in all cases, even when the solution was known to be linear. ReLu seemed to perform better in most of the cases. This is due to the fact that in some cases the semi-explicit solution has a piecewise linear behavior.

A.1. Possible enhancements

The deep learning literature offers a variety of architectural and methodological enhancements that could be used to further push the results that we obtained.

Table A1. Average relative errors between the losses and the theoretical infimum, their standard deviation, and the average L^2 error between $\hat{\Phi}_1$ and Φ_1 .

Entropic case – $\mathcal{U}[-1, 1]$ – Infimum = 0.03423			
	Avg. Rel. Error	Std. Rel. Error	Avg. L^2 Error $\hat{\Phi}_1$
Linear	% 2.466 · 10^{-4}	% 2.888 · 10^{-4}	1.207 · 10^{-8}
ReLu	%9.067 · 10^{-4}	%4.682 · 10^{-4}	1.802 · 10^{-4}
Tanh	%1.135 · 10^{-3}	%3.498 · 10^{-4}	1.077 · 10^{-4}
Entropic case – $\mathcal{N}(0, 1)$ – Infimum = 0.09656			
Linear	%1.800 · 10^{-5}	%1.172 · 10^{-5}	1.788 · 10^{-7}
ReLu	% 5.143 · 10^{-6}	%6.341 · 10^{-5}	4.096 · 10^{-5}
Tanh	%5.955 · 10^{-6}	% 1.814 · 10^{-6}	5.813 · 10^{-5}
Entropic case – <i>Beta</i> (2, 5) – Infimum = 0.2876			
Linear	%8.979 · 10^{-5}	% 2.128 · 10^{-5}	3.759 · 10^{-8}
ReLu	% 7.943 · 10^{-5}	%5.504 · 10^{-5}	3.146 · 10^{-4}
Tanh	%1.001 · 10^{-3}	%3.202 · 10^{-5}	1.299 · 10^{-4}

Table A2. Average relative errors between the losses and the theoretical infimum, their standard deviation, and the average L^2 error between $\hat{\Phi}_1$ and Φ_1 .

Expected Shortfall – $\mathcal{U}[-1, 1]$ – Infimum = 0.2006			
	Avg. Rel. Error	Std. Rel. Error	Avg. L^2 Error $\hat{\Phi}_1$
Linear	% 2.846 · 10^{-4}	% 2.270 · 10^{-3}	9.701 · 10^{-6}
ReLu	%1.379 · 10^{-2}	%8.171 · 10^{-3}	1.817 · 10^{-2}
Tanh	%2.742 · 10^{-2}	%3.301 · 10^{-3}	2.521 · 10^{-3}
Expected Shortfall – $\mathcal{N}(0, 1)$ – Infimum = 0.3459			
Linear	% 2.961 · 10^{-3}	% 1.070 · 10^{-3}	3.050 · 10^{-5}
ReLu	%1.910 · 10^{-2}	%1.036 · 10^{-2}	3.325 · 10^{-2}
Tanh	%1.530 · 10^{-1}	%1.867 · 10^{-2}	2.511 · 10^{-2}
Expected Shortfall – <i>Beta</i> (2, 5) – Infimum = 0.3343			
Linear	% 6.031 · 10^{-4}	% 5.253 · 10^{-4}	1.813 · 10^{-5}
ReLu	%1.455 · 10^{-3}	%5.449 · 10^{-4}	3.969 · 10^{-2}
Tanh	%1.017 · 10^{-1}	%1.506 · 10^{-2}	2.284 · 10^{-3}

One could include several other activation functions, such as GELU (Hendrycks and Gimpel (2016)) or ELU which obtained recent success in NLP (Brown *et al.* (2020)) and Image Classification (Clevert *et al.* (2015)), respectively. Similarly, one could try different optimizers which may converge to more stable regions of the loss landscape. For example, recent optimizers that found great success in NLP and Computer Vision are SAM and its variants. As detailed in Wen *et al.* (2023) and Compagnoni, Orvieto, *et al.* (2023), this class of optimizers drives the dynamics towards flatter regions of the landscapes which result in provenly more stable DNNs. Other possibilities include standard techniques such as Batch Normalization and Residual Connections which are proven to stabilize the optimization process.

Finally, since the functions we are learning are monotonic, an interesting approach, suggested by an anonymous referee, would be to enforce the monotonicity of the approximating functions. This could be attained by leveraging specific network structures such as in Daniels and Velikova (2010) or suitable penalization terms (Liu *et al.* 2020). While all our experiments reached convergence without the need of imposing monotonicity, this might be necessary in other cases where convergence is more elusive. As a side note, we remark that not enforcing a priori monotonicity allows for a further sanity check in the experiments, as we can check if the monotone behavior of the optima is learned without any external enforcement.

It is worth noting that, for all the architectural changes that would alter the DNNs, one should of course provide the proof of suitable versions of theorems 2.8, 2.11 and 3.3 for this very specific class of NNs. Since our experiments already achieved satisfactory results, there is no compelling reason to do so at the moment, and we leave these for future research.

Appendix 2. Modeling alternatives

As suggested by an anonymous referee, there might be other possible ways to successfully model the functions f and $\text{Id} - f$, for example, using a basis-based approach, such as Random Feature Models, Rahimi and Recht (2008) or using Kernel functions (Scholkopf and Smola 2018).

In the basis-based approach, it is required to fix (or randomly generate) a number of representations of the input and then to linearly combine them to fit the output via a linear layer. These techniques have proven to be effective and computationally cheap in many fields (Rahimi and Recht 2007). However, the key to their success is a careful design and selection of the (possibly random) features, an operation which is not always straightforward (Compagnoni,

Scampicchio, *et al.* 2023). Much differently, DNNs are able to learn and adapt the features during the optimization procedure.

The second approach is based on Reproducing Kernel Hilbert Space (RKHS), also known in the Machine Learning community as *kernel methods*. This is a very powerful set of techniques that maps the input data into a higher (possibly infinite) dimensional space, in which it is easier to separate data points respect to their native space. These methods found success in many applications (Scholkopf and Smola 2018) such as in Classification, Signal Detection (Kailath and Weinert 1975), and Function Emulation (Schölkopf *et al.* 2015). However, we find that the *kernel trick* (Theodoridis and Koutroumbas 2006) at the basis of these methods does not allow us to find a closed-form solution for our problem. Therefore, while this would allow us to face a convex optimization problem, we would still have to rely on an optimizer such as SGD (or Adam) to actually find the unique solution. In this regard, we recall that using RKHS requires calculating the Gramian matrix, which has a complexity of $\mathcal{O}(N^2)$, where N is the number of data points. Therefore, even just evaluating the loss function in equation (20) has a complexity of $\mathcal{O}(N^2)$, for each training epoch. This cost is additional to the computation of gradients and the update of the parameters in the optimization step, therefore, we expect a much higher computational cost and less scalability of RKHS-based techniques compared to that of DNN. From a theoretical point of view, consistency results for RKHS-based techniques in the literature are only available for the Supervised Learning case and it is not clear if they would be easily adapted to our Unsupervised Learning setting.

To conclude, while many alternatives are present, many of them present criticalities such as higher computational cost and design challenges, that DNNs easily avoid.

A.2. Basis-based approach

In the experiments of section 4, we considered DNN architectures with 3 hidden layers. As we discussed above, *basis-based approaches* offer a valid alternative that has been effectively used in many fields and comes at a cheap computational cost. To implement these models one has to extract some fixed *nonlinear static features* (the basis) and combine them linearly into the output. From a theoretical point of view, we expect these methods to approximate well the solution: they may be seen as a particular case of the general theory we developed in section 2–3, where one hidden layer is hard-coded to match the specific form of the basis. However, consistently with the literature, we expect such models to perform well only if the basis is carefully chosen and this requires some a priori knowledge of the sought solution; In our case, the functional shape of Φ_1 and Φ_2 . We confirm this expected downside by replicating the experiments of section 4 for the following methods[†]:

- (i) *RFNN*: We parametrize $\hat{\Phi}_1$ and $\hat{\Phi}_2$ with a *random feature neural network* whose base architecture is the same as the one in section 4, but where only the last layer is trained;
- (ii) *LCRF*: We parametrize $\hat{\Phi}_1$ and $\hat{\Phi}_2$ as Linear Combinations of **ReLU Functions**:

$$\Phi(x) = \sum_{i=0}^{100} \xi_i \text{ReLU}(x - K_i), \quad K_i := x_0 + (x_1 - x_0) \frac{i}{100}$$

where $[x_0, x_1]$ is the range of the random variable of interest X , and optimize only over the parameters ξ_i ;

- (iii) *LCGRBF*: We parametrize $\hat{\Phi}_1$ and $\hat{\Phi}_2$ as Linear Combinations of **Gaussian Radial Basis Functions**:

$$\Phi(x) = \sum_{i=0}^{100} \xi_i e^{-\lambda(x-K_i)^2}, \quad K_i := x_0 + (x_1 - x_0) \frac{i}{100}$$

with x_0, x_1 as above, and optimize only over the parameters ξ_i and λ .

[†] These examples were all suggested by an anonymous referee whom we thank.

Table A3. Average loss of the achieved training losses together with its standard deviation.

Case equation (23) – $-Beta(2, 5)$		
	Avg. Loss	Std. Loss
RFNN	0.3145099	$4.65897 \cdot 10^{-4}$
LCRF	0.3142185	$9.22487 \cdot 10^{-5}$
LCGRBF	0.3152364	$5.41497 \cdot 10^{-6}$
DNN-ReLu	0.3132883	$1.25900 \cdot 10^{-6}$

Since the loss function for these models is convex and admits a unique global minimum, we used Gradient Descent which is guaranteed to converge to such an optimum. This aspect is however not particularly decisive since, in all our experiments, we never experienced convergence problems for the general DNN algorithm. The main advantage is actually the interpretability of the solution: The form of the solution is fixed and only the coefficients of the linear combination are optimized.

For all models, we initialize the learning rate at $\eta = 0.1$, while the parameter λ for the LCGRBF model is equal to 1.0. To make the convergence even more stable, we used the *ReduceLRonPlateau* scheduler for the learning rate, with `patience` equal to 1000 and `threshold` equal to 10^{-6} , while all other parameters are as per default specification. Finally, all experiments have been run 3 times for a number of `epochs` equal to 300.

We first compare these three methods with each other and then we compare them to the ones based on DNNs from section 4. We present the results for the heterogeneous case of section 4.4 and for the entropic case of section 4.2 using $X \sim -Beta(2, 5)$. Analogous conclusions can be found for the other distributions.

Heterogeneous agents. Figure A1 shows the comparison of the loss functions of the three methods described above in the framework of section 4.4, case (23). To visualize this better, the right of such a figure shows the comparison toward the end of the training. The best performance is achieved by the LCRF model, followed by the RFNN model, and finally by the LCGRBF model; See also table A3, where we report the average and standard deviation of the loss at convergence. The left of figure A2 shows instead the comparison of the loss functions of the LCRF model against the DNN-based models. Once again, we provide a zoom towards the end of training on the right of such a figure. Together with the comparison of table A3 we confirm that a DNN with ReLU activations provides a better approximation of the solution, both in terms of average loss and in terms of the uncertainty of the estimate. Note that, although sub-optimal, LCRF achieves a loss level that is comparable with that of DNN, however, the variance is much larger.

Entropic case. While in the above experiment, the performance of the basis-based models was satisfying, in the entropic case this does not hold true. As above, we first identify the best of such methods by comparing their loss functions (see figure A3) and the average relative errors between the losses and the theoretical infimum, their standard deviation, and the average L^2 error between $\hat{\Phi}_1$ and Φ_1 (see table A4). In this case, RFNN exhibited the best performance. We next compare it against all the DNN-based ones. The striking result of figure A3 shows the limitations of basis-based models which underperformed all the DNN-based models: The average relative error with respect to the theoretical infimum achieved by RFNN is two orders of magnitude worse than the ReLU-DNN model (see table A4).

Discussion. The fact that deep networks outperform basis-based methods is not surprising as they are intrinsically more powerful: instead of *fixing* some static features, the algorithm will *learn* such features from the data. On the other hand, the convexity of the loss

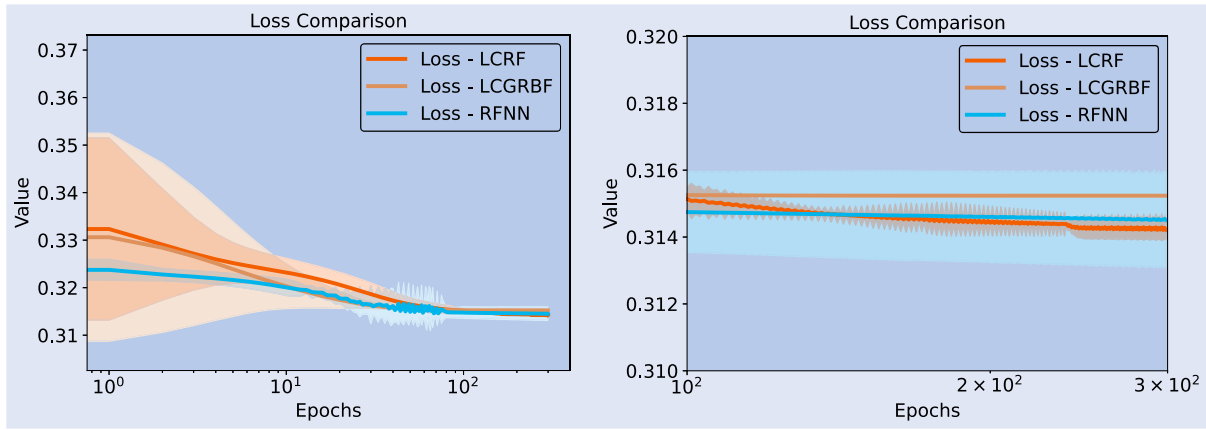


Figure A1. Heterogeneous agents – case equation (23) – beta distribution – we train all models over 3 trials and plot the average loss along with the ± 3 standard deviations. On the left, we compare the *basis-based models* among each other, and on the right, we provide a zoom towards the end of training.

function for basis-based methods allows Gradient Descent to find the optimal solution faster than Adam does for the DNNs. We conclude that DNN-based models are more flexible and achieve better performance on a greater variety of different setups. Nevertheless, both

are valid alternatives. The choice ultimately boils down to assessing the trade-off between accuracy of the solution, for which the DNN approach is better, and interpretability of the solution, for which the basis-based approach is better.

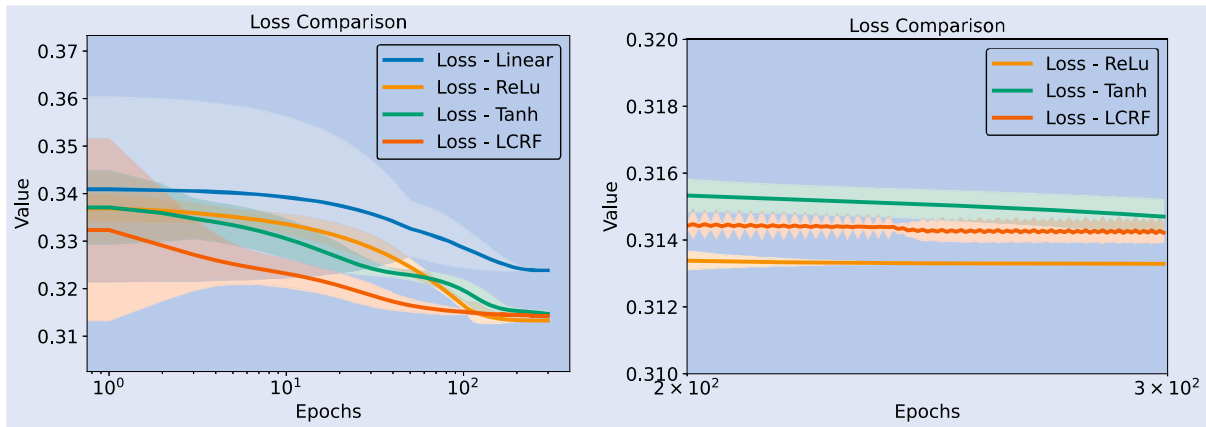


Figure A2. Heterogeneous agents – case equation (23) – beta distribution – we train all models over 3 trials and plot the average loss along with the ± 3 standard deviations. On the left, we compare the *DNN-based models* among each other together with the LCRF model. On the right, we provide a zoom towards the end of training.

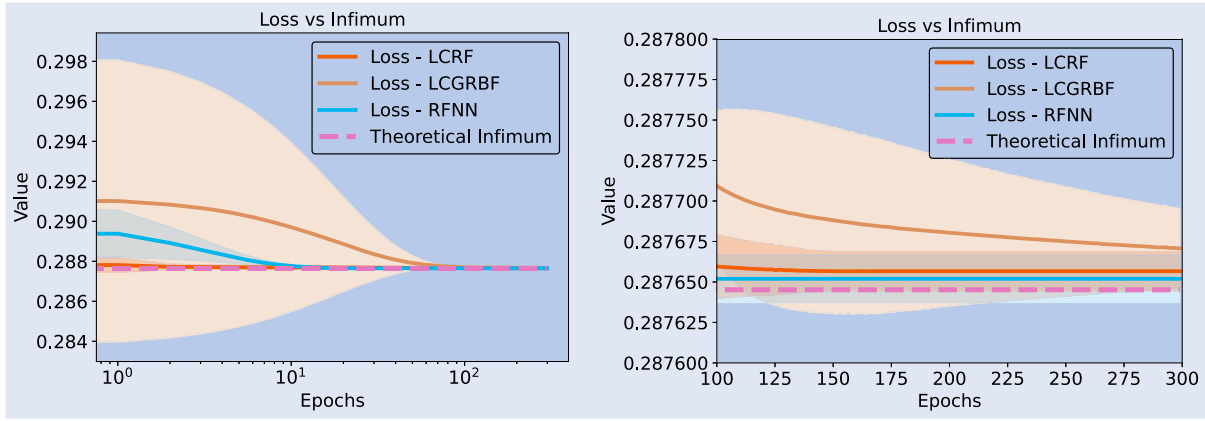


Figure A3. Entropic case equation (21) – beta distribution – we train all models over 3 trials and plot the average loss along with the ± 3 standard deviations. On the left, we compare the *basis-based models* among each other, and on the right, we provide a zoom towards the end of training.

Table A4. Average relative errors between the losses and the theoretical infimum, their standard deviation, and the average L^2 error between $\hat{\Phi}_1$ and Φ_1 .

Entropic case – $Beta(2, 5)$ – Infimum = 0.2876			
	Avg. Rel. Error	Std. Rel. Error	Avg. L^2 Error $\hat{\Phi}_1$
RFNN	$\%2.359 \cdot 10^{-3}$	$\%1.698 \cdot 10^{-3}$	$1.177 \cdot 10^{-3}$
LCRF	$\%3.999 \cdot 10^{-3}$	$\%1.394 \cdot 10^{-3}$	$4.904 \cdot 10^{-3}$
LCGRBF	$\%8.889 \cdot 10^{-3}$	$\%2.849 \cdot 10^{-3}$	$8.763 \cdot 10^{-2}$
DNN-ReLu	$\%7.943 \cdot 10^{-5}$	$\%5.504 \cdot 10^{-5}$	$3.146 \cdot 10^{-4}$

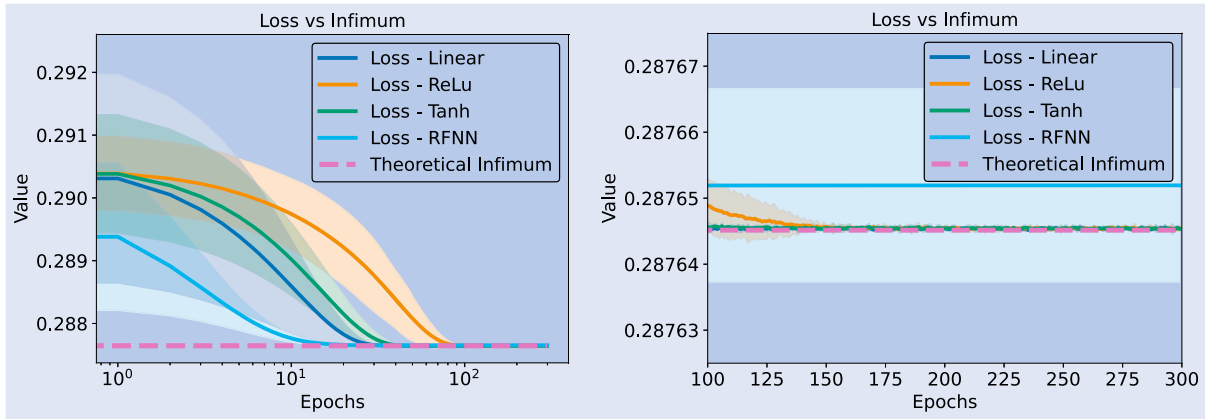


Figure A4. Entropic case equation (21) – beta distribution – we train all models over 3 trials and plot the average loss along with the ± 3 standard deviations. On the left, we compare the *DNN-based models* among each other together with the RFNN model. On the right, we provide a zoom towards the end of training.