Nendel, Max; Sgarabottolo, Alessandro

Working Paper

A parametric approach to the estimation of convex risk functionals based on Wasserstein distance

Center for Mathematical Economics Working Papers, No. 724

Provided in Cooperation with:
Center for Mathematical Economics (IMW), Bielefeld University

Suggested Citation: Nendel, Max; Sgarabottolo, Alessandro (2024) : A parametric approach to the estimation of convex risk functionals based on Wasserstein distance, Center for Mathematical Economics Working Papers, No. 724, Bielefeld University, Center for Mathematical Economics (IMW), Bielefeld,
https://nbn-resolving.de/urn:nbn:de:0070-pub-30052909

This Version is available at:
https://hdl.handle.net/10419/324290

Center for Mathematical Economics
Working Papers

724

August 2024

A parametric approach to the estimation of convex risk functionals based on Wasserstein distance

Max Nendel and Alessandro Sgarabottolo

Center for Mathematical Economics (IMW)
Bielefeld University
Universitätsstraße 25
D-33615 Bielefeld · Germany

e-mail: imw@uni-bielefeld.de
uni-bielefeld.de/zwe/imw/research/working-papers
ISSN: 0931-6558

A PARAMETRIC APPROACH TO THE ESTIMATION OF CONVEX RISK FUNCTIONALS BASED ON WASSERSTEIN DISTANCE

MAX NENDEL AND ALESSANDRO SGARABOTTOLO

Abstract. In this paper, we explore a static setting for the assessment of risk in the context of mathematical finance and actuarial science that takes into account model uncertainty in the distribution of a possibly infinite-dimensional risk factor. We study convex risk functionals that incorporate a safety margin with respect to nonparametric uncertainty by penalizing perturbations from a given baseline model using Wasserstein distance. We investigate to which extent this form of probabilistic imprecision can be approximated by restricting to a parametric family of models. The particular form of the parametrization allows to develop numerical methods based on neural networks, which give both the value of the risk functional and the worst-case perturbation of the reference measure. Moreover, we consider additional constraints on the perturbations, namely, mean and martingale constraints. We show that, in both cases, under suitable conditions on the loss function, it is still possible to estimate the risk functional by passing to a parametric family of perturbed models, which again allows for numerical approximations via neural networks.

Key words: Risk measure, model uncertainty, Wasserstein distance, martingale optimal transport, parametric estimation, neural network, measurable direction of steepest ascent

AMS 2020 Subject Classification: Primary 62G05; 90C31; Secondary 41A60; 68T07; 91G70

1. Introduction

In this article, we study a class of convex risk functionals which arises naturally in the context of mathematical finance and actuarial science, when dealing with expected values for a risk factor whose distribution is not perfectly known. Given a random variable Y on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in a separable Hilbert space H endowed with its Borel σ-algebra $\mathcal{B}(H)$, one is usually interested in expressions of the form

$$\mathbb{E}_{\mathbb{P}}\big[f(Y)\big] = \int_H f(y)\,\mu(\mathrm{d}y),$$

where $f\colon H \to \mathbb{R}$ is a, say, continuous loss or payoff function and $\mu = \mathbb{P} \circ Y^{-1}$ is the distribution of Y, i.e., a probability measure on $\mathcal{B}(H)$.

Obtaining precise knowledge of the distribution μ using estimation procedures is therefore of central importance in applications, and lies at the heart of statistics. In practice, however, one often has to deal with statistical imperfections, e.g., a lack of data or information about the dependence structure between single coordinates of Y, leading to a so-called *model calibration error* or *model specification error*, respectively. Therefore, in many situations, precise knowledge of the underlying distribution μ of Y may not be at hand, and only a rough estimate or an expert opinion suggesting a particular form of

Date: August 13, 2024.

The authors thank Daniel Bartl, Jonas Blessing, Stephan Eckstein, Michael Kupper, and Riccardo Mattivi for valuable comments and discussions related to this work. Financial support through the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1283/2 2021 – 317210226 is gratefully acknowledged.

reference distribution μ may be available. This is a special instance of *model uncertainty* appearing, for example, in the context of catastrophic risk in reinsurance or default risk within large credit portfolios in banking.

In the economic literature, model uncertainty is also referred to as *Knightian uncertainty*, and a standard way to deal with it is to look at worst case losses among a set of plausible probability distributions. In our study, we follow this approach, and estimate worst case losses over the set $\mathcal{P}_p$ of all Borel probability measures on H with finite moment of order $p \in (1,\infty)$, weighting the different measures via a penalization term depending on the *p-Wasserstein distance* from a reference model μ (which is assumed to have finite moment of order p as well). This leads to an expression of the form

$$\mathcal{I}f := \sup_{\nu\in\mathcal{P}_p}\bigg(\int_H f(z)\,\nu(\mathrm{d}z) - \varphi\big(\mathcal{W}_p(\mu,\nu)\big)\bigg). \tag{1.1}$$

Here, the penalty function $\varphi\colon [0,\infty)\to[0,\infty]$, which is assumed to be nondecreasing with $\varphi(0)=0$, reflects a degree of confidence in the reference measure, which could, for example, be related to the availability of data for the estimation of μ. The value $\varphi(a)=\infty$ for some $a>0$ corresponds to a rejection of every model ν with Wasserstein distance $\mathcal{W}_p(\mu,\nu)\geq a$, and the limit case $\varphi=\infty\cdot\mathbb{1}_{(0,\infty)}$ resembles perfect confidence in the measure μ.

Functionals of the form (1.1) belong to the class of convex risk measures and, under suitable conditions on the penalty function φ, to the class of coherent risk measures, cf. [2] and [18]. Moreover, they are widely studied in the context of distributionally robust optimization problems, see, for example, [5, 20, 26, 31, 33, 34], where the authors usually consider an additional optimization procedure, leading to an inf-sup-formulation.

A standard approach to tackle the infinite-dimensional optimization related to (1.1) is to look for a suitable dual formulation, for example, by transforming the primal problem into a superhedging problem. For example, in [5], the authors transform a class of robust optimized certainty equivalents (OCEs) into a one-dimensional optimization that leads to an explicit correction term. In general, however, this approach leads to a nested optimization problem, which can be numerically challenging.

We therefore look at this problem from a similar yet different angle, and aim to identify a parametric version of the functional (1.1) together with suitable optimizing directions. This idea is merely related to the paradigm of looking for extreme points in Wasserstein balls; a topic that has been explored in detail in the case where the reference measure is an empirical distribution (uniform over the samples) or, more generally, a convex combination of Dirac measures. In [33], it is shown that extreme points of Wasserstein balls centered in a measure supported on at most n points are supported on at most $n+3$ points. The paper [29] refines this result, showing that these extreme distributions are in fact supported on $n+2$ points. Finally, in [26], the authors show that, under stronger assumptions on the loss function, the infinite-dimensional optimization problem can be solved via a convex shifting of the support points and that the optimizing distribution is supported on the same number of points as the reference measure. In the case, where μ is a convex combination of Dirac measures, we get a similar result, but in a different fashion, cf. Section 3.1.

The key idea of our approach is to look for a parametric version of the functional (1.1) in terms of a first order approximation as the level of uncertainty related to the penalty function φ tends to zero. More precisely, we introduce a scaling parameter $h>0$, and substitute φ with the rescaled version $\varphi_h := h\varphi(\cdot/h)$, which allows to control the level of uncertainty in terms of h. We then consider the operator $\mathcal{I}(h)$, given by

$$\mathcal{I}(h)f := \sup_{\nu\in\mathcal{P}_p}\bigg(\int_H f(z)\,\nu(\mathrm{d}z) - \varphi_h\big(\mathcal{W}_p(\mu,\nu)\big)\bigg) \tag{1.2}$$

for a suitable class of continuous payoff function $f\colon H \to \mathbb{R}$, and look for a parametric version $\mathcal{I}_\Theta(h)$ that asymptotically coincides with $\mathcal{I}(h)$ up to a first order error as h tends to zero. To that end, we consider a *parameter set* Θ consisting of vector fields $\theta\colon H \to H$, which are p-integrable with respect to the reference measure μ. For each *parameter* $\theta \in \Theta$, we consider a probability measure μ_θ, which is a shifted version of μ, i.e.,

$$\int_H f(z)\,\mu_\theta(\mathrm{d}z) = \int_H f\big(y + \theta(y)\big)\,\mu(\mathrm{d}y), \tag{1.3}$$

and define the parametric version $\mathcal{I}_\Theta(h)$ of $\mathcal{I}(h)$, for $h > 0$, by

$$\mathcal{I}_\Theta(h)f := \sup_{\theta\in\Theta}\left(\int_H f(z)\,\mu_\theta(\mathrm{d}z) - \varphi_h\big(\|\theta\|_{L_p(\mu;H)}\big)\right).$$

In Theorem 2.7, we provide conditions on the parameter set Θ and the function f, ensuring that

$$\lim_{h\downarrow 0} \frac{\mathcal{I}(h)f - \mathcal{I}_\Theta(h)f}{h} = 0.$$

In Theorem 2.5, we compute a safety margin for asymptotically small $h > 0$, and show that an asymptotically optimal parameter θ can be found by looking at directions of steepest ascent for f.

A crucial step in the direction of studying the asymptotic behaviour of optimization problems of the form (1.2) was done in [3], see also [7] for its extension to a multi-period setting using adapted Wasserstein distance and [6, 19] for dynamic versions without an additional optimization. In [3], the authors study sensitivities for various forms of robust optimization problems over p-Wasserstein balls employing an elegant computational approach. In the proof of Theorem 2.5, we build on these methods, allowing for more general penalty functions φ and, at the same time, using weaker differentiability assumptions on f. In particular, we can drop the assumption of differentiability (μ-a.e.), introducing the concept of a *measurable direction of steepest ascent* as a generalization of the gradient, cf. Definition 2.3. This way, we can overcome differentiability issues related to the function f for reference measures μ, which are not regular. Most functions of interest in financial applications, e.g., call options, have a measurable direction of steepest ascent, so that our results can be applied without any restrictions on the reference measure μ, cf. Remark 2.6 b) for a more thorough discussion.

Following the lead of [3], in Section 2.2, we study an additional mean constraint in the optimization (1.2). This constraint enters naturally when dealing with risk-neutral pricing, where the mean of the underlying is assumed to be known (e.g., by standard non-arbitrage arguments). Thanks to our parametric description of the risk functional, cf. Theorem 2.8, we can show that uncertainty in the return of a financial position can be replicated by means of self-financing portfolios of call options or digital options, cf. Section 4.5.

In Section 2.3, we impose a so-called martingale constraint on the functional (1.2). That is, we restrict the optimization to a set of probability measures, which are given in terms of a martingale perturbation of the reference measure or, in different words, measures that admit a martingale coupling with μ. This setup is closely intertwined with the topics of martingale optimal transport (MOT) and model-free pricing in mathematical finance. For example, in [8], the authors study robust superhedging problems based on MOT, whereas [9] provides a precise description and fine properties of the optimal transport plan under a martingale constraint. We also refer to [28] for a class of robust estimators for superhedging prices based on martingale measures which are, up to a small perturbation in Wasserstein distance, equivalent to an empirical distribution.

In Theorem 2.9, we provide an asymptotic parametrization of the constrained version of (1.2) through a suitable *randomization* of the reference measure. Given a direction $\theta \in \Theta$, we include an additional coin flip in (1.3), which is independent of μ and determines the sign of θ. This leads to a parametric description in terms of measures $\mu_\theta^{\mathrm{Mart}} \in \mathcal{P}_p$, given by

$$\int_H f(z)\,\mu_\theta^{\mathrm{Mart}}(\mathrm{d}z) = \int_H \int_{\mathbb{R}} f\big(y + s\theta(y)\big)\,B_{\mathrm{sym}}(\mathrm{d}s)\,\mu(\mathrm{d}y), \tag{1.4}$$

where B_{sym} is the symmetric Bernoulli distribution on $\mathbb{R}$ with equal probabilities, i.e.,

$$B_{\mathrm{sym}}\big(\{-1\}\big) = B_{\mathrm{sym}}\big(\{1\}\big) = \frac{1}{2}.$$

In principle, B_{sym} could be replaced by any distribution on $\mathbb{R}$ with mean zero. However, the symmetric Bernoulli distribution has the identifying property that, apart from having mean zero, $\int_{\mathbb{R}} |s|^\alpha\,B_{\mathrm{sym}}(\mathrm{d}s) = 1$ for all $\alpha \in (0,\infty)$; a property that is of central importance for the asymptotic parametrization in this framework.

Apart from this, the symmetric Bernoulli distribution is fundamentally connected to the Brownian motion via Donsker's theorem. Having in mind that, in a Brownian filtration, all martingales can be represented as stochastic integrals with respect to the Brownian motion, our parametrization via measures of the form (1.4) can be seen as a microscopic version of a martingale representation theorem in a completely different and model-free setting. Moreover, the measure $\mu_\theta^{\mathrm{Mart}}$ can be represented via the explicit formula

$$\int_H f(z)\,\mu_\theta^{\mathrm{Mart}}(\mathrm{d}z) = \int_H \frac{f\big(y+\theta(y)\big) + f\big(y-\theta(y)\big)}{2}\,\mu(\mathrm{d}y), \tag{1.5}$$

which, after substracting $\int_H f(y)\,\mu(\mathrm{d}y)$, leads to a finite difference approximation of the second derivative of f in the direction θ using central differences. This links $\mu_\theta^{\mathrm{Mart}}$ also from an analytic perspective to a Brownian motion through its infinitesimal generator.

In view of numerical approximations, our construction of parametric versions of convex risk functionals of the form (1.2) leads to a description of asymptotically relevant models for the optimization in terms of Monge transports (p-integrable vector fields), which, in turn, suggests a numerical investigation of the risk functional in the spirit of [16], see also [17]. In Section 3.2, we develop a numerical scheme based on neural networks. Previous works on this topic provide approximations from above based on duality results, whereas our approach leads to an approximation from below, based on the restriction to a parametric family of models. As a byproduct, we also approximate the optimizing measure through the Monge transport generating it. Observing that, for small values of uncertainty, the optimal transport plan depends on the reference measure only through a multiplicative factor, we draw a connection to *transfer learning*, and discuss the robustness of our approach with respect to deviations in the reference measure.

In machine learning, transfer learning usually consists of taking a neural network trained on a previous dataset and training only its last layer on a new dataset, which is often significantly smaller than the first one, see [13] for references to seminal works on this topic. The idea is that if the two datasets share some common features, the part of the network that is inherited from the first training is able to extract most of these features, while the training of the last layer learns from specific characteristics of the new dataset. In our framework, we can explicitly distinguish between the common feature that can be extracted from the first training, namely the optimizing vector field, and the feature that is specific to each reference measure, i.e., the rescaling factor. In particular, once the approximation is performed on one measure, we can transfer it to a different measure at the price of a one-dimensional optimization, see Section 4.4 for further details.

The paper is organized as follows. In Section 2, we introduce the setup and state the main results on parametric versions of risk functions of the form (1.2). Section 3 provides numerical methods for the approximation of the parametric risk functional $\mathcal{I}_\Theta(h)$. We distinguish between two situations leading to different numerical schemes. One is based on a reduction to a finite-dimensional optimization, cf. Section 3.1, while the other uses an approximation via neural networks, cf. Section 3.2. In Section 4, we discuss applications, both, of our theoretical and numerical findings by means of examples from finance and insurance. Section 5 contains the proofs of the main theorems and, in the Appendix A, we provide a simple approximation result that helps to reduce the set of parameters Θ.

2. Setup and main results

In this section, we introduce the class of convex risk functionals that (together with additional constraints) form the center of our study, and we state our main results concerning their parametric estimation.

Throughout, let $p \in (1,\infty)$, $q := \frac{p}{p-1} \in (1,\infty)$ be the conjugate exponent of p, and $(H, \langle\cdot\,,\cdot\rangle)$ be a separable Hilbert space. As usual, we endow H with its canonical norm $\|\cdot\| := \sqrt{\langle\cdot\,,\cdot\rangle}$, and identify the topological dual space H' of H with H itself. We denote the set of all probability measures ν on the Borel σ-algebra $\mathcal{B}(H)$ of H with

$$|\nu|_p := \left(\int_H \|y\|^p \,\nu(\mathrm{d}y)\right)^{1/p} < \infty$$

by $\mathcal{P}_p = \mathcal{P}_p(H)$.

In the following, we consider a fixed probability measure $\mu \in \mathcal{P}_p$, which we will refer to as the *reference measure* or *baseline model*, and a *penalty function* $\varphi\colon [0,\infty) \to [0,\infty]$, which is assumed to be nondecreasing with $\varphi(0) = 0$.

For $\nu \in \mathcal{P}_p$, we denote the *p-Wasserstein distance* between the reference measure μ and ν by

$$\mathcal{W}_p(\mu,\nu) = \left(\inf_{\pi\in\mathrm{Cpl}(\mu,\nu)} \int_{H\times H} \|y-z\|^p \,\pi(\mathrm{d}y,\mathrm{d}z)\right)^{1/p}, \tag{2.1}$$

where $\mathrm{Cpl}(\mu,\nu)$ is the set of all couplings between μ and ν, i.e., the set of probability measures on the Borel σ-algebra $\mathcal{B}(H\times H)$ of the product space $H\times H$ with first and second marginal μ and ν, respectively. We refer to [1] and [32] for a detailed discussion on Wasserstein distances and, more generally, the topic of optimal transport. Here, we only recall that, for all $\nu \in \mathcal{P}_p$, there exists an optimal coupling $\pi^* \in \mathrm{Cpl}(\mu,\nu)$ that attains the infimum in (2.1), i.e.,

$$\mathcal{W}_p(\mu,\nu) = \left(\int_{H\times H} \|y-z\|^p \,\pi^*(\mathrm{d}y,\mathrm{d}z)\right)^{1/p}.$$

For $\varrho \in [1,\infty)$, we denote the space of all (μ-equivalence classes of) measurable functions $f\colon H \to \mathbb{R}$ with

$$\|f\|_{L_\varrho(\mu)} := \left(\int_H |f(y)|^\varrho \,\mu(\mathrm{d}y)\right)^{1/\varrho} < \infty$$

by $L_\varrho(\mu)$. In a similar fashion, $L_\varrho(\mu;H)$ denotes the space of all (μ-equivalence classes of) measurable functions $g\colon H \to H$ with

$$\|g\|_{L_\varrho(\mu;H)} := \left(\int_H \|g(y)\|^\varrho \,\mu(\mathrm{d}y)\right)^{1/\varrho} < \infty.$$

Recall that, by assumption, H is a separable Hilbert space, so that the image $g(H) \subset H$ of g is automatically separable.

2.1. **The unconstrained case.** Throughout this subsection, we assume that

$$\liminf_{v\to\infty} \frac{\varphi(v)}{v^p} > 0. \tag{2.2}$$

As a consequence, the *convex conjugate* $\varphi^*\colon [0,\infty) \to [0,\infty)$, given by

$$\varphi^*(u) := \sup_{v\geq 0} \Big(uv - \varphi(v) \Big) \quad \text{for all } u \in [0,\infty), \tag{2.3}$$

is well-defined, convex, and continuous with $\varphi^*(0) = 0$.

We denote by Lip_p the space of all functions $f\colon H \to \mathbb{R}$, for which there exists a constant $L_f \geq 0$ such that

$$|f(x_1) - f(x_2)| \leq L_f \big(1 + \max\{\|x_1\|, \|x_2\|\}\big)^{p-1} \|x_1 - x_2\| \quad \text{for all } x_1, x_2 \in H. \tag{2.4}$$

Let $f \in \mathrm{Lip}_p$. Choosing $x_1 = x \in H$ and $x_2 = 0$ in (2.4), we find that

$$|f(x)| \leq (|f(0)| + L_f)\big(1 + \|x\|\big)^p \quad \text{for all } x \in H. \tag{2.5}$$

Moreover, the *local Lipschitz constant* $|f|_{\mathrm{Lip}}\colon H \to \mathbb{R}$ of f, given by

$$|f|_{\mathrm{Lip}}(x) := \limsup_{h\downarrow 0} \sup_{\|u\|\in\{0,1\}} \frac{f(x+hu) - f(x)}{h} \quad \text{for all } x \in H, \tag{2.6}$$

is measurable due to the continuity of f, and satisfies

$$0 \leq |f|_{\mathrm{Lip}}(x) = \limsup_{h\downarrow 0} \sup_{\|u\|\in\{0,1\}} \frac{f(x+hu) - f(x)}{h} \leq L_f \big(1 + \|x\|\big)^{p-1} \quad \text{for all } x \in H. \tag{2.7}$$

For $h > 0$, we consider the operator $\mathcal{I}(h)$ that associates to every function $f \in \mathrm{Lip}_p$ the quantity

$$\mathcal{I}(h)f := \sup_{\nu\in\mathcal{P}_p} \left(\int_H f(z)\, \nu(\mathrm{d}z) - \varphi_h(\mathcal{W}_p(\mu,\nu)) \right) \in (-\infty, \infty], \tag{2.8}$$

where $\varphi_h(v) := h\varphi\left(\frac{v}{h}\right)$ for all $v \in [0,\infty)$. The fact that $f \in \mathrm{Lip}_p$ ensures that, at least for sufficiently small $h > 0$,

$$\mathcal{I}(h)f < \infty.$$

Remark 2.1. Considering the extreme case, where $\varphi = \infty \cdot \mathbb{1}_{(a,\infty)}$ for some $a \geq 0$, the operator $\mathcal{I}(h)$ simplifies to

$$\mathcal{I}(h)f := \sup_{\mathcal{W}_p(\mu,\nu)\leq ah} \int_H f(z)\, \nu(\mathrm{d}z) \quad \text{for all } f \in \mathrm{Lip}_p,$$

which corresponds to the consideration of worst case scenarios over Wasserstein balls with radius ah for $h > 0$. Thus, even in the simplest case, the operator $\mathcal{I}(h)$ leads to a rather complex optimization problem, which involves the determination of a Wasserstein ball. Moreover, due to the geometric structure (more precisely, the lack of finitely many extreme points) of Wasserstein balls, it is, in general, not trivial to characterize optimizers of $\mathcal{I}(h)$ and to analyze their dependence on the loss function f and the reference model μ. One aspect that triggers this problematic is the appearance of the Wasserstein distance in the penalty term. On the other hand, also the lack of concrete structure of the set $\mathcal{P}_p$ makes the problem intractable. In the following, we thus introduce a simplified version of the operator $\mathcal{I}(h)$, which depends on the reference measure in a parametric way, circumvents the use of the Wasserstein distance, and can, in many cases, be computed in an efficient manner.

Throughout, we consider a *parameter set* $\Theta \subseteq L_p(\mu; H)$. For every *parameter* $\theta \in \Theta$, we define a probability measure $\mu_\theta \in \mathcal{P}_p$ via

$$\int_H f(z)\, \mu_\theta(\mathrm{d}z) := \int_H f\big(y + \theta(y)\big)\, \mu(\mathrm{d}y) \quad \text{for all } f \in \mathrm{Lip}_p.$$

Note that, for all $\theta \in \Theta$, μ_θ is in fact an element of $\mathcal{P}_p$ since

$$|\mu_\theta|_p \leq |\mu|_p + \|\theta\|_{L_p(\mu;H)}.$$

For $h > 0$, we define the *parametric version* $\mathcal{I}_\Theta$ of $\mathcal{I}$ by

$$\mathcal{I}_\Theta(h) f := \sup_{\theta \in \Theta} \left(\int_H f(z)\, \mu_\theta(\mathrm{d}z) - \varphi_h\big(\|\theta\|_{L_p(\mu;H)}\big) \right) \quad \text{for all } f \in \mathrm{Lip}_p.$$

By definition, $\mathcal{I}_\Theta(h) f \leq \mathcal{I}(h) f$ for all $h > 0$ and $f \in \mathrm{Lip}_p$. In fact, as already observed, $\{\mu_\theta\}_{\theta \in \Theta} \subset \mathcal{P}_p$, and to bound the penalization term we choose a perfectly correlated coupling and get that $\mathcal{W}_p(\mu, \mu_\theta) \leq \|\theta\|_{L_p(\mu;H)}$. In conclusion, for $h > 0$, the operator $\mathcal{I}_\Theta(h)$ restricts the family of measures appearing in the optimization problem (2.8) and, at the same time, has a more explicit penalization term.

Remark 2.2. Assume that μ is nonatomic. Then, by a standard construction, there exists a measurable function $f\colon H \to [0,1]$ that is uniformly distributed under μ. In fact, since μ is nonatomic, for every set $B \in \mathcal{B}(H)$, there exists a set $A \in \mathcal{B}(H)$ with $A \subset B$ and $\mu(A) = \frac{1}{2}\mu(B)$, so that there exists an i.i.d. sequence $(f_n)_{n \in \mathbb{N}}$ of $\{0,1\}$-valued random variables with

$$\mu(f_n = 0) = \frac{1}{2} = \mu(f_n = 1) \quad \text{for all } n \in \mathbb{N}.$$

Choosing $f = \sum_{n=1}^\infty 2^{-n} f_n$, we find that $f\colon H \to [0,1]$ is uniformly distributed. On the other hand, since H is a separable Hilbert space, it is Borel isomorphic to $\mathbb{R}$, cf. [15, Theorem 13.1.1]. Therefore, every measure $\nu \in \mathcal{P}_p$ can be obtained as a push forward measure $\mu \circ g^{-1}$ for a measurable function $g\colon H \to H$. Choosing, $\theta(x) := g(x) - x$ for all $x \in H$, it follows that $\nu = \mu_\theta$. Since both μ and ν are elements of $\mathcal{P}_p$, we obtain that $\theta \in L_p(\mu; H)$. Hence, for every $\nu \in \mathcal{P}_p$, there exists some $\theta \in L_p(\mu; H)$ with $\nu = \mu_\theta$. On the other hand, if μ is even assumed to be regular, there exists some $\theta \in L_p(\mu; H)$ with $\nu = \mu_\theta$ and

$$\mathcal{W}_p(\mu, \nu) = \|\theta\|_{L_p(\mu;H)}, \tag{2.9}$$

see, e.g., [1, Theorem 6.2.10]. In conclusion, if μ is regular, then

$$\mathcal{I}(h) f = \mathcal{I}_{L_p(\mu;H)}(h) f \quad \text{for all } f \in \mathrm{Lip}_p \text{ and all } h > 0. \tag{2.10}$$

However, one should keep in mind the following two observations:

1) Since the starting point of our study is a somewhat imperfect knowledge of the reference measure μ, it cannot be assumed to be nonatomic. In particular, if μ is estimated from data as a, say, empirical distribution, it will be of the form $\mu = \sum_{i=1}^n \alpha_i \delta_{x_i}$ with $n \in \mathbb{N}$, $x_1, \ldots, x_n \in H$, and $\alpha_1, \ldots, \alpha_n \in [0,1]$ with $\sum_{i=1}^n \alpha_i = 1$, where δ_x denotes the Dirac measure with barycenter $x \in H$.
2) Even if μ is regular, and we have (2.10) at hand, it neither reveals a particular shape nor qualitative properties of the optimizer. Moreover, a priori, the optimizer may heavily depend on the reference measure μ, which is assumed not to be known precisely. Therefore, the computation might exhibit huge instabilities for small deviations from the reference measure.

We therefore aim towards a different direction, and try to come up with an asymptotic version of the equality (2.10) for infinitesimally small $h > 0$ and *all* reference measures μ. On the other hand, we are interested in understanding the structure of asymptotic optimizers of $\mathcal{I}(h)$ and their dependence on the reference measure μ for infinitesimally small levels of uncertainty $h > 0$.

Every element of Lip_p is locally Lipschitz continuous, and thus Gateaux differentiable outside a Gaussian null set, see [11, Theorem 5.11.1]. However, since we do not want to restrict our attention to regular measures, i.e., the ones that assign measure zero to every Gaussian null set, we make use of a similar yet slightly different notion of differentiability based on the idea that the gradient is the direction of steepest ascent. The following definition formalises this idea.

Definition 2.3. Let $f \in \mathrm{Lip}_p$. We say that a vector field $v\colon H \to H$ is a *measurable direction of steepest ascent* for f, if it is measurable, $\|v(x)\| \in \{0,1\}$, and

$$|f|_{\mathrm{Lip}}(x) = \lim_{h\downarrow 0} \frac{f\big(x + hv(x)\big) - f(x)}{h} \quad \text{for all } x \in H.$$

Remark 2.4. Let $f\colon H \to \mathbb{R}$ be Fréchet differentiable with continuous gradient $\nabla f\colon H \to H$ satisfying

$$L_f := \sup_{x\in H} \frac{\|\nabla f(x)\|}{(1+\|x\|)^{p-1}} < \infty.$$

Then,

$$\begin{aligned} \big|f(x_1) - f(x_2)\big| &\le \int_0^1 \big|\big\langle \nabla f\big(tx_1 + (1-t)x_2\big), x_1 - x_2\big\rangle\big|\,\mathrm{d}t \\ &\le L_f\big(1 + \max\{\|x_1\|, \|x_2\|\}\big)^{p-1}\|x_1 - x_2\| \quad \text{for all } x_1, x_2 \in H, \end{aligned}$$

so that $f \in \mathrm{Lip}_p$. In this case, a measurable direction of steepest ascent $v\colon H \to H$ is given by

$$v(x) := \begin{cases} \frac{\nabla f(x)}{\|\nabla f(x)\|}, & \text{if } \nabla f(x) \neq 0, \\ v(x) = 0, & \text{otherwise,} \end{cases} \quad \text{for all } x \in H, \tag{2.11}$$

and $|f|_{\mathrm{Lip}}(x) = \|\nabla f(x)\|$ for all $x \in H$.

The following theorem provides a first order approximation of the functional $\mathcal{I}(h)$ in terms of its parametric version $\mathcal{I}_\Theta(h)$, for every reference measure μ and infinitesimally small $h > 0$, under weak regularity assumptions on the function f. Additionally, it shows that an asymptotic optimizer can be computed explicitly and (almost) independently of the reference measure μ. The proof is delayed to Section 5.

Theorem 2.5. *Let $f \in \mathrm{Lip}_p$ with a measurable direction of steepest ascent $v\colon H \to H$. Define*

$$\theta(x) = v(x)\big(|f|_{\mathrm{Lip}}(x)\big)^{q-1} \quad \textit{for all } x \in H. \tag{2.12}$$

Then,

$$\lim_{h\downarrow 0} \frac{\mathcal{I}(h)f - \mathcal{I}_\Theta(h)f}{h} = 0,$$

whenever $\{\lambda\theta \,|\, \lambda \ge 0\} \subseteq \Theta$. Moreover,

$$\lim_{h\downarrow 0} \frac{\mathcal{I}(h)f - \mu f}{h} = \varphi^*\Big(\big\||f|_{\mathrm{Lip}}\big\|_{L_q(\mu)}\Big), \tag{2.13}$$

where $\mu f := \int_H f(y)\,\mu(\mathrm{d}y)$.

Remark 2.6. a) The previous result shows us that, when considering a fixed function in Lip_p with a measurable direction of steepest ascent, the optimization problem (2.8) asymptotically reduces to a one-dimensional optimization scheme as $h \downarrow 0$. Moreover, this scheme is independent of the reference measure μ. We exploit this property in Section 4.4 to develop a *transfer learning* technique, where we take the optimizing vector field of a given problem and search for optimizers with respect to different reference measures by simply rescaling the original vector field. This is quite remarkable in view of the concerns expressed in point 2) of Remark 2.2, since it indicates sort of a stability of the optimization problem $\mathcal{I}_\Theta(h)$ with respect to perturbations of the reference measure, at least for small $h > 0$. We further mention that the vector field θ, given by (2.12), satisfies

$$\sup_{x \in H} \frac{\|\theta(x)\|}{1 + \|x\|} < \infty.$$

b) The conditions of the previous theorem might, at first sight, seem a bit artificial, and, in view of Remark 2.4, a more natural formulation seems to be in terms of continuously differentiable functions $f\colon H \to \mathbb{R}$. However, for instance the payoff function of a European call option with *strike price* $K \in \mathbb{R}$ ($f(x) = (x - K)^+$) is not differentiable, since its derivative has a discontinuity at the strike price K. In [3], when computing the sensitivity of the robust optimization problem, the authors remark that the assumption of differentiability can be quite naturally relaxed when the measure μ is absolutely continuous with respect to the Lebesgue measure, assuming that the function f admits a weak derivative in the Sobolev sense, which is continuous μ-a.e. They also show how one can avoid the absolute continuity of μ with respect to the Lebesgue measure using a regularization procedure via convolution with normal distributions. Nevertheless, this solution does not solve the issue arising from f having a kink on a set which has strictly positive measure under μ. This is the case for a call option with strike K and a reference measure μ with $\mu(\{K\}) > 0$. In our framework, we can deal with this particular case observing that, if f is the profile of the call option with strike K,

$$|f|_{\mathrm{Lip}}(x) = \begin{cases} 0, & \text{for } x < K, \\ 1, & \text{for } x \geq K, \end{cases}$$

and a measurable direction of steepest ascent is given by $v(x) = \mathbb{1}_{[K,\infty)}(x)$ for all $x \in \mathbb{R}$.

If Θ is dense in $L_p(\mu; H)$ with $0 \in \Theta$, then

$$\mathcal{I}_\Theta(h)f = \mathcal{I}_{L_p(\mu;H)}(h)f \quad \text{for all } f \in \mathrm{Lip}_p. \tag{2.14}$$

A combination of Theorem 2.5 and this insight, cf. Lemma 5.2, is the essence of the proof of the following theorem, which is our second main result of this subsection.

Theorem 2.7. *Assume that Θ is dense in $L_p(\mu; H)$ with $0 \in \Theta$. Then,*

$$\lim_{h \downarrow 0} \frac{\mathcal{I}(h)f - \mathcal{I}_\Theta(h)f}{h} = 0$$

for every $f \in \mathrm{Lip}_p$ that has a measurable direction of steepest ascent.

In the Appendix A, we provide a simple sufficient condition for the density of Θ in $L_p(\mu; H)$, which allows to perform numerical computations of the operator $\mathcal{I}(h)$ in Section 3 via sufficiently small sets Θ and to make use of a neural network approximation of the operator $\mathcal{I}(h)$, for $h > 0$. In Section 4, we also discuss approximations, which are based on the use of portfolios of European call options or digital options.

2.2. **Mean constraint.** In this section, we consider a constrained version of the operator $\mathcal{I}(h)$, for $h > 0$, where the supremum is taken over all measures $\nu \in \mathcal{P}_p$ with

$$\int_H z\,\nu(\mathrm{d}z) = \int_H y\,\mu(\mathrm{d}y),$$

i.e., over all measures that have the same mean as the reference measure μ. This kind of constraint arises naturally in several contexts, for example, when pricing a financial instrument under the assumption that the expected value of the underlying is known.

Again, we assume that the penalty function satisfies (2.2). In the following, we restrict the operator $\mathcal{I}$ to measures, which satisfy the mean constraint, and consider, for $h > 0$ and $f \in \mathrm{Lip}_p$,

$$\mathcal{I}^{\mathrm{Mean}}(h)f := \sup_{\nu\in\mathcal{P}_p(\mu)} \left(\int_H f(z)\,\nu(\mathrm{d}z) - \varphi_h\big(\mathcal{W}_p(\mu,\nu)\big)\right),$$

where $\mathcal{P}_p(\mu)$ denotes the set of all $\nu \in \mathcal{P}_p$ with $\int_H z\,\nu(\mathrm{d}z) = \int_H y\,\mu(\mathrm{d}y)$.

We call C_p^1 the space of all continuously (Fréchet) differentiable functions $f\colon H \to \mathbb{R}$, for which there exists a constant $C \geq 0$ such that

$$\|\nabla f(x)\| \leq C(1 + \|x\|)^{p-1} \quad \text{for all } x \in H.$$

It is immediate to see that, for $\theta \in L^p(\mu; H)$, the measure μ_θ is an element of $\mathcal{P}_p(\mu)$ if and only if $\int_H \theta(y)\,\mu(\mathrm{d}y) = 0$.

Theorem 2.8. *Let $p = 2$ and $f \in C_p^1$. Then,*

$$\lim_{h\downarrow 0} \frac{\mathcal{I}^{\mathrm{Mean}}(h)f - \mathcal{I}_\Theta(h)f}{h} = 0,$$

where $\Theta := \{b\theta \,|\, b \geq 0\}$ with $\theta(x) := \nabla f(x) - \int_H \nabla f(y)\,\mu(\mathrm{d}y)$ for all $x \in H$. Moreover,

$$\lim_{h\downarrow 0} \frac{\mathcal{I}^{\mathrm{Mean}}(h)f - \mu f}{h} = \varphi^*\left(\left(\int_H \left\|\nabla f(y) - \int_H \nabla f(y)\,\mu(\mathrm{d}y)\right\|^2 \mu(\mathrm{d}y)\right)^{1/2}\right). \tag{2.15}$$

Equation (2.15) appears in a different form in [4]. There, the authors consider general linear constraints, whereas we focus on the mean constraint and provide an asymptotic formula for more general penalty functions. In that sense, Theorem 2.8 is a partial extension of [4, Theorem 7.7].

2.3. **Martingale constraint.** We now study an extension of the first order expansion given by Equation (2.13). In particular, we show how the inclusion of a martingale constraint improves the rate of convergence, leading to a *second order* parametrization of the operator $\mathcal{I}(h)$, for infinitesimally small $h > 0$. This parametrization will no longer be given in terms of Monge transports of the baseline measure μ, but by an actual randomization via an additional coin flip.

We start by introducing the notion of a *martingale coupling*, which is of fundamental importance for this section. Let $\nu \in \mathcal{P}_p$ and $\pi \in \mathrm{Cpl}(\mu,\nu)$. Then, π is called a *martingale coupling* between μ and ν if

$$\int_{H\times H} z \cdot \mathbb{1}_B(y)\,\pi(\mathrm{d}y,\mathrm{d}z) = \int_H y \cdot \mathbb{1}_B(y)\,\mu(\mathrm{d}y) \quad \text{for all } B \in \mathcal{B}(H).$$

The set of all $\nu \in \mathcal{P}_p$, for which there exists a martingale coupling between μ and ν, is denoted by $\mathcal{M}_p(\mu)$.

In the following, we use the notation $L(H)$ for the space of bounded linear operators $H \to H$ and we denote by $\|\cdot\| = \|\cdot\|_{L(H)}$ the usual operator norm on this space. Moreover,

given a bounded symmetric (or, equivalently, bounded self-adjoint) operator $S \in L(H)$, we define

$$|S|_{\mathrm{Max}} := \sup_{\|w\| \leq 1} \langle w, Sw \rangle. \tag{2.16}$$

Note that if S negative semidefinite, $|S|_{\mathrm{Max}} = 0$. Moreover, $|S|_{\mathrm{Max}} \leq \|S\|$ for all $S \in L(H)$, and the supremum in (2.16) can be taken over all $w \in H$ with $\|w\| \in \{0, 1\}$.

Throughout this section, we assume that the penalty function satisfies

$$\liminf_{v \to \infty} \frac{\varphi(v)}{v^{p/2}} > 0. \tag{2.17}$$

Again, this property implies that the convex conjugate $\varphi^* \colon [0, \infty) \to [0, \infty)$, given by (2.3), is well-defined, convex, and continuous with $\varphi^*(0) = 0$.

We call C_p^2 the space of all twice continuously (Fréchet) differentiable functions $f \colon H \to \mathbb{R}$, for which there exists a constant $C \geq 0$ such that

$$\|\nabla^2 f(x)\| \leq C(1 + \|x\|)^{p-2} \quad \text{for all } x \in H.$$

Note that C_p^2 is a subspace of C_p^1, in particular, it is contained in Lip_p.

We include the martingale constraint in our study by considering a slightly different version of the operator $\mathcal{I}(h)$ that we call $\mathcal{I}^{\mathrm{Mart}}(h)$ for $h > 0$. To that end, we define the *p-martingale Wasserstein distance* between μ and $\nu \in \mathcal{M}_p(\mu)$ as

$$\mathcal{W}_p^{\mathrm{Mart}}(\mu, \nu) := \left(\inf_{\pi \in \mathrm{Mart}(\mu,\nu)} \int_{H \times H} \|z - y\|^p \, \pi(\mathrm{d}y, \mathrm{d}z) \right)^{1/p},$$

where $\mathrm{Mart}(\mu, \nu)$ is the set of all martingale couplings between μ and ν. Then, for every function $f \in \mathrm{Lip}_p$ and every $h > 0$, we define

$$\mathcal{I}^{\mathrm{Mart}}(h) f := \sup_{\nu \in \mathcal{M}_p(\mu)} \left(\int_H f(z) \, \nu(dz) - \varphi_h\big(\mathcal{W}_p^{\mathrm{Mart}}(\mu, \nu)^2\big) \right), \tag{2.18}$$

where, again, $\varphi_h(v) = h\varphi(\frac{v}{h})$ for all $v \in [0, \infty)$. We point out that, apart from the additional martingale constraint in the Wasserstein distance, the penalty function is now applied to the squared (martingale) Wasserstein distance between μ and ν, whereas, in the previous sections, it was applied to the (usual) Wasserstein distance without an additional power.

Lemma 5.3 shows that the growth condition (2.17) allows us to restrict the supremum in (2.18) to a ball of radius $\sqrt{ah}$, for a suitable constant $a \geq 0$ and $h > 0$ sufficiently small, so that, in this case, uncertainty grows proportionally to the square root of h.

As in the unconstrained case, we are looking for a suitable parametric version of the operator $\mathcal{I}^{\mathrm{Mart}}(h)$ for $h > 0$. As in the previous subsections, let $\Theta \subset L_p(\mu; H)$. For $\theta \in \Theta$, we consider the probability measure $\mu_\theta^{\mathrm{Mart}} \in \mathcal{M}_p(\mu)$, which is given by

$$\int_H f(z) \, \mu_\theta^{\mathrm{Mart}}(\mathrm{d}z) := \int_H \frac{f\big(y + \theta(y)\big) + f\big(y - \theta(y)\big)}{2} \, \mu(\mathrm{d}y) \quad \text{for all } f \in \mathrm{Lip}_p. \tag{2.19}$$

Note that, by definition,

$$\int_H f(z) \, \mu_\theta^{\mathrm{Mart}}(\mathrm{d}z) = \int_H \int_{\mathbb{R}} f\big(y + s\theta(y)\big) \, B_{\mathrm{sym}}(\mathrm{d}s) \, \mu(\mathrm{d}y) \quad \text{for all } f \in \mathrm{Lip}_p, \tag{2.20}$$

where B_{sym} is the symmetric Bernoulli distribution on $\mathbb{R}$ with equal probabilities, i.e.,

$$B_{\mathrm{sym}}\big(\{-1\}\big) = B_{\mathrm{sym}}\big(\{1\}\big) = \frac{1}{2}.$$

For $h > 0$, we define the *parametric version* $\mathcal{I}_\Theta^{\mathrm{Mart}}(h)$ of $\mathcal{I}^{\mathrm{Mart}}(h)$ by

$$\mathcal{I}_\Theta^{\mathrm{Mart}}(h)f := \sup_{\theta\in\Theta}\left(\int_H f(z)\,\mu_\theta^{\mathrm{Mart}}(\mathrm{d}z) - \varphi_h\big(\|\theta\|^2_{L_p(\mu;H)}\big)\right) \quad \text{for all } f \in \mathrm{Lip}_p. \tag{2.21}$$

Choosing a perfectly correlated coupling, we find that

$$\begin{aligned}\mathcal{W}_p^{\mathrm{Mart}}\big(\mu,\mu_\theta^{\mathrm{Mart}}\big)^2 &\le \left(\int_H\int_{\mathbb{R}} \|s\theta(y)\|^p\, B_{\mathrm{sym}}(\mathrm{d}s)\,\mu(\mathrm{d}y)\right)^{2/p}\\ &= \left(\int_H \|\theta(y)\|^p\,\mu(\mathrm{d}y)\right)^{2/p} = \|\theta\|^2_{L_p(\mu;H)}.\end{aligned}$$

In particular, $\mathcal{I}_\Theta^{\mathrm{Mart}}(h)f \le \mathcal{I}^{\mathrm{Mart}}(h)f$ for all $h > 0$ and $f \in \mathrm{Lip}_p$.

We now state the convergence result for the operator $\mathcal{I}^{\mathrm{Mart}}(h)$ as $h \downarrow 0$. The proof is again relegated to Section 5.

Theorem 2.9. *Assume that Θ is dense in $L_p(\mu;H)$ with $0 \in \Theta$. Then,*

$$\lim_{h\downarrow 0}\frac{\mathcal{I}^{\mathrm{Mart}}(h)f - \mathcal{I}_\Theta^{\mathrm{Mart}}(h)f}{h} = 0 \quad \textit{for all } f \in C_p^2.$$

Moreover, for all $f \in C_p^2$,

$$\lim_{h\downarrow 0}\frac{\mathcal{I}^{\mathrm{Mart}}(h)f - \mu f}{h} = \varphi^*\left(\frac{1}{2}\left(\int_H \big(\big|\nabla^2 f(y)\big|_{\mathrm{Max}}\big)^{\frac{p}{p-2}}\,\mu(\mathrm{d}y)\right)^{\frac{p-2}{p}}\right). \tag{2.22}$$

Remark 2.10. The proof of the previous theorem is based on the fact that, for every function $f \in C_p^2$ and every $\varepsilon > 0$, there exists a measurable function $v\colon H \to H$ with $\|v(x)\| \in \{0,1\}$ and

$$\big|\nabla^2 f(x)\big|_{\mathrm{Max}}^{\frac{2}{p-2}}\big\langle v(x), \nabla^2 f(x)v(x)\big\rangle \ge \big|\nabla^2 f(x)\big|_{\mathrm{Max}}^{\frac{p}{p-2}} - \varepsilon \quad \text{for all } x \in H. \tag{2.23}$$

In fact, let $\varepsilon > 0$ and $\mathbb{S}_0 := \{w \in H \,|\, \|w\| = 1\} \cup \{0\}$ denote the unit sphere enriched by the neutral element in H. Then, the function $H \times \mathbb{S}_0 \to \mathbb{R}$, $(x,w) \mapsto \frac{1}{2}\langle w, \nabla^2 f(x)w\rangle$ is continuous, and we can apply a measurable selection argument, see [10, Proposition 7.34], in order to come up with a measurable function $v\colon H \to \mathbb{S}_0$ satisfying (2.23). If $H = \mathbb{R}^d$, the set $\mathbb{S}_0$ is compact, and the supremum in Equation (2.16) is indeed a maximum. In this case, we can invoke [10, Proposition 7.33] to obtain a *measurable direction of maximal curvature* v, i.e., a measurable map $v\colon H \to H$ with $\|v(x)\| \in \{0,1\}$ and

$$\big\langle v(x), \nabla^2 f(x)v(x)\big\rangle = \big|\nabla^2 f(x)\big|_{\mathrm{Max}} \quad \text{for all } x \in H.$$

Inspecting the proof of the previous theorem, we immediately get the following reduction of the parameter set if f has a measurable direction of maximal curvature.

Corollary 2.11. *Assume that $f \in C_p^2$ has a measurable direction of maximal curvature v, and define*

$$\theta(x) = v(x)\big(|\nabla^2 f(x)|_{\mathrm{Max}}\big)^{1/(p-2)} \quad \textit{for all } x \in H.$$

Then,

$$\lim_{h\downarrow 0}\frac{\mathcal{I}^{\mathrm{Mart}}(h)f - \mathcal{I}_\Theta^{\mathrm{Mart}}(h)f}{h} = 0,$$

whenever $\{\lambda\theta \,|\, \lambda \ge 0\} \subseteq \Theta$.

3. Numerical computation of the parametric risk functional

In this section, we discuss numerical methods to approximate the functional $\mathcal{I}_\Theta(h)$ for $h > 0$. We show how, under different hypotheses, the infinite-dimensional optimization problem related to the computation of $\mathcal{I}_\Theta(h)$ can be reduced to a finite-dimensional optimization for all $h > 0$. This echoes and complements the results in [26] and [16].

We distinguish between two relevant cases: the one where the baseline measure μ is a convex combination of Dirac measures, and the more general one, which also includes regular measures. In the first case, the operator $\mathcal{I}_\Theta(h)$ can be obtained through a finite-dimensional optimization scheme, whereas in the second case we approximate the operator $\mathcal{I}_\Theta(h)$ via a neural network approach for all $h > 0$. Moreover, we show how to extend these numerical schemes to the constrained parametric operators.

3.1. Convex combination of Dirac measures.

Throughout this subsection, we assume that $\mu = \sum_{i=1}^n a_i \delta_{x_i}$, where, for $i \in \{1, \dots, n\}$, $\alpha_i \in (0, \infty)$ and δ_{x_i} is the Dirac measure centered in $x_i \in H$. Moreover, since μ is a probability measure, $\sum_{i=1}^n \alpha_i = 1$ is implicitly assumed. Such a type of baseline model is particularly relevant when considering empirical distributions, including data driven problems where one observes historical events and assumes that the generating distribution is the uniform distribution on the sample points (i.e., $a_i = 1/n$ for all $i \in \{1, \dots, n\}$). In this case, the worst case loss, represented by $\mathcal{I}(h)f$ with uncertainty level $h > 0$, accounts for statistical errors present in the sample, cf. [26] for more details and real world applications of this type of estimates.

In this situation, two functions $g_1, g_2 \colon H \to H$ are μ-a.s. equal if and only if they coincide on the atoms of the distribution μ, i.e., if and only if $g_1(x_i) = g_2(x_i)$ for all $i \in \{1, \dots, n\}$. In particular, the set $L_p(\mu, H)$ is isomorphic to the n-fold Cartesian product $H^n := \Pi_{i=1}^n H$ of H with itself (endowed with the product topology). This leads to a drastic simplification of the related optimization problem, and if H is finite-dimensional, so is the parameter set. We summarize these observations in the following proposition.

Proposition 3.1. *Let $n \in \mathbb{N}$, $x_1, \dots, x_n \in H$, $\alpha_1, \dots \alpha_n \in (0, \infty)$ with $\sum_{i=1}^n \alpha_i = 1$, and*

$$\mu = \sum_{i=1}^n \alpha_i \delta_{x_i}.$$

Moreover, let $\mathcal{D} \subset H^n$ be a dense subset of H^n with $0 \in \mathcal{D}$, and define

$$\mathcal{I}_\mathcal{D}(h)f := \sup_{(\theta_1, \dots, \theta_n) \in \mathcal{D}} \left(\sum_{i=1}^n \alpha_i f(x_i + \theta_i) - \varphi_h \left(\left(\sum_{i=1}^n \alpha_i \|\theta_i\|^p \right)^{1/p} \right) \right) \quad \text{for all } f \in \mathrm{Lip}_p.$$

Then,

$$\lim_{h \downarrow 0} \frac{\mathcal{I}(h)f - \mathcal{I}_D(h)f}{h} = 0$$

for every $f \in \mathrm{Lip}_p$ with a measurable direction of steepest ascent.

Note that, in case of $H = \mathbb{R}^d$, this result is somewhat similar to [26, Theorem 4.4], where, working with an empirical distribution, the authors show that the worst case loss, for a function that is the pointwise maximum of concave functions (and satisfies some other conditions), can be computed via a convex shifting of the initial sample points. In our framework, we can widely relax the hypothesis on the loss function f, consider more general atomic measures, and come up with a result in the spirit of [26, Theorem 4.4] for infinitesimally small amounts of uncertainty. In conclusion, if the baseline measure is given by a convex combination of Dirac measures, for small levels of uncertainty, one does not have to spread away mass in order to find the worst case loss.

3.2. **Absolutely continuous measures and approximation by neural networks.** In this subsection, we provide a framework to build an approximation of the operator $\mathcal{I}_\Theta(h)$, for $h > 0$, in the case where $H = \mathbb{R}^d$ and the baseline measure μ is absolutely continuous with respect to the Lebesgue measure. In this case, the set $L_p(\mu; H)$ cannot be reduced to a finite-dimensional space. We therefore explore a numerical procedure, based on a neural network approach, to approximate the operator $\mathcal{I}_\Theta(h)$ for $h > 0$.

Following the method proposed in [16], the idea is to consider an increasing family of parameter sets, which are finite-dimensional and whose union is dense in $L_p(\mu; H)$.

The numerical method builds on the following result, which holds true for every baseline measure μ, even if we apply it mainly to reference measures which are absolutely continuous with respect to the Lebesgue measure.

Proposition 3.2. *Let $(\Theta^n)_{n\in\mathbb{N}}$ be a family of subsets of $L_p(\mu; H)$ with $\Theta^n \subseteq \Theta^{n+1}$ for all $n \in \mathbb{N}$. If the set $\bigcup_{n\in\mathbb{N}} \Theta^n$ is dense in $L_p(\mu; H)$ and contains 0, then*

$$\mathcal{I}_{\Theta^n}(h)f \nearrow \mathcal{I}_{L_p(\mu;H)}(h)f \quad \text{as } n \to \infty$$

for all $h > 0$ and $f \in \mathrm{Lip}_p$.

Proof. Let $h > 0$, $f \in \mathrm{Lip}_p$, and $\Theta := \bigcup_{n\in\mathbb{N}} \Theta^n$. The inclusion $\Theta^n \subseteq \Theta^{n+1}$ implies that $\mathcal{I}_{\Theta^n}(h)f \leq \mathcal{I}_{\Theta^{n+1}}(h)f$ for all $n \in \mathbb{N}$. Moreover,

$$\mathcal{I}_\Theta(h)f = \sup_{n\in\mathbb{N}} \mathcal{I}_{\Theta^n}(h)f \leq \mathcal{I}_{L_p(\mu;H)}(h)f.$$

The statement now follows from Lemma 5.2. □

Although the set $L_p(\mu; H)$ might, a priori, be very large, in particular if the reference measure μ is absolutely continuous with respect to the Lebesgue measure, Proposition 3.2 tells us that we can implement a numerical scheme to approximate $\mathcal{I}_{L_p(\mu;H)}(h)$ for all $h > 0$ based on an increasing sequence of subsets Θ^n of $L_p(\mu; H)$, for which we can compute $\mathcal{I}_{\Theta^n}(h)$.

Considering the case $H = \mathbb{R}^d$, [22, Theorem 1] ensures that the set of neural networks with one layer and arbitrary number of neurons is dense in $L_p(\mu)$ and, with a simple generalization, in $L_p(\mu; \mathbb{R}^d)$, cf. [23, Corollary 2.6 and Corollary 2.7]). Therefore, using Proposition 3.2, we can can approximate the quantity $\mathcal{I}_{L_p(\mu;H)}(h)f$ with the quantity $\mathcal{I}_{\Theta^n}(h)f$ for all $h > 0$, where Θ^n is the set of functions from $\mathbb{R}^d$ to $\mathbb{R}^d$ given by neural networks with one layer and n neurons, for $n \in \mathbb{N}$ sufficiently large. We provide a more precise framework, following closely what has been done in [16].

For $m \in \mathbb{N}$, we consider fully connected feed-forward neural networks from $\mathbb{R}^d$ to $\mathbb{R}^d$, which are maps of the form

$$x \mapsto A_m \circ \psi \circ A_{m-1} \circ \cdots \circ \psi \circ A_0(x),$$

where A_i are affine transoformations of the form $A_i = M_i x + b_i$, for a matrix M_i and a vector b_i, $\psi\colon \mathbb{R} \to \mathbb{R}$ is a nonlinear activation function, and the application of the *activation function* is to be understood componentwise, i.e., $\psi((x_1, \dots, x_d)) = (\psi(x_1), \dots, \psi(x_d))$. We say that the neural network has *depth* m and *width* $n \in \mathbb{N}$ (alternatively has m *layers* and n *neurons* per layer) if A_0 is a map from $\mathbb{R}^d$ to $\mathbb{R}^n$, $A_1, \dots, A_{m-1}$ are maps from $\mathbb{R}^n$ to $\mathbb{R}^n$, and A_m is a map from $\mathbb{R}^n$ to $\mathbb{R}^d$. The coefficients of the matrices and the vectors forming the affine transformations $A_1, \dots, A_m$ represent the parameters of the network and can be regarded as an element of $\mathbb{R}^N$, for some $N \in \mathbb{N}$ that depends on the structure of the network. We call $\mathfrak{N}_d^{m,n}$ the set of neural network from $\mathbb{R}^d$ to $\mathbb{R}^d$ with depth m and width n. In case we want to specify the activation function used in the network, we also write $\mathfrak{N}_d^{m,n}(\psi)$, where ψ is the activation function. Moreover, $\mathfrak{N}_d^m := \bigcup_{n\in\mathbb{N}} \mathfrak{N}_d^{m,n}$ denotes the set

of neural networks with m layers and arbitrary number of neurons. Proposition 3.2 can now be restated in a neural network framework.

Corollary 3.3. *For every integer $m \in \mathbb{N}$ and every bounded, nonconstant, Lipschitz continuous activation function ψ, we have*

$$\mathcal{I}_{\mathfrak{N}_d^{m,n}}(h)f \nearrow \mathcal{I}_{L_p(\mu;H)}(h)f \quad \text{as } n \to \infty$$

for all $h > 0$ and $f \in \mathrm{Lip}_p$.

Proof. The convergence for $m = 1$ is direct consequence of Proposition 3.2 and [22, Theorem 1]. Since the activation function ψ is Lipschitz continuous, the functions in $\mathfrak{N}_d^{m,n}$ are also Lipschitz continuous as a composition of Lipschitz continuous functions, so that $\mathfrak{N}_d^{m,n} \subseteq L_p(\mu; H)$. Moreover, $\mathfrak{N}_d^{1,n} \subseteq \mathfrak{N}_d^{m,n}$, and therefore

$$\mathcal{I}_{\mathfrak{N}_d^{1,n}}(h)f \leq \mathcal{I}_{\mathfrak{N}_d^{m,n}}(h)f \leq \mathcal{I}_{L_p(\mu;H)}(h)f \quad \text{for all } n \in \mathbb{N}.$$

□

Remark 3.4. a) As already stated in Remark 2.2, when the baseline measure μ is absolutely continuous w.r.t. the Lebesgue measure, the parametrized version $\mathcal{I}_{L_p(\mu;H)}(h)f$ corresponds in fact to $\mathcal{I}(h)f$, so that the previous corollary gives us an approximation scheme for $\mathcal{I}(h)f$ for all $h > 0$.

b) The fact that we get an approximation from below complements the approach in [16], where the authors obtain an approximation from above based on a version of the Daniell-Stone theorem, which leads to a dual representation in terms of a superhedging problem. Therefore, their functional is expressed as the infimum over a set of functions that they approximate via neural networks. Combining the two approaches, one can overcome the problem of an unknown rate of convergence as $h \downarrow 0$ by estimating the relevant quantity from below and from above. In practice, one can then train the two algorithms until the gap between the two estimates is below a certain threshold.

c) The assumption of a bounded activation function can be relaxed, and we can allow for unbounded Lipschitz continuous functions using more layers, cf. [22, Section 3] and [16, Remark 3.4]. In fact, let ζ be a bounded, nonconstant, and Lipschitz continuous activation function, and let ψ be an unbounded activation function such that $\zeta \in \mathfrak{N}_1^1(\psi)$. Then, for $m > 1$, we have $\mathfrak{N}_d^m(\zeta) \subset \mathfrak{N}_d^m(\psi)$ and hence $\mathfrak{N}_d^m(\psi)$ is dense in $L_p(\mu; \mathbb{R}^d)$, since $\mathfrak{N}_d^m(\psi) \subseteq L_p(\mu; \mathbb{R}^d)$ due to the Lipschitz continuity of ψ. This allows us to work, for example, with ReLU activation functions, i.e., $\psi(x) = \max\{x, 0\}$, since one can obtain the bounded function $\zeta(x) = \min\{\max\{x, 0\}, 1\}$ from a ReLU network. Alternatively, note that we can prove Corollary 3.3 for neural networks with ReLU activation function directly using the simple approximation result, cf. Proposition A.1, without invoking the universal approximation of [22, Theorem 1]. In fact, a single layer network with ReLU activation function and arbitrary number of neurons can reproduce the span of the set L_1 in Remark A.2 c).

d) One could also choose a different class of universal approximators, like polynomials. Our choice of employing a neural network framework comes from their recent applications to optimal transport and penalized problems, as, for example, in [17] and [16]. Moreover, in Section 4.5, we shall give a financial interpretation to the neural network approximation.

In the numerical implementation of the algorithm, we train the network using as loss function

$$L(\theta) := -\left(\int_{\mathbb{R}^d} f(z)\, \mu_\theta(\mathrm{d}z) - \varphi_h(\|\theta\|_{L_p(\mu;\mathbb{R}^d)})\right) \quad \text{for all } \theta \in \Theta,$$

where the integrals are computed via Monte Carlo simulations. Note that this corresponds to a classical stochastic gradient descent, since at each step of the optimization new samples are drawn from the distribution μ.

Finally, we remark that this approach can also be employed when working with empirical distributions. In particular, when the sample size is high, stochastic gradient descent might be preferred over the full-batch finite-dimensional optimization proposed in Section 3.1, see, for example, [12].

3.3. **Extension of neural network scheme to constrained settings.** In this section, we briefly explain how to extend the neural network scheme proposed in Section 3.2 to the settings with mean and martingale constraint, cf. Section 2.2 and Section 2.3.

In the case of a mean constraint, one should in principle restrict the optimization to functions in $L_p(\mu; H)$ with mean zero. Even if, a priori, it is not possible to impose this constraint on the neural network, we can obtain it simply by subtracting componentwise the Monte Carlo mean of the vector field. Therefore, at each step of the optimization, we only have to perform the computation of the quantity $\int_H \theta(y)\,\mu(\mathrm{d}y)$, which doesn't require any additional sampling.

The martingale constraint almost doesn't require any additional computational effort. As a matter of fact, in this framework we have

$$\int_H f(z)\,\mu_\theta(\mathrm{d}z) = \int_H \frac{f\big(y+\theta(y)\big) - f\big(y-\theta(y)\big)}{2}\,\mu(\mathrm{d}y) \quad \text{for all } \theta \in L_p(\mu; H),$$

and it is clear that we only have to perform one evaluation more in each step of the optimization.

4. Examples and applications

The aim of this section is to illustrate the results developed in the previous sections in several examples and applications.

4.1. **A toy model for an earthquake.** We start by investigating a toy model for an earthquake, i.e., a one-period model, where the baseline measure μ describes an a priori estimate for the location of the epicenter of an earthquake in $\mathbb{R}^2$. Suppose that an insurance company, selling a coverage for damages caused by earthquakes, wants to compute the worst case loss in terms of uncertainty regarding the precise location of the epicenter. For the sake of simplicity and analytical tractability, we choose a loss function that is given by the following compactly supported version of a Gaussian kernel:

$$f(x_1,x_2) = \begin{cases} c\exp\left(1+\frac{r^2}{(x_1-\bar{x}_1)^2+(x_2-\bar{x}_2)^2-r^2}\right), & \text{if } (x_1-\bar{x}_1)^2+(x_2-\bar{x}_2)^2 < r^2,\\ 0, & \text{otherwise,}\end{cases}$$

where $\bar{x} \in \mathbb{R}^2$ can, for example, represent the location of a city center, $c > 0$ and $r > 0$ are positive constants representing, respectively, the maximal loss level and the radius of the area covered by the insurance. The idea is that losses decrease as the epicenter of the earthquake moves away from the city center. When needed, we will write $f(c, r, \bar{x}_1, \bar{x}_2; x_1, x_2)$ to highlight the dependence on c, r, and $\bar{x}$. Figure 1 depicts this loss function for $r = 1.5$, $c = 1$, and the city center $\bar{x}$ located in the origin.

In the following, we work with the 2-Wasserstein distance, and numerically approximate the quantity

$$\mathcal{I}(h)f = \sup_{\nu\in\mathcal{P}_2}\left(\int_{\mathbb{R}^2} f(z_1,z_2)\,\nu(\mathrm{d}z_1,\mathrm{d}z_2) - \varphi_h(\mathcal{W}_2(\mu,\nu))\right)$$

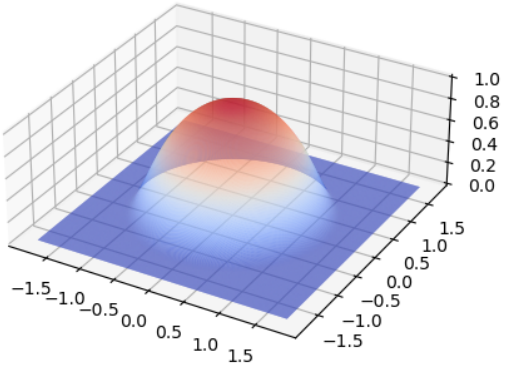


FIGURE 1. Loss function for the earthquake model

computing its parametric version

$$\mathcal{I}_{L_2(\mu;\mathbb{R}^2)}(h)f = \sup_{\theta\in L_2(\mu;\mathbb{R}^2)} \left(\int_{\mathbb{R}^2} f(z_1, z_2)\, \mu_\theta(\mathrm{d}z_1, \mathrm{d}z_2) - \varphi_h\big(\|\theta\|_{L_2(\mu;\mathbb{R}^2)}\big) \right)$$

through a procedure which is independent of the loss function f. We consider the penalty function $\varphi(v) = v^2$ for $v \geq 0$, and we explore two different frameworks, one leading to a finite-dimensional optimization and another one based on neural networks.

4.2. **Epicenter as a sum of Dirac measures.** Consider the loss function from Section 4.1 and a baseline model given by

$$\mu = \frac{1}{3}\delta_{x_1} + \frac{1}{3}\delta_{x_2} + \frac{1}{3}\delta_{x_3}$$

with $x_1 = (0.55,\, 0)$, $x_2 = (0,\, 0.85)$, and $x_3 = (-1.10,\, 0)$. The optimization suggested by Proposition 3.1 can be easily performed with *out of the box* tools that are nowadays integrated in many scientific programming languages. We implement the optimization using Python 3.9[1] and the optimizer included in the package Scipy (version 1.7.3), which applies the Quasi-Newton method BFGS, see [27, Chapter 6]. Figure 2 illustrates the initial situation, the optimizers, and the value of the operator $\mathcal{I}_{L_2(\mu;\mathbb{R}^2)}(h)$ at different uncertainty levels h ranging from 0 to 0.3. Looking at Figure 2 (b), we see how the worst case loss is obtained by shifting the three possible locations for the epicenter of the earthquake towards the city center, which means in the direction of the gradient of the loss function. This is in line with Theorem 2.5 and Remark 2.4, nevertheless it is worth noting that the optimization procedure is independent of the particular loss function. Moreover, note that, because of the penalization term, one has to pay a price to move the atoms of the measure; this is why the worst case loss is obtained by shifting more the points whose incremental contribute to the loss is higher, that is, those points which lie in a region where the loss function is steeper.

4.3. **Epicenter with absolutely continuous distribution.** Here, we exploit the neural network approach developed in Section 3.2 to approximate the value of $\mathcal{I}_\Theta(h)f$, for $h > 0$, in the case where the baseline measure μ is a two-dimensional Gaussian random variable with mean $(1,\, 0)$ and covariance matrix equals the identity matrix. Again, we implement the code in Python 3.9, using the package Pytorch [30] (version 1.11.0) to handle the neural network architecture.

[1]All source codes are available at https://github.com/sgarale/wasserstein_parametrization.

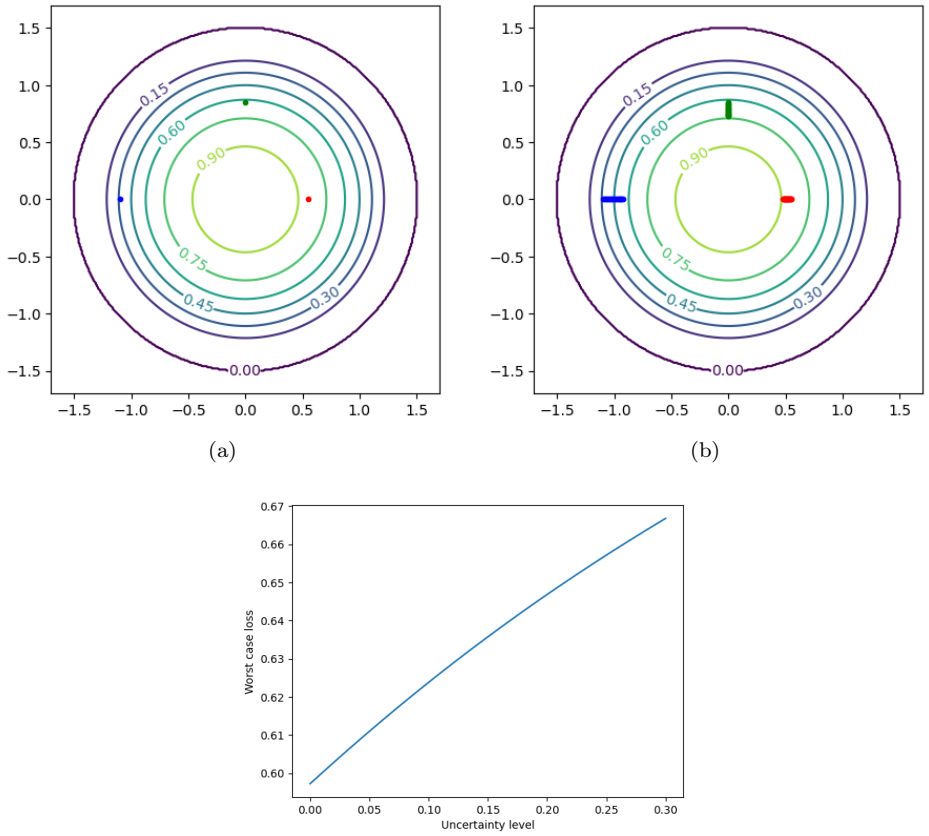
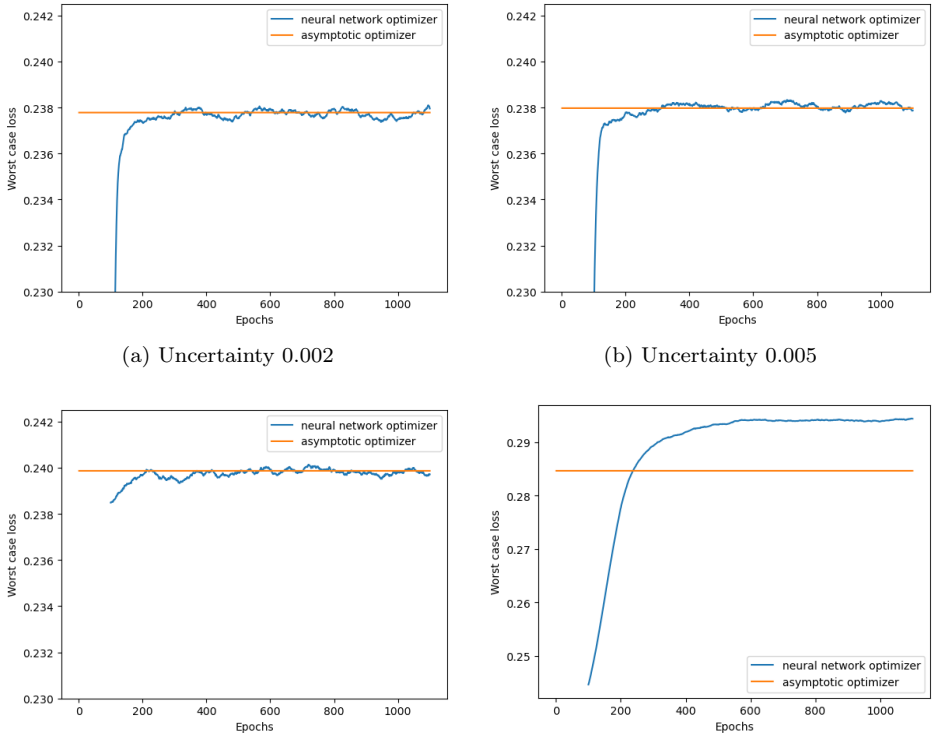


FIGURE 2. Calculation of the worst case loss for a reference model given by a convex sum of Dirac measures. Figure (a): loss function and atoms of the baseline measure. Figure (b): optimizers as the uncertainty level increases. Figure (c): worst case loss for different levels of uncertainty.

We use the built-in Adam optimizer, cf. [25], and present the results obtained from training a neural network with depth 4, width 20, and ReLU activation function $\psi(x) = \max\{x, 0\}$. For the optimization, we use a learning rate of 0.001 and a batch size of 32 768 samples (the numbers of points used to compute the integrals via Monte Carlo). The calculations are run on a laptop with Intel® Core™ i7-7600U processor (2.90GHz) and 16GB RAM. Computing the worst case loss at a fixed uncertainty level takes less than a minute.

Figure 3 shows the training phase of the algorithm at different uncertainty levels and Figure 4 shows the optimizers obtained through the training. Theorem 2.5 tells us that, asymptotically as the level of uncertainty tends to zero, the optimizer is a scalar multiple of the gradient of the loss function (we are working with $p = 2$), therefore we always compare the neural network optimization with the one-dimensional optimization computed using the gradient. Note how the optimizers resemble the gradient of the loss function for small levels of uncertainty, however, for higher values of h, for example $h = 0.5$, the neural network optimization outperforms the optimization along the gradient, see Figure 3 (d).

4.4. **An example of transfer learning.** We now apply the considerations of Remark 2.6 to our neural network framework. We have seen that, when the uncertainty $h > 0$ is small,



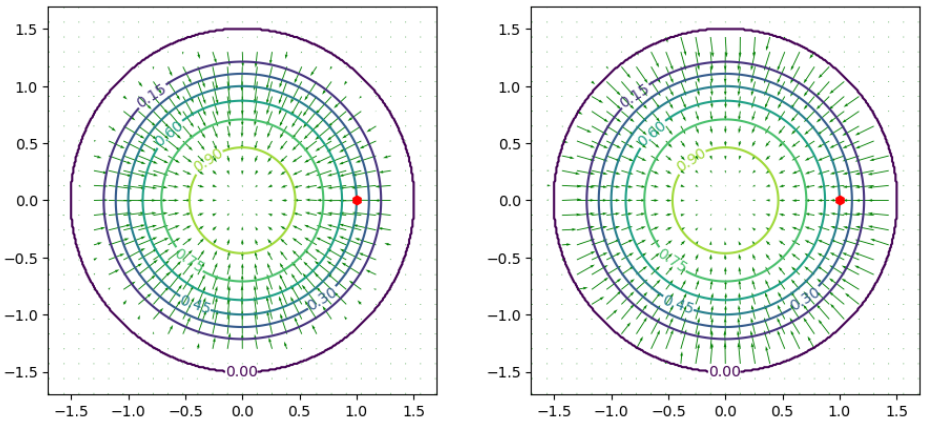
(a) Uncertainty 0.002

(b) Uncertainty 0.005

(c) Uncertainty 0.02

(d) Uncertainty 0.5

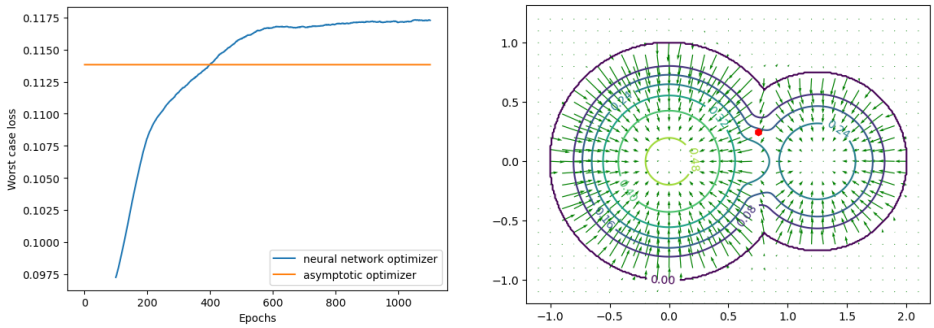
FIGURE 3. Traning of neural network for the calculation of the worst case loss for a reference model given by a Gaussian random variable. The yellow line in all the graphs represents the result of the one-dimensional optimization performed using the gradient of the loss function. All values are moving averages on a window of 100 epochs.



(a) Uncertainty 0.02

(b) Uncertainty 0.5

FIGURE 4. Optimizers obtained by training a neural network to compute the worst case loss for the gaussian reference model. The contour plots represent the loss function and the red dots represent the location of the mean of the distribution of the epicenter.



(a) Training phase (b) Optimizing vector field

FIGURE 5. Training phase and optimizer obtained for the loss function $\bar{f}$ in Section 4.4. The red dot is the location of the mean of the distribution of the epicenter.

the optimizer for $\mathcal{I}_\Theta(h)f$ is independent of the reference measure up to a one-dimensional optimization. We exploit this observation to provide a simplified framework to compute the worst case loss with respect to different reference measures. Since, for small values of h, the optimizers for different reference measures differ only by a multiplicative factor, we add a multiplicative layer to the network; formally we add the composition with a matrix $M = c\mathbb{1}_{2\times 2}$, where $\mathbb{1}_{2\times 2}$ is the two-dimensional identity matrix and $c \in \mathbb{R}$ is the parameter of this additional layer. The network then has the structure

$$x \mapsto M \circ A_m \circ \psi \circ A_{m-1} \circ \cdots \circ \psi \circ A_0(x).$$

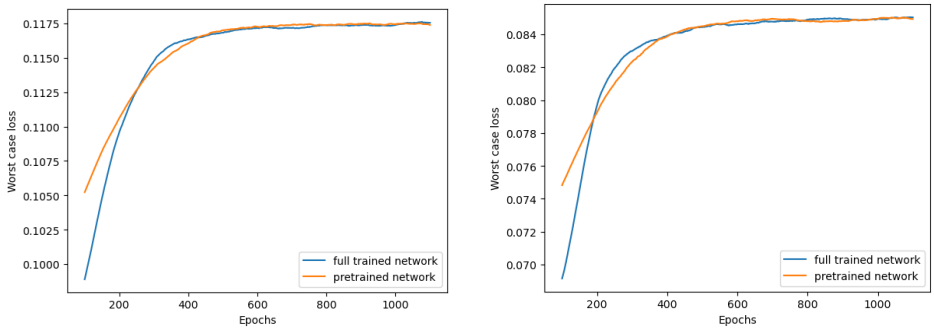
Once we have trained the neural network on a starting reference measure, we keep the parameters of the affine transformations $A_0, \ldots, A_m$ fixed, and allow only for the training of the last layer, which corresponds to performing a one-dimensional optimization. This drastically reduces the computation time of the worst case loss for different reference measures.

To avoid effects due to particular symmetries of the problem, we perform this exercise for the loss function

$$\bar{f}(x_1, x_2) := f(0.5, 1, 0, 0; x_1, x_2) + f(0.3, 0.75, 1.25, 0; x_1, x_2),$$

where the function f is defined as in Section 4.1. From an application perspective, one can think of two cities, a bigger one with the city center located in the origin and a smaller one with the city center located in $(1.25, 0)$. As reference measure, we consider a two-dimensional Gaussian random variable with mean located in $(0.75, 0.25)$ and the identity matrix as covariance matrix. Figure 5 depicts the training phase and the optimizer obtained by the training at the uncertainty level $h = 0.5$. Again, we observe that, for this level of uncertainty, the optimization of the full vector field outperforms the one-dimensional optimization obtained through the asymptotic optimizer.

Figure 6 shows the training phase of the partial training for different reference measures, where the vector field obtained by the training on the initial reference measure is optimally rescaled, in comparison with the full training on each reference measure. In this example, we move the mean of the reference model, which stays a Gaussian random variable with identity matrix as covariance matrix. These results suggest a strong stability with respect to the uncertainty level, even if, as seen before, the asymptotic optimizer is outperformed by the full neural network training for a high value of the uncertainty. In machine learning jargon, this technique is called *transfer learning*; a method that is used both to reduce



(a) Mean: (0.75, −0.25) (b) Mean: (−0.75, 0.50)

FIGURE 6. Comparison of full training and partial training (with fixed inner layers) when the mean of the reference model is moved to different locations. Mean of the reference measure for the pretrained network: (0.75, 0.25). Full training batch size: 65 536. Partial training batch size: 4 096.

computational costs and to avoid a large bias from a full training when only sparse data is available. We refer to [13] for a list of references on this topic.

If only sparse empirical data is available, this observation suggests to train the neural network for the approximation of the worst case loss on a randomly generated and sufficiently large dataset, and then to fine-tune it with a one-dimensional training on the available empirical data.

4.5. **A financial perspective: replicating uncertainty via portfolios of call options or digital options.** Here, we use a simple approximation result given in the Appendix A to give a financial interpretation to the parametric approximation of the operator $\mathcal{I}(h)$, for $h > 0$, and to motivate the use of a neural network approximation. We assume that $H = \mathbb{R}$, and consider the payoff function $f\colon \mathbb{R} \to \mathbb{R}$ of a financial contract depending on an underlying risk factor (typically an equity asset or index). The reference measure μ is then the distribution of this risk factor. Note that in order to include the case of an asset that takes only nonnegative values, one can simply pass to a log-scale, while maintaining the structure of a Hilbert space, cf. Section 4.6.

Taking $\Theta = L_p(\mu;\mathbb{R})$, we obtain a first order approximation of $\mathcal{I}(h)$ as $h \downarrow 0$ by Theorem 2.7. Moreover, Lemma 5.2 tells us that we can rather look at a dense subset of $L_p(\mu;\mathbb{R})$. As stated in Remark A.2 c), for example, portfolios of call options and portfolios of digital options are dense in $L_p(\mu;\mathbb{R})$. In fact, portfolios of digital options are the span of the set $L_0 := \{\mathbb{1}_{[K,\infty)} \,|\, K \in \mathbb{R}\}$ and portfolios of call options are the span of the set $L_1 := \{[x \mapsto (x-K)^+] \,|\, K \in \mathbb{R}\}$. This means that we can obtain a first order approximation via the parametric version $\mathcal{I}_\Theta(h)$, for small $h > 0$, where the set Θ is given by the profiles of all possible portfolios of call options or digital options.

As we remarked in Section 2.2, the hypothesis of absence of arbitrage, imposes a mean constraint on perturbations of the reference measure, which leads to so-called *self-financing* portfolios. As a matter of fact, the condition $\int_{\mathbb{R}} \theta(y)\,\mu(\mathrm{d}z) = 0$ corresponds to the fact that the portfolio with profile θ has price zero (we are working in a simple setting where the risk-free rate is zero, but this can be easily generalized), i.e., we can replicate the uncertainty on the price of a financial position without spending additional money.

As a final remark, we point out that an arbitrary portfolio of call options can be built using a single-layer neural network with ReLU activation function, which is exactly the profile of a call option with strike zero. Therefore, one can prove the density of ReLU networks

using Proposition A.1 without invoking universal approximation theorems. Moreover, as a byproduct of the training, we get a portfolio of call options which replicates the uncertainty; this insight may have interesting implications for the financial industry.

4.6. **An example for the martingale constraint: pricing of a bull spread shortly after a quoted maturity.** We conclude this section by providing a framework that illustrates the implementation of the martingale constraint.

Suppose that we want to price a bull call spread option with strikes $K_1, K_2 \in \mathbb{R}$ with $0 < K_1 < K_2$ and maturity $T + h$, for a small time $h > 0$, having quoted options for the shorter maturity $T > 0$. The payoff of the option is given by the sum of a long call option on the lower strike K_1 and a short call option on the upper strike K_2:

$$f(x) = \max(x - K_1, 0) - \max(x - K_2, 0).$$

This situation is complementary to a situation described in [14]. There, it is shown that pricing a call option on a maturity that is shorter than the first quoted maturity, using a diffusion model with deterministic volatility curve, naturally leads to volatility uncertainty. Here, we reverse the point of view, and assume that we know the distribution of the underlying at time T, and we look for the price of the option at time $T + h$. Since no information is provided by the market for the maturity $T+h$, we look for upper and lower price bounds based on Wasserstein uncertainty around the distribution μ of the underlying at time T, which can be seen as a best guess for the distribution of the underlying at time $T + h$. We assume μ to be a log-normal distribution coming from a Black-Scholes model, cf. [18, Chapter 5], with volatility $\sigma > 0$ estimated from market data. This choice is quite natural since the log-normal distribution is used by the market to quote options via the implied volatility surface. For simplicity, we assume that the risk-free rate is zero, so that the value of the option at time 0 for the maturity T is given by

$$V_{0,T}(f) := \int_{\mathbb{R}} f\big(e^y\big)\, N\big(-\tfrac{\sigma^2 T}{2}, \sigma^2 T\big)(\mathrm{d}y),$$

i.e., we are working with the reference measure μ on $\mathbb{R}$, which is the distribution of e^Y for $Y \sim N\big(-\frac{\sigma^2 T}{2}, \sigma^2 T\big)$. We make the following two assumptions:

a) the uncertainty in the returns grows as the product of the square root of time and the volatility σ, i.e., the uncertainty level at time $T + h$ is $\sigma\sqrt{h}$,
b) the market is free of arbitrage.

The first hypothesis is in line with the assumption of a log-normal distribution as a baseline measure. In fact, embedding this distribution in a dynamic setting, we get a Black-Scholes model, where the volatility of the diffusion process increases as the square root of time; we are then imposing the same scaling for the uncertainty. The second hypothesis is standard and, in the spirit of the fundamental theorem of asset pricing, it implies that the asset process is a martingale, see, for instance, [18, Theorem 1.7].

Under these hypotheses, we can bound the value $V_{0,T+h}(f)$ of the option at time 0 with maturity $T + h$ by

$$-\mathcal{I}^{\mathrm{Mart}}(h)(-f) \le V_{0,T+h}(f) \le \mathcal{I}^{\mathrm{Mart}}(h)f \quad \text{for all } h > 0,$$

where the penalty function in the definition of the operator $\mathcal{I}^{\mathrm{Mart}}$ is given by $\varphi = \infty \cdot \mathbb{1}_{(\sigma^2,\infty)}$. Therefore, for small values of h, we can approximate these bounds estimating the operator $\mathcal{I}^{\mathrm{Mart}}_{\Theta}(h)f$.

Using an adjusted neural network scheme as explained in Section 3.3 and thinking of the Black-Scholes price as the price without uncertainty, we depict the outcome of this estimate in Figure 7 with reference time $T = 1$ (one year) and intervals h ranging from 0 to 1/12 (one

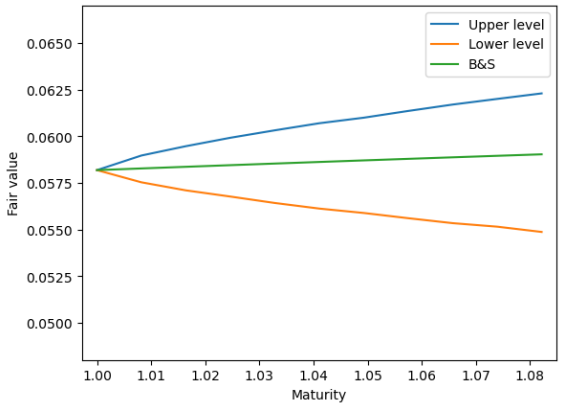


FIGURE 7. Black-Scholes option price with upper and lower bounds obtained by imposing the martingale constraint in the perturbation of the reference model. For the optimization, a neural network with width 20 and depth 4 has been used. Batch size: 2^{19} points.

month). We work with $p = 3$ and, in order to avoid numerical problems related to the value ∞ in the convex indicator function, we replace the penalization function $\varphi = \infty \cdot \mathbb{1}_{(\sigma^2,\infty)}$ by the real-valued function $\varphi(x) = \frac{1}{n}(x/\sigma^2)^n$ with $n = 5$. Heuristically, this is coherent with our assumptions, since, thanks to Lemma 5.3, we have the desired rescaling of the uncertainty up to a constant factor, and, moreover, this penalization converges pointwise to $\infty \cdot \mathbb{1}_{(\sigma^2,\infty)}$ as $n \to \infty$. Moreover, we remark that, although the payoff function of a bull spread is not twice continuously differentiable, at least on a discrete level, it is indistinguishable from a smoothened version of it, also in view of the fact that, in this example, the set of kinks has measure zero under μ.

5. PROOFS

In order to prove Theorem 2.5 and Theorem 2.7, we need the following two auxiliary results. The first one is a version of a standard result, cf. [6] and [19], specific to our setting. For the sake of a self-contained exposition, we provide a short proof.

Lemma 5.1. *Let $f \in \mathrm{Lip}_p$. Then, there exist $h_0 > 0$ and $a \geq 0$ (depending only on f and the penalty function φ) such that*

$$\mathcal{I}(h)f = \sup_{\mathcal{W}_p(\mu,\nu)\leq ah} \left(\int_H f(z)\,\nu(\mathrm{d}z) - \varphi_h\big(\mathcal{W}_p(\mu,\nu)\big) \right) \quad \text{for all } h \in (0, h_0]. \tag{5.1}$$

Proof. Since $f \in \mathrm{Lip}_p$, there exists a constant $L_f \geq 0$ such that

$$|f(x_1) - f(x_2)| \leq L_f\big(1 + \max\{\|x_1\|, \|x_2\|\}\big)^{p-1}\|x_1 - x_2\| \quad \text{for all } x_1, x_2 \in H.$$

By (2.2), there exist constants $a \geq 0$ and $h_0 > 0$ such that

$$\varphi(u) > 1 + L_f(1 + |\mu|_p + h_0 u)^{p-1}u \quad \text{for all } u \in (a, \infty). \tag{5.2}$$

Let $h \in (0, h_0]$, $\nu \in \mathcal{P}_p$ with

$$\int_H f(y)\,\mu(\mathrm{d}y) \leq \mathcal{I}(h)f \leq h + \int_H f(z)\,\nu(\mathrm{d}z) - \varphi_h\big(\mathcal{W}_p(\mu,\nu)\big),$$

and $\pi \in \mathrm{Cpl}(\mu,\nu)$ be an optimal coupling between μ and ν. Then, using Hölder's inequality and Minkowski's inequality,

$$\begin{aligned}\varphi\left(\frac{\mathcal{W}_p(\mu,\nu)}{h}\right) &\leq 1 + \int_{H\times H} \frac{f(z)-f(y)}{h}\,\pi(\mathrm{d}y,\mathrm{d}z)\\ &\leq 1 + L_f \int_{H\times H} \big(1+\max\{\|y\|,\|z\|\}\big)^{p-1}\frac{\|z-y\|}{h}\,\pi(\mathrm{d}y,\mathrm{d}z)\\ &\leq 1 + L_f\bigg(\int_{H\times H}\big(1+\|y\|+\|z-y\|\big)^p\,\pi(\mathrm{d}y,\mathrm{d}z)\bigg)^{1/q}\frac{\mathcal{W}_p(\mu,\nu)}{h}\\ &\leq 1 + L_f\bigg(1+|\mu|_p+h_0\frac{\mathcal{W}_p(\mu,\nu)}{h}\bigg)^{p-1}\frac{\mathcal{W}_p(\mu,\nu)}{h}.\end{aligned}$$

By (5.2), it follows that $\frac{\mathcal{W}_p(\mu,\nu)}{h} \leq a$. □

Lemma 5.2. *Let Θ be dense in $L_p(\mu;H)$ with $0\in\Theta$. Then,*

$$\mathcal{I}_\Theta(h)f = \mathcal{I}_{L_p(\mu;H)}(h)f \quad \textit{for all } f\in \mathrm{Lip}_p \textit{ and } h>0.$$

Proof. Let $h>0$ and $f\in\mathrm{Lip}_p$. Clearly, $\mathcal{I}_\Theta(h)f \leq \mathcal{I}_{L_p(\mu;H)}(h)f$ since $\Theta\subset L_p(\mu;H)$. To prove equality, fix $\varepsilon>0$, and let $\theta_0\in L_p(\mu;H)$ with

$$\mathcal{I}_{L_p(\mu;H)}(h)f \leq \int_H f(z)\,\mu_{\theta_0}(\mathrm{d}z) - \varphi_h(\|\theta_0\|_{L_p(\mu;H)}) + \frac{\varepsilon}{3}.$$

Then, by dominated convergence, there exists some $\lambda\in(0,1)$ such that

$$\int_H f(z)\,\mu_{\theta_0}(\mathrm{d}z) \leq \int_H f(z)\,\mu_{\lambda\theta_0}(\mathrm{d}z) + \frac{\varepsilon}{3}.$$

Again, by dominated convergence and since Θ is dense in $L_p(\mu;H)$ with $0\in\Theta$, there exists some $\theta\in\Theta$ with $\|\theta\|_{L_p(\mu;H)} \leq \|\theta_0\|_{L_p(\mu;H)}$ and

$$\int_H f(z)\,\mu_{\lambda\theta_0}(\mathrm{d}z) \leq \int_H f(z)\,\mu_{\theta}(\mathrm{d}z) + \frac{\varepsilon}{3}.$$

Therefore, since φ is nondecreasing,

$$\mathcal{I}_{L_p(\mu;H)}(h)f \leq \int_H f(z)\,\mu_\theta(\mathrm{d}z) - \varphi_h(\|\theta\|_{L_p(\mu;H)}) + \varepsilon \leq \mathcal{I}_\Theta(h)f + \varepsilon.$$

Letting $\varepsilon\downarrow 0$, we find that $\mathcal{I}_\Theta(h)f = \mathcal{I}_{L_p(\mu;H)}(h)f$. □

Proof of Theorem 2.5. The proof is divided in two steps. In a first step, we show that

$$\limsup_{h\downarrow 0}\frac{\mathcal{I}(h)f-\mu f}{h} \leq \varphi^*\Big(\big\||f|_{\mathrm{Lip}}\big\|_{L_q(\mu)}\Big), \tag{5.3}$$

and, in a second step, we prove

$$\liminf_{h\downarrow 0}\frac{\mathcal{I}_\Theta(h)f-\mu f}{h} \geq \varphi^*\Big(\big\||f|_{\mathrm{Lip}}\big\|_{L_q(\mu)}\Big), \tag{5.4}$$

whenever $\{b\theta\,|\,b\geq 0\}\subset\Theta$ for θ given by (2.12). In order to prove (5.3), let $h_0>0$ and $a\geq 0$ as in Lemma 5.1. For $\delta>0$, let $|f|_{\mathrm{Lip},\delta}\colon H\to\mathbb{R}$ be given by

$$|f|_{\mathrm{Lip},\delta}(x) := \sup_{\gamma\in(0,\delta)}\sup_{\|u\|\in\{0,1\}}\frac{f(x+\gamma u)-f(x)}{\gamma} \quad \text{for all } x\in H.$$

Then, $|f|_{\mathrm{Lip},\delta}$ is lower semicontinuous and thus measurable. Moreover, since $f \in \mathrm{Lip}_p$, for all $\delta > 0$,

$$|f|_{\mathrm{Lip},\delta}(x) \le \big(1 + \delta + \|x\|\big)^{p-1} \quad \text{for all } x \in H,$$

which shows that $|f|_{\mathrm{Lip},\delta} \in L^q(\mu)$. Let $\varepsilon > 0$. By dominated convergence, there exists some $\delta > 0$, such that

$$\int_H \big(|f|_{\mathrm{Lip},\delta}(y)\big)^q \,\mu(\mathrm{d}y) \le \int_H \big(|f|_{\mathrm{Lip}}(y)\big)^q \,\mu(\mathrm{d}y) + \varepsilon. \tag{5.5}$$

Let $h \in (0, h_0]$, $\nu \in \mathcal{P}_p$ with $\mathcal{W}_p(\mu,\nu) \le ah$ and

$$\mathcal{I}(h)f \le \varepsilon h + \int_H f(z)\,\nu(\mathrm{d}z) - \varphi_h\big(\mathcal{W}_p(\mu,\nu)\big),$$

and $\pi \in \mathrm{Cpl}(\mu,\nu)$ be an optimal coupling between μ and ν. Then, for arbitrary $\delta > 0$,

$$\begin{aligned}\frac{\mathcal{I}(h)f - \mu f}{h} &\le \varepsilon + \frac{1}{h}\int_{H\times H} f(z) - f(y)\,\pi(\mathrm{d}y,\mathrm{d}z) - \varphi\left(\frac{\mathcal{W}_p(\mu,\nu)}{h}\right)\\ &= \varepsilon + \frac{1}{h}\int_{\|z-y\|<\delta} f(z) - f(y)\,\pi(\mathrm{d}y,\mathrm{d}z) - \varphi\left(\frac{\mathcal{W}_p(\mu,\nu)}{h}\right)\\ &\quad + \frac{1}{h}\int_{\|z-y\|\ge\delta} f(z) - f(y)\,\pi(\mathrm{d}y,\mathrm{d}z).\end{aligned} \tag{5.6}$$

We start by estimating the last term on the right-hand side in (5.6). Since $f \in \mathrm{Lip}_p$,

$$\begin{aligned}\int_{\|z-y\|\ge\delta} f(z)-f(y)\,\pi(\mathrm{d}y,\mathrm{d}z) &\le L_f \int_{\|z-y\|\ge\delta} \big(1+\max\{\|y\|,\|z\|\}\big)^{p-1}\|z-y\|\,\pi(\mathrm{d}y,\mathrm{d}z)\\ &\le 2^{p-1}L_f\bigg(\int_{\|z-y\|\ge\delta}\big(1+\|y\|\big)^{p-1}\|z-y\|\,\pi(\mathrm{d}y,\mathrm{d}z) + \mathcal{W}_p(\mu,\nu)^p\bigg)\\ &\le 2^{p-1}L_f\mathcal{W}_p(\mu,\nu)\bigg(\int_{\|z-y\|\ge\delta}\big(1+\|y\|\big)^{p}\,\pi(\mathrm{d}y,\mathrm{d}z) + \mathcal{W}_p(\mu,\nu)^{p-1}\bigg),\end{aligned}$$

where, in the last step, we used Hölder's inequality together with the fact that $u^{1/q} \le u$ for all $u \in [1,\infty)$. By the monotone convergence theorem, there exists some some $\lambda \in (0,1)$ such that

$$\int_H \big(1+\|y\|\big)^p - \big(1+\|y\|\big)^{\lambda p}\,\mu(\mathrm{d}y) \le \frac{\varepsilon}{2}.$$

Using Hölder's inequality with $\frac{1}{\lambda}$ and $\frac{1}{1-\lambda}$, Minkowski's inequality, and Lemma 5.1, we find that

$$\begin{aligned}\int_{\|z-y\|\ge\delta}\big(1+\|y\|\big)^p\,\pi(\mathrm{d}y,\mathrm{d}z) &\le \int_{\|z-y\|\ge\delta}\big(1+\|y\|\big)^{\lambda p}\,\pi(\mathrm{d}y,\mathrm{d}z) + \frac{\varepsilon}{2}\\ &\le \frac{1}{\delta^{(1-\lambda)p}}\int_{H\times H}\big(1+\|y\|\big)^{\lambda p}\|z-y\|^{(1-\lambda)p}\,\pi(\mathrm{d}y,\mathrm{d}z) + \frac{\varepsilon}{2}\\ &\le \frac{1}{\delta^{(1-\lambda)p}}(1+|\mu|_p)^{\lambda p}\mathcal{W}_p(\mu,\nu)^{(1-\lambda)p} + \frac{\varepsilon}{2} \le \varepsilon\end{aligned}$$

for $h > 0$ sufficiently small. Again, by Lemma 5.1, we can estimate $\mathcal{W}_p(\mu,\nu)^{p-1} \le \varepsilon$, for $h > 0$ sufficiently small, and end up with

$$\int_{\|z-y\|\ge\delta} f(z) - f(y)\,\pi(\mathrm{d}y,\mathrm{d}z) \le 2^p L_f \varepsilon \mathcal{W}_p(\mu,\nu). \tag{5.7}$$

In order to estimate the first term on the right-hand side in (5.6), we use Hölder's inequality to observe that

$$\begin{aligned}\int_{\|z-y\|<\delta} f(z)-f(y)\,\pi(\mathrm{d}y,\mathrm{d}z) &\leq \int_{H\times H} |f|_{\mathrm{Lip},\delta}(y)\|z-y\|\,\pi(\mathrm{d}y,\mathrm{d}z)\\ &\leq \big\||f|_{\mathrm{Lip},\delta}\big\|_{L_q(\mu)}\mathcal{W}_p(\mu,\nu).\end{aligned}$$

Hence, a combination of (5.5), (5.6), and (5.7) yields that

$$\begin{aligned}\frac{\mathcal{I}(h)f-\mu f}{h} &\leq \varepsilon+\varphi^*\bigg(2^pL_f\varepsilon+\big\||f|_{\mathrm{Lip},\delta}\big\|_{L_q(\mu)}\bigg)\\ &\leq \varepsilon+\varphi^*\bigg((1+2^pL_f)\varepsilon+\big\||f|_{\mathrm{Lip}}\big\|_{L_q(\mu)}\bigg).\end{aligned}$$

The bound from above (5.3) is then proved thanks to the arbitrariness of $\varepsilon>0$ and the continuity of φ^*.

For the bound from below (5.4), consider the function $\theta\in L_p(\mu;H)$, which, for $x\in H$, is given by

$$\theta(x)=v(x)\big(|f|_{\mathrm{Lip}}(x)\big)^{q-1},$$

where $v\colon H\to H$ is a measurable direction of steepest ascent for f. We observe that, by definition of θ, $\int_H\|\theta(y)\|^p\,\mu(\mathrm{d}y)=\big\||f|_{\mathrm{Lip}}\big\|_{L_q(\mu)}^q$ and, for $b\geq 0$,

$$\frac{\mathcal{I}_\Theta(h)f-\mu f}{h}\geq\int_H\frac{f\big(y+bh\theta(y)\big)-f(y)}{h}\,\mu(\mathrm{d}y)-\varphi\big(b\|\theta\|_{L_p(\mu;H)}\big),$$

which, by Fatou's lemma, leads to

$$\begin{aligned}\liminf_{h\downarrow 0}\frac{\mathcal{I}_\Theta(h)f-\mu f}{h} &\geq b\int_H\big(|f|_{\mathrm{Lip}}(y)\big)^q\,\mu(\mathrm{d}y)-\varphi\big(b\|\theta\|_{L_p(\mu;H)}\big)\\ &= b\big\||f|_{\mathrm{Lip}}\big\|_{L_q(\mu)}^q-\varphi\Big(b\big\||f|_{\mathrm{Lip}}\big\|_{L_q(\mu)}^{q-1}\Big)\quad\text{for all } b\geq 0.\end{aligned}$$

We pass to the convex conjugate φ^* of φ by taking the supremum over all $b\geq 0$, and end up with (5.4).

□

Proof of Theorem 2.8. Consider the auxiliary function

$$\tilde f(x)=f(x)-\Big\langle\int_H\nabla f(y)\,\mu(\mathrm{d}y),x-\int_H y\,\mu(\mathrm{d}y)\Big\rangle\quad\text{for all } x\in H,$$

and note that $\nu f=\nu\tilde f$ for all $\nu\in\mathcal{P}_p(\mu)$. In particular, $\mathcal{I}^{\mathrm{Mean}}(h)f=\mathcal{I}^{\mathrm{Mean}}(h)\tilde f$ for all $h>0$. Since the supremum in the definition of $\mathcal{I}(h)$ runs over a larger set than the one in the definition of $\mathcal{I}^{\mathrm{Mean}}(h)$, we have $\mathcal{I}^{\mathrm{Mean}}(h)\tilde f\leq\mathcal{I}(h)\tilde f$ for all $h>0$. Define

$$\theta(x):=\nabla\tilde f(x)=\nabla f(x)-\int_H\nabla f(y)\,\mu(\mathrm{d}y)\quad\text{for all } x\in H,\tag{5.8}$$

and let $\Theta:=\{b\theta\,|\,b\geq 0\}$. By definition, $\int_H\theta(y)\,\mu(\mathrm{d}y)=0$, so that $\mathcal{I}_\Theta(h)f\leq\mathcal{I}^{\mathrm{Mean}}(h)f$ and $\mathcal{I}_\Theta(h)f=\mathcal{I}_\Theta(h)\tilde f$ for all $h>0$. This implies that

$$0\leq\frac{\mathcal{I}^{\mathrm{Mean}}(h)f-\mathcal{I}_\Theta(h)f}{h}\leq\frac{\mathcal{I}(h)\tilde f-\mathcal{I}_\Theta(h)\tilde f}{h}\quad\text{for all } h>0,$$

and therefore, by Theorem 2.7,

$$\lim_{h\downarrow 0}\frac{\mathcal{I}^{\mathrm{Mean}}(h)f-\mathcal{I}_\Theta(h)f}{h}=\lim_{h\downarrow 0}\frac{\mathcal{I}(h)\tilde f-\mathcal{I}_\Theta(h)\tilde f}{h}=0.$$

Again, by Theorem 2.7, it follows that

$$\lim_{h\downarrow 0}\frac{\mathcal{I}^{\mathrm{Mean}}(h)f-\mu f}{h}=\lim_{h\downarrow 0}\frac{\mathcal{I}(h)\tilde{f}-\mu\tilde{f}}{h}=\varphi^*\Big(\big\|\nabla\tilde{f}\big\|_{L_p(\mu;H)}\Big).$$

□

Before turning our focus on the proof of Theorem 2.9, we provide the following two modifications of Lemma 5.1 and Lemma 5.2.

Lemma 5.3. *Let $f\in C_p^2$. Then, there exist $h_0>0$ and constant $a\geq 0$ (depending only on f and the penalty function φ) such that*

$$\mathcal{I}^{\mathrm{Mart}}(h)f=\sup_{\mathcal{W}_p^{\mathrm{Mart}}(\mu,\nu)\leq\sqrt{ah}}\left(\int_H f(z)\,\nu(\mathrm{d}z)-\varphi_h\big(\mathcal{W}_p^{\mathrm{Mart}}(\mu,\nu)^2\big)\right)\quad \text{for all } h\in(0,h_0].$$

Proof. The proof is similar to the one of Lemma 5.1. Since $f\in C_p^2$, there exists a constant $C\geq 0$ such that

$$\big\|\nabla^2 f(x)\big\|\leq C(1+\|x\|)^{p-2}\quad\text{for all } x\in H.$$

Moreover, by Equation (2.17), there exist constants $a\geq 0$ and $h_0>0$ such that

$$\varphi(u^2)>1+\frac{C}{2}\Big(1+|\mu|_p+\sqrt{h_0}u\Big)^{p-2}u^2\quad\text{for all } u\in(\sqrt{a},\infty).\tag{5.9}$$

Let $h\in(0,h_0]$, $\nu\in\mathcal{P}_p$ with

$$\int_H f(y)\,\mu(\mathrm{d}y)\leq\mathcal{I}(h)f\leq h+\int_H f(z)\,\nu(\mathrm{d}z)-\varphi_h\big(\mathcal{W}_p^{\mathrm{Mart}}(\mu,\nu)^2\big),$$

and $\pi\in\mathrm{Cpl}(\mu,\nu)$ be an optimal martingale coupling between μ and ν. Using Taylor's theorem,

$$\begin{aligned}\varphi\left(\frac{\mathcal{W}_p^{\mathrm{Mart}}(\mu,\nu)^2}{h}\right)&\leq 1+\int_{H\times H}\frac{f(z)-f(y)}{h}\,\pi(\mathrm{d}y,\mathrm{d}z)\\&=1+\frac{1}{h}\bigg(\int_{H\times H}\big\langle\nabla f(y),z-y\big\rangle\,\pi(\mathrm{d}y,\mathrm{d}z)\\&\quad+\int_{H\times H}\int_0^1(1-t)\big\langle\nabla^2 f\big(y+t(z-y)\big)(z-y),z-y\big\rangle\,\mathrm{d}t\,\pi(\mathrm{d}y,\mathrm{d}z)\bigg).\end{aligned}$$

Thanks to the martingale constraint, the first term in the Taylor expansion vanishes, so that, using Fubini's theorem, Hölder's inequality, and Minkowski's inequality, we obtain

$$\begin{aligned}\varphi\left(\frac{\mathcal{W}_p^{\mathrm{Mart}}(\mu,\nu)^2}{h}\right)&\leq 1+\frac{1}{h}\int_0^1(1-t)\int_{H\times H}\big|\nabla^2 f(y+t(z-y))\big|_{\mathrm{Max}}\|z-y\|^2\,\pi(\mathrm{d}y,\mathrm{d}z)\,\mathrm{d}t\\&\leq 1+\int_0^1(1-t)\left(\int_{H\times H}\big|\nabla^2 f(y+t(z-y))\big|_{\mathrm{Max}}^{\frac{p}{p-2}}\,\pi(\mathrm{d}y,\mathrm{d}z)\right)^{\frac{p-2}{p}}\mathrm{d}t\,\frac{\mathcal{W}_p^{\mathrm{Mart}}(\mu,\nu)^2}{h}\\&\leq 1+\frac{C}{2}\left(\int_{H\times H}\big(1+\|y\|+\|z-y\|\big)^p\,\pi(\mathrm{d}y,\mathrm{d}z)\right)^{\frac{p-2}{p}}\frac{\mathcal{W}_p^{\mathrm{Mart}}(\mu,\nu)^2}{h}\\&\leq 1+\frac{C}{2}\left(1+|\mu|_p+\sqrt{h_0}\frac{\mathcal{W}_p^{\mathrm{Mart}}(\mu,\nu)}{\sqrt{h}}\right)^{p-2}\frac{\mathcal{W}_p^{\mathrm{Mart}}(\mu,\nu)^2}{h}.\end{aligned}$$

By Equation (5.9), it follows that $\frac{\mathcal{W}_p^{\mathrm{Mart}}(\mu,\nu)}{\sqrt{h}}\leq\sqrt{a}$. □

Lemma 5.4. *Let Θ be dense in $L_p(\mu;H)$ with $0\in\Theta$. Then,*

$$\mathcal{I}^{\mathrm{Mart}}_{\Theta}(h)f=\mathcal{I}^{\mathrm{Mart}}_{L_p(\mu;H)}(h)f\quad \text{for all } f\in \mathrm{Lip}_p \text{ and } h>0.$$

Proof. We proceed exactly as in the proof of Lemma 5.2. Let $h>0$ and $f\in C^2_p$. Again, $\mathcal{I}^{\mathrm{Mart}}_{\Theta}(h)f\le\mathcal{I}^{\mathrm{Mart}}_{L_p(\mu;H)}(h)f$, since $\Theta\subset L_p(\mu;H)$. Let $\varepsilon>0$ and $\theta_0\in L_p(\mu;H)$ with

$$\mathcal{I}^{\mathrm{Mart}}_{L_p(\mu;H)}(h)f\le\int_H f(z)\,\mu^{\mathrm{Mart}}_{\theta_0}(\mathrm{d}z)-\varphi_h(\|\theta_0\|^2_{L_p(\mu;H)})+\frac{\varepsilon}{3}.$$

By dominated convergence, we first find $\lambda\in(0,1)$ with

$$\int_H f(z)\,\mu^{\mathrm{Mart}}_{\theta_0}(\mathrm{d}z)\le\int_H f(z)\,\mu^{\mathrm{Mart}}_{\lambda\theta_0}(\mathrm{d}z)+\frac{\varepsilon}{3},$$

and then $\theta\in\Theta$ with $\|\theta\|_{L_p(\mu;H)}\le\|\theta_0\|_{L_p(\mu;H)}$ and

$$\int_H f(z)\,\mu^{\mathrm{Mart}}_{\lambda\theta_0}(\mathrm{d}z)\le\int_H f(z)\,\mu^{\mathrm{Mart}}_{\theta}(\mathrm{d}z)+\frac{\varepsilon}{3},$$

where we used the fact that Θ is dense in $L_p(\mu;H)$ and $0\in\Theta$. Since φ is nondecreasing,

$$\mathcal{I}^{\mathrm{Mart}}_{L_p(\mu;H)}(h)f\le\int_H f(z)\,\mu^{\mathrm{Mart}}_{\theta}(\mathrm{d}z)-\varphi_h(\|\theta\|^2_{L_p(\mu;H)})+\varepsilon\le\mathcal{I}^{\mathrm{Mart}}_{\Theta}(h)f+\varepsilon.$$

Letting $\varepsilon\downarrow 0$ the claim follows. □

Proof of Theorem 2.9. We use a similar approach as in the proof of Theorem 2.5, providing a bound from above for $\mathcal{I}^{\mathrm{Mart}}(h)f-\mu f$ and a bound from below for $\mathcal{I}^{\mathrm{Mart}}_{\Theta}(h)f-\mu f$ in the limit $h\downarrow 0$. We start with the bound from above writing a Taylor expansion for f as we did in the proof of Lemma 5.3. To that end, let $a\ge 0$ and $h_0>0$ as in Lemma 5.3. For all $h\in(0,h_0]$, let $\nu_h\in\mathcal{M}_p(\mu)$ with $\mathcal{W}^{\mathrm{Mart}}_p(\mu,\nu_h)\le\sqrt{ah}$ and

$$\mathcal{I}^{\mathrm{Mart}}(h)f\le h^2+\int_H f(z)\,\nu_h(\mathrm{d}z)-\varphi_h\big(\mathcal{W}^{\mathrm{Mart}}(\mu,\nu_h)^2\big),$$

and $\pi_h\in\mathrm{Mart}(\mu,\nu)$ be a martingale coupling between μ and ν_h that attains the infimum in definition of the p-martingale Wasserstein distance between μ and ν_h, cf. [9, Theorem 1.7]. Then, for all $h\in(0,h_0]$,

$$\begin{aligned}\frac{\mathcal{I}^{\mathrm{Mart}}(h)f-\mu f}{h}&\le h+\frac{1}{h}\int_{H\times H}f(z)-f(y)\,\pi_h(\mathrm{d}y,\mathrm{d}z)-\varphi\bigg(\frac{\mathcal{W}^{\mathrm{Mart}}_p(\mu,\nu_h)^2}{h}\bigg)\\&=h+\frac{1}{h}\int_{H\times H}\int_0^1(1-t)\big\langle\nabla^2 f\big(y+t(z-y)\big)(z-y),z-y\big\rangle\,\mathrm{d}t\,\pi_h(\mathrm{d}y,\mathrm{d}z)\\&\quad-\varphi\bigg(\frac{\mathcal{W}^{\mathrm{Mart}}_p(\mu,\nu_h)^2}{h}\bigg),\end{aligned}$$

where, thanks to the martingale constraint, the first order term in the Taylor expansion vanishes as in the proof of Lemma 5.3.

In the second order term, we renormalize the vector $z-y$ multiplying and dividing by $\|z-y\|$. Then, using Fubini's theorem and Hölder's inequality, we obtain, for all $h \in (0, h_0]$,

$$\begin{aligned}
&\frac{1}{h}\int_{H\times H}\int_0^1 (1-t)\big\langle \nabla^2 f\big(y+t(z-y)\big)(z-y), z-y\big\rangle \,\mathrm{d}t\,\pi_h(\mathrm{d}y,\mathrm{d}z)\\
&\qquad\le \frac{1}{h}\int_0^1\int_{H\times H}(1-t)\big|\nabla^2 f(y+t(z-y))\big|_{\mathrm{Max}}\|z-y\|^2\,\pi_h(\mathrm{d}y,\mathrm{d}z)\,\mathrm{d}t\\
&\qquad\le \int_0^1 (1-t)\left(\int_{H\times H}\big|\nabla^2 f(y+t(z-y))\big|_{\mathrm{Max}}^{\frac{p}{p-2}}\,\pi_h(\mathrm{d}y,\mathrm{d}z)\right)^{\frac{p-2}{p}}\mathrm{d}t\,\frac{\mathcal{W}_p^{\mathrm{Mart}}(\mu,\nu_h)^2}{h}.
\end{aligned}$$

Substituting the last inequality in the previous computations and passing to the convex conjugate of φ, we find that

$$\frac{\mathcal{I}^{\mathrm{Mart}}(h)f-\mu f}{h}\le h+\varphi^*\left(\int_0^1(1-t)\left(\int_{H\times H}\big|\nabla^2 f(y+t(z-y))\big|_{\mathrm{Max}}^{\frac{p}{p-2}}\,\pi_h(\mathrm{d}y,\mathrm{d}z)\right)^{\frac{p-2}{p}}\mathrm{d}t\right)$$

for all $h\in(0,h_0]$. By assumption, f is an element of C_p^2, so that $|\nabla^2 f|_{\mathrm{Max}}\colon H\to\mathbb{R}$ is continuous, and there exists a constant $C\ge 0$ such that

$$\big|\nabla^2 f(x)\big|_{\mathrm{Max}}^{\frac{p}{p-2}}\le\|\nabla^2 f(x)\|^{\frac{p}{p-2}}\le C(1+\|x\|)^p\quad\text{for all } x\in H.$$

Moreover,

$$\left(\int_{H\times H}\|y-z\|^p\,\pi_h(\mathrm{d}\mu,\mathrm{d}\nu_h)\right)^{1/p}=\mathcal{W}_p^{\mathrm{Mart}}(\mu,\nu_h)\le\sqrt{ah}\quad\text{for all } h\in(0,h_0].$$

Therefore, π_h concentrates on the diagonal as $h\downarrow 0$, i.e.,

$$\int_{H\times H}g(y,z)\,\pi_h(\mathrm{d}y,\mathrm{d}z)\to\int_{H\times H}g(y,y)\,\mu(\mathrm{d}y)\quad\text{as } h\downarrow 0$$

for all continuous functions $g\colon H\times H\to\mathbb{R}$ with $|g(x_1,x_2)|\le M(1+\|x_1\|+\|x_2\|)^p$ for all $x_1,x_2\in H$ and some constant $M\ge 0$, cf. [32, Theorem 6.9], and the continuity of φ^* implies that

$$\limsup_{h\downarrow 0}\frac{\mathcal{I}^{\mathrm{Mart}}(h)f-\mu f}{h}\le\varphi^*\left(\frac{1}{2}\left(\int_{H\times H}\big|\nabla^2 f(y)\big|_{\mathrm{Max}}^{\frac{p}{p-2}}\,\mu(\mathrm{d}y)\right)^{\frac{p-2}{p}}\right).$$

For the bound from below, let $\varepsilon>0$ and $v\colon H\to\mathbb{S}_0$ satisfying (2.23) in Remark 2.10, and define

$$\theta(x):=v(x)\big(\big|\nabla^2 f(x)\big|_{\mathrm{Max}}\big)^{\frac{1}{p-2}}.\tag{5.10}$$

Since $f\in C_p^2$, it follows that $\theta\in L_p(\mu;H)$, so that, for every $b\ge 0$,

$$\frac{\mathcal{I}_\Theta^{\mathrm{Mart}}(h)f-\mu f}{h}\ge\int_H\int_{\mathbb{R}}\frac{f\big(y+s\sqrt{bh}\theta(y)\big)-f(y)}{h}\,B_{\mathrm{sym}}(\mathrm{d}s)\,\mu(\mathrm{d}y)-\varphi\Big(b\|\theta\|_{L_p(\mu;H)}^2\Big).$$

Using Fatou's lemma, this leads to

$$\begin{aligned}
\liminf_{h\downarrow 0}\frac{\mathcal{I}_\Theta^{\mathrm{Mart}}(h)f-\mu f}{h}&\ge\int_H\frac{b}{2}|\nabla^2 f(y)|_{\mathrm{Max}}^{\frac{p}{p-2}}\,\mu(\mathrm{d}y)-\varphi\Big(b\|\theta\|_{L_p(\mu;H)}^2\Big)-\varepsilon\\
&=\frac{b}{2}\int_H|\nabla^2 f(y)|_{\mathrm{Max}}^{\frac{p}{p-2}}\,\mu(\mathrm{d}y)-\varphi\left(b\Big(\int_H|\nabla^2 f(y)|_{\mathrm{Max}}^{\frac{p}{p-2}}\,\mu(\mathrm{d}y)\Big)^{\frac{2}{p}}\right)-\varepsilon
\end{aligned}$$

for all $b \geq 0$. Taking the supremum over all $b \geq 0$, we can pass to the conjugate φ^* of φ, and obtain that

$$\liminf_{h \downarrow 0} \frac{\mathcal{I}_\Theta^{\mathrm{Mart}}(h)f - \mu f}{h} \geq \varphi^* \left(\frac{1}{2} \left(\int_H |\nabla^2 f(y)|_{\mathrm{Max}}^{\frac{p}{p-2}} \, \mu(\mathrm{d}y) \right)^{\frac{p-2}{p}} \right) - \varepsilon.$$

Taking the limit $\varepsilon \downarrow 0$ and invoking Lemma 5.4 completes the proof. □

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Appendix A. A simple but useful approximation result

Proposition A.1. *Let $(\Omega, \mathcal{F}, \mu)$ be a σ-finite measure space, $(S_n)_{n\in\mathbb{N}} \subset \mathcal{F}$ with $\mu(S_n) < \infty$ for all $n \in \mathbb{N}$ and $\Omega = \bigcup_{n\in\mathbb{N}} S_n$, X a Banach space, endowed with the Borel σ-algebra, and $\mathcal{A}$ an $\cap$-stable generator of $\mathcal{F}$ with $\Omega \in \mathcal{A}$. Let $p \in [1, \infty)$ and $L \subset L_p(\mu)$ with $\mathbb{1}_{A\cap S_n} \in \overline{\operatorname{spn} L}$ (closure of $\operatorname{spn} L$ in $L_p(\mu)$) for all $A \in \mathcal{A}$ and $n \in \mathbb{N}$. Then,*

$$\overline{X\text{-}\operatorname{spn} L} = L_p(\mu; X),$$

where $X\text{-}\operatorname{spn} L := \operatorname{spn}\{x \cdot f \,|\, x \in X,\, f \in L\}$.

Proof. Let $\mathcal{D} := \{B \in \mathcal{F} \,|\, \mathbb{1}_{B\cap A_n} \in \overline{\operatorname{spn} L} \text{ for all } n \in \mathbb{N}\}$. By assumption, $A \subset \mathcal{D}$ and, in particular, $\Omega \in \mathcal{D}$. Let $B \in \mathcal{D}$ and $n \in \mathbb{N}$. Then, there exist sequences $(f_k)_{k\in\mathbb{N}} \subset \operatorname{spn} L$ and $(g_k)_{k\in\mathbb{N}} \subset \operatorname{spn} L$ with

$$\lim_{k\to\infty} \|f_k - \mathbb{1}_{S_n}\|_{L_p(\mu)} = 0 \quad \text{and} \quad \lim_{k\to\infty} \|g_k - \mathbb{1}_{B\cap S_n}\|_{L_p(\mu)} = 0.$$

Then, $f_k - g_k \in \operatorname{spn} L$ for all $k \in \mathbb{N}$, and

$$\|f_k - g_k - \mathbb{1}_{B^c\cap S_n}\|_{L_p(\mu)} \leq \|f_k - \mathbb{1}_{S_n}\|_{L_p(\mu)} + \|g_k - \mathbb{1}_{B\cap S_n}\|_{L_p(\mu)} \to 0 \quad \text{as } k \to \infty.$$

Now, let $(A_k)_{k\in\mathbb{N}} \subset \mathcal{D}$ pairwise disjoint, $\varepsilon > 0$, $n \in \mathbb{N}$, $A_k^n := A_k \cap S_n$ for all $k \in \mathbb{N}$, and $A^n := \dot{\bigcup}_{k\in\mathbb{N}} A_k^n$. By σ-additivity of μ, there exists some $k \in \mathbb{N}$ with

$$\left\| \mathbb{1}_{A^n} - \sum_{i=1}^k \mathbb{1}_{A_i^n} \right\|_{L_p(\mu)} < 2^{-k}\varepsilon.$$

On the other hand, there exists $f_1, \ldots, f_n \in \operatorname{spn} L$ with $\|f_i - \mathbb{1}_{A_i^n}\|_{L_p(\mu)} < 2^{-i}\varepsilon$ for all $i \in \{1, \ldots, k\}$. Therefore,

$$\left\| \mathbb{1}_{A^n} - \sum_{i=1}^k f_i \right\|_{L_p(\mu)} \leq \left\| \mathbb{1}_{A^n} - \sum_{i=1}^k \mathbb{1}_{A_i^n} \right\|_{L_p(\mu)} + \sum_{i=1}^k \left\| \mathbb{1}_{A_i^n} - f_i \right\|_{L_p(\mu)} < \varepsilon.$$

This shows that $\dot{\bigcup}_{k\in\mathbb{N}} A_k \in \mathcal{D}$. By Dynkin's lemma, $\mathcal{D} = \mathcal{F}$. Now the statement follows from the fact that $L_p(\mu; X)$ is the closure of $\operatorname{spn}\{x \cdot \mathbb{1}_{B\cap S_n} \,|\, x \in X,\, B \in \mathcal{F},\, n \in \mathbb{N}\}$, which is a consequence of the dominated convergence theorem, cf. [24, Proposition 1.2.5], and [24, Lemma 1.2.19(1)]. □

As a consequence of the previous proposition, we obtain the following approximation results.

Remark A.2. Let $(\Omega, \mathcal{F}, \mu)$ be a σ-finite measure space, X a Banach space, endowed with the Borel σ-algebra, and $p \in [1, \infty)$.

a) As a consequence of Proposition A.1, we obtain the following well-known approximation result, see, e.g., [21, Lemma A.1] for a proof in the special case $X = \mathbb{R}$: Let Ω be a metric space with the Borel σ-algebra $\mathcal{F}$ and a finite measure μ. Then, $X\text{-}\operatorname{spn}\operatorname{Lip}_b(\Omega)$ is dense in $L_p(\mu; X)$, where $\operatorname{Lip}_b(\Omega)$ denotes the space of all bounded Lipschitz continuous functions $\Omega \to \mathbb{R}$. In particular, if Ω or, alternatively, X is assumed to be separable, the space $\operatorname{Lip}_b(\Omega; X)$ of all bounded Lipschitz continuous functions $\Omega \to X$ is a dense subset of $L_p(\mu; X)$. In order to deduce this approximation result from Proposition A.1, recall that, for every open set $U \subset \Omega$, $\mathbb{1}_U$ can be obtained as the pointwise monotone limit of bounded Lipschitz continuous functions using the so-called inf-convolution

$$f_\delta(\omega) := \inf_{\omega_0 \in \Omega} \left(\mathbb{1}_U(\omega_0) + \frac{d(\omega, \omega_0)}{\delta} \right) \quad \text{for } \omega \in \Omega \text{ and } \delta > 0.$$

b) If $\Omega = H$ is a separable Hilbert space, $\mathcal{F}$ is the Borel σ-algebra, and every ball in H has finite measure under μ, then $X\text{-}\operatorname{spn} C^\infty_{b,b}(H)$ is dense in $L_p(\mu; X)$, where $C^\infty_{b,b}(H)$ denotes the space of all bounded infinitely smooth functions $H \to \mathbb{R}$ with bounded support. In particular, the space $C^\infty_{b,b}(H; X)$ of all bounded infinitely smooth functions $H \to X$ with bounded support is dense in $L_p(\mu; X)$. This approximation result follows directly from Proposition A.1 together with the fact that the indicator function of every open ball with radius $r > 0$ and barycenter $x_0 \in H$ is the pointwise monotone limit of the functions

$$f_\delta(x) := \begin{cases} \exp\left(-\frac{\delta \|x - x_0\|^2}{r^2 - \|x - x_0\|^2}\right), & \|x - x_0\| < r, \\ 0, & \|x - x_0\| \geq r, \end{cases} \quad \text{for } x \in H \text{ and } \delta > 0,$$

and finite intersections of such balls are monotone limits of finite products of such functions.

c) Let $\Omega = \mathbb{R}$, $\mathcal{F}$ be the Borel σ-algebra, and μ be a finite measure. Then, for

$$L_0 := \left\{ 1_{[K,\infty)} \,\middle|\, K \in \mathbb{R} \right\},$$

the space $X\text{-}\operatorname{spn} L_0$ is dense in $L_p(\mu; X)$. If, additionally, $\int_{\mathbb{R}} |y|^p \, \mu(\mathrm{d}y) < \infty$ and

$$L_1 := \left\{ \left[x \mapsto (x - K)^+ \right] \,\middle|\, K \in \mathbb{R} \right\},$$

then also $X\text{-}\operatorname{spn} L_1$ is dense in $L_p(\mu; X)$.

d) Let μ be a finite measure and $\mathcal{A}$ an $\cap$-stable generator of $\mathcal{F}$. Then, $X\text{-}\operatorname{spn}\{\mathbb{1}_A \mid A \in \mathcal{A}\}$ is dense in $L_p(\mu; X)$. Note that this approximation holds for an arbitrary $\cap$-stable generator $\mathcal{A}$ without any additional assumptions. As a consequence, for $X = \mathbb{R}$, one obtains the well-known Carathéodory approximation: if μ is a finite measure and $\mathcal{A}$ is an algebra with $\sigma(\mathcal{A}) = \mathcal{F}$, then, for all $B \in \mathcal{F}$ and all $\varepsilon > 0$, there exists some $A \in \mathcal{A}$ with

$$\mu(A \Delta B) = \|\mathbb{1}_A - \mathbb{1}_B\|_{L_1(\mu)} < \varepsilon.$$

Indeed, let $f \in \operatorname{spn}\{\mathbb{1}_A \mid A \in \mathcal{A}\}$ with $\|\mathbb{1}_B - f\|_{L_1(\mu)} < \varepsilon/2$. Then, since $\mathcal{A}$ is an algebra, $A := \{\omega \in \Omega \mid f(\omega) > 1 - \varepsilon/2\}$ is an element of $\mathcal{A}$, and

$$\mu(A \Delta B) = \mu(A \setminus B) + \mu(B \setminus A) < \varepsilon.$$

References

[1] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

[2] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Math. Finance*, 9(3):203–228, 1999.

[3] D. Bartl, S. Drapeau, J. Obłój, and J. Wiesel. Sensitivity analysis of Wasserstein distributionally robust optimization problems. *Proc. R. Soc. A.*, 477(2256):Paper No. 20210176, 19, 2021.

[4] D. Bartl, S. Drapeau, J. Obłój, and J. Wiesel. Supplementary material from "Sensitivity analysis of Wasserstein distributionally robust optimization problems". 2021.
[5] D. Bartl, S. Drapeau, and L. Tangpi. Computational aspects of robust optimized certainty equivalents and option pricing. *Math. Finance*, 30(1):287–309, 2020.
[6] D. Bartl, S. Eckstein, and M. Kupper. Limits of random walks with distributionally robust transition probabilities. *Electron. Commun. Probab.*, 26:Paper No. 28, 13, 2021.
[7] D. Bartl and J. Wiesel. Sensitivity of multi-period optimization problems in adapted Wasserstein distance. *Preprint*, arXiv:2208.05656, 2022.
[8] M. Beiglböck, P. Henry-Labordère, and F. Penkner. Model-independent bounds for option prices—a mass transport approach. *Finance Stoch.*, 17(3):477–501, 2013.
[9] M. Beiglböck and N. Juillet. On a problem of optimal transport under marginal martingale constraints. *Ann. Probab.*, 44(1):42–106, 2016.
[10] D. P. Bertsekas and S. E. Shreve. *Stochastic optimal control*, volume 139 of *Mathematics in Science and Engineering*. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1978. The discrete time case.
[11] V. I. Bogachev. *Gaussian measures*, volume 62 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1998.
[12] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Physica-Verlag/Springer, Heidelberg, 2010.
[13] S. Bozinovski. Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica*, 44, 2020.
[14] R. Cont. Model uncertainty and its impact on the pricing of derivative instruments. *Math. Finance*, 16(3):519–547, 2006.
[15] R. M. Dudley. *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.
[16] S. Eckstein and M. Kupper. Computation of optimal transport and related hedging problems via penalization and neural networks. *Appl. Math. Optim.*, 83(2):639–667, 2021.
[17] S. Eckstein, M. Kupper, and M. Pohl. Robust risk aggregation with neural networks. *Math. Finance*, 30(4):1229–1272, 2020.
[18] H. Föllmer and A. Schied. *Stochastic finance*. De Gruyter Graduate. De Gruyter, Berlin, 2016. An introduction in discrete time, Fourth revised and extended edition.
[19] S. Fuhrmann, M. Kupper, and M. Nendel. Wasserstein perturbations of Markovian transition semigroups. *Preprint*, arXiv:2105.05655, 2021.
[20] R. Gao and A. J. Kleywegt. Distributionally Robust Stochastic Optimization with Wasserstein distance, 2016.
[21] S. Hanneke, A. Kontorovich, S. Sabato, and R. Weiss. Universal Bayes consistency in metric spaces. *Ann. Statist.*, 49(4):2129–2150, 2021.
[22] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.
[23] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.
[24] T. Hytönen, J. van Neerven, M. Veraar, and L. Weis. *Analysis in Banach spaces. Vol. I. Martingales and Littlewood-Paley theory*, volume 63 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer, Cham, 2016.
[25] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Confernence on Learning Representations*, 2014.
[26] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Math. Program.*, 171(1-2, Ser. A):115–166, 2018.
[27] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.
[28] J. Obłój and J. Wiesel. Robust estimation of superhedging prices. *Ann. Statist.*, 49(1):508–530, 2021.
[29] H. Owhadi and C. Scovel. Extreme points of a ball about a measure with finite support. *Commun. Math. Sci.*, 15(1):77–96, 2017.
[30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep

learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
[31] G. Pflug and D. Wozabal. Ambiguity in portfolio selection. *Quant. Finance*, 7(4):435–442, 2007.
[32] C. Villani. *Optimal transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.
[33] D. Wozabal. A framework for optimization under ambiguity. *Ann. Oper. Res.*, 193:21–47, 2012.
[34] C. Zhao and Y. Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Oper. Res. Lett.*, 46(2):262–267, 2018.

Center for Mathematical Economics, Bielefeld University
Email address: max.nendel@uni-bielefeld.de

Center for Mathematical Economics, Bielefeld University
Email address: alessandro.sgarabottolo@uni-bielefeld.de