PROCEEDINGS A

royalsocietypublishing.org/journal/rspa

Research

Cite this article: Bartl D, Drapeau S, Obłój J, Wiesel J. 2021 Sensitivity analysis of Wasserstein distributionally robust optimization problems. *Proc. R. Soc. A* **477**: 20210176.
https://doi.org/10.1098/rspa.2021.0176

Received: 25 February 2021
Accepted: 9 November 2021

Subject Areas:
applied mathematics, mathematical finance, statistics

Keywords:
robust stochastic optimization, sensitivity analysis, uncertainty quantification, non-parametric uncertainty, Wasserstein metric

Author for correspondence:
Jan Obłój
e-mail: jan.obloj@maths.ox.ac.uk

Sensitivity analysis of Wasserstein distributionally robust optimization problems

Daniel Bartl[1], Samuel Drapeau[2], Jan Obłój[3] and Johannes Wiesel[4]

[1]Department of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria
[2]School of Mathematical Sciences & Shanghai Advanced Institute of Finance, Shanghai Jiao Tong University, 211 West Huaihai Road, Shanghai 200030, People's Republic of China
[3]Mathematical Institute, University of Oxford, Woodstock Road, Oxford OX2 6GG, UK
[4]Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York,
NY 10027, USA

JO, 0000-0002-5686-5498

We consider sensitivity of a generic stochastic optimization problem to model uncertainty. We take a non-parametric approach and capture model uncertainty using Wasserstein balls around the postulated model. We provide explicit formulae for the first-order correction to both the value function and the optimizer and further extend our results to optimization under linear constraints. We present applications to statistics, machine learning, mathematical finance and uncertainty quantification. In particular, we provide an explicit first-order approximation for square-root LASSO regression coefficients and deduce coefficient shrinkage compared to the ordinary least-squares regression. We consider robustness of call option pricing and deduce a new Black–Scholes sensitivity, a non-parametric version of the so-called Vega. We also compute sensitivities of optimized certainty equivalents in finance and propose measures to quantify robustness of neural networks to adversarial examples.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5730987.

THE ROYAL SOCIETY
PUBLISHING

1. Introduction

We consider a generic stochastic optimization problem

$$\inf_{a\in\mathcal{A}}\int_{\mathcal{S}} f(x,a)\,\mu(\mathrm{d}x), \tag{1.1}$$

where $\mathcal{A}$ is the set of actions or choices, f is the loss function and μ is a probability measure over the state space $\mathcal{S}$. Such problems are found across the whole of applied mathematics. The measure μ is the crucial input and it could represent, for example, a dynamic model of the system, as is often the case in mathematical finance or mathematical biology, or the empirical measure of observed data points, or the training set, as is the case in statistics and machine learning applications. In virtually all the cases, there is a certain degree of uncertainty around the choice of μ coming from modelling choices and simplifications, incomplete information, data errors, finite sample error, etc. It is thus very important to understand the influence of changes in μ on (1.1), both on its value and on its optimizer. Often, the choice of μ is done in two stages: first a parametric family of models is adopted and then the values of the parameters are fixed. Sensitivity analysis of (1.1) with changing parameters is a classical topic explored in parametric programming and statistical inference, e.g. [1–3]. It also underscores a lot of progress in the field of uncertainty quantification, see [4]. Considering μ as an abstract parameter, the mathematical programming literature looked into qualitative and quantitative stability of (1.1). We refer to [5,6] and the references therein. When μ represents data samples, there has been a considerable interest in the optimization community in designing algorithms which are robust and, in particular, do not require excessive hypertuning, see [7] and the references therein.

A more systematic approach to model uncertainty in (1.1) is offered by the distributionally robust optimization problem

$$V(\delta):=\inf_{a\in\mathcal{A}} V(\delta,a):=\inf_{a\in\mathcal{A}}\sup_{\nu\in B_\delta(\mu)}\int_{\mathcal{S}} f(x,a)\,\nu(\mathrm{d}x), \tag{1.2}$$

where $B_\delta(\mu)$ is a ball of radius δ around μ in the space of probability measures, as specified below. Such problems greatly generalize more classical robust optimization and have been studied extensively in operations research and machine learning in particular; we refer the reader to [8] and the references therein. Our goal in this paper is to understand the behaviour of these problems for small δ. Our main results compute first-order behaviour of $V(\delta)$ and its optimizer for small δ. This offers a measure of sensitivity to errors in model choice and/or specification as well as points in the abstract direction, in the space of models, in which the change is most pronounced. We use examples to show that our results can be applied across a wide spectrum of science.

This paper is organized as follows. We first present the main results and then, in §3, explore their applications. Further discussion of our results and the related literature is found in §4, which is then followed by the proofs. The online appendix [9] contains many supplementary results and remarks, as well as some more technical arguments from the proofs.

2. Main results

Take $d,k\in\mathbb{N}$, endow $\mathbb{R}^d$ with the Euclidean norm $|\cdot|$ and write Γ^o for the interior of a set Γ. Assume that $\mathcal{S}$ is a closed convex subset of $\mathbb{R}^d$. Let $\mathcal{P}(\mathcal{S})$ denote the set of all (Borel) probability measures on $\mathcal{S}$. Further fix a seminorm $||\cdot||$ on $\mathbb{R}^d$ and denote by $||\cdot||_*$ its (extended) dual norm, i.e. $||y||_*:=\sup\{\langle x,y\rangle : ||x||\le 1\}$. In particular, for $||\cdot||=|\cdot|$ we also have $||\cdot||_*=|\cdot|$. For $\mu,\nu\in\mathcal{P}(\mathcal{S})$, we define the p-Wasserstein distance as

$$W_p(\mu,\nu)=\inf\left\{\int_{\mathcal{S}\times\mathcal{S}} ||x-y||_*^p\,\pi(\mathrm{d}x,\mathrm{d}y):\pi\in\mathrm{Cpl}(\mu,\nu)\right\}^{1/p},$$

where $\mathrm{Cpl}(\mu,\nu)$ is the set of all probability measures π on $\mathcal{S}\times\mathcal{S}$ with first marginal $\pi_1 := \pi(\cdot\times\mathcal{S}) = \mu$ and second marginal $\pi_2 := \pi(\mathcal{S}\times\cdot) = \nu$. Denote the Wasserstein ball

$$B_\delta(\mu) = \{\nu\in\mathcal{P}(\mathcal{S}) : W_p(\mu,\nu)\le\delta\},$$

of size $\delta \ge 0$ around μ. Note that, taking a suitable probability space $(\Omega,\mathcal{F},\mathbb{P})$ and a random variable $X\sim\mu$, we have the following probabilistic representation of $V(\delta,a)$:

$$\sup_{\nu\in B_\delta(\mu)}\int_{\mathcal{S}} f(x,a)\,\nu(\mathrm{d}x) = \sup_Z \mathbb{E}_{\mathbb{P}}[f(X+Z,a)],$$

where the supremum is taken over all Z satisfying $X+Z\in\mathcal{S}$ almost surely and $\mathbb{E}_{\mathbb{P}}[||Z||_*^p]\le\delta^p$. Wasserstein distances and optimal transport techniques have proved to be powerful and versatile tools in a multitude of applications, from economics [10,11] to image recognition [12]. The idea to use Wasserstein balls to represent model uncertainty was pioneered in [13] in the context of investment problems. When sampling from a measure with a finite pth moment, the measures converge to the true distribution and Wasserstein balls around the empirical measures have the interpretation of confidence sets, see [14]. In this set-up, the radius δ can then be chosen as a function of a given confidence level α and the sample size N. This yields finite sample guarantees and asymptotic consistency, see [15,16], and justifies the use of the Wasserstein metric to capture model uncertainty. The value $V(\delta,a)$ in (1.2) has a dual representation, see [17,18], which has led to significant new developments in distributionally robust optimization, e.g.[15,19–21].

Naturally, other choices for the distance on the space of measures are also possible: such as the Kullblack–Leibler divergence, see [22] for general sensitivity results and [23] for applications in portfolio optimization, or the Hellinger distance, see [24] for a statistical robustness analysis. We refer to §4 for a more detailed analysis of the state of the art in these fields. Both of these approaches have good analytic properties and often lead to theoretically appealing closed-form solutions. However, they are also very restrictive since any measure in the neighbourhood of μ has to be absolutely continuous with respect to μ. In particular, if μ is the empirical measure of N observations then measures in its neighbourhood have to be supported on those fixed N points. To obtain meaningful results, it is thus necessary to impose additional structural assumptions, which are often hard to justify solely on the basis of the data at hand and, equally importantly, create another layer of model uncertainty themselves. We refer to [17, sec. 1.1] for further discussion of potential issues with ϕ-divergences. The Wasserstein distance, while harder to handle analytically, is more versatile and does not require any such additional assumptions.

Throughout the paper, we take the convention that continuity and closure are understood w.r.t. $|\cdot|$. We assume that $\mathcal{A}\subset\mathbb{R}^k$ is convex and closed and that the seminorm $||\cdot||$ is strictly convex in the sense that for two elements $x,y\in\mathbb{R}^d$ with $||x||=||y||=1$ and $||x-y||\neq 0$, we have $||\frac{1}{2}x+\frac{1}{2}y||<1$ (note that this is satisfied for every l^s-norm $|x|_s := (\sum_{i=1}^d |x_i|^s)^{1/s}$ for $s>1$). We fix $p\in(1,\infty)$, let $q := p/(p-1)$ so that $1/p+1/q=1$, and fix $\mu\in\mathcal{P}(\mathcal{S})$ such that the boundary of $\mathcal{S}\subset\mathbb{R}^d$ has μ–zero measure and $\int_{\mathcal{S}}|x|^p\,\mu(\mathrm{d}x)<\infty$. Denote by $\mathcal{A}_\delta^\star$ the set of optimizers for $V(\delta)$ in (1.2).

Assumption 2.1. The loss function $f:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ satisfies

- — $x\mapsto f(x,a)$ is differentiable on $\mathcal{S}^o$ for every $a\in\mathcal{A}$. Moreover, $(x,a)\mapsto\nabla_x f(x,a)$ is continuous and for every $r>0$ there is $c>0$ such that $|\nabla_x f(x,a)|\le c(1+|x|^{p-1})$ for all $x\in\mathcal{S}$ and $a\in\mathcal{A}$ with $|a|\le r$.
- — For all $\delta\ge 0$ sufficiently small, we have $\mathcal{A}_\delta^\star\neq\emptyset$ and for every sequence $(\delta_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty}\delta_n=0$ and $(a_n^\star)_{n\in\mathbb{N}}$ such that $a_n^\star\in\mathcal{A}_{\delta_n}^\star$ for all $n\in\mathbb{N}$ there is a subsequence which converges to some $a^\star\in\mathcal{A}_0^\star$.

The above assumption is not restrictive: the first part merely ensures existence of $||\nabla_x f(\cdot,a^\star)||_{L^q(\mu)}$, while the second part is satisfied as soon as either $\mathcal{A}$ is compact or $V(0,\cdot)$ is coercive, which is the case in most examples of interest; see [9, lemma 7.15] for further comments.

Theorem 2.2. *If assumption 2.1 holds then $V'(0)$ is given by*

$$\Upsilon := \lim_{\delta\to 0}\frac{V(\delta)-V(0)}{\delta} = \inf_{a^\star\in\mathcal{A}_0^\star}\left(\int_{\mathcal{S}} \|\nabla_x f(x,a^\star)\|^q\,\mu(dx)\right)^{1/q}.$$

Remark. Inspecting the proof, defining

$$\tilde{V}(\delta) = \inf_{a^\star\in\mathcal{A}_0^\star}\ \sup_{\nu\in B_\delta(\mu)}\int_{\mathcal{S}} f(x,a^\star)\,\nu(\mathrm{d}x)$$

we obtain $\tilde{V}'(0)=V'(0)$. This means that for small $\delta>0$ there is no first-order gain from optimizing over all $a\in\mathcal{A}$ in the definition of $V(\delta)$ when compared with restricting simply to $a^\star\in\mathcal{A}_0^\star$, as in $\tilde{V}(\delta)$.

The above result naturally extends to computing sensitivities of robust problems, i.e. $V'(r)$, see [9, corollary 7.5], as well as to the case of stochastic optimization under linear constraints, see [9, theorem 7.7]. We recall that $V(0,a)=\int_{\mathcal{S}} f(x,a)\,\mu(\mathrm{d}x)$.

Assumption 2.3. Suppose the f is twice continuously differentiable, $a^\star\in\mathcal{A}_0^\star\cap\mathcal{A}^o$ and

- $\sum_{i=1}^k |\nabla_{a_i}\nabla_x f(x,a)| \le c(1+|x|^{p-1-\varepsilon})$ for some $\varepsilon>0$, $c>0$, all $x\in\mathcal{S}$ and all a close to $a^\star$.
- The function $a\mapsto V(0,a)$ is twice continuously differentiable in the neighbourhood of $a^\star$ and the matrix $\nabla_a^2 V(0,a^\star)$ is invertible.

Theorem 2.4. *Suppose $a^\star\in\mathcal{A}_0^\star$ and $a_\delta^\star\in\mathcal{A}_\delta^\star$ such that $a_\delta^\star\to a^\star$ as $\delta\to 0$ and assumptions 2.1 and 2.3 are satisfied. If $\nabla_x f(x,a^\star)\neq 0$ μ-a.e. or if $\nabla_x\nabla_a f(x,a^\star)=0$ μ-a.e., then*

$$\beth := \lim_{\delta\to 0}\frac{a_\delta^\star-a^\star}{\delta} = -\left(\int_{\mathcal{S}} \|\nabla_x f(x,a^\star)\|^q\,\mu(dx)\right)^{(1/q)-1}$$
$$\cdot\,(\nabla_a^2 V(0,a^\star))^{-1}\int_{\mathcal{S}}\frac{\nabla_x\nabla_a f(x,a^\star)\,h(\nabla_x f(x,a^\star))}{\|\nabla_x f(x,a^\star)\|^{1-q}}\,\mu(dx),$$

where $h:\mathbb{R}^d\setminus\{0\}\to\{x\in\mathbb{R}^d\ :\ \|x\|_*=1\}$ is the unique function satisfying $\langle\cdot,h(\cdot)\rangle=\|\cdot\|$, see [9, *Lemma* 6.2]. *In particular, $h(\cdot)=\cdot/|\cdot|$ if $\|\cdot\|=|\cdot|$.*

Above and throughout the convention is that $\nabla_x f(x,a)\in\mathbb{R}^{d\times 1}$, $\nabla_{a_i}\nabla_x f(x,a)\in\mathbb{R}^{d\times 1}$, $\nabla_a f(x,a)\in\mathbb{R}^{k\times 1}$, $\nabla_x\nabla_a f(x,a)\in\mathbb{R}^{k\times d}$ and $0/0=0$. The assumed existence and convergence of optimizers holds, e.g. with suitable convexity of f in a; see [9, lemma 7.14] for a worked out setting. In line with the financial economics practice, we gave our sensitivities letter symbols, Υ and $\beth$, loosely motivated by $\Upsilon\pi\acute{o}\delta\varepsilon\iota\gamma\mu\alpha$, the Greek for *Model*, and בקרה, the Hebrew for *control*.

3. Applications

We now illustrate the universality of theorems 2.2 and 2.4 by considering their applications in a number of different fields. Unless otherwise stated, $\mathcal{S}=\mathbb{R}^d$, $\mathcal{A}=\mathbb{R}^k$ and $\int$ means $\int_{\mathcal{S}}$.

(a) Financial economics

We start with the simple example of risk-neutral pricing of a call option written on an underlying asset $(S_t)_{t\le T}$. Here, $T,K>0$ are the maturity and the strike, respectively, $f(x,a)=(S_0x-K)^+$ and μ is the distribution of S_T/S_0. We set interest rates and dividends to zero for simplicity. In [25], the model μ is a lognormal distribution, i.e. $\log(S_T/S_0)\sim\mathcal{N}(-\sigma^2T/2,\sigma^2T)$ is Gaussian with mean $-\sigma^2T/2$ and variance σ^2T. In this case, $V(0)$ is given by the celebrated Black–Scholes formula.

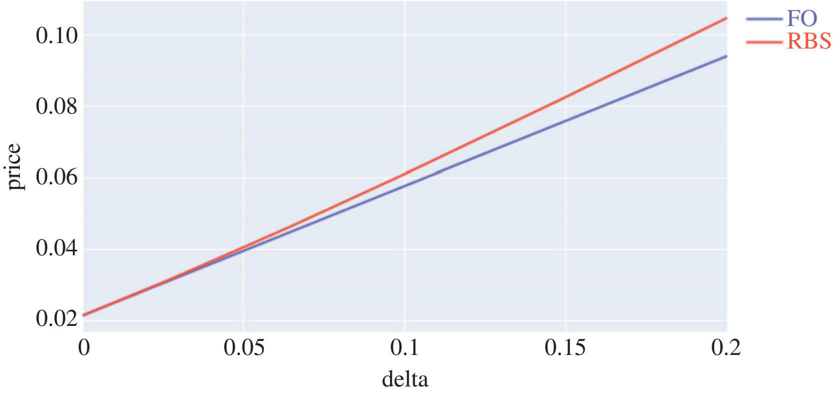

Figure 1. DRO value $\mathcal{R}\mathrm{BS}(\delta)$ versus the first order (FO) approximation $\mathcal{R}\mathrm{BS}(0)+\Upsilon\delta, S_0=T=1, K=1.2, \sigma=0.2$. (Online version in colour.)

Note that this example is particularly simple since f is independent of a. However, to ensure risk-neutral pricing, we have to impose a linear constraint on the measures in $B_\delta(\mu)$, giving

$$\sup_{\nu\in B_\delta(\mu):\int x\nu(\mathrm{d}x)=1}\int (S_0x-K)^+\nu(\mathrm{d}x). \tag{3.1}$$

To compute its sensitivity we encode the constraint using a Lagrangian and apply theorem 2.2, see [9, remark 7.3, theorem 7.7]. For $p=2$, letting $k=K/S_0$ and $\mu_k=\mu([k,\infty))$, the resulting formula, see [9, example 7.10], is given by

$$\Upsilon=S_0\sqrt{\int\left(\mathbf{1}_{x\geq k}-\mu_k\right)^2\mu(\mathrm{d}x)}=S_0\sqrt{\mu_k(1-\mu_k)}.$$

Let us specialize to the lognormal distribution of the Black–Scholes model above and denote the quantity in (3.1) as $\mathcal{R}BS(\delta)$. It may be computed exactly using methods from [26]. Figure 1 compares the exact value and the first-order approximation. We have $\Upsilon=S_0\sqrt{\Phi(d_-)(1-\Phi(d_-))}$, where $d_-=\log(S_0/K)-\sigma^2T/2/\sigma\sqrt{T}$ and Φ is the cdf of $\mathcal{N}(0,1)$ distribution. It is also insightful to compare Υ with a parametric sensitivity. If instead of Wasserstein balls, we consider $\{\mathcal{N}(-\tilde{\sigma}^2T/2,\tilde{\sigma}^2T):|\sigma-\tilde{\sigma}|\leq\delta\}$ the resulting sensitivity is known as the Black–Scholes Vega and given by $\mathcal{V}=S_0\Phi'(d_-+\sigma\sqrt{T})$. We plot the two sensitivities in figure 2. It is remarkable how, for the range of strikes of interest, the non-parametric model sensitivity Υ traces out the usual shape of $\mathcal{V}$ but shifted upwards to account for the idiosyncratic risk of departure from the lognormal family. More generally, given a book of options with payoff $f=f^+-f^-$ at time T, with $f^+,f^-\geq 0$, we could say that the book is Υ-neutral if the sensitivity Υ was the same for f^+ and for f^-. In analogy to Delta-Vega hedging standard, one could develop a non-parametric model-agnostic Delta-Upsilon hedging. We believe these ideas offer potential for exciting industrial applications and we leave them to further research.

We turn now to the classical notion of the optimized certainty equivalent (OCE) of [27]. It is a decision theoretic criterion designed to split a liability between today's and tomorrow's payments. It is also a convex risk measure in the sense of [28] and covers many of the popular risk measures such as expected shortfall or entropic risk, see [29]. We fix a convex monotone function $l:\mathbb{R}\to\mathbb{R}$ which is bounded from below and $g:\mathbb{R}^d\to\mathbb{R}$. Here, g represents the payoff of a financial position and l is the negative of a utility function, or a loss function. We take $||\cdot||=|\cdot|$ and refer to [9, lemma 7.14] for generic sufficient conditions for assumptions 2.1 and 2.3 to hold in this setup.

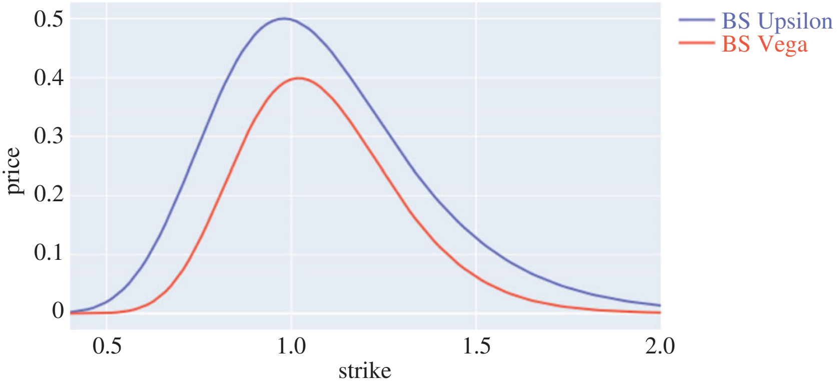

Figure 2. Black–Scholes model: Υ versus $\mathcal{V}$, $S_0 = T = 1, \sigma = 0.2$. (Online version in colour.)

The OCE corresponds to V in (1.1) for $f(x,a) = l(g(x) - a) + a$ and $\mathcal{A} = \mathbb{R}$, $\mathcal{S} = \mathbb{R}^d$. Theorems 2.2 and 2.4 yield the sensitivities

$$\Upsilon = \inf_{a^\star \in \mathcal{A}_0^\star} \left(\int \left| l'(g(x) - a^\star) \nabla g(x) \right|^q \mu(\mathrm{d}x) \right)^{1/q},$$

$$\beth = \left(\int |l'(g(x) - a^\star)\, \nabla g(x)|^2 \,\mu(\mathrm{d}z) \right)^{-1/2} \cdot \frac{\int l''(g(x) - a^\star)\, l'(g(x) - a^\star)\, (\nabla g(x))^2 \,\mu(\mathrm{d}x)}{\int l''(g(x) - a^\star)\, \mu(\mathrm{d}x)},$$

where, for simplicity, we took $p = q = 2$ for the latter.

A related problem considers hedging strategies which minimize the expected loss of the hedged position, i.e. $f(x,a) = l(g(x) + \langle a, x - x_0 \rangle)$, where $\mathcal{A} = \mathbb{R}^k$ and (x_0, x) represent today's and tomorrow's traded prices. We compute Υ as

$$\inf_{a^\star \in \mathcal{A}_0^\star} \left(\int \left| l'(g(x) + \langle a^\star, x - x_0 \rangle)(\nabla g(x) + a^\star) \right|^q \mu(\mathrm{d}x) \right)^{1/q}.$$

Furthermore we can combine loss minimization with OCE and consider $a = (H, m) \in \mathbb{R}^k \times \mathbb{R}$, $f(x,(h,m)) = l(g(x) + \langle H, x - x_0 \rangle + m) - m$. This gives $V'(0)$ as the infimum over $(H^\star, m^\star) \in \mathcal{A}_0^\star$ of

$$\left(\int \left| l'(g(x) + \langle H^\star, x - x_0 \rangle + m^\star)(\nabla g(x) + H^\star) \right|^q \mu(\mathrm{d}x) \right)^{1/q}.$$

The above formulae capture non-parametric sensitivity to model uncertainty for examples of key risk measurements in financial economics. To the best of our knowledge, this has not been achieved before.

Finally, we consider briefly the classical mean-variance optimization of [30]. Here μ represents the loss distribution across the assets and $a \in \mathbb{R}^d$, $\sum_{i=1}^d a_i = 1$ are the relative investment weights. The original problem is to minimize the sum of the expectation and γ standard deviations of returns $\langle a, X \rangle$, with $X \sim \mu$. Using the ideas in [31, Example 2] and considering measures on $\mathbb{R}^d \times \mathbb{R}^d$, we can recast the problem as (1.1). While [31] focused on the asymptotic regime $\delta \to \infty$, their non-asymptotic statements are related to our theorem 2.2 and either result could be used here to obtain that $V(\delta) \approx V(0) + \sqrt{1 - \gamma^2}\delta$ for small δ.

(b) Neural networks

We specialize now to quantifying robustness of neural networks (NN) to adversarial examples. This has been an important topic in machine learning since [32] observed that NN consistently misclassify inputs formed by applying small worst-case perturbations to a dataset. This produced a number of works offering either explanations for these effects or algorithms to create such adversarial examples, e.g. [33–39] to name just a few. The main focus of research works in this area, see [40], has been on faster algorithms for finding adversarial examples, typically leading to an overfit to these examples without any significant generalization properties. The viewpoint has been mainly pointwise, e.g. [32], with some generalizations to probabilistic robustness, e.g. [39].

In contrast, we propose a simple metric for measuring robustness of NN which is independent of the architecture employed and the algorithms for identifying adversarial examples. In fact, theorem 2.2 offers a simple and intuitive way to formalize robustness of NN: for simplicity consider a 1-layer neural network trained on a given distribution μ of pairs (x,y), i.e. $(A_1^\star, A_2^\star, b_1^\star, b_2^\star)$ solve

$$\inf \int |y - ((A_2(\cdot)+b_2)\circ\sigma\circ(A_1(\cdot)+b_1))(x)|^p\,\mu(\mathrm{d}x,\mathrm{d}y),$$

where the inf is taken over $a=(A_1,A_2,b_1,b_2)\in\mathcal{A}=\mathbb{R}^{k\times d}\times\mathbb{R}^{d\times k}\times\mathbb{R}^k\times\mathbb{R}^d$, for a given activation function $\sigma:\mathbb{R}\to\mathbb{R}$, where the composition above is understood componentwise. Set $f(x,y;A,b):=|y-(A_2(\cdot)+b_2)\circ\sigma\circ(A_1(\cdot)+b_1)(x)|^p$. Data perturbations are captured by $\nu\in B_\delta^p(\mu)$ and (1.2) offers a robust training procedure. The first-order quantification of the NN sensitivity to adversarial data is then given by

$$\left(\int |\nabla f(x,y;A^\star,b^\star)|^q\,\mu(\mathrm{d}x,\mathrm{d}y)\right)^{1/q}.$$

A similar viewpoint, capturing robustness to adversarial examples through the optimal transport lens, has been recently adopted by other authors. The dual formulation of (1.2) was used by [21] to reduce the training of neural networks to tractable linear programs. [41] modified (1.2) to consider a penalized problem $\inf_{a\in\mathcal{A}}\sup_{\nu\in\mathcal{P}(\mathcal{S})}\int_{\mathcal{S}} f(x,a)\,\nu(\mathrm{d}x)-\gamma W_p(\mu,\nu)$ to propose new stochastic gradient descent algorithms with inbuilt robustness to adversarial data.

(c) Uncertainty quantification

In the context of UQ, the measure μ represents input parameters of a (possibly complicated) operation G in a physical, engineering or economic system. We consider the so-called *reliability* or *certification problem*: for a given set E of undesirable outcomes, one wants to control $\sup_{\nu\in\mathcal{P}}\nu(G(x)\in E)$, for a set of probability measures $\mathcal{P}$. The distributionally robust adversarial classification problem considered recently by [42] is also of this form, with Wasserstein balls $\mathcal{P}$ around an empirical measure of N samples. Using the dual formulation of [18], they linked the problem to minimization of the conditional value-at-risk and proposed a reformulation, and numerical methods, in the case of linear classification. We propose instead a regularized version of the problem and look for

$$\delta(\alpha):=\sup\left\{\delta\geq 0:\ \inf_{\nu\in B_\delta(\mu)}\int \mathrm{d}(G(x),E)\,\nu(\mathrm{d}x)\geq\alpha\right\},$$

for a given safety level α. We thus consider the average distance to the undesirable set, $d(G(x),E):=\inf_{e\in E}|G(x)-e|$, and not just its probability. The quantity $\delta(\alpha)$ could then be used to quantify the implicit uncertainty of the certification problem, where higher δ corresponds to less uncertainty. Taking statistical confidence bounds of the empirical measure in Wasserstein distance into account, see [14], δ would then determine the minimum number of samples needed to estimate the empirical measure.

Assume that E is convex. Then $x\mapsto d(x,E)$ differentiable everywhere except at the boundary of E with $\nabla_x d(x,E)=0$ for $x\in E^o$ and $|\nabla_x d(x,E)|=1$ for all $x\in\bar{E}^c$. Furthermore, assume μ is absolutely continuous w.r.t. Lebesgue measure on $\mathcal{S}$. Theorem 2.2, using [9, remark 7.3], gives a first-order expansion for the above problem:

$$\inf_{\nu\in B_\delta(\mu)}\int \mathrm{d}(G(x),E)\,\nu(\mathrm{d}x)=\int \mathrm{d}(G(x),E)\,\mu(\mathrm{d}x)-\left(\int |\nabla_x d(G(x),E)\nabla_x G(x)|^q\,\mu(\mathrm{d}x)\right)^{1/q}\delta+o(\delta).$$

In the special case $\nabla_x G(x) = cI$ this simplifies to

$$\int d(G(x), E)\,\mu(\mathrm{d}x) - c(\mu(G(x) \notin E))^{1/q}\delta + o(\delta),$$

and the minimal measure ν pushes every point $G(x)$ not contained in E in the direction of the orthogonal projection. This recovers the intuition of [43, theorem 1], which in turn relies on [17, corollary 2, example 7]. Note however that our result holds for general measures μ. We also note that such an approximation could provide an ansatz for dimension reduction, by identifying the dimensions for which the partial derivatives are negligible and then projecting G on to the corresponding lower-dimensional subspace (thus providing a simpler surrogate for G). This would be an alternative to a basis expansion (e.g. in orthogonal polynomials) used in UQ and would exploit the interplay between the properties of G and μ simultaneously.

(d) Statistics

We discuss two applications of our results in the realm of statistics. We start by highlighting the link between our results and the so-called *influence curves* (IC) in robust statistics. For a functional $\mu \mapsto T(\mu)$ its IC is defined as

$$\mathrm{IC}(y) = \lim_{t\to 0} \frac{T(t\delta_y + (1-t)\mu) - T(\mu)}{t}.$$

Computing the IC, if it exists, is in general hard and closed form solutions may be unachievable. However, for the so-called M-estimators, defined as optimizers for $V(0)$,

$$T(\mu) := \mathrm{argmin}_a \int f(x,a)\mu(\mathrm{d}x),$$

for some f (e.g. $f(x,a) = |x-a|$ for the median), we have

$$\mathrm{IC}(y) = \frac{\nabla_a f(y, T(\mu))}{-\int \nabla_a^2 f(s, T(\mu))\,\mu(\mathrm{d}s)},$$

under suitable assumptions on f, see [44, section 3.2.1]. In comparison, writing T^δ for the optimizer for $V(\delta)$, theorem 2.4 yields

$$\lim_{\delta\to 0} \frac{T^\delta - T(\mu)}{\delta} = \frac{\int \nabla_x\nabla_a f(x, T(\mu))\nabla_x f(x, T(\mu))\,\mu(\mathrm{d}x)}{-\int \nabla_a^2 f(s, T(\mu))\,\mu(\mathrm{d}s)}, \tag{3.2}$$

under assumption 2.3 and normalization $||\nabla_x f(x, T(\mu))||_{L^p(\mu)} = 1$. To investigate the connection let us Taylor-expand $\mathrm{IC}(y)$ around x to obtain

$$\mathrm{IC}(y) - \mathrm{IC}(x) = \frac{\nabla_a\nabla_x f(x, T(\mu))}{-\int \nabla_a^2 f(s, T(\mu))\,\mu(\mathrm{d}s)}(y-x).$$

Choosing $y = x + \delta\nabla f_x(x, T(\mu))$ and integrating both sides over μ and dividing by δ, we obtain the asymptotic equality

$$\int \frac{\mathrm{IC}(x + \delta\nabla_x f(x, T(\mu))) - \mathrm{IC}(x)}{\delta}\,\mu(\mathrm{d}x) \approx \frac{T^\delta - T(\mu)}{\delta},$$

for $\delta \to 0$ by (3.2). We conclude that considering the average directional derivative of IC in the direction of $\nabla f_x(x, T(\mu))$ gives our first-order sensitivity. For an interesting conjecture regarding the comparison of influence functions and sensitivities in KL-divergence, we refer to [45, Section 7.3] and [22, Section 3.4.2].

Our second application in statistics exploits the representation of the LASSO/Ridge regressions as robust versions of the standard linear regression. We consider $\mathcal{A} = \mathbb{R}^k$ and $\mathcal{S} = \mathbb{R}^{k+1}$.

If instead of the Euclidean metric we take $||(x,y)||_* = |x|_r \mathbf{1}_{\{y=0\}} + \infty \mathbf{1}_{\{y\neq 0\}}$, for some $r > 1$ and $(x,y) \in \mathbb{R}^k \times \mathbb{R}$, in the definition of the Wasserstein distance, then [19] showed that

$$\inf_{a\in\mathbb{R}^k} \sup_{\nu\in B_\delta(\mu)} \int (y - \langle x,a\rangle)^2 \, \nu(\mathrm{d}x,\mathrm{d}y) = \inf_{a\in\mathbb{R}^k} \left(\sqrt{\int (y - \langle a,x\rangle)^2 \, \mu(\mathrm{d}x,\mathrm{d}y)} + \delta |a|_s \right)^2 \tag{3.3}$$

holds, where $1/r + 1/s = 1$. The $\delta = 0$ case is the ordinary least-squares regression. For $\delta > 0$, the r.h.s. for $s = 2$ is directly related to the Ridge regression, while the limiting case $s = 1$ is called the square-root LASSO regression, a regularized variant of linear regression well known for its good empirical performance. Closed-form solutions to (3.3) do not exist in general and it is a common practice to use numerical routines to solve it approximately. Theorem 2.4 offers instead an explicit first-order approximation of $a^\star_\delta$ for small δ. We denote by $a^\star$ the ordinary least-squares estimator and by I the $k \times k$ identity matrix. Note that the first-order condition on $a^\star$ implies that $\int (y - \langle a^\star, x\rangle) x_i \mu(\mathrm{d}x,\mathrm{d}y) = 0$ for all $1 \le i \le k$. In particular, $V(0) = \int (y^2 - \langle a^\star, x\rangle y)\mu(\mathrm{d}x,\mathrm{d}y)$ and $a^\star = D^{-1} \int yx\mu(\mathrm{d}x,\mathrm{d}y)$, where we assume the system is overdetermined so that $D = \int xx^T \, \mu(\mathrm{d}x,\mathrm{d}y)$ is invertible. A direct computation, see [9, example 8.2], yields

$$a^\star_\delta \approx a^\star - \sqrt{V(0)} D^{-1} \, h(a^\star)\delta. \tag{3.4}$$

For $s = 2$, $h(a^\star) = a^\star / |a^\star|_2$ and for $s = 1$, $h(a^\star) = \mathrm{sign}(a^\star)$ and hence[1] $a^\star_\delta$ is approximately

$$\left(1 - \frac{\sqrt{V(0)}}{|a^\star|_2} D^{-1}\delta\right) a^\star \quad \text{and} \quad a^\star - \sqrt{V(0)} D^{-1} \mathrm{sign}(a^\star)\delta, \tag{3.5}$$

respectively. This corresponds to parameter shrinkage: proportional for square-root Ridge and a shift towards zero for square-root LASSO. To the best of our knowledge, these are first such results and we stress that our formulae are valid in a general context and, in particular, parameter shrinkage depends on the direction through the D^{-1} factor. Figure 3 compares the first-order approximation with the actual results and shows a remarkable fit. Furthermore, our results agree with what is known in the canonical test case for the (standard) Ridge and LASSO, see [46]. When $\mu = \mu_N$ is the empirical measure of N i.i.d. observations, the data are centred and the covariates are orthogonal, i.e. $D = (1/N)I$. In that case, (3.5) simplifies to

$$\left(1 - \delta \sqrt{N\left(\frac{1}{R^2} - 1\right)}\right) a^\star \quad \text{and} \quad a^\star - \sqrt{N}\, |y| \sqrt{1 - R^2} \, \mathrm{sign}(a^\star)\delta,$$

where R^2 is the usual coefficient of determination.

The case of μ_N is naturally of particular importance in statistics and data science and we continue to consider it in the next subsection. In particular, we characterize the asymptotic distribution of $\sqrt{N}(a^\star_{1/\sqrt{N}} - a^\star)$, where $a^\star_\delta \in \mathcal{A}^\star_\delta(\mu_N)$ and $a^\star \in \mathcal{A}^\star_0(\mu_\infty)$ is the optimizer of the non-robust problem for the data-generating measure. This recovers the central limit theorem of [47], a link we explain further in §4b.

(e) Out-of-sample error

A benchmark of paramount importance in optimization is the so-called *out-of-sample error*, also known as the *prediction error* in statistical learning. Consider the setup above when μ_N is the empirical measure of N i.i.d. observations sampled from the 'true' distribution $\mu = \mu_\infty$ and take, for simplicity, $|| \cdot || = | \cdot |_s$, with $s > 1$. Our aim is to compute the optimal $a^\star$ which solves the original problem (1.1). However, we only have access to the training set, encoded via μ_N. Suppose we solve the distributionally robust optimization problem (1.2) for μ_N and denote the robust

[1] In the case $s = 1$, inspecting the proof, we see that theorem 2.4 still holds since $a^\star$ does not have zero components μ-a.s., which are the only points of discontinuity of h.

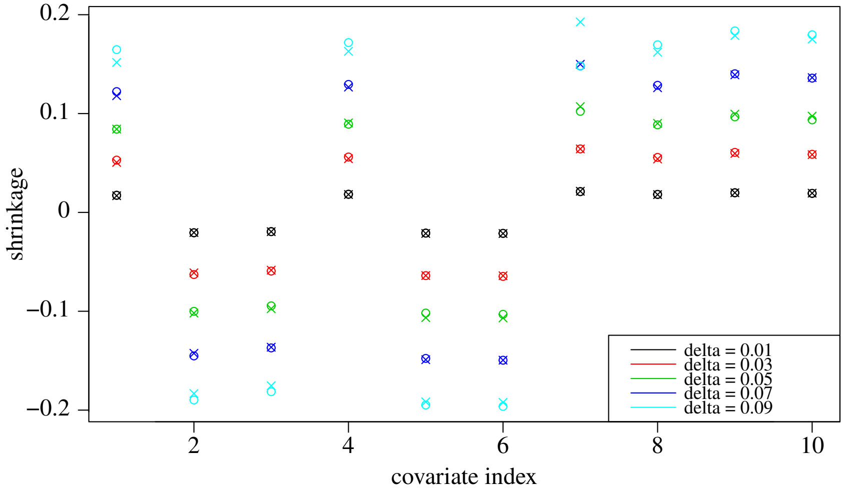

Figure 3. Square-root LASSO parameter shrinkage $a^\star_\delta - a^\star_0$: exact (o) and the first-order approximation (*x*) in (3.5). 2000 observations generated according to $Y = 1.5X_1 - 3X_2 - 2X_3 + 0.3X_4 - 0.5X_5 - 0.7X_6 + 0.2X_7 + 0.5X_8 + 1.2X_9 + 0.8X_{10} + \varepsilon$ with all X_i, ε i.i.d. $\mathcal{N}(0, 1)$. (Online version in colour.)

optimizer $a^{\star,N}_\delta$. Then the *out-of-sample error*

$$V(0, a^{\star,N}_\delta) - V(0, a^\star) = \int f(x, a^{\star,N}_\delta)\,\mu(\mathrm{d}x) - \int f(x, a^\star)\,\mu(\mathrm{d}x)$$

quantifies the error from using $a^{\star,N}_\delta$ as opposed to the true optimizer $a^\star$.

While this expression seems to be hard to compute explicitly for finite samples, theorem 2.4 offers a way to find the asymptotic distribution of a (suitably rescaled version of) the out-of-sample error. We suppose the assumptions in theorem 2.4 are satisfied and note that the first order condition for $a^\star$ gives $\nabla_a V(0, a^\star) = 0$. Then, a second-order Taylor expansion gives

$$V(0, a^{\star,N}_\delta) - V(0, a^\star) = \tfrac{1}{2}(a^{\star,N}_\delta - a^\star)^T \nabla^2_a V(0, \tilde{a})(a^{\star,N}_\delta - a^\star), \tag{3.6}$$

for some $\tilde{a}$ (coordinate-wise) between $a^\star$ and $a^{\star,N}_\delta$. Now we write

$$a^{\star,N}_\delta - a^\star = a^{\star,N}_\delta - a^{\star,N} + a^{\star,N} - a^\star,$$

where we define $a^{\star,N}$ as the optimizer of the non-robust problem (1.1) with μ replaced by μ_N. In particular, the δ-method for M-estimators implies that

$$\sqrt{N}(a^{\star,N} - a^\star) \Rightarrow (\nabla^2_a V(0, a^\star))^{-1} H, \tag{3.7}$$

where $H \sim \mathcal{N}(0, \int (\nabla_a f(x, a^\star))^T \nabla_a f(x, a^\star)\,\mu(\mathrm{d}x))$ and $\Rightarrow$ denotes the convergence in distribution. On the other hand, for a fixed $N \in \mathbb{N}$, theorem 2.4 applied to μ_N yields

$$\begin{aligned} a^{\star,N}_\delta - a^{\star,N} = &-\left(\int |\nabla_x f(x, a^{\star,N})|^q_s\,\mu_N(\mathrm{d}x)\right)^{(1/q)-1} \cdot \left(\int \nabla^2_a f(x, a^{\star,N})\,\mu_N(\mathrm{d}x)\right)^{-1} \\ &\cdot \int \frac{\nabla_x \nabla_a f(x, a^{\star,N})\, h(\nabla_x f(x, a^{\star,N}))}{|\nabla_x f(x, a^{\star,N})|^{1-q}_s}\,\mu_N(\mathrm{d}x) \cdot \delta + o(\delta) \qquad (3.8) \\ = &-((\nabla^2_a V(0, a^\star))^{-1} \Theta + \Delta_N) \cdot \delta + o(\delta), \qquad (3.9) \end{aligned}$$

where

$$\Theta := \left(\int |\nabla_x f(x,a^\star)|_s^q \,\mu(\mathrm{d}x)\right)^{(1/q)-1} \cdot \int \frac{\nabla_x \nabla_a f(x,a^\star)\, h(\nabla_x f(x,a^\star))}{|\nabla_x f(x,a^\star)|_s^{1-q}}\,\mu(\mathrm{d}x),$$

$$\Delta_N := \left(\int |\nabla_x f(x,a^{\star,N})|_s^q \,\mu_N(\mathrm{d}x)\right)^{(1/q)-1} \cdot \left(\int \nabla_a^2 f(x,a^{\star,N})\,\mu_N(\mathrm{d}x)\right)^{-1}$$
$$\cdot \int \frac{\nabla_x \nabla_a f(x,a^{\star,N})\, h(\nabla_x f(x,a^{\star,N}))}{|\nabla_x f(x,a^{\star,N})|_s^{1-q}}\,\mu_N(\mathrm{d}x) - (\nabla_a^2 V(0,a^\star))^{-1}\Theta.$$

Almost surely (w.r.t. sampling of μ_N), we know that $\mu_N \to \mu$ in W_p as $N \to \infty$, and under the regularity and growth assumptions on f in [9, equation (8.2)] we check that $\Delta_N \to 0$ a.s., see [9, example 8.4] for details. In particular, taking $\delta = 1/\sqrt{N}$ and combining the above with (3.7) we obtain

$$\sqrt{N}\left(a^{\star,N}_{1/\sqrt{N}} - a^\star\right) \Rightarrow (\nabla_a^2 V(0,a^\star))^{-1}(H - \Theta). \qquad (3.10)$$

This recovers the central limit theorem of [47], as discussed in more detail in §4b below. Together, (3.6) and (3.9) give us the a.s. asymptotic behaviour of the out-of-sample error

$$V(0,a^{\star,N}_\delta) - V(0,a^\star) = \frac{1}{2N}(H-\Theta)^T(\nabla_a^2 V(0,a^\star))^{-1}(H-\Theta) + o\left(\frac{1}{N}\right). \qquad (3.11)$$

These results also extend and complement [48, Prop. 17]. [48] investigate when the distributionally robust optimizers $a^{\star,N}_\delta$ yield, on average, better performance than the simple in-sample optimizer $a^{\star,N}$. To this end, they consider the expectation, over the realizations of the empirical measure μ_N of

$$V(0,a^{\star,N}_\delta) - V(0,a^{\star,N}) = \int f(x,a^{\star,N}_\delta)\,\mu(\mathrm{d}x) - \int f(x,a^{\star,N})\,\mu(\mathrm{d}x).$$

This is closely related to the out-of-sample error and our derivations above can be easily modified. The first-order term in the Taylor expansion no longer vanishes and, instead of (3.6), we now have

$$V(0,a^{\star,N}_\delta) - V(0,a^{\star,N}) = \nabla_a V(0,a^{\star,N})(a^{\star,N}_\delta - a^{\star,N}) + o(|a^{\star,N}_\delta - a^{\star,N}|),$$

which holds, e.g. if for any $r > 0$, there exists $c > 0$ such that $\sum_{i=1}^k |\nabla_a \nabla_{a_i} f(x,a)| \le c(1+|x|^p)$ for all $x \in \mathcal{S}$, $|a| \le r$. Combined with (3.8), this gives asymptotics in small δ for a fixed N. For quadratic f and taking $q \uparrow \infty$, we recover the result in [48, Prop. 17], see [9, example 8.4] for details.

4. Further discussion and literature review

We start with an overview of related literature and then focus specifically on a comparison of our results with the CLT of [47] mentioned above.

(a) Discussion of related literature

Let us first remark, that while theorem 2.2 offers some superficial similarities to a classical maximum theorem, which is usually concerned with continuity properties of $\delta \mapsto V(\delta)$, in this work, we are instead interested in the exact first derivative of the function $\delta \mapsto V(\delta)$. Indeed, the convergence $\lim_{\delta\to 0}\sup_{\nu\in B_\delta(\mu)} \int f(x)\,\nu(\mathrm{d}x) = \int f(x)\,\mu(\mathrm{d}x)$ follows for all f satisfying $f(x) \le c(1+|x|^p)$ directly from the definition of convergence in Wasserstein metric (e.g. [49, Def. 6.8]). In conclusion, the main issue is to quantify the rate of this convergence by calculating the first derivative $V'(\delta)$.

Our work investigates model uncertainty broadly conceived: it includes errors related to the choice of models from a particular (parametric or not) class of models as well as the mis-specification of such a class altogether (or indeed, its absence). In the decision theoretic literature, these aspects are sometimes referred to as model ambiguity and model mis-specification, respectively, see [50]. However, seeing our main problem (1.2) in decision theoretic terms is not necessarily helpful as we think of f as given and not coming from some latent expected utility type of problem. In particular, our actions $a \in \mathcal{A}$ are just constants.

In our work, we decided to capture the uncertainty in the specification of μ using neighbourhoods in the Wasserstein distance. As already mentioned, other choices are possible and have been used in past. Possibly, the most often used alternative is the relative entropy, or the Kullblack–Leibler divergence. In particular, it has been used in this context in economics, see [51]. To the best of our knowledge, the only comparable study of sensitivities with respect to relative entropy balls is [22], see also [45] allowing for additional marginal constraints. However, this only considered the specific case $f(x,a) = f(x)$ where the reward function is independent of the action. Its main result is

$$\sup_{\nu \in B_\delta^{KL}(\mu)} \int f(x)\,\nu(\mathrm{d}x) = \int f(x)\,\mu(\mathrm{d}x) + \sqrt{2\,\mathrm{Var}_\mu(f(X))}\delta + \frac{1}{3}\frac{\kappa_3(f(X))}{\mathrm{Var}_\mu(f(X))}\delta^2 + O(\delta^3),$$

where $B_\delta^{KL}(\mu)$ is a ball of radius δ^2 centred around μ in KL-divergence, $\mathrm{Var}_\mu(f(X))$ and $\kappa_3(f(X))$ denote the variance and kurtosis of f under the measure μ respectively. In particular, the first-order sensitivity involves the function f itself. By contrast, our theorem 2.2 states $V'(\delta) = (\int (f'(x))^2\,\mu(\mathrm{d}x))^{1/2}$ and involves the first derivative f'. In the trivial case of a point mass $\mu = \delta_x$, we recover the intuitive sensitivity $V'(\delta) = |f'(x)|$, while the results of [22] do not apply for this case. We also note that [22] requires exponential moments of the function f under the baseline measure μ, while we only require polynomial moments. In particular, in applications in econometrics (or any field in which μ typically has fat tails), the scope of application of the corresponding results might then be decisively different. We remark however, that this requirement can be substantially weakened (to the existence of polynomial moments) when replacing KL-divergences by α-divergences, e.g. [52,53]. We expect a sensitivity analysis similar to [22] to hold in this setting. However, to the best of our knowledge no explicit results seem to be available in the literature.

To understand the relative technical difficulties and merits, it is insightful to go into the details of the statements. In fact, in the case of relative entropy and the one-period set-up we are considering, the exact form of the optimizing density can be determined exactly (see [22, Proposition 3.1]) up to a one-dimensional Langrange parameter. This is well known and is the reason behind the usual elegant formulae obtained in this context. But this then reduces the problem in [22] to a one-dimensional problem, which can be well-approximated via a Taylor approximation. By contrast, when we consider balls in the Wasserstein distance, the form of the optimizing measure is not known (apart from some degenerate cases). In fact, a key insight of our results is that the optimizing measure can be approximated by a deterministic shift in the direction $(x + f'(x)\delta)_*\mu$ (this is, in general, not exact but only true as a first-order approximation). The reason for these contrastive starting points of the analyses is the fact that Wasserstein balls contain a more heterogeneous set of measures, while in the case of relative entropy, exponentiating f will always do the trick. We remark however that this is not true for the finite-horizon problems considered in [22, Section 3.2] any more, where the worst-case measure is found using an elaborate fixed-point equation.

A point which further emphasizes the fact that the topology introduced by the Wasserstein metric is less tractable is the fact that

$$W_p^p(\mu,\nu) = \lim_{\varepsilon \to 0} \inf_{\pi \in \Pi(\mu,\nu)} \int |x-y|^p\,\pi(\mathrm{d}x,\mathrm{d}y) + \varepsilon H(\pi \mid \mu \otimes \nu) = \lim_{\varepsilon \to 0} \varepsilon \inf_{\pi \in \Pi(\mu,\nu)} H(\pi \mid R^\varepsilon),$$

where $H(\pi \mid R^\varepsilon) = \int \log(\frac{d\pi}{dR^\varepsilon})\,d\pi$ is the relative entropy and

$$\mathrm{d}R^\varepsilon = c_0 \exp\left(-\frac{|x-y|^p}{\varepsilon}\right)\mathrm{d}(\mu \otimes \nu),$$

for some normalizing constant $c_0 > 0$ (e.g. [54]). This is known as the entropic optimal transport formulation and has received considerable interest in the ML community in the past years (e.g. [55]). In particular, the Wasserstein distance can be approximated by relative entropy, but only with respect to reference measures on the product space. As we consider optimization over ν

above it amounts to changing the reference measure. In consequence, the topological structure imposed by Wasserstein distances is more intricate compared to relative entropy, but also more flexible.

The other well-studied distance is the Hellinger distance. [24] calculates influence curves for the minimum Hellinger distance estimator $a^{\mathrm{Hell},\star}$ on a countable sample space. Their main result is that for the choice $f(x,a)=\log(\ell(x,a))$ (where $(\ell(x,a))_{a\in\mathcal{A}}$ is a collection of parametric densities)

$$IC(x)=-(\nabla_a^2 V(0,a^{\mathrm{Hell},\star}))^{-1}\nabla_a \log(\ell(x,a^{\mathrm{Hell},\star})),$$

the product of the inverse Fisher information matrix and the score function, which is the same as for the classical maximum-likelihood estimator. Denote by μ_N the empirical measure of N data samples and by $a^{\mathrm{Hell},\star}(N)$ the corresponding minimum Hellinger distance estimator for μ_N. In particular, the above result then implies the same CLT as for M-estimators given by

$$N^{1/2}(a^{\mathrm{Hell},\star}(N)-a^{\mathrm{Hell},\star})\Rightarrow(\nabla_a^2 V(0,a^{\mathrm{Hell},\star}))^{-1}H,$$

where $H\sim\mathcal{N}(0,\int\nabla_a f(x,a^{\mathrm{Hell},\star})^T\nabla_a f(x,a^{\mathrm{Hell},\star})\,\mu(\mathrm{d}x))$. As we discuss in the next section, our theorem 2.4 yields a similar CLT, namely

$$N^{1/2}(a^{\star,N}_{1/\sqrt{N}}-a^\star)\Rightarrow(\nabla_a^2 V(0,a^\star))^{-1}\cdot\left(H-\nabla_a\sqrt{\int|\nabla_x f(x,a^\star)|_s^2\,\mu(\mathrm{d}x)}\right).$$

Thus the Wasserstein worst-case approach leads to a shift of the mean of the normal distribution in the direction

$$-\nabla_a\sqrt{\int|\nabla_x f(x,a^\star)|_s^2\,\mu(\mathrm{d}x)},$$

compared to the non-robust case. In the simple case $\mu=\mathcal{N}(0,\sigma^2)$ with standard deviation $\sigma>0$, we obtain the MLE $\sigma^{\star,N}=\frac{1}{N}\sum_{k=1}^N X_i^2$. We can directly compute (for $a=\sigma$) that

$$\nabla_\sigma\sqrt{\int\left|\nabla_x\left(\mathrm{const.}+\log\left(\exp\left(-\frac{x^2}{2(\sigma^\star)^2}\right)\right)\right)\right|_s^2\mu(\mathrm{d}x)}=\nabla_\sigma\sqrt{\int\frac{x^2}{(\sigma^\star)^4}\,\mu(\mathrm{d}x)}$$
$$=\nabla_\sigma\frac{\sigma^\star}{(\sigma^\star)^2}=\nabla_\sigma\frac{1}{\sigma^\star}=-\frac{1}{(\sigma^\star)^2}.$$

Thus the robust approach accounts for a shift of $1/(\sigma^\star)^2$ (of order 1 if mulitplied with inverse Fisher information) to account for a possibly higher variance in the underlying data. In particular, in our approach, the so-called neutral spaces considered (e.g. [56], eqn (21)]) as

$$\{a:\ -(a-a^\star)^T\nabla_a^2 V(0,a^\star)(a-a^\star)\le\delta\}$$

should also take this shift into account, i.e. their definition should be adjusted to

$$\left\{a:-\left(a-a^\star+\nabla_a\sqrt{\int|\nabla_x f(x,a^\star)|_s^2\,\mu(\mathrm{d}x)}\right)^T\nabla_a^2 V(0,a^\star)\right.$$
$$\left.\cdot\left(a-a^\star+\nabla_a\sqrt{\int|\nabla_x f(x,a^\star)|_s^2\,\mu(\mathrm{d}x)}\right)\le\delta\right\}.$$

Lastly, let us mention another situation when our approach provides directly interpretable insights in the context of a parametric family of models. Namely, if one considers a family of models $\mathcal{P}$ such that the worst-case model in the Wasserstein ball remains in $\mathcal{P}$, i.e. $(x+f'(x)\delta)_*\mu\in\mathcal{P}$, then considering (the first-order approximation to) model uncertainty over Wasserstein balls actually reduces to considerations within the parametric family. While uncommon, such a situation would arise, for example, for a scale-location family $\mathcal{P}$, with $\mu\in\mathcal{P}$ and a linear/quadratic f.

(b) Link to the central limit theorem of [47]

As observed in §3e above, theorem 2.4 allows to recover the main results in [47]. We explain this now in detail. Set $||\cdot|| = |\cdot|_s$, $p = q = 2$, $\mathcal{S} = \mathbb{R}^d$. Let μ_N denote the empirical measure of N i.i.d. samples from μ. We impose the assumptions on μ and f from [47], including Lipschitz continuity of gradients of f and strict convexity. These, in particular, imply that the optimizers $a^{\star,N}_\delta, a^{\star,N}$ and $a^\star$, as defined in §3e are well defined and unique, and further $a^{\star,N}_{1/\sqrt{N}} \to a^\star$ as $N \to \infty$. [47, Thm. 1] implies that, as $N \to \infty$,

$$\sqrt{N}(a^{\star,N}_{1/\sqrt{N}} - a^\star) \Rightarrow (\nabla^2_a V(0,a^\star))^{-1} \cdot \left(H - \nabla_a \sqrt{\int |\nabla_x f(x,a^\star)|_s^2 \, \mu(\mathrm{d}x)} \right), \tag{4.1}$$

where $H \sim \mathcal{N}(0, \int \nabla_a f(x,a^\star)^T \nabla_a f(x,a^\star) \, \mu(\mathrm{d}x))$. We note that for $||\cdot|| = |\cdot|_s$ we have

$$h(x) = (\mathrm{sign}(x_1) \, |x_1|^{s-1}, \ldots, \mathrm{sign}(x_k) \, |x_k|^{s-1}) \cdot |x|_s^{1-s} = \nabla_x |x|_s.$$

Thus

$$\nabla_a \sqrt{\int |\nabla_x f(x,a^\star)|_s^2 \, \mu(\mathrm{d}x)} = \frac{\int |\nabla_x f(x,a^\star)|_s h(\nabla_x f(x,a^\star)) \nabla_x \nabla_a f(x,a^\star) \, \mu(\mathrm{d}x)}{\sqrt{\int |\nabla_x f(x,a^\star)|_s^2 \, \mu(\mathrm{d}x)}},$$

and (4.1) agrees with (3.10) which is justified by the Lipschitz growth assumptions on f, $\nabla_x f(x,a)$ and $\nabla_a \nabla_x f(x,a)$ from [47], see [9, equation (8.2)]. In particular, theorem 2.4 implies (4.1) as a special case. While this connection is insightful to establish[2] it is also worth stressing that the proofs in [47] pass through the dual formulation and are thus substantially different from ours. Furthermore, while theorem 2.4 holds under milder assumptions on f than those in [47], the last argument in our reasoning above requires the stronger assumptions on f. It is thus not clear if our results could help to significantly weaken the assumptions in the central limit theorems of [47].

5. Proofs

We consider the case $\mathcal{S} = \mathbb{R}^d$ and $||\cdot|| = |\cdot|$ here. For the general case and additional details, we refer to [9]. When clear from the context, we do not indicate the space over which we integrate.

Proof of theorem 2.2. For every $\delta \geq 0$, let $C_\delta(\mu)$ denote those $\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ which satisfy

$$\pi_1 = \mu \quad \text{and} \quad \left(\int |x-y|^p \, \pi(\mathrm{d}x, \mathrm{d}y) \right)^{1/p} \leq \delta.$$

As the infimum in the definition of $W_p(\mu, \nu)$ is attained (see [49, Theorem 4.1, p. 43]) one has $B_\delta(\mu) = \{\pi_2 : \pi \in C_\delta(\mu)\}$.

We start by showing the '$\leq$' inequality in the statement. For any $a^\star \in \mathcal{A}^\star_0$, one has $V(\delta) \leq \sup_{\nu \in B_\delta(\mu)} \int f(y,a^\star) \, \nu(\mathrm{d}y)$ with equality for $\delta = 0$. Therefore, differentiating $f(\cdot, a^\star)$ and using both Fubini's theorem and Hölder's inequality, we obtain that

$$\begin{aligned} V(\delta) - V(0) &\leq \sup_{\pi \in C_\delta(\mu)} \int f(y,a^\star) - f(x,a^\star) \, \pi(\mathrm{d}x,\mathrm{d}y) \\ &= \sup_{\pi \in C_\delta(\mu)} \int_0^1 \int \langle \nabla_x f(x + t(y-x), a^\star), (y-x) \rangle \, \pi(\mathrm{d}x,\mathrm{d}y) \, \mathrm{d}t \\ &\leq \delta \sup_{\pi \in C_\delta(\mu)} \int_0^1 \left(\int |\nabla_x f(x + t(y-x), a^\star)|^q \pi(\mathrm{d}x,\mathrm{d}y) \right)^{1/q} \mathrm{d}t. \end{aligned}$$

Any choice $\pi^\delta \in C_\delta(\mu)$ converges in p-Wasserstein distance on $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ to the pushforward measure of μ under the mapping $x \mapsto (x,x)$, which we denote $[x \mapsto (x,x)]_* \mu$. This can be seen by,

[2] We thank Jose Blanchet for pointing out the possible link and encouraging us to explore it.

for example, considering the coupling $[(x,y)\mapsto(x,y,x,x)]_*\pi^\delta$ between π^δ and $[x\mapsto(x,x)]_*\mu$. Now note that $q=p/(p-1)$ and the growth assumption on $\nabla_x f(\cdot,a^\star)$ implies

$$|\nabla_x f(x+t(y-x),a^\star)|^q \le c(1+|x|^p+|y|^p), \tag{5.1}$$

for some $c>0$ and all $x,y\in\mathbb{R}^d$, $t\in[0,1]$. In particular, $\int|\nabla_x f(x+t(y-x),a^\star)|^q\,\pi^\delta(\mathrm{d}x,\mathrm{d}y)\le C$ for all $t\in[0,1]$ and small $\delta>0$, for another constant $C>0$. As further $(x,y)\mapsto|\nabla_x f(x+t(y-x),a^\star)|^q$ is continuous for every t, the p-Wasserstein convergence of π^δ to $[x\mapsto(x,x)]_*\mu$ implies that

$$\int|\nabla_x f(x+t(y-x),a^\star)|^q\,\pi^\delta(\mathrm{d}x,\mathrm{d}y)\to\int|\nabla_x f(x,a^\star)|^q\,\mu(\mathrm{d}x),$$

for every $t\in[0,1]$ for $\delta\to0$, see [9, lemma 7.13]. Dominated convergence (in t) then yields '$\le$' in the statement of the theorem.

We turn now to the opposite '$\ge$' inequality. As $V(\delta)\ge V(0)$ for every $\delta>0$, there is no loss of generality in assuming that the right-hand side is not equal to zero. Now take any, for notational simplicity not relabelled, subsequence of $(\delta)_{\delta>0}$ which attains the liminf in $(V(\delta)-V(0))/\delta$ and pick $a^\star_\delta\in\mathcal{A}^\star_\delta$. By assumption, for a (again not relabelled) subsequence, one has $a^\star_\delta\to a^\star\in\mathcal{A}^\star_0$. Further note that $V(0)\le\int f(x,a^\star_\delta)\,\mu(\mathrm{d}x)$ which implies

$$V(\delta)-V(0)\ge\sup_{\pi\in C_\delta(\mu)}\int f(y,a^\star_\delta)-f(x,a^\star_\delta)\,\pi(\mathrm{d}x,\mathrm{d}y).$$

Now define $\pi^\delta:=[x\mapsto(x,x+\delta T(x))]_*\mu$, where

$$T(x):=\frac{\nabla_x f(x,a^\star)}{|\nabla_x f(x,a^\star)|^{2-q}}\left(\int|\nabla_x f(z,a^\star)|^q\,\mu(\mathrm{d}z)\right)^{1/q-1}$$

for $x\in\mathbb{R}^d$ with the convention $0/0=0$. Note that the integral is well defined since, as before in (5.1), one has $|\nabla_x f(x,a^\star)|^q\le C(1+|x|^p)$ for some $C>0$ and the latter is integrable under μ. Using that $pq-p=q$ it further follows that

$$\begin{aligned}\int|x-y|^p\,\pi^\delta(\mathrm{d}x,\mathrm{d}y)&=\delta^p\int|T(x)|^p\,\mu(\mathrm{d}x)\\&=\delta^p\frac{\int|\nabla_x f(x,a^\star)|^{p+pq-2p}\,\mu(\mathrm{d}x)}{\left(\int|\nabla_x f(z,a^\star)|^q\,\mu(\mathrm{d}z)\right)^{p(1-1/q)}}=\delta^p.\end{aligned}$$

In particular, $\pi^\delta\in C_\delta(\mu)$ and we can use it to estimate from below the supremum over $C_\delta(\mu)$ giving

$$\begin{aligned}\frac{V(\delta)-V(0)}{\delta}&\ge\frac{1}{\delta}\int f(x+\delta T(x),a^\star_\delta)-f(x,a^\star_\delta)\,\mu(\mathrm{d}x)\\&=\int_0^1\int\langle\nabla_x f(x+t\delta T(x),a^\star_\delta),T(x)\rangle\,\mu(\mathrm{d}x)\,\mathrm{d}t.\end{aligned}$$

For any $t\in[0,1]$, with $\delta\to0$, the inner integral converges to

$$\int\langle\nabla_x f(x,a^\star),T(x)\rangle\,\mu(\mathrm{d}x)=\left(\int|\nabla_x f(x,a^\star)|^q\,\mu(\mathrm{d}x)\right)^{1/q}.$$

The last equality follows from the definition of T and a simple calculation. To justify the convergence, first note that $\langle\nabla_x f(x+t\delta T(x),a^\star_\delta),T(x)\rangle\to\langle\nabla_x f(x,a^\star),T(x)\rangle$ for all $x\in\mathbb{R}^d$ by continuity of $\nabla_x f$ and since $a^\star_\delta\to a^\star$. Moreover, as before in (5.1), one has $|T(x)|\le c(1+|x|)$ for some $c>0$, hence $|\langle\nabla_x f(x+t\delta T(x),a^\star),T(x)\rangle|\le C(1+|x|^p)$ for some $C>0$ and all $t\in[0,1]$. The latter is integrable under μ; hence convergence of the integrals follows from the dominated convergence theorem. This concludes the proof. ■

Proof of theorem 2.4. We first show that

$$\lim_{\delta\to 0}\frac{-\nabla_{a_i}V(0,a^\star_\delta)}{\delta}=\int \nabla_x\nabla_{a_i}f(x,a^\star)\frac{\nabla_x f(x,a^\star)}{|\nabla_x f(x,a^\star)|^{2-q}}\,\mu(\mathrm{d}x)\cdot\left(\int |\nabla_x f(x,a^\star)|^q\,\mu(\mathrm{d}x)\right)^{1/q-1} \tag{5.2}$$

for all $i\in\{1,\ldots,k\}$. We start with the '$\leq$' inequality. For any $a\in\mathcal{A}^o$, we have

$$\nabla_a f(y,a)-\nabla_a f(x,a)=\int_0^1 \nabla_x\nabla_a f(x+t(y-x),a)(y-x)\,\mathrm{d}t.$$

Let $\delta>0$ and recall that $a^\star_\delta\in\mathcal{A}^\star_\delta$ converge to $a^\star\in\mathcal{A}^\star_0$. Let $B^\star_\delta(\mu,a^\star_\delta)$ denote the set of $\nu\in B_\delta(\mu)$ which attain the value: $\int f(x,a^\star_\delta)\,\nu(\mathrm{d}x)=V(\delta)$. This is non-empty by assumption 2.3 and [9, lemma 7.16]. By [9, lemma 8.5] the function $a\mapsto V(\delta,a)$ is (one-sided) directionally differentiable at $a^\star_\delta$ for all $\delta>0$ small and thus for all $i\in\{1,\ldots,k\}$

$$\sup_{\nu\in B^\star_\delta(\mu,a^\star_\delta)}\int \nabla_{a_i}f(x,a^\star_\delta)\,\nu(\mathrm{d}x)\geq 0.$$

Then, using Lagrange multipliers to encode the optimality of $B^\star_\delta(\mu,a^\star_\delta)$ in $B_\delta(\mu)$, we obtain

$$\begin{aligned}-\nabla_{a_i}V(0,a^\star_\delta)&\leq \sup_{\nu\in B^\star_\delta(\mu,a^\star_\delta)}\int \nabla_{a_i}f(y,a^\star_\delta)\nu(\mathrm{d}y)-\nabla_{a_i}V(0,a^\star_\delta)\\
&=\sup_{\nu\in B_\delta(\mu)}\inf_{\lambda\in\mathbb{R}}\left(\int [\nabla_{a_i}f(y,a^\star_\delta)+\lambda(f(y,a^\star_\delta)-V(\delta))]\nu(\mathrm{d}y)\right.\\
&\qquad\left.-\int [\nabla_{a_i}f(x,a^\star_\delta)+\lambda(f(x,a^\star_\delta)-V(0,a^\star_\delta))]\mu(\mathrm{d}x)\right)\\
&=\inf_{\lambda\in\mathbb{R}}\left(\sup_{\pi\in C_\delta(\mu)}\int_0^1\int \Big\langle \nabla_x\nabla_{a_i}f(x+t(y-x),a^\star_\delta)\right.\\
&\qquad+\lambda\nabla_x f(x+t(y-x),a^\star_\delta),y-x\Big\rangle\,\pi(\mathrm{d}x,\mathrm{d}y)\,\mathrm{d}t\\
&\qquad\left.-\lambda\sup_{\pi\in C_\delta(\mu)}\int_0^1\int \langle \nabla_x f(x+t(y-x),a^\star_\delta,y-x\rangle\,\pi(\mathrm{d}x,\mathrm{d}y)\,\mathrm{d}t\right),\end{aligned}$$

where we used a minimax argument as well as Fubini's theorem. We note that the functions above satisfy the assumptions of theorem 2.2 for a fixed λ. In particular, using exactly the same arguments as in the proof of theorem 2.2 (i.e. Hölder's inequality and a specific transport attaining the supremum) we obtain by exchanging the order of lim sup and inf that

$$\limsup_{\delta\to 0}\frac{-\nabla_{a_i}V(0,a^\star_\delta)}{\delta}\leq \inf_{\lambda\in\mathbb{R}}\left(\left(\int |\nabla_x\nabla_{a_i}f(x,a^\star)+\lambda\nabla_x f(x,a^\star)|^q\,\mu(\mathrm{d}x)\right)^{1/q}-\lambda\left(\int |\nabla_x f(x,a^\star)|^q\,\mu(\mathrm{d}x)\right)^{1/q}\right). \tag{5.3}$$

For $q=2$, the infimum can be computed explicitly and equals

$$\frac{\int \langle \nabla_x\nabla_{a_i}f(x,a^\star),\nabla_x f(x,a^\star)\rangle\,\mu(\mathrm{d}x)}{\sqrt{\int |\nabla_x f(x,a^\star)|^2\,\mu(\mathrm{d}x)}}.$$

For the general case, we refer to [9, lemma 8.6], noting that by assumption $\nabla_x f(x,a^\star)\neq 0$, we see that the r.h.s. above is equal to the r.h.s. in (5.2).

The proof of the '$\geq$' inequality in (5.2) follows by the very same arguments. Indeed, [9, lemma 8.5] implies that

$$\inf_{\nu\in B^{\star}_{\delta}(\mu,a^{\star}_{\delta})} \int \nabla_{a_i} f(x,a^{\star}_{\delta})\,\nu(\mathrm{d}x) \leq 0,$$

for all $i\in\{1,\dots,k\}$ and we can write

$$\begin{aligned}
-\nabla_{a_i} V(0,a^{\star}_{\delta}) &\geq \inf_{\nu\in B^{\star}_{\delta}(\mu,a^{\star}_{\delta})} \int \nabla_{a_i} f(y,a^{\star}_{\delta})\,\nu(\mathrm{d}y) - \nabla_{a_i} V(0,a^{\star}_{\delta})\\
&= \inf_{\nu\in B_{\delta}(\mu)} \sup_{\lambda\in\mathbb{R}} \Bigg(\int \big[\nabla_{a_i} f(y,a^{\star}_{\delta}) + \lambda(f(y,a^{\star}_{\delta}) - V(\delta))\big]\nu(\mathrm{d}y)\\
&\qquad - \int \big[\nabla_{a_i} f(x,a^{\star}_{\delta}) + \lambda(f(x,a^{\star}_{\delta}) - V(0,a^{\star}_{\delta}))\big]\mu(\mathrm{d}x)\Bigg).
\end{aligned}$$

From here on, we argue as in the '$\leq$' inequality and conclude that indeed (5.2) holds.

By assumption, the matrix $\nabla^2_a V(0,a^{\star})$ is invertible. Therefore, in a small neighbourhood of $a^{\star}$, the mapping $\nabla_a V(0,\cdot)$ is invertible. In particular, $a^{\star}_{\delta} = (\nabla_a V(0,\cdot))^{-1}(\nabla_a V(0,a^{\star}_{\delta}))$ and by the first-order condition $a^{\star} = (\nabla_a V(0,\cdot))^{-1}(0)$. Applying the chain rule and using (5.2) gives

$$\begin{aligned}
\lim_{\delta\to 0} \frac{a^{\star}_{\delta} - a^{\star}}{\delta} &= (\nabla^2_a V(0,a^{\star}))^{-1} \cdot \lim_{\delta\to 0} \frac{\nabla_a V(0,a^{\star}_{\delta})}{\delta}\\
&= -(\nabla^2_a V(0,a^{\star}))^{-1} \left(\int |\nabla_x f(z,a^{\star})|^{q}\,\mu(\mathrm{d}z) \right)^{1/q-1} \cdot \int \frac{\nabla_x \nabla_a f(x,a^{\star}) \nabla_x f(x,a^{\star})}{|\nabla_x f(x,a^{\star})|^{2-q}}\,\mu(\mathrm{d}x).
\end{aligned}$$

This completes the proof. ■

Data accessibility. The codes used to generate figures in the paper are available on GitHub: http://github.com/JanObloj/Robust-uncertainty-sensitivity-analysis.

Authors' contributions. D.B., S.D., J.O. and J.W. formulated the mathematical problem, carried out the analysis, established the main results and drew conclusions. J.O. and J.W. wrote the first draft of the paper. D.B. and J.W. wrote the first draft of the appendix. S.D. and J.W. performed the numerical analysis. All the authors proof read and corrected the manuscript, gave final approval for publication and agree to be held accountable for the work performed therein.

Competing interests. The authors declare no competing interests.

Funding. This work was supported by the European Research Council [7th FP/ERC grant agreement no. 335421], the Vienna Science and Technology Fund (WWTF) [project MA16-021], the Austrian Science Fund (FWF) [project P28661] and the National Science Foundation of China (grant nos 11971310 and 11671257).

Acknowledgements. We thank Jose Blanchet, Mike Giles, Daniel Kuhn and Peyman Mohajerin Esfahani for their helpful comments on an earlier draft of this paper.

References

1. Armacost RL, Fiacco AV. 1974 Computational experience in sensitivity analysis for nonlinear programming. *Math. Program.* **6**, 301–326. (doi:10.1007/BF01580247)
2. Vogel S. 2007 Stability results for stochastic programming problems. *Optimization* **19**, 269–288. (doi:10.1080/02331938808843343)
3. Bonnans JF, Shapiro A. 2013 *Perturbation analysis of optimization problems*. New York, NY: Springer.
4. Ghanem R, Higdon D, Owhadi H eds. 2017 *Handbook of uncertainty quantification*. Cham, Switzerland: Springer.
5. Dupacova J. 1990 Stability and sensitivity analysis for stochastic programming. *Ann. Oper. Res.* **27**, 115–142. (doi:10.1007/BF02055193)
6. Romisch W. 2003 Stability of stochastic programming problems. In *Stochastic programming*, pp. 483–554. Amsterdam, The Netherlands: Elsevier. (doi:10.1016/S0927-0507(03)10008-4)
7. Asi H, Duchi JC. 2019 The importance of better models in stochastic optimization. *Proc. Natl Acad. Sci. USA* **116**, 22 924–22 930. (doi:10.1073/pnas.1908018116)
8. Rahimian H, Mehrotra S. 2019 Distributionally robust optimization: a review. (http://arxiv.org/abs/1908.05659)

9. Bartl D, Drapeau S, Obłój J, Wiesel J. 2021 Supplementary material from "Sensitivity analysis of Wasserstein distributionally robust optimization problems". The Royal Society. Collection. (https://doi.org/10.6084/m9.figshare.c.5730987)
10. Chiappori PA, McCann RJ, Nesheim L. 2010 Hedonic price equilibria, stable matching, and optimal transport: equivalence, topology, and uniqueness. *Econ. Theory* **42**, 317–354. (doi:10.1007/s00199-009-0455-z)
11. Carlier G, Ekeland I. 2010 Matching for teams. *Econ. Theory* **42**, 397–418. (doi:10.1007/s00199-008-0415-z)
12. Peyré G, Cuturi M. 2019 Computational optimal transport. *Found. Trends Mach. Learn.* **11**, 355–607. (doi:10.1561/2200000073)
13. Pflug G, Wozabal D. 2007 Ambiguity in portfolio selection. *Quant. Finance* **7**, 435–442. (doi:10.1080/14697680701455410)
14. Fournier N, Guillin A. 2014 On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Relat. Fields* **162**, 707–738. (doi:10.1007/s00440-014-0583-7)
15. Mohajerin Esfahani P, Kuhn D. 2018 Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Math. Program.* **171**, 115–166. (doi:10.1007/s10107-017-1172-1)
16. Obłój J, Wiesel J. 2021 Robust estimation of superhedging prices. *Ann. Stat.* **49**, 508–530. (doi:10.1214/20-AOS1966)
17. Gao R, Kleywegt AJ. 2016 Distributionally robust stochastic optimization with Wasserstein distance. (http://arxiv.org/abs/1604.02199)
18. Blanchet J, Murthy K. 2019 Quantifying distributional model risk via optimal transport. *Math. Oper. Res.* **44**, 565–600. (doi:10.1287/moor.2018.0936)
19. Blanchet J, Kang Y, Murthy K. 2019 Robust Wasserstein profile inference and applications to machine learning. *J. Appl. Probab.* **56**, 830–857. (doi:10.1017/jpr.2019.49)
20. Kuhn D, Esfahani PM, Nguyen VA, Shafieezadeh-Abadeh S. 2019 Wasserstein distributionally robust optimization: theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pp. 130–166. INFORMS. (doi:10.1287/educ.2019.0198)
21. Shafieezadeh-Abadeh S, Kuhn D, Esfahani PM. 2019 Regularization via mass transportation. *J. Mach. Learn. Res.* **20**, 1–68.
22. Lam H. 2016 Robust sensitivity analysis for stochastic systems. *Math. Oper. Res.* **41**, 1248–1275. (doi:10.1287/moor.2015.0776)
23. Calafiore GC. 2007 Ambiguous risk measures and optimal robust portfolios. *SIAM J. Optim.* **18**, 853–877. (doi:10.1137/060654803)
24. Lindsay BG. 1994 Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Ann. Stat.* **22**, 1081–1114. (doi:10.1214/aos/1176325512)
25. Black F, Scholes M. 1973 The pricing of options and corporate liabilities. *J. Political Econ.* **81**, 637–654. (doi:10.1086/260062)
26. Bartl D, Drapeau S, Tangpi L. 2020 Computational aspects of robust optimized certainty equivalents and option pricing. *Math. Finance* **30**, 287–309. (doi:10.1111/mafi.12203)
27. Ben Tal A, Teboulle M. 1986 Expected utility, penalty functions, and duality in stochastic nonlinear programming. *Manage. Sci.* **32**, 1445–1466. (doi:10.1287/mnsc.32.11.1445)
28. Artzner P, Delbaen F, Eber J, Heath D. 1999 Coherent measures of risk. *Math. Finance* **9**, 203–228. (doi:10.1111/1467-9965.00068)
29. Ben Tal A, Teboulle M. 2007 An old-new concept of convex risk measures: the optimized certainty equivalent. *Math. Finance* **17**, 449–476. (doi:10.1111/j.1467-9965.2007.00311.x)
30. Markowitz H. 1952 Portfolio selection. *J. Finance* **7**, 77–91. (doi:10.2307/2975974)
31. Pflug GC, Pichler A, Wozabal D. 2012 The 1/N investment strategy is optimal under high model ambiguity. *J. Bank. Finance* **36**, 410–417. (doi:10.1016/j.jbankfin.2011.07.018)
32. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. 2013 Intriguing properties of neural networks. (http://arxiv.org/abs/1312.6199)
33. Goodfellow IJ, Shlens J, Szegedy C. 2014 Explaining and harnessing adversarial examples. (http://arxiv.org/abs/1412.6572)
34. Li L, Zhong Z, Li B, Xie T. 2019 Robustra: training provable robust neural networks over reference adversarial space. In *Proc. 28th Int. Joint Conf. on Artificial Intelligence*, pp. 4711–4717. AAAI Press. (doi:10.24963/ijcai.2019/654)
35. Carlini N, Wagner D. 2017 Towards evaluating the robustness of neural networks. In *2017 IEEE Symp. on Security and Privacy (SP)*, pp. 39–57. IEEE. (doi:10.1109/SP.2017.49)

36. Wong E, Kolter JZ. 2017 Provable defenses against adversarial examples via the convex outer adversarial polytope. (http://arxiv.org/abs/1711.00851)
37. Weng TW, Zhang H, Chen PY, Yi J, Su D, Gao Y, Hsieh CJ, Daniel L. 2018 Evaluating the robustness of neural networks: an extreme value theory approach. (http://arxiv.org/abs/1801.10578)
38. Araujo A, Pinot R, Negrevergne B, Meunier L, Chevaleyre Y, Yger F, Atif J. 2019 Robust neural networks using randomized adversarial training. (http://arxiv.org/abs/1903.10219)
39. Mangal R, Nori AV, Orso A. 2019 Robustness of neural networks: a probabilistic and practical approach. In *Proc. 41st Int. Conf. on Software Engineering: New Ideas and Emerging Results*, pp. 93–96. IEEE Press. (doi:10.1109/ICSE-NIER.2019.00032)
40. Bastani O, Ioannou Y, Lampropoulos L, Vytiniotis D, Nori A, Criminisi A. 2016 Measuring neural net robustness with constraints. (https://arxiv.org/abs/1605.07262)
41. Sinha A, Namkoong H, Volpi R, Duchi J. 2020 Certifying some distributional robustness with principled adversarial training. (http://arxiv.org/abs/1710.10571v5)
42. Ho-Nguyen N, Wright SJ. 2020 Adversarial classification via distributional robustness with wasserstein ambiguity. (http://arxiv.org/abs/2005.13815)
43. Chen Z, Kuhn D, Wiesemann W. 2018 Data-driven chance constrained programs over Wasserstein balls. (http://arxiv.org/abs/1809.00210)
44. Huber P, Ronchetti E. 1981 *Robust statistics*. Wiley Series in Probability and Mathematical Statistics, vol. 52. New York, NY: Wiley-IEEE.
45. Lam H. 2018 Sensitivity to serial dependency of input processes: a robust approach. *Manage. Sci.* **64**, 1311–1327. (doi:10.1287/mnsc.2016.2667)
46. Tibshirani R. 1996 Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B. Stat. Methodol.* **58**, 267–288.
47. Blanchet J, Murthy K, Si N. 2019 Confidence regions in Wasserstein distributionally robust estimation. (http://arxiv.org/abs/1906.01614)
48. Anderson EJ, Philpott AB. 2019 Improving sample average approximation using distributional robustness. *Optimization Online*. See http://www.optimization-online.org/DB_HTML/2019/10/7405.html.
49. Villani C. 2009 *Optimal transport: old and new*, Berlin, Germany: Springer.
50. Hansen LP, Marinacci M. 2016 Ambiguity aversion and model misspecification: an economic perspective. *Stat. Sci.* **31**, 511–515. (doi:10.1214/16-STS570)
51. Hansen LP, Sargent T. 2007 *Robustness*. Princeton, NJ: Princeton University Press.
52. Atar R, Chowdhary K, Dupuis P. 2015 Robust bounds on risk-sensitive functionals via Rényi divergence. *SIAM/ASA J. Uncertain. Quantif.* **3**, 18–33. (doi:10.1137/130939730)
53. Glasserman P, Xu X. 2014 Robust risk measurement and model risk. *Quant. Finance* **14**, 29–58. (doi:10.1080/14697688.2013.822989)
54. Carlier G, Duval V, Peyré G, Schmitzer B. 2017 Convergence of entropic schemes for optimal transport and gradient flows. *SIAM J. Math. Anal.* **49**, 1385–1418. (doi:10.1137/15M1050264)
55. Peyré G, Cuturi M. 2019 Computational optimal transport: with applications to data science. *Found. Trends Mach. Learn.* **11**, 355–607. (doi:10.1561/2200000073)
56. Komorowski M, Costa MJ, Rand DA, Stumpf MP. 2011 Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proc. Natl Acad. Sci. USA* **108**, 8645–8650. (doi:10.1073/pnas.1015814108)