# APPENDIX TO

# SENSITIVITY ANALYSIS OF WASSERSTEIN DISTRIBUTIONALLY ROBUST OPTIMIZATION PROBLEMS

DANIEL BARTL, SAMUEL DRAPEAU, JAN OBŁÓJ AND JOHANNES WIESEL

## 1. Preliminaries

*Please note that this is a supplementary file which should be read in conjunction with the main manuscript* [1].

We recall and further explain the setting from the main body of the paper [1]. Take $d, k \in \mathbb{N}$, endow $\mathbb{R}^d$ with the Euclidean norm $|\cdot|$. Throughout the paper we take the convention that topological properties, such as continuity or closure, are understood w.r.t. $|\cdot|$. We let $\Gamma^o, \bar{\Gamma}, \partial\Gamma, \Gamma^c$ denote respectively the interior, the closure, the boundary and the complement of a set $\Gamma \subset \mathbb{R}^d$. We denote the set of all probability measures on $\Gamma$ by $\mathcal{P}(\Gamma)$. For a variable $\gamma \in \Gamma$, we will denote the optimizer by $\gamma^\star$ and the set of optimizers by $\Gamma^\star$.

Fix a seminorm $\|\cdot\|$ on $\mathbb{R}^d$ and denote by $\|\cdot\|_*$ its (extended) dual norm, i.e. $\|y\|_* := \sup\{\langle x, y\rangle : \|x\| \le 1\}$. Let us define the equivalence relation $x \sim y$ if and only if $\|x - y\| = 0$. Furthermore let us set $U := \{x \in \mathbb{R}^d \ : \ \|x\| = 0\}$ and write $[x] = x + U$. With this notation, the quotient space $\mathbb{R}^d/U = \{[x] \ : \ x \in \mathbb{R}^d\}$ is a normed space for $\|\cdot\|$. Furthermore, by the triangle inequality for $\|\cdot\|$ and equivalence of norms on $\mathbb{R}^d$, there exists $c > 0$ such that $\|x\| \le c|x|$ and $|x| \le c\|x\|_*$ for all $x \in \mathbb{R}^d$. As $|\cdot|$ is Hausdorff, this immediately implies that $\|\cdot\|_*$ is Hausdorff as well. Furthermore we conclude, that $\|\cdot\|$ is continuous and $\|\cdot\|_*$ is lower semicontinuous w.r.t. $|\cdot|$ (as the supremum over continuous functions $\langle x, \cdot\rangle$). Lastly we make the convention that $B_\delta(x)$ denotes the ball of radius $\delta$ around $x$ in $|\cdot|$. As our setup is slightly non-standard, we state the following lemmas for completeness:

**Lemma 5.** *For every $x \in \mathbb{R}^d$ we have that $\|x\| = \sup\{\langle x, y\rangle \ : \ \|y\|_* \le 1\}$.*

*Proof.* As $\{x \in \mathcal{S} \ : \ \|x\| \le 1\}$ is convex and closed, this follows directly from the bipolar theorem. □

**Lemma 6.** *Assume that $\|\cdot\|_*$ is strictly convex. Then the following hold:*

*(i) For all $x \in \mathbb{R}^d$ there exists $h(x) \in \mathbb{R}^d$ such that $\|h(x)\|_* = 1$ and $\|x\| = \langle x, h(x)\rangle$. If $x \neq 0$, then $h(x)$ is unique.*
*(ii) The map $h : \mathbb{R}^d \setminus \{0\} \to \mathbb{R}^d$ is continuous.*

*Proof.* Fix $x \in \mathbb{R}^d \setminus \{0\}$. The existence of $h(x) \in \mathbb{R}^d$ in *(i)* follows from Lemma 5. Assume towards a contradiction that there exists another $\tilde{h}(x) \in \mathbb{R}^d$ with $\|\tilde{h}(x)\|_* = 1$, $\langle x, \tilde{h}(x)\rangle = \|x\|$ and $\tilde{h}(x) \neq h(x)$. Defining $\bar{h}(x) = (h(x) + \tilde{h}(x))/2$ we have $\langle x, \bar{h}(x)\rangle = (\langle x, h(x)\rangle + \langle x, \tilde{h}(x)\rangle)/2 = \|x\|$. On the other hand, by the Hausdorff property of $\|\cdot\|_*$, we have $\|h(x) - \tilde{h}(x)\|_* \neq 0$ and thus, by strict convexity of $\|\cdot\|_*$, $\|\bar{h}(x)\|_* < 1$. Using again Lemma 5, we conclude $\|x\| \ge \langle x, \bar{h}(x)/\|\bar{h}(x)\|_*\rangle > \|x\|$, a contradiction.
For *(ii)* we assume towards a contradiction that for some sequence $(x_n)_{n\in\mathbb{N}}$ in $\mathbb{R}^d$ we have $\lim_{n\to\infty} x_n = x \in \mathbb{R}^d \setminus \{0\}$, but $\lim_{n\to\infty} h(x_n) \neq h(x)$. As remarked above, we have $\{\|\cdot\|_* \le 1\} \subseteq B_c(0)$, in particular $\lim_{n\to\infty} h(x_n) = y \in \mathbb{R}^d$ after taking a subsequence. Recalling that $h(x) \neq y$ and $\|\cdot\|_*$ is lower semicontinuous, we conclude that $\|y\|_* \le 1$ and in particular $\|x\| > \langle x, y\rangle$ by Lemma 5 and *(i)*. Finally

$$\|x\| = \lim_{n\to\infty} \|x_n\| = \lim_{n\to\infty} \langle x_n, h(x_n)\rangle = \langle x, y\rangle,$$

which leads to a contradiction. □

**Lemma 7.** *If $\|\cdot\|$ is strictly convex, then $\|\cdot\|_*$ is strictly convex as well.*

*Date*: November 6, 2021.

*Proof.* Fix $y \in \mathbb{R}^d \setminus \{0\}$. We first note that

$$k(y) := \{x \in \mathbb{R}^d \;:\; \|x\| = 1,\, \|y\|_* = \langle x, y\rangle\}/U$$

is uniquely defined. Indeed, this follows from applying the exact same arguments as in the proof of Lemma 6, adjusting for $U$. Take now $y, y' \in \mathbb{R}^d$ such that $\|y\|_* = \|y'\|_* = 1$ and $\|y - y'\|_* \neq 0$. Set $\bar{y} = (y + y')/2$ and note that $\|\bar{y} - y\|_*, \|\bar{y} - y'\|_* \neq 0$. Then $\|\bar{y}\|_* = (\langle [k(\bar{y})], y\rangle + \langle [k(\bar{y})], y'\rangle)/2 < 1$. This shows the claim. □

Let $\mathcal{S}$ denote the state space which is a closed convex subset of $\mathbb{R}^d$. Fix $p > 1$ and take $q = p/(p-1)$ so that $1/p + 1/q = 1$. For probability measures $\mu$ and $\nu$ on $\mathcal{S}$, we define their $p$-Wasserstein distance as

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

where $\mathrm{Cpl}(\mu, \nu)$ is the set of all probability measures $\pi \in \mathcal{P}(\mathcal{S} \times \mathcal{S})$ with first marginal $\pi_1 := \pi(\cdot \times \mathcal{S}) = \mu$ and second marginal $\pi_2 := \pi(\mathcal{S} \times \cdot) = \nu$. In the proofs we sometimes also use the $p$-Wasserstein distance with respect to the Euclidean norm $|\cdot|$ given by

$$W_p^{|\cdot|}(\mu, \nu) = \inf \left\{ \int_{\mathcal{S}\times\mathcal{S}} |x - y|^p \,\pi(dx, dy) \colon \pi \in \mathrm{Cpl}(\mu, \nu) \right\}^{1/p}.$$

Recall that $|\cdot| \le c\|\cdot\|_*$ for some constant $c > 0$, which in turn implies that $W_p^{|\cdot|}(\cdot, \cdot) \le cW_p(\cdot, \cdot)$. A Wasserstein ball of size $\delta \ge 0$ around $\mu$ is denoted

$$B_\delta(\mu) := \left\{ \nu \in \mathcal{P}(\mathbb{R}^d) : W_p(\mu, \nu) \le \delta \right\}.$$

From now on, we fix $\mu \in \mathcal{P}(\mathcal{S})$ such that $\mu(\partial\mathcal{S}) = 0$ and $\int_{\mathcal{S}} |x|^p \,\mu(dx) < \infty$. Let $\mathcal{A}$ denote the action (decision) space which is a convex and closed subset of $\mathbb{R}^k$. We consider robust stochastic optimization problem [1.2]:

$$V(\delta) := \inf_{a\in\mathcal{A}} V(\delta, a) := \inf_{a\in\mathcal{A}} \sup_{\nu\in B_\delta(\mu)} \int_{\mathcal{S}} f\,(x, a)\;\nu(dx).$$

In accordance with our conventions, we write $a^\star$ for an optimizer: $V(\delta) = V(\delta, a^\star)$ and $\mathcal{A}^\star_{\subset}\mathcal{A}$ for the set of such optimizers. We also let $B^\star_\delta(\mu, a)$ denote the set of measures $\nu^\star$ such that $V(\delta, a) = \int_{\mathcal{S}} f(x, a)\,\nu^\star(dx)$ and sometimes write $B^\star_\delta(\mu)$ for $B^\star_\delta(\mu, a^\star)$ if $a^\star \in \mathcal{A}^\star_\delta$ is fixed.

## 2. Discussion, extensions and proofs related to Theorem 2

We complement now the discussion of Theorem 2. We start with some remarks, extensions and further examples before proceeding with the proofs, including a complete proof of Theorem 2 for general seminorms $\|\cdot\|$.

### 2.1. **Discussion and extensions of Theorem 2.**

*Remark* 8. Theorem 2 may fail for $p = 1$. Indeed take $d = 1$, $\|\cdot\| = |\cdot|$ and $f(x) = x^2$, $\mathcal{S} = [-1, 1]$, $\mu$ the point mass in zero, $\mu = \delta_0$. Then $\nabla_x f(x) = 2x$ and the $\mu(dx)$–essential supremum of $|\nabla_x f(x)|$ is equal to 0. However $\nu_\lambda := \lambda\delta_1 + (1 - \lambda)\delta_0 \in B_\lambda(\mu)$ for all $\lambda \in [0, 1]$ and it is easy to see $V(\delta) = \delta$ and thus $V'(0) = 1$. The point where the proof of Theorem 2 breaks down is that the map $\delta$ to the $\nu_\delta$–essential supremum of $|\nabla_x f(x)|$ is not continuous at $\delta = 0$.

*Remark* 9. Let $p > 2$. In addition to Assumption 1, suppose that $f$ is twice continuously differentiable and that for ever $r \ge 0$ there is $c \ge 0$ such that $|\nabla_x^2 f(x, a)| \le c(1 + |x|^{p-2})$ for all $x \in \mathcal{S}$ and all $a \in \mathcal{A}$ with $|a| \le r$. Then, the same arguments as in the proof of Theorem 2 but with a second order Taylor expansion yield

$$\begin{aligned} V(\delta) \le V(0) &+ \delta \left( \int_{\mathcal{S}} \|\nabla_x f(a^\star, x)\|^q \,\mu(dx) \right)^{1/q} \\ &+ \delta^2 \left( \int_{\mathcal{S}} \lambda_{\max}\Big(\frac{1}{2}\nabla_x^2 f(a^\star, x)\Big)^r \mu(dx) \right)^{1/r} + o(\delta^2), \end{aligned}$$

for small $\delta \ge 0$, where $\lambda_{\max}$ denotes the largest eigenvalue of the Hessian taken w.r.t. the norm $\|\cdot\|_*$ and $r = p/(p-2)$ is such that $2/p + 1/r = 1$.

In particular, this means that if the term in front of $\delta^2$ is the same order of magnitude as the term in front of $\delta$, then the first order approximation is quite accurate for small $\delta$. Note that larger $p$ implies smaller $r$ and therefore a smaller term in front of the $\delta^2$ term.

*Remark* 10. We believe that Assumption 1 lists natural sufficient conditions for differentiability of $V(\delta)$ in zero. In particular all these conditions are used in the proof of Theorem 2. Relaxing Assumption 1 seems to require a careful analysis of the interplay between (the space explored by balls around) $\mu$ and the functions $f, \nabla_x f$. We state here a straightforward extension to the case where $f$ is only weakly differentiable and leave more fundamental extensions (e.g., to manifolds) for future research.
Specifically, in case that the baseline distribution $\mu$ is absolutely continuous w.r.t. the Lebesgue measure and $\|\cdot\| = |\cdot|$, Theorem 2 remains true if we merely assume that $f(\cdot, a)$ has a weak derivative (in the Sobolev sense) on $\mathcal{S}^o$ for all $a \in \mathcal{A}$ and replace $\nabla_x f(\cdot, a)$ by the weak derivative of $f(\cdot, a)$ in Assumption 1. More concretely the first point of Assumption 1 should read:

- The weak derivative $(x, a) \mapsto g(x, a)$ of $f(\cdot, a)$ is continuous at every point $(x, a) \in N \times \mathcal{A}^\star(0)$, where $N$ is a Lebesgue-null set, and for every $r > 0$ there is $c > 0$ such that $|g(x, a)| \le c(1 + |x|^{p-1})$ for all $x \in \mathcal{S}$ and $|a| \le r$.

*Proof of Remark 10.* For notational simplicity we only consider the case $\mathcal{S} = \mathbb{R}^d$. Note that by, e.g., [2][Theorem 8.2] we can assume that $f(\cdot, a)$ is continuous and satisfies

$$f(y, a) - f(x, a) = \int_0^1 \langle g(x + t(y - x), a), y - x\rangle \, dt$$

for all $x, y \in \mathbb{R}^d$ and all $a \in \mathcal{A}$. Furthermore

$$\sup_{\nu \in B_\delta(\mu)} \int_{\mathcal{S}} f(x, a)\, \nu(dx) = \sup_{\nu \in B_\delta(\mu),\ \nu \ll \mathrm{Leb}} \int_{\mathcal{S}} f(x, a)\, \nu(dx), \tag{10}$$

where $\nu \ll \mathrm{Leb}$ means that $\nu$ is absolutely continuous w.r.t. the Lebesgue measure. Indeed, let us take $\nu \in B_\delta(\mu)$ and set $\tilde\nu = \tilde\nu(t, \varepsilon) = (1-t)(\nu * N(0, \varepsilon)) + t\mu$, where $N(0, \varepsilon)$ denotes the multivariate normal distribution with covariance $\varepsilon\mathbf{I}$, $\varepsilon > 0$ and $*$ denotes the convolution operator. For every $0 < t < 1$, by convexity of $W_p^p(\cdot, \cdot)$ and the triangle inequality for $W_p$, we have

$$\begin{aligned} W_p^p(\mu, \tilde\nu) &\le (1-t) W_p^p(\nu * N(0, \varepsilon), \mu) + t W_p^p(\mu, \mu) \\ &= (1-t) W_p^p(\nu * N(0, \varepsilon), \mu) \\ &\le (1-t) \left( W_p(\nu * N(0, \varepsilon), \nu) + W_p(\nu, \mu) \right)^p . \end{aligned}$$

By assumption $W_p(\nu, \mu) \le \delta$ and one can check that $W_p(\nu * N(0, \varepsilon), \nu) \to 0$ as $\varepsilon \to 0$. Hence, for every $t < 1$ there exists small $\varepsilon = \varepsilon(t) > 0$ such that $W_p(\mu, \tilde\nu) \le \delta$. As further $\lim_{t \to 1} \int_{\mathcal{S}} f(x, a)\, \tilde\nu(dx) = \int_{\mathcal{S}} f(x, a)\, \nu(dx)$, this shows (10). The proof of the remark now follows by the exact same arguments as in the proof of Theorem 2. ☐

A natural example, which highlights the importance of Remark 10 is the following:

*Example* 11. We let $\mu$ be a model for a vector of returns $X \in \mathcal{S} = \mathbb{R}^d$ and assume that $\mu$ is absolutely continuous with respect to Lebesgue measure. Let further $\|\cdot\| = |\cdot|$ and let $z \in B \subset \mathbb{R}^d$ denote a portfolio. We then consider the average value at risk at level $\alpha \in (0, 1)$ of the portfolio wealth $\langle z, X\rangle$, which can be written as

$$\mathrm{AV@R}_\alpha(\langle z, X\rangle) = \frac{1}{\alpha} \int_{1-\alpha}^1 \mathrm{V@R}_u(\langle z, X\rangle) du,$$

where $\mathrm{V@R}_u(\langle z, X\rangle)$ is the value at risk at level $u \in (0, 1)$ defined as

$$\mathrm{V@R}_u(\langle z, X\rangle) = \inf\{x \in \mathbb{R}^d \ : \mu(\langle z, x\rangle) \ge u\}.$$

We note that the average value at risk is an example for an optimized certainty equivalent (OCE), when choosing $l(x, a) = a + \frac{1}{\alpha}(x - a)^+$ in [1, p. 3]. We can thus rewrite the optimization problem

$$V(0) = \inf_{z \in B} \mathrm{AV@R}_\alpha(\langle z, X\rangle)$$

as

$$V(0) = \inf_{z\in B, m\in\mathbb{R}} \left( m + \frac{1}{\alpha}\int_{\mathcal{S}} \left(\langle z, x\rangle - m\right)^+ \mu(dx)\right).$$

Set $\mathcal{A} = B\times\mathbb{R}$ and assume that there exists a unique minimiser $(z^\star, m^\star)\in\mathcal{A}_0^\star$ of $V(0)$. Then $m^\star$ is given by $\mathrm{V@R}(\langle z^\star, X\rangle)$. The robust version of $V(0)$ reads

$$V(\delta) = \inf_{(z,m)\in\mathcal{A}} \sup_{\nu\in B_\delta(\mu)} \left( m + \frac{1}{\alpha}\int_{\mathcal{S}} \left(\langle z, x\rangle - m\right)^+ \nu(dx)\right).$$

Note that the function $x\mapsto x^+$ is weakly differentiable with weak derivative $\mathbf{1}_{\{x\geq 0\}}$. In conclusion $f(x,(z,m)) = m + \frac{1}{\alpha}\left(\langle z, x\rangle - m\right)^+$ has weak derivative

$$g(x,(z,m)) = \frac{1}{\alpha}\mathbf{1}_{\{\langle z,x\rangle - m\geq 0\}},$$

which is continuous at $(x,(h^\star,m^\star))$ except on the lower-dimensional set $\{x\in\mathcal{S}\ :\ \langle z^\star, x\rangle - m^\star = 0\}$, which is in particular a Lebesgue null set. Remark 10 thus yields

$$V'(0) = |z^\star|\left(\frac{1}{\alpha^q}\int_{\mathcal{S}} \mathbf{1}_{\{\langle z^\star x\rangle\geq \mathrm{V@R}_\alpha(\langle z^\star, X\rangle)\}}\mu(dx)\right)^{\frac{1}{q}} = \frac{|z^\star|}{\alpha^{1/p}}$$

and thus

$$V(\delta) = \mathrm{AV@R}_\alpha\left(\langle z^\star, X\rangle\right) + \frac{|z^\star|}{\alpha^{1/p}}\delta + o(\delta).$$

Comparing with [3, Table 1], we see that this approximation is actually exact for $p = 1, 2$.

We now mention two extensions of Theorem 2. The first one concerns the derivative of $V(\delta)$ for $\delta > 0$.

**Corollary 12.** *Fix $r > 0$ and in addition to the assumptions of Theorem 2 assume that*

- *$\mathcal{A}^\star_{r+\delta}\neq\emptyset$ for $\delta\geq 0$ small enough 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 $a_n^\star\in\mathcal{A}^\star_{r+\delta_n}$ there is a subsequence which converges to some $a^\star\in\mathcal{A}^\star_r$.*
- *there exists $\varepsilon > 0$ such that for all $\gamma > 0$ and every $a\in\mathcal{A}$ with $|a|\leq\gamma$ one has $|\nabla_x f(x,a)|\leq c(1+|x|^{p-1-\varepsilon})$ for all $x\in\mathcal{S}$ and some constant $c > 0$.*

*Then*

$$V'(r+) = \lim_{\delta\to 0}\frac{V(r+\delta) - V(r)}{\delta} = \inf_{a^\star\in\mathcal{A}^\star_r}\sup_{\nu\in B_r^\star(\mu,a^\star)}\left(\int_{\mathcal{S}} \|\nabla_x f(x,a^\star)\|^q\,\nu(dx)\right)^{1/q},$$

*where we recall that $B_r^\star(\mu,a^\star)$ is the set of all $\nu\in B_r(\mu)$ for which $\int_{\mathcal{S}} f(x,a^\star)\,\nu(dx) = V(r)$.*

*Remark* 13. Recall the notation $V(\delta,a)$ in (1.2). Inspecting the proof of the above Corollary, it is clear the main difficulty is in showing that

$$\lim_{\delta\to 0}\frac{V(r+\delta,a) - V(r,a)}{\delta} = \sup_{\nu\in B_r^\star(\mu,a)}\left(\int_{\mathcal{S}} \|\nabla_x f(x,a)\|^q\,\nu(dx)\right)^{1/q}.$$

In this way, the final statement of Corollary 12, or indeed of Theorem 2, can be interpreted as an instance of the envelope theorem.

The second extension of Theorem 2 offers a more specific sensitivity result by including additional constraints on the ball $B_\delta(\mu)$ of measures considered. Let $m\in\mathbb{N}$ and let $\Phi = (\Phi_1,\dots,\Phi_m)\colon\mathcal{S}\to\mathbb{R}^m$ be a family of $m$ functions and assume that $\mu$ is calibrated to $\Phi$ in the sense that $\int_{\mathcal{S}}\Phi(x)\,\mu(dx) = 0$. Consider the set

$$B_\delta^\Phi(\mu) := \left\{\nu\in B_\delta(\mu) : \int_{\mathcal{S}}\Phi(x)\,\nu(dx) = 0\right\}$$

and the corresponding optimization problem

$$V^\Phi(\delta) := \inf_{a\in\mathcal{A}}\sup_{\nu\in B_\delta^\Phi(\mu)}\int_{\mathcal{S}} f\left(x,a\right)\,\nu(dx).$$

We have the following result.

**Theorem 14** (Sensitivity of $V(\delta)$ under linear constraints)**.** *In addition to the assumptions of Theorem 2, assume that there is some small $\varepsilon > 0$ such that for every $a \in \mathcal{A}$ one has $|f(x,a)| \leq c(1+|x|^{p-\varepsilon})$ for all $x \in \mathbb{R}^d$ and some constant $c > 0$. Further assume that $\Phi_i$, $i \leq m$, are continuously differentiable with $|\Phi_i(x)| \leq c(1+|x|^{p-\varepsilon})$, $|\nabla_x \Phi_i(x)| \leq c(1+|x|^{p-1})$ and that the non-degeneracy condition*

$$\inf\left\{ \int_{\mathcal{S}} \left\| \sum_{i=1}^m \lambda_i \nabla_x \Phi_i(x) \right\|^q \mu(dx) : \lambda \in \mathbb{R}^d,\ |\lambda| = 1 \right\} > 0 \tag{11}$$

*holds. Then*

$$(V^\Phi)'(0) = \inf_{a^\star \in \mathcal{A}_0^\star} \inf_{\lambda \in \mathbb{R}^m} \left( \int_{\mathcal{S}} \left\| \nabla_x f(x,a^\star) + \sum_{i=1}^m \lambda_i \nabla_x \Phi_i(x) \right\|^q \mu(dx) \right)^{1/q}.$$

*Remark* 15. Note that if $\|\cdot\|$ is a norm and $\mu$ has full support, the above non-degeneracy condition (11) can be made without loss of generality. Indeed, as the unit circle is compact and the function $\lambda \mapsto \int_{\mathcal{S}} \|\sum_{i=1}^m \lambda_i \nabla_x \Phi_i(x)\|^q \, \mu(dx)$ is continuous, the infimum in (11) is attained. In particular, if

$$\inf\left\{ \int_{\mathcal{S}} \left\| \sum_{i=1}^m \lambda_i \nabla_x \Phi_i(x) \right\|^q \mu(dx) : |\lambda| = 1 \right\} = 0,$$

then $\sum_{i=1}^m \lambda_i \nabla_x \Phi_i = 0$ $\mu$-a.s. for some $\lambda$ in the unit circle. As $\mu$ has full support this implies that $\sum_{i=1}^m \lambda_i \nabla_x \Phi_i = 0$ on $\mathcal{S}$. Thus $\nabla_x \Phi_1, \dots, \nabla_x \Phi_m$ are linearly dependent functions on $\mathcal{S}$. Deleting all linearly dependent coordinates and calling the resulting vector $\tilde{\Phi}$, we have $V^\Phi(\delta) = V^{\tilde{\Phi}}(\delta)$ for every $\delta \geq 0$. Moreover, the non-degeneracy condition (11) holds for $\tilde{\Phi}$.

*Remark* 16. We can relax the conditions of Theorem 14 in the spirit of Remark 10: more specifically, assume that the baseline distribution $\mu$ is absolutely continuous w.r.t. the Lebesgue measure and $\|\cdot\| = |\cdot|$. Then Theorem 14 remains true if we merely assume that $f(\cdot,a)$ and $\Phi_i$ have a weak derivative (in the Sobolev sense) on $\mathcal{S}^o$ for all $a \in \mathcal{A}$ and replace $\nabla_x f(\cdot,a)$ and $\nabla \Phi_i$ by the weak derivative of $f(\cdot,a)$ and of $\Phi_i$ respectively. More concretely the assumption should read:

- The weak derivatives $(x,a) \mapsto g(x,a)$ of $f(\cdot,a)$ and $x \mapsto g_i(x)$ of $\Phi_i$ are continuous at every point $(x,a) \in N \times \mathcal{A}^\star(0)$, where $N$ is a Lebesgue-null set, and for every $r > 0$ there is $c > 0$ such that $|g_i(x,a)| \leq c(1+|x|^{p-1})$ and $|g_i(x)| \leq c(1+|x|^{p-1})$ for all $x \in \mathcal{S}$, $i = 1, \dots, m$ and $|a| \leq r$.

*Example* 17 (Martingale constraints). Let $d = 1$, $\mathcal{S} = \mathbb{R}$, $\|\cdot\| = |\cdot|$, $p = 2$, and let $\Phi_1(x) := x - x_0$ and $\Phi := \{\Phi_1\}$, i.e., $B_\delta^\Phi(\mu)$ corresponds to the measures $\nu \in B_\delta(\mu)$ satisfying the martingale (barycentre preservation) constraint $\int_{\mathbb{R}} x \, \nu(dx) = x_0$. Clearly the assumptions on $\Phi$ of Theorem 14 are satisfied. It remains to solve the optimization problem over $\lambda \in \mathbb{R}$ and plug in the optimizer. We then obtain

$$(V^\Phi)'(0) = \inf_{a^\star \in \mathcal{A}_0^\star} \left( \int_{\mathbb{R}} \left( \nabla_x f(x,a^\star) - \int_{\mathbb{R}} \nabla_x f(y,a^\star)\, \mu(dy) \right)^2 \mu(dx) \right)^{1/2},$$

i.e., $(V^\Phi)'(0)$ is the standard deviation of $\nabla_x f(\cdot,a^\star)$ under $\mu$. In line with the previous remark, this results extend to the case of the call option pricing discussed in the main body of the paper.

*Example* 18 (Covariance constraints). Let $d = 2$, $\mathcal{S} = \mathbb{R}^2$, $\|\cdot\| = |\cdot|$, $p = 2$. Further let $\Phi_1(x_1,x_2) := x_1 x_2 - b$ for some $b \in \mathbb{R}$ and $\Phi := \{\Phi_1\}$, i.e., we want to optimize over measures $\nu \in B_\delta(\mu)$ satisfying the covariance constraint $\int_{\mathbb{R}^2} x_1 x_2 \, \nu(dx) = b$. Assume that there exists no $\lambda \in \mathbb{R} \setminus \{0\}$ such that $\mu$-a.s. $x_1 = \lambda x_2$. Clearly the assumptions on $\Phi$ of Theorem 14 are satisfied. Note that

$$\begin{aligned} &\int_{\mathbb{R}^2} |\nabla_x f(x,a) + \lambda_1 \nabla_x \Phi_1(x)|^2 \, \mu(dx) \\ &= \int_{\mathbb{R}^2} (\nabla_{x_1} f(x,a) + \lambda_1 x_2)^2 + (\nabla_{x_2} f(x,a) + \lambda_1 x_1)^2 \, \mu(dx), \end{aligned}$$

so in particular the optimal $\lambda$ in the definition of $(V^\Phi)'(0)$ is given by

$$\lambda_1 = \frac{-\int_{\mathbb{R}^2} \nabla_{x_1} f(x,a)x_2 + \nabla_{x_2} f(x,a)x_1 \, \mu(dx)}{\int_{\mathbb{R}^2} x_1^2 + x_2^2 \, \mu(dx)}.$$

Plugging this in gives

$$\begin{aligned}\int_{\mathbb{R}^2} |\nabla_x f(x,a) + \lambda_1 \nabla_x \Phi_1(x)|^2 \, \mu(dx) &= \int_{\mathbb{R}^2} (\nabla_{x_1} f(x,a))^2 + (\nabla_{x_2} f(x,a))^2 \, \mu(dx) \\ &\quad + 2\lambda_1 \int_{\mathbb{R}^2} (\nabla_{x_1} f(x,a)x_2 + \nabla f_{x_2} f(x,a)x_1)\mu(dx) + \lambda_1^2 \int_{\mathbb{R}^2} x_2^2 + x_1^2 \, \mu(dx) \\ &= \int_{\mathbb{R}^2} (\nabla_{x_1} f(x,a))^2 + (\nabla_{x_2} f(x,a))^2 \, \mu(dx) - \frac{\left(\int_{\mathbb{R}^2} (\nabla_{x_1} f(x,a)x_2 + \nabla_{x_2} f(x,a)x_1) \, \mu(dx)\right)^2}{\int_{\mathbb{R}^2} (x_1^2 + x_2^2) \, \mu(dx)}.\end{aligned}$$

It follows that

$$\begin{aligned}(V^\Phi)'(0) = \inf_{a^\star \in \mathcal{A}_0^\star} \Bigg( &\int_{\mathbb{R}^2} |\nabla_x f(x,a^\star)|^2 \, \mu(dx) \\ &- \frac{\left(\int_{\mathbb{R}^2} \nabla_{x_1} f(x,a^\star)x_2 + \nabla_{x_2} f(x,a^\star)x_1 \, \mu(dx)\right)^2}{\int_{\mathbb{R}^2} |x|^2 \, \mu(dx)} \Bigg)^{1/2}.\end{aligned}$$

*Example* 19 (Calibration). Consider the function $f((T,K),a) = (E_{\mathbb{P}_a}[(S_T - K)^+] - C((T,K))^2$, the discrete measure $\mu$ formalises grid points for which option data $C(T,K)$ is available, $\mathcal{S} \subset \mathbb{R}_+ \times \mathbb{R}_+$ is the set of maturities and strikes of interest and $\{\mathbb{P}_a, a \in \mathcal{A}\}$, for a given compact set $\mathcal{A}$, is a class of parametric models (e.g., Heston). A Wasserstein ball around $\mu$ can then be seen as a plausible formalisation of market data uncertainty. Derivatives in $T$ and $K$ correspond to classical pricing sensitivities, which are readily available for most common parametric models. These have to be only evaluated for one model $\mathbb{P}_{a^\star}$. Changing the class of parametric models $\{\mathbb{P}_a, a \in A\}$ and computing the sensitivity in Theorem 2 could then yield insights into when a calibration procedure can be considered reasonably robust.

## 2.2. **Proofs and auxiliary results related to Theorem 2.**

*Proof of Theorem 2.* We present now a complete proof of Theorem 2 for general state space $\mathcal{S}$ and semi-norm $\|\cdot\|$. All the essential ideas have already been outlined in [1] but for the convenience of the reader we repeat all of the steps as opposed to only detailing where the general case differs from the one treated in [1].

**Step 1:** Let us first assume that $\mathcal{S} = \mathbb{R}^d$. For every $\delta \geq 0$ let $C_\delta(\mu)$ denote those $\pi \in \mathcal{P}(\mathcal{S} \times \mathcal{S})$ which satisfy

$$\pi_1 = \mu \text{ and } \left( \int_{\mathcal{S} \times \mathcal{S}} \|x - y\|_*^p \, \pi(dx,dy) \right)^{1/p} \leq \delta.$$

Note that the dual norm $\|\cdot\|_*$ is lower semicontinuous, which implies that the infimum in the definition of $W_p(\mu,\nu)$ is attained (see [4, 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 $a^\star \in \mathcal{A}_0^\star$ one has $V(\delta) \leq \sup_{\nu \in B_\delta(\mu)} \int_{\mathcal{S}} f(y,a^\star) \, \nu(dy)$ with equality for $\delta = 0$. Therefore, differentiating $f(\cdot,a^\star)$ and using Fubini's theorem, we obtain that

$$\begin{aligned}V(\delta) - V(0) &\leq \sup_{\pi \in C_\delta(\mu)} \int_{\mathcal{S} \times \mathcal{S}} f(y,a^\star) - f(x,a^\star) \, \pi(dx,dy) \\ &= \sup_{\pi \in C_\delta(\mu)} \int_0^1 \int_{\mathcal{S}} \langle \nabla_x f(x + t(y-x), a^\star), (y-x) \rangle \, \pi(dx,dy)dt.\end{aligned}$$

Now recall that $\langle x,y\rangle \le \|x\|\|y\|_*$ for every $x,y\in\mathbb{R}^d$, whence for any $\pi\in C_\delta(\mu)$ and $t\in[0,1]$, we have that

$$\begin{aligned}&\int_{\mathcal{S}} \langle \nabla_x f(x+t(y-x),a^\star),(y-x)\rangle\,\pi(dx,dy)\\ &\le \int_{\mathcal{S}} \|\nabla_x f(x+t(y-x),a^\star)\|\|y-x\|_*\,\pi(dx,dy)\\ &\le \Big(\int_{\mathcal{S}} \|\nabla_x f(x+t(y-x),a^\star)\|^q\,\pi(dx,dy)\Big)^{1/q}\Big(\int_{\mathcal{S}}\|y-x\|^p\,\pi(dx,dy)\Big)^{1/p},\end{aligned}$$

where we used Hölder's inequality to obtain the last inequality. By definition of $C_\delta(\mu)$ the last integral is smaller than $\delta$ and we end up with

$$V(\delta)-V(0)\le \delta \sup_{\pi\in C_\delta(\mu)}\int_0^1\Big(\int_{\mathcal{S}}\|\nabla_x f(x+t(y-x),a^\star)\|^q\pi(dx,dy)\Big)^{1/q}dt.$$

It remains to show that the last term converges to the integral under $\mu$. To that end, note that any choice $\pi^\delta\in C_\delta(\mu)$ converges in $W_p^{|\cdot|}$ on $\mathcal{P}(\mathcal{S}\times\mathcal{S})$ to the pushforward measure of $\mu$ under the mapping $x\to(x,x)$, which we denote $[x\mapsto(x,x)]_*\mu$. This can be seen by, e.g., considering the coupling $[(x,y)\mapsto(x,y,x,x)]_*\pi^\delta$ between $\pi^\delta$ and $[x\mapsto(x,x)]_*\mu$. Now note that, together with growth restriction on $\nabla_x f$ of Assumption 1, $q=p/(p-1)$ implies

$$\|\nabla_x f(x+t(y-x),a^\star)\|^q\le c(1+|x|^p+|y|^p) \tag{12}$$

for some $c>0$ and all $x,y\in\mathbb{R}^d$, $t\in[0,1]$. Recall that there furthermore exists $\tilde c>0$ such that $\|x\|\le\tilde c|x|$, in particular $\int_{\mathcal{S}}\|\nabla_x f(x+t(y-x),a^\star)\|^q\,\pi^\delta(dx,dy)\le C$ for all $t\in[0,1]$ and small $\delta>0$, for another constant $C>0$. As Assumption 1 further yields continuity of $(x,y)\mapsto\|\nabla_x f(x+t(y-x),a^\star)\|^q$ for every $t$, the $p$-Wasserstein convergence of $\pi^\delta$ to $[x\mapsto(x,x)]_*\mu$ implies that

$$\int_{\mathcal{S}}\|\nabla_x f(x+t(y-x),a^\star)\|^q\,\pi(dx,dy)\to\int_{\mathcal{S}}\|\nabla_x f(x,a^\star)\|^q\,\mu(dx)$$

for every $t\in[0,1]$, see Lemma 20. 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 in 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 the second part of Assumption 1, for a (again not relabelled) subsequence, one has $a^\star_\delta\to a^\star\in\mathcal{A}^\star_0$. Further note that $V(0)\le\int_{\mathcal{S}} f(x,a^\star_\delta)\,\mu(dx)$ which implies

$$V(\delta)-V(0)\ge\sup_{\pi\in C_\delta(\mu)}\int_{\mathcal{S}\times\mathcal{S}} f(y,a^\star_\delta)-f(x,a^\star_\delta)\,\pi(dx,dy).$$

By Lemma 6 there exists a function $h\colon\mathbb{R}^d\mapsto\{x\in\mathbb{R}^d:\|x\|_*=1\}$ such that $\|x\|=\langle x,h(x)\rangle$ for every $x\in\mathbb{R}^d$. Now define

$$\begin{aligned}\pi^\delta&:=[x\mapsto(x,x+\delta T(x))]_*\mu,\quad\text{where}\\ T(x)&:=\frac{h(\nabla_x f(x,a^\star))}{\|\nabla_x f(x,a^\star)\|^{1-q}}\Big(\int_{\mathcal{S}}\|\nabla_x f(z,a^\star)\|^q\,\mu(dz)\Big)^{1/q-1}\end{aligned}$$

for $x\in\mathbb{R}^d$ with the convention $h(\cdot)/0=0$. Note that the integral is well defined since, as before in (12), one has $\|\nabla_x f(x,a^\star)\|^q\le C(1+|x|^p)$ for some $C>0$ and the latter is integrable under $\mu$. Using that $pq-p=q$ it further follows that

$$\begin{aligned}&\int_{\mathcal{S}\times\mathcal{S}}\|x-y\|_*^p\,\pi^\delta(dx,dy)=\delta^p\int_{\mathcal{S}}\|T(x)\|_*^p\,\mu(dx)\\ &=\delta^p\frac{\int_{\mathcal{S}}\|\nabla_x f(x,a^\star)\|^{pq-p}\,\mu(dx)}{\big(\int_{\mathcal{S}}\|\nabla_x f(z,a^\star)\|^q\,\mu(dz)\big)^{p(1-1/q)}}=\delta^p.\end{aligned}$$

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

$$\begin{aligned}\frac{V(\delta)-V(0)}{\delta} &\geq \frac{1}{\delta}\int_{\mathcal{S}} f(x+\delta T(x), a_\delta^\star) - f(x, a_\delta^\star)\,\mu(dx)\\ &= \int_0^1 \int_{\mathcal{S}} \langle \nabla_x f(x+t\delta T(x), a_\delta^\star), T(x)\rangle\,\mu(dx)\,dt.\end{aligned}$$

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

$$\int_{\mathcal{S}} \langle \nabla_x f(x,a^\star), T(x)\rangle\,\mu(dx) = \Big(\int_{\mathcal{S}} \|\nabla_x f(x,a^\star)\|^q\,\mu(dx)\Big)^{1/q}.$$

The last equality follows from the definition of $T$ and a simple calculation. To justify the convergence, first note that

$$\langle \nabla_x f(x+t\delta T(x), a_\delta^\star), T(x)\rangle \to \langle \nabla_x f(x,a^\star), T(x)\rangle$$

for all $x\in\mathbb{R}^d$ by continuity of $(a,x)\mapsto \nabla_x f(x,a)$ and since $a_\delta^\star\to a^\star$. Moreover, as before in (12), one has

$$\|\langle \nabla_x f(x+t\delta T(x), a^\star), T(x)\rangle\| \leq C(1+|x|^p)$$

for some $C>0$ and all $t\in[0,1]$. The latter is integrable under $\mu$, hence convergence of the integrals follows from the dominated convergence theorem.

**Step 2:** We now extend the proof to the case, where $\mathcal{S}\subset\mathbb{R}^d$ is closed convex and its boundary has zero measure under $\mu$.
Note that the proof of the "$\leq$"-inequality remains unchanged. We modify the proof of the "$\geq$"-inequality as follows: let us first define

$$\mathcal{S}^\varepsilon := \{x\in\mathcal{S} \;:\; |x-z|\geq\varepsilon \text{ for all } z\in\mathcal{S}^c\}$$

for all $\varepsilon>0$, so that in particular $\bigcup_{\varepsilon>0}\mathcal{S}^\varepsilon = \mathcal{S}^o$. We now redefine

$$\pi^\delta := \Big[x\mapsto \Big(x, x+\delta T(x)\mathbf{1}_{\{x\in\mathcal{S}^{\sqrt{\delta}}\}}\mathbf{1}_{\{|T(x)|\leq 1/\sqrt{\delta}\}}\Big)\Big]_*\mu.$$

Then $\pi^\delta\in\mathcal{P}(\mathcal{S}\times\mathcal{S})$ and in particular $\pi^\delta\in C_\delta(\mu)$ as in Step 1. Noting that

$$\lim_{\delta\to 0} T(x)\mathbf{1}_{\{x\in\mathcal{S}^{\sqrt{\delta}}\}}\mathbf{1}_{\{|T(x)|\leq 1/\sqrt{\delta}\}} = T(x)\mathbf{1}_{\{x\in\mathcal{S}^o\}},$$

the remaining steps of the proof follow as in Step 1. This concludes the proof. □

**Lemma 20.** *Let $p\in[1,\infty)$, let $a_0\in\mathcal{A}$ and assume that $f$ is continuous and, for some constant $c>0$, satisfies $|f(x,a)|\leq c(1+|x|^p)$ for all $x\in\mathcal{S}$ and all $a$ in a neighborhood of $a_0$. Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of probability measures which converges to some $\mu$ w.r.t. $W_p^{|\cdot|}$ and $(a_n)_{n\in\mathbb{N}}$ be a sequence which converges to $a_0$. Then $\int_{\mathcal{S}} f(x,a_n)\,\mu_n(dx)\to\int_{\mathcal{S}} f(x,a_0)\,\mu(dx)$ as $n\to\infty$.*

*Proof.* Let $K$ be a small neighborhood of $a_0$ such that $|f(x,a)|\leq c(1+|x|^p)$ for all $x\in\mathcal{S}$ and $a\in K$. The measures $\mu_n\otimes\delta_{a_n}$ converge in $W_p^{|\cdot|}$ to the measure $\mu\otimes\delta_{a_0}$. As $\int_{\mathcal{S}} f(x,a_n)\,\mu_n(dx) = \int_{\mathcal{S}\times K} f(x,a)\,(\mu_n\otimes\delta_{a_n})(d(x,a))$ and similarly for $\mu\otimes\delta_{a_0}$, the claim follows from [4, Lemma 4.3, p.43]. □

The following lemma relates to the financial economics applications described in [1]. We focus on a sufficient condition for the second part of Assumption 1. For this, we assume that $\mu$ does not contain any redundant assets, i.e. $\mu(\{x\in\mathbb{R}^d : \langle a, x-x_0\rangle>0\})>0$ for every $a\neq 0$. If $\mu$ satisfies this condition, we call it non-degenerate. Note that this condition is slightly stronger than no-arbitrage. However, if $\mu$ satisfies no arbitrage, then one can always delete the redundant dimensions in $\mu$ similarly to the remark after Theorem 14, so that the modified measure satisfies $\mu(\{x\in\mathbb{R}^d : \langle a, x-x_0\rangle>0\})>0$ for every $a\neq 0$.

**Lemma 21.** *Assume that $l\colon\mathbb{R}\to\mathbb{R}$ is convex, increasing, bounded from below and $f(x,a):=l(g(x)+\langle a,x\rangle)$ satisfies the first part of Assumption 1. Furthermore assume that $\mu$ is non-degenerate in the above sense. Then for every $\delta\geq 0$ there exists an optimizer $a_\delta^\star\in\mathbb{R}^d$ for $V(\delta)$, i.e.,*

$$V(\delta) = \sup_{\nu\in B_\delta(\mu)} \int_{\mathbb{R}^d} l(g(x)+\langle a_\delta^\star, x-x_0\rangle)\,\nu(dx) < \infty.$$

*Furthermore, if $l$ is strictly convex, the optimizer $a^\star$ of $V(0)$ is unique and $a^\star_\delta \to a^\star$ as $\delta \to 0$. In particular, Assumption 1 is satisfied.*

*Proof.* The first statement is trivially true if $l$ is constant, so assume otherwise in the following. Moreover, note by the first part of Assumption 1 we have $V(\delta) < \infty$ for all $\delta \geq 0$. Now fix $\delta \geq 0$, and let $(a_n)_{n\in\mathbb{N}}$ be a minimizing sequence, i.e.

$$V(\delta) = \lim_{n\to\infty} \sup_{\nu \in B_\delta(\mu)} \int_{\mathbb{R}^d} l(g(x) + \langle a_n, x - x_0\rangle)\, \nu(dx).$$

If $(a_n)_{n\in\mathbb{N}}$ is bounded, then after passing to a subsequence there is a limit, and Fatou's lemma shows that this limit is a minimizer. It remains to argue why $(a_n)_{n\in\mathbb{N}}$ is bounded. Heading for a contradiction, assume that $|a_n| \to \infty$ as $n \to \infty$. After passing to a (not relabeled) subsequence, there is $\tilde{a} \in \mathbb{R}^d$ with $|\tilde{a}| = 1$ such that $a_n/|a_n| \to \tilde{a}$ as $n \to \infty$. By our assumption we have $\mu(\{x \in \mathbb{R}^d : \langle \tilde{a}, x - x_0\rangle > 0\}) > 0$. As $l$ is bounded below this shows that

$$\sup_{\nu \in B_\delta(\mu)} \int_{\mathbb{R}^d} l(g(x) + \langle a_n, x - x_0\rangle)\, \nu(dx) \geq \int_{\mathbb{R}^d} l(g(x) + \langle a_n, x - x_0\rangle)\, \mu(dx) \to \infty,$$

as $n \to \infty$, a contradiction.

To prove the second claim note that strict convexity of $l$ readily implies that $V(0)$ admits a unique minimizer $a^\star$. Now, heading for a contraction, assume that there exists a subsequence $(\delta_n)_{n\in\mathbb{N}}$ converging to zero, such that $a^\star_{\delta_n}$ does not converge to $a^\star$. The exact same reasoning as above shows that $(a^\star_{\delta_n})_{n\in\mathbb{N}}$ is bounded, hence (possibly after passing to a not relabeled subsequence) there is a limit $\tilde{a} \neq a^\star$. Using Fatou's lemma once more implies

$$\begin{aligned} V(0) &< \int_{\mathbb{R}^d} l(g(x) + \langle \tilde{a}, x - x_0\rangle)\, \mu(dx) \\ &\leq \liminf_{n\to\infty} \int_{\mathbb{R}^d} l(g(x) + \langle a^\star_{\delta_n}, x - x_0\rangle)\, \mu(dx) \leq \liminf_{n\to\infty} V(\delta_n). \end{aligned}$$

On the other hand, plugging $a^\star$ into $V(\delta)$ implies

$$\limsup_{n\to\infty} V(\delta_n) \leq \limsup_{n\to\infty} \sup_{\nu \in B_{\delta_n}(\mu)} \int_{\mathbb{R}^d} l(g(x) + \langle a^\star, x - x_0\rangle)\, \nu(dx) = V(0),$$

which follows from as $l(g(x)+\langle a^\star, x-x_0\rangle) \leq c(1+|x|^p)$ and that any $\nu_n \in B_{\delta_n}(\mu)$ converges in $W_p^{|\cdot|}$ to $\mu$ by definition. This gives the desired contraction. ◻

In analogy to the above result, the following summarizes simple sufficient conditions for the second part of Assumption 1.

**Lemma 22.** *Assume that either $\mathcal{A}$ is compact or that $a \mapsto V(0,a)$ is coercive, in the sense that $V(0,a_n) \to \infty$ if $|a_n| \to \infty$. Moreover, assume that $f$ is continuous, such that $f(x,a) \leq c(1+|x|^p)$ for some $c \geq 0$. Then the second part of Assumption 1 is satisfied.*

*Proof.* Let us first note that for fixed $\delta \geq 0$ the function $a \mapsto V(\delta, a)$ is lower semiconinuous as a supremum of continuous functions $a \mapsto \int f(x,a)\, \nu(dx)$ for $\nu \in B_\delta(\mu)$. Next we note that $\mathcal{A}^\star(\delta) \neq \emptyset$. Indeed, if $\mathcal{A}$ is compact, this directly follows from lower semicontinuity of $a \mapsto V(\delta, a)$. Otherwise, the fact that $V(\delta, a) \geq V(0, a)$ for all $a \in \mathcal{A}$ and coercivity imply that any minimising sequence $(a_n)_{n\in\mathbb{N}}$ is bounded. Lastly, we show that any accumulation point of such a sequence is an element of $\mathcal{A}^\star(0)$. By the above we can assume (by taking a subsequence without relabelling if necessary) that $\lim_{n\to\infty} a_n = a \in \mathcal{A}$. If $a \notin \mathcal{A}^\star$, then

$$\liminf_{n\to\infty} V(\delta_n, a_n) \geq \lim_{n\to\infty} V(0, a_n) = V(0, a) > V(0, a^\star) = \lim_{n\to\infty} V(\delta_n, a^\star)$$

for any $a^\star \in \mathcal{A}^\star(0)$. This contradicts $a_n \in \mathcal{A}^\star(\delta_n)$ for all $n \in \mathbb{N}$ and concludes the proof. ◻

*Proof of Corollary 12.* We start with the "$\le$"-inequality. First, note that for any $\delta > 0$, $a^r \in \mathcal{A}_r^\star$, and $\nu^{r+\delta} \in B_{r+\delta}^\star(\mu, a^r)$, we have

$$V(r+\delta) \le V(r+\delta, a^r) = \int_{\mathcal{S}} f(x, a^r)\, \nu^{r+\delta}(dx),$$
$$V(r) \ge \sup_{\nu \in B_r(\mu)\cap B_\delta(\nu^{r+\delta})} \int_{\mathcal{S}} f(x, a^r)\, \nu(dx).$$

This implies that

$$\begin{aligned} V(r+\delta) - V(r) &\le \sup_{\pi \in C_\delta(\nu^{r+\delta})} \int_{\mathcal{S}\times\mathcal{S}} f(x,a^r) - f(y,a^r)\, \pi(dx,dy) \\ &= \sup_{\pi \in C_\delta(\nu^{r+\delta})} \int_0^1 \int_{\mathcal{S}\times\mathcal{S}} \langle \nabla_x f(y+t(x-y), a^r), (x-y)\rangle\, \pi(dx,dy)\, dt \\ &\le \delta \sup_{\pi \in C_\delta(\nu^{r+\delta})} \int_0^1 \left( \int_{\mathcal{S}\times\mathcal{S}} \|\nabla_x f(y+t(x-y), a^r)\|^q\, \pi(dx,dy)\right)^{1/q} dt. \end{aligned} \tag{13}$$

Note that the assumption $|\nabla_x f(x,a)| \le c(1+|x|^{p-1-\varepsilon})$ implies $|\nabla_x f(x,a)|^q \le c(1+|x|^{\frac{p(p-1-\varepsilon)}{p-1}})$ ( for some new constant $c$). To simplify notation let us thus define $\tilde\varepsilon = (p-1-\varepsilon)/(p-1) < 1$ and recall that $B_{r+1}(\mu)$ is compact w.r.t. $W_{p\tilde\varepsilon}^{|\cdot|}$ by Lemma 23, hence there is $\tilde\nu^r \in B_r(\mu)$ such that (after passing to a subsequence) $\nu^{r+\delta} \to \tilde\nu^r$ w.r.t. $W_{p\tilde\varepsilon}^{|\cdot|}$ as $\delta \to 0$. The same arguments as in the proof of Theorem 2 show that (13) (divided by $\delta$) converges to $\left(\int_{\mathcal{S}} \|\nabla_x f(x,a^r)\|^q\, \tilde\nu^r(dx)\right)^{1/q}$ when $\delta \to 0$. So, to conclude the "$\le$"-part, all that is left to do is show that $\tilde\nu^r \in B_r^\star(\mu, a^r)$, which follows as

$$V(r) \le \lim_{\delta\to 0} V(r+\delta) \le \lim_{\delta \to 0} \int_{\mathcal{S}} f(x,a^r)\nu^{r+\delta}(dx) = \int_{\mathcal{S}} f(x,a^r)\, \tilde\nu^r(dx) \le V(r).$$

We now turn to the proof of the "$\ge$"-inequality. To that end, let $(a^{r+\delta})_{\delta>0}$ be a sequence of optimizers, i.e. $a^{r+\delta} \in \mathcal{A}_{r+\delta}^\star$ for all $\delta > 0$. Then by assumption there exists $a^r \in \mathcal{A}_r^\star$ such that (after passing to a subsequence) $\lim_{\delta\to 0} a^{r+\delta} = a^r$. Let $\nu^r \in B_r^\star(\mu, a^r)$ be arbitrary. As $B_\delta(\nu^r) \subset B_{r+\delta}(\mu)$ (by the triangle inequality) we have

$$V(r+\delta) \ge \sup_{\nu \in B_\delta(\nu^r)} \int_{\mathcal{S}} f(x, a^{r+\delta})\, \nu(dx).$$

As further (trivially) $V(r) \le \int_{\mathcal{S}} f(x,a^{r+\delta})\, \nu^r(dx)$ we conclude

$$\begin{aligned} \frac{V(\delta+r) - V(r)}{\delta} &\ge \sup_{\nu\in B_\delta(\nu^r)} \frac{1}{\delta} \int_{\mathcal{S}} f(x,a^{r+\delta})\nu(dx) - \int_{\mathcal{S}} f(x,a^{r+\delta})\nu^r(dx) \\ &\to \left( \int_{\mathcal{S}} \|\nabla_x f(x,a^r)\|^q\, \nu^r(dx)\right)^{1/q}, \end{aligned}$$

as $\delta \to 0$, where the the last equality follows from the exact same arguments as presented int he proof of Theorem 2. As $\nu^r \in B_r^\star(\mu, a^r)$ was arbitrary, the claim follows. $\square$

*Proof of Theorem 14.* We start by showing the easier estimate

$$\begin{aligned} &\limsup_{\delta\to 0} \frac{V^\Phi(\delta) - V^\Phi(0)}{\delta} \\ &\le \inf_{a^\star \in \mathcal{A}_0^\star} \inf_{\lambda\in\mathbb{R}^m} \left( \int_{\mathcal{S}} \left\| \nabla_x f(x,a^\star) + \sum_{i=1}^m \lambda_i \nabla_x \Phi_i(x)\right\|^q \mu(dx)\right)^{1/q}. \end{aligned} \tag{14}$$

To that end, let $a^\star \in \mathcal{A}_0^\star$ and $\lambda \in \mathbb{R}^m$ by arbitrary. Then $V^\Phi(0) = \int_{\mathcal{S}} f(x,a^\star) + \sum_{i=1}^m \lambda_i \Phi_i(x)\, \mu(dx)$. Moreover, as $B_\delta^\Phi(\mu) \subset B_\delta(\mu)$, it further follows that $V^\Phi(\delta) \le \sup_{\nu\in B_\delta(\mu)} \int_{\mathcal{S}} f(y,a^\star) + \sum_{i=1}^m \lambda_i \Phi_i(y)\, \nu(dy)$. Therefore (14) is a consequence of Theorem 2 (applied to the function $\tilde f(x,a) := f(x,a^\star) + \sum_{i=1}^m \lambda_i \Phi_i(x)$).

To show the other direction, i.e. that

$$\begin{aligned}\liminf_{\delta\to 0} \frac{V^\Phi(\delta)-V^\Phi(0)}{\delta} \\ \geq \inf_{a^\star\in\mathcal{A}_0^\star}\inf_{\lambda\in\mathbb{R}^m}\left(\int_{\mathcal{S}\times\mathcal{S}}\left\|\nabla_x f(x,a^\star)+\sum_{i=1}^m\lambda_i\nabla_x\Phi_i(x)\right\|^q\mu(dx)\right)^{1/q}.\end{aligned} \tag{15}$$

pick a (not relabeled) subsequence of $(\delta)_{\delta>0}$ which converges to the liminf. For $a_\delta^\star\in\mathcal{A}_\delta^\star$, there is another (again not relabeled) subsequence which converges to some $a^\star\in\mathcal{A}_0^\star$. From now on stick to this subsequence. In a first step, notice that

$$\begin{aligned}V^\Phi(\delta)&=\sup_{\nu\in B_\delta(\mu)}\inf_{\lambda\in\mathbb{R}^m}\int_{\mathcal{S}} f(y,a_\delta^\star)+\sum_{i=1}^m\lambda_i\Phi_i(y)\,\nu(dy)\\ &=\inf_{\lambda\in\mathbb{R}^m}\sup_{\nu\in B_\delta(\mu)}\int_{\mathcal{S}} f(y,a_\delta^\star)+\sum_{i=1}^m\lambda_i\Phi_i(y)\,\nu(dy).\end{aligned} \tag{16}$$

Indeed, this follows from a minimax theorem (see [5, Cor. 2, p. 411]) and appropriate compactness of $B_\delta(\mu)$ as stated in Lemma 23. For notational simplicity let $\lambda_\delta^\star$ be an optimizer for (16). Then

$$\begin{aligned}&\frac{V^\Phi(\delta)-V^\Phi(0)}{\delta}\\ &\geq\frac{1}{\delta}\sup_{\pi\in C_\delta(\mu)}\int_{\mathcal{S}\times\mathcal{S}} f(y,a_\delta^\star)-f(x,a_\delta^\star)+\sum_{i=1}^m\lambda_{\delta,i}^\star(\Phi_i(y)-\Phi_i(x))\,\pi(dx,dy),\end{aligned} \tag{17}$$

where we used that $V^\Phi(0)\leq\int_{\mathcal{S}} f(x,a_\delta^\star)+\sum_{i=1}^m\lambda_{\delta,i}^\star\Phi_i(x)\,\mu(dx)$. Now, in case that $\lambda_\delta^\star$ is uniformly bounded for all small $\delta>0$, after passing to a subsequence, it converges to some $\lambda^\star$. Then it follows from the exact same arguments as used in the proof of Theorem 2 that

$$\begin{aligned}&\liminf_{\delta\to 0}\frac{1}{\delta}\sup_{\pi\in C_\delta(\mu)}\int_{\mathcal{S}\times\mathcal{S}} f(y,a_\delta^\star)-f(x,a_\delta^\star)+\sum_{i=1}^m\lambda_{\delta,i}^\star(\Phi_i(y)-\Phi_i(x))\,\pi(dx,dy)\\ &\geq\Big(\int_{\mathcal{S}}\Big\|\nabla_x f(x,a^\star)+\sum_{i=1}^m\lambda_i^\star\nabla_x\Phi_i(x)\Big\|^q\mu(dx)\Big)^{1/q}\end{aligned}$$

which shows (15). It remains to argue why $\lambda_\delta^\star$ is bounded for small $\delta>0$. By (17) and the estimate "$\sup(A+B)\geq\sup A+\inf B$" we have

$$\begin{aligned}\frac{V^\Phi(\delta)-V^\Phi(0)}{\delta}&\geq\frac{1}{\delta}\sup_{\pi\in C_\delta(\mu)}\int_{\mathcal{S}\times\mathcal{S}}\sum_{i=1}^m\lambda_{\delta,i}^\star(\Phi_i(y)-\Phi_i(x))\,\pi(dx,dy)\\ &\quad+\frac{1}{\delta}\inf_{\pi\in C_\delta(\mu)}\int_{\mathcal{S}\times\mathcal{S}} f(y,a_\delta^\star)-f(x,a_\delta^\star)\,\pi(dx,dy).\end{aligned}$$

The second term converges to $-(\int_{\mathcal{S}}\|\nabla_x f(x,a^\star)\|^q\,\mu(dx))^{1/q}$ (see the proof of Theorem 2), in particular it is bounded for all $\delta>0$ small. On the other hand by (11) and continuity as well as growth of $x\mapsto\nabla_x\Phi_i(x)$, the first term is larger than $c|\lambda_\delta^\star|$ for some $c>0$. By (14) this implies that $(\lambda_\delta^\star)_{\delta>0}$ must be bounded for small $\delta>0$. ☐

We have used the following lemma:

**Lemma 23.** *Let $p,q\in[1,\infty)$ such that $q<p$ and let $\mu$ be a probability measure on $\mathcal{S}$. Then $p$-Wasserstein ball $B_\delta(\mu)$ is compact w.r.t. $W_q^{|\cdot|}$.*

*Proof.* We recall that $\|\cdot\|_*$ is lower semicontinuous and there exists $c>0$ such that $|x|\leq c\|x\|_*$ for all $x\in\mathbb{R}^d$. As $\int_{\mathcal{S}}|x|^p\,\mu(dx)<\infty$ by assumption, an application of Prokhorov's theorem shows that $B_\delta(\mu)$ is weakly precompact (recall the convention that we continuity is defined for $(\mathbb{R}^d,|\cdot|)$). Hence, for every sequence of measures $(\nu_n)_{n\in\mathbb{N}}$ in $B_\delta(\mu)$ there exists a subsequence, which we also call $(\nu_n)_{n\in\mathbb{N}}$ and a measure $\nu$ such that $\nu_n$ converges weakly to

$\nu$. As $W_p$ is weakly lower semicontinuous (see [4, Lemma 4.3, p.43]), this implies $\nu \in B_\delta(\mu)$. Applying the same argument to the tight sequence $(\tilde{\nu}_n)_{n\in\mathbb{N}}$ defined via

$$\tilde{\nu}_n(dx) := \frac{|x|^q}{\int_{\mathcal{S}} |y|^q \, \nu_n(dy)} \nu_n(dx)$$

we conclude that there exists another subsequence of $(\nu_n)_{n\in\mathbb{N}}$ which also converges in $W_q^{|\cdot|}$. This concludes the proof. □

## 3. Discussion, proofs and auxiliary results related to Theorem 4

3.1. **Further discussion of Theorem 4.** We note that a natural way to compute the sensitivity of $a^\star_\delta$ would be by combining Theorem 2 with chain rule and differentiation of the function $V(a,\delta)$. This cannot however be rigorously justified as the following remark demonstrates.

*Remark* 24. Let us point out that it is not true that $a \mapsto V(a,\delta)$ is differentiable for $\delta > 0$ under the sole assumption that $(x,a) \mapsto f(x,a)$ is sufficiently smooth and $\nabla_a^2 f \neq 0$.

To give an example, let $\mathcal{S} = \mathbb{R}$, $\|\cdot\| = |\cdot|$, $\mathcal{A} = \mathbb{R}$ and take $f(x,a) := ax + a^2$ and $\mu = \delta_0$. A quick computation shows $V(\delta,a) = \delta|a| + a^2$ (independently of $p$). In particular $V(\delta) = 0$ and $a^\star_\delta = a^\star = 0$ for all $\delta > 0$ and $a \mapsto V(\delta,a)$ is clearly not differentiable in $a = 0$.

Instead, we use a more involved argument, combining differentiability of $a \mapsto V(0,a)$ with a Lagrangian approach. This however requires slightly stricter growth assumptions than the ones imposed in Assumption 1, which are specified in Assumption 3.

*Example* 25. We provide detailed computations behind the square-root LASSO/Ridge regression example discussed in [1]. We consider $\mathcal{A} = \mathbb{R}^k$, $\mathcal{S} = \mathbb{R}^{k+1}$. We fix norms $\|(x,y)\| = |x|_s$, $\|(x,y)\|_* = |x|_r \mathbf{1}_{\{y=0\}} + \infty \mathbf{1}_{\{y\neq 0\}}$, for some $s > 1$, $1/s + 1/r = 1$ and $(x,y) \in \mathbb{R}^k \times \mathbb{R}$. We recall than then (3.3) holds and we can apply our methodology for $f((x,y),a) := (y - \langle x,a\rangle)^2$. In general we have

$$\nabla_{(x,y)} f((x,y),a^\star) = (-2(y - \langle x,a^\star\rangle)a^\star, 2(y - \langle x,a^\star\rangle))$$

$\nabla_a^2 V(0,a^\star) = 2D$ and

$$\begin{aligned}\left(\int_{\mathbb{R}^{k+1}} \|\nabla_{(x,y)} f((x,y),a^\star)\|^2 \, \mu(dx,dy)\right)^{1/2} &= 2|a^\star|_s \left(\int_{\mathbb{R}^{k+1}} (y - \langle x,a^\star\rangle)^2 \, \mu(dx,dy)\right)^{1/2} \\ &= 2|a^\star|_s \sqrt{V(0)}.\end{aligned}$$

Recalling the convention that $\nabla_{(x,y)}\nabla_a f \in \mathbb{R}^{k\times(d+1)}$ is given by

$$\begin{bmatrix} \nabla_{x_1}\nabla_{a_1} f & \dots & \nabla_{x_d}\nabla_{a_1} f & \nabla_y \nabla_{a_1} f \\ \nabla_{x_1}\nabla_{a_2} f & \dots & \nabla_{x_d}\nabla_{a_2} f & \nabla_y \nabla_{a_2} f \\ \vdots & \vdots & \vdots & \vdots \\ \nabla_{x_1}\nabla_{a_k} f & \dots & \nabla_{x_d}\nabla_{a_k} f & \nabla_y \nabla_{a_k} f \end{bmatrix}$$

we conclude

$$\nabla_{(x,y)}\nabla_a f((x,y),a^\star) = 2\left(-y\mathbf{I} + x(a^\star)^T + (\mathbf{I}a^\star)(\mathbf{I}x), -x\right),$$

where $\mathbf{I}$ is the $k\times k$ identity matrix. Recall furthermore that $\int_{\mathbb{R}^{k+1}} (y - \langle a^\star, x\rangle) x_i \mu(dx,dy) = 0$ for all $1 \le i \le k$ and in particular $V(0) = \int_{\mathbb{R}^{k+1}} (y^2 - \langle a^\star, x\rangle y)\mu(dx,dy)$. Set now

$$h((x,y)) := (\mathrm{sign}(x_1)\,|x_1|^{s-1}, \dots, \mathrm{sign}(x_k)\,|x_k|^{s-1}, 0) \cdot |x|_s^{1-s}.$$

Then $\langle (x,y), h((x,y))\rangle = |x|_s$ and $|h(x,y)|_r = 1$ for $(x,y) \in \mathcal{S} \setminus U$. As $h$ does not depend on the last coordinate, we also write simply $h(x)$ for $h((x,y))$. As $q = 2$ we have in particular

$$\begin{aligned}
&\int_{\mathbb{R}^{k+1}} \nabla_{(x,y)} \nabla_a f((x,y),a^\star) \frac{h(\nabla_{(x,y)} f((x,y),a^\star))}{\|\nabla_{(x,y)} f((x,y),a^\star)\|^{-1}}\, \mu(dx,dy) \\
&= 4 \int_{\mathbb{R}^{k+1}} \big[ - y\mathbf{I} + x(a^\star)^T + (\mathbf{I}a^\star)(\mathbf{I}x)\big]\, h(-(y - \langle x, a^\star\rangle)a^\star)\, |a^\star|_s |y - \langle x, a^\star\rangle|\, \mu(dx,dy) \\
&= -4|a^\star|_s \int_{\mathbb{R}^{k+1}} \big[ - y\mathbf{I} + x(a^\star)^T + (\mathbf{I}a^\star)(\mathbf{I}x)\big]\, (y - \langle x, a^\star\rangle) h(a^\star)\, \mu(dx,dy) \\
&= 4|a^\star|_s V(0)\, h(a^\star).
\end{aligned}$$

In conclusion

$$\begin{aligned}
a^\star_\delta &\approx a^\star - \Big( \int_{\mathbb{R}^{k+1}} \|\nabla_{(x,y)} f((x,y),a^\star)\|^2\, \mu(dx,dy)\Big)^{-1/2} (\nabla_a^2 V(0,a^\star))^{-1} \\
&\qquad \cdot \int_{\mathbb{R}^{k+1}} \frac{\nabla_{(x,y)}\nabla_a f((x,y),a^\star)\, h(\nabla_{(x,y)} f((x,y),a^\star))}{\|\nabla_{(x,y)} f((x,y),a^\star)\|^{-1}}\, \mu(dx,dy)\cdot\delta \\
&= a^\star - \frac{1}{4|a^\star|_s \sqrt{V(0)}}\, D^{-1}\, 4|a^\star|_s V(0)\, h(a^\star)\cdot\delta \\
&= a^\star - \sqrt{V(0)} D^{-1}\, h(a^\star)\cdot\delta.
\end{aligned}$$

Let us now specialise to the typical statistical context and let $\mu = \mu_N$ equal to the empirical measure of $N$ data samples, i.e., $\mu_N = \frac{1}{N}\sum_{i=1}^N \delta_{(x_i,y_1)}$ for some points $x_1,\dots,x_N \in \mathbb{R}^d$ and $y_1,\dots,y_N \in \mathbb{R}$. Let us write $x_i = (x_{i,1},\dots,x_{i,d})$ and $X = (x_{i,j})_{i=1,\dots,N}^{j=1,\dots,d}$. Then in particular

$$D = \int_{\mathbb{R}^{k+1}} xx^T\, \mu_N(dx,dy) = \frac{1}{N} X^T X$$

and we recover the notation common in statistics. In particular, $a^\star = (X^T X)^{-1} X^T y$. If we now assume that $X^T X = \mathbf{I}$ (and hence $D^{-1} = N\mathbf{I}$), then we can easily compute

$$\begin{aligned}
V(0) &= \frac{1}{N}(y - Xa^\star)^T (y - Xa^\star) = \frac{1}{N}(y - XX^T y)^T (y - XX^T y) \\
&= \frac{1}{N} y^T (\mathbf{I} - XX^T)^T (\mathbf{I} - XX^T) y = \frac{1}{N} y^T (\mathbf{I} - XX^T - XX^T + XX^T XX^T) y \\
&= \frac{1}{N} y^T (\mathbf{I} - XX^T) y
\end{aligned}$$

Note that, under the assumption that $\sum_{i=1}^N y_i = 0$, $R^2$ is defined as

$$R^2 = 1 - \frac{y^T(\mathbf{I} - XX^T)y}{y^T y} = \frac{y^T y - y^T(\mathbf{I} - XX^T)y}{y^T y} = \frac{y^T XX^T y}{y^T y}.$$

Thus in the case $s = 1$ we have

$$\begin{aligned}
a^\star_\delta &\approx a^\star - \sqrt{V(0)} D^{-1}\, \mathrm{sign}(a^\star)\cdot\delta = a^\star - \sqrt{N}\, \sqrt{y^T y - y^T XX^T y}\, \mathrm{sign}(a^\star)\cdot\delta \\
&= a^\star - \sqrt{N}\, \sqrt{y^T y} \sqrt{1 - \frac{y^T XX^T y}{y^T y}}\, \mathrm{sign}(a^\star)\cdot\delta \\
&= a^\star - \sqrt{N}\, |y|\, \sqrt{1 - R^2}\, \mathrm{sign}(a^\star)\cdot\delta.
\end{aligned}$$

Furthermore, in the case $s=2$ we have

$$
\begin{aligned}
a^\star_\delta &\approx a^\star - D^{-1}\frac{\sqrt{V(0)}}{|a^\star|_2}a^\star\delta = a^\star\left(1 - N\frac{\sqrt{y^T(\mathbf{I}-XX^T)y}}{\sqrt{N}|a^\star|_2}\,\delta\right)\\
&= a^\star\left(1 - \frac{\sqrt{N\,y^T(\mathbf{I}-XX^T)y}}{|a^\star|_2}\,\delta\right). \qquad (18)
\end{aligned}
$$

We also have

$$|a| = \sqrt{\langle a^\star, a^\star\rangle} = \sqrt{y^TXX^Ty},$$

so (18) simplifies to

$$
\begin{aligned}
a^\star_\delta &\approx a^\star\left(1 - \frac{\sqrt{N\,y^T(\mathbf{I}-XX^T)y}}{\sqrt{y^TXX^Ty}}\,\delta\right) = a^\star\left(1-\delta\sqrt{N\left(\frac{y^Ty}{y^TXX^Ty}-1\right)}\right)\\
&= a^\star\left(1-\delta\sqrt{N\left(\frac{1}{R^2}-1\right)}\right).
\end{aligned}
$$

*Remark* 26. While $|\cdot|_1$ is not strictly convex, the above example can still be adapted to cover this case under the additional assumption, that $a^\star$ has no entries which are equal to zero. Indeed, we note that $x\mapsto h(x,y)$ is continuous (even constant) at every point $x$ except if a component of $x$ is equal to zero. Thus the proof of Lemma 29 still applies if we assume that $g$ has $\mu$-a.s. no components which are equal to zero instead of merely assuming that $g\neq 0$ $\mu$-a.s..

*Example* 27. We provide further details and discussion to complement the out-of-sample error example in [1]. First, we recall the remainder term obtained therein:

$$
\begin{aligned}
\Delta_N &:= \Big(\int |\nabla_x f(x,a^{\star,N})|_s^q\,\mu_N(dx)\Big)^{\frac{1}{q}-1}\cdot\Big(\int \nabla_a^2 f(x,a^{\star,N})\,\mu_N(dx)\Big)^{-1}\\
&\quad\cdot\int\frac{\nabla_x\nabla_a f(x,a^{\star,N})\,h(\nabla_x f(x,a^{\star,N}))}{|\nabla_x f(x,a^{\star,N})|_s^{1-q}}\,\mu_N(dx) - (\nabla_a^2 V(0,a^\star))^{-1}\Theta, \quad \text{where}\\
\Theta &:= \Big(\int |\nabla_x f(x,a^\star)|_s^q\,\mu(dx)\Big)^{\frac{1}{q}-1}\cdot\int\frac{\nabla_x\nabla_a f(x,a^\star)\,h(\nabla_x f(x,a^\star))}{|\nabla_x f(x,a^\star)|_s^{1-q}}\,\mu(dx).
\end{aligned}
$$

Recall that $\mu_N\to\mu$ in $W_p$ holds a.s. We suppose that Assumptions 1 and 3 hold, and that for any $r>0$, there exists $c>0$ such that the following hold uniformly for all $|a|\le r$:

$$
\begin{aligned}
&\sum_{i=1}^k\Big|\nabla_a\nabla_{a_i} f(x,a)\Big| \le c(1+|x|^p),\\
&\left|\frac{\nabla_x\nabla_a f(x,a^\star)\,h(\nabla_x f(x,a^\star))}{|\nabla_x f(x,a^\star)|^{1-q}}\right| \le c(1+|x|^p).
\end{aligned} \qquad (19)
$$

Recall from (3.7) that we already know that $a^{\star,N}\to a^\star$ a.s. Under the above integrability assumption, Lemma 20 gives

$$\left|\int\nabla_{a_i}\nabla_{a_j} f(x,a^{\star,N})\,\mu_N(dx) - \int\nabla_{a_i}\nabla_{a_j} f(x,a^\star)\,\mu(dx)\right|\to 0,$$

with analogous convergence for the other two terms in $\Delta_N$. We conclude that $\Delta_N\to 0$ a.s. and that (3.10) and (3.11) hold.

We now show how the arguments above can be adapted to extend and complement [6, Prop. 17]. Therein, the authors study VRS($\delta$) which is the expectation over realisations of $\mu_N$ of

$$\int f(x,a^{\star,N})\,\mu(dx) - \int f(x,a^{\star,N}_\delta)\,\mu(dx).$$

If $\mathrm{VRS}(\delta) > 0$ then, on average, the robust problem offers an improved performance, i.e., finds a better approximation to the true optimizer $a^\star$ than the classical non-robust problem. If we work with the difference above, then we look at first order Taylor expansion and obtain

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

which holds under the first condition in (19). This can be compared with [6, Lemma 1] which was derived under a Lipschitz continuity assumption on $a \mapsto f(x,a)$. For the quadratic case of [6, Prop. 17] we have $f(x,a) = 1/2a^2 - g(x)a$, where we took $d = 1$ for notational simplicity. We then have $\nabla_x f(xa) = -g'(x)a$, $\nabla_a^2 f(x,a) = 1$ and $\nabla_x \nabla_a f(x,a) \nabla_x f(x,a) = (g'(x))^2 a$. Specialising (3.8) to this setting, with $s = 2$, gives

$$\begin{aligned} a_\delta^{\star,N} - a^{\star,N} &\approx -\left(\nabla_a^2 V(0,a^\star)\right)^{-1} \left(\int |\nabla_x f(x,a^\star)|^q \, \mu_N(dx)\right)^{1/q-1} \cdot \int \frac{\nabla_x \nabla_a f(x,a^\star)\, \nabla_x f(x,a^\star)}{|\nabla_x f(x,a^\star)|^{2-q}} \mu_N(dx) \\ &= -|a^\star|^{1-q} \left(\int |g'(x)|^q \, \mu_N(dx)\right)^{1/q-1} \int \frac{(g'(x))^2 a^\star}{|g'(x)a^\star|^{2-q}} \, \mu_N(dx) \\ &= -\mathrm{sign}(a^\star) \left(\int |g'(x)|^q \, \mu_N(dx)\right)^{1/q}. \end{aligned}$$

While our results work for $p > 1$, see Remark 8, we can formally let $q \uparrow \infty$. The last term then converges to $-\mathrm{sign}(a^\star)\|g'\|_{L^\infty(\mu)}$ which recovers [6, Prop. 17], taking into account that $\mathrm{sign}(a^\star) = \mathrm{sign}\left(\int g(x)\, \mu(dx)\right)$.

## 3.2. **Proofs and auxiliary results related to Theorem 4.**

**Lemma 28.** *Let $f\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be differentiable such that $(x,a) \mapsto \nabla_a f(x,a)$ is continuous, fix $a \in \mathcal{A}^o$, and assume that for some $\varepsilon > 0$ we have that $|\nabla_x f(x,\tilde a)| \le c(1 + |x|^{p-1-\varepsilon})$ and $|\nabla_a f(x,\tilde a)| \le c(1+|x|^{p-\varepsilon})$ for some $c > 0$, all $x \in \mathcal{S}$ and all $\tilde a \in \mathcal{A}$ close to $a$. Further fix $\delta \ge 0$ and recall that $B_\delta^\star(\mu,a)$ is the set of maximizing measures given the strategy $a$. Then the (one-sided) directional derivative of $V(\delta,\cdot)$ at $a$ in the in direction $b \in \mathbb{R}^k$ is given by*

$$\lim_{h\to 0} \frac{V(\delta, a+hb) - V(\delta,a)}{h} = \sup_{\nu \in B_\delta^\star(\mu,a)} \int_{\mathcal{S}} \langle \nabla_a f(x,a), b\rangle \, \nu(dx).$$

*Proof.* Fix $b \in \mathbb{R}^k$. We start by showing that

$$\liminf_{h\to 0} \frac{V(\delta, a+hb) - V(\delta,a)}{h} \ge \sup_{\nu \in B_\delta^\star(\mu,a)} \int_{\mathcal{S}} \langle \nabla_a f(x,a), b\rangle \, \nu(dx). \tag{20}$$

To that end, let $\nu \in B_\delta^\star(\mu,a)$ and $h > 0$ be arbitrary. By definition of $B_\delta^\star(\mu,a)$ one has $V(\delta,a) = \int_{\mathcal{S}} f(x,a)\, \nu(dx)$. Moreover $B_\delta^\star(\mu,a) \subseteq B_\delta(\mu)$ implies that $V(\delta, a+hb) \ge \int_{\mathcal{S}} f(x, a+hb)\, \nu(dx)$. Note that the assumption $|\nabla_x f(x,\tilde a)| \le c(1 + |x|^{p-1-\varepsilon})$ implies

$$\begin{aligned} |f(x,\tilde a) - f(0,\tilde a)| &= \left| \int_0^1 \langle \nabla_x f(tx,\tilde a), x\rangle dt \right| \\ &\le \int_0^1 c(1 + |tx|^{p-1-\varepsilon})|x| dt \le c(1 + |x|^{p-\varepsilon} \vee |x|). \end{aligned}$$

Therefore, by dominated convergence, one has

$$\begin{aligned} \liminf_{h\to 0} \frac{V(\delta, a+hb) - V(\delta,a)}{h} &\ge \liminf_{h\to 0} \int_{\mathcal{S}} \frac{f(x,a+hb) - f(x,a)}{h} \, \nu(dx) \\ &= \int_{\mathcal{S}} \lim_{h\to 0} \frac{f(x,a+hb) - f(x,a)}{h} \, \nu(dx) \\ &= \int_{\mathcal{S}} \langle \nabla_a f(x,a), b\rangle \, \nu(dx) \end{aligned}$$

and as $\nu \in B_\delta^\star(\mu,a)$ was arbitrary, this shows (20).

We proceed to show that

$$\limsup_{h\to 0} \frac{V(\delta, a+hb) - V(\delta,a)}{h} \leq \sup_{\nu\in B^\star_\delta(\mu,a)} \int_{\mathcal{S}} \langle \nabla_a f(x,a), b\rangle \,\nu(dx). \tag{21}$$

For every sufficiently small $h>0$ let $\nu^h \in B^\star_\delta(\mu, a+hb)$ such that $V(\delta, a+hb) = \int_{\mathcal{S}} f(x,a+hb)\,\nu^h(dx)$. The existence of such $\nu^h$ is guaranteed by Lemma 23, which also guarantees that (possibly after passing to a subsequence) there is $\tilde\nu \in B_\delta(\mu)$ such that $\nu^h \to \tilde\nu$ in $W^{|\cdot|}_{p-\varepsilon}$. We claim that $\tilde\nu \in B^\star_\delta(\mu,a)$. By Lemma 20 one has

$$\lim_{h\to 0} V(\delta, a+hb) = \int_{\mathcal{S}} f(x,a)\,\tilde\nu(dx) \leq V(\delta,a).$$

On the other hand, for any choice $\tilde\nu \in B^\star_\delta(\mu,a)$ one has

$$\lim_{h\to 0} V(\delta, a+hb) \geq \lim_{h\to 0} \int_{\mathcal{S}} f(x,a+hb)\,\tilde\nu(dx) = \int_{\mathcal{S}} f(x,a)\,\tilde\nu(dx) = V(\delta,a).$$

This implies $V(\delta,a) = \int_{\mathcal{S}} f(x,a)\,\tilde\nu(dx)$ and in particular $\tilde\nu \in B^\star_\delta(\mu,a)$. At this point expand

$$f(x,a+hb) = f(x,a) + \int_0^1 \langle \nabla_a f(x,a+thb), hb\rangle\,dt$$

so that

$$\begin{aligned}
&V(\delta,a+hb) - V(\delta,a)\\
&= \int_{\mathcal{S}} \Big( f(x,a) + \int_0^1 \langle \nabla_a f(x,a+thb), hb\rangle\,dt\Big)\,\nu^h(dx) - \int_{\mathcal{S}} f(x,a)\,\tilde\nu(dx)\\
&\leq \int_{\mathcal{S}} \int_0^1 \langle \nabla_a f(x,a+thb), hb\rangle\,dt\,\nu^h(dx)
\end{aligned}$$

where we used $\tilde\nu \in B^\star_\delta(\mu,a)$ for the last inequality. Recall that $\nu^h$ converges to $\tilde\nu$ in $W^{|\cdot|}_{p-\varepsilon}$ and by assumption $|\nabla_a f(x,\tilde a)| \leq c(1+|x|^{p-\varepsilon})$ for all $\tilde a \in \mathcal{A}$ close to $a$. In particular

$$\frac{1}{h}\langle \nabla_a f(x,a+thb), hb\rangle \leq |\nabla_a f(x,a+thb)||b| \leq c(1+|x|^{p-\varepsilon})$$

for $h$ sufficiently small. As furthermore $(x,a)\mapsto \nabla_a f(x,a)$ is continuous, we conclude by Lemma 20 that

$$\lim_{h\to 0} \frac{1}{h} \int_{\mathcal{S}} \langle \nabla_a f(x,a+thb), hb\rangle\,dt\,\nu^h(dx) = \int_{\mathcal{S}} \langle \nabla_a f(x,a), b\rangle\,\tilde\nu(dx).$$

Lastly, by by Fubini's theorem and dominated convergence (in $t$)

$$\frac{1}{h}\int_{\mathcal{S}} \int_0^1 \langle \nabla_a f(x,a+thb), hb\rangle\,dt\,\nu^h(dx) \to \int_{\mathcal{S}} \langle \nabla_a f(x,a), b\rangle\,\tilde\nu(dx)$$

as $h\to 0$, which ultimately shows (21). □

**Lemma 29.** *Let $q\in(1,\infty)$ and let $f,g\colon \mathcal{S}\to\mathbb{R}^d$ be measurable such that $\int_{\mathcal{S}} \|f(x)\|^q + \|g\|^q\,\mu(dx) < \infty$ and such that $g\neq 0$ $\mu$-a.s.. Then we have that*

$$\begin{aligned}
&\inf_{\lambda\in\mathbb{R}} \left( \left( \int_{\mathcal{S}} \|f(x)+\lambda g(x)\|^q\,\mu(dx)\right)^{1/q} - \lambda \left( \int_{\mathcal{S}} \|g(x)\|^q\,\mu(dx)\right)^{1/q}\right)\\
&= \int_{\mathcal{S}} \frac{\langle f(x), h(g(x))\rangle}{\|g(x)\|^{1-q}}\,\mu(dx)\cdot \Big( \int_{\mathcal{S}} \|g(x)\|^q\,\mu(dx)\Big)^{1/q-1},
\end{aligned} \tag{22}$$

*where $h\colon \mathbb{R}^d\setminus\{0\}\to\mathbb{R}^d$ was defined in Lemma 6.*

*Proof.* First recall that $h$ is continuous and satisfies $\|x\| = \langle x, h(x)\rangle$ for every $x\neq 0$. Now define

$$G(x) := \frac{h(g(x))}{\|g(x)\|^{1-q}} \Big( \int_{\mathcal{S}} \|g(z)\|^q\,\mu(dz)\Big)^{1/q-1} \quad \text{for } x\in\mathcal{S}.$$

Similarly, define $G^\lambda$ by replacing $g$ in the definition of $G$ by $g^\lambda := f + \lambda g$. As in the proof of Theorem 2 we compute

$$\int_{\mathcal{S}} \|G(x)\|_*^p \,\mu(dx) = 1 \quad \text{and} \quad \Big(\int_{\mathcal{S}} \|g(x)\|^q \,\mu(dx)\Big)^{1/q} = \int_{\mathcal{S}} \langle g(x), G(x)\rangle \,\mu(dx).$$

This remains true when $g$ and $G$ are replaced by $g^\lambda$ and $G^\lambda$, respectively. Moreover, Hölder's inequality implies that

$$\Big(\int_{\mathcal{S}} \|g^\lambda(x)\|^q \,\mu(dx)\Big)^{1/q} \geq \int_{\mathcal{S}} \langle g^\lambda(x), G(x)\rangle \,\mu(dx),$$
$$\Big(\int_{\mathcal{S}} \|g(x)\|^q \,\mu(dx)\Big)^{1/q} \geq \int_{\mathcal{S}} \langle g(x), G^\lambda(x)\rangle \,\mu(dx).$$

The first of these two inequalities immediately implies that the left hand side in (22) is larger than the right hand side.

To show the other inequality, note that $h$ is continuous and satisfies $h(\lambda x) = h(x)$ for $\lambda > 0$, hence $h(g(x)) = \lim_{\lambda\to\infty} h(g^\lambda(x))$ for all $x \in \mathcal{S}$ such that $g(x) \neq 0$. Consequently one quickly computes $G(x) = \lim_{\lambda\to\infty} G^\lambda(x)$ for all $x \in \mathcal{S}$ such that $g(x) \neq 0$. By dominated convergence we conclude that

$$\begin{aligned}
&\inf_{\lambda\in\mathbb{R}} \bigg(\Big(\int_{\mathcal{S}} \|f(x)+\lambda g(x)\|^q \,\mu(dx)\Big)^{1/q} - \lambda\Big(\int_{\mathcal{S}} \|g(x)\|^q \,\mu(dx)\Big)^{1/q}\bigg)\\
&\leq \lim_{\lambda\to\infty} \Big(\int_{\mathcal{S}} \langle f(x)+\lambda g(x), G^\lambda(x)\rangle \,\mu(dx) - \lambda \int_{\mathcal{S}} \langle g(x), G^\lambda(x)\rangle \,\mu(dx)\Big)\\
&= \int_{\mathcal{S}} \langle f(x), G(x)\rangle \,\mu(dx)
\end{aligned}$$

and the claim follows. □

Let us lastly give the proof of Theorem 4 for general seminorms.

*Proof of Theorem 4.* Recall the convention that $\nabla_x\nabla_a f(x,a) \in \mathbb{R}^{k\times d}$ and $\nabla_x f(x,a) \in \mathbb{R}^{d\times 1}$, $\nabla_a f(x,a) \in \mathbb{R}^{k\times 1}$ as well as $h(\cdot)/0 = 0$. Further recall that $a^\star \in \mathcal{A}^\star(0)$ and $a^\star_\delta \in \mathcal{A}^\star(\delta)$ converge to $a^\star$ as $\delta \to 0$. In order to show

$$\begin{aligned}
\lim_{\delta\to 0} \frac{a^\star_\delta - a^\star}{\delta} = &-\Big(\int_{\mathcal{S}} \|\nabla_x f(z,a^\star)\|^q \,\mu(dz)\Big)^{\frac{1}{q}-1} (\nabla_a^2 V(0,a^\star))^{-1}\\
&\cdot \int_{\mathcal{S}} \frac{\nabla_x\nabla_a f(x,a^\star)\, h(\nabla_x f(x,a^\star))}{\|\nabla_x f(x,a^\star)\|^{1-q}} \,\mu(dx),
\end{aligned}$$

we first show that for every $i \in \{1,\ldots,k\}$

$$\begin{aligned}
\lim_{\delta\to 0} \frac{-\nabla_{a_i} V(0,a^\star_\delta)}{\delta} = &\int_{\mathcal{S}} \nabla_x \nabla_{a_i} f(x,a^\star) \frac{h(\nabla_x(f(x,a^\star))}{\|\nabla_x f(x), a^\star)\|^{1-q}} \,\mu(dx)\\
&\cdot \Big(\int_{\mathcal{S}} \|\nabla_x f(x,a^\star)\|^q \,\mu(dx)\Big)^{1/q-1},
\end{aligned} \tag{23}$$

where we recall that $\nabla_{a_i} V(0,a^\star_\delta)$ is the $i$-th coordinate of the vector $\nabla_a V(0,a^\star_\delta)$. We start with the "$\leq$"-inequality in (23). For any $a \in \mathcal{A}^o$, the fundamental theorem of calculus implies that

$$\nabla_a f(y,a) - \nabla_a f(x,a) = \int_0^1 \nabla_x\nabla_a f(x+t(y-x),a)(y-x)\,dt.$$

Moreover, by Lemma 28 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 $i \in \{1,\ldots,k\}$

$$\sup_{\nu\in B^\star_\delta(\mu,a^\star_\delta)} \int_{\mathcal{S}} \nabla_{a_i} f(x,a^\star_\delta)\,\nu(dx) \geq 0, \tag{24}$$

where we recall $B^\star_\delta(\mu, a^\star_\delta)$ is the set of all $\nu \in B_\delta(\mu)$ for which $\int_{\mathcal{S}} f(x, a^\star_\delta)\,\nu(dx) = V(\delta, a^\star_\delta) = V(\delta)$. We now encode the optimality of $\nu$ in $B^\star_\delta(\mu, a^\star_\delta)$ via a Lagrange multiplier to obtain

$$\begin{aligned}&\sup_{\nu\in B^\star_\delta(\mu,a^\star_\delta)} \int_{\mathcal{S}} \nabla_{a_i} f(x, a^\star_\delta)\,\nu(dx)\\ &= \sup_{\nu\in B_\delta(\mu)} \inf_{\lambda\in\mathbb{R}} \int_{\mathcal{S}} \big[\nabla_{a_i} f(y, a^\star_\delta) + \lambda(f(y, a^\star_\delta) - V(\delta))\big]\nu(dy).\end{aligned} \tag{25}$$

In a similar manner, we trivially have

$$\int_{\mathcal{S}} \nabla_{a_i} f(x, a^\star_\delta)\,\mu(dx) = \int_{\mathcal{S}} \big[\nabla_{a_i} f(x, a^\star_\delta) + \lambda(f(x, a^\star_\delta) - V(0, a^\star_\delta))\big]\mu(dx) \tag{26}$$

for any $\lambda \in \mathbb{R}$, as $\int_{\mathcal{S}} f(x, a^\star_\delta)\,\mu(dx) = V(0, a^\star_\delta)$. Applying (24) and then (25), (26) we thus conclude for $i \in \{1, \dots, k\}$

$$\begin{aligned}-\nabla_{a_i} V(0, a^\star_\delta) &\le \sup_{\nu\in B^\star_\delta(\mu)} \int_{\mathcal{S}} \nabla_{a_i} f(y, a^\star_\delta)\,\nu(dy) - \nabla_{a_i} V(0, a^\star_\delta)\\ &= \sup_{\nu\in B_\delta(\mu)} \inf_{\lambda\in\mathbb{R}} \bigg( \int_{\mathcal{S}} \big[\nabla_{a_i} f(y, a^\star_\delta) + \lambda(f(y, a^\star_\delta) - V(\delta))\big]\,\nu(dy)\\ &\quad - \int_{\mathcal{S}} \big[\nabla_{a_i} f(x, a^\star_\delta) + \lambda(f(x, a^\star_\delta) - V(0, a^\star_\delta))\big]\mu(dx)\bigg)\\ &= \sup_{\nu\in B_\delta(\mu)} \inf_{\lambda\in\mathbb{R}} \bigg( \int_{\mathcal{S}} \Big[\nabla_{a_i} f(y, a^\star_\delta) + \lambda f(y, a^\star_\delta)\Big]\,\nu(dy)\\ &\quad - \int_{\mathcal{S}} \Big[\nabla_{a_i} f(x, a^\star_\delta) + \lambda f(x, a^\star_\delta)\Big]\,\mu(dx) - \lambda(V(\delta) - V(0, a^\star_\delta))\bigg).\end{aligned} \tag{27}$$

As in the proof of Lemma 28 we note that $B_\delta(\mu)$ is compact in $W^{|\cdot|}_{p-\varepsilon}$ and both terms inside the $\nu(dy)$ grow at most as $c(1 + |y|^{p-\varepsilon})$ by Assumption 3. Thus using [5, Cor. 2, p. 411] we can interchange the infimum and supremum in the last line above. Recall that

$$V(\delta) = \sup_{\nu\in B_\delta(\mu)} \int_{\mathcal{S}} f(y, a^\star_\delta)\,\nu(dy),$$

whence (27) is equal to

$$\begin{aligned}&\inf_{\lambda\in\mathbb{R}} \bigg( \sup_{\pi\in C_\delta(\mu)} \int_{\mathcal{S}\times\mathcal{S}} \Big[\nabla_{a_i} f(y, a^\star_\delta) - \nabla_{a_i} f(x, a^\star_\delta) + \lambda(f(y, a^\star_\delta) - f(x, a^\star_\delta))\Big]\,\pi(dx, dy)\\ &- \lambda \sup_{\pi\in C_\delta(\mu)} \int_{\mathcal{S}\times\mathcal{S}} f(y, a^\star_\delta) - f(x, a^\star_\delta)\,\pi(dx, dy)\bigg).\end{aligned}$$

For every fixed $\lambda \in \mathbb{R}$ we can follow the arguments in the proof of Theorem 2 to see that, when divided by $\delta$, the term inside the infimum converges to

$$\left(\int_{\mathcal{S}} \|\nabla_x \nabla_{a_i} f(x, a^\star) + \lambda \nabla_x f(x, a^\star)\|^q \,\mu(dx)\right)^{1/q} - \lambda \left(\int_{\mathcal{S}} \|\nabla_x f(x, a^\star)\|^q \,\mu(dx)\right)^{1/q} \tag{28}$$

as $\delta \to 0$. Note that following these arguments requires the following properties, which are a direct consequence of Assumptions 1 and Assumption 3:

- $(x, a) \mapsto f(x, a)$ is differentiable on $\mathcal{S}^o \times \mathcal{A}^o$,
- $x \mapsto \nabla_{a_i} f(x, a)$ is differentiable on $\mathcal{S}^o$ for every $a \in \mathcal{A}$,
- $(x, a) \mapsto \nabla_x f(x, a)$ is continuous,
- $(x, a) \mapsto \nabla_x \nabla_{a_i} f(x, a)$ is continuous,
- for every $r > 0$ there is $c > 0$ such that $|\lambda \nabla_x f(x, a)| \le c(1 + |x|^{p-1})$ for all $x \in \mathcal{S}$ and $a \in \mathcal{A}$ with $|a| \le r$.
- for every $r > 0$ there is $c > 0$ such that $|\nabla_x \nabla_{a_i} f(x, a)| \le c(1 + |x|^{p-1})$ for all $x \in \mathcal{S}$ and $a \in \mathcal{A}$ with $|a| \le r$.
- For all $\delta \ge 0$ sufficiently small we have $\mathcal{A}^\star_\delta \ne \emptyset$ and for every sequence $(\delta_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} \delta_n = 0$ and $(a^\star_n)_{n\in\mathbb{N}}$ such that $a^\star_n \in \mathcal{A}^\star_{\delta_n}$ for all $n \in \mathbb{N}$ there is a subsequence which converges to some $a^\star \in \mathcal{A}^\star_0$.

Suppose first that $\nabla_x \nabla_{a_i} f(x,a^\star)=0$ $\mu$-a.s.. Then the right hand side of (23) is equal to zero. Moreover, taking $\lambda=0$ in (28), we also have that $\nabla_{a_i} V(0,a_\delta^\star)\le 0$, which proves that indeed the left hand side in (23) is smaller than the right hand side.

Now suppose that $\nabla_x f(x,a^\star)\neq 0$ $\mu$-a.s.. Then, using the inequality "$\limsup_\delta \inf_\lambda \le \inf_\lambda \limsup_\delta$" and Lemma 29 to compute the last term (noting that $\nabla_x f(x,a^\star)\neq 0$ by assumption), we conclude that indeed the

$$\limsup_{\delta\to 0} \frac{-\nabla_{a_i} V(0,a_\delta^\star)}{\delta} \le \int_{\mathcal{S}} \nabla_x \nabla_{a_i} f(x,a^\star) \frac{h(\nabla_x (f(x,a^\star))}{\|\nabla_x f(x), a^\star)\|^{1-q}}\, \mu(dx) \cdot \Big(\int_{\mathcal{S}} \|\nabla_x f(x,a^\star)\|^q \,\mu(dx)\Big)^{1/q-1}.$$

To obtain the reverse "$\ge$"-inequality in (23) follows by the very same arguments. Indeed, Lemma 28 implies that

$$\inf_{\nu\in B_\delta^\star(\mu,a_\delta^\star)} \int_{\mathcal{S}} \nabla_{a_i} f(x,a_\delta^\star)\,\nu(dx)\le 0$$

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

$$\begin{aligned}
-\nabla_{a_i} V(0,a_\delta^\star) &\ge \inf_{\nu\in B_\delta^\star(\mu)} \int_{\mathcal{S}} \nabla_{a_i} f(y,a_\delta^\star)\,\nu(dy) - \int_{\mathcal{S}} \nabla_{a_i} f(y,a_\delta^\star)\,\mu(dx)\\
&= \inf_{\nu\in B_\delta(\mu)} \sup_{\lambda\in\mathbb{R}} \int_{\mathcal{S}} \big[\nabla_{a_i} f(y,a_\delta^\star)+\lambda(f(y,a_\delta^\star)-V(\delta))\big]\,\nu(dy) - \int_{\mathcal{S}} \nabla_{a_i} f(x,a_\delta^\star)\,\mu(dx)
\end{aligned}$$

From here on we argue as in the "$\le$"-inequality to conclude that (23) holds.

By assumption the matrix $\nabla_a^2 V(0,a^\star)$ is invertible. Therefore, in a small neighborhood of $a^\star$, the mapping $\nabla_a V(0,\cdot)$ is invertible. In particular

$$a_\delta^\star = (\nabla_a V(0,\cdot))^{-1}(\nabla_a V(0,a_\delta^\star)) \quad\text{and}\quad a^\star=(\nabla_a V(0,\cdot))^{-1}(0),$$

where the second equality holds by the first order condition for optimality of $a^\star$. Applying the chain rule and using (23) gives

$$\begin{aligned}
\lim_{\delta\to 0}\frac{a_\delta^\star - a^\star}{\delta} &= (\nabla_a^2 V(0,a^\star))^{-1}\cdot \lim_{\delta\to 0}\frac{\nabla_a V(0,a_\delta^\star)}{\delta}\\
&= -(\nabla_a^2 V(0,a^\star))^{-1}\Big(\int_{\mathcal{S}} \|\nabla_x f(z,a^\star)\|^q\,\mu(dz)\Big)^{1/q-1}\\
&\quad\cdot \int_{\mathcal{S}} \frac{\nabla_x\nabla_a f(x,a^\star) h(\nabla_x f(x,a^\star))}{\|\nabla_x f(x,a^\star)\|^{1-q}}\,\mu(dx).
\end{aligned}$$

This completes the proof. □

## References


[1] Bartl D, Drapeau S, Obłój J, Wiesel J. 2021 Sensitivity analysis of Wasserstein distributionally robust optimization problems. *TBC.*

[2] Brezis H. 2010 *Functional analysis, Sobolev spaces and partial differential equations.* Springer Science & Business Media.

[3] Bartl D, Drapeau S, Tangpi L. 2020 Computational aspects of robust optimized certainty equivalents and option pricing. *Math. Finance* **30**, 287–309.

[4] Villani C. 2008 *Optimal transport: old and new* vol. 338. Springer Science & Business Media.

[5] Terkelsen F. 1973 Some minimax theorems. *Math. Scand.* **31**, 405–413.

[6] Anderson EJ, Philpott AB. 2019 Improving sample average approximation using distributional robustness. *Optimization Online.*