A first-order regularized approach to the order-value optimization problem*

G. Q. Álvarez† E. G. Birgin‡

August 16, 2023§

Abstract

Minimization of the order-value function is part of a large family of problems involving functions whose value is calculated by sorting values from a set or subset of other functions. The order-value function has as particular cases the minimum and maximum functions of a set of functions and is well suited for applications involving robust estimation. In this paper, a first order method with quadratic regularization to solve the problem of minimizing the order-value function is proposed. An optimality condition for the problem and theoretical results of iteration complexity and evaluation complexity for the proposed method are presented. The applicability of the problem and method to parameter estimation problems with outliers is illustrated.

Keywords: Order-value optimization, regularized models, complexity, algorithms, applications.

Mathematics Subject Classification: 90C30, 65K05, 49M37, 90C60, 68Q25.

1 Introduction

Generalized order-value functions, systematized in [17], are functions whose value $f(x)$, for a given x in the domain, depends on order relations on a set of the form $\{f_i(x)\}_{i\in I}$. One such function is the order-value function of order p defined in [3, 4]. Given m functions $f_1, \ldots, f_m$, the value of the pth order-value function f at a point x in the domain corresponds to the value at the pth position when the values $f_1(x), f_2(x), \ldots, f_m(x)$ are ordered from smallest to largest. It is important to note that, even if all f_i are differentiable, f may be non-differentiable.

*This work has been partially supported by the Brazilian agencies FAPESP (grants 2013/07375-0, 2022/05803-3, and 2023/08706-1) and CNPq (grant 302073/2022-1).

†Department of Applied Mathematics, Institute of Mathematics and Statistics, University of São Paulo, Rua do Matão, 1010, Cidade Universitária, 05508-090, São Paulo, SP, Brazil. e-mail: gdavid@ime.usp.br

‡Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo, Rua do Matão, 1010, Cidade Universitária, 05508-090, São Paulo, SP, Brazil. e-mail: egbirgin@ime.usp.br

§Revision made on September 23, 2024.

For the particular choices $p = 1$ and $p = m$, we have that $f(x) = \min\{f_1(x), f_2(x), \ldots, f_m(x)\}$ and $f(x) = \max\{f_1(x), f_2(x), \ldots, f_m(x)\}$, respectively. If x is a vector of portfolio positions and $f_i(x)$ represents the expected loss for choosing x under scenario i, then the order-value function is the discrete value-at-risk (VaR) function, which is widely used in risk analysis; see [16]. If each function f_i represents the precision with which a certain model that depends on unknown parameters x fits the ith observation, minimizing the pth order-value function is equivalent to fitting the model's x parameters by discarding the poorest fitted $m - p$ observations. In general, the order-value function is well suited for applications involving robust estimation, i.e., estimation techniques that are not affected by slight deviations in the data or from the idealized premises.

In [3], it was introduced a steepest descent type method for the minimization of the pth order-value function restricted to a closed and convex set. Convergence to points that satisfy a weak optimality conditions was proven. In [4], stronger optimality conditions and a nonlinear programming reformulation with equilibrium constraints of the problem were given. In [5], it was introduced a quasi-Newton method that generalizes the method proposed in [3]. In [7], it was proposed a global optimization strategy that combines multistart and a tunneling approach. In [20], it was proved that the minimization of the order-value function is an NP-hard problem in the strong sense in the case that constraints are given by a polytope.

In 2006, [19] introduced the idea of computational complexity in continuous optimization. Since then, algorithms with complexity results have been developed for a wide variety of continuous optimization problems. See, for example, [9, 10]. In 2022, the first book [13] specifically dedicated to the subject was released. This paper contributes to this line of research by proposing a method that possesses complexity results for the problem of minimizing the order-value function with box constraints.

The rest of this paper is organized as follows. In Section 2, the problem of minimizing the order-value function and the proposed regularized method are defined. Section 3 is devoted to the definition of an adequate optimality condition and to prove the well-definiteness, convergence, and complexity results of the method. Illustrative numerical experiments are given in Section 4. Conclusions and final remarks are given in the last section.

Notation. The symbol $\|\cdot\|$ denotes the Euclidean norm. For $i = 1, \ldots, n$, $e^i \in \mathbb{R}^n$ denotes the ith column of the identity matrix in $\mathbb{R}^{n\times n}$.

2 Quadratically regularized first-order method

Let $f_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, m$ be given. For a given $p \in \{1, 2, \ldots, m\}$, the pth-order-value function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$f(x) \equiv f_{i_p(x)}(x), \tag{1}$$

where the indices $\{i_1(x), i_2(x), \ldots, i_m(x)\} = \{1, 2, \ldots, m\}$ are such that

$$f_{i_1(x)}(x) \leq f_{i_2(x)}(x) \leq \cdots \leq f_{i_m(x)}(x), \tag{2}$$

that is, f is such that $f(x)$ corresponds to the value $f_i(x)$ which, when the values $f_1(x), f_2(x), \ldots, f_m(x)$ are ordered from smallest to largest, is ranked in the pth position. In the present work,

we consider the order-value optimization (OVO) problem given by

$$\text{Minimize } f(x) \text{ subject to } x \in \Omega, \tag{3}$$

where $\Omega \equiv \{x \in \mathbb{R}^n \mid \ell \leq x \leq u\}$, $\ell, u \in \mathbb{R}^n$ and $\ell_i < u_i$ for $i = 1, \ldots, n$.

We introduce hereafter a first-order method to tackle problem (3) which, at each step, minimizes a quadratically regularized linear model of $f(x)$. The specification of the algorithm requires the following definitions. Given $\delta > 0$, for all $x \in \Omega$, we define

$$I(x, \delta) \equiv \{i \in \{1, 2, \ldots, m\} \mid f(x) - \delta \leq f_i(x) \leq f(x) + \delta\}. \tag{4}$$

For further reference, we define, for all $x \in \Omega$, $I(x) \equiv I(x, 0)$. In addition to $\delta > 0$, for given $\sigma > 0$ and $\bar{x} \in \Omega$, we also define

$$\Psi(x; \bar{x}, \delta, \sigma) \equiv \max_{i \in I(\bar{x}, \delta)} \left\{\nabla f_i(\bar{x})^T (x - \bar{x})\right\} + \frac{\sigma}{2}\|x - \bar{x}\|^2. \tag{5}$$

The proposed method follows below.

Algorithm 2.1: Let $\delta > 0$, $\sigma_{\min} > 0$, $\alpha \in (0, 1)$, $\gamma > 1$, and $x^0 \in \Omega$ be given. Initialize $k \leftarrow 0$.

Step 1. Initialize $j \leftarrow 0$ and $\sigma_{k,j} = \sigma_{\min}$.

Step 2. Compute $x_{\text{trial}}^{k,j}$ as a solution to

$$\text{Minimize } \Psi(x; x^k, \delta, \sigma_{k,j}) \text{ subject to } x \in \Omega. \tag{6}$$

Step 3. Consider condition

$$f(x) \leq f(x^k) - \alpha\|x - x^k\|^2. \tag{7}$$

If (7) with $x \equiv x_{\text{trial}}^{k,j}$ does not hold, then set $\sigma_{k,j+1} = \gamma\sigma_{k,j}$, update $j \leftarrow j + 1$, and go to Step 2.

Step 4. Define $x^{k+1} = x_{\text{trial}}^{k,j}$, $\sigma_k = \sigma_{k,j}$, $j_k = j$, update $k \leftarrow k + 1$ and go to Step 1.

Remark. Theory requires $x_{\text{trial}}^{k,j}$ to be a stationary point of (6) such that $\Psi(x_{\text{trial}}^{k,j}; x^k, \delta, \sigma_{k,j}) \leq 0$. This can be achieved by any iterative method that generates a sequence with decreasing value of the objective function, starting from x^k, since $\Psi(x^k; x^k, \delta, \sigma_{k,j}) = 0$. Moreover, $\Psi(x; x^k, \delta, \sigma_{k,j})$ is convex (a piecewise linear convex function plus a convex quadratic) and the constraints are given by bounded box constraints. Therefore, every stationary point of (6) is a global minimizer. In other words, since the subproblem is simple, what the theory requires is equivalent to asking that $x_{\text{trial}}^{k,j}$ be a global minimizer of (6), which is guaranteed to exist.

3 Convergence and complexity

In this section, we introduce an optimality condition $C(\delta, \epsilon)$ for problem (3) that depends on the parameter δ and an optimality tolerance ϵ. In the sequence, we show that Algorithm 2.1 is well defined and present complexity results for obtaining an iterate that satisfies the optimality condition $C(\delta, \epsilon)$ for prescribed values of $\delta > 0$ and $\epsilon > 0$.

The three theorems that follow (Theorems 3.1, 3.2, and 3.3) show that if x^* is a local minimizer of (3), then it is also a local minimizer of minimizing $\Psi(x; x^*, 0, 0)$, $\Psi(x; x^*, 0, \sigma)$, and $\Psi(x; x^*, \delta, \sigma)$ subject to $x \in \Omega$ for any $\delta > 0$ and $\sigma > 0$, respectively. These results will be used in the construction of the optimality condition $C(\delta, \epsilon)$ for problem (3).

Assumption A1. *Functions $f_1, f_2, \ldots, f_m$ are continuously differentiable for all $x \in \Omega$.*

Theorem 3.1. *Suppose that Assumption A1 holds. Let x^* be a local minimizer of* (3) *and consider the problem minimize $\Psi(x; x^*, 0, 0)$ subject to $x \in \Omega$, i.e.*

$$\textit{Minimize} \max_{i \in I(x^*)} \left\{ \nabla f_i(x^*)^T (x - x^*) \right\} \textit{ subject to } x \in \Omega. \tag{8}$$

Then, x^* is a solution to (8).

Proof. Assume that the thesis is not true. Then there exists $x \in \Omega$ such that

$$\max_{i \in I(x^*)} \left\{ \nabla f_i(x^*)^T (x - x^*) \right\} < \max_{i \in I(x^*)} \left\{ \nabla f_i(x^*)^T (x^* - x^*) \right\} = 0.$$

This means that $\nabla f_i(x^*)^T (x - x^*) < 0$ for all $i \in I(x^*)$. By Assumption A1, f_i is differentiable for all $i \in I(x^*)$, then we have that

$$\lim_{t \to 0} \frac{f_i(x^* + t(x - x^*)) - f_i(x^*)}{t} = \nabla f_i(x^*)^T (x - x^*) < 0,$$

for all $i \in I(x^*)$. Therefore, there exists $\bar{t}_i > 0$ such that $f_i(x^* + t(x - x^*)) < f_i(x^*)$ for all $t \in (0, \bar{t}_i]$. Taking $\bar{t} = \min_{i \in I(x^*)} \{\bar{t}_i\}$, we obtain that

$$f_i(x^* + t(x - x^*)) < f_i(x^*) = f(x^*) \text{ for all } i \in I(x^*) \text{ and } t \in (0, \bar{t}]. \tag{9}$$

Moreover, for all $j \in \{1, 2, \ldots, m\} \setminus I(x^*)$, if t is sufficiently small, by the continuity of the functions $f_1, \ldots, f_m$, one has that

$$f_i(x^* + t(x - x^*)) < f_j(x^* + t(x - x^*)) \text{ whenever } i \in I(x^*) \text{ and } f(x^*) < f_j(x^*) \tag{10}$$

and

$$f_i(x^* + t(x - x^*)) > f_j(x^* + t(x - x^*)) \text{ whenever } i \in I(x^*) \text{ and } f(x^*) > f_j(x^*). \tag{11}$$

Inequalities (10) say that if $j \notin I(x^*)$, $i \in I(x^*)$, and $f_j(x^*) > f_i(x^*) = f(x^*)$ then, for t small enough, $f_j(x^* + t(x - x^*)) > f_i(x^* + t(x - x^*))$, i.e. the (strict) inequality is preserved from x^* to $x^* + t(x - x^*)$. Inequalities (11) are analogous for the case $f_j(x^*) < f_i(x^*) = f(x^*)$. This

means that the set of positions occupied by the values $\{f_i(x^*+t(x-x^*))\}_{i\in I(x^*)}$, in the smallest to largest ranking of the values $\{f_i(x^*+t(x-x^*))\}_{i=1}^m$, is equal to the set of positions occupied by the values $\{f_i(x^*)\}_{i\in I(x^*)}$ in the smallest to largest ranking of the values $\{f_i(x^*)\}_{i=1}^m$. This set includes $i_p(x^*+t(x-x^*))$. This implies that for every t small enough there exists $i \in I(x^*)$ such that $i = i_p(x^*+t(x-x^*))$, i.e. $i \in I(x^*)$ such that

$$f_i(x^*+t(x-x^*)) = f_{i_p(x^*+t(x-x^*))}(x^*+t(x-x^*)) = f(x^*+t(x-x^*)).$$

Therefore, for all t small enough, by (9), $f(x^*+t(x-x^*)) < f(x^*)$. Since x and x^* belong to Ω, which is convex, $x^*+t(x-x^*) \in \Omega$ for all $t \in [0,1]$ and, in particular, for all t sufficiently small. Hence, x^* can not be a local minimizer of (3). □

Theorem 3.2. *Suppose that Assumption A1 holds. Let x^* be a local minimizer of (3), $\sigma \geq 0$, and consider the problem minimize $\Psi(x; x^*, 0, \sigma)$ subject to $x \in \Omega$, i.e.*

$$\textit{Minimize} \max_{i\in I(x^*)} \left\{\nabla f_i(x^*)^T(x-x^*)\right\} + \frac{\sigma}{2}\|x-x^*\|^2 \textit{ subject to } x \in \Omega. \tag{12}$$

Then, x^* is a solution to (12).

Proof. Assume that the thesis is not true. Then there exists $x \in \Omega$ such that

$$\max_{i\in I(x^*)} \left\{\nabla f_i(x^*)^T(x-x^*)\right\} + \frac{\sigma}{2}\|x-x^*\|^2 < \max_{i\in I(x^*)} \left\{\nabla f_i(x^*)^T(x^*-x^*)\right\} + \frac{\sigma}{2}\|x^*-x^*\|^2 = 0,$$

that is,

$$\max_{i\in I(x^*)} \left\{\nabla f_i(x^*)^T(x-x^*)\right\} < -\frac{\sigma}{2}\|x-x^*\|^2 \leq 0.$$

In other words, there exists $x \in \Omega$ such that $\Psi(x; x^*, 0, 0) < 0$. But this is impossible because, by the Theorem 3.1, x^* is a solution to (8) and $\Psi(x^*; x^*, 0, 0) = 0$. □

Theorem 3.3. *Suppose that Assumption A1 holds. Let x^* be a local minimizer of (3), $\sigma \geq 0$, $\delta \geq 0$, and consider the problem minimize $\Psi(x; x^*, \delta, \sigma)$ subject to $x \in \Omega$, i.e.*

$$\textit{Minimize} \max_{i\in I(x^*,\delta)} \left\{\nabla f_i(x^*)^T(x-x^*)\right\} + \frac{\sigma}{2}\|x-x^*\|^2 \textit{ subject to } x \in \Omega. \tag{13}$$

Then, x^* is a solution to (13).

Proof. Assume that the thesis is not true. Then there exists $x \in \Omega$ such that

$$\max_{i\in I(x^*,\delta)} \left\{\nabla f_i(x^*)^T(x-x^*)\right\} + \frac{\sigma}{2}\|x-x^*\|^2 < 0.$$

Therefore, for all $i \in I(x^*, \delta)$,

$$\nabla f_i(x^*)^T(x-x^*) + \frac{\sigma}{2}\|x-x^*\|^2 < 0,$$

which, since $I(x^*) \subseteq I(x^*, \delta)$, implies that

$$\nabla f_i(x^*)^T(x-x^*) + \frac{\sigma}{2}\|x-x^*\|^2 < 0,$$

for all $i \in I(x^*)$. But, by Theorem 3.2 this is impossible. □

The following theorem (Theorem 3.4), which requires some technical lemmas and assumptions, is the theorem that motivates the definition of the optimality condition $C(\delta, \epsilon)$ for problem (3).

Definition 3.1. *Let $c : \mathbb{R}^n \to \mathbb{R}^m$ be continuously differentiable and consider the feasible set $\mathcal{C} = \{z \in \mathbb{R}^n \mid c(z) \leq 0\}$. We say that $z \in \mathcal{C}$ verifies the Mangasarian-Fromovitz Constraint Qualification (MFCQ) if there exists $d \in \mathbb{R}^n$ such that*

$$\nabla c_i(z)^T d < 0 \text{ for all } i \in \{1, \ldots, m\} \text{ such that } c_i(z) = 0.$$

Definition 3.2. *Given $q \in \mathbb{N}$, we define the unit simplex $\sum_q \subset \mathbb{R}^q$ by*

$$\sum\nolimits_q = \left\{\lambda \in \mathbb{R}^q \mid \sum_{j=1}^q \lambda_j = 1 \text{ and } \lambda_j \geq 0 \text{ for } j = 1, \ldots, q\right\}.$$

Lemma 3.1. *Consider the problem*

$$\text{Minimize } \max_{i \in \mathcal{I}} \varphi_i(x) \text{ subject to } x \in \Omega, \tag{14}$$

where $\mathcal{I} \subset \mathbb{N}$ is a finite set of indices and $\varphi_i : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable for all $i \in \mathcal{I}$. Assume that $x^* \in \Omega$ is a local minimizer of (14). Then, there exist $\mu \in \sum_{|\mathcal{I}|}$ and $\nu^\ell, \nu^u \in \mathbb{R}^n_+$ such that $\sum_{i \in \mathcal{I}} \mu_i \nabla \varphi_i(x^*) + \sum_{i=1}^n (\nu_i^u - \nu_i^\ell) e^i = 0$ and $(\ell_i - x_i^*)\nu_i^\ell = (x_i^* - u_i)\nu_i^u = 0$ for $i = 1, \ldots, n$.

Proof. Let x^* be a local minimizer of (14). Then, $(x^*, \max_{i \in \mathcal{I}} \varphi_i(x^*)) \in \mathbb{R}^{n+1}$ is the solution to

$$\text{Minimize } y \text{ subject to } \varphi_i(x) \leq y \text{ for all } i \in \mathcal{I} \text{ and } x \in \Omega.$$

Let $d \in \mathbb{R}^{n+1}$ be given by

$$d_j = \begin{cases} -1, & \text{if } x_j^* = \ell_j, \\ 1, & \text{if } x_j^* = u_j, \\ 0, & \text{otherwise,} \end{cases}$$

for $j = 1, \ldots, n$ plus

$$d_{n+1} = \max_{i \in \mathcal{I}} \left\{ [\nabla \varphi_i(x^*)]^T (d_1, \ldots, d_n) \right\} + 1.$$

This d shows that $(x^*, \max_{i \in \mathcal{I}} \varphi_i(x^*))$ satisfies MFCQ with respect to $c : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^{|\mathcal{I}|}$ given by $c_i(x, y) = \varphi_i(x) - y$ for all $i \in \mathcal{I}$. Thus, the thesis follows using the KKT conditions for the problem above. □

The corollary below will be used later in the complexity results.

Corollary 3.1. *Suppose that Assumption A1 holds. Then, for every k and $j = 0, \ldots, j_k$, there exist $\mu \in \sum_{|I(x^k, \delta)|}$ and $\nu^\ell, \nu^u \in \mathbb{R}^n_+$ such that $(\ell_i - [x^{k,j}_{\text{trial}}]_i)\nu_i^\ell = ([x^{k,j}_{\text{trial}}]_i - u_i)\nu_i^u = 0$ for $i = 1, \ldots, n$ and*

$$x^{k,j}_{\text{trial}} - x^k = \frac{1}{\sigma_{k,j}} \left[\sum_{i=1}^n \left(\nu_i^\ell - \nu_i^u \right) e^i - \sum_{i \in I(x^k, \delta)} \mu_i \nabla f_i(x^k) \right]. \tag{15}$$

Proof. The thesis follows from Lemma 3.1 considering $\varphi_i(x) \equiv \nabla f_i(x^k)^T(x - x^k) + (\sigma_{k,j}/2)\|x - x^k\|^2$. Assumption A1 is used to guarantee the existence of $\nabla f_i(x^k)$ for $i = 1, \ldots, m$. □

Theorem 3.4. *Suppose that Assumption A1 holds. Assume that x^* is a local minimizer of (3). Given $\delta \geq 0$, there exist $\mu \in \sum_{|I(x^*,\delta)|}$ and $\nu^\ell, \nu^u \in \mathbb{R}^n_+$ such that $(\ell_i - x_i^*)\nu_i^\ell = (x_i^* - u_i)\nu_i^u = 0$ for $i = 1, \ldots, n$ and*

$$\sum_{i \in I(x^*,\delta)} \mu_i \nabla f_i(x^*) + \sum_{i=1}^{n} \left(\nu_i^u - \nu_i^\ell\right) e^i = 0. \tag{16}$$

Proof. Let x^* a local minimizer of (3) and $\delta \geq 0$. By Theorem 3.3, x^* is a solution to

$$\text{Minimize } \Psi(x; x^*, \delta, \sigma) \text{ subject to } x \in \Omega,$$

for any $\sigma > 0$. Then, by Lemma 3.1, there exist $\mu \in \sum_{|I(x^*,\delta)|}$ and $\nu^\ell, \nu^u \in \mathbb{R}^n_+$ such that $(\ell_i - x_i^*)\nu_i^\ell = (x_i^* - u_i)\nu_i^u = 0$ for $i = 1, \ldots, n$ and

$$\sum_{i \in I(x^*,\delta)} \mu_i \nabla \left[\nabla f_i(x^*)^T(x - x^*) + \frac{\sigma}{2}\|x - x^*\|^2\right]\Big|_{x=x^*} + \sum_{i=1}^{n} \left(\nu_i^u - \nu_i^\ell\right) e^i = 0,$$

from which (16) follows. □

Theorem 3.4 leads to the definition of the following approximate necessary optimality condition.

Definition 3.3. *We say that x satisfies the approximate optimality condition $C(\delta, \epsilon)$ if there exist $\mu \in \sum_{|I(x,\delta)|}$ and $\nu^\ell, \nu^u \in \mathbb{R}^n_+$ such that $(\ell_i - x_i)\nu_i^\ell = (x_i - u_i)\nu_i^u = 0$ for $i = 1, \ldots, n$ and*

$$\left\| \sum_{i \in I(x,\delta)} \mu_i \nabla f_i(x) + \sum_{i=1}^{n} \left(\nu_i^u - \nu_i^\ell\right) e^i \right\| \leq \epsilon. \tag{17}$$

From here to the end of the section, we are devoted to show an upper bound for the cost of Algorithm 2.1, in terms of iterations and function evaluations, to, given $\epsilon > 0$, find an iterate x^k that satisfies the approximate optimality condition $C(\delta, \epsilon)$. We will also show that Algorithm 2.1 is well defined in the sense that the inner loop defined by Steps 2 and 3 terminates in a finite number of steps that does not depend on either k or ϵ. A few assumptions and technical lemmas precede the main results.

Assumption A2. *For all k and $j = 0, \ldots, j_k$, the associated Lagrange multipliers ν^ℓ and ν^u of Corollary 3.1 are such that $\max_{\{i=1,\ldots,n\}}\{|\nu_i^\ell - \nu_i^u|\}$ is bounded by a constant c_ν which depends neither on k nor on j.*

Note that MFCQ guarantees that, for every k and $j \in \{0, \ldots, j_k\}$, the associated Lagrange multipliers ν^ℓ and ν^u of Corollary 3.1 are bounded by a constant (see [15]), which in turn implies that $\max_{\{i=1,\ldots,n\}}\{|\nu_i^\ell - \nu_i^u|\}$ is bounded by a constant. Assumption A2 says that there exists a constant for all k and $j \in \{0, \ldots, j_k\}$ which depends neither on k nor on j.

Assumption A3. *$\|\nabla f_1(x)\|, \|\nabla f_2(x)\|, \ldots, \|\nabla f_m(x)\|$ are bounded from above by a constant c_∇ for all $x \in \Omega$.*

Lemma 3.2. *Suppose that Assumptions A1, A2, and A3 hold. Then, for all k and $j = 0, \ldots, j_k$, there exist $c_x > 0$, which depends neither on k nor on j, such that*

$$\|x^{k,j}_{\text{trial}} - x^k\| \leq c_x/\sigma_{k,j}. \tag{18}$$

Proof. By Corollary 3.1 and Assumption A3 we have that

$$\|x^{k,j}_{\text{trial}} - x^k\| \leq \frac{1}{\sigma_{k,j}} \left[\sum_{i=1}^{n} \left| \nu_i^\ell - \nu_i^u \right| + c_\nabla \right].$$

By Assumption A2, (18) holds with $c_x = n\, c_\nu + c_\nabla$. □

Assumption A4. *All the gradients ∇f_i satisfy a Lipschitz condition, that is, there exists $L > 0$ such that, for $i = 1, \ldots, m$ and all $x, y \in \Omega$,*

$$\|\nabla f_i(y) - \nabla f_i(x)\| \leq L\|y - x\|. \tag{19}$$

As a consequence of Assumption A4, for $i = 1, \ldots, m$ and all $x, y \in \Omega$,

$$\left| f_i(y) - \left[f_i(x) + \nabla f_i(x)^T (y - x) \right] \right| \leq \frac{L}{2}\|y - x\|^2. \tag{20}$$

In particular, for $i = 1, \ldots, m$ and all $x, y \in \Omega$,

$$f_i(y) \leq f_i(x) + \nabla f_i(x)^T (y - x) + \frac{L}{2}\|y - x\|^2. \tag{21}$$

See, for example, [10].

Lemma 3.3. *([3, Lemma 2.1]) Let $a_1, \ldots, a_r \in \mathbb{R}$, $b_1, \ldots, b_r \in \mathbb{R}$, $\beta > 0$, and $\{i_1, \ldots, i_r\} = \{1, \ldots, r\}$ be such that $a_1 \leq \cdots \leq a_r$, $b_j \leq a_j - \beta$ for $j = 1, \ldots, r$, and $b_{i_1} \leq \cdots \leq b_{i_r}$. Then, $b_{i_q} \leq a_q - \beta$ for $q = 1, \ldots, r$.*

Proof. By hypothesis, we have that, for any $q \in \{1, \ldots, r\}$,

$$\begin{array}{rcl} b_{i_q} & \leq & a_{i_q} - \beta \\ b_{i_q} & \leq & b_{i_{q+1}} \leq a_{i_{q+1}} - \beta \\ & \vdots & \\ b_{i_q} & \leq & b_{i_{q+1}} \leq \cdots \leq b_{i_r} \leq a_{i_r} - \beta. \end{array}$$

Therefore, $b_{i_q} \leq \min\{a_{i_q}, a_{i_{q+1}}, \ldots, a_{i_r}\} - \beta$. Since the set $\{a_{i_q}, a_{i_{q+1}}, \ldots, a_{i_r}\}$ has $r - q + 1$ elements, then there exist $\tilde{q} \in \{1, \ldots, q\}$ such that $a_{\tilde{q}} \in \{a_{i_q}, a_{i_{q+1}}, \ldots, a_{i_r}\}$. Thus,

$$b_{i_q} \leq a_{\tilde{q}} - \beta \leq a_q - \beta,$$

as we wanted to prove. □

Lemma 3.4. *Suppose that Assumptions A1 and A4 hold. For every k and $j = 0, \ldots, j_k$, if $\sigma_{k,j} \geq L + 2\alpha$, then*

$$f_i(x_{\text{trial}}^{k,j}) \leq f_i(x^k) - \alpha \|x_{\text{trial}}^{k,j} - x^k\|^2, \tag{22}$$

for all $i \in I(x^k, \delta)$.

Proof. By (21), which is implied by Assumption A4, if $i \in I(x^k, \delta)$, then

$$\begin{aligned} f_i(x_{\text{trial}}^{k,j}) &\leq f_i(x^k) + \nabla f_i(x^k)^T (x_{\text{trial}}^{k,j} - x^k) + \frac{L}{2} \|x_{\text{trial}}^{k,j} - x^k\|^2 \\ &= f_i(x^k) + \nabla f_i(x^k)^T (x_{\text{trial}}^{k,j} - x^k) + \frac{\sigma_{k,j}}{2} \|x_{\text{trial}}^{k,j} - x^k\|^2 - \frac{\sigma_{k,j}}{2} \|x_{\text{trial}}^{k,j} - x^k\|^2 + \frac{L}{2} \|x_{\text{trial}}^{k,j} - x^k\|^2. \end{aligned}$$

But the objective function of (6), defined in (5), vanishes if $x = x^k$. Therefore,

$$\nabla f_i(x^k)^T (x_{\text{trial}}^{k,j} - x^k) + \frac{\sigma_{k,j}}{2} \|x_{\text{trial}}^{k,j} - x^k\|^2 \leq 0 \text{ for all } i \in I(x^k, \delta).$$

Thus,

$$f_i(x_{\text{trial}}^{k,j}) \leq f_i(x^k) - \frac{\sigma_{k,j}}{2} \|x_{\text{trial}}^{k,j} - x^k\|^2 + \frac{L}{2} \|x_{\text{trial}}^{k,j} - x^k\|^2 \text{ for all } i \in I(x^k, \delta).$$

So, if $\sigma_{k,j} \geq L + 2\alpha$, then we have that

$$f_i(x_{\text{trial}}^{k,j}) \leq f_i(x^k) - \alpha \|x_{\text{trial}}^{k,j} - x^k\|^2 \text{ for all } i \in I(x^k, \delta).$$

□

The next theorem, together with the fact that, for all k, the initial value of the regularization parameter is equal to $\sigma_{\min} > 0$ and, whenever a new value is calculated, its value is multiplied by $\gamma > 1$, is what shows that the loop defined by Steps 2 and 3 is executed a finite number of times per iteration.

Theorem 3.5. *Suppose that Assumptions A1, A3 and A4 hold. Then, for all k and $j = 0, \ldots, j_k$, if*

$$\sigma_{k,j} \geq \max\left\{L + 2\alpha, 9c_x^2/(2\delta)\right\} \tag{23}$$

then $x_{\text{trial}}^{k,j}$ satisfies (7).

Proof. Assume that $\sigma_{k,j} \geq L + 2\,\alpha$. By the Cauchy-Schwarz inequality, Assumptions A3 and A4, and (18) in Lemma 3.2,

$$\left| f_i(x_{\text{trial}}^{k,j}) - f_i(x^k) \right| \leq \frac{c_x^2}{\sigma_{k,j}} + \frac{L\, c_x^2}{2\, \sigma_{k,j}^2} \text{ for } i = 1, \ldots, m.$$

Note that $L \leq \sigma_{k,j}$. Then,

$$\left| f_i(x_{\text{trial}}^{k,j}) - f_i(x^k) \right| \leq \frac{3\, c_x^2}{2\, \sigma_{k,j}} \text{ for } i = 1, \ldots, m.$$

Therefore, if $\sigma_{k,j} \geq \max\{L + 2\,\alpha, 9\,c_x^2/(2\,\delta)\}$, then we have that

$$\left| f_i(x_{\text{trial}}^{k,j}) - f_i(x^k) \right| \leq \frac{\delta}{3} \text{ for } i = 1, \ldots, m. \tag{24}$$

Thus, for $j = 1, \ldots, p$,

$$f_{i_j(x^k)}(x_{\text{trial}}^{k,j}) \leq f_{i_j(x^k)}(x^k) + \frac{\delta}{3} \leq f(x^k) + \frac{\delta}{3}$$

and, for $j = p, \ldots, m$,

$$f_{i_j(x^k)}(x_{\text{trial}}^{k,j}) \geq f_{i_j(x^k)}(x^k) - \frac{\delta}{3} \geq f(x^k) - \frac{\delta}{3}.$$

This means that p elements of the set $\{f_1(x_{\text{trial}}^{k,j}), f_2(x_{\text{trial}}^{k,j}), \ldots, f_m(x_{\text{trial}}^{k,j})\}$ are less than or equal to $f(x^k) + \delta/3$ and that $m - p + 1$ elements of that set are greater than or equal to $f(x^k) - \delta/3$. Then, at least one element satisfies both inequalities and, when ranked from smallest to largest, one of those values satisfying the two inequalities must be the value at the p-th position. As a consequence, $f_{i_p(x_{\text{trial}}^{k,j})}(x_{\text{trial}}^{k,j}) = f(x_{\text{trial}}^{k,j})$ satisfies both inequalities, i.e.

$$f(x^k) - \frac{\delta}{3} \leq f(x_{\text{trial}}^{k,j}) \leq f(x^k) + \frac{\delta}{3}. \tag{25}$$

By (25) and the definition of $I(\cdot, \cdot)$ in (4), $i_p(x_{\text{trial}}^{k,j}) \in I(x^k, \delta)$. Let us write

$$I(x^k, \delta) = \{i_1, \ldots, i_r\} = \{i'_1, \ldots, i'_r\},$$

where

$$f_{i_1}(x^k) \leq \cdots \leq f_{i_r}(x^k) \text{ and } f_{i'_1}(x_{\text{trial}}^{k,j}) \leq \cdots \leq f_{i'_r}(x_{\text{trial}}^{k,j}).$$

Let j be such that $f_j(x^k) < f(x^k) - \delta$. Then, by (24), $f_j(x_{\text{trial}}^{k,j}) \leq f_j(x^k) + \delta/3 < f(x^k) - 2\delta/3 < f(x^k)$. This means that the indices $j \notin I(x^k, \delta)$ such that $f_j(x^k) < f(x^k)$ are the same as the indices $j \notin I(x^k, \delta)$ such that $f_j(x_{\text{trial}}^{k,j}) < f(x^k)$. Analogously, the indices $j \notin I(x^k, \delta)$ such that $f_j(x^k) > f(x^k)$ are the same as the indices $j \notin I(x^k, \delta)$ such that $f_j(x_{\text{trial}}^{k,j}) > f(x^k)$. Therefore, if $q \in \{1, \ldots, r\}$ is such that $i_p(x^k) = i_q$, then $i_p(x_{\text{trial}}^{k,j}) = i'_q$.

By Lemma 3.4,

$$f_{i_j}(x_{\text{trial}}^{k,j}) \leq f_{i_j}(x^k) - \alpha \|x_{\text{trial}}^{k,j} - x^k\|^2$$

for $j = 1, \ldots, r$. Therefore, by Lemma 3.3 taking $\beta = \alpha \|x_{\text{trial}}^{k,j} - x^k\|^2$, $a_j = f_{i_j}(x^k)$ and $b_j = f_{i_j}(x_{\text{trial}}^{k,j})$ for $j = 1, \ldots, r$, we have that

$$f_{i'_j}(x_{\text{trial}}^{k,j}) \leq f_{i_j}(x^k) - \alpha \|x_{\text{trial}}^{k,j} - x^k\|^2$$

for $j = 1, \ldots, r$. In particular, it holds for the index $q \in \{1, \ldots, r\}$ of the previous paragraph such that $i_p(x^k) = i_q$ and $i_p(x_{\text{trial}}^{k,j}) = i'_q$. Therefore,

$$f(x_{\text{trial}}^{k,j}) \leq f(x^k) - \alpha \|x_{\text{trial}}^{k,j} - x^k\|^2$$

as we wanted to prove. □

The theorem below shows that Algorithm 2.1 requires $O(\delta^{-2}\epsilon^{-2})$ iterations and $O(|\log(\delta)|)$ functional evaluations per iteration to find a point that satisfies the $C(\delta,\epsilon)$ optimality condition of problem (3).

Theorem 3.6. *Suppose that Assumptions A1, A3 and A4 hold and there exists $f_{\text{low}} \in \mathbb{R}$ such that $f(x) \geq f_{\text{low}}$ for all $x \in \Omega$. Then, σ_k is such that*

$$\sigma_k \leq \gamma \max\{L + 2\alpha, 9c_x^2/(2\delta)\}, \tag{26}$$

where c_x is a constant that depends on c_ν and c_∇, and at most

$$\left\lfloor 1 + \log_\gamma \left(\frac{\sigma_k}{\sigma_{\min}} \right) \right\rfloor \tag{27}$$

functional evaluations are done to get (7). Moreover, the number of iterations k at which $C(\delta,\epsilon)$ is not satisfied by x^k is bounded above by

$$\left\lfloor \left(\frac{\gamma^2 \max\{L + 2\alpha, 9c_x^2/(2\delta)\}^2}{\alpha} \right) \left(\frac{f(x^0) - f_{\text{low}}}{\epsilon^2} \right) \right\rfloor. \tag{28}$$

Proof. Applying Theorem 3.5, (26) and (27) follow from (23) and the fact that, at Step 3, Algorithm 2.1 updates the regularization parameter by multiplying its value by γ if (7) does not hold.

For the second part, let $K \subset \mathbb{N}$ be the set of indices k such that $C(\delta,\epsilon)$ is not satisfied by x^{k+1}. By the mechanism of Algorithm 2.1, $x^{k+1} = x_{\text{trial}}^{k,j}$, where $x_{\text{trial}}^{k,j}$ is a solution to (6) and satisfies (7). Then, on the one hand, by Corollary 3.1, for each $k \in K$, there exist $\mu \in \sum_{|I(x^k,\delta)|}$ and $\nu^\ell, \nu^u \in \mathbb{R}^n_+$ such that $(\ell_i - x_i^{k+1})\nu_i^\ell = (x_i^{k+1} - u_i)\nu_i^u = 0$ for $i = 1, \ldots, n$ and

$$x^{k+1} - x^k = \frac{1}{\sigma_k} \left[\sum_{i=1}^n \left(\nu_i^\ell - \nu_i^u \right) e^i - \sum_{i \in I(x^k,\delta)} \mu_i \nabla f_i(x^k) \right],$$

i.e.

$$\|x^{k+1} - x^k\| = \frac{1}{\sigma_k} \left\| \sum_{i \in I(x^k,\delta)} \mu_i \nabla f_i(x^k) + \sum_{i=1}^n \left(\nu_i^u - \nu_i^\ell \right) e^i \right\|,$$

and, by (26) and the fact that $C(\delta,\epsilon)$ does not hold at x^{k+1}, it holds

$$\|x^{k+1} - x^k\| \geq \frac{\epsilon}{\gamma \max\{L + 2\alpha, 9\, c_x^2/(2\delta)\}}.$$

On the other hand, since for each $k \in K$, x^{k+1} satisfies (7), we have that

$$f(x^{k+1}) \leq f(x^k) - \alpha \left(\frac{\epsilon}{\gamma \max\{L + 2\alpha, 9\, c_x^2/(2\delta)\}} \right)^2.$$

Summing for all $k \in K$,

$$\sum_{k \in K} \left(f(x^k) - f(x^{k+1}) \right) \geq |K| \alpha \left(\frac{\epsilon}{\gamma \max\{L + 2\alpha, 9\, c_x^2/(2\delta)\}} \right)^2.$$

Since $f(x) \geq f_{\text{low}}$ for all $x \in \mathbb{R}^n$,

$$f(x^0) - f_{\text{low}} \geq |K|\alpha \left(\frac{\epsilon}{\gamma \max\{L + 2\alpha, 9\, c_x^2/(2\delta)\}} \right)^2,$$

from which (28) follows. □

Algorithm 2.1 defines at each iteration k a regularization parameter $\sigma_k \geq \sigma_{\min} > 0$, i.e., bounded away from zero. Specifically, the first trial $\sigma_{k,0}$ is an arbitrary value not smaller than $\sigma_{\min}$ that is then successively multiplied by γ. In practice, it may be adequate the first trial $\sigma_{k,0}$ at iteration $k > 1$ to be a fraction of σ_{k-1}. In this case, each $\sigma_k \geq \sigma_{\min}^k$, but it may be the case that $\sigma_{\min}^k \to 0$. For such a modified version of Algorithm 2.1, with a slightly different analysis than the one performed in Theorem 3.6, similar complexity bounds can also be obtained; see [8, §4].

4 Numerical illustration

In this section, we illustrate how the OVO problem, and in particular the method proposed in this paper to solve it, can be used for model fitting in the case where observations contain outliers. Let $y(t, x)$ be a model with unknown parameters $x \in \Omega \subseteq \mathbb{R}^n$ and let (t_i, y_i) for $i = 1, \ldots, m$ be a data set containing an unknown number $o \leq m$ of outliers. Let us define

$$f_i(x) = \frac{1}{2} \left(y(t_i, x) - y_i \right)^2,$$

for $i = 1, \ldots, m$. For o given, the OVO problem consisting of solving (3) with $f(x)$ defined as in (1) with $p = m - o$ corresponds to finding parameters $x \in \Omega$ that minimize the largest squared error of p observations by discarding o observations considered outliers. Given that the number of outliers is unknown, the methodology consists of solving a sequence of OVO problems with increasing values of o, starting from a known lower bound. It will be seen that a sudden drop in the optimal value of f as a function of o will clearly identify the number of outliers. In Section 4.1, we consider the epidemiological model introduced in [14]. In Section 4.2, we consider the model analyzed in [6, §7.2.1]. In Section 4.3, we consider the model analyzed in [3, §4]. The experiments in Sections 4.2 and 4.3 allow to compare the introduced method with other existing alternatives. In addition, the experiments in Section 4.3 show that the proposed method is scalable for increasing amounts of data.

As suggested in the proof of Lemma 3.1, subproblem (6) of Step 2 is reformulated as

$$\text{Minimize } y \text{ subject to } \nabla f_i(x^k)^T (x - x^k) + \frac{\sigma_{k,j}}{2} \|x - x^k\|^2 \leq y \text{ for all } i \in I(x^k, \delta) \text{ and } x \in \Omega. \quad (29)$$

We opted for this reformulation because its resolution provides, besides a solution $x_{\text{trial}}^{k,j}$, the associated Lagrange multipliers μ, ν^ℓ, and ν^u required to check the optimality condition $C(\delta, \epsilon)$. As stopping criterion, we checked the satisfaction of the optimality condition $C(\delta, \epsilon)$ in-between Steps 3 and 4. For the stopping criterion, we considered $\epsilon = 10^{-4}$. The ϵ tolerance value is standard when using first order methods. In Algorithm 2.1, we considered $\sigma_{\min} = 0.1$, $\alpha = 10^{-8}$,

and $\gamma = 5$. The value of these three parameters is quite standard in the literature of methods using regularized models and the method is not very sensitive to variations in these parameters. The choice of δ is more difficult. It is dimensional, depends on the problem, and was chosen by trial and error for each of the three applications separately. Basically, if δ is "small", few functions f_i will be considered in the computation of the next iterate, and this may cause the next iterate to fail the sufficient descent criterion (in which other f_i not considered in its computation will have an effect). On the other hand, if δ is "large", many f_i are considered in the computation of the next iterate, and since it is more difficult to find a descent direction for many f_i simultaneously, the next iterate stays very close to the current iterate. Either way, it is possible that the use of an "inappropriate" δ will produce short steps far from the solution and this will increase the total number of iterations. Therefore, the choice of an "appropriate" δ is based on observing the number of iterations the algorithm performs for different δ choices in $\{10^{-1}, 10^{-2}, 10^{-3}\}$. In Sections 4.1 and 4.2 we ended up considering $\delta = 10^{-3}$, while in Section 4.3 we considered $\delta = 0.1$.

Algorithm 2.1 was implemented in Fortran. Problem (29) is a smooth nonlinear programming problem and we chose to solve it with Algencan. Algencan [2, 11, 12] is a safeguarded augmented Lagrangian method introduced in [1, 2]. Its convergence theory, properties and usage are described in detail in [11]. Complexity results and an extensive numerical comparison with another state-of-the-art method for nonlinear programming can be found in [12]. In this work we use Algencan with all its default parameters.

Codes were implemented in Fortran 90. Tests were conducted on a computer with a 5.2 GHz Intel Core i9-12900 processor and 128GB 3200 MHz DDR4 RAM memory, running Linux (Ubuntu 22.04.4 LTS). Code was compiled by the GNU Fortran compiler (version 11.4.0) with the -O3 optimization directive enabled.

4.1 Epidemiological model

The epidemiological model considered in the present section was developed in [14] with the purpose of modeling a serological data set of 8,870 people before the introduction of measles, mumps and rubella vaccine in United Kingdom. The model aims to describe the rate at which susceptible individuals acquire infection by the diseases mentioned above at different ages. The data in Table 1, taken from [14], show the estimated proportion of seropositive in the unvaccinated segment of the sample divided into 29 age groups.

The model we wish to fit to the data in Table 1 is given by

$$y(t, x) = 1 - \exp\left\{\frac{x_1}{x_2} t e^{-x_2 t} + \frac{x_1}{x_2}\left(\frac{x_1}{x_2} - x_3\right)\left(e^{-x_2 t} - 1\right) - x_3 t\right\}, \tag{30}$$

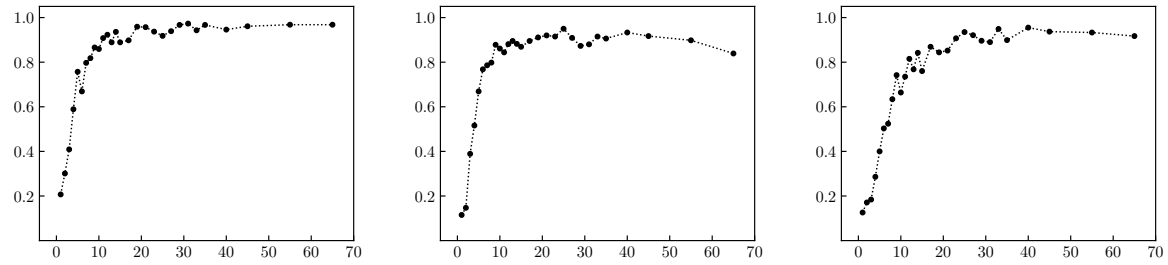
where x_1, x_2, x_3 are non-negative unknown parameters. Therefore, we define $\Omega = \{(x_1, x_2, x_3) \in \mathbb{R}^3 \mid (x_1, x_2, x_3) \geq 0\}$. The amount of data is $m = 29$, and we wish to estimate the parameters x_1, x_2, x_3 of model (30) for each of the three diseases separately. That is, we consider three independent problems. We consider each t_i as the left limit of each age range $[t_{\min}, t_{\max})$ and y_i as the corresponding observation. (Considering $t_i = (t_{\min} + t_{\max})/2$ would also be another valid alternative.) Figure 1 shows a graphical representation of the data in Table 1, with the definition of t_i mentioned above. To illustrate the result of tackling a parameter fitting

Age group (years)	Proportion seropositive		
	Measles	Mumps	Rubella
$[1, 2)$	0.207	0.115	0.126
$[2, 3)$	0.301	0.147	0.171
$[3, 4)$	0.409	0.389	0.184
$[4, 5)$	0.589	0.516	0.286
$[5, 6)$	0.757	0.669	0.400
$[6, 7)$	0.669	0.768	0.503
$[7, 8)$	0.797	0.786	0.524
$[8, 9)$	0.818	0.798	0.634
$[9, 10)$	0.866	0.878	0.742
$[10, 11)$	0.859	0.861	0.664
$[11, 12)$	0.908	0.844	0.735
$[12, 13)$	0.923	0.881	0.815
$[13, 14)$	0.889	0.895	0.768
$[14, 15)$	0.936	0.882	0.842
$[15, 17)$	0.889	0.869	0.760

Age group (years)	Proportion seropositive		
	Measles	Mumps	Rubella
$[17, 19)$	0.898	0.895	0.869
$[19, 21)$	0.959	0.911	0.844
$[21, 23)$	0.957	0.920	0.852
$[23, 25)$	0.937	0.915	0.907
$[25, 27)$	0.918	0.950	0.935
$[27, 29)$	0.939	0.909	0.921
$[29, 31)$	0.967	0.873	0.896
$[31, 33)$	0.973	0.880	0.890
$[33, 35)$	0.943	0.915	0.949
$[35, 40)$	0.967	0.906	0.899
$[40, 45)$	0.946	0.933	0.955
$[45, 55)$	0.961	0.917	0.937
$[55, 65)$	0.968	0.898	0.933
$[65, +\infty)$	0.968	0.839	0.917

Table 1: Proportion of seropositive for measles, mumps and rubella by age group.

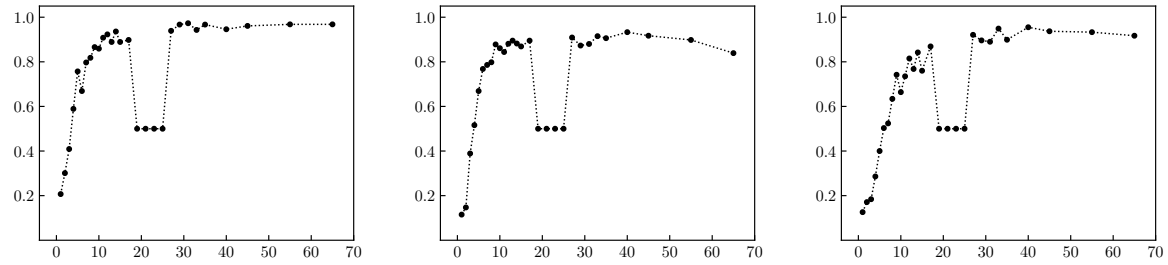
problem in the presence of outliers using the OVO approach, we contaminated the observations of the age groups $[19, 21)$, $[21, 23)$, $[23, 25)$, and $[25, 27)$, replacing the corresponding observation with 0.5. The modified observations are shown in Figure 2. As initial guess $x^0 \in \Omega$, we considered the least squares solution using the data with the inclusion of outliers, namely, $x^0 \approx (0.379029, 0.500859, 0.016986)^T$ for measles, $x^0 \approx (0.285745, 0.424520, 0.005894)^T$ for mumps, and $x^0 \approx (0.117309, 0.341322, 0.026605)^T$ for rubella.



(a) measles (b) mumps (c) rubella

Figure 1: Observed proportion of seropositive for the three considered diseases.

We solved the OVO problem (3) with $o \in \{0, 1, \ldots, 10\}$. Table 2 and Figure 3 show the results. The table shows, for each value of o, the smallest value found for the OVO function (column $f(x^*)$) and, as a measure of Algorithm 2.1 performance, the number of iterations (column "#it"), the number of functional evaluations (column "#fcnt"), and the CPU time in



(a) measles (b) mumps (c) rubella

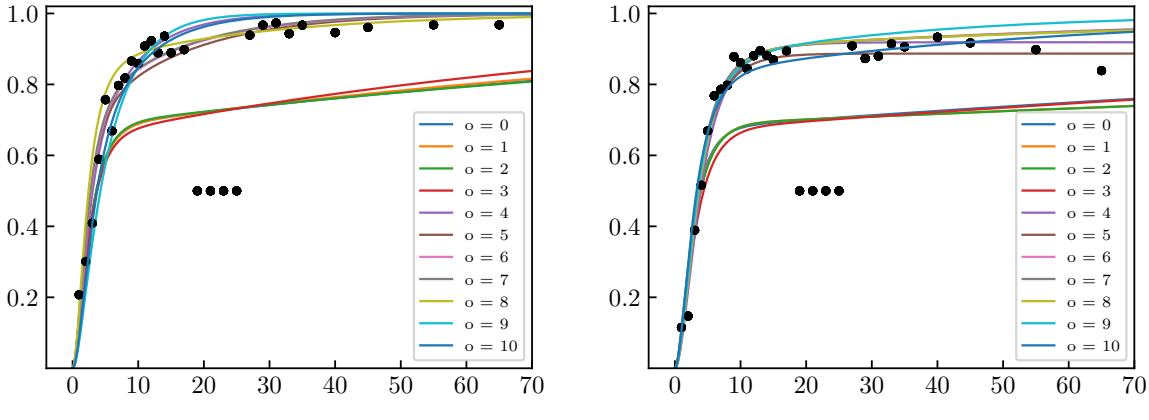
Figure 2: Observed proportion of seropositive for the three considered diseases after the inclusion of outliers.

seconds (column "Time") that were necessary to meet the stopping criterion. The figures in the table show that the smallest value found for the objective function of the OVO problem is on the order of 10^{-2} when $o \leq 3$ and drops by an order of magnitude when $o \geq 4$. This shows that this approach might be used to automatically detect the number of outliers contained in the data. In the figure, some of the curves appear overlapped, but as expected the models whose parameters were fitted considering $0 \leq o \leq 3$ fail to reproduce the observed data.

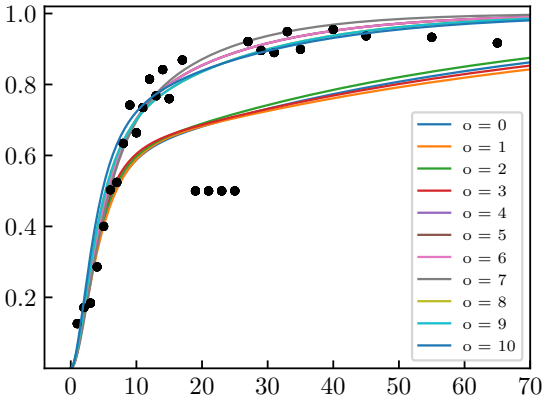
Figure 4 shows, on the left, the models adjusted when considering $o \in \{4, 5, 6\}$. It is not entirely clear that the model found by considering $o = 4$ is "the best"; and comparing the values of $f(x^*)$ obtained in the three cases does not help to decide, since it is natural that the more observations are left out, the better (smaller) is the value found. This suggests that, assuming model (30) is "correct", there are already outliers in the observed data available in [14]. Figure 4 shows on the right side the fitted models considering $o = 10$. In these plots, the observations that the optimal solution of the OVO problem points out as outliers are highlighted in red. It is clear that choosing these observations manually would be practically impossible.

o	measles				mumps				rubella			
	$f(x^*)$	#it	#fcnt	Time	$f(x^*)$	#it	#fcnt	Time	$f(x^*)$	#it	#fcnt	Time
0	2.688E−02	8	40	3.463E−03	2.161E−02	9	28	2.811E−03	2.161E−02	7	33	2.102E−03
1	2.638E−02	5	19	1.735E−03	2.125E−02	5	12	1.963E−03	2.151E−02	5	27	1.397E−03
2	2.609E−02	5	19	1.874E−03	2.107E−02	4	11	1.578E−03	1.969E−02	23	56	3.256E−03
3	2.550E−02	8	30	3.086E−03	2.087E−02	4	13	2.114E−03	2.017E−02	7	17	1.478E−03
4	3.496E−03	16	34	1.997E−03	3.180E−03	10	34	3.162E−03	3.172E−03	4	21	2.127E−03
5	2.871E−03	8	15	2.500E−03	1.760E−03	6	18	1.295E−03	2.999E−03	4	11	1.782E−03
6	2.084E−03	3	7	1.165E−03	1.356E−03	5	12	2.235E−03	2.825E−03	4	11	1.818E−03
7	1.651E−03	7	12	1.483E−03	1.315E−03	5	10	1.830E−03	1.983E−03	3	9	1.053E−03
8	1.136E−03	6	12	2.211E−03	1.086E−03	6	13	2.614E−03	2.617E−03	4	11	1.127E−03
9	2.286E−03	3	4	9.820E−04	1.113E−03	4	8	1.928E−03	2.492E−03	4	20	1.419E−03
10	1.187E−03	2	3	9.740E−04	1.065E−03	4	9	1.877E−03	1.751E−03	5	12	1.809E−03

Table 2: Details of applying Algorithm 2.1 for solving the OVO problem of Section 4.1 with $p = m - o$ and $o \in \{0, 1, \ldots, 10\}$.



(a) measles (b) mumps



(c) rubella

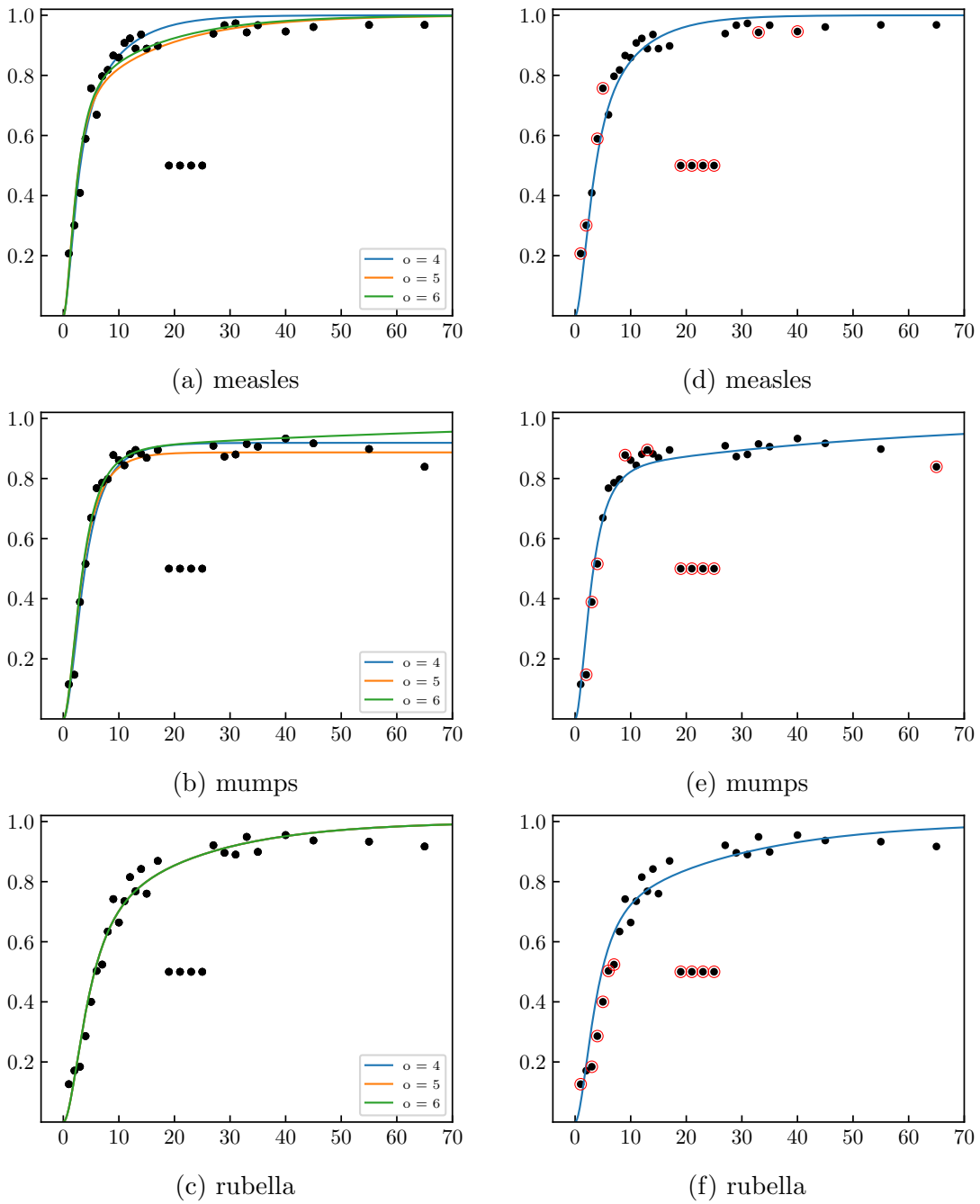
Figure 3: Models adjusted by solving the OVO problem of Section 4.1 with $p = m - o$ and $o \in \{0, 1, \ldots, 10\}$.

4.2 Osborne II function

In the present section, following [6], we consider a variation of problem Osborne II from [18]. The problem consists in finding parameters $x_1, \ldots, x_{11}$ to fit the model

$$y(t, x) = x_1 \exp(-tx_5) + x_2 \exp(-(t - x_9)^2 x_6) + x_3 \exp(-(t - x_{10})^2 x_7) + x_4 \exp(-(t - x_{11})^2 x_8)$$

to data (t_i, y_i) for $i = 1, \ldots, 65$ reported in [18, p.25]. The variation consists in introducing 13 additional data representing outliers. The outliers were taken from [6]. Considering the original data plus the 13 outliers, we arrive to $m = 78$. There are no constraints in the unknown parameters and, therefore, $\Omega = \mathbb{R}^n$ in this case. As initial guess x^0, we considered the least squares solution using the data with the inclusion of outliers, namely, $x^0 \approx (1.312197, 0.367105, 0.551044, 0.642714, 0.596455, 2.306908, 0.365859, 8.197276, 2.016720, 4.339855, 5.686341)^T$. We solved the OVO problem (3) with $o \in \{0, 1, \ldots, 15\}$. Table 3 shows



(a) measles (d) measles

(b) mumps (e) mumps

(c) rubella (f) rubella

Figure 4: On the left side, the models fitted with $p = m - o$ and $o \in \{4, 5, 6\}$. On the right side, the models fitted with $o = 10$, highlighting the observations that the optimal solution to the OVO problem points to as outliers.

the results. The clear drop in the value of $f(x^*)$ when $o = 13$ shows that the method correctly identified the number of outliers. The slightly increase in the value of $f(x^*)$ from $o = 6$ to $o = 7$, from $o = 7$ to $o = 8$, and from $o = 13$ to $o = 14$ shows that the method may be finding local non-global solutions. Anyway, it does not effect the identification of the correct number of outliers neither the adequate model fit for the case $o = 13$.

Figure 5 illustrates the models found for the cases $o \in \{0, 3, 8, 13\}$. It is clear that the cases with $o \neq 13$ are very poor due to the presence of the outliers, while the case $o = 13$ corresponds to a model that correctly represents the real data. In [6], a different problem is solved to fit the model: the function being minimized is the LOVO function defined as $f(x) = \sum_{j=1}^{p} f_{i_j}(x)$ instead of the OVO function defined as $f(x) = f_{i_p}(x)$ and considered in the present work. Moreover, since the problem is different, a different method is employed, for which the stopping criterion is not mentioned. Anyway, comparing Table 3 with [6, p. 18, Table 1], it is possible to see that both methods employ a similar number of iterations and function evaluations. Additionally, a comparison between Figure 5 and [6, p. 18, Figure 2] shows that solutions found are qualitative equivalent.

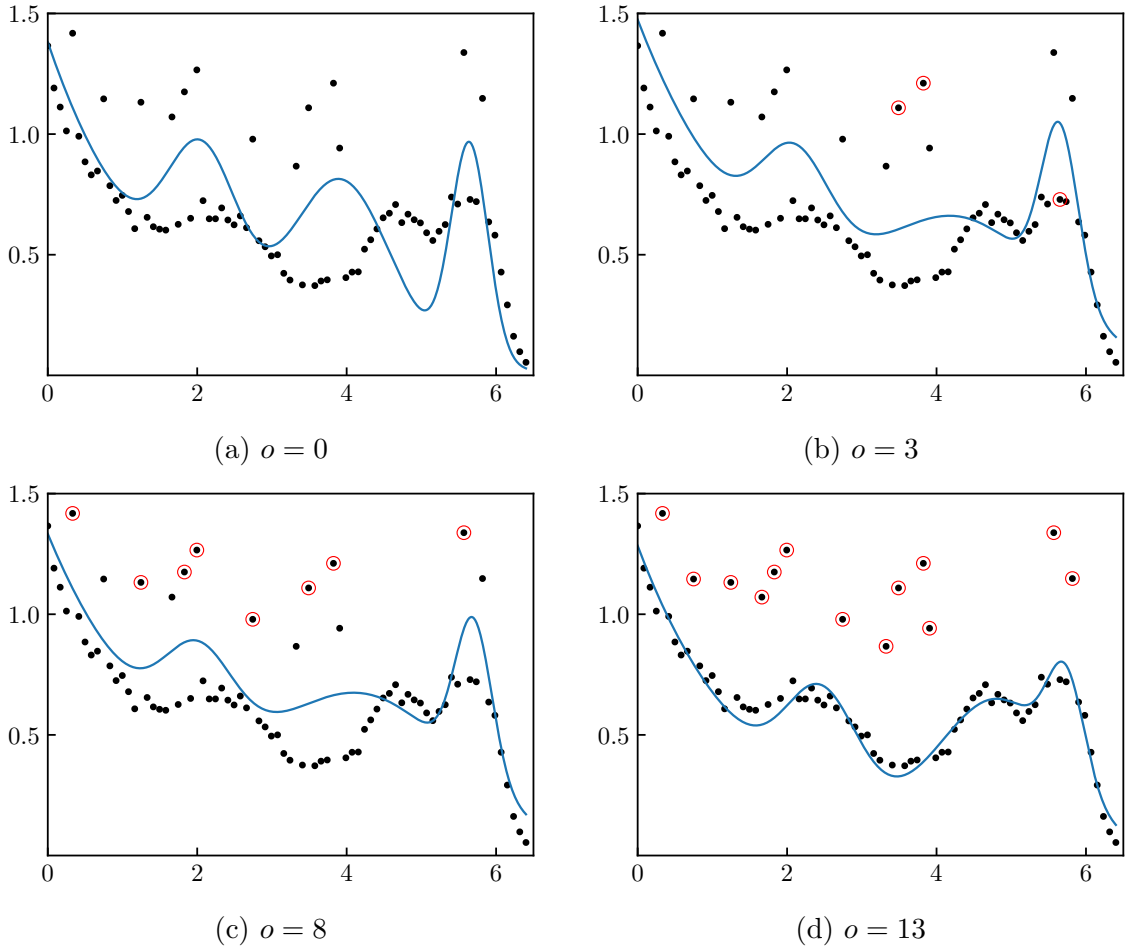
o	$f(x^*)$	#it	# fcnt	Time
0	8.055e−02	85	139	3.561e−02
1	6.788e−02	17	44	5.022e−03
2	4.640e−02	30	100	1.441e−02
3	4.585e−02	28	82	1.543e−02
4	3.917e−02	13	43	8.359e−03
5	3.705e−02	18	58	1.114e−02
6	3.633e−02	30	44	1.360e−02
7	3.651e−02	7	18	5.089e−03
8	3.694e−02	10	31	4.082e−03
9	3.673e−02	28	37	8.291e−03
10	3.640e−02	5	12	2.610e−03
11	2.615e−02	12	38	3.815e−03
12	2.306e−02	7	27	2.659e−03
13	3.714e−03	26	62	1.269e−02
14	3.828e−03	22	48	1.406e−02
15	2.804e−03	35	69	2.797e−02

Table 3: Details of applying Algorithm 2.1 for solving the OVO problem of Section 4.2 with $p = m - o$ and $o \in \{0, 1, \ldots, 15\}$.

4.3 Third degree polynomial in the canonical basis

In the present section, following [3], we consider the model

$$y(t, x) = x_1 + x_2 t + x_3 t^2 + x_4 t^3,$$



(a) $o = 0$ (b) $o = 3$

(c) $o = 8$ (d) $o = 13$

Figure 5: Models adjusted by solving the OVO problem of Section 4.2 with $p = m - o$ and $o \in \{0, 3, 8, 13\}$.

where x_1, x_2, x_3, x_4 are unknown parameters satisfying $-10 \leq x_i \leq 10$ for $i = 1, \ldots, 4$. Therefore, we define Ω accordingly. A set of $m = 47$ data (t_i, y_i) containing 10 outliers was taken from [3]. The least squares solution using the data with the inclusion of outliers is given by $\bar{x} \approx (6.460187, 2.707182, -7.541815, 2.160429)^T$. As the OVO problem may have many non-global local minimizers, in this experiment we considered 100 different starting points given by x^0 with components $x_i^0 = \bar{x}_i + r|x_i^0|$, where $r \in [-0.5, 0.5]$ is a random number with uniform distribution. For each considered value of $o \in \{0, \ldots, 12\}$, we report the best minimizer found. Table 4 shows the results while Figure 6 illustrates the solutions obtained for the cases $o \in \{0, 4, 8, 10\}$. In the table, the sudden drop of $f(x^*)$ from the case $o = 9$ to the case $o = 10$ shows that the method is able to clearly detect the amount of outliers contained in the data. The table also shows that from case $o = 0$ to case $o = 10$ the value of $f(x^*)$ decreases monotonically, as expected. The cases $o = 11$ and $o = 12$ are not very relevant and the values of $f(x^*)$ are

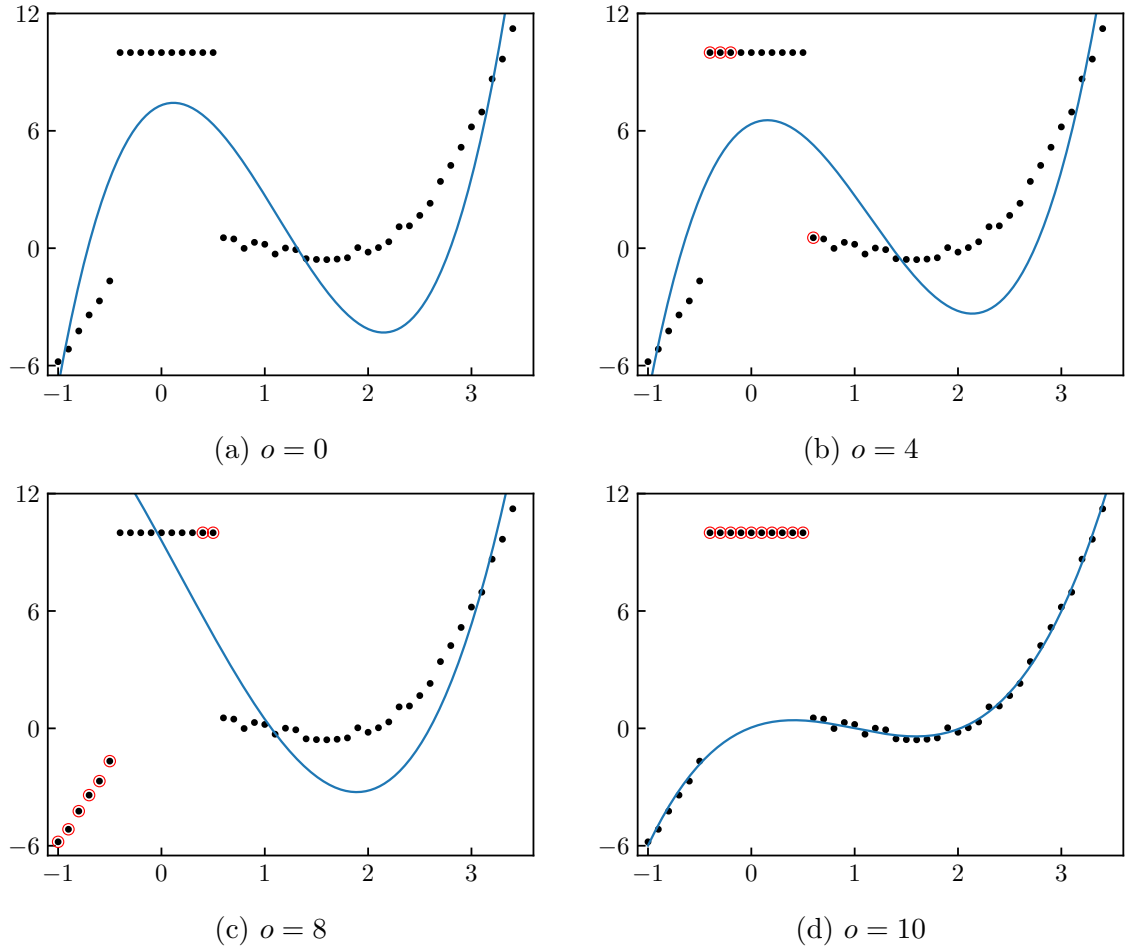
similar to the case $o = 10$. But the fact that they are slightly larger than the value of case $o = 10$ shows that the algorithm found local solutions of good quality but not corresponding to a global minimizer. Figure 6(d) shows that, when the number of outliers is identified, the model fit is very good. Figures 6(a-c) are interesting because they show the type of fit that is possible when the data are considered to contain fewer outliers than they actually do.

In [3], a first-order method for the OVO problem was proposed. The proposed method is similar to the one presented in the present work. The main difference is that in subproblem (6), instead of penalizing $\|x - \bar{x}\|^2$, a box constraint of the form $\|x - \bar{x}\|_\infty \leq \Delta$ is considered. This difference is what allows the method proposed in the present work to pose a worst-case analysis that provides an upper bound for the number of iterations and function evaluations that are necessary to find a solution with a prescribed accuracy. On the opposite side, the method proposed in [3] presents a classical asymptotic convergence theory. If on the one hand the two methods have such a difference in their theoretical properties, it is not expected that they behave very differently in practice. This is exactly what Table 4 shows. In [3] neither the stopping criterion nor the number of function evaluations is reported, but [3, p. 398, Table 2] shows a number of iterations compatible with (in fact, a slightly larger than) the number of iterations of our method, as shown in Table 4.

o	$f(x^*)$	#it	# fcnt	Time
0	1.363e+01	27	105	1.483e−02
1	1.145e+01	37	170	3.047e−02
2	1.004e+01	40	194	1.381e−02
3	9.535e+00	39	190	2.557e−02
4	9.013e+00	40	164	3.945e−02
5	8.455e+00	38	170	2.586e−02
6	7.436e+00	44	225	2.646e−02
7	6.921e+00	38	206	2.109e−02
8	5.503e+00	43	210	2.453e−02
9	4.176e+00	25	117	2.020e−02
10	3.120e−02	31	154	2.421e−02
11	3.452e−02	28	145	1.852e−02
12	3.228e−02	50	262	2.080e−02

Table 4: Details of applying Algorithm 2.1 for solving the OVO problem of Section 4.3 with $p = m - o$ and $o \in \{0, 1, \ldots, 12\}$.

We end this section by showing that the proposed method can be applied to problems with an increasing amount of data. For this purpose we consider problems with $m \in \{10^2, 10^3, \ldots, 10^6\}$ observations. Let $t_{\min} = -1$, $t_{\max} = 3.5$, $t_i = t_{\min} + \left(\frac{i-1}{m-1}\right)(t_{\max} - t_{\min})$ for $i = 1, \ldots, m$, and $\tilde{x} = (0, 2, -3, -1)^T$. Each y_i is considered an outlier with probability 0.1. If the data is not an outlier, y_i corresponds to $y(t_i, \tilde{x})$ plus noise $r \in [-0.5, 0.5]$. If the data is considered an outlier, it is larger than $y(t_i, \tilde{x})$ with probability 0.8 and smaller with probability 0.2. If greater, it is a random value between $y(t_i, \tilde{x})$ and 15. If less, it is a random value between -6 and $y(t_i, \tilde{x})$.



(a) $o = 0$ (b) $o = 4$

(c) $o = 8$ (d) $o = 10$

Figure 6: Models fitted solving the OVO problem of Section 4.3 with $p = m - o$ and $o \in \{0, 4, 8, 10\}$.

Figure 7 shows the data for the case $m = 1{,}000$ as an example. This way of generating data is similar to the one used in [3] to generate the problem with $m = 47$ presented above. The figure shows that some of the data generated as outliers are identical to data generated as "correct". Therefore, discovering the exact number of generated outliers is impossible. Or, in other words, not all data generated as outliers are in fact outliers.

For the case $m = 100$, we solved OVO problems with o varying from 5 to 15 in increments of 1. (The expected number of outliers is 10 in this case). For the case $m = 1{,}000$, which has an expected number of outliers of 100, we vary o from 50 to 150 in increments of 1. For the case $m = 10{,}000$, we vary o from 500 to 1,500 in increments of 10. For the case $m = 100{,}000$, we vary o from 5,000 to 15,000 in increments of 100. For the case $m = 1{,}000{,}000$, we vary o from 50,000 to 150,000 in increments of 1,000. For each value of o, we solved the problem starting from 100 different starting points, as in the case with $m = 47$ presented above. For each value

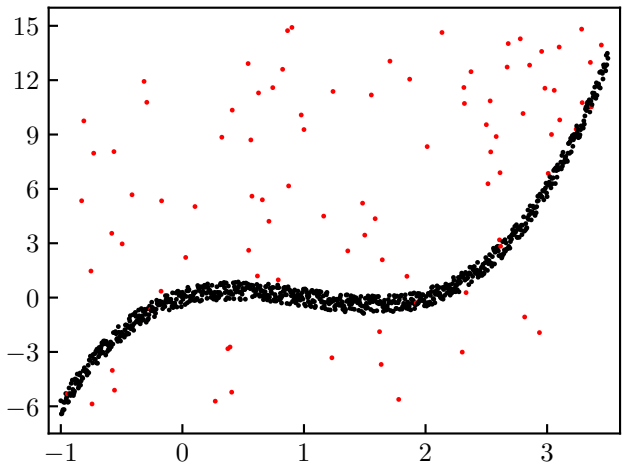
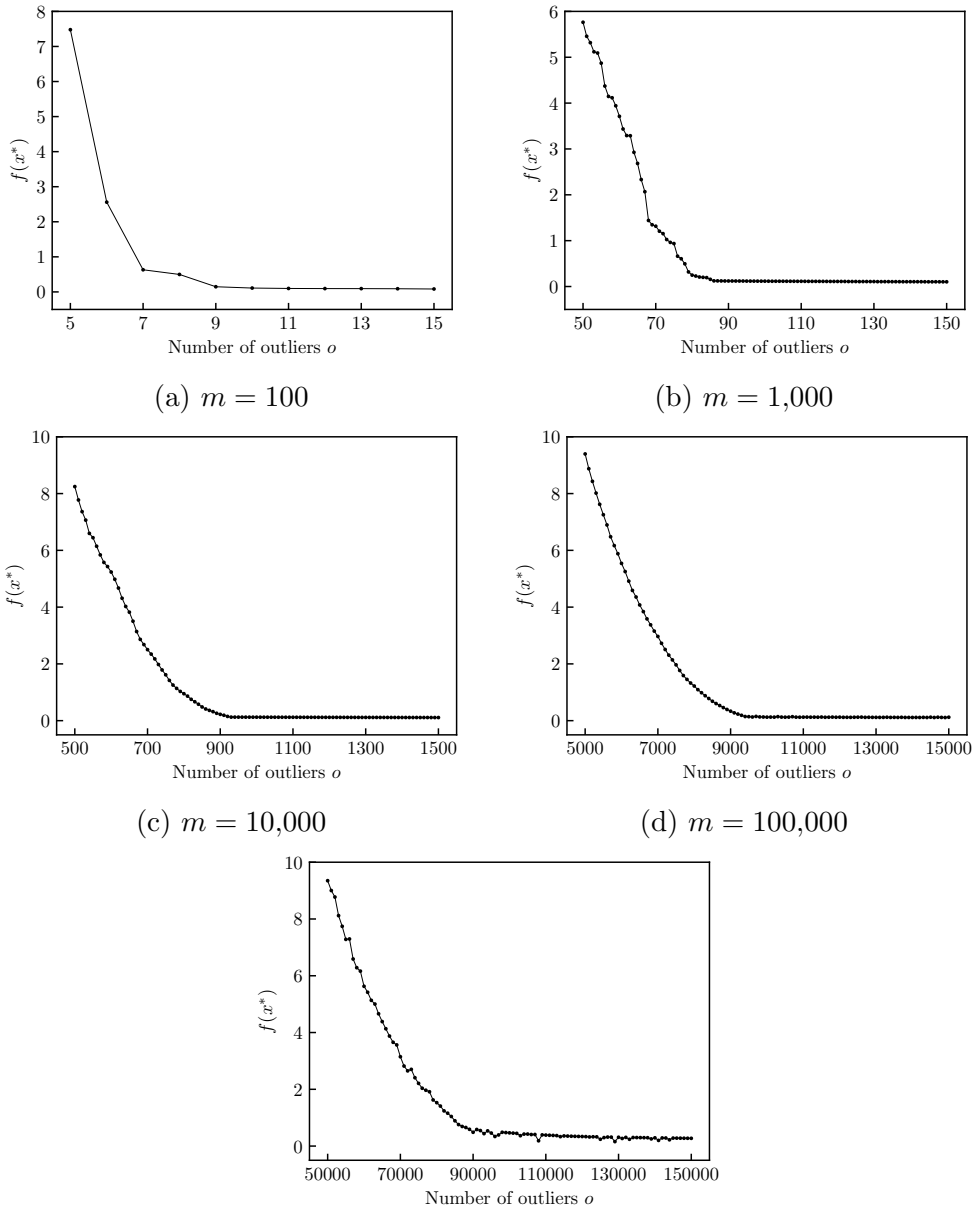


Figure 7: Data with outliers for the case $m = 1{,}000$. For those readers who are viewing the figure with colors, the black dots are the data generated as "correct" (contaminated with noise), while the red dots are the data generated as outliers.

of m, Figure 8 shows the value of $f(x^*)$ as a function of o (for each o, the best value of the 100 runs is shown). The figure shows that in all cases there is a sharp drop of the value of $f(x^*)$ when an approximately correct number of outliers is considered. Note that the exact number of outliers is unknown because of the way they are generated and because when a data is generated as an outlier, it may in some cases look very similar to a real data. Moreover, not all values of o are tested in the experiment. Even under these conditions, the figures show that the method succeeds perfectly well in identifying and disregarding outliers. In Table 5 we show, for each value of m, the amount $\bar{o}$ of data that were generated as outliers, and the smallest value of o for which a value of $f(x^*)$ considered "small" is obtained (i.e. the first value of o after the drop in the value of $f(x^*)$). For these selected cases, performance metrics of the method (number of iterations, number of function evaluations, and CPU time in seconds) are shown. The column showing the time per function evaluation suggests that the cost of the function evaluations increases linearly with the number of data m, although it includes a $m \log(m)$ ordering of the f_i for $i = 1, \ldots, m$.

5 Final remarks

In this paper we introduced a method for the problem of minimizing the order-value function with box constraints. The method is of first order and uses quadratic regularization. As lines of future work we can mention the development of methods for problems with more general constraints and methods using higher order models. Generalized order-value functions are functions whose evaluation at a point x of the domain depends on the order relation of the elements in a set of the form $f_i(x)_{i\in\mathcal{I}}$, where $\mathcal{I} \subset \mathbb{N}$ is a finite set of indices. Problems that include such



(a) $m = 100$ (b) $m = 1{,}000$

(c) $m = 10{,}000$ (d) $m = 100{,}000$

(e) $m = 1{,}000{,}000$

Figure 8: For each value of o within a predefined interval, we solve an OVO problem with fixed o using 100 different starting points and keep the best x^*. These plots show $f(x^*)$ as a function of o. The graphs show that $f(x^*)$ decreases in an accentuated way up to a certain value of o and, from that point on, it decreases very slowly. The smallest value of o for which $f(x^*)$ decreases slowly is the value detected by the strategy as "amount of outliers contained in the data".

m	$\bar{o}$	o	$f(x^*)$	#it	#fcnt	Time	Time / #fcnt
100	10	11	9.852E−02	65	467	2.486E−02	5.324E−05
1,000	92	85	1.567E−01	72	469	5.224E−02	1.114E−04
10,000	980	910	1.905E−01	38	224	1.511E−01	6.747E−04
100,000	10024	9300	1.887E−01	21	90	6.726E−01	7.473E−03
1,000,000	100083	108000	1.863E−01	4	13	1.070E+00	8.233E−02

Table 5: Details of the performance of Algorithm 2.1 when solving OVO problems with increasing amount of data m.

functions in their definition, either in the objective function or in the constraints, are called generalized order-value optimization (GOVO) problems [17]. The problem considered in the present work belongs to this family of problems. Proposing methods with complexity results for other problems of the GOVO family is also a possible line of future work.

Disclosure statement: The authors report there are no competing interests to declare.

References

[1] R. Andreani, E. G. Birgin, J. M. Martínez and M. L. Schuverdt, Augmented Lagrangian methods under the Constant Positive Linear Dependence constraint qualification, *Mathematical Programming* 111, pp. 5–32, 2008.

[2] R. Andreani, E. G. Birgin, J. M. Martínez and M. L. Schuverdt, On Augmented Lagrangian methods with general lower-level constraints, *SIAM Journal on Optimization* 18, pp. 1286–1309, 2008.

[3] R. Andreani, C. Dunder and J. M. Martínez, Order-Value Optimization: Formulation and solution by means of a primal Cauchy method, *Mathematical Methods of Operations Research* 58, pp. 387–399, 2003.

[4] R. Andreani, C. Dunder and J. M. Martínez, Nonlinear-programming reformulation of the order-value optimization problem, *Mathematical Methods of Operations Research* 61, pp. 365–384, 2005.

[5] R. Andreani, J. M. Martínez, M. Salvatierra, and F. S. Yano, Quasi-Newton methods for Order-Value Optimization and Value-at-Risk calculations, *Pacific Journal of Optimization* 2, pp. 11–33, 2006.

[6] R. Andreani, J. M. Martínez, L. Martínez, and F. S. Yano, Low order-value optimization and applications, *Journal of Global Optimization* 43, pp. 1–22, 2009.

[7] R. Andreani, J. M. Martínez, M. Salvatierra, and F. S. Yano, Global order-value optimization by means of a multistart harmonic oscillator tunneling strategy, in *Global Optimization:*

From Theory to Implementation, L. Liberti and N. Maculan (eds.), Springer, Boston, MA, 2006, pp. 379–404.

[8] E. G. Birgin, J. L. Gardenghi, J. M. Martínez, and S. A. Santos, On the use of third-order models with fourth-order regularization for unconstrained optimization, *Optimization Letters* 14, pp. 815–838, 2020.

[9] E. G. Birgin, J. L. Gardenghi, J. M. Martínez, S. A. Santos, and Ph. L. Toint, Evaluation complexity for nonlinear constrained optimization using unscaled KKT conditions and high-order models, *SIAM Journal on Optimization* 26, pp. 951–967, 2016.

[10] E. G. Birgin, J. L. Gardenghi, J. M. Martínez, S. A. Santos, and Ph. L. Toint, Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models, *Mathematical Programming* 163, pp. 359–368, 2017.

[11] E. G. Birgin and J. M. Martínez, *Practical Augmented Lagrangian Methods for Constrained Optimization*, Society for Industrial and Applied Mathematics, Philadelphia, 2014.

[12] E. G. Birgin and J. M. Martínez, Complexity and performance of an Augmented Lagrangian algorithm, *Optimization Methods and Software* 35, pp. 885–920, 2020.

[13] C. Cartis, N. I. M. Gould, and Ph. L. Toint, *Evaluation Complexity of Algorithms for Nonconvex Optimization: Theory, Computation, and Perspectives*, SIAM, Philadelphia, PA, 2022.

[14] C. P. Farrington, Modelling forces of infection for measles, mumps and rubella, *Statistics in Medicine* 9, pp. 953–967, 1990.

[15] J. Gauvin, A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming, *Mathematical Programming* 12, pp. 136–138.

[16] P. Jorion, *Value at Risk: The new benchmark for managing financial risk*, 3rd. ed., McGraw-Hill, 2009.

[17] J. M. Martínez, Generalized Order-Value Optimization, *TOP* 20, pp. 75–98, 2012.

[18] J. J. Moré, B. S. Garbow, and K. E. Hillstrom, Testing unconstrained optimization software, *ACM Transactions on Mathematical Software* 7, pp. 17–41, 1981.

[19] Y. Nesterov and B. T. Polyak, Cubic regularization of Newton method and its global performance, *Mathematical Programming* 108, 177–205, 2006.

[20] Z. Jiang, Q. Hu, and X. Zheng, Optimality condition and complexity of order-value optimization problems and low order-value optimization problems, *Journal of Global Optimization* 69, pp. 511–523, 2017.