A first-order regularized algorithm with complexity properties for the unconstrained and the convexly constrained low order-value optimization problem

G. Q. Álvarez[*] E. G. Birgin[†] J. M. Martínez[‡]

May 23, 2024

Abstract

In this paper we consider the minimization of the unconstrained low order-value function. We also consider the case in which the feasible region is given by a closed convex set, assuming that the projection operation is affordable. For both cases, we introduce regularized first-order algorithms and prove worst-case iteration and evaluation complexity results. Asymptotic convergence results are also presented. The proposed algorithm for the case of constraints given by an arbitrary closed convex set has the classical projected gradient method as a particular case. The algorithms are implemented and several numerical experiments illustrate their application.

Keywords: Low order-value optimization, regularized models, convex constraints, projected gradient, complexity, algorithms.

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

1 Introduction

The low order-value optimization (LOVO) problem, introduced in [7], is part of a family of optimization problems in which the evaluation of the objective function at a given point depends on the sorting of functional values at that point. Given $f_1, \dots, f_m$, with $f_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \dots, m$, examples of such problems include minimizing $\min\{f_1(x), \dots, f_m(x)\}$, minimizing $\max\{f_1(x), \dots, f_m(x)\}$, and, given $1 \le q \le m$ minimizing $f_{i_q(x)}(x)$, where for every $x \in \mathbb{R}^n$, $i_q(x)$ is such that $f_{i_1(x)}(x) \le f_{i_2(x)}(x) \le \cdots \le f_{i_m(x)}(x)$. The latter problem is known as the order-value optimization (OVO) problem [1, 4, 5, 8, 9]. In the LOVO problem the goal is to

[*] 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

[‡] Department of Applied Mathematics, Institute of Mathematics, Statistics and Scientific Computing, State University of Campinas, CP 6065, 13081-970 Campinas SP, Brazil. email: martinez@ime.unicamp.br

minimize $\sum_{j=1}^{q} f_{i_j(x)}(x)$. An overview and generalization of OVO and LOVO type problems is presented in [20].

In [7], two methods were introduced for the unconstrained LOVO problem, each of them with asymptotic convergence to so called weakly critical points and strongly critical points. In addition, a method based on augmented Lagrangians [2, 3, 13, 14] for the constrained LOVO problem was also introduced in [7]. This method has asymptotic convergence to weakly critical points. In [19], it was shown that the minimization of the low order-value function is an NP-hard problem in the case where all f_i are affine functions and the constraints are given by a polytope. There is no method in the literature with worst-case complexity analysis for the LOVO problem. In this paper we introduce two first-order regularized algorithms for the unconstrained LOVO problem and for the LOVO problem in which the feasible region is given by a closed convex set. For both algorithms we prove worst-case iteration and evaluation complexity for convergence to so called approximate strongly critical points with precision $\epsilon > 0$. Unfortunately, convergence to approximate strongly critical points does not imply asymptotic convergence to strongly critical points. Then, we also prove asymptotic convergence to weakly critical points for the two methods. The proposed algorithm for the case of constraints given by an arbitrary closed convex set has the classical projected gradient method as a particular case when $m = q = 1$.

The LOVO problem with $q = m$ and a suitable choice of functions f_i is a generalization of the classical least squares problem and, as such, has a wide range of applications. Moreover, with the same choice of functions and $q < m$, it can be used to disregard the influence of outliers in the data. Thus, the LOVO problem has interesting applications for robust parameter estimation [16]. The LOVO problem can also be used to find hidden structures in data and one of its most successful applications has been the protein alignment problem [7]. In [15], the LOVO problem was related to a multiple fitting strategy for the estimation of parameters in supercritical fluid extraction models. In [10], a portfolio optimization problems with a constraint on the admissible Value at Risk was modeled as a problem in which a low order-value function appears in the constraints. Algorithms for this type of problem were introduced and portfolio optimization problems with transaction costs were solved.

The remainder of this work is organized as follows. In Section 2, we formally define the LOVO problem and present the concepts of weakly critical and strongly critical points. In Sections 3 and 4, we present the methods and their convergence theory for the unconstrained case and the case where the feasible region is determined by a closed and convex set, respectively. Numerical experiments illustrating the applicability of the proposed methods are presented in Section 5. The final section presents conclusions and possibilities for future work.

Notation. $B_\epsilon(x)$ denotes the closed ball in $\mathbb{R}^n$ centered at the point x and with radius ϵ, that is, $B_\epsilon(x) = \{y \in \mathbb{R}^n \mid \|x - y\| \leq \epsilon\}$. For $a \in \mathbb{R}$, $a_+ = \max\{0, a\}$. For a matrix $A \in \mathbb{R}^{n\times n}$, $\lambda_{\min}(A)$ represents its smallest eigenvalue.

2 Problem definition and preliminaries

Let $f_i : \Omega \to \mathbb{R}$ for $i = 1, \dots, m$ be given, where $\Omega = \mathbb{R}^n$ or $\Omega \subset \mathbb{R}^n$ is a closed and convex set. For a given $q \in \{1, 2, \dots, m\}$ the qth low order-value function $S_q : \Omega \to \mathbb{R}$ is defined as

$$S_q(x) := \sum_{j=1}^{q} f_{i_j(x)}(x),$$

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

$$f_{i_1(x)}(x) \le f_{i_2(x)}(x) \le \cdots \le f_{i_m(x)}(x),$$

that is, S_q is such that, for all $x \in \Omega$, $S_q(x)$ corresponds to the sum of the q smallest values among $\{f_1(x), f_2(x), \dots, f_m(x)\}$. If $q = 1$, then $S_q(x) = \min\{f_1(x), f_2(x), \dots, f_m(x)\}$. This makes it clear that even if all f_i are differentiable, it is very likely that the low order-value function is not. In the present work, we consider the low order-value optimization (LOVO) problem given by

$$\text{Minimize } S_q(x) \text{ subject to } x \in \Omega. \tag{1}$$

There exist exactly

$$r = \binom{m}{q} = \frac{m!}{q!\,(m-q)!}$$

subsets $\mathcal{C}_1, \dots, \mathcal{C}_r$ of the set $\{1, \dots, m\}$ such that $|\mathcal{C}_i| = q$ for $i = 1, \dots, r$. For all $i = 1, \dots, r$, we define

$$F_i(x) = \sum_{j \in \mathcal{C}_i} f_j(x)$$

and

$$F_{\min}(x) = \min\{F_1(x), \dots, F_r(x)\}. \tag{2}$$

Clearly, for every $x \in \Omega$, the permutation $\{i_1(x), \dots, i_q(x)\}$ such that

$$f_{i_1(x)}(x) \le f_{i_2(x)}(x) \le \cdots \le f_{i_q(x)}(x) \le f_{i_{q+1}(x)}(x) \le \cdots \le f_{i_m(x)}(x)$$

coincides with one of the subsets $\mathcal{C}_1, \dots, \mathcal{C}_r$. Therefore, $F_{\min}(x) = S_q(x)$ for all $x \in \Omega$ and, in consequence, the LOVO problem (1) can be rewritten as

$$\text{Minimize } F_{\min}(x) \text{ subject to } x \in \Omega. \tag{3}$$

For all $x \in \Omega$, we define

$$I_{\min}(x) := \{i \in \{1, \dots, r\} \,|\, F_i(x) = F_{\min}(x)\}.$$

Theorem 2.1. *[7, Thm. 2.1] Let x^* be a local minimizer of* (3) *and for $i \in I_{\min}(x^*)$ consider the problem*

$$\textit{Minimize } F_i(x) \textit{ subject to } x \in \Omega. \tag{4}$$

Then, x^* is a local minimizer of (4) *for all $i \in I_{\min}(x^*)$.*

Proof. Assume that, for some $i \in I_{\min}(x^*)$, x^* is not a local minimizer of (4). Then, for every $\epsilon > 0$, there exist $\bar{x} \in B_\epsilon(x^*) \cap \Omega$ such that $F_i(\bar{x}) < F_i(x^*)$. Then, by the definitions of $F_{\min}$ and $I_{\min}$ we have that

$$F_{\min}(\bar{x}) \leq F_i(\bar{x}) < F_i(x^*) = F_{\min}(x^*),$$

which contradicts the fact that x being a local minimizer of (3). □

Theorem 2.2. *[7, Prop. 2.1] Assume that x^* is a local minimizer of* (4) *for all $i \in I_{\min}(x^*)$ and that F_i is continuous at x^* for all $i \notin I_{\min}(x^*)$. Then, x^* is a local minimizer of* (3).

Proof. Since x^* is a local minimizer of (4), there exist $\epsilon_1 > 0$ such that for all $i \in I_{\min}(x^*)$ and for all $x \in B_{\epsilon_1}(x^*) \cap \Omega$

$$F_i(x) \geq F_i(x^*) = F_{\min}(x^*). \tag{5}$$

On the other hand, if $i \notin I_{\min}(x^*)$, then $F_i(x^*) > F_{\min}(x^*)$. Therefore, by the continuity of F_i at x^* for all $i \notin I_{\min}(x^*)$, there exist $\epsilon_2 > 0$ such that for all $i \notin I_{\min}(x^*)$ and for all $x \in B_{\epsilon_2}(x^*) \cap \Omega$

$$F_i(x) > F_{\min}(x^*). \tag{6}$$

Therefore, taken $\epsilon = \min\{\epsilon_1, \epsilon_2\}$, by (5) and (6), we have that, for all $x \in B_\epsilon(x^*) \cap \Omega$ and for all $i = 1, 2, \ldots, r$, $F_i(x) \geq F_{\min}(x^*)$. Thus,

$$F_{\min}(x) \geq F_{\min}(x^*),$$

for all $x \in B_\epsilon(x^*) \cap \Omega$, and we conclude that x^* is a local minimizer of (3). □

Theorem 2.1 states that x^* being a local minimizer of problem (4) for all $i \in I_{\min}(x^*)$ is a necessary condition for x^* being a local minimizer of (3). As a consequence, clearly, being a local minimizer of problem (4) for *some* $i \in I_{\min}(x^*)$ is a (weaker) necessary condition as well. Additionally, it is also clear that satisfying a necessary optimality condition (NOC) of problem (4) for *some* or for *all* $i \in I_{\min}(x^*)$ is also a NOC for problem (3). Thus, following [7], given a NOC for problem (4), we say that $x^* \in \Omega$ is a *weakly critical point* of problem (3) (with respect to the given NOC), if x^* satisfies the given NOC for problem (4) for some $i \in I_{\min}(x^*)$; and we say that $x^* \in \Omega$ is a *strongly critical point* of problem (3), if x^* satisfies the given NOC for problem (4) for all $i \in I_{\min}(x^*)$.

3 Regularized first-order method for the unconstrained case

In this section, we consider $\Omega = \mathbb{R}^n$ and introduce a regularized first-order method associated with the necessary optimality condition

$$\nabla F_i(x) = 0 \tag{7}$$

of problem (4). For (7) to be a necessary optimality condition of (4), it is enough to ask F_i to be continuously differentiable and for that it is enough to ask $f_1, \ldots, f_m$ to be continuously differentiable. We therefore place below our first assumption.

Assumption A1. *Functions $f_1, \ldots, f_m$ are continuously differentiable for all $x \in \mathbb{R}^n$.*

Given $x^0 \in \mathbb{R}^n$, the method generates iterates $x^1, x^2, \ldots$. The sequence of iterates is finite if, for some k, $\nabla F_i(x^k) = 0$ for all $i \in I_{\min}(x^k)$. Otherwise, the method generates an infinite sequence $\{x^k\}$. Each iteration k of the method is based on the minimization of a first-order regularized model of $F_{\nu_k}(x)$, for some $\nu_k \in I_{\min}(x^k)$. The algorithm follows below.

Algorithm 3.1: Let $\sigma_{\min} > 0$, $\theta \in (0,1]$, $M > 0$, $\gamma > 1$, $\alpha \in (0,1)$, and $x^0 \in \mathbb{R}^n$ be given. Initialize $k \leftarrow 0$.

Step 1. Initialize $j \leftarrow 0$ and $\sigma_{k,0} = 0$ and choose $\nu_k \in I_{\min}(x^k)$ such that

$$\|\nabla F_{\nu_k}(x^k)\| \geq \theta \|\nabla F_i(x^k)\| \text{ for all } i \in I_{\min}(x^k).$$

If $\nabla F_{\nu_k}(x^k) = 0$, stop.

Step 2. Choose $B_{k,j} \in \mathbb{R}^{n \times n}$ symmetric, positive definite, and such that $\|B_{k,j}\| \leq M$, and compute $x^{k,j}_{\text{trial}}$ as a solution to the problem

$$\text{Minimize } \nabla F_{\nu_k}(x^k)^T (x - x^k) + \frac{1}{2}(x - x^k)^T B_{k,j} (x - x^k) + \frac{\sigma_{k,j}}{2}\|x - x^k\|^2. \tag{8}$$

Step 3. Consider condition

$$F_{\min}(x) \leq F_{\min}(x^k) - \alpha \|x - x^k\|^2. \tag{9}$$

If (9) with $x \equiv x^{k,j}_{\text{trial}}$ does not hold, then set $\sigma_{k,j+1} = \max\{\sigma_{\min}, \gamma\sigma_{k,j}\}$, update $j \leftarrow j+1$, and go to Step 2.

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

Remark. At Step 2 of Algorithm 3.1, $x^{k,j}_{\text{trial}}$ is given by $x^{k,j}_{\text{trial}} = x^k - (B_{k,j} + \sigma_{k,j} I)^{-1} \nabla F_{\nu_k}(x^k)$.

Next, we show that Algorithm 3.1 is well defined. For this purpose, the following assumption is needed.

Assumption A2. *There exist $L > 0$ such that, for all $k \geq 0$ and $j = 0, 1, \ldots, j_k$,*

$$F_{\nu_k}(x) \leq F_{\nu_k}(x^k) + \nabla F_{\nu_k}(x^k)^T (x - x^k) + \frac{L}{2}\|x - x^k\|^2 \tag{10}$$

holds with $x = x^{k,j}_{\text{trial}}$, where $x^{k,j}_{\text{trial}}$ is the trial point computed at Step 2 of Algorithm 3.1.

Theorem 3.1. *Suppose that Assumption A1 and A2 hold. Then, if Algorithm 3.1 does not stop at x^k, iteration k is well-defined and*

$$\sigma_k \leq \sigma_{\max} := \max\{\sigma_{\min}, \gamma(2\alpha + L)\}. \tag{11}$$

Proof. Assume that Algorithm 3.1 does not stop at x^k. If $\sigma_{k,j} \geq 2\alpha + L$, then, by the definition (2) of $F_{\min}$, (10) in Assumption A2, $B_{k,j}$ being positive definite, and the fact the objective value at the solution of (8) being necessarily non-positive implies $\nabla F_{\nu_k}(x^k)^T(x^{k,j}_{\text{trial}} - x^k) \leq -\frac{1}{2}(x^{k,j}_{\text{trial}} - x^k)^T B_{k,j}(x^{k,j}_{\text{trial}} - x^k) - \frac{\sigma_{k,j}}{2}\|x^{k,j}_{\text{trial}} - x^k\|^2$, we have that

$$\begin{aligned}
F_{\min}(x^{k,j}_{\text{trial}}) &\leq F_{\nu_k}(x^{k,j}_{\text{trial}}) \\
&\leq F_{\nu_k}(x^{k,j}_{\text{trial}}) + \nabla F_{\nu_k}(x^k)^T(x^{k,j}_{\text{trial}} - x^k) + \frac{L}{2}\|x^{k,j}_{\text{trial}} - x^k\|^2 \\
&\leq F_{\min}(x^k) - \frac{1}{2}(x^{k,j}_{\text{trial}} - x^k)^T B_{k,j}(x^{k,j}_{\text{trial}} - x^k) - \frac{\sigma_{k,j}}{2}\|x^{k,j}_{\text{trial}} - x^k\|^2 + \frac{L}{2}\|x^{k,j}_{\text{trial}} - x^k\|^2 \\
&\leq F_{\min}(x^k) + \left(\frac{L - \sigma_{k,j}}{2}\right)\|x^{k,j}_{\text{trial}} - x^k\|^2 \\
&\leq F_{\nu_k}(x^k) - \alpha\|x^{k,j}_{\text{trial}} - x^k\|^2.
\end{aligned}$$

Thus, $x^{k,j}_{\text{trial}}$ satisfies the descent condition (9) and iteration k is over. That means that if (9) does not hold, we must have $\sigma_{k,j} < 2\alpha + L$. Therefore, the initialization rule $\sigma_{k,0} = 0$ and the updating rule $\sigma_{k,j+1} = \max\{\sigma_{\min}, \gamma\sigma_{k,j}\}$ guarantee that either $\sigma_k = \sigma_{k,0} = 0$, $\sigma_k = \sigma_{k,1} = \sigma_{\min}$, or $\sigma_k = \sigma_{k,j}$ for some $j \geq 2$ and $\sigma_{k,j} < \gamma(2\alpha + L)$ because $\sigma_{k,j-1} < 2\alpha + L$, from which (11) follows. □

The next lemma is an intermediate result that is used in both the asymptotic convergence to weak critical points and the complexity result for finding strong critical points with precision ϵ.

Lemma 3.1. *Suppose that Assumption A1 and A2 hold. Then, for all $k \geq 0$,*

$$\|x^{k+1} - x^k\| \geq \frac{\|\nabla F_{\nu_k}(x^k)\|}{M + \sigma_{\max}}, \tag{12}$$

where $\sigma_{\max}$ is given by (11)*.*

Proof. By definition of Algorithm 3.1,

$$\nabla\left[\nabla F_{\nu_k}(x^k)^T(x - x^k) + \frac{1}{2}(x - x^k)^T B_{k,j}(x - x^k) + \frac{\sigma_{k,j}}{2}\|x - x^k\|^2\right]\Bigg|_{x=x^{k+1}} = 0,$$

i.e.

$$\nabla F_{\nu_k}(x^k) + (B_{k,j} + \sigma_{k,j}I)(x^{k+1} - x^k) = 0.$$

Then

$$\|\nabla F_{\nu_k}(x^k)\| \leq (M + \sigma_{\max})\|x^{k+1} - x^k\|,$$

as we wanted to prove. □

The following theorem shows the worst-case iteration complexity of the proposed method for finding a strongly critical point with precision $\epsilon > 0$, i.e. an iterate x^k such that $\|\nabla F_i(x^k)\| \leq \epsilon$ for all $i \in I_{\min}(x^k)$.

Theorem 3.2. *Suppose that Assumptions A1 and A2 hold and that there exists $F_{\text{low}} \in \mathbb{R}$ such that $F_{\min}(x) \geq F_{\text{low}}$ for all $x \in \mathbb{R}^n$. Given $\epsilon > 0$, the number of iterations k at which $\frac{1}{\theta}\|\nabla F_{\nu_k}(x^k)\| > \epsilon$ is bounded above by*

$$\left\lfloor \left(\frac{(M + \sigma_{\max})^2}{\alpha\,\theta^2}\right)\left(\frac{F_{\min}(x^0) - F_{\text{low}}}{\epsilon^2}\right)\right\rfloor \tag{13}$$

Proof. Let $K = \{k_0, k_1, k_2, \dots\}$ with $k_0 < k_1 < k_2 < \cdots$ be a set of indices such that $\frac{1}{\theta}\|\nabla F_{\nu_{k_j}}(x^{k_j})\| > \epsilon$. Then, by (12) in Lemma 3.1,

$$\|x^{k_j+1} - x^{k_j}\| > \frac{\epsilon\,\theta}{M + \sigma_{\max}}.$$

Therefore, since (9) holds with $x = x^{k_j+1}$, we have that

$$F_{\min}(x^{k_j}) - F_{\min}(x^{k_j+1}) \geq \alpha\|x^{k_j+1} - x^{k_j}\|^2 > \alpha\left(\frac{\epsilon\,\theta}{M + \sigma_{\max}}\right)^2.$$

Summing for all $k_j \in K$, we obtain

$$\sum_{k_j \in K}[F_{\min}(x^{k_j}) - F_{\min}(x^{k_j+1})] > |K|\alpha\left(\frac{\epsilon\,\theta}{M + \sigma_{\max}}\right)^2.$$

The sum on the left-hand side is the sum of the improvements obtained over the iterations $k_j \in K$. But the method generates a sequence $\{x^k\}$ for which $\{F_{\min}(x^k)\}$ is monotonically decreasing. Then, all improvements are positive, their sum is bounded by $F(x^0) - F_{\text{low}}$, and the desired result follows. □

Theorem 3.3. *Suppose that Assumptions A1 and A2 hold and that there exists $F_{\text{low}} \in \mathbb{R}$ such that $F_{\min}(x) \geq F_{\text{low}}$ for all $x \in \mathbb{R}^n$. Given $\epsilon > 0$, in at most $k_{\max} + 1 = O(\epsilon^{-2})$ iterations, where $k_{\max}$ is given by* (13)*, the algorithm finds an iterate x^k such that $\|\nabla F_i(x^k)\| \leq \epsilon$ for all $i \in I_{\min}(x^k)$, i.e. a strongly critical point of problem* (3) *with precision ϵ with respect to the NOC given by* (7)*.*

Proof. This is a direct consequence of Theorem 3.2 and the way ν_k is chosen at Step 1 of Algorithm 3.1. □

Remark. By the definition of Algorithm 3.1, $\sigma_k = \sigma_{k,0} = 0$ or $\sigma_k = \sigma_{k,j_k} = \gamma^{j_k-1}\sigma_{\min}$ for some $j_k > 0$. But Theorem 3.1 guarantees that $\sigma_k \leq \sigma_{\max}$. Then, $j_k \leq \lfloor \log_\gamma(\sigma_{\max}/\sigma_{\min})\rfloor + 1$. By the definition of Algorithm 3.1, the number of functional evaluations per iteration is exactly $j_k + 1$. Therefore, the number of evaluations of $F_{\min}$ at any iteration k is limited by $\lfloor \log_\gamma(\sigma_{\max}/\sigma_{\min})\rfloor + 2$. Combining the limitation on the number of evaluations of $F_{\min}$ per iteration with the limitation on the number of iterations given by Theorem 3.3, we obtain a limitation on the total number of evaluations of $F_{\min}$, i.e. we obtain a worst-case evaluation complexity result.

Unfortunately, the complexity result of Theorem 3.3 for convergence to strongly critical points with precision $\epsilon > 0$ does not imply an asymptotic convergence result to strongly critical points. Consider the problem with $m = 2$, $f_1(x) = x^2$, $f_2(x) = (x-1)^2 - 1$, and $q = 1$, i.e. the problem corresponds to minimizing $\min\{f_1(x), f_2(x)\}$. The point $x_\epsilon = -\epsilon/2$ is a strongly critical point with precision ϵ, because $I_{\min}(x_\epsilon) = \{1\}$ and $|f_1'(x_\epsilon)| = \epsilon$. However, $\lim_{\epsilon\to 0} x_\epsilon = \bar{x} = 0$ is a weakly critical point, because $I_{\min}(\bar{x}) = \{1, 2\}$, $|f_1'(\bar{x})| = 0$, and $|f_2'(\bar{x})| = 2 \neq 0$. (The strongly critical point of this problem is $x^* = 1$, for which $I_{\min}(x^*) = \{2\}$ and $|f_2'(x^*)| = 0$.) The next theorem shows asymptotic convergence to weakly critical points.

Theorem 3.4. *Suppose that Assumptions A1 and A2 hold. Let K be an infinite set of indices such that $i = \nu_k \in I_{\min}(x^k)$ and $\lim_{k\in K} x^k = x^*$. Then $i \in I_{\min}(x^*)$, $\nabla F_i(x^*) = 0$, and*

$$\lim_{k\in K} \left\|\nabla F_{\nu_k}(x^k)\right\| = 0.$$

Proof. Let $K = \{k_0, k_1, k_2, \dots\}$ with $k_0 < k_1 < k_2 < \cdots$ be an infinite set of indices such that $i = \nu_{k_j} \in I_{\min}(x^{k_j})$ and $\lim_{j\to\infty} x^{k_j} = x^*$. By the continuity of F_i,

$$\lim_{j\to\infty} F_i(x^{k_j}) = F_i(x^*). \tag{14}$$

As, for all $j \in \mathbb{N}$, $i = \nu_{k_j}$ and $\nu_{k_j} \in I_{\min}(x^{k_j})$, then, for all $j \in \mathbb{N}$, we have that $F_i(x^{k_j}) \le F_\ell(x^{k_j})$ for all $\ell \in \{1,\dots,r\}$. Then, taking limits for $j \to \infty$, by (14), we see that $F_i(x^*) \le F_\ell(x^*)$ for all $\ell \in \{1,\dots,r\}$. Therefore,

$$i \in I_{\min}(x^*).$$

On the other hand, since $k_{j+1} \ge k_j + 1$, we have:

$$F_i(x^{k_{j+1}}) = F_{\min}(x^{k_{j+1}}) \le F_{\min}(x^{k_j+1}) \le F_{\min}(x^{k_j}) - \alpha\|x^{k_j+1} - x^{k_j}\|^2 \le F_{\min}(x^{k_j}) = F_i(x^{k_j})$$

for all $j \in \mathbb{N}$. Taking limits for $j \to \infty$, we have that $\lim_{j\to\infty} \|x^{k_j+1} - x^{k_j}\|^2 = 0$. Therefore,

$$\lim_{j\to\infty} \|x^{k_j+1} - x^{k_j}\| = 0.$$

Now, by (12), we have that

$$\|\nabla F_i(x^{k_j})\| \le (M + \sigma_{\max})\|x^{k_j+1} - x^{k_j}\|.$$

So, $\lim_{j\to\infty} \|\nabla F_i(x^{k_j})\| = 0$ as we wanted to prove. □

4 Regularized first-order method for the constrained case

In this section, we consider Ω is a closed and convex set and introduce a regularized first-order method associated with the necessary optimality condition

$$P_\Omega(x - \nabla F_i(x)) - x = 0 \tag{15}$$

of problem (4), where P_Ω represents the projector operator onto Ω. For (15) to be a necessary optimality condition of (4), it is enough to ask F_i to be continuously differentiable and for that it is enough to ask $f_1, \dots, f_m$ to be continuously differentiable. We therefore place below our first assumption of the current section.

Assumption A3. *Functions $f_1, \dots, f_m$ are continuously differentiable for all x in an open set that contains Ω.*

Given $x^0 \in \Omega$, the method generates iterates $x^1, x^2, \dots$. The sequence of iterates is finite if, for some k, $P_\Omega(x^k - \nabla F_i(x^k)) - x^k = 0$ for all $i \in I_{\min}(x^k)$. Otherwise, the method generates an infinite sequence $\{x^k\}$. Each iteration k of the method is based on the minimization of a first-order regularized model of $F_{\nu_k}(x)$ subject to $x \in \Omega$, for some $\nu_k \in I_{\min}(x^k)$. The algorithm follows below.

Algorithm 4.1: Let $\sigma_{\min} > 0$, $\theta \in (0,1]$, $\gamma > 1$, $\alpha \in (0,1)$, and $x^0 \in \Omega$ be given. Initialize $k \leftarrow 0$.

Step 1. Initialize $j \leftarrow 0$ and $\sigma_{k,0} = \sigma_{\min}$ and choose $\nu_k \in I_{\min}(x^k)$ such that

$$\|P_\Omega(x^k - \nabla F_{\nu_k}(x^k)) - x^k\| \geq \theta \|P_\Omega(x^k - \nabla F_i(x^k)) - x^k\| \text{ for all } i \in I_{\min}(x^k).$$

If $P_\Omega(x^k - \nabla F_{\nu_k}(x^k)) - x^k = 0$, stop.

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

$$\text{Minimize } \nabla F_{\nu_k}(x^k)^T(x - x^k) + \frac{\sigma_{k,j}}{2}\|x - x^k\|^2 \text{ subject to } x \in \Omega. \tag{16}$$

Step 3. Consider condition

$$F_{\min}(x) \leq F_{\min}(x^k) - \alpha\|x - x^k\|^2. \tag{17}$$

If (9) 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 1. At Step 2, $x_{\text{trial}}^{k,j}$ is given by $x_{\text{trial}}^{k,j} = P_\Omega\left(x^k - \dfrac{1}{\sigma_{k,j}}\nabla F_{\nu_k}(x^k)\right)$.

Remark 2. In Algorithm 4.1, when compared to Algorithm 3.1, the second-order term in the subproblem considered at Step 2 is limited to the regularization term, i.e. there is no positive definite matrix $B_{k,j}$. For that reason, in Step 1 of Algorithm 4.1, $\sigma_{k=0} = \sigma_{\min}$ instead of $\sigma_{k=0} = 0$ as in Algorithm 3.1.

Next, we show that Algorithm 4.1 is well defined. For this purpose, the following assumption is needed.

Assumption A4. *There exist $L > 0$ such that, for all $k \geq 0$ and $j = 0, 1, \dots, j_k$,*

$$F_{\nu_k}(x) \leq F_{\nu_k}(x^k) + \nabla F_{\nu_k}(x^k)^T(x - x^k) + \frac{L}{2}\|x - x^k\|^2 \tag{18}$$

holds with $x = x_{\text{trial}}^{k,j}$, where $x_{\text{trial}}^{k,j}$ is the trial point computed at Step 2 of Algorithm 4.1.

Theorem 4.1. *Suppose that Assumption A3 and A4 hold. Then, if Algorithm 4.1 does not stop at x^k, iteration k is well-defined and*

$$\sigma_k \leq \sigma_{\max} := \max\{\sigma_{\min}, \gamma(2\alpha + L)\}. \tag{19}$$

Proof. Assume that the algorithm does not stop at x^k. If $\sigma_{k,j} \geq 2\alpha + L$, then, by the definition (2) of $F_{\min}$, (18) in Assumption A4, and the fact the objective value at the solution of (16) being necessarily non-positive implies $\nabla F_{\nu_k}(x^k)^T(x^{k,j}_{\text{trial}} - x^k) \leq -\frac{\sigma_{k,j}}{2}\|x^{k,j}_{\text{trial}} - x^k\|^2$, we have that

$$\begin{aligned} F_{\min}(x^{k,j}_{\text{trial}}) &\leq F_{\nu_k}(x^{k,j}_{\text{trial}}) \\ &\leq F_{\nu_k}(x^{k,j}_{\text{trial}}) + \nabla F_{\nu_k}(x^k)^T(x^{k,j}_{\text{trial}} - x^k) + \frac{L}{2}\|x^{k,j}_{\text{trial}} - x^k\|^2 \\ &\leq F_{\min}(x^k) + \left(\frac{L - \sigma_{k,j}}{2}\right)\|x^{k,j}_{\text{trial}} - x^k\|^2 \\ &\leq F_{\nu_k}(x^k) - \alpha\|x^{k,j}_{\text{trial}} - x^k\|^2. \end{aligned}$$

Thus, $x^{k,j}_{\text{trial}}$ satisfies the descent condition (17) and iteration k is over. That means that if (17) does not hold, we must have $\sigma_{k,j} < 2\alpha + L$. Therefore, the initialization rule $\sigma_{k,0} = \sigma_{\min}$ and the updating rule $\sigma_{k,j+1} = \gamma\sigma_{k,j}$ guarantee that either $\sigma_k = \sigma_{k,0} = \sigma_{\min}$ or $\sigma_k = \sigma_{k,j}$ for some $j \geq 1$ and $\sigma_{k,j} < \gamma(2\alpha + L)$ because $\sigma_{k,j-1} < 2\alpha + L$, from which (19) follows. □

The two theorems that follow shows the worst-case iteration complexity of the proposed method for finding an iterate x^k such that $\|P_\Omega(x^k - \nabla F_i(x^k)) - x^k\| \leq \epsilon$ for all $i \in I_{\min}(x^k)$.

Theorem 4.2. *Suppose that Assumptions A3 and A4 hold and that there exists $F_{\text{low}} \in \mathbb{R}$ such that $F_{\min}(x) \geq F_{\text{low}}$ for all $x \in \Omega$. Given $\delta > 0$, the number of iterations k at which $\|P(x^k - \frac{1}{\sigma_k}\nabla F_{\nu_k}(x^k)) - x^k\| > \delta$ is bounded above by*

$$k_{\max} := \left\lfloor \left(\frac{1}{\alpha}\right)\left(\frac{F_{\min}(x^0) - F_{\text{low}}}{\delta^2}\right)\right\rfloor \tag{20}$$

Proof. Let $K = \{k_0, k_1, k_2, \dots\}$ with $k_0 < k_1 < k_2 < \cdots$ be a set of indices such that $\|P_\Omega(x^{k_j} - \frac{1}{\sigma_{k_j}}\nabla F_{\nu_{k_j}}(x^{k_j})) - x^{k_j}\| > \delta$. Then, by the definition of the algorithm,

$$\|x^{k_j+1} - x^{k_j}\| = \left\|P_\Omega\left(x^{k_j} - \frac{1}{\sigma_{k_j}}\nabla F_{\nu_{k_j}}(x^{k_j})\right) - x^{k_j}\right\| > \delta.$$

Therefore, since (17) holds with $x = x^{k_j+1}$, we have that

$$F_{\min}(x^{k_j}) - F_{\min}(x^{k_j+1}) \geq \alpha\|x^{k_j+1} - x^{k_j}\|^2 > \alpha\,\delta^2.$$

Summing for all $k_j \in K$, we obtain

$$\sum_{k_j \in K}[F_{\min}(x^{k_j}) - F_{\min}(x^{k_j+1})] > |K|\alpha\,\delta^2.$$

The sum on the left-hand side is the sum of the improvements obtained over the iterations $k_j \in K$. But the algorithm generates a sequence $\{x^k\}$ for which $\{F_{\min}(x^k)\}$ is monotonically decreasing. Then, all improvements are positive, their sum is bounded by $F(x^0) - F_{\text{low}}$, and the desired result follows. □

Theorem 4.3. *Suppose that Assumptions A3 and A4 hold and that there exists* $F_{\rm low} \in \mathbb{R}$ *such that* $F_{\min}(x) \geq F_{\rm low}$ *for all* $x \in \Omega$*. Given* $\epsilon > 0$*, in at most*

$$\left\lfloor \left(\frac{\max\{1, \sigma_{\max}^2\}}{\alpha\, \theta^2} \right) \left(\frac{F_{\min}(x^0) - F_{\rm low}}{\epsilon^2} \right) \right\rfloor + 1$$

iterations, the algorithm finds an iterate x^k *such that* $\|P_\Omega(x^k - \nabla F_i(x^k)) - x^k\| \leq \epsilon$ *for all* $i \in I_{\min}(x^k)$*, i.e. a strongly critical point of problem* (3) *with precision* ϵ *with respect to the NOC given by* (15)*.*

Proof. Consider $\delta = \epsilon\, \theta / \max\{1, \sigma_{\max}\}$. By Theorem 4.2, the number of iterations at which $\left\| P_\Omega \left(x^k - \frac{1}{\sigma_k} \nabla F_{\nu_k}(x^k) \right) - x^k \right\| > \epsilon\, \theta / \max\{1, \sigma_{\max}\}$ is limited by $k_{\max}$ defined in (20) with $\delta = \epsilon\, \theta / \max\{1, \sigma_{\max}\}$. On the other hand, for any iteration k, from [18, Ex. 3.2.16] and (19), it follows that

$$\left\| P_\Omega \left(x^k - \nabla F_{\nu_k}(x^k) \right) - x^k \right\| \leq \max\{1, \sigma_{\max}\} \left\| P_\Omega \left(x^k - \frac{1}{\sigma_k} \nabla F_{\nu_k}(x^k) \right) - x^k \right\|.$$

This means that the number of iterations at which $\left\| P_\Omega \left(x^k - \nabla F_{\nu_k}(x^k) \right) - x^k \right\| > \epsilon\, \theta$ is limited by $k_{\max}$ in (20) with $\delta = \epsilon\, \theta / \max\{1, \sigma_{\max}\}$ and the result follows from the way ν_k is chosen at Step 1. □

Remark. By the definition of Algorithm 4.1, $\sigma_k = \sigma_{k,j_k} = \gamma^{j_k} \sigma_{\min}$ for some $j_k \geq 0$. But Theorem 4.1 guarantees that $\sigma_k \leq \sigma_{\max}$. Then, $j_k \leq \lfloor \log_\gamma(\sigma_{\max}/\sigma_{\min}) \rfloor$. By the definition of Algorithm 4.1, the number of functional evaluations per iteration is exactly $j_k + 1$. Therefore, the number of evaluations of $F_{\min}$ at any iteration k is limited by $\lfloor \log_\gamma(\sigma_{\max}/\sigma_{\min}) \rfloor + 1$. Combining the limitation on the number of evaluations of $F_{\min}$ per iteration with the limitation on the number of iterations given by Theorem 4.3, we obtain a limitation on the total number of evaluations of $F_{\min}$, i.e. we obtain a worst-case evaluation complexity result.

The next theorem shows asymptotic convergence to weakly critical points.

Theorem 4.4. *Suppose that Assumptions A3 and A4 hold. Let* K *be an infinite set of indices such that* $i = \nu_k \in I_{\min}(x^k)$ *and* $\lim_{k \in K} x^k = x^*$*. Then* $i \in I_{\min}(x^*)$*,* $P_\Omega(x^* - \nabla F_i(x^*)) - x^* = 0$*, and*

$$\lim_{k \in K} \left\| P_\Omega(x^k - \nabla F_{\nu_k}(x^k)) - x^k) \right\| = 0.$$

Proof. Let $K = \{k_0, k_1, k_2, \dots\}$ with $k_0 < k_1 < k_2 < \cdots$ be an infinite set of indices such that $i = \nu_{k_j} \in I_{\min}(x^{k_j})$ and $\lim_{j \to \infty} x^{k_j} = x^*$. By the continuity of F_i,

$$\lim_{j \to \infty} F_i(x^{k_j}) = F_i(x^*). \tag{21}$$

As, for all $j \in \mathbb{N}$, $i = \nu_{k_j}$ and $\nu_{k_j} \in I_{\min}(x^{k_j})$, then, for all $j \in \mathbb{N}$, we have that $F_i(x^{k_j}) \leq F_\ell(x^{k_j})$ for all $\ell \in \{1, \dots, r\}$. Then, taking limits for $j \to \infty$, by (21), we see that $F_i(x^*) \leq F_\ell(x^*)$ for all $\ell \in \{1, \dots, r\}$. Therefore,

$$i \in I_{\min}(x^*).$$

On the other hand, since $k_{j+1} \geq k_j + 1$, we have:

$$F_i(x^{k_{j+1}}) = F_{\min}(x^{k_{j+1}}) \leq F_{\min}(x^{k_j+1}) \leq F_{\min}(x^{k_j}) - \alpha\|x^{k_j+1} - x^{k_j}\|^2 \leq F_{\min}(x^{k_j}) = F_i(x^{k_j})$$

for all $j \in \mathbb{N}$. Taking limits for $j \to \infty$, we have that $\lim_{j\to\infty} \|x^{k_j+1} - x^{k_j}\|^2 = 0$. Therefore,

$$\lim_{j\to\infty} \|x^{k_j+1} - x^{k_j}\| = 0.$$

But, by the definition of the algorithm and [18, Ex. 3.2.16], we have that

$$\begin{aligned}\|x^{k_j+1} - x^{k_j}\| &= \left\| P_\Omega\left(x^{k_j} - \tfrac{1}{\sigma_{k_j}} \nabla F_{\nu_{k_j}}(x^{k_j})\right) - x^{k_j} \right\| \\ &\geq \frac{1}{\max\{1, \sigma_{\max}\}} \left\| P_\Omega\left(x^{k_j} - \nabla F_{\nu_{k_j}}(x^{k_j})\right) - x^{k_j} \right\|\end{aligned}$$

So, $\lim_{j\to\infty} \left\| P_\Omega\left(x^{k_j} - \nabla F_{\nu_{k_j}}(x^{k_j})\right) - x^{k_j} \right\| = 0$ as we wanted to prove. □

5 Numerical experiments

In this section, we illustrate with numerical experiments how parameter estimation problems in the presence of outliers can be solved by modeling them as LOVO type problems and solving them with Algorithm 3.1 or 4.1. In Section 5.1, we consider a simple problem of parameter fitting with a small excerpt of data from the COVID-19 pandemic. In section 5.2, assuming that both test and training data contain outliers, we consider a parameter fitting problem in which the objective is to find an underlying hidden function. In Section 5.3, we consider a simple problem of parameter fitting of an epidemiological model in which the model parameters have constraints.

We implemented Algorithms 3.1 and 4.1 in Fortran 90. For Algorithm 3.1, we implemented as a stopping criterion, in Step 1, $\|\nabla F_{\nu_k}(x^k)\|_\infty \leq \epsilon$, while for Algorithm 4.1, we implemented, also in Step 1, $\|P_\Omega(x^k - \nabla F_{\nu_k}(x^k)) - x^k\|_\infty \leq \epsilon$. In the numerical experiments, based on preliminary experiments, we considered $\sigma_{\min} = 0.1$, $\theta = 1$, $\gamma = 10$, and $\alpha = 10^{-8}$ for both algorithms. For the stopping criteria of both algorithms, we set $\epsilon = 10^{-8}$. In Algorithm 3.1, we considered $B_{k,j} = \nabla^2 F_{\nu_k}(x^k) + \left(-\lambda_{\min}(\nabla^2 F_{\nu_k}(x^k)) + \sqrt{\epsilon_{\text{mach}}}\right)_+ I$ for all k and all j. Condition $\|B_{k,j}\|_F \leq M$ hold in every conducted experiment for $M = 1000$.

Tests were conducted on a computer with a 3.9 GHz AMD Ryzen 5 5600G processor and 32GB 3200 MHz DDR3 RAM memory, running Windows 11 Pro and a Windows Subsystem for Linux with Debian GNU/Linux 11. Code was compiled by the GNU Fortran compiler (version 10.2.1) with the -O3 optimization directive enabled.

5.1 Experiments with a small excerpt of data from COVID-19 pandemic

In this experiment, we considered $m = 30$ data $(t_i, \tilde{y}_i)$, $i = 1, \ldots, m$, corresponding to the 7-day rolling average of deaths per million inhabitants in Italy during the COVID-19 pandemic from August 6 to September 4, 2020, taken from `https://ourworldindata.org`. Figure 1 shows the data. We chose these data because they apparently contain 7 outliers. We divided the m data into training data $(t_i, \tilde{y}_i)$, $i = 1, \dots, \bar{m}$, and test data $(t_i, \tilde{y}_i)$, $i = \bar{m}+1, \dots, m$, with $\bar{m} = 27$. The objective is to fit the model

$$y(t; x) = \tilde{y}_{\bar{m}} + x_1 \left(\frac{t - t_{\bar{m}}}{t_{\bar{m}}} \right) + x_2 \left(\frac{t - t_{\bar{m}}}{t_{\bar{m}}} \right)^2 + x_3 \left(\frac{t - t_{\bar{m}}}{t_{\bar{m}}} \right)^3$$

to the training data and then use it to simulate the prediction of the test data. It is worth noting that the model has only three parameters (x_1, x_2, x_3) since it is imposed $y(t_{\bar{m}}; x) = y_{\bar{m}}$ for all x. We defined the functions $f_i(x) = \frac{1}{2}(y(t_i; x) - \tilde{y}_i)^2$, $i = 1, \dots, \bar{m}$, and solved eleven different LOVO problems with $q = \bar{m} - o$ and the number of outliers $o \in \{0, 1, \dots, 10\}$ using Algorithm 3.1. In all cases we used the least squares solution as the initial solution. Table 1 shows the value of the objective function $S_q(x^*)$, the mean error of the training data given by $\frac{1}{\bar{m}} \sum_{i=1}^{\bar{m}} |y(t_i; x^*) - \tilde{y}_i|$ and the mean error of the test data given by $\frac{1}{m - \bar{m}} \sum_{i=\bar{m}+1}^{m} |y(t_i; x^*) - \tilde{y}_i|$. Additionally, it shows, for the testing data, $\tilde{y}_i$, $y(t_i' x^*)$, and $E_i(x^*) = |y(t_i; x^*) - \tilde{y}_i|$, $i = 28, 29, 30$. In all problems Algorithm 3.1 satisfied the stopping criterion using a maximum of 4 iterations, a single function evaluation per iteration and a CPU time of less than 0.01 seconds. Figure 2 shows the fitted models for the different values of o. The table shows a clear drop of $S_q(x^*)$ and the training and testing errors for $o \geq 7$. It is clear that the strategy is able to identify outliers in the data and that the models give reasonable predictions when $o \geq 7$. Table 1 shows that in these cases the error in the test data is small. It should be noted that this is because the test data set contains no outliers.

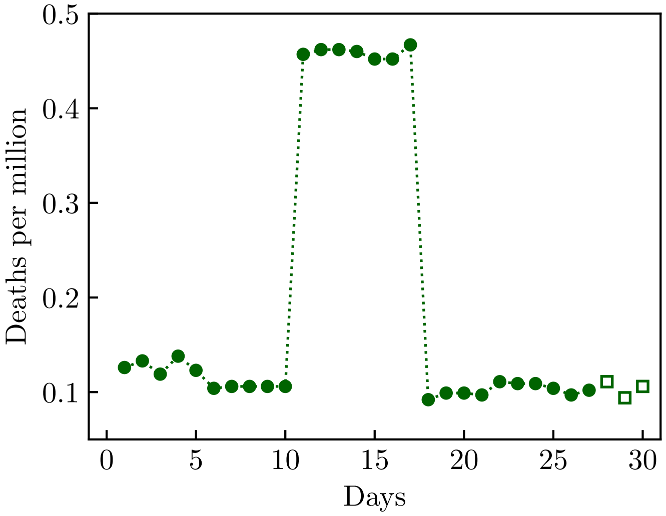
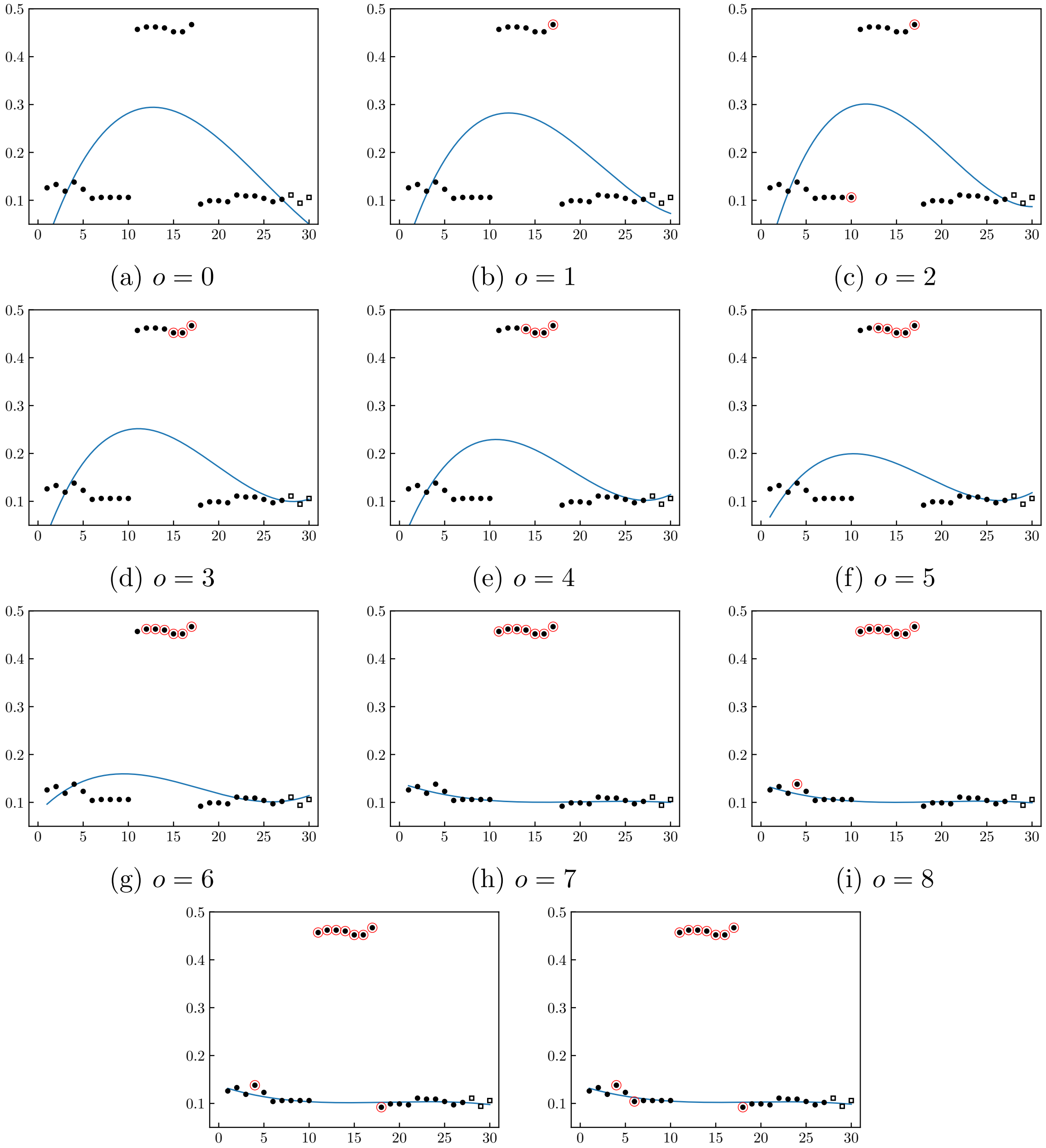


Figure 1: 7-day rolling average of deaths per million inhabitants in Italy during the COVID-19 pandemic from August 6 to September 4, 2020, taken from `https://ourworldindata.org`. We assume that the 7 values above 0.4 are outliers.



(a) $o = 0$

(b) $o = 1$

(c) $o = 2$

(d) $o = 3$

(e) $o = 4$

(f) $o = 5$

(g) $o = 6$

(h) $o = 7$

(i) $o = 8$

(j) $o = 9$

(k) $o = 10$

Figure 2: Graphical representation of the fitted models considering the number of outliers $o \in \{0, 1, \ldots, 10\}$. The solid black dots are the training data and the transparent small squares are the test data. The data circled in red are the data that the method chose to consider outliers and, therefore, excluded from the fitting process.

o	$S_q(x^*)$	$\tilde{y}_{28} = 0.111$		$\tilde{y}_{29} = 0.094$		$\tilde{y}_{30} = 0.106$		Training	Testing
		$y(t_i; x^*)$	$E_i(x^*)$	$y(t_i; x^*)$	$E_i(x^*)$	$y(t_i; x^*)$	$E_i(x^*)$	error	error
0	2.167E-01	0.084	0.027	0.067	0.027	0.050	0.056	0.112	0.037
1	1.952E-01	0.090	0.021	0.080	0.014	0.072	0.034	0.105	0.023
2	1.792E-01	0.094	0.017	0.089	0.005	0.087	0.019	0.103	0.014
3	1.532E-01	0.100	0.011	0.100	0.006	0.104	0.002	0.088	0.006
4	1.264E-01	0.103	0.008	0.107	0.013	0.114	0.008	0.076	0.010
5	9.405E-02	0.105	0.006	0.110	0.016	0.118	0.012	0.058	0.011
6	5.338E-02	0.104	0.007	0.108	0.014	0.114	0.008	0.036	0.010
7	4.509E-04	0.102	0.009	0.101	0.007	0.100	0.006	0.005	0.007
8	2.667E-04	0.101	0.010	0.100	0.006	0.099	0.007	0.004	0.008
9	2.223E-04	0.101	0.010	0.099	0.005	0.098	0.008	0.004	0.008
10	1.936E-04	0.101	0.010	0.100	0.006	0.098	0.008	0.004	0.008

Table 1: Details of the model fitting process when we consider different values $o \in \{0, 1, \ldots, 10\}$ for the presumed number of outliers in the training data.

5.2 Hidden underlying function

In the experiment of the previous section we applied the classical strategy of separating the available data into training and test data, fitting the model using only the training data and testing it using the test data. The small error in both, the training data (disregarding the outliers) and the testing data was used to infer the accuracy of the fitted models. But how to assess the quality of a prediction if the test data are also contaminated with outliers? To overcome this issue, in the experiment of the present section, we considered a ground truth model. With this model we generated data and inserted noise and outliers into the data. We then separated the data into training data and test data and fit the model by solving a LOVO problem with the training data only. At the end, we compared the obtained model with the ground truth model, displaying the test data as a reference only.

The considered ground truth model is given by

$$y(t; c) = c_0 + c_1 t + c_2 t^2 + c_3 t^3, \tag{22}$$

where $(c_0, c_1, c_2, c_3) = (1, 1, -3, 1)$. We initially considered the data set (t_i, y_i), $i = 1, \ldots, m$, with $y_i = y(t_i; c)$, $m = 100$, $t_i = a + (b - a)(i - 1)/(m - 1)$, $a = -1$ and $b = 3$. From these data, we considered the data with noise given by $\tilde{y}_i = y_i + r_i$, where $r_i \in [-0.1, 0.1]$ is a random variable with uniform distribution. Finally, to include outliers, with probability 0.1 we redefined y_i as $y_i = y_i \pm s_i$, where $s_i \in [0.2\, y^{\max}, 0.5\, y^{\max}]$, with $y^{\max} = \max\{|y_i|\}$, is a random variable with uniform distribution, where we considered probability 0.8 that $\tilde{y}_i = y_i - s_i$ and probability 0.2 that $\tilde{y}_i = y_i + s_i$. Figure 3(a) shows the model, the $\tilde{y}_i$ data with noise and outliers and, by way of illustration, the least squares solution.

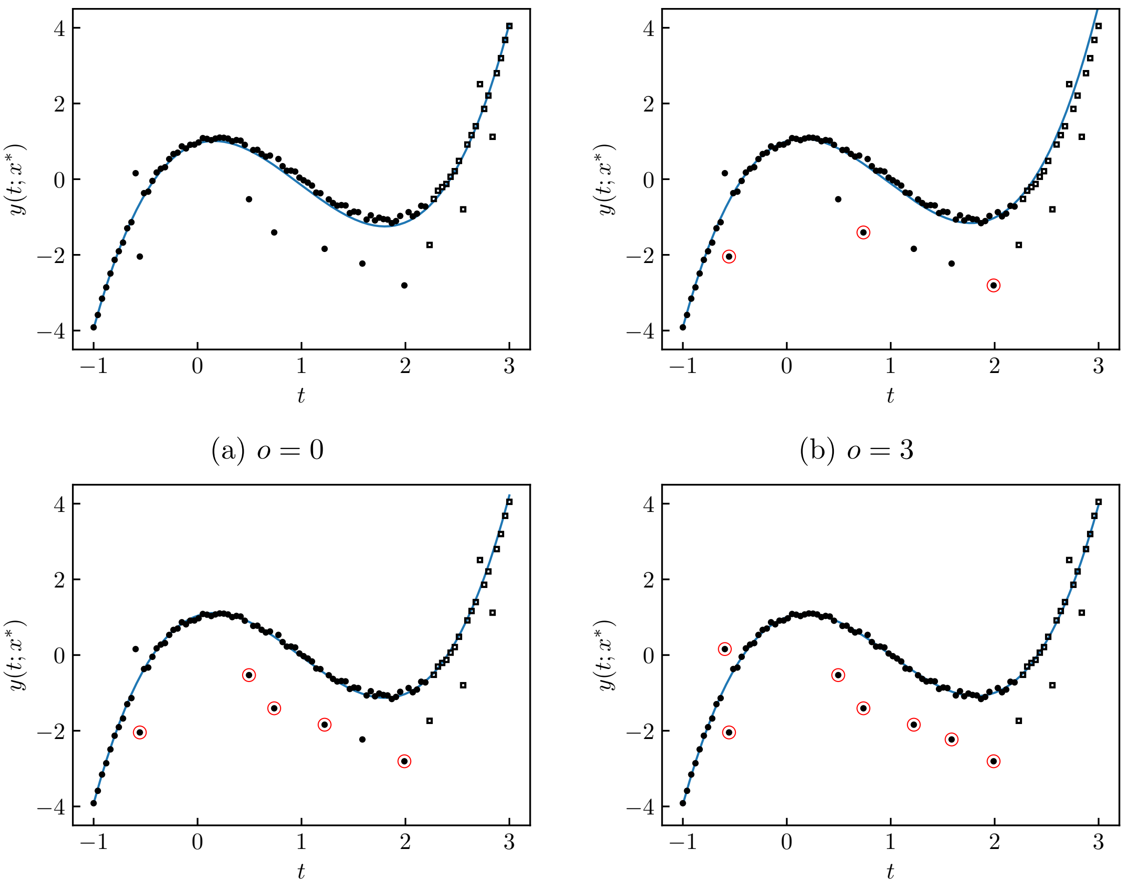
In the fitting process, we considered $\bar{m} = 80$, used the data $(t_i, \tilde{y}_i)$, $i = 1, \ldots, \bar{m}$, as training data and left the data $(t_i, \tilde{y}_i)$, $i = \bar{m} + 1, \ldots, m$, as test data. (Figure 3(a) shows that the training data have 7 outliers, while the test data have 4 outliers.) We defined the functions $f_i(x) = \frac{1}{2}(y(t_i; x) - \tilde{y}_i)^2$, $i = 1, \ldots, \bar{m}$, and solved eleven different LOVO problems with $q = \bar{m} - o$ and the number of outliers $o \in \{0, 1, \ldots, 10\}$ using Algorithm 3.1. In all cases we used the least squares solution as the initial solution. Table 2 shows the solution found x^*, the value of the objective function $S_q(x^*)$ and the distance $\|x^* - c\|_\infty$ from the solution found x^* to the true coefficients c. The table also shows the mean error of the training data and the test data. In all problems Algorithm 3.1 satisfied the stopping criterion using a maximum of 5 iterations, a single function evaluation per iteration and a CPU time of less than 0.01 seconds. Figure 4 shows the value of $S_q(x^*)$ as a function of the number of outliers o. The figure, in logarithmic scale, makes it evident that the method employed is capable of detecting the exact number of outliers contained in the training data. Figure 3(a–d) shows the fitted models $y(t; x^*)$ for some values of o. The observation of the figure makes clear the good fitting of the data with noise in the case where the correct amount of outliers is considered ($o = 7$). It is worth noting that, as anticipated, the mean error of the test data, shown in Table 2, does not help to determine the number of outliers contained in the data and to choose the best model. This is due to the fact that the error is dominated by the error related to the outliers present in the test data. The training error decreases monotonically as the number of outliers considered increases (and, consequently, the amount of fitted data decreases). It is noted that, as well as in the value of $S_q(x^*)$, there is a noticeable drop in the training error for $o \geq 7$, i.e. $q \leq \bar{m} - o = 73$.

o	x_1^*	x_2^*	x_3^*	x_4^*	$S_q(x^*)$	$\|c - x^*\|_\infty$	Training error	Testing error
0	0.942	0.873	-2.993	1.018	7.254	0.127	0.219	0.295
1	0.970	0.921	-3.003	1.006	5.634	0.079	0.181	0.356
2	0.971	0.895	-3.015	1.034	4.181	0.105	0.148	0.437
3	1.006	0.848	-3.023	1.047	3.145	0.152	0.133	0.511
4	1.043	0.873	-3.060	1.053	2.218	0.127	0.108	0.480
5	1.038	0.918	-3.019	1.022	1.335	0.082	0.078	0.332
6	1.026	0.946	-2.976	0.999	0.576	0.054	0.059	0.283
7	1.004	0.981	-2.977	0.993	0.091	0.023	0.043	0.280
8	1.002	0.978	-2.976	0.994	0.086	0.024	0.042	0.279
9	1.005	0.975	-2.978	0.995	0.082	0.025	0.042	0.280
10	1.005	0.977	-2.977	0.993	0.077	0.023	0.041	0.280

Table 2: Details of the solutions found in the problem of discovering a hidden function, when we consider different values $o \in \{0, 1, \ldots, 10\}$ for the presumed number of outliers in the training data.

5.3 Parameter fitting of an epidemiological model

In this section, we examine the epidemiological model devised in [17] to simulate a serological data set including 8870 individuals prior to the implementation of the measles, mumps, and rubella vaccination in the United Kingdom. This problem was recently considered in [1], where it was modeled as an OVO type problem. The only difference between what was done in [1] and what is shown in the present section is that, with the OVO function, the largest quadratic error between the model and the samples that are not considered outliers is minimized, while



(a) $o = 0$

(b) $o = 3$

(c) $o = 5$

(d) $o = 7$

Figure 3: Graphical representation of the fitted models considering the number of outliers $o \in \{0, 3, 5, 7\}$. The solid black dots are the training data and the transparent small squares are the test data. The data circled in red are the data that the method chose to consider outliers and, therefore, excluded from the fitting process.

with the LOVO function the sum of the same quadratic errors is minimized. For this particular experiment, it is expected the two approaches to generate similar solutions. This problem is included in the present study because the model parameters have constraints and, therefore, Algorithm 4.1 will be used in the optimization process instead of Algorithm 3.1. The model seeks to characterize the rate at which vulnerable people become infected with the aforementioned illnesses at various ages. The estimated proportion of seropositive individuals in the unvaccinated part of the sample, broken down into $m = 29$ age categories, is displayed in Table 3, which is derived from [17]. We polluted the observations of the age groups $[19, 21)$, $[21, 23)$, $[23, 25)$, and $[25, 27)$, replacing the corresponding observation with the value 0.5 (numbers colored red in Table 3).

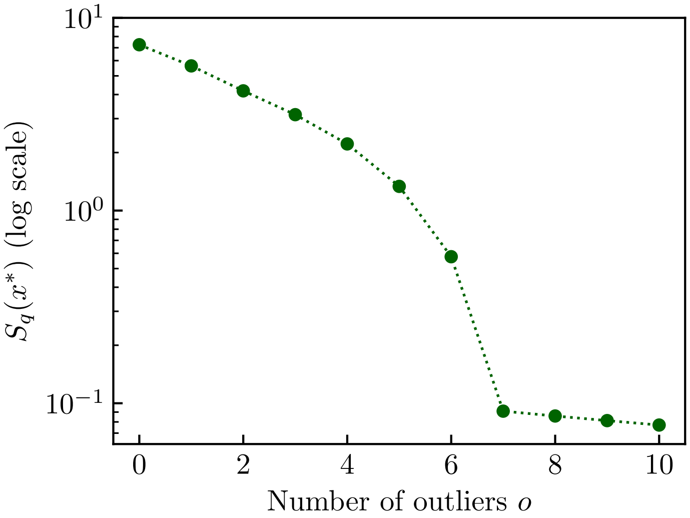


Figure 4: Optimal value $S_q(x^*)$ as a function of the number of outliers o considered in the optimization process. (Recall that $q = \bar{m} - o$.) The clear drop at $o = 7$ shows that the procedure is able to detect the number of outliers in the training data.

Age group	Proportion seropositive		
(years)	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
$[17, 19)$	0.898	0.895	0.869
$[19, 21)$	0.500	0.500	0.500
$[21, 23)$	0.500	0.500	0.500
$[23, 25)$	0.500	0.500	0.500
$[25, 27)$	0.500	0.500	0.500
$[27, 29)$	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 3: Proportion of seropositive for measles, mumps and rubella by age group.

Equation

$$y(t, x) = 1 - \exp\left\{\frac{x_1}{x_2} t e^{-x_2 t} + \frac{x_1}{x_2}\left(\frac{x_1}{x_2} - x_3\right)\left(e^{-x_2 t} - 1\right) - x_3 t\right\}, \tag{23}$$

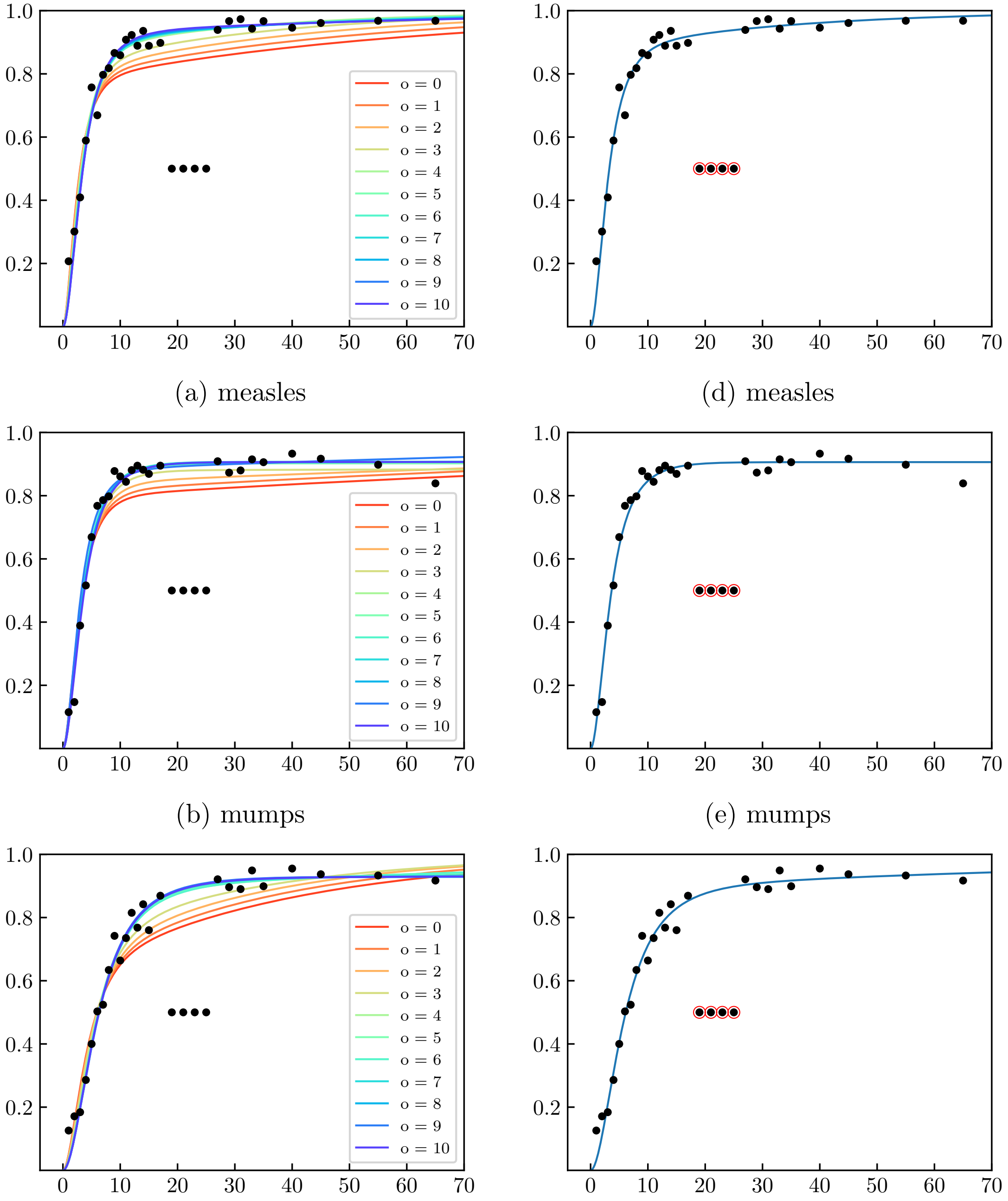
presents the model we want to fit to the data in Table 3, where x_1, x_2, and x_3 are non-negative unknown parameters. We want to estimate the parameters x_1, x_2, and x_3 of model (23) for each of the three illnesses individually, i.e. we have three separate problems. In order to translate the model parameter fitting problem into a LOVO type problem, we define, once again, $f_i(x) = \frac{1}{2}\left(y(t_i, x) - y_i\right)^2$, $i = 1, \dots, m$, where t_i denotes the left limit of an age range $[t_{\min}, t_{\max})$ and y_i denotes the associated observation. We also define $\Omega = \{x \in \mathbb{R}^3 \mid 0 \le x_i \le M_x \text{ for } i = 1, 2, 3\}$, where M_x is a sufficiently large number.

Table 4 shows the result of applying Algorithm 4.1 with $q = m - o$ and $o \in \{0, 1, \dots, 10\}$. The table shows the value of $S_q(x^*)$, the number of iterations, the number of functional evaluations and the CPU time in seconds. Looking at the values of $S_q(x^*)$ for different presumed numbers o of outliers, it is clear that the value drops by an order of magnitude from the case $o = 3$ to the case $o = 4$. This is the only drastic drop and shows that the performed experiment allows to identify the amount of outliers in the data. For values $o \ge 4$, the value of $S_q(x^*)$ drops gradually, but this is expected since it corresponds to a sum of a decreasing number of squared errors. Figure 5 shows, on the left, the obtained models for the different values of o in the three different illnesses. On the right of the figure, the case $o = 4$ is highlighted, showing that the method identified correctly the introduced outliers in all cases.

Table 4 shows that, for $o \ge 4$, the number of iterations of the method is relatively small, while the average number of function evaluations per iteration is of the order of 20. This number seems relatively large and is due to the choice of the initial value of the regularization parameter $\sigma_{\min}$. On the one hand, we could fine-tune this parameter to find one that decreases the number of function evaluations. On the other hand, the reasonable strategy would be to use the information of what was the value of σ_{k-1} to choose the initial value of the regularization parameter in iteration k. This is in fact possible and was done in [11]. It just makes it a little cumbersome to calculate the complexity of the algorithms and we chose not to include this detail in the present work for ease of exposition. The relatively high number of iterations, in relation to the experiments of the two previous sections, is also related to the rather stringent stopping criteria for a first order method such as Algorithm 4.1.

o	measles				mumps				rubella			
	$f(x^*)$	#it	#fcnt	Time	$f(x^*)$	#it	#fcnt	Time	$f(x^*)$	#it	#fcnt	Time
0	3.101E-01	95	478	4.130E-04	2.695E-01	405	2355	1.565E-03	2.278E-01	591	3456	2.364E-03
1	2.455E-01	717	7351	2.545E-03	2.154E-01	283	1575	1.129E-03	1.810E-01	316	1790	1.306E-03
2	1.758E-01	656	5227	2.280E-03	1.559E-01	71	358	3.210E-04	1.315E-01	173	949	7.380E-04
3	9.996E-02	183	1801	6.510E-04	8.915E-02	45	223	2.240E-04	7.816E-02	129	697	5.590E-04
4	1.610E-02	44	879	1.690E-04	1.351E-02	47	216	1.900E-04	1.772E-02	153	866	6.270E-04
5	9.974E-03	36	1013	1.420E-04	8.151E-03	40	178	1.570E-04	1.328E-02	205	1167	8.160E-04
6	6.508E-03	30	1286	1.110E-04	6.007E-03	39	171	1.580E-04	1.045E-02	213	1206	8.450E-04
7	3.821E-03	34	1347	1.280E-04	4.915E-03	37	175	1.500E-04	8.005E-03	241	1364	9.630E-04
8	3.156E-03	30	1485	1.090E-04	3.716E-03	144	755	5.350E-04	5.642E-03	257	1472	1.044E-03
9	2.640E-03	23	1566	8.400E-05	2.445E-03	707	3647	2.207E-03	4.550E-03	91	504	3.940E-04
10	2.055E-03	28	623	1.070E-04	2.620E-03	10	40	4.100E-05	3.801E-03	152	839	6.260E-04

Table 4: Details of applying Algorithm 4.1 for solving the LOVO problem with $q = m - o$ and $o \in \{0, 1, \dots, 10\}$.



(a) measles

(d) measles

(b) mumps

(e) mumps

(c) rubella

(f) rubella

Figure 5: The models fitted with $o \in \{0, 1, \ldots, 10\}$ are on the left. The models fitted with $o = 4$ on the right side show the observations that the LOVO problems' optimal solutions indicate are outliers.

6 Conclusions

In this paper, we introduced first-order methods for the unconstrained LOVO problem and for the LOVO problem restricted to a closed and convex set. For these methods, we developed iteration and evaluation worst-case complexity theory for convergence to approximate strongly critical points and asymptotic theory for convergence to weakly critical points. Regarding the development of algorithms with complexity results for the LOVO problem, there are several possible directions to follow. Analyzing the methods introduced in [7], it would be interesting to try to develop methods with both, complexity and asymptotic convergence theory, to strongly critical points. Moreover, considering the augmented Lagrangian-based method for the general constrained LOVO problem introduced in [7], and the augmented Lagrangian complexity theory developed in [14], it could be attempted to develop a method with complexity theory for the LOVO problem with general constraints. Finally, regularized high-order methods [11, 12] for the LOVO problem could be developed.

Conflict of interest statement: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Author contributions: All authors contributed equally to all phases of the development of this work.

Data availability: The datasets generated during and/or analysed during the current study are available at `http://ime.usp.br/~egbirgin/`.

Funding: This work has been partially supported by the Brazilian agencies FAPESP (grants 2013/07375-0, 2022/05803-3, and 2023/08706-1) and CNPq (grants 302073/2022-1 and 302538/2019-4).

References

[1] G. Q. Álvarez and E. G. Birgin, A first-order regularized approach to the order-value optimization problem, *submitted*, available at `https://www.ime.usp.br/~egbirgin/publications/albmovo.pdf`.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] R. Andreani and J. M. Martínez and L. Martínez and F. S. Yano, Continuous optimization methods for structure alignments, *Mathematical Programming* 112, pp. 93–124, 2008.

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

[8] 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.

[9] 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.

[10] E. G. Birgin, L. F. Bueno, N. Krejić, and J. M. Martínez, Low order-value approach for solving VaR-constrained optimization problems, *Journal of Global Optimization* 51, pp. 715–742, 2011.

[11] 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.

[12] E. G. Birgin and J. L. Gardenghi and J. M. Martínez and 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.

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

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

[15] E. P. Carvalho, F. Pisnitchenko, N. Mezzomo, S. R. S. Ferreira, J. M. Martínez, and J. Martínez, Low order-value multiple fitting for supercritical fluid extraction models, *Computers & Chemical Engineering* 40, pp. 148–156, 2012.

[16] E. V. Castelani, R. Lopes, W. V. I. Shirabayashi, and F. N. C. Sobral, A robust method based on LOVO functions for solving least squares problems, *Journal of Global Optimization* 80, pp. 387–414, 2021.

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

[18] A. Izmailov and M. Solodov, *Otimização Vol. 1 - Condições de otimalidade, elementos de análise convexa e de dualidade*, 4th ed., IMPA, Rio de Janeiro, RJ, Brazil, 2020.

[19] 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.

[20] J. M. Martínez, Generalized order-value optimization, *TOP* 20, pp. 75–98, 2012.