SMOP: Stochastic trust region method for multi-objective problems

Nataša Krejić[*], Nataša Krklec Jerinkić [†], Luka Rutešić [‡§]

January 10, 2025

Abstract

The problem considered is a multi-objective optimization problem, in which the goal is to find an optimal value of a vector function representing various criteria. The aim of this work is to develop an algorithm which utilizes the trust region framework with probabilistic model functions, able to cope with noisy problems, using inaccurate functions and gradients. We prove the almost sure convergence of the proposed algorithm to a Pareto critical point if the model functions are good approximations in probabilistic sense. Numerical results demonstrate effectiveness of the probabilistic trust region by comparing it to competitive stochastic multi-objective solvers. The application in supervised machine learning is showcased by training non discriminatory Logistic Regression models on different size data groups. Additionally, we use several test examples with irregularly shaped fronts to exhibit the efficiency of the algorithm.

Key words: Multi-objective optimization, Pareto-optimal points, Probabilistically fully linear models, Trust-region method, Almost sure convergence.

1 Introduction

Multi-objective optimization problems arise in many real-world applications, such as finance, scientific computing, social sciences, engineering, and be-

[*]Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia. e-mail: natasak@uns.ac.rs

[†]Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia. e-mail: natasa.krklec@dmi.uns.ac.rs

[‡]Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia. e-mail: luka.rutesic@dmi.uns.ac.rs

[§]Corresponding author

yond. These problems are characterized by the need to simultaneously optimize multiple, often conflicting objectives, which significantly complicates the decision making process. Whether you are maximizing efficiency while minimizing computational cost or minimizing risk while maximizing income, identifying the optimal trade-offs is far from straightforward. The complexity comes from the competing nature of the objectives, where improving one criterion comes at the expense of other. The problem we are solving can formally be stated as

$$\min_x f(x) = \min_x (f_1(x), ..., f_q(x))$$

where $f : R^n \to R^q$. The main goal of multi-objective optimization, unlike in the scalar optimization, is to identify a Pareto critical point. A Pareto critical point is a solution that cannot be strictly improved (or dominated) in terms of all objective function values within a local neighborhood, see [16], [23]. When extending this concept to global solutions, we can construct a Pareto front, a set of globally optimal points where no point in the set dominates (strictly improves) another. By finding Pareto critical points, it is possible to find the entire Pareto front, through standard procedures, see [14]. The insight into the structure of the entire set of solutions can be crucial in the decision making process, hence it is important for the model to be able to find the entire front.

Trust region methods for solving this kind of problems work within the standard trust region framework, building a model for each function f_i, generating a direction by solving a multi-model optimization problem and performing the acceptance check as in the classical one dimensional case, see [30]. Therein it is shown that the method converges to a Pareto critical point under standard assumptions. The convergence towards a stationary point is a common main result of papers dealing with multi-objective problems. The complexity of the problem greatly increases if the functions involved are costly. Computing efficiency and high cost of obtaining exact information play an important role and motivation in opting for the stochastic and derivative free approaches. When creating models within a trust region framework, it is possible to use inexact gradient information. Such derivative free trust region approach can be seen in [27]. In the mentioned paper, one criterion is assumed to be a black box function with a difficulty to calculate derivative, while other functions and their derivatives can be easily computed. The convergence towards a Pareto critical point is proved. Another version of a derivative free multi-objective trust region approach is discussed in [4], where radial basis surrogate models are used.

It is also possible to approach this problem within a line search framework. Deterministic approach using Armijo-like conditions with the steepest descent and Newton direction is discussed in [17]. In [17] authors also analyze the projected gradient method for constrained cases. Stochastic multi-gradient multi-criteria approach can be found in [21]. The authors of [21]

successfully extend the classical stochastic gradient (SG, see [22]) method for single-objective optimization to a multi criteria method, and prove sublinear convergence for convex and strongly convex functions.

Random models are also frequently used within the trust region framework in the case of a single objective function, i.e., for the case $q = 1$. A number of approaches are available in literature. Probabilistic trust region method which uses approximate models can be seen in [1]. It is shown there that with probability one the method converges towards a stationary point, if the models are accurate enough with high probability. Trust region method for scalar optimization problems utilizing both approximate functions and gradients can be found in [12]. The analysis therein requires that the model and function estimates are sufficiently accurate with fixed, sufficiently high probability. These probabilities are predetermined and constant throughout the optimization process and almost sure convergence towards a stationary point is proved. Additionally, an adaptive subsampling technique for problems involving functions expressed as finite sums, which are common in applications such as machine learning, is proposed therein. Unlike the traditional subsampling techniques with monotonically increasing size, that method adjusts the size based on the progress. The literature also covers methods specifically designed for optimization of finite sums, which exploit the form through the use of different subsampling strategies, and other various techniques. Some papers in the literature on this topic are [3],[5],[8]-[11],[20], [25],[26].

The method we propose here is based on probabilistically fully linear (quadratic) models, introduced in [1] and used later on in [12, 6]. The concept of full (probabilistic) linearity is extended to vector function in a natural way as explained further on.

Having a fully linear (quadratic) model, one has to deal with the fact that at each step of the trust region method we compute the ratio function using approximations of the function values at subsequent steps. Therefore we can not rely on the true model reduction and the decreasing monotonicity. Thus some additional conditions are needed to control the errors. One possibility is to assume that we work with sufficiently small ε_F accurate values as done in [12]. We propose a different assumption here, see ahead Assumption 3, motivated by the applications from machine learning problems. Roughly speaking we are assuming that the approximate gradient g_i is close enough to the true gradient of the approximate objective $\tilde{f}_i$, which is common in the case of finite sums where one subsamples functional values and takes the approximate gradient as the true gradient of the subsampled function, see [25]. The assumption also holds if one approximates the gradient by finite differences for example.

The quality of approximate models is controlled by a probability sequence α_k which is approaching 1 sufficiently fast. This way one can take advantage of relatively poor model at the beginning of iterative process,

hoping to save some computational costs and yet achieve good approximate solution at the end using high quality models.

Pareto optimal points can be characterized as zeros of the so called marginal functions, see [16]. This characterization reduces to the usual first order optimality conditions (gradient equal to zero) in the case of $q = 1$. The concept of marginal function is used in [30] to define the trust region method. However, as we work with the approximate functions and gradients, an approximate marginal function is used together with the corresponding scalar representation, see [27].

To summarize, the main contribution of this paper are the following. We propose the trust region algorithm which uses approximate function and gradient values to solve multi-criteria optimization problem. The almost sure convergence towards a Pareto stationary point is proved under certain conditions. Numerical experiments are presented, which demonstrate the benefits of the stochastic approach. Using the adaptive subsampling strategy, we manage to solve machine learning problems efficiently with a first order algorithm and we also discuss the use of second order information. At the end of this paper, the standard procedure of finding the Pareto front using the proposed algorithm is implemented and presented.

2 Preliminaries

As mentioned, we are solving the following unconstrained multiobjective minimization problem:

$$\min_x f(x) = \min_x (f_1(x), ..., f_q(x)) \tag{1}$$

where $f : R^n \to R^q$. It is assumed that the functions $f_j, j = 1, ..., q$ are smooth with Lipschitz continuous gradients. Assuming that the explicit evaluation of these functions and its derivatives is unavailable, we will rely on approximating them with $\tilde{f}_i(x)$ and $g_i(x)$, respectively.

For problem (1) one can define efficient and weakly efficient solution as follows.

Definition 1. A point $x^* \in \mathbb{R}^n$ is called (an) efficient (solution) for (1) (or Pareto optimal) if there exists no point $x \in \mathbb{R}^n$ satisfying $f_i(x) \leq f_i(x^*)$ for all $i \in \{1, 2, ..., q\}$ and $f(x) \neq f(x^*)$. A point $x^* \in \mathbb{R}^n$ is called (a) weakly efficient (solution) for (1) (or weakly Pareto optimal) if there exists no point $x \in \mathbb{R}^n$ satisfying $f_i(x) < f_i(x^*)$ for all $i \in \{1, 2, ..., q\}$.

In other words, Pareto point is such that for every direction $d \in \mathbb{R}^n$, there exists a component function f_i with nonnegative directional derivative in that direction d, i.e.,

$$\langle \nabla f_i(x^*), d \rangle \geq 0.$$

In order to mimic the scalar problem's stationarity condition, the marginal function related to (1) is defined

$$\omega(x) = -\min_{\|d\|\leq 1}\left(\max_{i\in\{1,\ldots,q\}}\langle\nabla f_i(x), d\rangle\right). \tag{2}$$

The marginal function plays the role of the norm of the gradient, and it gives us the information about the Pareto optimality. Indeed, if $q = 1$, the optimal direction is $d^{opt}(x) = \nabla f(x)/\|\nabla f(x)\|$, and $\omega(x) = \|\nabla f(x)\|$. The following lemma is proved in [16].

Lemma 1. [16] Let $\mathcal{D}(x)$ be the set of solutions of (2). Then

a) $w(x) \geq 0$, for every $x \in \mathbb{R}^n$;

b) If x is Pareto critical for (1) then $0 \in \mathcal{D}(x)$ and $w(x) = 0$;

c) If x is not Pareto critical of (1) then $w(x) > 0$ and any $d \in \mathcal{D}(x)$ is a descent direction for (1);

d) The mapping $x \to w(x)$ is continuous.

The scalar representation of the multiobjective problem (1) (MOP) is

$$\min_x \phi(x), \quad \phi(x) = \max_{i\in\{1,\ldots,q\}} f_i(x).$$

We assume that $-\infty < \inf\{\phi(x) : x \in \mathbb{R}^n\}$. This problem is not equivalent to problem (1), but every solution of this scalar problem is a Pareto optimal point.

Given that we work with approximate functions $\tilde{f}_i$ and approximate gradients $g_i, i = 1, \ldots, q$, we will follow [27], and consider the approximate marginal function

$$\omega_m(x) = -\min_{\|d\|\leq 1}\left(\max_{i\in\{1,\ldots,q\}}\langle g_i(x), d\rangle\right). \tag{3}$$

where g_i are approximations of ∇f_i. Then the corresponding scalar problem is given by

$$\min_x \tilde{\phi}(x), \quad \tilde{\phi}(x) = \max_{i\in\{1,\ldots,q\}} \tilde{f}_i(x). \tag{4}$$

The quadratic model with which ϕ is approximated by locally in deterministic trust region method is

$$m_k^{true}(d) = \max_{i\in\{1,\ldots,q\}}\{f_i(x_k) + \langle\nabla f_i(x_k), d\rangle\} + \frac{1}{2}\langle d, H_k d\rangle,$$

where H_k is a Hessian approximation. Since we will not operate with true gradients nor functions, we define the approximate quadratic models of ϕ as

$$\tilde{m}_k(d) = \max_{i\in\{1,\ldots,q\}} \tilde{m}_{k,i}(d),$$

$$\tilde{m}_{k,i}(d) = \max_i \{\tilde{f}_i(x_k) + \langle g_i(x_k), d\rangle\} + \frac{1}{2}\langle d, H_k d\rangle.$$

Notice that for each $i = 1, ..., q$, $\nabla \tilde{m}_{k,i}(0) = g_i(x_k)$, and $\tilde{m}_{k,i}(0) = \tilde{f}_i(x_k)$. Our main motivation comes from observing machine learning problems where the functions f_j are in the form of finite sums. In that case, the functions are usually approximated by random subsampling which induces randomness in the optimization process, yielding random sequence of iterates.

Notation. Let $B(x, \delta)$ denote a ball centered at x with the radius δ. An arbitrary vector norm and its induced matrix norm will be denoted by $\|\cdot\|$.

3 Algorithm

Let us define the measure of proximity for the models used to define the trust region algorithm. We will rely on the standard definition of full linearity.

Definition 2. Function $\tilde{m}_{k,i}$ is (c_f, c_g) fully linear (FL) model of f_i on $B(x_k, \delta_k)$, if for every $d \in B(0, \delta_k)$ the following two inequalities hold

$$|f_i(x_k + d) - \tilde{m}_{k,i}(d)| \leq c_f \delta_k^2 \tag{5}$$

$$\|\nabla f_i(x_k + d) - \nabla \tilde{m}_{k,i}(d)\| \leq c_g \delta_k \tag{6}$$

Model $\tilde{m}_k$ is (c_f, c_g) fully linear if for every $i = 1, ..., q$, $\tilde{m}_{k,i}$ is (c_f, c_g) fully linear model of f_i.

The algorithm is constructed in such a way that, at each iteration k, the FL conditions (5)-(6) are satisfied with some high enough probability α_k for each $i = 1, ..., q$. Let us denote by $\mathcal{F}_k$ the σ-algebra generated by $x_0, ..., x_k$. Also, let us define the following events

$$I_{k,i} = \{\tilde{m}_{k,i} \text{ is } (c_f, c_g) \text{ fully linear model of } f_i \text{ on } B_k(x_k, \delta_k)\}.$$

The definition of probabilistically fully linear models [1] is stated as follows.

Definition 3. The random sequence $\{\tilde{m}_{k,i}\}$ is α_k probabilistically (c_f, c_g) fully linear (PFL) if $P(I_{k,i}|\mathcal{F}^k) \geq \alpha_k$ for all k. The model sequence $\{\tilde{m}_k\}$ is α_k probabilistically (c_f, c_g) fully linear if $\{\tilde{m}_{k,i}\}$ is α_k probabilistically (c_f, c_g) fully linear for all $i = 1, ..., q$.

Let us denote by I_k the event that $\{\tilde{m}_k\}$ is α_k probabilistically (c_f, c_g) fully linear, i.e., we have

$$I_k = \bigcap_{j=1}^{q} I_{k,j}. \tag{7}$$

We will assume that $I_{k,i}, i = 1, ..., q$ are independent for all k and thus

$$P(I_k|\mathcal{F}^k) = \prod_{i=1}^{q} P(I_{k,i}|\mathcal{F}^k) \geq \alpha_k^q.$$

This is true if, for instance, the functions $f_i, i = 1, ..., q$ are in the form of mathematical expectation and approximated by independent sampling.

The stochastic trust region method for multiobjective problems (SMOP) we proposed is as follows. Let us denote $\beta_k := 1 + \|H_k\|$.

Algorithm 1. *(SMOP)*

Step 0. Input parameters: $x_0 \in \mathbb{R}^n, \Theta > 0, \delta_{max} > 0, \delta_0 \in (0, \delta_{max})$, $\gamma_1, \eta_1 \in (0, 1), \gamma_2 = 1/\gamma_1, \{\alpha_k\}_k$.

Step 1. Form an α_k probabilistically (c_f, c_g) fully linear model

$$\tilde{m}_k(d) = \max_{i \in \{1,...,p\}} \{\tilde{f}_i(x_k) + \langle g_i(x_k), d\rangle\} + \frac{1}{2}\langle d, H_k d\rangle.$$

Step 2. Find a step $d_k \in B(0, \delta_k)$ such that:

$$\tilde{m}_k(0) - \tilde{m}_k(d_k) \geq \frac{1}{2}\omega_m(x_k) \min\{\delta_k, \frac{\omega_m(x_k)}{\beta_k}\}. \tag{8}$$

Step 3. Compute

$$\rho_k = \frac{\tilde{\phi}(x_k) - \tilde{\phi}(x_k + d_k)}{\tilde{m}_k(0) - \tilde{m}_k(d_k)}$$

If $\rho_k \geq \eta_1$ and $\omega_m(x_k) > \Theta\delta_k$, set $x_{k+1} = x_k + d_k$ and $\delta_{k+1} = \min\{\delta_{\max}, \gamma_2\delta_k\}$.
Else, set $x_{k+1} = x_k$ and $\delta_{k+1} = \gamma_1\delta_k$.

Step 4. Set $k = k + 1$ and go to Step 1.

The following lemma shows that the algorithm is well defined.

Lemma 2. For all k, there exists d_k such that (8) holds.

Proof. We will prove that the condition (8) holds for the Cauchy direction, $d_k^c = \alpha_k d_k^*$, where d_k^* is a solution of the problem stated in (3), i.e.,

$$\omega_m(x_k) = -\min_{\|d\| \leq 1} \left(\max_i \langle g_i(x_k), d\rangle \right) = -\max_i \langle g_i(x_k), d_k^*\rangle$$

and $\alpha_k = argmin_{0 \leq \alpha \leq \delta_k}\{\tilde{m}_k(\alpha d_k^*)\}$. Since $\|d_k^*\| \leq 1$, we have $\alpha_k d_k^* \in B_k(0, \delta_k)$. Notice that

$$\alpha_k = argmin_{0 \leq \alpha \leq \delta_k}\{\tilde{m}_k(\alpha d_k^*)\} = argmax_{0 \leq \alpha \leq \delta_k}\{\tilde{m}_k(0) - \tilde{m}_k(\alpha d_k^*)\}.$$

Next, we lower bound $\tilde{m}_k(0) - \tilde{m}_k(\alpha d_k^*)$ by a quadratic function of α.

$$\begin{aligned}
\tilde{m}_k(0) - \tilde{m}_k(\alpha d_k^*) &= \max_i \tilde{f}_i(x_k) - \max_i\{\tilde{f}_i(x_k) + \langle g_i(x_k), \alpha d_k^*\rangle\} - \frac{1}{2}\alpha^2\langle d_k^*, H_k d_k^*\rangle \\
&\geq -\alpha \max_i \langle g_i(x_k), d_k^*\rangle - \frac{1}{2}\alpha^2\langle d_k^*, H_k d_k^*\rangle \\
&\geq \alpha\omega_m(x_k) - \frac{1}{2}\alpha^2\|d_k^*\|^2 \cdot \|H_k\| \geq \alpha\omega_m(x_k) - \frac{1}{2}\alpha^2\beta_k.
\end{aligned}$$

Thus, we conclude that

$$\tilde{m}_k(0) - \tilde{m}_k(d_k^c) = \max_{0\leq\alpha\leq\delta_k}\{\tilde{m}_k(0) - \tilde{m}_k(\alpha d_k^*)\} \geq \max_{0\leq\alpha\leq\delta_k}\{\alpha\omega_m(x_k) - \frac{1}{2}\alpha^2\beta_k\}. \tag{9}$$

Notice that the solution of the problem at the right-hand side of (9) is given by $\alpha^* = \min\{\frac{\omega_m(x_k)}{\beta_k}, \delta_k\}$. If $\frac{\omega_m(x_k)}{\beta_k} \leq \delta_k$, then we have

$$\max_{0\leq\alpha\leq\delta_k}\{\alpha\omega_m(x_k) - \frac{1}{2}\alpha^2\beta_k\} = \frac{\omega_m(x_k)^2}{\beta_k} - \frac{1}{2}\frac{\omega_m(x_k)^2}{\beta_k^2}\beta_k = \frac{\omega_m(x_k)^2}{2\beta_k}.$$

Else, if $\frac{\omega_m(x_k)}{\beta_k} > \delta_k$, we obtain

$$\begin{aligned}\max_{0\leq\alpha\leq\delta_k}\{\alpha\omega_m(x_k) - \frac{1}{2}\alpha^2\beta_k\} &= \delta_k\omega_m(x_k) - \frac{1}{2}\delta_k^2\beta_k \\ &\geq \delta_k\omega_m(x_k) - \frac{1}{2}\delta_k\omega_m(x_k) \\ &= \frac{1}{2}\delta_k\omega_m(x_k).\end{aligned}$$

Thus, having in mind both cases and using (9) we obtain

$$\tilde{m}_k(0) - \tilde{m}_k(d_k^c) \geq \frac{1}{2}\omega_m(x_k)\min\{\frac{\omega_m(x_k)}{\beta_k}, \delta_k\},$$

which completes the proof.

■

We state here the following two theorem needed for the convergence analysis of the proposed method.

Theorem 1. [24] Let $U_k, \beta_k, \xi_k, \rho_k \geq 0$ be $\mathcal{F}_k$ measurable random variables such that

$$E(U_{k+1}|\mathcal{F}_k) \leq (1+\beta_k)U_k + \xi_k - \rho_k$$

If $\sum\beta_k < \infty$ and $\sum\xi_k < \infty$ then $U_k \to U$ a.s. and $\sum\rho_k < \infty$ a.s.

Theorem 2. [15] Let G_k be a sequence of integrable random variables such that $E(G_k|\mathcal{V}_{k-1}) \geq G_{k-1}$, where $\mathcal{V}_{k-1}$ is a σ-algebra generated by $G_0, ..., G_{k-1}$. Assume further that $|G_k - G_{k-1}| \leq M < \infty$ for every k. Consider the random events $C = \{\lim_{k\to\infty} G_k \text{ exists and is finite}\}$ and $D = \{\limsup_{k\to\infty} G_k = \infty\}$. Then $P(C \cup D) = 1$

4 Convergence analysis

We start this section by stating some of the assumptions needed for the analysis.

Assumption 1. Functions $f_i, i = 1, ..., q$ are twice continuously differentiable and bounded from bellow.

Assumption 2. There exists a positive constant c_h such that for all $x \in \mathbb{R}^n$ and $i = 1, ..., q$ there holds $\|\nabla^2 f_i(x)\| \leq c_h$. Furthermore there exists a positive constant c_b such that $\beta_k = 1 + \|H_k\| \leq c_b$ for every k.

Assumption 3. Approximate functions $\tilde{f}_i, i = 1, ..., q$ are continuously-differentiable with L-Lipschitz continuous gradients satisfying the following inequality $\|\nabla \tilde{f}_i(x_k) - g_i(x_k)\| \leq c_a \delta_k$ with some $c_a > 0$.

Assumption 3 is satisfied in many applications. For instance, subsampling strategies for finite sums usually yield $\nabla \tilde{f}_i(x_k) = g_i(x_k)$. Alternatively, one can apply finite differences to approximate the relevant gradients with a controllable accuracy. The following lemma quantifies the distance between the true and approximate functions on $B(x_k, \delta_k)$.

Lemma 3. Assume that A1-A3 hold. Suppose that $\tilde{m}_k$ is (c_f, c_g)-fully linear model of f. Then there exists $c_e > 0$ such that for all $d_k \in B(0, \delta_k)$ and $i = 1, .., q$ there holds

$$|\tilde{f}_i(x_k) - f_i(x_k)| \leq c_e \delta_k^2 \text{ and } |\tilde{f}_i(x_k + d_k) - f_i(x_k + d_k)| \leq c_e \delta_k^2. \tag{10}$$

Proof. Let us observe an arbitrary $i \in \{1, 2, ..., q\}$. Putting $d = 0$ in (5) and using the fact that $\tilde{m}_{k,i}(0) = \tilde{f}_i(x_k)$ we obtain

$$|\tilde{f}_i(x_k) - f_i(x_k)| \leq c_f \delta_k^2 \tag{11}$$

Now, let us take any $d_k \in B(0, \delta_k)$, i.e., any d_k satisfying $\|d_k\| \leq \delta_k$. Then there exists $\tau_k^i, v_k^i \in In(x_k, x_k + d_k)$ such that

$$\begin{aligned}
&|\tilde{f}_i(x_k + d_k) - f_i(x_k + d_k)| = |\tilde{f}_i(x_k) + \nabla \tilde{f}_i(\tau_k^i) d_k - f_i(x_k + d_k)| \\
&= |\tilde{f}_i(x_k) + \nabla^T \tilde{f}_i(\tau_k^i) d_k - f_i(x_k) - \nabla^T f_i(x_k) d_k - \frac{1}{2} d_k^T \nabla^2 f_i(v_k^i) d_k| \\
&\leq |\tilde{f}_i(x_k) - f_i(x_k)| + \|\nabla \tilde{f}_i(\tau_k^i) - \nabla f_i(x_k)\| \|d_k\| + \frac{1}{2} \|d_k\|^2 \|\nabla^2 f_i(v_k^i)\| \\
&\leq c_f \delta_k^2 + \|\nabla \tilde{f}_i(\tau_k^i) - \nabla f_i(x_k)\| \delta_k + \frac{1}{2} \delta_k^2 c_h.
\end{aligned}$$

Moreover, by using (6) and the fact that $g_i(x_k) = \nabla \tilde{m}_{k,i}(0)$, we can upper bound $\|\nabla \tilde{f}_i(\tau_k^i) - \nabla f_i(x_k)\|$ as follows.

$$\begin{aligned}
&\|\nabla \tilde{f}_i(\tau_k^i) - \nabla f_i(x_k)\| = \|\nabla \tilde{f}_i(\tau_k^i) - \nabla f_i(x_k) + g_i(x_k) - g_i(x_k) + \nabla \tilde{f}_i(x_k) - \nabla \tilde{f}_i(x_k)\| \\
&\leq \|\nabla \tilde{f}_i(\tau_k^i) - \nabla \tilde{f}_i(x_k)\| + \|\nabla f_i(x_k) - g_i(x_k)\| + \|\nabla \tilde{f}_i(x_k) - g_i(x_k)\| \\
&\leq L \|\tau_k^i - x_k\| + c_g \delta_k + c_a \delta_k \leq L \delta_k + c_g \delta_k + c_a \delta_k = (L + c_g + c_a) \delta_k.
\end{aligned}$$

Thus we conclude that

$$|\tilde{f}_i(x_k + d_k) - f_i(x_k + d_k)| \leq \delta_k^2 c_e,$$

where $c_e = c_f + L + c_g + c_a + c_h/2$, which completes the proof. ■

For the purpose of the convergence analysis, let us denote by $J_{k,i}$ the event (10) and define

$$J_k = \bigcap_{j=1}^{q} J_{k,j}. \tag{12}$$

Then, under assumptions A1-A3, according to Lemma 3 there holds $P(J_k|I_k, \mathcal{F}^k) = 1$. Moreover,

$$P(I_k, J_k|\mathcal{F}^k) = P(I_k|\mathcal{F}^k)P(J_k|I_k, \mathcal{F}^k) \geq \alpha_k^q 1 = \alpha_k^q \tag{13}$$

and we can also conclude that $P(I_k, \bar{J}_k|\mathcal{F}^k) = 0$, $P(\bar{I}_k, J_k|\mathcal{F}^k) \leq 1 - \alpha_k^q$ and $P(\bar{I}_k, \bar{J}_k|\mathcal{F}^k) \leq 1 - \alpha_k^q$.

Now we state the conditions under which the difference between the true and the approximate marginal function is of the order δ_k.

Lemma 4. Suppose that A1-A3 hold and that the model $\tilde{m}_k$ is fully linear. Then

$$|\omega(x_k) - \omega_m(x_k)| \leq \delta_k c_g \tag{14}$$

Proof. Let us define $\tilde{h}_k(d) = \max_i \langle g_i(x_k), d\rangle$ and $h_k(d) = \max_i \langle \nabla f_i(x_k), d\rangle$. Furthermore, let d_k^* be the solution of the problem stated in (3), i.e., we have $\|d_k^*\| \leq 1$ and

$$\omega_m(x_k) = -\min_{\|d\|\leq 1} \max_i \langle g_i(x_k), d\rangle = -\min_{\|d\|\leq 1} \tilde{h}_k(d) = -\tilde{h}_k(d_k^*).$$

Similarly, let $d_k^{*,true}$ be the solution of the problem stated in (2). Then we have $\|d_k^{*,true}\| \leq 1$ and

$$\omega(x_k) = -\min_{\|d\|\leq 1} \max_i \langle \nabla f_i(x_k), d\rangle = -\min_{\|d\|\leq 1} h_k(d) = -h_k(d_k^{*,true}).$$

Notice that there holds

$$\max_i \langle \nabla f_i(x_k), d_k^{*,true}\rangle = h_k(d_k^{*,true}) \leq h_k(d_k^*) = \max_i \langle \nabla f_i(x_k), d_k^*\rangle$$

and

$$\max_i \langle g_i(x_k), d_k^*\rangle = \tilde{h}_k(d_k^*) \leq \tilde{h}_k(d_k^{*,true}) = \max_i \langle g_i(x_k), d_k^{*,true}\rangle.$$

Next, depending on the sign of $\omega(x_k) - \omega_m(x_k)$ we analyze separately two cases.

First, suppose that $\omega(x_k) > \omega_m(x_k)$, then

$$\begin{aligned}
|\omega(x_k) - \omega_m(x_k)| &= \max_i \langle g_i(x_k), d_k^*\rangle - \max_i \langle \nabla f_i(x_k), d_k^{*,true}\rangle \\
&= \tilde{h}_k(d_k^*) - h_k(d_k^{*,true}) \\
&\leq \tilde{h}_k(d_k^{*,true}) - h_k(d_k^{*,true}) \\
&= \max_i \langle g_i(x_k), d_k^{*,true}\rangle - \max_i \langle \nabla f_i(x_k), d_k^{*,true}\rangle.
\end{aligned}$$

Let us denote $r_k = argmax_i \langle g_i(x_k), d_k^{*,true} \rangle$. By noticing that $\max_i \langle \nabla f_i(x_k), d_k^{*,true} \rangle \geq \langle \nabla f_{r_k}(x_k), d_k^{*,true} \rangle$ we obtain

$$\begin{aligned} |\omega(x_k) - \omega_m(x_k)| &\leq \langle g_{r_k}(x_k), d_k^{*,true} \rangle - \langle \nabla f_{r_k}(x_k), d_k^{*,true} \rangle \\ &\leq \|g_r(x_k) - \nabla f_r(x_k)\| \cdot \|d_k^{*,true}\| \leq c_g \delta_k, \end{aligned} \tag{15}$$

where the last inequality follows from the FL condition (6).

Consider now the case $\omega(x_k) \leq \omega_m(x_k)$ We have

$$\begin{aligned} |\omega_m(x_k) - \omega(x_k)| &= \max_i \langle \nabla f_i(x_k), d_k^{*,true} \rangle - \max_i \langle g_i(x_k), d_k^* \rangle \\ &= h_k(d_k^{*,true}) - \tilde{h}_k(d_k^*) \\ &\leq h_k(d_k^*) - \tilde{h}_k(d_k^*) \\ &= \max_i \langle \nabla f_i(x_k), d_k^* \rangle - \max_i \langle g_i(x_k), d_k^* \rangle \end{aligned}$$

Denoting $l_k = argmax_i \langle \nabla f_i(x_k), d_k^* \rangle$ and noticing that $\max_i \langle g_i(x_k), d_k^* \rangle \geq \langle g_{l_k}(x_k), d_k^* \rangle$, by employing the FL condition (6) once again we obtain

$$\begin{aligned} |\omega_m(x_k) - \omega(x_k)| &= \langle \nabla f_{l_k}(x_k), d_k^* \rangle - \langle g_{l_k}(x_k), d_k^* \rangle \\ &\leq \|\nabla f_{l_k}(x_k) - g_{l_k}(x_k)\| \cdot \|d_k^*\| \leq c_g \delta_k. \end{aligned}$$

Putting both cases together we get the statement.

■

The following lemma shows that the distance of the model from the approximate scalar representation of f is small enough under the stated assumptions.

Lemma 5. Suppose that A1-A3 hold, the model $\tilde{m}_k$ is fully linear and $d_k \in B(0, \delta_k)$. Then

$$|\phi(x_k + d_k) - \tilde{m}_k(d_k)| \leq c_f \delta_k^2 \tag{16}$$

and

$$|\phi(x_k) - \tilde{m}_k(0)| \leq c_f \delta_k^2. \tag{17}$$

Additionally, if A3 holds, then

$$|\tilde{\phi}(x_k + d_k) - \phi(x_k + d_k)| \leq c_e \delta_k^2 \tag{18}$$

and

$$|\tilde{\phi}(x_k + d_k) - \tilde{m}_k(d_k)| \leq c_{\tilde{\Phi}} \delta_k^2 \tag{19}$$

Proof. The full linearity condition (5) implies that for any $d \in B(0, \delta_k)$ there holds

$$-c_f \delta_k^2 \leq f_i(x_k + d) - \tilde{m}_{k,i}(d) \leq c_f \delta_k^2$$

and therefore

$$-c_f\delta_k^2 + \max_i \tilde{m}_{k,i}(d) \leq \max_i f_i(x_k + d) \leq \max_i \tilde{m}_{k,i}(d) + c_f\delta_k^2.$$

Hence, putting $d = d_k$ and $d = 0$ we obtain (16) and (17), respectively.

If additionally $A3$ holds, then (10) is true. Thus for every i,

$$-c_e\delta_k^2 + f_i(x_k + d_k) \leq \tilde{f}_i(x_k + d_k) \leq f_i(x_k + d_k) + c_e\delta_k^2$$

So, taking the maximum, we get

$$|\tilde{\phi}(x_k + d_k) - \phi(x_k + d_k)| \leq c_e\delta_k^2 \tag{20}$$

Using (16) and (20), we get that

$$|\tilde{\phi}(x_k + d_k) - \tilde{m}_k(d_k)| \leq |\tilde{\phi}(x_k + d_k) - \phi(x_k + d_k)| + |\phi(x_k + d_k) - \tilde{m}_k(d_k)| \leq$$

$$\leq c_e\delta_k^2 + c_f\delta_k^2 =: c_{\tilde{\Phi}}\delta_k^2.$$

■

Lemma 6. Suppose that A1-A3 hold, the model $\tilde{m}_k$ is fully linear and d_k satisfies (8). Then $\rho_k \geq \eta_1$ provided that

$$\delta_k \leq \min\{\frac{\omega_m(x_k)}{c_b}, \frac{\omega_m(x_k)(1-\eta_1)}{2c_{\tilde{\Phi}}}\} \tag{21}$$

Proof. From (8) it follows

$$\begin{aligned}\tilde{m}_k(0) - \tilde{m}_k(d_k) &\geq \frac{1}{2}\omega_m(x_k)\min\{\frac{\omega_m(x_k)}{\beta_k}, \delta_k\} \\ &\geq \frac{1}{2}\omega_m(x_k)\min\{\frac{\omega_m(x_k)}{c_b}, \delta_k\} \\ &= \frac{1}{2}\omega_m(x_k)\delta_k\end{aligned}$$

Furthermore, using $\tilde{\phi}(x_k) = \tilde{m}_k(0)$ and (19), we obtain

$$|\rho_k - 1| = |\frac{\tilde{\phi}(x_k + d_k) - \tilde{\phi}(x_k) - \tilde{m}_k(d_k) + \tilde{m}_k(0)}{\tilde{m}_k(d_k) - \tilde{m}_k(0)}| =$$

$$\leq |\frac{\tilde{\phi}(x_k + d_k) - \tilde{m}_k(d_k)}{\tilde{m}_k(d_k) - \tilde{m}_k(0)}| \leq \frac{2c_{\tilde{\Phi}}\delta_k^2}{\omega_m(x_k)\delta_k} =\leq 1 - \eta_1,$$

and thus we conclude that $\rho_k \geq \eta_1$. ■

To continue with the convergence analysis, let us define an auxiliary Lapynov function as usual in this type of analysis, [12]

$$\psi_k := \nu\phi(x_k) + (1-\nu)\delta_k^2, \quad \nu \in (0,1).$$

We are going to show that we can choose the algorithm parameters such that the following inequality holds

$$E[\psi_{k+1} - \psi_k | \mathcal{F}_k] \leq -\sigma\delta_k^2 + (1-\alpha_k^q)\tilde{\sigma}, \quad k = 0, 1, ... \tag{22}$$

for some $\sigma, \tilde{\sigma} > 0$. Let us denote by S_k the event of successful iteration and the complementary event (unsuccessful iteration) by $\bar{S}_k$. Notice that if the iteration is not successful we have $x_{k+1} = x_k$ and

$$\psi_{k+1} - \psi_k = (1-\nu)(\gamma_1^2 - 1)\delta_k^2 =: -c_1\delta_k^2, \tag{23}$$

for some $c_1 > 0$. Thus, (22) holds in the case of $\bar{S}_k$ and we focus on the case of successful iterations S_k in the following lemmas. In that case we have

$$\psi_{k+1} - \psi_k = \nu(\phi(x_{k+1}) - \phi(x_k)) + (1-\nu)(\gamma_2^2 - 1)\delta_k^2$$

The proof of the following lemma resembles the analysis of [12]. However, having the multi-objective problem requires nontrivial modifications.

Lemma 7. Suppose that A1-A3 hold, the sequence $\{x_k\}$ is bounded and there exists $\bar{\alpha} > 0$ such that $\alpha_k \geq \bar{\alpha}$ for all k. Then there exist positive constants c_6, c_7 such that the following holds for all k

$$E(\psi_{k+1} - \psi_k | \mathcal{F}_k, S_k) \leq -c_6\delta_k^2 + c_7(1-\alpha_k^q), \tag{24}$$

provided that

$$\theta \geq \max\{c_b, 5c_f, \frac{4c_e}{\eta_1}\}. \tag{25}$$

Proof. The analysis will be performed with respect to the events I_k and J_k defined in (7) and (12), conditioned on S_k, i.e., under assumption of successful iterations. First of all notice that if I_k happened, the assumptions A1-A3 imply that Lemma 3 holds and hence J_k is true as well. Thus we have to consider 3 cases overall, $U_k^1 := I_k, U_k^2 := \bar{I}_k \bigcap J_k$ and $U_k^3 := \bar{I}_k \bigcap \bar{J}_k$.

a)U_k^1 happened. Note that in this case the model is fully linear. Moreover, the iteration is successful hence $\omega_m(x_k) \geq \theta\delta_k$ which together with (19), and Lemma 2 implies

$$\begin{aligned}
\phi(x_{k+1}) - \phi(x_k) &= \phi(x_{k+1}) - \tilde{m}_k(d_k) + \tilde{m}_k(0) - \phi(x_k) + \tilde{m}_k(d_k) - \tilde{m}_k(0) \\
&\leq 2c_f\delta_k^2 - \frac{1}{2}w_m(x_k)\min\{\delta_k, \frac{w_m(x_k)}{c_b}\} \\
&\leq 2c_f\delta_k^2 - \frac{1}{2}w_m(x_k)\delta_k \leq 2c_f\delta_k^2 - \frac{1}{2}\Theta\delta_k^2 \qquad (26) \\
&< -\frac{1}{2}c_f\delta_k = -c_1\delta_k^2, \qquad (27)
\end{aligned}$$

for $\Theta \geq \max\{c_b, 5c_f\}$ and $c_1 = \frac{1}{2}c_f > 0$. This further implies

$$\psi_{k+1} - \psi_k = \nu(\phi(x_{k+1}) - \phi(x_k)) + (1-\nu)(\gamma_2^2 - 1)\delta_k^2 \leq [-\nu c_1 + (1-\nu)(\gamma_2^2 - 1)]\delta_k^2,$$

and thus by choosing ν such that

$$\frac{\nu}{1-\nu} \geq \frac{2\gamma_2^2 - 1}{c1}$$

we can obtain

$$\psi_{k+1} - \psi_k \leq -\gamma_2^2 \delta_k^2 = -c_2 \delta_k^2, \tag{28}$$

with $c_2 = \gamma_2^2 > 0$.

b)$U_k^2 := \overline{I}_k \bigcap J_k$ happened. Using (18) and (8), and the fact the step is accepted ($\rho_k \geq \eta_1$) we get

$$\begin{aligned}
\phi(x_{k+1}) - \phi(x_k) &= \phi(x_{k+1}) - \tilde{\phi}(x_{k+1}) + \tilde{\phi}(x_k) - \phi(x_k) + \tilde{\phi}(x_{k+1}) - \tilde{\phi}(x_k) \\
&\leq 2c_e\delta_k^2 + \tilde{\phi}(x_{k+1}) - \tilde{\phi}(x_k) = 2c_e\delta_k^2 - \rho_k(\tilde{m}_k(d_k) - \tilde{m}_k(0)) \\
&\leq 2c_e\delta_k^2 - \eta_1(\tilde{m}_k(d_k) - \tilde{m}_k(0)) \leq 2c_e\delta_k^2 - \frac{\eta_1 \omega_m(x_k)}{2}\min\{\frac{\omega_m(x_k)}{c_b}, \delta_k\} \\
&\leq [2c_e - \frac{\eta_1\Theta}{2}\min\{\frac{\Theta}{c_b}, 1\}]\delta_k^2 = [2c_e - \frac{\eta_1\Theta}{2}]\delta_k^2 \leq -\frac{1}{2}c_e\delta_k^2 = -c_3\delta_k^2.
\end{aligned}$$

for $\Theta \geq \frac{5c_e}{\eta_1}$, and $c_3 = \frac{1}{2}c_e > 0$. Again, for ν such that

$$\frac{\nu}{1-\nu} \geq \frac{2\gamma_2^2 - 1}{c_3}$$

we get that

$$\psi_{k+1} - \psi_k \leq [-\nu c_3 + (1-\nu)(\gamma_2^2 - 1)]\delta_k^2 \leq -\gamma_2^2\delta_k^2 = -c_4\delta_k^2. \tag{29}$$

for $c_4 = \gamma_2^2 > 0$

c)$U_k^3 := \overline{I}_k \bigcap \overline{J}_k$ happened. Since the iteration is successful, we have $\omega_m(x_k) \geq \Theta\delta_k$, but an increase of the function Ψ can happen. However, using Taylor expansion, A2, and the Cauchy Schwartz inequality, we can bound the increase as follows.

$$\begin{aligned}
\phi(x_{k+1}) - \phi(x_k) &= \max_i f_i(x_{k+1}) - \max_i f_i(x_k) \\
&= \max_i\{f_i(x_k) + \nabla^T f_i(x_k)d_k + \frac{1}{2}d_k^T\nabla^2 f_i(\tau_k)d_k\} - \max_i f_i(x_k) \\
&\leq \max_i \nabla^T f_i(x_k)d_k + \frac{1}{2}\delta_k^2 c_h.
\end{aligned}$$

Since the iterates are assumed to be bounded, the continuity of the gradients implies the existence of $G > 0$ such that $\max_i \|\nabla f_i(x_k)\| \leq G$. Since $\delta_k \leq \delta_{max}$ there exists a constant c_5 such that

$$\phi(x_{k+1}) - \phi(x_k) \leq G\delta_{max} + \frac{1}{2}\delta_{max}^2 c_h = c_5 \tag{30}$$

and thus

$$\Psi_{k+1} - \Psi_k \leq c_5 + (1-\nu)(\gamma_2^2 - 1)\delta_k^2. \tag{31}$$

Now, we combine inequalities (28),(29) and (31) to estimate $E(\Psi_{k+1} - \Psi_k|\mathcal{F}_k, S_k)$. Using the total probability formula we obtain

$$\begin{aligned} & E(\Psi_{k+1} - \Psi_k|\mathcal{F}_k, S_k) \qquad (32)\\ = & \sum_{i=1}^{3} P(U_k^i|\mathcal{F}_k, S_k) E(\Psi_{k+1} - \Psi_k|\mathcal{F}_k, S_k, U_k^i) \\ \leq & P(U_k^1|\mathcal{F}_k, S_k) E(\Psi_{k+1} - \Psi_k|\mathcal{F}_k, S_k, U_k^1) \\ + & P(U_k^3|\mathcal{F}_k, S_k) E(\Psi_{k+1} - \Psi_k|\mathcal{F}_k, S_k, U_k^4), \end{aligned}$$

where the last inequality follows from the fact that $E(\Psi_{k+1} - \Psi_k|\mathcal{F}_k, S_k, U_k^2) \leq -c_3\delta_k^2 < 0$. Moreover, notice that (28) implies $E(\Psi_{k+1} - \Psi_k|\mathcal{F}_k, S_k, U_k^1) \leq -c_2\delta_k^2 < 0$ and that the conditional expectation $E(\Psi_{k+1} - \Psi_k|\mathcal{F}_k, S_k, U_k^3)$ is upper bounded by the positive quantity given in (31). Thus, by (13) we obtain

$$E(\Psi_{k+1} - \Psi_k|\mathcal{F}_k, S_k) \leq -\alpha_k^q c_2 \delta_k^2 + (1 - \alpha_k^q)(c_5 + (1-\nu)(\gamma_2^2 - 1)\delta_k^2)$$

and the result follows with $c_6 = \bar{\alpha}^q c_2$ and $c_7 = c_5 + (1-\nu)(\gamma_2^2 - 1)\delta_{\max}$ due to $\alpha_k \geq \bar{\alpha}$ and $\delta_k \leq \delta_{max}$.

■

Remark 1. The choice of ν for Lyapunov function can be formulated as

$$\frac{\nu}{1-\nu} \geq \frac{4\gamma_2^2 - 2}{\min\{c_e, c_f\}},$$

which means that we can always find $\nu \in (0, 1)$ large enough, i.e., close enough to 1, so that the proof holds.

Now we show that the sequence of trust region radii is square sumable under the folllowing assumption.

Assumption 4. The sequence $\{\alpha_k\}_k$ satisfies $\sum_{k=0}^{\infty}(1 - \alpha_k^q) \leq c_\alpha < \infty$.

Theorem 3. Suppose that A1-A4, (25) hold and that the sequence $\{x_k\}$ is bounded. Then the sequence $\{\Psi_k\}_k$ converges a.s. and there holds

$$\sum_{k=0}^{\infty} \delta_k^2 < \infty \quad \text{a.s.} \tag{33}$$

Proof. Assumption 4 implies that $\lim_{k\to\infty} \alpha_k = 1$, so without loss of generality we can assume that $\alpha_k \geq \bar{\alpha} > 0$ for all k. Then, according to (23) and Lemma 7 we obtain

$$\begin{aligned} & E(\Psi_{k+1} - \Psi_k|\mathcal{F}_k) \\ = \; & E(\Psi_{k+1} - \Psi_k|\mathcal{F}_k, S_k)P(S_k|\mathcal{F}_k) + E(\Psi_{k+1} - \Psi_k|\mathcal{F}_k, \bar{S}_k)P(\bar{S}_k|\mathcal{F}_k) \\ \leq \; & (-c_6\delta_k^2 + c_7(1-\alpha_k^q))P(S_k|\mathcal{F}_k) - c_1\delta_k^2 P(\bar{S}_k|\mathcal{F}_k) \leq \\ \leq \; & -\min\{c_1, c_6\}(P(S_k|\mathcal{F}_k) + P(\bar{S}_k|\mathcal{F}_k))\delta_k^2 + c_7(1-\alpha_k^q) \\ =: \; & -c_8\delta_k^2 + c_7(1-\alpha_k^q) \end{aligned} \tag{34}$$

Since Ψ_k is bounded from bellow by Ψ^*, by adding and subtracting Ψ^* in the conditional expectation above and using the fact that Ψ_k is $\mathcal{F}_k$-measurable we obtain

$$E(\Psi_{k+1} - \Psi^*|\mathcal{F}_k) \leq \Psi_k - \Psi^* - c_8\delta_k^2 + c_7(1-\alpha_k^q)$$

and the result follows from Theorem 1. ■

Now we show that under the states conditions a.s. there exists an infinite sequence of iterations with fully linear models.

Theorem 4. Suppose that the assumptions of Theorem 3 hold. Then a.s. there exists an infinite $K \subseteq \mathbb{N}$ such that the model $\tilde{m}_k$ is fully linear for all $k \in K$.

Proof. Notice that assumption A4 implies the existence of $\bar{k}$ such that $\alpha_k^q > 0.5$ for all $k \geq \bar{k}$. Let us define a random variable

$$W_k = \sum_{s=\bar{k}}^{k} V_s, \tag{35}$$

where $V_k = 1$ if I_k happens and $V_k = -1$ otherwise. Moreover,

$$\begin{aligned} E(V_{k+1}|\mathcal{F}_k) &= P(I_k|\mathcal{F}_k) - P(\bar{I}_k|\mathcal{F}_k) = P(I_k|\mathcal{F}_k) - (1 - P(I_k|\mathcal{F}_k)) \\ &= 2P(I_k|\mathcal{F}_k) - 1 \geq 2\alpha_k^q - 1 > 0. \end{aligned} \tag{36}$$

This implies $E(W_{k+1}|\mathcal{F}_k) = W_k + E(V_{k+1}|\mathcal{F}_k) > W_k$. We also have $|W_{k+1} - W_k| = |V_{k+1}| = 1$ and thus the conditions of Theorem 2 are satisfied with $G_k = W_k$ and $M = 1$. Moreover, $|W_{k+1} - W_k| = 1$ also indicates that the sequence of W_k cannot be convergent and thus Theorem 2 implies that a.s.

$$\limsup_{k\to\infty} W_k = \infty. \tag{37}$$

The statement to be proved is that I_k happens infinitely many times a.s. Assume that this is not true. Then, there exists $\tilde{k}$ such that for each $k \geq \tilde{k}$ the event $\bar{I}_k$ happens so $V_k = -1$. As $W_k = W_{\tilde{k}} + (k - \tilde{k})V_k$ we get $\lim_{k\to\infty} W_k = -\infty$, which is in contradiction with (37).

■

Theorem 5. Suppose that the assumptions of Theorem 3 hold. Then a.s.

$$\liminf_{k\to\infty} \omega(X_k) = 0$$

Proof. Let us denote by Ω all the sample paths of SMOP algorithm. Suppose the contrary, that with positive probability none of the subsequences converges to 0. In other words there exists $\hat{\Omega} \subset \Omega$ such that $P(\hat{\Omega}) > 0$ and $\omega(x_k(v))$ is bounded away from zero for all $v \in \hat{\Omega}$. Let us observe an arbitrary $v \in \hat{\Omega}$ and the corresponding $\omega(x_k) = \omega(x_k(v))$. We know that there exists $\epsilon > 0$ and $k_1 = k_1(v)$, such that $\omega(x_k) \geq \epsilon > 0,\ k \geq k_1$. Moreover,Theorem 4 implies that there exists $K = K(v) \subseteq N$ such that for all $k \in K$ model $m_k = m_k(v)$ is fully linear. We denote by $\overline{\Omega}$ the subset of $\hat{\Omega}$ for which there exists such K. Notice that $P(\overline{\Omega}) > 0$. Further, let us observe an arbitrary $v \in \overline{\Omega}$. We will omit writing v further on for the sake of simplicity. Since δ_k tends to 0 a.s., without loss of generality we can assume that $\lim_{k\to\infty} \delta_k = 0$ hence there exists k_2 such that for $k \geq k_2$,

$$\delta_k < b := \min\{\frac{\epsilon}{2c_g}, \frac{\epsilon}{2\Theta}, \frac{\epsilon}{2c_b}, \frac{\epsilon(1-\eta_1)}{4c_{\tilde{\Phi}}}\} \tag{38}$$

Let us denote by $\hat{K}$ the set of all indices from K such that $k \geq k_3 = \max\{k_1, k_2\}$. Thus, for all $k \in \hat{K}$ there holds that m_k is fully linear, $\omega(x_k) \geq \epsilon$ and δ_k is small enough. From Lemma 4 and (38),

$$|\omega(x_k) - \omega_m(x_k)| \leq c_g\delta_k \leq \frac{\epsilon}{2}$$

hence $\omega_m(x_k) \geq \frac{\epsilon}{2} \geq \delta_k\Theta$ (once again from (38)). For these $k \in \hat{K}$, the condition from Lemma 6 is satisfied

$$\delta_k < \min\{\frac{\epsilon}{2c_g}, \frac{\epsilon}{2\Theta}, \frac{\epsilon}{2c_b}, \frac{\epsilon(1-\eta_1)}{4c_{\tilde{\Phi}}}\} \leq \min\{\frac{\omega_m(x_k)}{c_b}, \frac{\omega_m(x_k)(1-\eta_1)}{2c_{\tilde{\Phi}}}\}$$

thus we have that $\rho_k \geq \eta_1$ which together with $\omega_m(x_k) \geq \delta_k\Theta$ implies that such iterations are successful. Therefore for all $k \in \hat{K}$ there holds $\delta_{k+1} = \delta_k\gamma_2 > \delta_k$. Let us define

$$r_k := \log_{\gamma_2}(b^{-1}\delta_k)$$

where b is defined in (38). Notice that for $k \geq k_3, \delta_k < b$, hence $\gamma_2^{r_k} < 1$ and $r_k < 0$. Moreover,

$$r_{k+1} = \log_{\gamma_2}(b^{-1}\delta_{k+1}) = \begin{cases} r_k + 1 & if \quad \delta_{k+1} = \gamma_2\delta_k \\ r_k - 1 & if \quad \delta_{k+1} = \frac{\delta_k}{\gamma_2} \end{cases}$$

Since all iterations $k \in \hat{K}$ are successful, we have $\delta_{k+1} = \gamma_2\delta_k$ and thus $r_{k+1} = r_k + 1$ for all $k \in \hat{K}$. The increase of r_k can also happen even if k is

not in $\hat{K}$, i.e. if k is large enough and m_k is not fully linear. However, for W_k as in (35), we know that it only increases if m_k is fully linear. Thus for all $k > k_3, r_k$ has increased at least as many times as W_k, hence we get

$$r_k - r_{k_3} \geq W_k - W_{k_3}.$$

According to (37) we conclude that $\limsup_{k\to\infty} r_k = \infty$ which contradicts the fact that $r_k < 0$ for all $k \geq k_3$. ■

Proposition 1. Suppose that the assumptions of Theorem 5 hold. If there exists an infinite subsequence $K \subseteq \mathbb{N}$ such that $\omega(x_k) \geq \varepsilon > 0$ for all $k \in K$ then there holds

$$E(\sum_{k\in K} \delta_k) < \infty.$$

Moreover, $\sum_{k\in K} \delta_k < \infty$ a. s.

Proof. Let us observe iterations $k \in K$. We distinguish two scenarios: I_k and $\bar{I}_k$.

If I_k happens, then the model is fully linear. Moreover, we know that $\lim_{k\to\infty} \delta_k = 0$ a.s. according to (33) which together with Lemma 4 implies the existence of $\tilde{\varepsilon} > 0$ such that $\omega_m(x_k) \geq \tilde{\varepsilon}$ for each $k \in K$ sufficiently large. The above further implies that $\omega_m(x_k) \geq \Theta\delta_k > c_b\delta_k$ for each $k \in K$ sufficiently large, and thus, due to Lemma 6, $\rho_k \geq \eta_1$ for each $k \in K$ sufficiently large. Without loss of generality, let us assume that K contains only those sufficiently large iterations such that all the above holds. Then, for each $k \in K$ we conclude that the iteration is successful if I_k happens. Thus, due to (26), for each $k \in K$ there holds

$$\begin{aligned} \phi(x_{k+1}) - \phi(x_k) &\leq 2c_f\delta_k^2 - \frac{1}{2}w_m(x_k)\min\{\delta_k, \frac{w_m(x_k)}{c_b}\} \\ &\leq 2c_f\delta_k^2 - \frac{1}{2}\tilde{\varepsilon}\delta_k = -\delta_k(\frac{\tilde{\varepsilon}}{2} - 2c_f\delta_k). \end{aligned} \tag{39}$$

Once again, assuming that $k \in K$ are all sufficiently large, we obtain $\phi(x_{k+1}) - \phi(x_k) \leq -c_9\delta_k$, where $c_9 = \frac{\tilde{\varepsilon}}{2} - 2c_f\delta_k > 0$, and thus we conclude

$$\psi_{k+1} - \psi_k = \nu(\phi(x_{k+1}) - \phi(x_k)) + (1-\nu)(\gamma_2^2 - 1)\delta_k^2 \leq -c_{10}\delta_k + c_{11}\delta_k^2,$$

where $c_{10} = \nu c_9 > 0$, and $c_{11} = (1-\nu)(\gamma_2^2 - 1) > 0$, i.e. for all $k \in K$ there holds

$$E(\Psi_{k+1} - \Psi_k | I_k, \mathcal{F}_k) \leq -c_{10}\delta_k + c_{11}\delta_k^2. \tag{40}$$

In the case of $\bar{I}_k$, considering (29) and (31) we conclude that

$$\phi(x_{k+1}) - \phi(x_k) \leq c_5 \tag{41}$$

and thus

$$E(\Psi_{k+1} - \Psi_k | \bar{I}_k, \mathcal{F}_k) \leq c_5 + (1-\nu)(\gamma_2^2 - 1)\delta_k^2 = c_5 + c_{11}\delta_k^2. \tag{42}$$

Now, combining both cases regarding I_k we conclude that for all $k \in K$ there holds

$$\begin{aligned} E(\Psi_{k+1} - \Psi_k | \mathcal{F}_k) &= P(I_k|\mathcal{F}_k) E(\Psi_{k+1} - \Psi_k | I_k, \mathcal{F}_k) \\ &+ P(\bar{I}_k|\mathcal{F}_k) E(\Psi_{k+1} - \Psi_k | \bar{I}_k, \mathcal{F}_k) \\ &\leq P(I_k|\mathcal{F}_k)(-c_{10}\delta_k + c_{11}\delta_k^2) \\ &+ P(\bar{I}_k|\mathcal{F}_k)(c_5 + c_{11}\delta_k^2) \\ &\leq -\bar{\alpha}^q c_{10}\delta_k + c_{11}\delta_k^2 + c_5(1-\alpha_k^q) \\ &=: -c_{12}\delta_k + c_{11}\delta_k^2 + c_5(1-\alpha_k^q). \end{aligned} \tag{43}$$

where $c_{12} = \bar{\alpha}^q c_{10} > 0$. Applying the expectation we conclude that for all $k \in K$ there holds

$$E(\Psi_{k+1} - \Psi_k) \leq -c_{12}E(\delta_k) + c_{11}E(\delta_k^2) + c_5(1-\alpha_k^q). \tag{44}$$

On the other hand, (34) holds in all the iterations $k \in \mathbb{N}$ and applying the expectation we obtain

$$E(\Psi_{k+1} - \Psi_k) \leq -c_8 E(\delta_k^2) + c_7(1-\alpha_k^q) \leq c_7(1-\alpha_k^q). \tag{45}$$

Let us denote $\{k\}_{k\in K} = \{k_{(j)}\}_{j\in\mathbb{N}}$. Then, for each $j \in \mathbb{N}$ there holds

$$\begin{aligned} E(\Psi_{k_{(j+1)}} - \Psi_{k_{(j)}}) &= E(\Psi_{k_{(j)}+1} - \Psi_{k_{(j)}}) + \sum_{i=k_{(j)+1}}^{k_{(j+1)}-1} E(\Psi_{i+1} - \Psi_i) \\ &\leq -c_{12}E(\delta_{k_{(j)}}) + c_{11}E(\delta_{k_{(j)}}^2) + c_5(1-\alpha_{k_{(j)}}^q) \\ &+ c_7 \sum_{i=k_{(j)+1}}^{k_{(j+1)}-1} (1-\alpha_i^q) \\ &\leq -c_{12}E(\delta_{k_{(j)}}) + c_{11}E(\delta_{k_{(j)}}^2) \\ &+ c_{13} \sum_{i=k_{(j)}}^{k_{(j+1)}-1} (1-\alpha_i^q), \end{aligned} \tag{46}$$

where $c_{13} = \max\{c_5, c_7\}$. Therefore, for every $m \in \mathbb{N}$ there holds

$$E(\Psi_{k_{(m)}} - \Psi_{k_{(0)}}) \leq -c_{12}E(\sum_{j=0}^{m-1} \delta_{k_{(j)}}) + c_{11}E(\sum_{k=0}^{\infty} \delta_k^2) + c_{13}c_\alpha. \tag{47}$$

Letting m tend to infinity and using (33) together with the assumption of Ψ being bounded from below, we conclude that

$$E(\sum_{k\in K}\delta_k) = E(\sum_{j=0}^{\infty}\delta_{k_{(j)}}) < \infty.$$

Finally, assuming that $\sum_{k\in K}\delta_k = \infty$ with some positive probability yields the contradiction with the previous inequality and we conclude that $\sum_{k\in K}\delta_k < \infty$ a.s, which completes the proof. ■

Theorem 6. Suppose that the assumptions of Theorem 5 hold. Then a.s.

$$\lim_{k\to\infty}\omega(x_k) = 0.$$

Proof. Suppose the contrary, that with positive probability there exists a subsequence $\omega(x_k)$ which does not converge to zero. More precisely, there exists a subset of all possible outcomes of the algorithm $\hat{\Omega}\subset\Omega$ such that $P(\hat{\Omega}) > 0$ and for all $v\in\hat{\Omega}$ there exist $\epsilon > 0$ and $K\subseteq\mathbb{N}$, both dependent on sample path v, such that for all $k\in K$ there holds

$$\omega(x_k)\geq 2\epsilon.$$

On the other hand, Theorem 5 implies the existence of $K_l\subset\mathbb{N}$ such that $\lim_{k\in K_l}\omega(x_k) = 0$ for almost every $v\in\Omega$. Therefore, without loss of generality, we assume that $\omega(x_k) < \epsilon$ for all $k\in K_l$. Since both K and K_l are infinite, there exists $K_s\subseteq K_l$ such that for each $k\in K_s$ we have both

$$\omega(x_k) < \epsilon \quad \text{and} \quad \omega(x_{k+1})\geq\epsilon.$$

In other words, we observe the subsequence K_s of K_l such that $k\in K_l$ and the subsequent iteration does not belong to K_l, i.e., $k+1\notin K_l$. Furthermore, let us observe the pairs $(k_{j,1}, k_{j,2}), j = 1, 2, ...$, where $k_{j,1}\in K_s$ and $k_{j,2}$ is the first $k > k_{j,1}$ that belongs to K, i.e.,

$$\omega(x_{k_{j,1}}) < \epsilon \quad \text{and} \quad \omega(x_{k_{j,2}})\geq 2\epsilon, \quad j = 1, 2, ...$$

This also implies that for each $j\in\mathbb{N}$ there holds

$$|\omega(x_{k_{j,1}}) - \omega(x_{k_{j,2}})|\geq\epsilon. \tag{48}$$

Notice that, by the construction of the relevant subsequences, $k_{j,1}$ represents the last iteration prior to $k_{j,2}$ such that $\omega(x_{k_{j,1}}) < \epsilon$. Therefore, if $k_{j,2}\neq k_{j,1}+1$, for all the intermediate iterations $k\in\{k_{j,1}+1, ..., k_{j,2}-1\}$ and all $j\in\mathbb{N}$ there holds

$$\omega(x_k)\geq\epsilon.$$

Moreover, Proposition 1 implies that a.s.

$$\sum_{j=1}^{\infty} \sum_{i=k_{j,1}+1}^{k_{j,2}-1} \delta_i < \infty. \tag{49}$$

Notice that $k_{j,1}$ must be a successful iteration for all $j \in \mathbb{N}$, since the marginal function changes only when the step is accepted, and thus $\delta_{k_{j,1}+1} = \gamma_2 \delta_{k_{j,1}} > \delta_{k_{j,1}}$. Therefore, for all $j \in \mathbb{N}$, we have

$$\begin{aligned} \|x_{k_{j,1}} - x_{k_{j,2}}\| &= \|x_{k_{j,1}} - x_{k_{j,1}+1} + x_{k_{j,1}+1} - \ldots - x_{k_{j,2}}\| \leq \sum_{i=k_{j,1}}^{k_{j,2}-1} \|x_i - x_{i+1}\| \quad (50) \\ &\leq \sum_{i=k_{j,1}}^{k_{j,2}-1} \delta_i = \delta_{k_{j,1}} + \sum_{i=k_{j,1}+1}^{k_{j,2}-1} \delta_i \leq \delta_{k_{j,1}+1} + \sum_{i=k_{j,1}+1}^{k_{j,2}-1} \delta_i \leq 2 \sum_{i=k_{j,1}+1}^{k_{j,2}-1} \delta_i. \end{aligned}$$

Thus, summing over j we conclude that a.s. $\sum_{j=1}^{\infty} \|x_{k_{j,1}} - x_{k_{j,2}}\| < \infty$ due to (49). This further implies that $\lim_{j\to\infty} \|x_{k_{j,1}} - x_{k_{j,2}}\| = 0$ a.s However, this further implies that a.s. $\lim_{j\to\infty} |\omega(x_{k_{j,1}}) - \omega(x_{k_{j,2}})| = 0$ due to Lemma 1, d), which is a contradiction with (48). ■

5 Numerical results

5.1 Experiment overview

Several experiments are reported in this paper in order to demonstrate the efficiency of the SMOP algorithm. The first experiment utilizes benchmark test problems from [28]. We employ SMOP by adding noise to both function and gradient values, thus simulating noisy conditions. Two multi-objective problems with different properties are considered. Simulations are made for different variance of the noise, which showcases the behavior of SMOP under varying inaccuracy.

The second set of experiments focuses on the machine learning application, and the concept of model fairness, as discussed in [21]. The notion is to create a fair model which makes unbiased decisions, an important feature when the prediction outcome affects individuals or groups of people. The goal is to prevent discrimination based on particularly sensitive attributes. In our tests, the problem of minimizing the logistic regression loss function is reformulated into a multi-objective optimization problem by splitting the dataset based on such sensitive features. This setup allows the analysis of fairness of model predictions by treating each subgroup as a separate objective. We compare the performance of several algorithms using appropriate metrics. The experiments emphasize the benefits of the stochastic models approach, and the algorithm's ability to navigate to the solution with

less information and lower costs. Additionally, we show the effectiveness of utilizing second order information within the algorithm.

The third and final set of experiments focuses on the visualization and characterization of the Pareto front for the first two experiments. The SMOP algorithm is integrated into a standard procedure, which approximates the Pareto front and closely identifies it through iterations. We provide a comprehensive representation of the Pareto front for both convex and nonconvex case.

5.2 Test problems with added noise

Test 1. Firstly, we go over a function with a convex Pareto front, [28]. Let the problem be defined as follows:

$$\min_{x\in\mathbb{R}^2} f(x) = \min_{x\in\mathbb{R}^2}(f_1(x), f_2(x)) \tag{51}$$

where,

$$f_1(x) = x_1^2 + x_2^2$$
$$f_2(x) = (x_1 - 5)^2 + (x_2 - 5)^2.$$

To achieve randomness and emulate our assumptions, we add the noise to function and gradient values in the following way:

$$\tilde{f}_i(x_k) = f_i(x_k) + \varepsilon_i \delta_k^2$$
$$g_i(x_k) = \nabla f_i(x_k) + \varepsilon \delta_k$$

where $\varepsilon = [\varepsilon_1, \varepsilon_2]^T$, $\varepsilon_i : N(0, \sigma^2)$, $i = 1, 2$. Noises are scaled respectively with the trust region radius δ_k, which emulates the probabilistic full linearity in a way. We posed 3 different scenarios, in which we changed the variance of the noise. Each of the runs had 10 independent simulations, with fixed 500 iterations.

The following figures show how modifying the variance of the noise 'increases' randomness, and changes the location of the last iteration, however it doesn't change the fact that the algorithm finds a Pareto critical point. By increasing the variance of the noise, we can notice that last iterations of simulations disperse throughout the set of Pareto criticality, which for this problem is on the Pareto front.

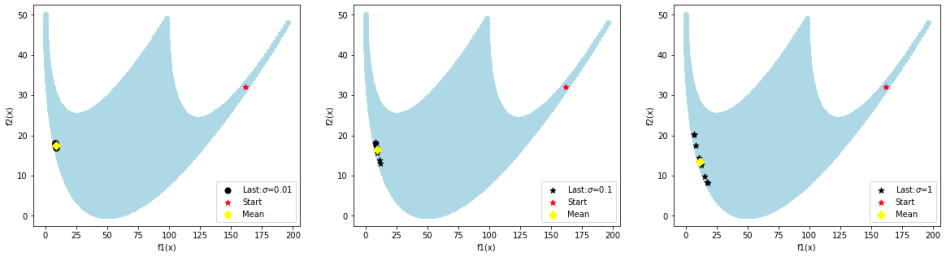

Figure 1: Pareto front for problem (51) for different levels of noise $\sigma \in \{0.01, 0.1, 1\}$. "Start" indicates function values for starting point $f(x_0)$. "Last" indicates all values $f(x^*)$ obtained in the last iterations x^* for 10 simulations. "Mean" indicates $f(x^*_{mean})$, where x^*_{mean} is the mean of last iterations of simulations. Model parameters are: $x_0 = (9, 9), k_{max} = 500, \theta = 10^{-4}, \gamma_1 = 0.5, \gamma_2 = 2, \eta_1 = 10^{-4}$.

Test 2. The second example involves two functions, which together generate a non convex front, [28]. We are solving:

$$\min_{x \in \mathbb{R}^2} f(x) = \min_{x \in \mathbb{R}^2} (f_1(x), f_2(x)) \tag{52}$$

where,

$$f_1(x) = \sin x_2$$
$$f_2(x) = 1 - e^{(-(x_1 - \frac{1}{2})^2 - (x_2 - \frac{1}{2})^2)}.$$

As in the first example the noise is generated in the following way:

$$\tilde{f}_i(x_k) = f_i(x_k) + \varepsilon \delta_k^2$$
$$g_i(x_k) = \nabla f_i(x_k) + \varepsilon \delta_k$$

where $\varepsilon = [\varepsilon_1, \varepsilon_2]^T, \varepsilon_i : N(0, \sigma^2)$, $i = 1, 2$. Three separate experiments have been conducted, with 10 simulations each and 500 iterations. It can be seen that even though the standard deviation increases the SMOP algorithm successfully identifies a Pareto critical point for problems with both convex and non convex front.

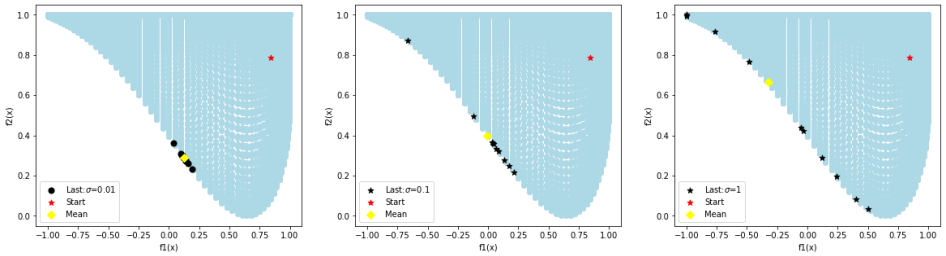

Figure 2: Pareto front for problem (52) for different levels of noise $\sigma \in \{0.01, 0.1, 1\}$. "Start" indicates function values for starting point $f(x_0)$. "Last" indicates all values $f(x^*)$ obtained in the last iterations x^* for 10 simulations. "Mean" indicates $f(x^*_{mean})$, where x^*_{mean} is the mean of last iterations of simulations. Model parameters: $x_0 = (-0.5, 1), k_{max} = 500, \theta = 0.4, \gamma_1 = 0.5, \gamma_2 = 2, \eta_1 = 0.4$.

5.3 Machine learning (logistic regression)

When handling sensitive data in machine learning, there is a significant risk of developing models that exhibit discriminatory behavior. Unfairness emerges when the performance of a model measured in terms of accuracy or another metric, varies across subgroups of data that are categorized based on sensitive attributes, such as race, gender or age [21]. For more on this topic, see [2],[18], [31],[32],[33]. Such differences can have real-world consequences, particularly in applications where the subgroups represent actual individuals, such as in hiring systems, healthcare diagnostics, or judicial decision-making. These biases can create systemic inequalities and harm marginalized communities.

Achieving fairness in machine learning involves ensuring that the predictive performance of a model is consistent across different subgroups. Hence, fairness can be defined as the condition where no subgroup experiences significantly worse outcomes. To address this issue, one effective approach is to formulate the scalar optimization problem as a multi-objective one. By splitting the data based on the sensitive attributes and treating the performance on each subgroup as a separate objective, it is possible to train models that find more balanced optimum values.

The problem we consider is optimization of a regularized logistic regression loss function, as in [21]:

$$\min_x f(x) := \frac{1}{N} \sum_{j \in N} \log(1 + e^{(-y_j(x^T a_j))}) + \frac{\lambda}{2} \|\hat{x}\|^2, i = 1, 2. \tag{53}$$

where x represents model coefficients we are trying to find, $\hat{x}$ coefficient vector without the intercept, a_j the feature vector of $j - th$ sample, y_j its respective label and N the training set size.

In order to create a multi-objective problem, we choose a feature, split the data with respect to it, and create a function for each subgroup. For the sake of simplicity, for each dataset, two subgroups were made. The loss functions for such problem are:

$$f_i(x) = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \log(1 + e^{(-y_j(x^T a_j))}) + \frac{\lambda_i}{2}\|\hat{x}\|^2, i = 1, 2.$$

where $|\mathcal{N}_i|$ is the size of the $i-th$ sample subgroup, and hence the problem becomes:

$$\min_x (f_1(x), f_2(x)) \tag{54}$$

The approximation of function and gradient values is done by an adaptive subsampling strategy $\mathcal{N}_i^k \subseteq \mathcal{N}_i$, $i = 1, 2$ motivated by the result from [25] (Lemma 4). Namely, for each subgroup we get $P(|f_i(x_k) - \tilde{f}_i(x_k)| \leq \delta_k^2) \geq \alpha_k$ provided that

$$|\mathcal{N}_i^k| \geq \frac{F_i(x_k)^2}{\delta_k^4} \left(1 + \sqrt{8 \log(\frac{1}{1-\alpha_k})}\right)^2 \tag{55}$$

where $F_i(x_k)$ is the upper bound of $|f_i(x_k)|$. Although this kind of bound is not easy to obtain in general, for logistic regression problems it is possible to use e.g.

$$F_i(x_k) = e^{\|x_k\|} \max_j \|a_j\| + \log(2) + \frac{\lambda_i}{2}\|x_k\|^2.$$

Similar bound can be derived for the gradients where we get $P(\|\nabla f_i(x_k) - \nabla \tilde{f}_i(x_k)\|| \leq \delta_k) \geq \alpha_k$ provided that

$$|\mathcal{N}_i^k| \geq \frac{G_i(x_k)^2}{\delta_k^2} \left(1 + \sqrt{8 \log(\frac{1}{1-\alpha_k})}\right)^2 \tag{56}$$

with

$$G_i(x_k) = \max_j \|a_j\| + \lambda_i \|x_k\|.$$

In our tests, we use only estimated bounds where the upper-bounds F_i, G_i are replaced by some constants and the sample size behaves like

$$\frac{1}{\delta_k^4} \left(1 + \sqrt{8 \log(\frac{1}{1-\alpha_k})}\right)^2.$$

Approximate functions are created in the following way:

$$\tilde{f}_i(x_k) = \frac{1}{|\mathcal{N}_i^k|} \sum_{j \in \mathcal{N}_i^k} \log(1 + e^{(-y_j(x^T a_j))}) + \frac{\lambda_i}{2}\|\hat{x}\|^2, i = 1, 2.$$

Gradients of the approximate functions can then be used to approximate gradients of the true functions, i.e. $g_i(x_k) = \nabla \tilde{f}_i(x_k)$, hence Assumption 3

is satisfied. The probabilities α_k should theoretically converge to 1, however choosing $\alpha_k = \sqrt{0.5}$ in our implementations demonstrated strong performance. This way, the probability that both function approximations $\tilde{f}_1(x)$ and $\tilde{f}_2(x)$ are δ_k^2-close to $f_1(x)$ and $f_2(x)$ respectively is greater than 0.5. Using these functions we create probabilistically FL models with high enough probability. This has theoretical implications, as seen in the proof of Theorem (4).

We compared the SMOP algorithm with the deterministic trust region [30] (DMOP), and the stochastic multi-gradient [21] (SMG). The evaluation focuses on the algorithms' performance by measuring the value of true marginal function $\omega(x_k)$, and scalar representation $\phi(x_k)$ in terms of number of scalar products at each iteration. It is important to notice that for each function evaluation of $f(x_k) = (f_1(x_k), f_2(x_k))$, we have $|\mathcal{N}_1^k| + |\mathcal{N}_2^k|$ scalar products. In this manner, we demonstrate how the SMOP algorithm achieves significant improvements in performance while utilizing only a fraction of the available data. This highlights the algorithms effectiveness in leveraging limited resources, making it particularly valuable in scenarios where data collection is expensive.

The comparison for covtype [7] dataset is shown, which has 450000 training size, and attribute dimension 54. The data is split in a way that $|N_1| = 200000$ and $|N_2| = 250000$. Due to the stochasticity of the algorithms and the nature of the multiobjective problem, the resulting Pareto critical points can be different. This implies that the values of $\phi(x^*)$ will not necessarily be the same for different algorithms, or even for subsequent simulations of the same algorithm. However, the fact that $\omega(x_k)$ converges to zero shows that the algorithms finds a critical value. The efficiency of the algorithm is emphasized when considering the number of scalar products necessary to reach the near-optimality threshold, as seen in Figure3. Notice how the resulting ϕ values are different for each algorithm.

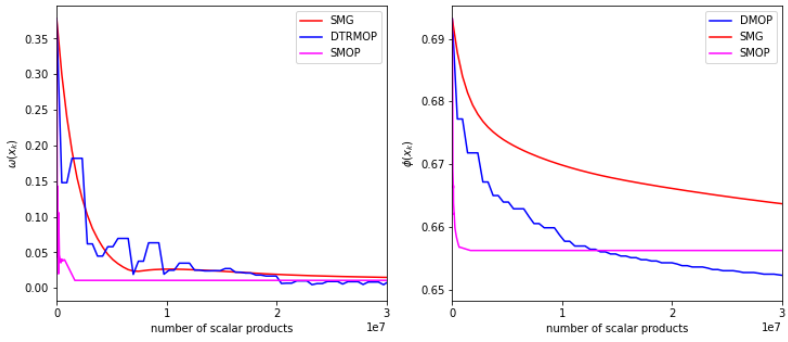

Figure 3: Comparison of SMOP with DMOP [30] and SMG [21], in terms of $\omega(x_k)$, $\phi(x_k)$ and number of scalar products. Model parameters: $x_0 = (0, ..., 0), k_{max} = 150, \theta = 0.01, \gamma_1 = 0.5, \gamma_2 = 2, \eta_1 = 0.25$

5.4 First vs second order algorithm

Although the models m_k are designed to have Hessian values, in the previous applications we used only the first order information, i.e. $H_k = 0$. The algorithm uses models to calculate the trust region ratio, hence the quadratic model should more precisely determine whether the step is accepted, and whether the radius should be increased or decreased. Using the second order information, we expect our algorithm to be more precise and stable in the long run. We compare the first and the second order versions of SMOP, using the heart dataset [19]. The heart dataset has 242 samples, and the attribute dimension is 14. In a similar way, by splitting the data into two sets $|\mathcal{N}_1| = 121, |\mathcal{N}_2| = 121$, we convert the optimization problem (53) into a multi-objective one (54). We approximate the Hessian with the subsampled Hessians using the same subsample as for the functions approximations (55). The following Figure 4 shows the convergence to a critical point of both versions, and ilustrates the comparative advantage of the second order method.

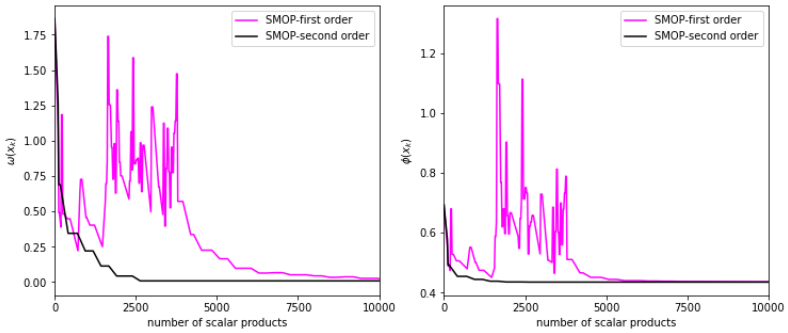

Figure 4: Algorithm comparison. Model parameters: $x_0 = (0, .., 0), k_{max} = 150, \theta = 0.01, \gamma_1 = 0.5, \gamma_2 = 2, \eta_1 = 0.25$

5.5 Finding Pareto front

As previously mentioned, the goal of the SMOP algorithm is to identify a Pareto critical point. However, in some cases, we may wish to better understand the structure of the entire Pareto front, or even find a specific optimal point within it. Hence, approximating the Pareto front is an important feature to implement. Using the standard front finding technique, from [14], we successfully find the Pareto front for different problems. The notion of this Pareto finding method is to approximate the Pareto front using a set of random points. At each iteration this approximation set is expanded by generating perturbed points around the existing elements in the set. Predetermined number of SMOP iterations is then applied to the existing points, after which the results are also added to the approximation set. To refine the

approximation, all dominated points are removed from the set, leaving only the non dominated points to serve as the updated Pareto approximation for the next iteration. Point $x \in \mathcal{L}$ is said to be dominated if there exists $y \in \mathcal{L}$, such that $f(y) < f(x)$, i.e. $f_i(y) < f_i(x)$ for $i = 1, ..., q$. The procedure is described in the following way:

Algorithm 2. *(Pareto front SMOP)*

Step 0. Generate intial Pareto front $\mathcal{L}_0$. Select parameters $n_p, n_q, n_r \in \mathbb{N}$.

Step 1. Set $\mathcal{L}_{k+1} = \mathcal{L}_k$. For each point x in $\mathcal{L}_{k+1}$, add n_r points to $\mathcal{L}_{k+1}$ from the neighborhood of x.

Step 2. For each point x in $\mathcal{L}_{k+1}$, repeat n_p times: Apply n_q iterations of SMOP with x as a starting point. Add the final iteration to $\mathcal{L}_{k+1}$

Step 3. Remove all dominated points from $\mathcal{L}_{k+1}$. Go to Step 1.

The procedure was implemented for convex and nonconvex case. The following figures show the Pareto front for Logistic regression example (54) on heart dataset [19] and for Test problem 2 (52). By choosing different n_p, n_q, n_r it is possible to get a sparser front with less details, hence depending on the problem, the cost of finding the front can be reduced at the expense of Pareto front information.

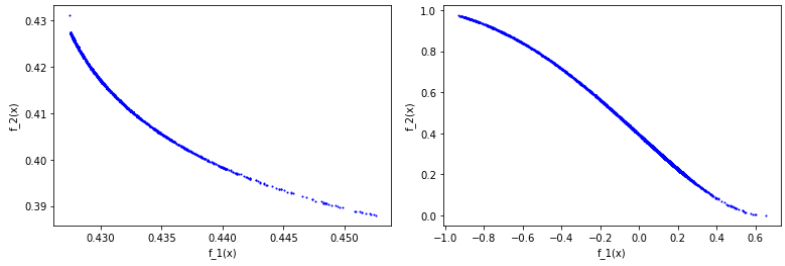

Figure 5: Pareto front for logistic regression (left), and test example 2 (right)

5.6 Conclusion

We have proposed a multi-objective trust region method which utilizes approximate functions, gradients and Hessians. The method is designed to operate under the key assumptions that the approximation accuracy of models is achieved with high enough probability and that the approximate gradients are close enough to the gradients of the approximate functions. The theoretical contribution of this work is the proof of almost sure convergence to a Pareto critical point. We presented several numerical experiments that showcase the algorithm's efficient practical performance. We highlighted its capability to effectively solve multi-objective optimization problems while

maintaining low computational cost. Additionally, we implemented a Pareto finding routine in order to find the Pareto front. Future work could include the generalization of fully quadratic models or techniques such as additional sampling.

Acknowledgments

This work was supported by the Science Fund of the Republic of Serbia, Grant no. 7359, Project LASCADO.

References

[1] Bandeira, A.S., Scheinberg, K. & Vicente, L.N. (2014) Convergence of trust-region methods based on probabilistic models, *SIAM Journal on Optimization, 24(3), 1238-1264.*

[2] Barocas, S., Hardt, M. & Narayanan, A. (2017) Fairness in machine learning. *NIPS Tutorial, 1.*

[3] Bellavia, S., Krejić, N. & Morini, B. (2020) Inexact restoration with subsampled trust-region methods for finite-sum minimization. *Comput Optim Appl 76, 701–736.*

[4] Berkemeier, M. & Peitz, S. (2021) Derivative-Free Multiobjective Trust Region Descent Method Using Radial Basis Function Surrogate Models, *Math. Comput. Appl., 26, 31.*

[5] Berahas A. S., Bollapragada R. & Nocedal J. (2020) An Investigation of Newton-Sketch and Subsampled Newton Methods, *Optimization Methods and Software, 35(4), 661–680.*

[6] Blanchet, J., Cartis, C., Menickelly, M. & Scheinberg, K. (2019) Convergence Rate Analysis of a Stochastic Trust-Region Method via Supermartingales, *INFORMS Journal on Optimization, 1(2), 92-119.*

[7] Blackard, J. (1998) Covertype [Dataset]. *UCI Machine Learning Repository.*

[8] Bollapragada, R., Byrd, R. & Nocedal, J. (2019) Exact and Inexact Subsampled Newton Methods for Optimization, *IMA Journal of Numerical Analysis, 39(20), 545-578.*

[9] Bottou, L., Curtis F.E., Nocedal, J. (2018) Optimization Methods for LargeScale Machine Learning, *SIAM Review, 60(2), 223-311.*

[10] Byrd, R.H., Hansen, S.L., Nocedal, J. & Singer, Y. (2016) A Stochastic QuasiNewton Method for Large-Scale Optimization, *SIAM Journal on Optimization, 26(2), 1008-1021*

[11] Byrd, R.H., Chin, G.M., Nocedal, J. & Wu, Y. (2012) Sample size selection in optimization methods for machine learning, *Mathematical Programming, 134(1), 127-155.*

[12] Chen, R., Menickelly, M. & Scheinberg, K. (2018) Stochastic optimization using a trust-region method and random models, *Math. Program., 169, 447–487.*

[13] Curtis, F.E., Scheinberg, K. & Shi R. (2018) A Stochastic Trust Region Algorithm Based on Careful Step Normalization, *INFORMS Journal on Optimization, 1(3), 200-220.*

[14] Custodio, A.L., Madeira, J.A., Vaz, A.I.F. & Vicente, L. N. (2011) Direct multisearch for multiobjective optimization. *SIAM J. Optim, 21, 1109-1140.*

[15] Durrett, R. (2010) Probability: Theory and Examples. *Cambridge Series in Statistical and Probabilistic Mathemat- ics. Cambridge University Press, Cambridge, fourth edition.*

[16] Fliege, J. & Svaiter, B.F. (2000) Steepest Descent Methods for Multicriteria Optimization, *Mathematical Methods of Operations Research, 51, 479-494.*

[17] Fukuda, E. H. & Drummond, L. M. G. (2014) A survey on multiobjective descent methods. *Pesq. Oper. 34 (3).*

[18] Hardt, M., Price, E. & Srebro, N. (2016) Equality of opportunity in supervised learning. *Advances in neural information processing systems, 3315-3323.*

[19] Janosi, A., Steinbrunn, W., Pfisterer, M. & Detrano, R. (1989) Heart Disease [Dataset]. *UCI Machine Learning Repository.*

[20] Krejić, N., Krklec Jerinkić, N., Martínez, A. & Yousefi, M. (2024) A non-monotone trust-region method with noisy oracles and additional sampling. *Comput Optim Appl 89, 247–278.*

[21] Liu, S. & Vicente, L.N. (2024) The stochastic multi-gradient algorithm for multi-objective optimization and its application to supervised machine learning, *Ann Oper Res, 339, 1119–1148.*

[22] Robbins, H. & Monro, S. (2011) A Stochastic Approximation Method *SIAM J. Optim, 21, 1109-1140.*

[23] Sawaragi, Y., Nakayama, H. & Tanino, T. (1985) Theory of multiobjective optimization, *Elsevier, MR 807529*

[24] Robbins, H. & Siegmund, D. (1971) A convergence theorem for non negative almost supermartingales and some applications, *Optimizing Methods in Statistics, 233-257.*

[25] Roosta-Khorasani, F. & Mahoney, M. W. (2016) Sub-Sampled Newton Methods I: Globally Convergent Algorithms, *arXiv:1601.04737*

[26] Tanabe, H., Fukuda, E.H. & Yamashita, N.(2019) Proximal gradient methods for multiobjective optimization and their applications *Comput. Optim. Appl. 72, 339–361.*

[27] Thomann, J. & Eichfelder, G. (2019) A Trust-Region Algorithm for Heterogeneous Multiobjective Optimization, *SIAM Journal on Optimization, 29, 1017 - 1047.*

[28] Thomann, J. & Eichfelder, G. (2019) Numerical results for the multiobjective trust region algorithm MHT, *Data in Brief, 25.*

[29] Tripuraneni, N., Stern, M., Jin, C., Regier, J. & Jordan, M.I. (2018) Stochastic Cubic Regularization for Fast Nonconvex Optimization *Advances in Neural Information Processing Systems 31.*

[30] Villacorta, K.D.V., Oliveira, P.R. & Soubeyran, A, (2014) A Trust-Region Method for Unconstrained Multiobjective Problems with Applications in Satisficing Processes, *J Optim Theory Appl 160, 865–889.*

[31] Woodworth, B., Gunasekar, S., Ohannessian, M.I. & Srebro, N. (2017) Learning non-discriminatory predictors. *Conference on Learning Theory, 1920-1953.*

[32] Zafar, M.B., Valera, I., Gomez Rodriguez, M. & Gummadi, K.P. (2017) Fairness constraints: Mechanisms for fair classiffication. *Artiffficial Intelligence and Statistics, 962-970.*

[33] Zemel, R., Wu, Y., Swersky, K., Pitassi, T. & Dwork, C. (2013) Learning fair representations. *International Conference on Machine Learning, 325-333.*