SMOP: Stochastic trust region method for multi-objective problems

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

May 14, 2026

Abstract

The problem we consider 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. The key novelty is approximation of each function in the multiobjective problem with probabilistically fully linear model which yields the composite model defined by max operator as a satisfactory approximation for the nonsmooth scalarized objective function. We prove the almost sure convergence of the proposed algorithm to a Pareto critical point. 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.

[*]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

1 Introduction

Multi-objective optimization problems arise in many real-world applications, such as finance, scientific computing, social sciences, engineering, and beyond. These problems are characterized by the need to simultaneously optimize multiple, often conflicting objectives, which significantly complicates the decision making process. Whether one is 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 considered problem can formally be stated as

$$\min_x f(x) = \min_{x \in \mathbb{R}^n} (f_1(x), ..., f_q(x))^T \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}^q$. The main goal of multi-objective optimization is to identify the set of Pareto optimal points. A locally Pareto optimal point is a point such that there exists a neighborhood around it in which no other point improves all objective function values simultaneously, see [21], [31]. If the point can not be improved on the entire domain, the point is globally Pareto optimal. When extending this concept to global solutions, the Pareto front is defined as the set of nondominated objective values corresponding to Pareto optimal points. The algorithms for solving problem (1) are designed to find a broader set of Pareto critical points; points for which no common descent direction exists that improves all objectives simultaneously. By finding Pareto critical points, it is possible by localize the stationary form of the Pareto front, see [18],[27]. 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 approximate the 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 [39]. 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 [36]. 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 [5], where radial basis surrogate models are used.

It is also possible to approach this problem within a line search framework. Armijo-like condition with the steepest descent and Newton direction is discussed in [22]. In [22] authors also analyze the projected gradient method for constrained cases. Stochastic multi-gradient multi-criteria approach can be found in [27]. The authors of [27] successfully extend the classical stochastic gradient (SG, see [30]) 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 [14]. 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 [4],[7],[10]-[13],[26], [33],[34].

The method we propose here is based on probabilistically fully linear models for each function f_i separately, as introduced in [1] and used later on in [14, 8]. The concept of full (probabilistic) linearity is extended to vector function in a natural way as explained further on.

Having a fully linear 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 [14]. 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 function $\tilde{f}_i, i = 1, \ldots, q$ 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 [33]. 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 [21]. 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 [39] 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 [36].

1.1 Contributions

We propose a trust region algorithm for solving multiobjective optimization problem. The problem is first transformed into composite optimization problem with max operator yielding a nonsmooth objective function. We proceed by considering random models per function $f_i, i = 1, \ldots, q$. The standard property of these per function models is assumed - α- probabilistic full linearity. Despite the fact that the scalarization function is nonsmooth we prove that the aggregate random model has sufficiently good agreement with the scalarized function under reasonable assumptions. The trust region method is then defined exploiting the random structure in an asymmetric way - the criteria for search direction is slightly weaker as it is based on approximate stationarity measure of an approximate model. On the other hand the acceptance criteria is slightly stronger than usual in trust region. This asymmetry seems to work well, taking into account randomness in the models and at the same time allowing us to prove theoretically strong result of almost sure convergence. The problem we analyse in detail is the multiobjective problem with finite sums. Hence the random models are based on subsampling of functions and gradients. Numerical results are presented, for the case of per function random models of the first order. These experiments demonstrate the advantages of the proposed approach, in partuclar for the case of large dimensions and large data sets. Full Pareto front is also considered. The proposed method is tested against the state-of-the-art SMG method [27] and deterministic trust region for multiobjective problem from [39].

2 Preliminaries

The considered problem is unconstrained multiobjective minimization problem (1) where $f : \mathbb{R}^n \to \mathbb{R}^q$ and functions $f_j, j = 1, ..., q$ are smooth. Assuming that the explicit evaluation of these functions and its gradients and Hessians is unavailable or too costly, we will rely on approximating them with $\tilde{f}_i(x)$, $g_i(x)$ and $H^i(x)$, respectively.

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

Definition 1. [31, Definition 3.1.2]. 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 $i \in \{1, 2, ..., q\}$.

A point x^* is Pareto critical if and only if there is no direction along which all objective functions decrease simultaneously. In other words, for every direction $d \in \mathbb{R}^n$, there exists at least one component function f_i such that the directional derivative satisfies

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

Pareto optimality implies Pareto criticality, however the converse is not necessarily true.

A stationarity condition for (1) can be derived exploiting the marginal function

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

It plays a similar role to that of the norm of the gradient of the objective function for single objective problems. In fact, if $q = 1$, one gets $\omega(x) = \|\nabla f(x)\|$.

Let us define

$$\mathcal{D}(x) = \arg\min_{\|d\| \leq 1} \left(\max_{i \in \{1,...,q\}} \langle \nabla f_i(x), d \rangle \right).$$

The following lemma will be used for further considerations.

Lemma 1. [21, Lemma 3]. The following statements hold:

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.

One possible scalar representation of the multiobjective problem (1) is

$$\min_{x\in\mathbb{R}^n} \phi(x), \quad \phi(x) = \max_{i\in\{1,\dots,q\}} f_i(x). \tag{3}$$

This problem is not equivalent to problem (1), but every solution of this scalar problem is a Pareto critical point. In order to see that, let us recall that subdifferential of ϕ is given by $\partial\phi(x) = co\{\nabla f_i(x) : i \in I_f(x)\}$, where $I_f(x) := \{i \in \{1, ..., q\} : f_i(x) = \phi(x)\}$ and co denotes the convex hull of the stated vectors. The corresponding stationarity condition is $0 \in \partial\phi(x)$, where 0 represents vector of zeros of dimension n. Thus, a stationarity measure can be defined as

$$\omega_\phi(x) = -\min_{\|d\|\le 1} \left(\max_{i\in I_f(x)} \langle \nabla f_i(x), d\rangle \right)$$

Therefore, it follows that $0 \le \omega(x) \le \omega_\phi(x)$ for every $x \in \mathbb{R}^n$ and thus $\omega_\phi(\tilde{x}) = 0$ implies $\omega(\tilde{x}) = 0$.

Let $B(x, \delta)$ denote a closed ball centered at x with radius δ. In deterministic trust region methods, at iteration k, ϕ is usually approximated locally (on a ball $B(x_k, \delta_k)$, where δ_k represents trust region radius), by a quadratic model

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

where H_k^i approximates the Hessian of function f_i for $i = 1, ..., q$. A measure of proximity for the models is defined as follows, .

Definition 2. [16, Definition 6.1] Function m_k is (c_f, c_g) fully linear (FL) model of function h on $B(x_k, \delta_k)$ if for every $d \in B(0, \delta_k)$ the following two inequalities hold

$$|h(x_k + d) - m_k(d)| \le c_f \delta_k^2 \tag{4}$$

$$\|\nabla h(x_k + d) - \nabla m_k(d)\| \le c_g \delta_k. \tag{5}$$

Given that we assume that the computation of exact functions and their derivatives is not feasible, approximate functions are to be used in general. Therefore we define the approximate quadratic model for $\tilde{\phi}$ analogously, i.e.,

$$\tilde{m}_k(d) = \max_{i\in\{1,\dots,q\}} \tilde{m}_{k,i}(d), \tag{6}$$

where

$$\tilde{m}_{k,i}(d) = \tilde{f}_i(x_k) + \langle g_i(x_k), d\rangle + \frac{1}{2}\langle d, H_k^i d\rangle, \ i = 1, ..., q. \tag{7}$$

Notice that $\nabla \tilde{m}_{k,i}(0) = g_i(x_k)$ and $\tilde{m}_{k,i}(0) = \tilde{f}_i(x_k)$ for each $i = 1, ..., q$. Furthermore, following [36], we consider the approximate marginal function

$$\omega_m(x) = -\min_{\|d\| \leq 1} \left(\max_{i \in \{1,...,q\}} \langle g_i(x), d \rangle \right) \tag{8}$$

as a stationarity measure of the approximate multi-objective problem

$$\min_{x \in \mathbb{R}^n} \tilde{f}(x) = \min_{x \in \mathbb{R}^n} (\tilde{f}_1(x), ..., \tilde{f}_q(x))^T.$$

Analogously to (3) we denote by $\omega_{\tilde{\phi}}$ stationarity measure of the following scalar problem

$$\min_{x \in \mathbb{R}^n} \tilde{\phi}(x), \quad \tilde{\phi}(x) = \max_{i \in \{1,...,q\}} \tilde{f}_i(x)$$

and conclude that $0 \leq \omega_m(x) \leq \omega_{\tilde{\phi}}(x)$ for every $x \in \mathbb{R}^n$. These approximate versions are going to be used within the algorithm proposed in the next section, while the convergence analysis will relay on the true marginal function ω.

2.1 Stochastic framework

Our main motivation comes from observing machine learning problems where the functions $f_i, i = 1, ..., q$ 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. We will use upper case letters to emphasize random quantities where appropriate e.g., X_k for random iterates, and lowercase letters to denote the corresponding realizations e.g., x_k. To be more precise, let us denote by $(\Omega, \mathcal{F}, P)$ the probability space where: Ω represents the set of all possible outcomes, i.e., all possible sample paths of the algorithm to be stated; $\mathcal{F}$ is a σ-algebra on Ω; and P is a probability function on a measurable space $(\Omega, \mathcal{F})$.

We assume that the stochastic influence comes exclusively from random choices of approximate functions and its derivatives. The stochastic counterparts of $\tilde{f}_i, g_i$ and H^i will be denoted by $\tilde{F}_i, G_i$ and χ^i, respectively. These random objects constitute the model functions (7) which are also random and thus denoted by $\tilde{M}_{k,i}, i = 1, .., q$. The corresponding agregate model function is denoted accordingly by $\tilde{M}_k$. Since the iterates update will be based on random models, we will also have X_k as random vectors. The same is true for the trust region radius whose stochastic counterpart will be denoted by Δ_k. Although random sampling is an original generator of stochastic influence within the considered framework, we set $\{X_k\}_{k \in \mathbb{N}}$ as a representative stochastic process as common in the literature. Then, we denote by $\mathcal{F}_k$ the sub-σ-algebra of $\mathcal{F}$ generated by $X_1, ..., X_k$. Thus, $\{\mathcal{F}_k\}_{k \in \mathbb{N}}$ is the natural filtration of $\mathcal{F}$ with respect to $\{X_k\}_{k \in \mathbb{N}}$ and there

holds $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq ... \subseteq \mathcal{F}$. We denote by $E(\cdot|\mathcal{F}_k)$ the conditional expectation at iteration k, and by $E(\cdot)$ unconditional expectation with respect to all possible sample paths $v \in \Omega$.

In convergence analysis we will use the concept of probabilistic fully linear models [1]. Instead of fixing the probability parameter α, we introduce a sequence of relevant probabilities $\alpha := \{\alpha_k\}_{k\in\mathbb{N}}$ and give the following definition.

Definition 3. [1, Definition 3.2] A sequence of random models $\{\tilde{M}_{k,i}\}_{k\in\mathbb{N}}$ is α -probabilistically (c_f, c_g) fully linear with respect to the corresponding sequence of $B(X_k, \Delta_k)$ if the events

$$I_{k,i} = \{\tilde{M}_{k,i} \text{ is } (c_f, c_g) \text{ fully linear model of } f_i \text{ on } B(X_k, \Delta_k)\}$$

satisfy the condition $P(I_{k,i}|\mathcal{F}_k) \geq \alpha_k$ for all k.

Since the multi-objective setup requires multiple models, we will introduce the following definition of Jointly Independent Probabilistically Fully Linear models.

Definition 4. (JIPFL condition) We say that a sequence of multiple random models $\{\tilde{M}_{k,1}, ..., \tilde{M}_{k,q}\}_{k\in\mathbb{N}}$ is jointly independent α -probabilistically (c_f, c_g) fully linear with respect to the corresponding sequence of $B(X_k, \Delta_k)$ if the sequence of random models $\{\tilde{M}_{k,i}\}_{k\in\mathbb{N}}$ is α -probabilistically (c_f, c_g) fully linear with respect to the corresponding sequence of $B(X_k, \Delta_k)$ for each $i = 1, ..., q$ and the events $I_{k,1}, ..., I_{k,q}$ are mutually independent conditionally on $\mathcal{F}_k$ for all $k \in \mathbb{N}$.

Notice that the JIPFL condition implies

$$P(I_k|\mathcal{F}_k) := P(\bigcap_{i=1}^{q} I_{k,i}|\mathcal{F}_k) = \prod_{i=1}^{q} P(I_{k,i}|\mathcal{F}_k) \geq \alpha_k^q, \quad \text{for all } k \in \mathbb{N}. \tag{9}$$

The above stated condition of independence is often fulfilled in the finite sum setup since each function f_i is usually approximated by independent random sampling. The main point of the above definition is to allow us to connect the per function individual probabilistically fully linear models with the scalarization function ϕ, which is nonsmooth and hence full linearity of the multiple random models with respect to ϕ can not even be defined. However we will see later on that JIPFL condition allow us to prove that the multiple random models are good enough for almost sure convergence.

3 Algorithm

We state the algorithm as follows. Although a vast majority of objects in the algorithm is stochastic, we present them in small letters for readability.

Algorithm 1.
(SMOP: Stochastic trust region method for Multi-Objective Problems)

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

Step 1. Form a model $\tilde{m}_k(d)$ by (7) and (6).

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

$$\tilde{m}_k(0) - \tilde{m}_k(d_k) \geq \frac{1}{2}\omega_m(x_k)\min\{\delta_k, \frac{\omega_m(x_k)}{\beta_k}\}, \tag{10}$$

where $\beta_k := 1 + \max_{i\in\{1,\dots,q\}} \|H_k^i\|$.

Step 3. Compute

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

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

In the initialization we choose a starting point together with several hyperparameters. The parameters Θ and η are used to define successful iterations. These parameters also influence the trust region radius update, while the intensity of the update is controlled by parameter γ_1. Moreover, we set the initial value and the upper bound of the trust region radius to δ_0 and δ_{max}, respectively.

In Step 1 we form approximate random quadratic models for each function $f_i, i = 1, ..., q$ and the aggregate model by (7) and (6). Considering the stochastic framework, the inflow of randomness happens here since the constituting approximate functions $(\tilde{f}_i)$ and the derivatives (g_i, H_k^i) are constructed/sampled within this step. Later on we will see that the JIPFL condition ensures the almost sure convergence. JIPFL condition also guides the sampling approach from two perspectives. First, the samples are to be drawn independently across functions $f_i, i = 1, ..., q$. Beside that, JIPFL condition guides the sample size update since it influences the condition (9).

In Step 2 we search for a suitable direction which provides a sufficient decrease of the approximate aggregate model function. A suitable decrease is defined through δ_k, β_k and the stationarity measure ω_m (8). In Lemma 2 we prove that it is possible to find such direction and therefore Step 2 is well defined.

In Step 3 we calculate an agreement between the approximate model reduction and the reduction of approximate scalarization function. If the

agreement is sufficiently large and the stationarity measure $\omega_m(x_k)$ is relatively large with respect to δ_k, then we accept the proposed iterate update and increase the trust region radius if possible. We will refer to these iterations as successful iterations in the sequel. Otherwise, if the iteration is not successful, we reject the update and decrease the trust region radius in order to give better chances to the model to be a good representative of the function $\tilde{\phi}$ on smaller region in the next iteration.

One can notice that we alternate between the approximate multi-objective problem and its scalarization since the models are targeting $\tilde{\phi}$, while we use the measure of stationarity ω_m of an approximate multiobjective problem instead of the stationarity measure of its scalarized version $\omega_{\tilde{\phi}}$. Since it is known that $\omega_{\tilde{\phi}} \geq \omega_m$, this results in relaxed condition (10). The reasoning behind this includes the fact that we are dealing with approximate (stochastic) versions in general. Imposing strict conditions while having possibly poor approximations of the objective functions is usually far from beneficial. On the other hand, this kind of relaxation usually needs to be compensated. We compensate in Step 3 where ω_m is used within the acceptance criteria instead of $\omega_{\tilde{\phi}}$. This seems to provide a good balance on average.

The following lemma shows that the algorithm is well defined.

Lemma 2. For all k, there exists d_k such that (10) holds.

Proof. We will prove that the condition (10) holds for the Cauchy direction, $d_k^c = \alpha_k d_k^*$, where d_k^* is a solution of the problem stated in (8), i.e.,

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

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

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

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

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

Thus, we conclude that

$$\tilde{m}_k(0) - \tilde{m}_k(d_k^c) = \max_{0\leq\alpha\leq\delta_k}\{\tilde{m}_k(0) - \tilde{m}_k(\alpha d_k^*)\} \geq \max_{0\leq\alpha\leq\delta_k}\{\alpha\omega_m(x_k) - \frac{1}{2}\alpha^2\beta_k\}. \tag{11}$$

Notice that the solution of the problem at the right-hand side of (11) 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 (11) 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.

■

4 Convergence analysis

This section provides convergence analysis of the proposed method. We start the analysis by stating some basic assumptions. In Lemmas 3-5 we provide some bounds that hold under assumption of full linearity. More precisely, we show that if the realizations of random models $\tilde{m}_{k,i} = \tilde{M}_{k,i}(v), i = 1, ..., q$, for some $v \in \Omega$ are fully linear models of $f_i, i = 1, ..., q$, respectively, on $B(x_k, \delta_k)$, then the distance between the true function f_i and approximate function $\tilde{f}_i$ is controllable by δ_k^2 on $B(x_k, \delta_k)$ for every $i = 1, ..., q$. Then, we show that the same is true for the distance between the function ϕ and aggregate approximate model $\tilde{m}_k$. Under the same settings, we also show that the distance between the marginal function $\omega(x_k)$ and its approximation $\omega_m(x_k)$ is controllable by δ_k. In Lemma 6 we also consider the same setup, but we prove that one of the acceptance criteria ($\rho_k \geq \eta_1$) is satisfied provided that the trust region radius is small enough.

The analysis is continued by introducing Lyapunov function Ψ_k that combines $\phi(X_k)$ and Δ_k^2. Under uniformly bounded iterates assumption and JIPFL condition, we prove that the sequence $\{\Psi_k\}_{k\in\mathbb{N}}$ converges almost surely and that the sequence $\{\Delta_k\}_{k\in\mathbb{N}}$ is square sumable almost surely, provided that the sequence $\{\alpha_k\}_{k\in\mathbb{N}}$ tends to 1 fast enough (Theorem 2). Theorems 4-5 provide further properties and yield the main result stated in Theorem 6 where we prove that $\{\omega(X_k)\}_{k\in\mathbb{N}}$ tends to zero almost surely.

The remaining lemmas and theorems provide some important intermediate results.

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

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 $\beta_k \leq c_b$ 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 2 is strong, but it is fulfilled in some important classes of machine learning problems such as logistic regression and linear least squares. Assumption 3 is also 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.

Lemma 3. Assume that A1-A3 hold. Suppose that $v \in \Omega$ is such that $\tilde{m}_{k,i} = \tilde{M}_{k,i}(v)$ is (c_f, c_g)-fully linear model of f_i on $B(x_k, \delta_k)$, where $x_k = X_k(v)$ and $\delta_k = \Delta_k(v)$. Then there exists $c_e > 0$ such that for all $d_k \in B(0, \delta_k)$ there holds

$$|\tilde{f}_i(x_k + d_k) - f_i(x_k + d_k)| \leq c_e \delta_k^2 \tag{12}$$

where $\tilde{f}_i = \tilde{F}_i(v)$.

Proof. Let us take any $d_k \in B(0, \delta_k)$, i.e., any d_k satisfying $\|d_k\| \leq \delta_k$. Then there exist τ_k^i and v_k^i on the line segment between x_k and $x_k + d_k$ 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 the second full linearity condition (5) and the fact that $g_i(x_k) = \nabla \tilde{m}_{k,i}(0)$, where $g_i = G_i(v)$, 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. ■

Lemma 4. Assume that A1 holds Suppose that $v \in \Omega$ is such that $\tilde{m}_{k,i} = \tilde{M}_{k,i}(v)$ is (c_f, c_g)-fully linear model of f_i on $B(x_k, \delta_k)$ for all $i = 1, .., q$, where $x_k = X_k(v)$ and $\delta_k = \Delta_k(v)$. Then

$$|\omega(x_k) - \omega_m(x_k)| \leq \delta_k c_g \tag{13}$$

Proof. By using the following notation $h_k(d) := \max_{i\in\{1,\dots,q\}} \langle \nabla f_i(x_k), d\rangle$, $\tilde{h}_k(d) := \max_{i\in\{1,\dots,q\}} \langle g_i(x_k), d\rangle$ we get

$$\begin{aligned}
|\omega(x_k) - \omega_m(x_k)| &= \left| \min_{\|d\|\leq 1} h_k(d) - \min_{\|d\|\leq 1} \tilde{h}_k(d) \right| \\
&\leq \sup_{\|d\|\leq 1} \left| h_k(d) - \tilde{h}_k(d) \right| \\
&\leq \sup_{\|d\|\leq 1} \max_{i\in\{1,\dots,q\}} |\langle \nabla f_i(x_k) - g_i(x_k), d\rangle| \\
&\leq \max_{i\in\{1,\dots,q\}} \|\nabla f_i(x_k) - g_i(x_k)\| \sup_{\|d\|\leq 1} \|d\| \\
&\leq c_g \delta_k,
\end{aligned}$$

where the second full linearity condition (5) is used for the the final inequality. ■

In the next lemma we prove that although ϕ is nonsmooth the multidimensional random model approximates ϕ and $\tilde{\phi}$ with the order of δ_k^2.

Lemma 5. Assume that A1 -A3 hold. Suppose that $v \in \Omega$ is such that $\tilde{m}_{k,i} = \tilde{M}_{k,i}(v)$ is (c_f, c_g)-fully linear model of f_i on $B(x_k, \delta_k)$ for all $i = 1, .., q$, where $x_k = X_k(v)$ and $\delta_k = \Delta_k(v)$. Then the following inequalities hold for all $d_k \in B(0, \delta_k)$:

$$|\phi(x_k + d_k) - \tilde{m}_k(d_k)| \leq c_f \delta_k^2, \tag{14}$$

$$|\tilde{\phi}(x_k + d_k) - \phi(x_k + d_k)| \leq c_e \delta_k^2, \tag{15}$$

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

where $c_{\tilde{\Phi}} = \max\{c_f, c_e\}$.

Proof. Inequality (14) is obtained by using the first inequality of fully linear models (4) as follows

$$\begin{aligned} |\phi(x_k+d_k)-\tilde{m}_k(d_k)| &= \left| \max_{i\in\{1,\dots,q\}} f_i(x_k+d_k) - \max_{i\in\{1,\dots,q\}} \tilde{m}_{k,i}(d_k) \right| \\ &\leq \max_{i\in\{1,\dots,q\}} |f_i(x_k+d_k)-\tilde{m}_{k,i}(d_k)| \\ &\leq \max_{i\in\{1,\dots,q\}} c_f\delta_k^2 = c_f\delta_k^2. \end{aligned}$$

We obtain (15) by using similar arguments and (12) instead of (4), while (16) is obtained by using the fact that

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

and applying (14) and (15). ■

Lemma 6. Assume that A1 -A3 hold. Suppose that $v \in \Omega$ is such that $\tilde{m}_{k,i} = \tilde{M}_{k,i}(v)$ is (c_f, c_g)-fully linear model of f_i on $B(x_k, \delta_k)$ for all $i = 1, .., q$, where $x_k = X_k(v)$ and $\delta_k = \Delta_k(v)$. Suppose that d_k satisfies (10). Then $\rho_k \geq \eta_1$ provided that

$$\delta_k \leq \min\{\frac{\omega_m(x_k)}{c_b}, \frac{\omega_m(x_k)(1-\eta_1)}{2c_{\tilde{\Phi}}}\}. \tag{17}$$

Proof. From (10) 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 (16), we obtain

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

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

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

To continue with the convergence analysis, let us define an auxiliary Lyapunov function as usual in this type of analysis, [14]

$$\Psi_k := \nu\phi(X_k) + (1-\nu)\Delta_k^2, \quad \nu \in (0,1).$$

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

$$E[\Psi_{k+1} - \Psi_k | \mathcal{F}_k] \leq -\sigma \Delta_k^2 + (1 - \alpha_k^q)\tilde{\sigma}, \quad k = 0, 1, ...$$

for some $\sigma, \tilde{\sigma} > 0$. Let us define the event of successful iteration k as

$$S_k = \{\mathcal{R}_k \geq \eta_1 \text{ and } \omega_m(X_k) > \Theta \Delta_k\},$$

where $\mathcal{R}_k$ denotes the stochastic counterpart of ρ_k. We also define the complementary event (unsuccessful iteration k) by

$$\bar{S}_k = \{\mathcal{R}_k < \eta_1 \text{ or } \omega_m(X_k) \leq \Theta \Delta_k\}.$$

Notice that

$$E(\Psi_{k+1} - \Psi_k | \mathcal{F}_k, \bar{S}_k) = (1 - \nu)(\gamma_1^2 - 1)\Delta_k^2 =: -c_1 \Delta_k^2, \tag{18}$$

for some $c_1 > 0$. Thus, in the subsequent lemma we focus our attention on estimating

$$E(\Psi_{k+1} - \Psi_k | \mathcal{F}_k, S_k) = E(\nu(\phi(X_{k+1}) - \phi(X_k)) + (1 - \nu)(\gamma_2^2 - 1)\Delta_k^2 | \mathcal{F}_k, S_k).$$

The proof of the following lemma resembles the analysis of [14]. However, having the multi-objective problem requires nontrivial modifications due to the fact that the random models are defined per function f_i and that ϕ is nonsmooth. Two additional assumptions are needed for the analysis. The first one is JIPFL property defined in Section 3, while the second assumption states that the iterative sequence is uniformly bounded. Although relatively strong, the later assumption is often used in stochastic optimization, [11], [16], [6],[9],[29].

Assumption 4. The sequence of multiple random models $\{\tilde{M}_{k,1}, ..., \tilde{M}_{k,q}\}_{k \in \mathbb{N}}$ satisfies JIPFL condition with respect to the corresponding sequence of $B(X_k, \Delta_k)$.

Assumption 5. The sequence $\{X_k\}_{k \in \mathbb{N}}$ is uniformly bounded.

For the purpose of the convergence analysis, let us define the following events

$$J_{k,i} = \{|\tilde{F}_i(X_k + d_k) - f_i(X_k + d_k)| \leq c_e \Delta_k^2 \text{ for all } d_k \in B_k(0, \Delta_k)\},$$

$i = 1, ..., q$ and

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

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

$$P(I_k,J_k|\mathcal{F}_k)=P(I_k|\mathcal{F}_k)P(J_k|I_k,\mathcal{F}_k)\geq\alpha_k^q 1=\alpha_k^q \tag{19}$$

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 $P(\bar{I}_k,\bar{J}_k|\mathcal{F}_k)\leq 1-\alpha_k^q$, where $\bar{I}_k$ and $\bar{J}_k$ denote the complementary events of I_k and J_k, respectively. We also use D_k to denote stochastic counterpart of the step size d_k determined in Step 2 of the SMOP algorithm.

Lemma 7. Suppose that A1-A5 hold 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

$$E(\Psi_{k+1}-\Psi_k|\mathcal{F}_k,S_k)\leq -c_6\Delta_k^2+c_7(1-\alpha_k^q),$$

if

$$\Theta\geq\max\{c_b,5c_f,\frac{4c_e}{\eta_1}\}\text{ and }\frac{\nu}{1-\nu}\geq\frac{4\gamma_2^2-2}{\min\{c_e,c_f\}} \tag{20}$$

Proof. Given $\mathcal{F}_k\bigcap S_k$, the following events make mutually exclusive and collectively exhaustive events at iteration k

$$U_k^1:=I_k,\quad U_k^2:=\bar{I}_k\bigcap J_k,\quad U_k^3:=\bar{I}_k\bigcap\bar{J}_k.$$

We analyze these three cases separately and gather them together at the end of the proof to obtain the result.

a)Let us consider $E(\Psi_{k+1}-\Psi_k|\mathcal{F}_k,S_k,U_k^1)$ first. Since I_k implies J_k, and S_k implies that $\omega_m(X_k)\geq\Theta\Delta_k$, using (14), and Lemma 2 we obtain

$$\begin{aligned}
& E(\phi(X_{k+1})-\phi(X_k)|\mathcal{F}_k,S_k,U_k^1) \qquad (21)\\
= \;& E(\phi(X_{k+1})-\tilde{M}_k(D_k)+\tilde{M}_k(0)-\phi(X_k)+\tilde{M}_k(D_k)-\tilde{M}_k(0)|\mathcal{F}_k,S_k,U_k^1)\\
\leq \;& E(2c_f\Delta_k^2-\frac{1}{2}w_m(X_k)\min\{\Delta_k,\frac{w_m(X_k)}{c_b}\}|\mathcal{F}_k,S_k,U_k^1)\\
\leq \;& E(2c_f\Delta_k^2-\frac{1}{2}w_m(X_k)\Delta_k|\mathcal{F}_k,S_k,U_k^1)\\
\leq \;& E(2c_f\Delta_k^2-\frac{1}{2}\Theta\Delta_k^2|\mathcal{F}_k,S_k,U_k^1)\\
< \;& E(-\frac{1}{2}c_f\Delta_k^2|\mathcal{F}_k,S_k,U_k^1)=-c_1\Delta_k^2,
\end{aligned}$$

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

$$\begin{aligned}
& E(\Psi_{k+1}-\Psi_k|\mathcal{F}_k,S_k,U_k^1)\\
= \;& E(\nu(\phi(X_{k+1})-\phi(X_k))+(1-\nu)(\gamma_2^2-1)\Delta_k^2|\mathcal{F}_k,S_k,U_k^1)\\
\leq \;& [-\nu c_1+(1-\nu)(\gamma_2^2-1)]\Delta_k^2,
\end{aligned}$$

and thus by choosing ν as in (20) we obtain

$$\frac{\nu}{1-\nu} \geq \frac{2\gamma_2^2 - 1}{c_1}$$

and

$$E(\Psi_{k+1} - \Psi_k | \mathcal{F}_k, S_k, U_k^1) \leq -\gamma_2^2 \Delta_k^2 = -c_2 \Delta_k^2, \tag{22}$$

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

b)Now let us consider $E(\Psi_{k+1} - \Psi_k | \mathcal{F}_k, S_k, U_k^2)$. Using (15) and (10), and the fact that S_k implies $\mathcal{R}_k \geq \eta_1$ we get

$$\begin{aligned}
& E(\phi(X_{k+1}) - \phi(X_k) | \mathcal{F}_k, S_k, U_k^2) \\
= \; & E(\phi(X_{k+1}) - \tilde{\phi}(X_{k+1}) + \tilde{\phi}(X_k) - \phi(X_k) + \tilde{\phi}(X_{k+1}) - \tilde{\phi}(X_k) | \mathcal{F}_k, S_k, U_k^2) \\
\leq \; & 2E(c_e \Delta_k^2 + \tilde{\phi}(X_{k+1}) - \tilde{\phi}(X_k) | \mathcal{F}_k, S_k, U_k^2)) \\
= \; & E(2c_e \Delta_k^2 - \mathcal{R}_k(\tilde{M}_k(D_k) - \tilde{M}_k(0)) | \mathcal{F}_k, S_k, U_k^2) \\
\leq \; & E(2c_e \Delta_k^2 - \eta_1(\tilde{M}_k(D_k) - \tilde{M}_k(0)) | \mathcal{F}_k, S_k, U_k^2) \\
\leq \; & E(2c_e \Delta_k^2 - \frac{\eta_1 \omega_m(X_k)}{2} \min\{\frac{\omega_m(X_k)}{c_b}, \Delta_k\} | \mathcal{F}_k, S_k, U_k^2) \\
\leq \; & [2c_e - \frac{\eta_1 \Theta}{2} \min\{\frac{\Theta}{c_b}, 1\}] \Delta_k^2 = [2c_e - \frac{\eta_1 \Theta}{2}] \Delta_k^2 \leq -\frac{1}{2} c_e \Delta_k^2 = -c_3 \Delta_k^2.
\end{aligned}$$

for $\Theta \geq \frac{5c_e}{\eta_1}$, and $c_3 = \frac{1}{2} c_e > 0$. Again, from (20) we obtain

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

and we get that

$$E(\Psi_{k+1} - \Psi_k | \mathcal{F}_k, S_k, U_k^2) \leq [-\nu c_3 + (1-\nu)(\gamma_2^2 - 1)] \Delta_k^2 \leq -\gamma_2^2 \Delta_k^2 = -c_4 \Delta_k^2. \tag{23}$$

for $c_4 = \gamma_2^2 > 0$

c)Finally, let us consider $E(\Psi_{k+1} - \Psi_k | \mathcal{F}_k, S_k, U_k^3)$. Again, S_k implies $\omega_m(X_k) \geq \Theta \Delta_k$, but an increase of Ψ_k can happen. However, using Taylor expansion, A2, and the Cauchy Schwartz inequality, we obtain the following bound regardless of the scenario U_k^3.

$$\begin{aligned}
\phi(X_{k+1}) - \phi(X_k) &= \max_{i \in \{1,\dots,q\}} f_i(X_{k+1}) - \max_{i \in \{1,\dots,q\}} f_i(X_k) \\
&\leq \max_{i \in \{1,\dots,q\}} |\nabla^T f_i(x_k) d_k + \frac{1}{2} D_k^T \nabla^2 f_i(T_k) D_k| \\
&\leq \max_{i \in \{1,\dots,q\}} (\|\nabla f_i(X_k)\| \|D_k\| + \frac{1}{2} \Delta_k^2 c_h).
\end{aligned}$$

Since the iterates are assumed to be bounded, the continuity of the gradients implies the existence of $G > 0$ such that $\max_{i \in \{1,\dots,q\}} \|\nabla f_i(X_k)\| \leq G$. Since $\Delta_k \leq \delta_{max}$ there exists a constant c_5 such that

$$\phi(X_{k+1}) - \phi(X_k) \leq G \delta_{max} + \frac{1}{2} \delta_{max}^2 c_h =: c_5$$

and thus

$$E(\Psi_{k+1} - \Psi_k | \mathcal{F}_k, S_k, U_k^3) \leq c_5 + (1-\nu)(\gamma_2^2 - 1)\Delta_k^2. \tag{24}$$

Now, we combine inequalities (22),(23) and (24) to estimate $E(\Psi_{k+1} - \Psi_k | \mathcal{F}_k, S_k)$. Using the total probability formula we obtain

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

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 (22) 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 (24). Thus, by (19) 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 $\delta_k \leq \delta_{max}^2$.

■

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

Assumption 6. The sequence $\{\alpha_k\}_k$ satisfies $\sum_{k=0}^{\infty}(1 - \alpha_k^q) \leq c_\alpha < \infty$.

We also use the following result for further analysis.

Theorem 1. [32, Theorem 1] 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. Suppose that A1-A6 and (20) hold. Then the sequence $\{\Psi_k\}_{k\in\mathbb{N}}$ converges a.s. and there holds

$$\sum_{k=0}^{\infty} \Delta_k^2 < \infty \quad \text{a.s.} \tag{25}$$

Proof. Assumption 6 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 (18) and Lemma 7 we obtain

$$\begin{aligned} & E(\Psi_{k+1} - \Psi_k|\mathcal{F}_k) \\ = \; & E(\Psi_{k+1} - \Psi_k|\mathcal{F}_k, S_k)P(S_k|\mathcal{F}_k) + E(\Psi_{k+1} - \Psi_k|\mathcal{F}_k, \bar{S}_k)P(\bar{S}_k|\mathcal{F}_k) \\ \leq \; & (-c_6\Delta_k^2 + c_7(1-\alpha_k^q))P(S_k|\mathcal{F}_k) - c_1\Delta_k^2 P(\bar{S}_k|\mathcal{F}_k) \leq \\ \leq \; & -\min\{c_1, c_6\}(P(S_k|\mathcal{F}_k) + P(\bar{S}_k|\mathcal{F}_k))\Delta_k^2 + c_7(1-\alpha_k^q) \\ =: \; & -c_8\Delta_k^2 + c_7(1-\alpha_k^q) \end{aligned} \tag{26}$$

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 stated conditions a.s. there exists an infinite sequence of iterations with fully linear models. We employ the following result.

Theorem 3. [20, Theorem 5.3.1]. [20] 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 $P(C \cup D) = 1$

Theorem 4. Suppose that the assumptions of Theorem 2 hold. Then a.s. there exists an infinite $K \subseteq \mathbb{N}$ such that $\tilde{M}_{k,i}$ is (c_f, c_g) fully linear model of f_i on $B(X_k, \Delta_k)$ for all $i = 1, ..., q$ and all $k \in K$.

Proof. Notice that assumption A6 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

$$W_k = \sum_{s=\bar{k}}^{k} V_s, \tag{27}$$

where $V_k|I_k, \mathcal{F}_k = 1$ and $V_k|\bar{I}_k, \mathcal{F}_k = -1$ otherwise. Moreover,

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

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 3 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 3 implies that a.s.

$$\limsup_{k\to\infty} W_k = \infty. \tag{28}$$

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 (28).

■

Theorem 5. Suppose that the assumptions of Theorem 2 hold. Then a.s.

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

Proof. Recall that Ω stands for the set of all possible sample paths of SMOP algorithm. Suppose the contrary, that with positive probability none of the subsequences of $\{\omega(X_k(v))\}_{k\in\mathbb{N}}$ converges to 0. In other words there exists $\hat{\Omega} \subset \Omega$ such that $P(\hat{\Omega}) > 0$ and $\{\omega(X_k(v))\}_{k\in\mathbb{N}}$ is bounded away from zero for all $v \in \hat{\Omega}$.

Let us observe an arbitrary $v \in \hat{\Omega}$ and the corresponding realization $\{\omega(X_k(v))\}_{k\in\mathbb{N}}$. Under the current assumption we know that there exists $\epsilon(v) > 0$ and $k_1(v)$, such that $\omega(X_k(v)) \geq \epsilon(v) > 0$ for all $k \geq k_1(v)$. Moreover, Theorem 4 implies that there exists $\tilde{\Omega} \subseteq \Omega$ such that $P(\tilde{\Omega}) = 1$ and for every $v \in \tilde{\Omega}$ there exists $K(v) \subseteq \mathbb{N}$ such that for all $k \in K(v)$, $\tilde{M}_{k,i}(v)$ are (c_f, c_g) fully linear model of f_i on $B(X_k(v), \Delta_k(v))$ for all $i = 1, ..., q$.

Now, let us observe $\overline{\Omega} := \tilde{\Omega} \bigcap \hat{\Omega}$. Notice that $P(\overline{\Omega}) > 0$. Moreover, since Δ_k tends to 0 a.s. according to Theorem 2, without loss of generality we can assume that $\lim_{k\to\infty} \Delta_k(v) = 0$ for all $v \in \overline{\Omega}$.

Let us take an arbitrary $v \in \overline{\Omega}$. There exists $k_2(v)$ such that for $k \geq k_2(v)$,

$$\Delta_k(v) < b(v) := \min\{\frac{\epsilon(v)}{2c_g}, \frac{\epsilon(v)}{2\Theta}, \frac{\epsilon(v)}{2c_b}, \frac{\epsilon(v)(1-\eta_1)}{4c_{\tilde{\Phi}}}\} \tag{29}$$

Let us denote by $\hat{K}(v)$ the set of all indices from $K(v)$ such that $k \geq k_3(v) = \max\{k_1(v), k_2(v)\}$. Thus, for all $k \in \hat{K}(v)$ we have fully linear models, $\omega(X_k(v)) \geq \epsilon(v)$ and $\Delta_k(v)$ is small enough. Furthermore, from Lemma 4 and (29) we obtain,

$$|\omega(X_k(v)) - \omega_m(X_k(v))| \leq c_g\Delta_k(v) \leq \frac{\epsilon(v)}{2}$$

and $\omega_m(X_k(v)) \geq \frac{\epsilon(v)}{2} \geq \Delta_k(v)\Theta$. Moreover, for all $k \in \hat{K}(v)$, the condition

(17) is satisfied

$$\begin{aligned}\Delta_k(v) &< \min\{\frac{\epsilon(v)}{2c_g}, \frac{\epsilon(v)}{2\Theta}, \frac{\epsilon(v)}{2c_b}, \frac{\epsilon(v)(1-\eta_1)}{4c_{\tilde{\Phi}}}\} \\ &\leq \min\{\frac{\omega_m(X_k(v))}{c_b}, \frac{\omega_m(X_k(v))(1-\eta_1)}{2c_{\tilde{\Phi}}}\}.\end{aligned}$$

Thus we conclude that $\mathcal{R}_k(v) \geq \eta_1$ which together with $\omega_m(X_k(v)) \geq \Delta_k(v)\Theta$ implies that all iterations in $\hat{K}(v)$ are successful iterations. Therefore for all $k \in \hat{K}(v)$ there holds $\Delta_{k+1}(v) = \Delta_k(v)\gamma_2 > \Delta_k(v)$. Let us define

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

where $b(v)$ is defined in (29). Notice that for $k \geq k_3(v), \Delta_k(v) < b(v)$, hence $\gamma_2^{r_k(v)} < 1$ and $r_k(v) < 0$. Moreover,

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

Therefore, we have $r_{k+1}(v) = r_k(v) + 1$ for all $k \in \hat{K}$. Notice that the increase of $r_k(v)$ can also happen for some $k \geq k_3(v)$ even if $k \notin \hat{K}(v)$. On the other hand, the increase of $W_k(v)$ defined in (27) is possible if and only if the models $\tilde{M}_{k,i}(v), i = 1, ..., q$, are fully linear. That means that for all $k \geq k_3(v)$, the increase $W_{k+1}(v) = W_k(v) + 1$ happens only if $k \in \hat{K}(v)$, while in the remaining iterations $k \geq k_3(v), k \notin \hat{K}(v)$ we have $W_{k+1}(v) = W_k(v) - 1$. Thus for all $k > k_3(v)$ the increase of $r_k(v)$ happens in the same or bigger number of iterations than the increase of $W_k(v)$ and we conclude that the following must hold

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

Since (28) holds almost surely, without loss of generality we conclude that $\limsup_{k\to\infty} W_k(v) = \infty$ and thus $\limsup_{k\to\infty} r_k(v) = \infty$ which contradicts the fact that $r_k(v) < 0$ for all $k \geq k_3(v)$. ■

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 analyze the two possible scenarios regarding full linearity separately.

Let us consider $E(\Psi_{k+1} - \Psi_k | I_k, \mathcal{F}_k)$ first. According to (25) we have $\lim_{k\to\infty} \Delta_k = 0$ a.s. which, conditioning on I_k, 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, $\mathcal{R}_k \geq \eta_1$ for each $k \in K$ sufficiently large. Without loss of generality, let us assume that K contains only those sufficiently large k such that all the above holds. Then, an arbitrary iteration $k \in K$ under I_k is a successful iteration of SMOP. In other words, I_k implies S_k. Thus, due to (21), for each $k \in K$ there holds

$$\begin{aligned} & E(\phi(X_{k+1}) - \phi(X_k)|I_k, \mathcal{F}_k) \\ \leq\ & E(2c_f\Delta_k^2 - \frac{1}{2}w_m(X_k)\min\{\Delta_k, \frac{w_m(X_k)}{c_b}\}|I_k, \mathcal{F}_k) \\ \leq\ & 2c_f\Delta_k^2 - \frac{1}{2}\tilde{\varepsilon}\Delta_k = -\Delta_k(\frac{\tilde{\varepsilon}}{2} - 2c_f\Delta_k). \end{aligned}$$

Once again, assuming that $k \in K$ are all sufficiently large, we obtain $E(\phi(X_{k+1}) - \phi(X_k)|I_k, \mathcal{F}_k) \leq -c_9\Delta_k$, where $c_9 = \frac{\tilde{\varepsilon}}{4} > 0$, and thus we conclude

$$\begin{aligned} E(\psi_{k+1} - \psi_k|I_k, \mathcal{F}_k) &= E(\nu(\phi(X_{k+1}) - \phi(X_k)) + (1-\nu)(\gamma_2^2 - 1)\Delta_k^2|I_k, \mathcal{F}_k) \\ &\leq -c_{10}\Delta_k + c_{11}\Delta_k^2, \end{aligned}$$

where $c_{10} = \nu c_9 > 0$, and $c_{11} = (1-\nu)(\gamma_2^2 - 1) > 0$.

Now, let us consider $E(\Psi_{k+1} - \Psi_k|\bar{I}_k, \mathcal{F}_k)$. Considering (23) and (24) we conclude that

$$E(\phi(x_{k+1}) - \phi(x_k)|\bar{I}_k, \mathcal{F}_k) \leq c_5$$

and thus

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

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

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

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

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

On the other hand, (26) holds in all the iterations $k \in \mathbb{N}$ and applying the expectation we obtain

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

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

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

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

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

Letting m tend to infinity and using (25) 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 $\hat{\Omega} \subset \Omega$ such that $P(\hat{\Omega}) > 0$ and for all $v \in \hat{\Omega}$ there exist $\epsilon(v) > 0$ and $K(v) \subseteq \mathbb{N}$ such that for all $k \in K(v)$ there holds

$$\omega(X_k(v)) \geq 2\epsilon(v).$$

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

$$\omega(X_k(v)) < \epsilon(v) \quad \text{and} \quad \omega(X_{k+1}(v)) \geq \epsilon(v).$$

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

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

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

$$|\omega(X_{k_{j,1}}(v)) - \omega(X_{k_{j,2}}(v))| \geq \epsilon(v). \tag{30}$$

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

$$\omega(X_k(v)) \geq \epsilon(v).$$

Moreover, Proposition 1 implies that

$$\sum_{j=1}^{\infty} \sum_{i=k_{j,1}(v)+1}^{k_{j,2}(v)-1} \Delta_i(v) < \infty \tag{31}$$

holds for almost every $v \in \Omega$, and thus almost every $v \in \hat{\Omega}$. Notice that $k_{j,1}(v)$ must be successful iteration of SMOP for all $j \in \mathbb{N}$, since the marginal function changes only when the step is accepted, and thus $\Delta_{k_{j,1}+1}(v) = \gamma_2 \Delta_{k_{j,1}}(v) > \Delta_{k_{j,1}}(v)$. Therefore, for all $j \in \mathbb{N}$, we have

$$\begin{aligned}
& \|X_{k_{j,1}}(v) - X_{k_{j,2}}(v)\| \qquad (32)\\
= \; & \|X_{k_{j,1}}(v) - X_{k_{j,1}+1}(v) + X_{k_{j,1}+1}(v) - ... - X_{k_{j,2}}(v)\| \\
\leq \; & \sum_{i=k_{j,1}(v)}^{k_{j,2}(v)-1} \|X_i(v) - X_{i+1}(v)\| \leq \sum_{i=k_{j,1}(v)}^{k_{j,2}(v)-1} \Delta_i(v) = \Delta_{k_{j,1}}(v) + \sum_{i=k_{j,1}(v)+1}^{k_{j,2}(v)-1} \Delta_i(v) \\
\leq \; & \Delta_{k_{j,1}+1}(v) + \sum_{i=k_{j,1}(v)+1}^{k_{j,2}(v)-1} \Delta_i(v) \leq 2 \sum_{i=k_{j,1}(v)+1}^{k_{j,2}(v)-1} \Delta_i(v).
\end{aligned}$$

Thus, summing over j we conclude that $\sum_{j=1}^{\infty} \|X_{k_{j,1}}(v) - X_{k_{j,2}}(v)\| < \infty$ for almost every $v \in \hat{\Omega}$ due to (31). This further implies that $\lim_{j\to\infty} \|x_{k_{j,1}}(v) - x_{k_{j,2}}(v)\| = 0$ for almost every $v \in \hat{\Omega}$. However, this further implies that for almost every $v \in \hat{\Omega}$ we have $\lim_{j\to\infty} |\omega(X_{k_{j,1}}(v)) - \omega(X_{k_{j,2}}(v))| = 0$ due to Lemma 1, d), which is a contradiction with (30). ■

5 Numerical results

5.1 Experiment overview

Several experiments are reported in this paper in order to demonstrate the efficiency of the first order SMOP algorithm. The first set of experiments focuses on the machine learning (ML) application, and the concept of model fairness, as discussed in [27]. 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 sensitive features. Here, the SMOP algorithm from Section 3 is compared with the deterministic trust region [39] (DMOP), and the stochastic multi-gradient [27] (SMG). Additionally, a version SMOP-S which uses a subsampling technique which does not satisfy the theoretical assumption is considered. This version proves to be rather efficient as the theoretical bounds are rather demanding. The Pareto front finding technique, which can be seen in [18], [27] is also employed. This allows the comparison of SMOP/SMOP-S with a deterministic trust region DMOP [39], which is presented using different metrics.

The second set of experiments focuses on multi objective benchmark test problems from [37] and [18]. Four different problems are considered, each one having a differently shaped Pareto front. Comprehensive representation of the Pareto front for convex, disconnected, mixed, and concave shapes is provided. In this part, a thorough comparison has been made between stochastic and deterministic trust region, and several different metrics have been used to access the quality of the resulting Pareto front.

5.2 Experimental configuration

All considered experimental problems have two component functions that are in the form of a finite sum. Thus, the approximation of function and gradient values is achieved by exploiting its structure using an adaptive subsampling strategy. We demonstrate the first order algorithm, hence we assume that $H_k^i = 0$ for all $i = 1, ..., q$. Randomly, a subset $\mathcal{N}_i^k \subseteq \mathcal{N}_i$, $i = 1, 2$ is sampled following a rule motivated by the result from [33, 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

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

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.

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

In our tests, the upper-bound F_i is replaced by a constant, such that $|\mathcal{N}_i^0| = N_i^{min} = \max\{0.01|\mathcal{N}_i|, 2\}$, hence for SMOP the sample size behaves like

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

The probability parameters are defined as $\alpha_k = \sqrt{1 - 0.99^k}$, which satisfies theoretical assumptions. We have set $\delta_{min} = 10^{-4}$, $\delta_{max} = 8$, $\gamma_1 = 0.5, \gamma_2 = 2$ and $\delta_0 = 1$ in all experiments. This strategy showed large updates in subsampling size, due to δ_k^4 being present in the denominator, hence we introduced SMOP-S, a version of SMOP which controls the increases and decreases of the updates. The subsampling sizes of SMOP-S follow the update rule:

$$|\mathcal{N}_i^k| = \max\{\min\{j_k S_i, |\mathcal{N}_i|\}, N_i^{min}\}, \tag{33}$$

where $\delta_k^4 = 0.5^{j_k}$, i.e. at each iteration j_k is calculated as $j_k = \log_2 \frac{1}{\delta_k^4}$ for $\delta_k \leq 1$, and $S_i = |\mathcal{N}_i|/16$. For $\delta_k \geq 1$ the subsampling size is minimal, $N_i^{min} = \max\{0.01|\mathcal{N}_i|, 2\}$. This method connects the trust region radius to the stochastic average approximation error, that is, the approximation error follows the movement of δ_k.

The randomization is done by randomly and uniformly shuffling the indices $1, ..., |\mathcal{N}_i|$ without replacement at the start, and then slicing the first N_i^k elements at each iteration.

In Step 2 of Algorithm 1, a descent direction needs to be calculated using approximate function and gradient values. A common strategy in finding a descent direction in multi-objective optimization is to solve the problem

$$\min_{\beta\in\mathbb{R}, d\in\mathbb{R}^n} \beta + \frac{1}{2}\|d\|^2, \quad s.t. \quad \langle \nabla f_i(x_k), d\rangle \leq \beta, \quad i = 1, ..., q.$$

If x_k is Pareto critical, then the solution is $d_k = 0$, $\beta_k = 0$, and if it is not Pareto critical, then $\langle \nabla f_i(x_k), d_k\rangle \leq \beta < 0$ for all i, see [21, Lemma 1]. The dual of this problem can be written as

$$\min_{c_1,c_2\in\mathbb{R}} \|c_1\nabla f_1(x_k) + c_2\nabla f_2(x_k)\|^2, \quad s.t. \quad c_1, c_2 \geq 0, \quad c_1 + c_2 = 1,$$

see [21, Subsection 7]. This form offers an easily computable solution for two component functions, hence it is convenient for implementation. Since the

true gradients are unknown, by replacing these values with approximations, the dual problem becomes

$$\min_{c_1,c_2\in\mathbb{R}} \|c_1 g_1(x_k) + c_2 g_2(x_k)\|^2, \quad s.t. \quad c_1, c_2 \geq 0, \quad c_1 + c_2 = 1. \tag{34}$$

This dual problem produces a stochastic multi-gradient direction $d_k = -c_1^* g_1(x_k) - c_2^* g_2(x_k)$, see [27]. In all of our experiments, scaling this stochastic multi-gradient direction onto the trust region, i.e. using $\frac{d_k}{\|d_k\|}\delta_k$ as the direction, showed good performance. The direction calculated by solving (34) satisfies (10), since $\tilde{m}_k(d_k) \leq \tilde{m}_k(d_k^c)$, where d_k^c is the Cauchy direction defined in Lemma 2. Projecting it onto the trust-region may however only achieve a fraction of the Cauchy decrease (10).

When testing whether the second condition in Step 3 holds, i.e. $\omega_m(x_k) \geq \Theta\delta_k$, it is convenient to check instead if $-\max_i \langle g_i(x_k), \frac{d_k}{\|d_k\|}\rangle > \Theta\delta_k$. Since $\omega_m(x_k) \geq -\max_i \langle g_i(x_k), \frac{d_k}{\|d_k\|}\rangle$ holds, if the right hand side of the inequality is large enough, the desired condition is implied. This allows SMOP to skip the calculation of $\omega_m(x_k)$ entirely.

The implementation of SMG was done using the configuration showed in [27] and the code available on GitHub page [24] of the same authors. For DMOP we used the same configuration as in SMOP.

5.3 Finding the Pareto front

Using a front finding technique, [18], [27], we approximate the Pareto front for different problems. First, a set of random points $\mathcal{L}$ is taken as an approximation of the front. At each iteration, it is expanded by generating perturbed points around the existing elements in the set. The predetermined number of iterations of the chosen algorithm (SMOP, SMOP-S or DMOP) is then applied to the existing points, in order to come closer to the set of optimal points, after which the results are also added to the approximation set $\mathcal{L}$. To refine the approximation, all dominated points are removed from the set, leaving only the non dominated points to serve as the Pareto approximation. The 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 formally described in the following way:

Algorithm 2.[18]
(Pareto front procedure)

Step 0. Generate the initial 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 chosen method with x as a starting point. Add the final iteration to $\mathcal{L}_{k+1}$

Step 3. Remove all dominated points from $\mathcal{L}_{k+1}$. Set $k = k + 1$ and go to Step 1.

The only difference between the procedures is in Step 2, where we choose either SMOP, SMOP-S or DMOP as the underlying algorithm. All three procedures were implemented; however, since the SMOP-S version was superior to SMOP in terms of time while yielding similar quality fronts, only the comparison between the procedure with SMOP-S and DMOP will be presented and further discussed. By selecting different configurations n_p, n_q and n_r, it is possible to control the resulting front and get either a sparser front with fewer details or a denser one. In our experiments, the following parameters were set the same for SMOP-S and DMOP: $n_q = 5, n_p = 1, n_r = 10$. In order to reduce the number of points generated each iteration, a strategy from [27] is employed. At Step 1, n_r points is generated only for the pair of points with largest holes in the Pareto front. This requires sorting values of both component functions, and finding the original indices corresponding to the largest differences of consecutive function values in the sorted list.

The starting Pareto size was always 30 points, while the exiting criteria was set to be 1500 points, which was always the terminating factor. In both procedures, the true values of the functions were available to determine whether the points are dominated or not in Step 3, however in SMOP-S procedure, when applying n_q iterations in Step 2, the true function values were not visible, i.e. the n_q steps were calculated using the approximated values. For all problems, we measured the full time needed to find the Pareto front. Since the trust region radius was initialized as $\delta_0 = 1$ everywhere, after 5 iterations of each procedure, it is halved. This helped both SMOP-S and DMOP reach higher quality fronts.

To measure the quality of the approximated front, we have opted for three different metrics, as seen in [18], [27]. The Purity estimates the percentage of true nondominated points in the Pareto approximation. It does so by comparing the observed front to the combined front of all available Pareto approximations . If $\mathcal{L}_S$ and $\mathcal{L}_D$ are fronts generated by SMOP-S and DMOP respectively, and $\mathcal{L}_{SD}$ is the front we get by removing the nondominated points from $\mathcal{L}_S \cup \mathcal{L}_D$, then the Purity is given by:

$$P_S := \frac{|\mathcal{L}_S \cap \mathcal{L}_{SD}|}{|\mathcal{L}_S|}, \quad P_D := \frac{|\mathcal{L}_D \cap \mathcal{L}_{SD}|}{|\mathcal{L}_D|}. \tag{35}$$

The closer to 1 this ratio is the better for the procedure, since a higher percentage of points is 'truly' on the Pareto front. The Spread metrics are used to determine how well the points are spread throughout the Pareto

front approximation, as the name suggests. The Γ-spread metric measures the maximum size of the hole in the Pareto front. To calculate it the function values within the front approximations are sorted in a nondecreasing order for both components, $f_i(x^0) \leq ... \leq f_i(x^{|\mathcal{L}|})$. The measure is then calculated as:

$$\Gamma_S := \max_{i=1,2} \big(\max_{j=1,...,|\mathcal{L}_S|} l_{i,j} \big), \quad \Gamma_D := \max_{i=1,2} \big(\max_{j=1,...,|\mathcal{L}_D|} l_{i,j} \big) \tag{36}$$

where $l_{i,j} = f_i(x^{j+1}) - f_i(x^j)$. The second Δ-spread metric measures how well the points are distributed in the approximated front. It is computed as:

$$\Delta_S := \max_{i=1,2} \Big(\frac{l_{i,0} + l_{i,|\mathcal{L}_S|} + \sum_{j=1}^{|\mathcal{L}_S|} |l_{i,j} - \bar{l}_i|}{l_{i,0} + l_{i,|\mathcal{L}_S|} + (|\mathcal{L}_S| - 1)\bar{l}_i} \Big), \tag{37}$$

$$\Delta_D := \max_{i=1,2} \Big(\frac{l_{i,0} + l_{i,|\mathcal{L}_D|} + \sum_{j=1}^{|\mathcal{L}_D|} |l_{i,j} - \bar{l}_i|}{l_{i,0} + l_{i,|\mathcal{L}_D|} + (|\mathcal{L}_D| - 1)\bar{l}_i} \Big) \tag{38}$$

where $\bar{l}_i$ is the mean of $l_{i,j}$ for the respective procedure. For both Spread metrics, the lower the value, the better the distribution of the Pareto front is.

5.4 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. 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. 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 are more balanced. For more on this topic, and different ways to measure fairness, see [3],[23], [41],[42],[43].

The problem we consider is minimization of a regularized logistic regression loss function, as in [27]:

$$\min_x f(x) := \frac{1}{N} \sum_{j \in N} \log(1 + e^{(-y_j(x^T a_j))}) + \frac{\lambda}{2} \|\hat{x}\|^2, i = 1, 2. \tag{39}$$

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, N the training set size, and λ is the regularizer.

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 G_1 and G_2 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 $i-th$ and hence the problem becomes:

$$\min_x (f_1(x), f_2(x))^T \tag{40}$$

In the experiment the regularization was set to $\lambda_i = 10^{-3}$, for $i = 1, 2$, as in [27]. We consider six datasets which vary in size, namely two larger: covtype and mnist, and four smaller sets: svmguide, german, australian and heart, see [15]. The baseline unfairness is demonstrated by solving the scalar problem (39) and evaluating the logistic regression model on the entire dataset, and on groups G_1 and G_2 separately. Minimization of (39) was done in Python, using stochastic gradient descent, with maximum of 1000 iterations, and diminishing step size. The following Table 1 showcases tested unfairness, and how different Pareto points can create fair models. The "Split" column indicates the critical column of the dataset, by which the split was made. "Full" column shows the accuracies of the scalar logistic regression, trained on the full dataset, and evaluated on the training set. The column "G_1/G_2" represents the accuracies of of the same model evaluated on groups G_1 and G_2 respectively. We also present the accuracies of the multi-objective logistic regression models associated with three representative Pareto critical points produced by the SMOP-S. For each model, the training accuracy is evaluated separately for each group, and the results are reported in distinct rows. It can be seen that moving around the Pareto front reduction or increase of disparities between groups is possible for the tested datasets and hence desired group accuracies can be achieved. For larger datasets, the front was more difficult to evaluate, hence for covtype dataset, a minimal reduction of disparity is achieved. The split for australian, heart, svmguide and german was done using the known split technique, as in [27], however for mnist and covtype the split has been done by experimentally finding a column which creates groups with accuracy disparity.

As mentioned previously, we first tested SMOP, against DMOP, SMG and SMOP-S in finding critical points. The performance of algorithms is evaluated by measuring the value of the true marginal function $\omega(x_k)$ (2) in terms of the number of function evaluations FEV and time in seconds. In SMOP, at each iteration we evaluate the function approximation two times, in point x_k and $x_k + d_k$, hence for each function evaluation of $\tilde{f}$, we account $|\mathcal{N}_1^k| + |\mathcal{N}_2^k|$ evaluations. For SMG, at each iteration the number of function evaluation depends on the line-search technique, and achieving the sufficient descent. Both DMOP and SMG use full function values, hence for these

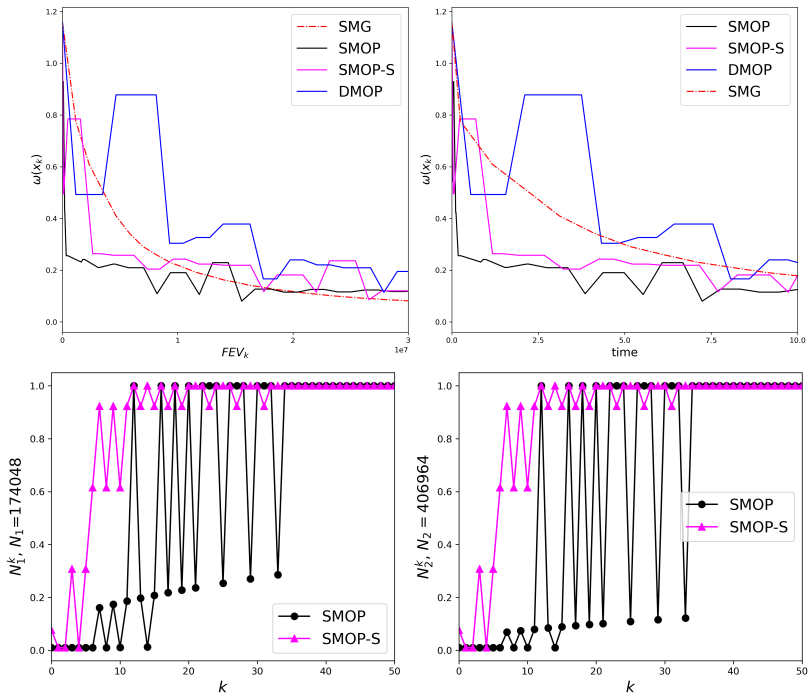
	N_1/N_2	Split	Full	G_1/G_2	P_1	P_2	P_3
covtype	174048 406964	1	92.4%	84.6% 95.8%	83.5% 96.1%	83.8% 95.8%	84.8% 94.8%
mnist	13919 47	106	97.6%	98.0% 91.5%	97.5% 93.6%	97.4% 95.7%	97.1% 97.5%
svmguide3	1182 61	10	82.1%	83.5% 55.7%	83.2% 65.5%	80.1% 73.7%	74.3% 75.4%
german	630 370	24	78.3%	80.4% 74.5%	80.4% 76.2 %	79.2% 75.9%	78.8% 77.2%
australian	468 222	1	86.5%	84.8% 90.1%	85.1% 91.4%	86.5% 90.1%	86.9% 89.1%
heart	183 87	2	85.5%	81.9% 93.1%	79.7% 94.1%	82.5% 91.3%	83.1% 90.8%

Table 1: Dataset parameters and classification accuracies.

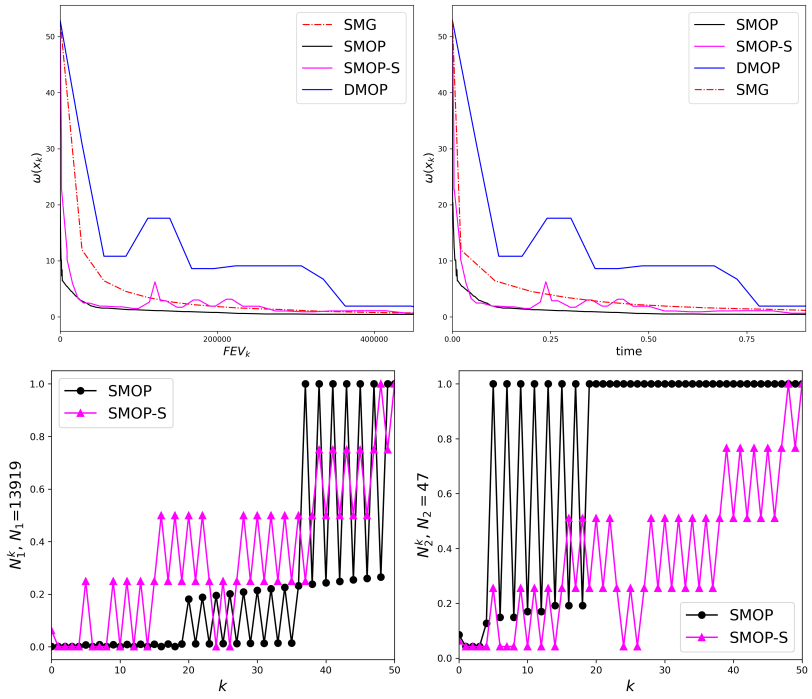
algorithms $|\mathcal{N}_1|+|\mathcal{N}_2|$ is added when the function is evaluated. The number of function evaluation showed to be a dominant factor for large datasets and the most time consuming operation for all algorithms, which can be seen in figures for mnist and covtype. However, for smaller datasets, FEV wasn't the only significant operation. For example, the difference between SMOP and DMOP, was also in the acceptance criteria in Step 3. of Algorithm 1, which involves additional scalar products in the stochastic case.

In this manner, we demonstrate how SMOP and SMOP-S algorithms achieve significant improvements in performance while utilizing only a fraction of the available data for large datasets, and stay comparable with DMOP in the case of smaller datasets. This highlights the stochastic algorithm effectiveness in leveraging limited resources, making it particularly valuable in scenarios where data collection is expensive. The parameters for SMOP and SMOP-S have been chosen the same for all problems: $\eta_1 = 0.25, \Theta = 0.25$, and the starting point $x_0 = (0.1, 0.1, ..., 0.1)$, while the rest of the parameters can be seen in Experiment configuration subsection. The following Figures 1, 2 and 3 show the behavior of the four algorithms for the datasets from Table 1. For large datasets, Figure 1 a) and b), the time agrees with the FEV measure. Here, both SMOP and SMOP-S perform efficiently in terms of FEV and time. For datasets seen in Figure 2 c) and d), the advantage is visible mostly in terms of FEV. It can be seen that for smallest datasets, Figure 3 e) and f), both versions of SMOP demonstrate slight advantage in terms of FEV and time at start, but later slow down. The time needed to reach near optimality for all algorithms for c)-f) is less than 0.005 seconds. The subsampling sizes of SMOP show considerable changes, unlike SMOP-S, where the growth is more controlled.

Further, a Pareto front finding technique from the previous section was

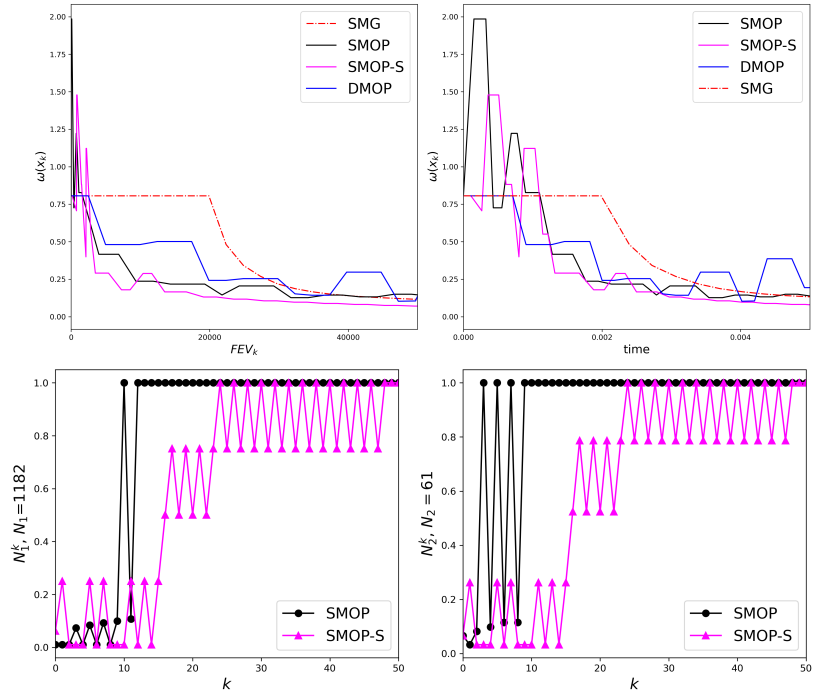


a)

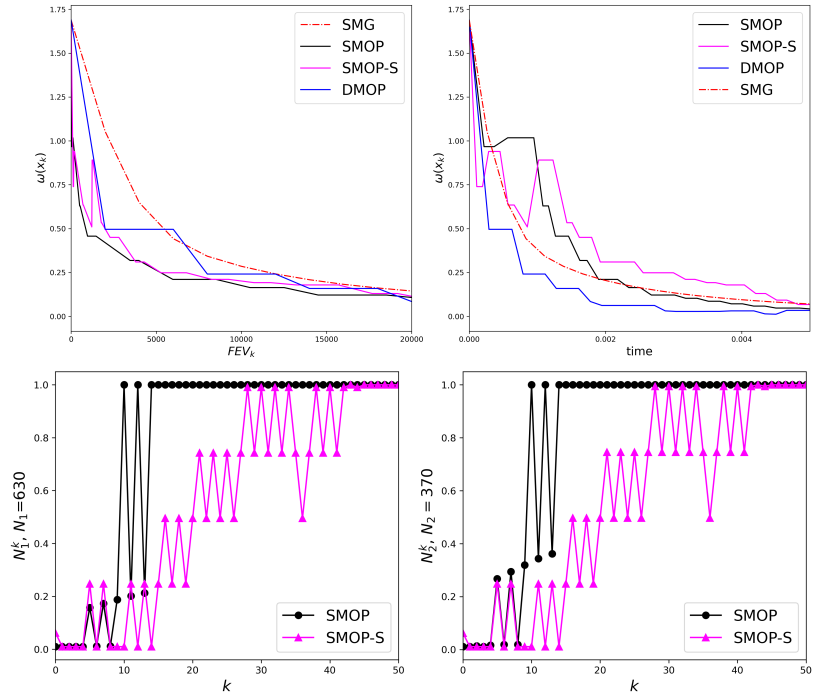


b)

Figure 1: Performance comparison - $\omega(x_k)$ in terms of FEV/time, subsample sizes through iterations for datasets: a) covtype, b) mnist.

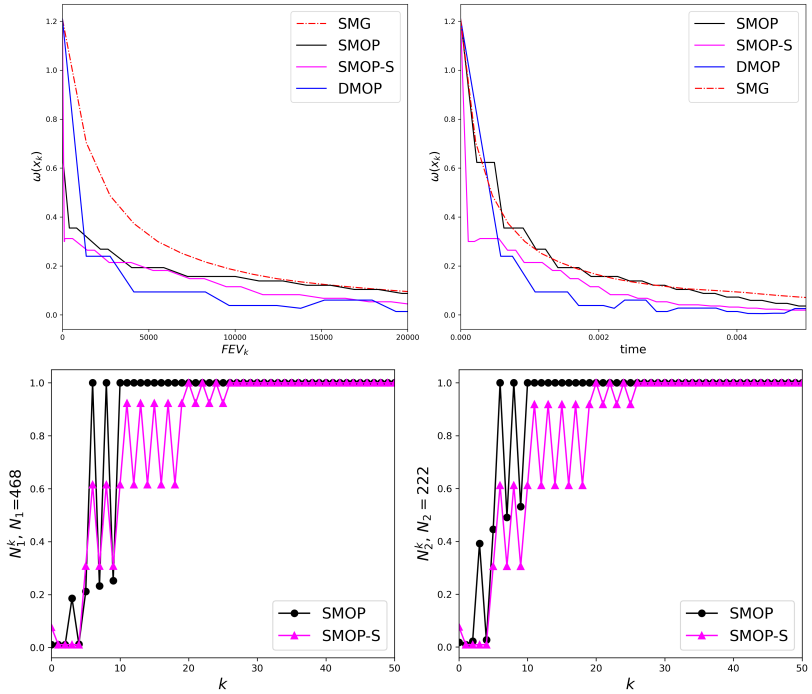


c)

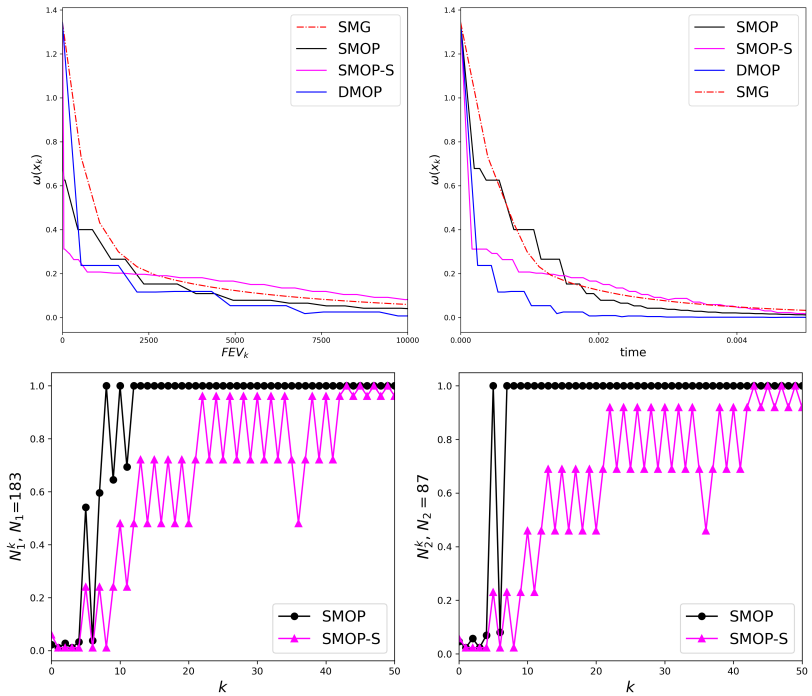


d)

Figure 2: Performance comparison - $\omega(x_k)$ in terms of FEV/time, subsample sizes through iterations for datasets: c) svmguide, d) german.



e)



f)

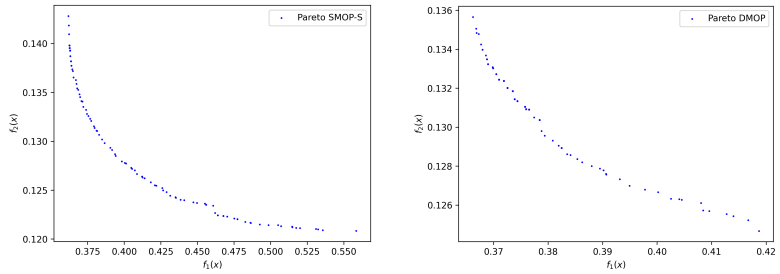
Figure 3: Performance comparison - $\omega(x_k)$ in terms of FEV/time, subsample sizes through iterations for datasets: e) australian, f) heart.

employed, and the comparison was made between the SMOP-S and DMOP procedures. We provide a table with the results, showing Purity, Γ and Δ spread, together with the sizes of the approximated fronts, and the time needed in seconds for finding the front. In order to show fair values, all presented results are the averages of 5 simulations, except for the covtype, which was calculated once due to high dimensions. For covtype the starting Pareto front was 3000 random points, instead of 30, and n_r was changed to 100. For mnist, we do not provide the Pareto front analysis, as it was to large; however the Pareto critical points we get from running SMOP-S multiple times provide us fair models, which can be seen in Table 1. From Table 2, we can see that most of the time DMOP produces higher quality approximations of the Pareto front, having larger Purity, and smaller Spread metrics. However, SMOP-S approximated the front more efficiently in terms of time, with DMOP requiring 2, 1.1, 1.4, 1.2, and 0.9 times the execution time of SMOP-S across the respective datasets. For heart dataset, DMOP was faster, however for the other problems SMOP-S had time advantage. From tests, we have noticed that DMOP procedure makes fewer iterations which add more points per iteration.

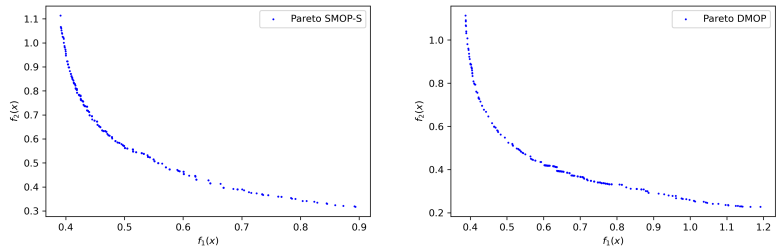
	Algorithm	Purity	Γ	Δ	$\lvert\mathcal{L}\rvert$	time (s)	#iter
covtype	SMOP-S	0.96	0.02	1.88	2709	31491.12	105
	DMOP	0.99	0.001	1.89	1878	62301.16	66
svmguide3	SMOP-S	0.92	0.044	1.79	1728	5.58	26
	DMOP	0.98	0.024	1.78	1878	6.12	18
german	SMOP-S	0.98	0.006	1.61	2088	4.68	26
	DMOP	0.99	0.005	1.87	1995	6.56	17
australian	SMOP-S	0.91	0.005	1.67	1831	5.24	27
	DMOP	0.97	0.004	1.81	1788	6.15	23
heart	SMOP-S	0.93	0.009	1.80	1947	2.11	35
	DMOP	0.92	0.017	1.87	2145	1.92	15

Table 2: Comparison metrics between resulting Pareto front approximations from SMOP and DMOP on different datasets

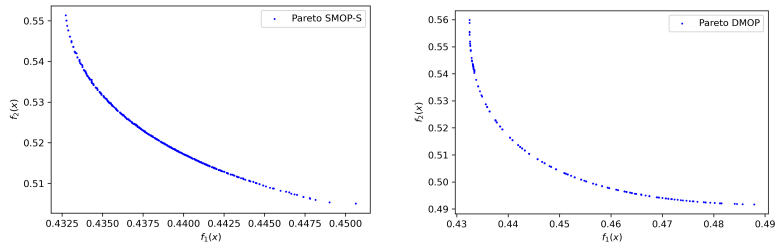
Additionally, we show Pareto front approximations for both SMOP-S and DMOP in Figure 4. The tests showed the SMOP-S iterations are faster in the beginning, and slow down as the points go closer to the Pareto front, since the trust region radius approaches 0 in later iterations. For homogeneous problems, these fast iterations can be explained as generating points closer to the front in early iterations, whereas for heterogeneous problems, these fast approximations act as an additional randomization factor which perturbs the points, and not necessarily improving the front.



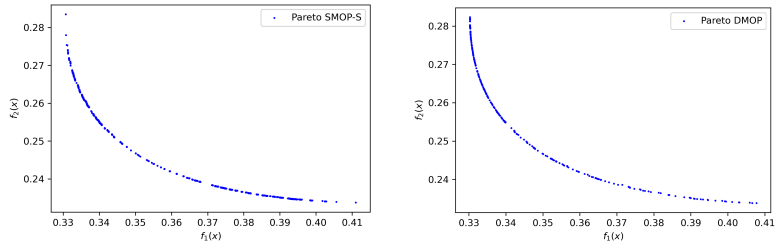
a)



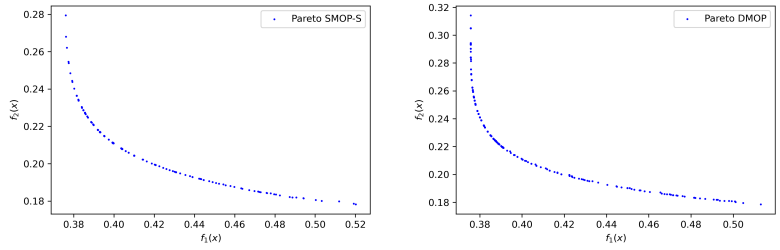
b)



c)



d)



e)

Figure 4: Pareto front approximations using SMOP (left) and DMOP (right): a) covtype b) svmguide c) german d) australian, e) heart,

5.5 Test problems with added noise

The second set of experiments consists of unconstrained multi objective problems gathered from [18] and [37]. The presented problems are low dimension and they differ in terms of the shape of the Pareto front. Namely, four shapes are presented: convex, concave, mixed (neither convex nor concave) and disconnected. The problem we are solving is the following:

$$\min_x f(x) = \min_x (f_1(x), f_2(x))^T \tag{41}$$

In order to simulate stochasticity we choose to perturb points at each iteration, as done in [27],[28], which turns the problem into

$$\min_x F(x) = \min_x (E[f_1(x+\omega)], E[f_2(x+\omega)])^T \tag{42}$$

where ω is a random uniform vector with mean zero. Using Stochastic Average Approximation, (42) is transformed into the finite sum problem.

$$\min_x F(x) \approx \min_x (\frac{1}{N}\sum_{i=1}^{N} f_1(x+\omega_i), \frac{1}{N}\sum_{i=1}^{N} f_2(x+\omega_i))^T \tag{43}$$

where ω_i are i.i.d. samples of the random variable ω. Similarly as in the previous experiment, we test SMOP-S against DMOP Pareto front finding procedure on the problem (43). The SMOP-S configuration follows the one explained in previous sections. For each problem, we generate $N = 500$ i.i.d. uniformly sampled vectors with mean zero and length of the interval equal to 0.1. The following Table 3 showcases the basic information for each considered problem. Exponential functions and its derivatives in FF1 and T2 are considered harder to calculate [37], hence the evaluation of these functions was a time consuming factor, similarly as in the logistic regression.

	n	Functions	Shape
SP1	2	$f_1(x) = (x_1-1)^2 + (x_1-x_2)^2$ $f_2(x) = (x_2-3)^2 + (x_1-x_2)^2$	convex
SK1	1	$f_1(x) = x^4 + 3x^3 - 10x^2 - 10x - 10$ $f_2(x) = 0.5x^4 - 2x^3 - 10x^2 + 10x - 5$	disconnected
FF1	2	$f_1(x) = 1 - e^{-(x_1-1)^2-(x_2+1)^2}$ $f_2(x) = 1 - e^{-(x_1+1)^2-(x_2-1)^2}$	concave
T2	2	$f_1(x) = \sin(x_2)$ $f_2(x) = 1 - e^{-(x_1-\frac{1}{\sqrt{2}})^2-(x_2-\frac{1}{\sqrt{2}})^2}$	mixed

Table 3: Problem description for (43)

As in the previous experiments, the results show that the DMOP generates fronts with better purity. Nonetheless, SMOP-S demonstrated the

ability to find the front more quickly with high enough quality. The average metrics of 5 simulations can be seen in Table (4).

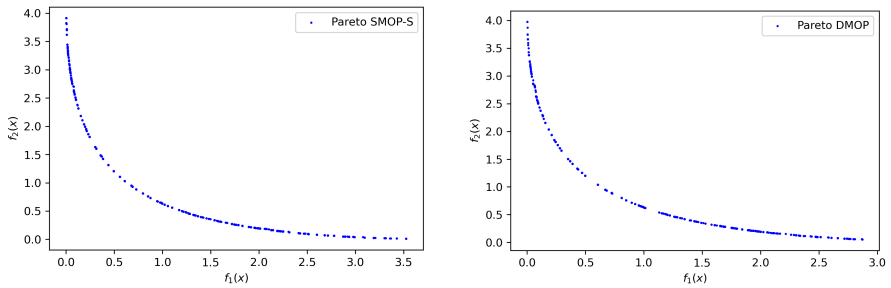
	Algorithm	Purity	Γ	Δ	$\lvert\mathcal{L}_k\rvert$	time (s)	#iter
SP1	SMOP-S	0.95	0.17	1.86	2475	38.4	6
	DMOP	0.96	0.16	1.84	2377	81.1	5
SK1	SMOP-S	0.99	24.68	1.86	2434	184	4
	DMOP	1.00	29.25	1.76	1682	253	4
FF1	SMOP-S	0.94	0.9	1.82	2208	42.4	6
	DMOP	0.95	0.07	1.80	1958	97.0	5
T2	SMOP-S	0.91	0.05	1.84	2953	84.1	5
	DMOP	0.94	0.05	1.86	3303	208.1	6

Table 4: Performance profile for problems from Table 3

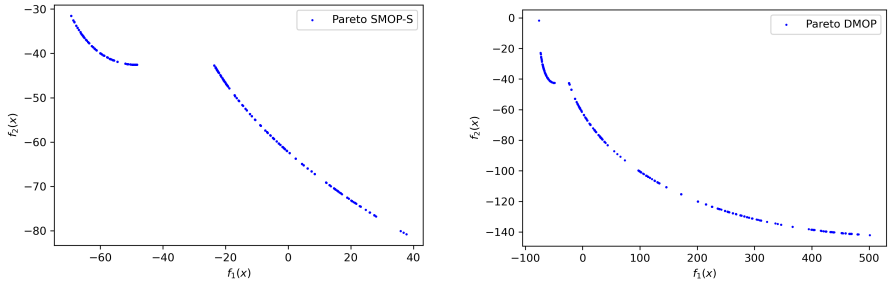
The Purity of the fronts generated by SMOP-S is lower than DMOP, which implicates that SMOP-S generated some points which were dominated by the those generated by DMOP. The Spread confirms that the gaps in SMOP-S fronts are larger in most problems. For problem SK1 the Γ spread is large since the front is disconnected, and it takes the disconnected part into the calculation. The average time needed to calculate the front showed that SMOP-S procedure was 2.1, 1.3, 2.2, and 2.5 times faster than DMOP respectively. The following Figure 5 illustrates the approximate Pareto front for both procedures.

6 Conclusion

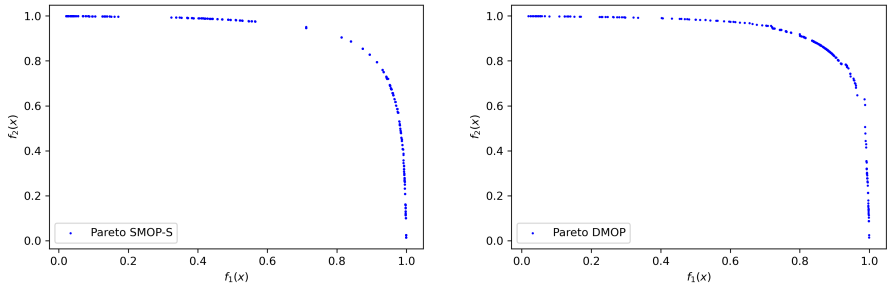
We have proposed a trust region method for multiobjective problems based on probabilistically fully linear models for each function. The method is designed to operate under the key assumptions that the approximation accuracy of per function models is achieved with high enough probability. The concept of Jointly Independent Fully Linear Probabilistic models is introduced to deal with the fact that the objective scalarization function is nonsmooth and hence the concept of full linearity can not be extended. However we prove that probabilistically fully linear models per function yield a satisfactory random model for a nonsmooth scalarization function ϕ. The theoretical contribution of this work is the proof of almost sure convergence to a Pareto critical point. Possible applications go beyond the scope of multiobjective optimization. We presented several numerical experiments that showcase the algorithm's efficient practical performance, especially for large scale problems and large data sets. Additionally, we implemented a Pareto finding routine, and made a thorough comparison between the stochastic and deterministic approach. Future work could include the generalization of fully quadratic models or techniques such as additional sampling.



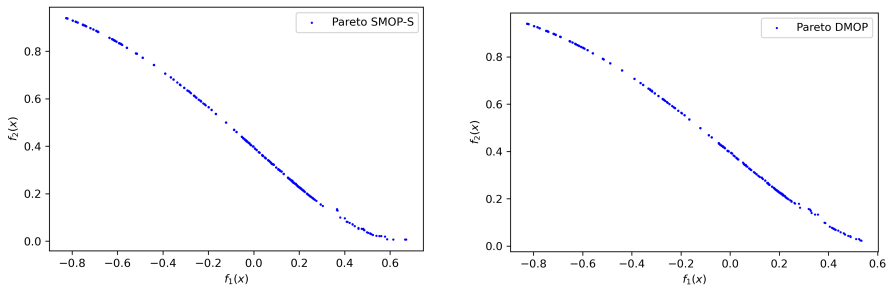
a)



b)



c)



d)

Figure 5: Pareto front approximations using SMOP (left) and DMOP (right) on different problems from Table 3: a) SP1, b) SK1, c) FF1, d) T2.

Acknowledgments

We are very grateful to the anonymous referees whose insightful comments helped us a lot to improve the paper.

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] Bannert, T. (1994) A trust region algorithm for nonsmooth optimization. *Mathematical Programming 67, 247–264*

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

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

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

[6] Berahas, A.S., Curtis, F.E., Robinson, D. & Zhou, B. (2021) Sequential Quadratic Optimization for Nonlinear Equality Constrained Stochastic Optimization *SIAM Journal on Optimization*

[7] 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.*

[8] 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.*

[9] Blanchet, J., Cartis, C., Menickelly, M. & Scheinberg, K. Convergence Rate Analysis of a Stochastic Trust Region Method for Nonconvex Optimization *https://web.stanford.edu/~jblanche/papers/Stochastic_Trust_Region.pdf.*

https://github.com/sul217

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

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

[12] 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*

[13] 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.*

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

[15] Chang, C. C, & Lin. C. J. (2011) LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST), 2:27*

[16] Conn, A.R., Scheinberg, K. & Vicente, L.N. (2009) Introduction to Derivative-Free Optimization. *Society for Industrial and Applied Mathematics, Philadelphia, PA, US*

[17] 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.*

[18] 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.*

[19] Davar, D. & Grapiglia, G. N. TRFD: A Derivative-Free Trust-Region Method Based on Finite Differences for Composite Nonsmooth Optimization *SIAM Journal on Optimization, 35, 3, 1792 - 1821*

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

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

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

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

[24] S. Liu and L. N. Vicente. (2020) Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach.ISE Technical Report 20T-016, Lehigh University, 2020. *https://github.com/sul217/MOO_Fairness*

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

[26] 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.*

[27] 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.*

[28] Mercier, Q., Poirion, F. & Desideri, J.A. (2018) A stochastic multiple gradient descent algorithm. *European Journal of Operational Research, pp.10.*

[29] *Orvieto, A., Lacoste-Julien, S. & Loizou, N.* Dynamics of SGD with Stochastic Polyak Stepsizes: Truly Adaptive Variants and Convergence to Exact Solution

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

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

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

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

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

[35] Tran-Dinh, Q., Pham, N.H., Phan, D.T. & Nguyen, L.M. (2022) A hybrid stochastic optimization framework for composite nonconvex optimization. *Math. Program. 191, 1005–1071* (2022).

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

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

[38] 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.*

[39] 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.*

[40] Wen, M., Zhang, Y., Tang, Y., Cui., A. & Peng, J. (2025) A stochastic primal–dual algorithm for composite optimization with a linear operator *Expert Systems with Applications, 267*

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

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

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