ASMOP: Additional sampling stochastic trust region method for multi-objective problems

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

June 12, 2025

Abstract

We consider an unconstrained multi-criteria optimization problem with finite sum objective functions. The proposed algorithm belongs to a non-monotone trust-region framework where additional sampling approach is used to govern the sample size and the acceptance of a candidate point. Depending on the problem, the method can result in a mini-batch or an increasing sample size approach. Therefore, this work can be viewed as an extension of additional sampling trust region method for scalar finite sum function minimization presented in the literature. We show stochastic convergence of the proposed scheme for twice continuously-differentiable, but possibly non-convex objective functions, under assumptions standard for this framework. The experiments on logistic regression and least squares problems show the efficiency of the proposed scheme and its competitiveness with the relevant state-of-the-art methods for the considered problems.

Key words: Additional sampling, non-monotone trust region, adaptive sample size, multi-objective optimization, Pareto critical points, stochastic convergence.

1 Introduction

Machine learning (ML) and deep learning (DL) have been widely researched in optimization, and many crafty methods have been created to solve problems that arise in the domain. These problems are often nonlinear, nonconvex, and large-scale; hence it is an important task to create algorithms which

[*]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

can find the solution in an efficient manner. Multi-objective optimization problems arise in ML and pose a great computational and logistic challenge. Utilizing the stochastic approach, in the multi-criteria scenarios, showed to be convenient considering the time and cost of obtaining the data and finding the optimal result. Moreover, in ML applications often we encounter finite sum problems, which due to their specific form allow us to employ various subsampling techniques. This is crucial for our work, as it motivates us to find a way to reduce costs by exploiting the problem's structure. The problem we are solving can be stated as

$$\min_{x\in\mathbb{R}^n} f(x) := (f^1(x), ..., f^q(x)) \tag{1}$$

where $f : R^n \to R^q$ and each component function is assumed to be smooth with Lipschitz-continuous gradients. Let $\mathcal{N}^i$, $i = 1, ..., q$ be the respective index sets and $N^i = |\mathcal{N}^i| = N$, for all $i = 1, ..., q$. We assume that each component function has the following finite sum form

$$f^i(x) := \frac{1}{N} \sum_{j\in\mathcal{N}^i} f^i_j(x),\ i = 1, ..., q. \tag{2}$$

In ML terms, each component function $f^i(x)$ can be seen as a distinct average loss function, where $x \in R^n$ is the vector of trainable parameters for input-label pairs $\{(a^i_j, y^i_j)\}^N_{j=1}$ of the training dataset, i.e.,

$$f^i_j(x) := L^i(a^i_j, y^i_j; x), \quad j = 1, .., N, i = 1, ..., q,$$

where $L^i(\cdot)$ measures the prediction error.

Since we are solving a multi-criteria problem, the points of interest are Pareto critical points which cannot be locally improved in terms of all component function values [12],[17]. Finding Pareto critical points yields possibility of finding an entire Pareto front - a set of globally optimal points [18]. This is extremely important as in some applications the representation of the entire front can provide crucial information. Pareto optimal (critical) points can be characterized as zeros of the marginal functions (see [12] for details). In the scalar case this concept is reduced to the well known first order optimality conditions, i.e., finding a stationary point where gradient is equal to zero. Furthermore, in the stochastic setup, where only approximate values of the functions and the gradients are available, the approximate marginal function with the corresponding scalar representation of the problem plays a significant role [13],[11]. The common result within this framework is (stochastic) convergence of the marginal function to zero.

Both the line search and the trust region approach has been researched in multi-objective optimization, resulting in a number of deterministic and stochastic algorithms. Deterministic multi-criteria steepest descent and

Newton method, together with a projected gradient method for the constrained case has been discussed in [26]. The stochastic multi-gradient, an extension of the classical stochastic gradient SG [16], can be found in [22], in which sublinear convergence for convex and strongly convex functions is shown.

In [14], the marginal function is utilized to define the trust region method, and therein a convergence towards a critical point is shown. The complexity of multi-objective problems motivates the development of the stochastic and derivative free approaches. Models within the trust region framework can utilize the inexact information, which can save time and reduce the computational cost with adequate methods. In [13] one criterion is a black box function, where the derivative is unknown. Therein it is showed that using true function values and gradient approximations to create models yields a convergence to a Pareto optimal point. It is also possible to use approximate function and gradient values if the estimates are sufficiently accurate with high probability (probabilistically fully linear), see [11], which is a generalization of [15]. Therein, an adaptive subsampling technique is used which depends on the trust region radius. It successfully reduced the computational cost by using less data when the radius is larger.

The literature also provides methods designed for problems with the finite sum objective functions. These methods exploit the structure of the function and their advantage lies in the subsampling techniques. It is shown that subsampling can help in reducing the costs of deterministic schemes where the full sample set is needed at all iterations, yielding excessive optimization costs. Some papers on this topic are [1],[2],[3],[4],[5],[6],[10]. In [7] an additional sampling technique is employed within a non-monotone trust-region framework, aiming to solve single-criterion problems. The idea of non-monotonicity withing trust region can be found in [19],[20] and [21]. In [19] fixed size subsampling batches are proposed, whereas in [20] a relaxed trust region ratio is utilized. The idea of relaxed trust region conditions in a stochastic setting is what we will also benefit from. Our work will extend the approach of [7] to vector functions, as we will explain further on. The additional sampling technique which is prominent here, is also discussed in [23],[25], and [24] in different framework and settings.

In this work, we propose a stochastic trust region algorithm for solving multiobjective problems. At each iteration we employ subsampled functions and gradients to find a candidate subsequent point. The acceptance of that point is based on additional sampling technique which also governs the subsampling strategy. This results in an adaptive subsampling technique. Knowing that we are dealing with noisy approximations, we do not pose strict monotonicity of the objective function through iterations, as mentioned. Similar as in [7] we rely on additional sampling as an independent controlling factor, and we utilize it to determine the next subsampling set and iteration. This means that together with model and approximate ob-

jective function decrease, the behavior of independent subsampled functions is monitored and used as a decision making criterion. By adaptively choosing the sample size for each component, we also handle the sample average approximation error for each function. Consequently, this leads to two different sample size scenarios: 1) mini-batch scenario where at least one of the objective functions is inaccurate during the whole optimization process, i.e., the full sample is not reached; 2) full sample scenario where the full sample is reached for all the objective functions eventually. Regardless of the scenario, we prove almost sure convergence to a Pareto critical point under some standard assumptions for the stochastic framework.

The paper is organized as follows. Some basic concepts are covered in the following section. Section 3 presents the proposed algorithm. Within Section 4 the stochastic convergence of the proposed method is analyzed, while Section 5 is devoted to numerical results. Some conclusions are drawn in Section 6.

2 Preliminaries

We start this section by defining efficient and weakly efficient solutions of problem (1).

Definition 1. [12] 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 $i \in \{1, 2, ..., q\}$.

Thus, a Pareto point is such that for every direction $d \in \mathbb{R}^n$, there exists a function $f^i(x)$ with a nonnegative directional derivative in that direction d, i.e., $\langle \nabla f^i(x^*), d \rangle \geq 0$. The scalar problem's stationarity condition - gradient equal to zero - is replaced with another metric related to marginal function defined as follows

$$\omega(x) = -\min_{\|d\| \leq 1} \left(\max_{i \in \{1,...,q\}} \langle \nabla f^i(x), d \rangle \right). \tag{3}$$

The marginal function generalizes the gradient norm in multi-objective settings - notice that $\omega(x) = \|\nabla f(x)\|$ if $q = 1$ since the solution of problem (3) is $d^{opt}(x) = \nabla f(x)/\|\nabla f(x)\|$ in that case. In general, marginal function characterizes Pareto critical points as stated in the following lemma.

Lemma 1. [12] Let $\mathcal{D}(x)$ be the set of solutions of (3). Then

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

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

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

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

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

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

This problem is not equivalent to problem (1), but it can be shown that every solution this problem is a Pareto optimal point of problem (1).

Since we are dealing with finite sum objective functions (2), at each iteration we form a sample average approximation functions and the corresponding gradients as follows

$$f^i_{\mathcal{N}^i_k}(x) = \frac{1}{N^i_k} \sum_{j\in\mathcal{N}^i_k} f^i_j(x), \quad \nabla f^i_{\mathcal{N}^i_k}(x) = \frac{1}{N^i_k} \sum_{j\in\mathcal{N}^i_k} \nabla f^i_j(x), \tag{4}$$

where $\mathcal{N}^i_k \subseteq \mathcal{N}^i$ and $N^i_k = |\mathcal{N}^i_k|$. Given that we work with approximate functions, we consider the approximate marginal functions [13]

$$\omega_{\mathcal{N}_k}(x) = -\min_{\|d\|\leq 1} \left(\max_{i\in\{1,...,q\}} \langle \nabla f^i_{\mathcal{N}^i_k}(x), d\rangle \right). \tag{5}$$

where $\mathcal{N}_k = (\mathcal{N}^1_k, ..., \mathcal{N}^q_k) \subseteq \mathcal{N} = (\mathcal{N}^1, ..., \mathcal{N}^q)$ is the set q-tuple. The corresponding scalar problem is then given by

$$\min_{x\in\mathbb{R}^n} \phi_{\mathcal{N}_k}(x), \quad \phi_{\mathcal{N}_k}(x) := \max_{i\in\{1,...,q\}} f^i_{\mathcal{N}^i_k}(x). \tag{6}$$

In deterministic second order trust region framework, the quadratic model of $\phi(x)$ is given by

$$m_k(d) := \max_{i\in\{1,...,q\}} \{f^i(x_k) + \langle \nabla f^i(x_k), d\rangle + \frac{1}{2}\langle d, H^i_k d\rangle\},$$

where H^i_k is a Hessian approximation of the respective component function. In general, we will use approximate functions and the gradients as well and thus we relate our quadratic model to (6) as follows

$$m_{\mathcal{N}_k}(d) = \max_{i\in\{1,...,q\}} m_{\mathcal{N}^i_k}(d), \tag{7}$$

$$m_{\mathcal{N}^i_k}(d) := f^i_{\mathcal{N}^i_k}(x_k) + \langle \nabla f^i_{\mathcal{N}^i_k}(x_k), d\rangle + \frac{1}{2}\langle d, H^i_k d\rangle.$$

Notice that for each $i = 1, ..., q$, $\nabla m_{\mathcal{N}^i_k}(0) = \nabla f^i_{\mathcal{N}^i_k}(x_k)$, and $m_{\mathcal{N}^i_k}(0) = f^i_{\mathcal{N}^i_k}(x_k)$. In SMOP [11], we assumed that the approximations are accurate

enough with a high probability. This allowed us to control the approximations using adaptive subsampling which depends on the trust region radius. Here, as in [7], we will rely on the independently sampled subsets and a nondecreasing, adaptive subsampling strategy. The subsampling strategy is constructed in such way to avoid employing the entire sample set for the functions f^i which are homogeneous, i.e., whose subsampled values do not noticeably differ from the full sample values. This will create, as mentioned, two possible scenarios: "mini-batch" (MB) and "full sample" (FS). Recall that in the MB scenario, there exists at least one component function $f^i(x)$ for which the subsampling size N_k^i is strictly less than the sample size N throughout the algorithm. On the other hand, FS does not imply fully deterministic approach, but rather an increasing sample mode where the full sample is reached eventually. More formally, let us define

$$M_b := \{i \in \{1, ..., q\} \mid N_k^i < N, \forall k \in \mathbb{N}\}. \tag{8}$$

Thus, $M_b \neq \emptyset$ implies MB scenario, while $M_b = \emptyset$ implies FS scenario.

3 Algorithm

Within this section we describe the proposed Additional Sampling algorithm for Multi-Objective Problems - ASMOP. As mentioned earlier, at each iteration k, a quadratic model $m_{\mathcal{N}_k}(d)$ is formed using the subsampled values (4) and Hessian approximations H_k^i. We find the direction by approximately solving the problem

$$\min_{\|d\| \leq \delta_k} m_{\mathcal{N}_k}(d) \tag{9}$$

in a such way to ensure the Cauchy decrease condition

$$m_{\mathcal{N}_k}(0) - m_{\mathcal{N}_k}(d_k) \geq \frac{1}{2}\omega_{\mathcal{N}_k}(x_k) \min\{\delta_k, \frac{\omega_{\mathcal{N}_k}(x_k)}{\beta_k}\} \tag{10}$$

with

$$\beta_k = 1 + \max_{i \in \{1,...,q\}} \|H_k^i\|. \tag{11}$$

It can be shown that such direction exists (see Lemma 2 of [11] with $H_k = \max_{i \in \{1,...,q\}} H_k^i$ for instance). Similar to [7], we will take a trial point $x_t = x_k + d_k$, and check the ratio between the decrease of the scalar function and the quadratic model. Since we are dealing with noisy approximations in general, we adopt non-monotone trust region strategy to avoid imposing a strict decrease and define the ratio as follows

$$\rho_{\mathcal{N}_k} := \frac{\phi_{\mathcal{N}_k}(x_t) - \phi_{\mathcal{N}_k}(x_k) - \delta_k t_k}{m_{\mathcal{N}_k}(d_k) - m_{\mathcal{N}_k}(0)} \tag{12}$$

where $t_k > 0$ for all k and

$$\sum_{k=0}^{\infty} t_k \leq t < \infty. \tag{13}$$

The role of t_k is to control a potential increase of the scalar function. Furthermore, throughout the whole MB phase of the algorithm, we perform an additional sampling to form $\mathcal{D}_k = (\mathcal{D}_k^1, ..., \mathcal{D}_k^q)$ with $\mathcal{D}_k^i \subset \mathcal{N}^i$ such that $D_k^i = |\mathcal{D}_k^i| < N$ for all $i = 1, ..., q$. This sampling is done independently of $\mathcal{N}_k$ and it is used to calculate $f^i_{\mathcal{D}_k^i}(x_k)$, $f^i_{\mathcal{D}_k^i}(x_t)$ and $\nabla f^i_{\mathcal{D}_k^i}(x_k)$ by the following formulas

$$f^i_{\mathcal{D}_k^i}(x) = \frac{1}{D_k^i} \sum_{j \in \mathcal{D}_k^i} f_j^i(x), \ \nabla f^i_{\mathcal{D}_k^i}(x) = \frac{1}{D_k^i} \sum_{j \in \mathcal{D}_k^i} \nabla f_j^i(x), \quad i = 1, ..., q. \tag{14}$$

Keep in mind that it is possible that some component functions reached the full sample, while the others did not. In such case we set $\mathcal{D}_k^i = \mathcal{N}^i$ for each $i \in \{1, ..., q\}$ such that $\mathcal{N}_k^i = \mathcal{N}^i$ to avoid unnecessary computations. For all other components such that $N_k^i < N$ it is possible to use a single-element subsample, i.e., $D_k^i = 1$, which minimizes the computational cost of additional sampling. The following ratio acts as an additional measure of the adequacy of the trial point

$$\rho_{\mathcal{D}_k} := \frac{\phi_{\mathcal{D}_k}(x_t) - \phi_{\mathcal{D}_k}(x_k) - \delta_k \bar{t}_k}{-\max_i \|\nabla f^i_{\mathcal{D}_k^i}(x_k)\|} \tag{15}$$

where $\bar{t}_k > 0$ and

$$\sum_{k=0}^{\infty} \bar{t}_k \leq \bar{t} < \infty. \tag{16}$$

Notice that if $\rho_{\mathcal{D}_k} \geq \nu$ then we have

$$\phi_{\mathcal{D}_k}(x_t) \leq \phi_{\mathcal{D}_k}(x_k) + \delta_k \bar{t}_k - \nu \max_i \|\nabla f^i_{\mathcal{D}_k^i}(x_k)\| \tag{17}$$

which is an Armijo-like condition. Throughout the MB phase of the algorithm, the trial point is accepted if both $\rho_{\mathcal{N}_k}$ and $\rho_{\mathcal{D}_k}$ are big enough. On the other hand, if the FS phase is reached, only $\rho_{\mathcal{N}_k} = \rho_{\mathcal{N}}$ is considered as in the deterministic version of the multi objective trust region [14]. If the sampling of $\mathcal{D}_k$ is done uniformly and randomly, with replacements, then $f^i_{\mathcal{D}_k^i}(x_t)$ represents a conditionally unbiased estimator of $f^i(x_t)$. The ratio $\rho_{\mathcal{D}_k}$ is also used to control the subsampling size. If $\phi_{\mathcal{D}_k}$ increases a lot, more precisely, if $\rho_{\mathcal{D}_k} < \nu$ and thus $\phi_{\mathcal{D}_k}(x_t) > \phi_{\mathcal{D}_k}(x_k) + \delta_k \bar{t}_k - \nu \max_i \|\nabla f^i_{\mathcal{D}_k^i}(x_k)\|$, the components which haven't reached the full sample N need to increase

the subsampling size N_k^i and choose a new subsample for the subsequent iteration. Using the stochastic average approximation error estimate

$$h_k^i := \frac{N - N_k^i}{N} \tag{18}$$

we can also increase the subsampling size N_k^i if we get too close to the Pareto optimal point of the approximate problem, i.e., if $\omega_{\mathcal{N}_k}(x_k)$ gets relatively close to 0. In other words, if for some i, $N_k^i << N$, and $\omega_{\mathcal{N}_k}$ is extremely small, we increase N_k^i so that the algorithm does not get stuck near the stationary point of the wrong function. In order to facilitate the exposition of the algorithm, let us define

$$M_b^k := \{i \in \{1, ..., q\} \mid N_k^i < N\}. \tag{19}$$

Notice that $M_b = \emptyset$ implies that the algorithm is in the FS phase, while $M_b \neq \emptyset$ implies the MB phase. The algorithm can be stated as follows.
Algorithm 1. *(ASMOP)*

Step 0. *Initialization.*
Choose $x_0 \in \mathbb{R}^n, \delta_0 \in (0, \delta_{max})$, $\gamma_1 \in (0,1), \gamma_2 = 1/\gamma_1, \nu, \varepsilon > 0, \eta \in (0, \frac{3}{4})$, $\{t_k\}$ satisfying (13) and $\{\bar{t}_k\}$ satisfying (16), $\mathcal{N}_0 = (\mathcal{N}_0^1, ..., \mathcal{N}_0^q)$.
Set $k = 0$.

Step 1. *Candidate point.*
Form the model (7) and find the step such that (10) holds.
Calculate $\rho_{\mathcal{N}_k}$ by (12) and $\omega_{\mathcal{N}_k}(x_k)$ by (5).
Set $x_t = x_k + d_k$.

Step 2. *Sample update.*
if $M_b^k = \emptyset$ **then**
go to Step 3.
else
For all $i \in M_b^k$:
Choose $\mathcal{D}_k^i$ randomly and uniformly, with replacement, from $\mathcal{N}^i$ and calculate $\rho_{\mathcal{D}_k}$ by (15).
if $\omega_{\mathcal{N}_k}(x_k) < \varepsilon h_k^i$ **then**
Choose $N_{k+1}^i \in (N_k^i, N]$ and choose $\mathcal{N}_{k+1}^i$.
else
if $\rho_{\mathcal{D}_k} < \nu$ **then**
Choose $N_{k+1}^i \in (N_k^i, N]$ and choose $\mathcal{N}_{k+1}^i$.
else
if $\rho_{\mathcal{N}_k} < \eta$ **then**
Set $N_{k+1}^i = N_k^i$ and $\mathcal{N}_{k+1}^i = \mathcal{N}_k^i$.
else
Set $N_{k+1}^i = N_k^i$, and choose $\mathcal{N}_{k+1}^i$.

end if
end if
end if
end if

Step 3. *Iterate update.*

if $M_b^k \neq \emptyset$ **then**
if $\rho_{\mathcal{N}_k} \geq \eta$ and $\rho_{\mathcal{D}_k} \geq \nu$ **then**
$x_{k+1} = x_t$
else
$x_{k+1} = x_k$
end if
else
if $\rho_{\mathcal{N}_k} \geq \eta$ **then**
$x_{k+1} = x_t$
else
$x_{k+1} = x_k$
end if
end if

Step 4. *Radius update.*

if $\rho_{\mathcal{N}_k} \geq \eta$ **then**
$\delta_{k+1} = \min\{\delta_{max}, \gamma_2 \delta_k\}$
else
$\delta_{k+1} = \gamma_1 \delta_k$
end if

Step 5. *Counter update.* Set $k = k + 1$ and go to Step 1.

The algorithm is such that the sample sizes are nondecreasing, hence once the full sample is reached, it does not change anymore. At Step 2, if $\rho_{\mathcal{D}_k} \geq \nu$ and $\rho_{\mathcal{N}_k} < \eta$, the sample stays the same, however the radius decreases in such case, hence we try to find new direction with the same approximate functions, but different radius which improves the quadratic model's accuracy. If the decrease (or more precisely - the controlled increase) happens for both $\phi_{\mathcal{N}_k}$ and $\phi_{\mathcal{D}_k}$, we find a new subsample of the same size, and increase the trust region radius. Recall that we calculate $\rho_{\mathcal{D}_k}$ only during the MB phase of the algorithm and there are no restrictions on how we choose $\mathcal{D}_k^i$ - in fact $D_k^i = 1$ is a common choice in the additional sampling strategy that we use and it is also the choice which we set for numerical experiments presented in Section 5.

4 Convergence analysis

Within this section we prove almost sure convergence of a subsequence of marginal functions $\omega(x_k)$, which is an equivalent of vanishing subsequence of gradients in the scalar case ($q = 1$). We start the analysis by imposing some standard assumptions for the considered framework.

Assumption 1. All the functions f_j^i, $j \in \mathcal{N}^i$, $i = 1, ..., q$ are twice continuously differentiable and bounded from bellow.

Notice that this assumption implies that all the subsampled functions $f^i_{\mathcal{N}_k^i}, i = 1, ..., q$ are twice continuously differentiable and bounded from bellow as well.

Assumption 2. There exists a positive constant c_h such that

$$\|\nabla^2 f_j^i(x)\| \leq c_h \quad \text{for all} \quad x \in \mathbb{R}^n,\ j \in \mathcal{N}^i,\ i = 1, ..., q,$$

and the sequence of β_k defined by (11) is uniformly bounded, i.e., there exists a positive constant c_b such that

$$\beta_k = 1 + \max_i \|H_k^i\| \leq c_b \quad \text{for all} \quad k \in \mathbb{N}.$$

This assumption implies that all the Hessians $\|\nabla^2 f^i_{\mathcal{N}_k^i}(x)\|$ are uniformly bounded as well with the same constant c_h.

The next lemma shows that the error of the approximate model $m_{\mathcal{N}_k}$ can be controlled by the trust region radius. The proof is similar to the proof of Proposition 5.1 in [14] and we state it for the sake of completeness.

Lemma 2. Suppose that Assumptions 1 and 2 hold. Then there exists a positive constant $c_f > 0$ such that for each $k \in \mathbb{N}$

$$|\phi_{\mathcal{N}_k}(x_k + d_k) - m_{\mathcal{N}_k}(d_k)| \leq c_f \delta_k^2.$$

Proof. Due to Assumption 1 and Taylor's expansion, for each i we obtain

$$f^i_{\mathcal{N}_k^i}(x_k + d_k) = f^i_{\mathcal{N}_k^i}(x_k) + \langle \nabla f^i_{\mathcal{N}_k^i}(x_k), d_k \rangle + \frac{1}{2}\langle d_k, \nabla^2 f^i_{\mathcal{N}_k^i}(x_k + \alpha_i d_k) d_k \rangle, \quad (20)$$

where $\alpha_i \in (0, 1)$. By adding and subtracting $\frac{1}{2}\langle d_k, H_k^i d_k \rangle$ on the right-hand side and using maximum over i we obtain

$$\begin{aligned}\max_{i \in \{1,...,q\}} f^i_{\mathcal{N}_k^i}(x_k + d_k) \leq & \max_{i \in \{1,...,q\}} \{f^i_{\mathcal{N}_k^i}(x_k) + \langle \nabla f^i_{\mathcal{N}_k^i}(x_k), d_k \rangle + \frac{1}{2}\langle d_k, H_k^i d_k \rangle\} \\ & + \max_{i \in \{1,...,q\}} \{\frac{1}{2}\langle d_k, (\nabla^2 f^i_{\mathcal{N}_k^i}(x_k + \alpha_i d_k) - H_k^i) d_k \rangle\},\end{aligned}$$

which is equivalent to

$$\phi_{\mathcal{N}_k}(x_k + d_k) - m_{\mathcal{N}_k}(d_k) \leq \max_{i \in \{1,...,q\}} \{\frac{1}{2}\langle d_k, (\nabla^2 f^i_{\mathcal{N}_k^i}(x_k + \alpha_i d_k) - H_k^i) d_k \rangle\}.$$

By using the fact that $\|d_k\| \leq \delta_k$ and Assumption 2, we conclude that for each i

$$\begin{aligned}\frac{1}{2}\langle d_k, (\nabla^2 f^i_{\mathcal{N}^i_k}(x_k+\alpha_i d_k) - H^i_k)d_k\rangle &\leq \frac{1}{2}\|d_k\|^2(\|\nabla^2 f^i_{\mathcal{N}^i_k}(x_k+\alpha_i d_k)\| + \|H^i_k\|) \\ &\leq \frac{1}{2}\delta_k^2(c_h + c_b) := c_f\delta_k^2.\end{aligned}$$

Thus, we obtain

$$\phi_{\mathcal{N}_k}(x_k+d_k) - m_{\mathcal{N}_k}(d_k) \leq c_f\delta_k^2. \tag{21}$$

Moreover, by rearranging (20) we obtain

$$f^i_{\mathcal{N}^i_k}(x_k) + \langle \nabla f^i_{\mathcal{N}^i_k}(x_k), d_k\rangle = f^i_{\mathcal{N}^i_k}(x_k+d_k) - \frac{1}{2}\langle d_k, \nabla^2 f^i_{\mathcal{N}^i_k}(x_k+\alpha_i d_k)d_k\rangle.$$

Adding $\frac{1}{2}\langle d_k, \nabla^2 H^i_k d_k\rangle$ on both sides and using maximum over i we obtain

$$m_{\mathcal{N}_k}(d_k) \leq \phi_{\mathcal{N}_k}(x_k+d_k) + \max_{i\in\{1,\ldots,q\}}\{-\frac{1}{2}\langle d_k, (\nabla^2 f^i_{\mathcal{N}^i_k}(x_k+\alpha_i d_k) - H^i_k)d_k\rangle\}$$

Since for each i there holds

$$-\frac{1}{2}\langle d_k, (\nabla^2 f^i_{\mathcal{N}^i_k}(x_k+\alpha_i d_k) - H^i_k)d_k\rangle \leq |-\frac{1}{2}\langle d_k, (\nabla^2 f^i_{\mathcal{N}^i_k}(x_k+\alpha_i d_k) - H^i_k)d_k\rangle| \leq c_f\delta_k^2,$$

we get

$$m_{\mathcal{N}_k}(d_k) - \phi_{\mathcal{N}_k}(x_k+d_k) \leq c_f\delta_k^2 \tag{22}$$

and combining this with (21) we obtain the result. ■

Remark 1. Notice that if $M^k_b = \emptyset$, i.e., if the algorithm is in the FS phase, Lemma 2 implies

$$|\phi(x_k+d_k) - m_k(d_k)| \leq c_f\delta_k^2. \tag{23}$$

The further analysis is conducted by observing the two possible outcomes of the algorithm with respect to sample size behavior (MB and FS) separately. However, at the end we combine all the possible outcomes to form the final result stated in Theorem 4.2. In order to do that, we follow the similar path as in other papers dealing with the considered additional sampling approach. Let us denote by $\mathcal{D}_k^+$ the subset of all possible outcomes of $\mathcal{D}_k$ such that $\rho_{\mathcal{D}_k} \geq \nu$, i.e.,

$$\mathcal{D}_k^+ = \{\mathcal{D}_k \subset \mathcal{N} \mid \phi_{\mathcal{D}_k}(x_t) \leq \phi_{\mathcal{D}_k}(x_k) + \delta_k\bar{t}_k - \nu \max_{i\in\{1,..,q\}} \|\nabla f^i_{\mathcal{D}^i_k}(x_k)\|\}. \tag{24}$$

If $\mathcal{D}_k \in \mathcal{D}_k^+$ and $\rho_{\mathcal{N}_k} \geq \eta$ then $x_{k+1} = x_t$, otherwise we have $x_{k+1} = x_k$. Similarly, we denote the set of outcomes where an increase occurs as

$$\mathcal{D}_k^- = \{\mathcal{D}_k \subset \mathcal{N} \mid \phi_{\mathcal{D}_k}(x_t) > \phi_{\mathcal{D}_k}(x_k) + \delta_k\bar{t}_k - \nu \max_{i\in\{1,..,q\}} \|\nabla f^i_{\mathcal{D}^i_k}(x_k)\|\}. \tag{25}$$

Notice that if $\mathcal{D}_k \in \mathcal{D}_k^-$, then $x_{k+1} = x_k$ Now, we proceed by observing the MB scenario first. The following lemma states that, within this scenario, we have $\rho_{\mathcal{D}_k} \geq \nu$ for all k large enough.

Lemma 3. If $M_b \neq \emptyset$ then there exists random, finite iteration $k_0 \in \mathbb{N}$ such that $\mathcal{D}_k^- = \emptyset$ for all $k \geq k_0$.

Proof. Since the sample sizes are nondecreasing, for each $i \in M_b$ there exist a corresponding $\overline{N}^i < N$ and $k_0^i \in \mathbb{N}$ such that $N_k^i = \overline{N}^i$ for all $k \geq k_0^i$. Without loss of generality, let us assume that for all $j \notin M_b$ there holds $N_k^j = N$ for all $k \geq k_0 := \max_{i \in \{1,..,q\}} k_0^i$, i.e., for all $k \geq k_0$ the subsample sizes reached their upper limit.

Let us assume the contrary, that there exist an infinite subsequence of iterations $K \subset \mathbb{N}$ such that $\mathcal{D}_k^- \neq \emptyset$ for all $k \in K$. That means that for all $k \in K$ there exists at least one possible choice of $\mathcal{D}_k$ such that $\rho_{\mathcal{D}_k} < \nu$. Without loss of generality, assume that for all $k \in K$ we have that $k \geq k_0$. Since $D_k^i \leq N - 1$, there are finitely many combinations with repetitions for $\mathcal{D}_k^i$ and thus there are finitely many choices for q-tuples $\mathcal{D}_k$. [1] Therefore there exists $\tilde{p} \in (0, 1)$ such that $P(\mathcal{D}_k \in \mathcal{D}_k^-) \geq \tilde{p}$, i.e., $P(\mathcal{D}_k \in \mathcal{D}_k^+) \leq 1 - \tilde{p} = p < 1$. Hence

$$P(\mathcal{D}_k \in \mathcal{D}_k^+, \forall k \in K) \leq \prod_{k \in K} p = 0,$$

i.e., almost surely there exists $k \geq k_0$, such that $\rho_{\mathcal{D}_k} < \nu$. However, according to Step 2 of the algorithm, the sample size is increased for all $i \in M_b^k$ in that case, which is a contradiction with the assumption that $N_k^i = \overline{N}^i$ for all $i \in M_b$ and $k \geq k_0$. This completes the proof. ■

The following lemma will help us to show that the marginal function tends to zero in the MB scenario. In order to prove the convergence result, we define an auxiliary function Φ_{fix}

$$\Phi_{fix}(x) := \frac{1}{N} \sum_{j=1}^{N} \max_{i \in \{1,\ldots,q\}} f_j^i(x). \tag{26}$$

The result is as follows.

Lemma 4. Suppose that Assumptions 1 holds and $M_b \neq \emptyset$. Then

$$\Phi_{fix}(x_t) \leq \Phi_{fix}(x_k) - \nu\omega(x_k) + \delta_k \bar{t}_k$$

holds for all $k \geq k_0$, where k_0 is as in Lemma 3.

[1] More precisely, the number of possible choices for $\mathcal{D}_k$ is $S(D_k^i) \leq \bar{S}^i := (2N-2)!/((N-1)!)^2$ where the upper bound follows from the combinatorics of unordered sampling with replacement. Thus, the number of choices for q-tuples $\mathcal{D}_k$ is also finite and bounded by $\bar{S} = \prod_{i \in M_b} \bar{S}^i$.

Proof. Lemma 3 implies that we have $\rho_{\mathcal{D}_k} \geq \nu$, i.e.,

$$\phi_{\mathcal{D}_k}(x_t) \leq \phi_{\mathcal{D}_k}(x_k) + \delta_k \bar{t}_k - \nu \max_{i \in \{1,...,q\}} \|\nabla f^i_{\mathcal{D}^i_k}(x_k)\|,$$

for all $k \geq k_0$ and for every possible choice of $\mathcal{D}_k$. Since the choice of each $\mathcal{D}^i_k$ is uniform and random with replacements, this further implies[2] that the previous inequality also holds for all the single-element choices of $\mathcal{D}^i_k$ and all their possible combinations forming $\mathcal{D}_k$. Further, observing the combinations of the form $\mathcal{D}_k = (j, ..., j)$ for $j = 1, ..., N$ we obtain

$$\max_{i \in \{1,...,q\}} f^i_j(x_t) \leq \max_{i \in \{1,...,q\}} f^i_j(x_k) + \delta_k \bar{t}_k - \nu \max_{i \in \{1,...,q\}} \|\nabla f^i_j(x_k)\|.$$

Now, summing over j and dividing by N we obtain

$$\Phi_{fix}(x_t) \leq \Phi_{fix}(x_k) + \delta_k \bar{t}_k - \nu \frac{1}{N} \sum_{j=1}^{N} \max_{i \in \{1,...,q\}} \|\nabla f^i_j(x_k)\|. \tag{27}$$

Further, let us observe the marginal function and let us denote by d^*_k the solution of (3) at iteration k, i.e.,

$$\omega(x_k) = - \max_{i \in \{1,..,q\}} \langle \nabla f^i(x_k), d^*_k \rangle.$$

Then for every $i \in \{1, .., q\}$ there holds

$$-\omega(x_k) = \max_{i \in \{1,..,q\}} \langle \nabla f^i(x_k), d^*_k \rangle \geq \langle \nabla f^i(x_k), d^*_k \rangle.$$

Furthermore, by using the Cauchy-Schwartz inequality and the fact that $\|d^*_k\| \leq 1$ we obtain

$$-\omega(x_k) \geq \langle \nabla f^i(x_k), d^*_k \rangle \geq -\|\nabla f^i(x_k)\| \|d^*_k\| \geq -\|\nabla f^i(x_k)\|.$$

Equivalently, $\omega(x_k) \leq \|\nabla f^i(x_k)\|$ for every i and thus we obtain

$$\begin{aligned} \omega(x_k) &\leq \max_{i \in \{1,..,q\}} \|\nabla f^i(x_k)\| = \max_{i \in \{1,..,q\}} \|\frac{1}{N} \sum_{j=1}^{N} \nabla f^i_j(x_k)\| \\ &\leq \max_{i \in \{1,..,q\}} \frac{1}{N} \sum_{j=1}^{N} \|\nabla f^i_j(x_k)\| \leq \frac{1}{N} \sum_{j=1}^{N} \max_{i \in \{1,..,q\}} \|\nabla f^i_j(x_k)\|. \end{aligned} \tag{28}$$

Combining this with (27) we obtain the result. ■

The following result states that in the MB case, eventually all the iterates remain within a random level set of the function Φ_{fix}.

[2]Additional sampling is such that it is possible to have $\mathcal{D}^i_k = \{j, .., j\}$ which is in fact equivalent of choosing one-element $\mathcal{D}^i_k = \{j\}$ due to sample average approximations.

Corollary 4.1. *Suppose that the assumptions of Lemma 4 hold. Then the following holds for all* $k \in \mathbb{N}$

$$\Phi_{fix}(x_{k_0+k}) \leq \Phi_{fix}(x_{k_0}) + \delta_{max}\bar{t}.$$

Proof. Since x_{k+1} is either x_k or x_t, Lemma 4 and nonnegativity of $\omega(x_k)$ (Lemma 1, a)) imply that in either case we have $\Phi_{fix}(x_{k+1}) \leq \Phi_{fix}(x_k)+\delta_k\bar{t}_k$ for every $k \geq k_0$. Then, the results follows from summability of $\bar{t}_k$ (16) and the fact that the sequence of δ_k is uniformly bounded by δ_{max}. ■

Similar result can be obtained for the FS scenario as well.

Lemma 5. Suppose that the Assumptions 1 and 2 hold. If $M_b = \emptyset$, then there exists a finite, random iteration k_1 such that the following holds for all $k \in \mathbb{N}$

$$\phi(x_{k_1+k}) \leq \phi(x_{k_1}) + \delta_{max}t.$$

Proof. Since $M_b = \emptyset$, there exists a random, finite iteration $k_1 \in \mathbb{N}$ such that for all $k \geq k_1$ the full sample is reached, i.e., $N_k^i = N$ for all $i = 1, ..., q$, and thus we have $\phi_{\mathcal{N}_k}(x_k) = \phi(x_k)$, $m_{\mathcal{N}_k}(d_k) = m_k(d_k)$ and $\omega_{\mathcal{N}_k}(x_k) = \omega(x_k)$. Let us observe iterations $k \geq k_1$ and denote $\rho_k := \rho_{\mathcal{N}}$. Then, according to the algorithm, the step is accepted if and only if $\rho_k \geq \eta$. So, either $\phi(x_{k+1}) = \phi(x_k)$ or $\phi(x_{k+1}) = \phi(x_t)$ and $\rho_k \geq \eta$, i.e.,

$$\begin{aligned}\phi(x_{k+1}) &\leq \phi(x_k) + t_k\delta_k + \eta(m_k(d_k) - m_k(0)) \\ &\leq \phi(x_k) + t_k\delta_k - \frac{\eta}{2}\omega(x_k)\min\{\delta_k, \frac{\omega(x_k)}{\beta_k}\} \\ &\leq \phi(x_k) + t_k\delta_{max}\end{aligned}$$

Hence the result follows from (13). ■

In order to prove the main result, we assume that the expected value of any local cost function f_j^i is uniformly bounded. For instance, this is true under the bounded iterates assumption which is common is stochastic framework. More precisely, we assume the following.

Assumption 3. There exists a positive constant C such that for every $i = 1, ..., q$ and $j \in \mathcal{N}^i$ we have

$$\mathbb{E}(|f_j^i(x_{k_0})| \mid MB) \leq C \quad \text{and} \quad \mathbb{E}(|f_j^i(x_{k_1})| \mid FS) \leq C,$$

where MB (FS) represents all possible mini-batch (full sample) sample paths of the algorithm, respectively.

Notice that Assumption 3 implies that

$$\mathbb{E}(\Phi_{fix}(x_{k_0}) \mid MB) \leq C \quad \text{and} \quad \mathbb{E}(\phi_{fix}(x_{k_1}) \mid FS) \leq C. \tag{29}$$

In the sequel we will show the convergence result for ASMOP algorithm. The analysis is conducted in a similar way as in the proof of Theorem 1 of [7], but adapted to fit the multi-objective framework. For the sake of readability, we observe separately MB and FS case. First we consider the MB scenario and show that $\liminf_{k\to\infty} \omega(x_k) = 0$ almost surely.

Theorem 1. Suppose that Assumptions 1, 2 and 3 hold. Then

$$P(\liminf_{k\to\infty} \omega(x_k) = 0 \mid MB) = 1.$$

Proof. Let us consider the MB scenario, i.e., the sample paths such that $M_b \neq \emptyset$. Then, Lemma 3 implies that we have $\rho_{\mathcal{D}_k} \geq \nu$, for all $k \geq k_0$ and for every possible choice of $\mathcal{D}_k$. Moreover, we know that for each $i \in M_b$ we have $N_k^i = \overline{N}^i < N$, and for each $i \notin M_b$ we have $N_k^i = N$ for all $k \geq k_0$. Moreover, according to Step 2 of the algorithm, we have

$$\omega_{\mathcal{N}_k}(x_k) \geq \varepsilon h_k^i \geq \varepsilon \frac{1}{N} =: \varepsilon_N > 0.$$

for all $i \in M_b$ and $k \geq k_0$. Now, we will show that there exists an infinite subset of iterations at which the trial point is accepted, i.e., that $\rho_{\mathcal{N}_k} \geq \eta$ occurs infinite number of times.

Suppose the contrary, that there exists $k_3 \geq k_0$ such that $\rho_{\mathcal{N}_k} < \eta$ for all $k \geq k_3$. This further implies that $\lim_{k\to\infty} \delta_k = 0$ due to Step 3 of the algorithm. Moreover, in this scenario, we also have a fixed sample $\mathcal{N}_k = \mathcal{N}_{k_3}$ for all $k \geq k_3$. Furthermore, since $m_{\mathcal{N}_k}(0) = \phi_{\mathcal{N}_k}(x_k)$, from Lemma 2 for all $k \geq k_3$ we have that

$$\begin{aligned} |\rho_{\mathcal{N}_k} - 1| &= |\frac{\phi_{\mathcal{N}_k}(x_t) - \phi_{\mathcal{N}_k}(x_k) - t_k\delta_k}{m_{\mathcal{N}_k}(d_k) - m_{\mathcal{N}_k}(0)} - 1| = |\frac{\phi_{\mathcal{N}_k}(x_t) - t_k\delta_k - m_{\mathcal{N}_k}(d_k)}{m_{\mathcal{N}_k}(d_k) - m_{\mathcal{N}_k}(0)}| \\ &\leq \frac{c_f\delta_k^2 + t_k\delta_k}{\frac{\omega_{\mathcal{N}_k}(x_k)}{2}\min\{\delta_k, \frac{\omega_{\mathcal{N}_k}(x_k)}{\beta_k}\}} \leq \frac{c_f\delta_k^2 + t_k\delta_k}{\frac{\varepsilon_N}{2}\min\{\delta_k, \frac{\varepsilon_N}{c_b}\}} \end{aligned}$$

Since $\lim_{k\to\infty} \delta_k = 0$, there exists $k_4 \geq k_3$ such that $\delta_k < \varepsilon_N/c_b$ for all $k \geq k_4$ and thus

$$|\rho_{\mathcal{N}_k} - 1| \leq \frac{c_f\delta_k^2 + t_k\delta_k}{\frac{\varepsilon_N\delta_k}{2}} = \frac{2c_f\delta_k + 2t_k}{\varepsilon_N}.$$

Letting $k \to \infty$ and using the fact that $\lim_{k\to\infty} t_k = 0$ due to (13), we obtain $\lim_{k\to\infty} \rho_{\mathcal{N}_k} = 1$, which is in contradiction with the assumption of $\rho_{\mathcal{N}_k} < \eta < \frac{3}{4}$ for all $k \geq k_3$.

Thus, we have just shown that there exists an infinite subsequence $K_2 \subset \mathbb{N}$ such that $\rho_{\mathcal{N}_k} \geq \eta$ for all $k \in K_2$. Let $K_3 = K_2 \cap \{k_0, k_0+1, ...\} := \{k_j\}_{j\in\mathbb{N}}$. Then, for all $k \in K_3$ we have $\rho_{\mathcal{D}_k} \geq \nu$ and $\rho_{\mathcal{N}_k} \geq \eta$, and thus $x_{k+1} = x_t$.

Notice that for all the intermediate iterations, i.e., for all $k \geq k_0$, $k \notin K_3$ we have $x_{k+1} = x_k$. Thus, Lemma 4 implies

$$\Phi_{fix}(x_{k_{j+1}}) = ... = \Phi_{fix}(x_{k_j+1}) \leq \Phi_{fix}(x_{k_j}) - \nu\omega(x_{k_j}) + \delta_{k_j}\bar{t}_{k_j}$$

and thus for each $j \in \mathbb{N}$

$$\Phi_{fix}(x_{k_j}) \leq \Phi_{fix}(x_{k_0}) - \nu \sum_{l=0}^{j-1} \omega(x_{k_l}) + \delta_{max}\bar{t}.$$

Applying the conditional expectation $\mathbb{E}(\cdot \mid MB)$ and using Assumptions 1 which implies that $\Phi_{fix}(x_{k_j})$ is bounded from below for any j, by employing Assumption 3 and letting $j \to \infty$ we conclude that

$$\sum_{l=0}^{\infty} \mathbb{E}(\omega(x_{k_l}) \mid MB) < \infty.$$

Now, for any ϵ, from Markov's inequality we obtain

$$\sum_{l=0}^{\infty} P(\omega(x_{k_l}) \geq \epsilon \mid MB) \leq \frac{1}{\epsilon} \sum_{l=0}^{\infty} \mathbb{E}(\omega(x_{k_l}) \mid MB) < \infty.$$

Finally, Borel-Cantelli Lemma implies that $\lim_{l\to\infty} \omega(x_{k_l}) = 0$ which completes the proof. ■

Next, we show the same result for the FS scenario.

Theorem 2. Suppose that Assumptions 1, 2 and 3 hold. Then

$$P(\liminf_{k\to\infty} \omega(x_k) = 0 \mid FS) = 1.$$

Proof. Let us consider the FS scenario, i.e., the sample paths such that $M_b = \emptyset$. Then, as in the proof of Lemma 5, for all $k \geq k_1$ we have $N_k^i = N$ for all $i = 1, ..., q$, and thus $\phi_{\mathcal{N}_k}(x_k) = \phi(x_k)$, $m_{\mathcal{N}_k}(d_k) = m_k(d_k)$, $\omega_{\mathcal{N}_k}(x_k) = \omega(x_k)$ and $\rho_k := \rho_{\mathcal{N}}$.

Let us suppose the the statement of this theorem is not true, i.e., that there exists $\varepsilon > 0$ and $k_2 \geq k_1$ such that $\omega(x_k) > \varepsilon$ for all $k \geq k_2$. Since (13) implies that $\lim_{k\to\infty} t_k = 0$, without loss of generality, assume that $t_k < c_f$ for all $k \geq k_2$. Then it can be shown that the sequence of δ_k is uniformly bounded away from zero. Indeed, if at any iteration $k \geq k_2$ the value of δ_k falls below $\hat{\delta} := \varepsilon/(20\max\{1, c_f, c_b\})$, then Lemma 2 implies

$$\begin{aligned} |\rho_k - 1| = |\frac{\phi(x_t) - m_k(d_k) - t_k\delta_k}{m_k(d_k) - m_k(0)}| &\leq \frac{c_f\delta_k^2 + t_k\delta_k}{\frac{\omega(x_k)}{2}\min\{\delta_k, \frac{\omega(x_k)}{\beta_k}\}} \\ &\leq \frac{c_f\delta_k^2 + t_k\delta_k}{\frac{\varepsilon}{2}\min\{\delta_k, \frac{\varepsilon}{c_b}\}} \leq \frac{c_f\delta_k^2 + t_k\delta_k}{\frac{\varepsilon}{2}\delta_k} < \frac{c_f\hat{\delta} + c_f\hat{\delta}}{\frac{\varepsilon}{2}} < \frac{1}{4} \end{aligned}$$

which further implies that $\rho_k > \frac{3}{4} > \eta$, and, according to the algorithm, the radius is increased, i.e., $\delta_{k+1} > \delta_k$. Therefore, there exists $\tilde{\delta} > 0$, such that $\delta_k > \tilde{\delta}$, for all $k \geq k_2$.

On the other hand, the existence of k_3 such that $\rho_k < \eta$ for every $k \geq k_3$ would imply $\lim_{k\to\infty} \delta_k = 0$ due to Step 4 of the algorithm, which is in contradiction with $\delta_k > \tilde{\delta}$, for all $k \geq k_2$. Thus, we conclude that there must exist an infinite number of iterations $K_1 \subset \mathbb{N}$ such that for all $k \in K_1$ there holds $k \geq k_2$ and $\rho_k \geq \eta$. Therefore, for all $k \in K_1$ we have

$$\begin{aligned}
\phi(x_{k+1}) &\leq \phi(x_k) + \delta_k t_k + \eta(m_k(d_k) - m_k(0)) \\
&\leq \phi(x_k) + \delta_k t_k - \eta \frac{\omega(x_k)}{2} \min\{\delta_k, \frac{\omega(x_k)}{\beta_k}\} \\
&\leq \phi(x_k) + \delta_k t_k - \eta \frac{\varepsilon}{2} \min\{\tilde{\delta}, \frac{\varepsilon}{c_b}\} \\
&=: \phi(x_k) + \delta_k t_k - \hat{c}.
\end{aligned}$$

Again, without loss of generality, we can assume that for all $k \in K_1$ there holds $\delta_k t_k \leq \hat{c}/2$ and thus

$$\phi(x_{k+1}) \leq \phi(x_k) - \frac{\hat{c}}{2}, \quad k \in K_1.$$

Furthermore, by denoting $K_1 = \{k_j\}_{j\in\mathbb{N}}$ and using the fact that for all $k \geq k_2$ such that $k \notin K_1$ the trial point is rejected and thus $\phi(x_{k+1}) = \phi(x_k)$, we conclude that for all $j \in \mathbb{N}$

$$\phi(x_{k_{j+1}}) = ... = \phi(x_{k_j+1}) \leq \phi(x_{k_j}) - \frac{\hat{c}}{2}.$$

Applying the conditional expectation and using Assumption 3 we obtain that for every $j \in \mathbb{N}$

$$\mathbb{E}(\phi(x_{k_j}) \mid FS) \leq C - j\frac{\hat{c}}{2}$$

and letting $j \to \infty$ we obtain $\lim_{j\to\infty} \mathbb{E}(\phi(x_{k_j}) \mid FS) = -\infty$. This is in contradiction with Assumption 1 which implies that the function ϕ is bounded from below. This completes the proof. ■

Finally, we obtain the following result as a direct consequence of the previous two theorems and Lemma 1.

Corollary 4.2. *Suppose that Assumptions 1, 2 and 3 hold. Then the sequence of iterates $\{x_k\}_{k\in\mathbb{N}}$ generated by the ASMOP algorithm satisfies*

$$\liminf_{k\to\infty} \omega(x_k) = 0 \quad a.s.$$

Moreover, if the sequence of iterates is bounded, then a.s. there exists an accumulation point of the sequence $\{x_k\}_{k\in\mathbb{N}}$ which is a Pareto critical point for problem (1).

5 Numerical results

In this section we showcase numerical results obtained on machine learning problems. The proposed algorithm - ASMOP is compared to the state-of-the-art stochastic multi-gradient method SMG [22] for multi-objective problems. We also compare ASMOP to SMOP [11] which also uses function approximations, but different sample size guidance. Additionally, we will compare different parameter configurations of ASMOP and demonstrate algorithm's behavior under different sample size increments.

5.1 Logistic regression

In our experiments we created multi objective problems using different methodologies. The first set of experiments consists of minimizing the regularized logistic regression loss function with the CIFAR10 and MNIST dataset. CIFAR10 is a dataset for an image classification problem, and it consists of $N = 5 \times 10^4$ RGB images of 32×32 pixels in 3 hues, which are in 10 categories, hence the dimension of the problem is $n = 32 \times 32 \times 3 = 3072$. We create a model which differentiates category 0 (airplane) from 1 (car), and also category 2 (bird) from 3 (cat) at the same time. Assuming $\mathcal{N}$ is the set of indices of the training dataset, we split the dataset into two subsets by creating two index subgroups $\mathcal{N}^1$ and $\mathcal{N}^2$ such that $\mathcal{N}^1$ are indices of the samples which are in categories 0 and 1, and $\mathcal{N}^2$ are indices of the samples in categories 2 and 3. The second dataset we used was the MNIST dataset which consists of $N = 7 \times 10^4$ samples of gray-scale images of handwritten digits with 32×32 pixels, hence the dimension is $n = 1024$. Once again, we created two datasets from MNIST, the first containing samples with labels 0 and 8, and the second with labels 1 and 4. This way the model differentiates digits 0 or 8, and digits 1 and 4 at the same time, similarly as with CIFAR10. We made both subgroups contain the same amount of elements $N = 10^4$. As mentioned, we are minimizing a regularized logistic regression loss function

$$\min_{x \in \mathbb{R}^n} f(x) := (f^1(x), f^2(x)), \tag{30}$$

where the function components are defined as follows:

$$f^i(x) = \frac{1}{N} \sum_{j \in \mathcal{N}^i} \log(1 + e^{(-y_j(x^T a_j))}) + \frac{\lambda_i}{2} \|\hat{x}\|^2, i = 1, 2. \tag{31}$$

Here, $x \in \mathbb{R}^n$ represents model coefficients we are trying to find, $\hat{x}$ coefficient vector without the intercept, a_j the feature vector and y_j the relevant label.

There are several factors and parameter choices which influence the behavior of the algorithm. For this experiment, we set the initial subsampling size $N_0^i = 0.05N$, and we update the size at Step 2 when needed by a small amount, $\Delta N_k^i = 0.001N$. For the nonmonotonicity parameters we

set $t_k = \frac{1}{(k+1)^{1.1}}$, and $\bar{t}_k = \frac{1}{(k+1)^{1.1}}$, which satisfy (13) and (16). Since the power of the denominator is slightly larger than 1, the algorithm will slowly and gradually decrease the tolerance towards the rise of the function value. Both additional sampling sizes D_k^1 and D_k^2 are set to 1 for all iterations in order to keep the additional sampling computationally cheap. We compare algorithms' true marginal function values $\omega(x_k)$ as a measure of stationarity for multi-objective problems. We calculate $\omega(x_k)$ at each iteration and plot it against the computational costs modeled by the number of scalar products. This serves as a reliable way to track the computational cost of the algorithms since scalar products represent a dominant cost in the process of evaluating the functions. The following Figures 1 and 2 compare the three mentioned algorithms for the fixed budget of 10^6 scalar products for CIFAR10 dataset and $5 \cdot 10^5$ for MNIST dataset. We have also included the subsampling sizes of SMOP and ASMOP for both function components in both figures. It is evident that ASMOP uses small amount of information to achieve large function reduction. We have noticed that if we set 5% of samples as a starting size, and increase the size of both subsamples by 1% of the respective maximum sample size, the subsampling sizes update similarly for both criteria f^1 and f^2.

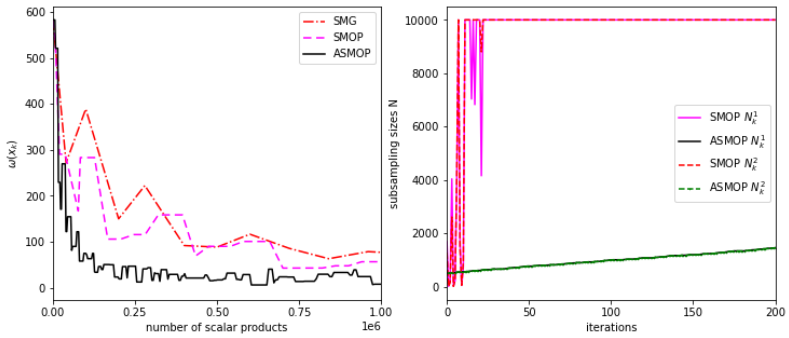

Figure 1: CIFAR10 dataset, problem (31), $N = 10^4, n = 3072$. Optimality measure against computational cost (left) and sample sizes behaviour (right). Parameters: $x_0 = (0.1, 0.1, ..., 0.1), \delta_0 = 1, \delta_{max} = 8, \gamma_1 = 0.5, \gamma_2 = 2, \nu = 10^{-4}, \eta = 0.25, \varepsilon = 10^{-5}$.

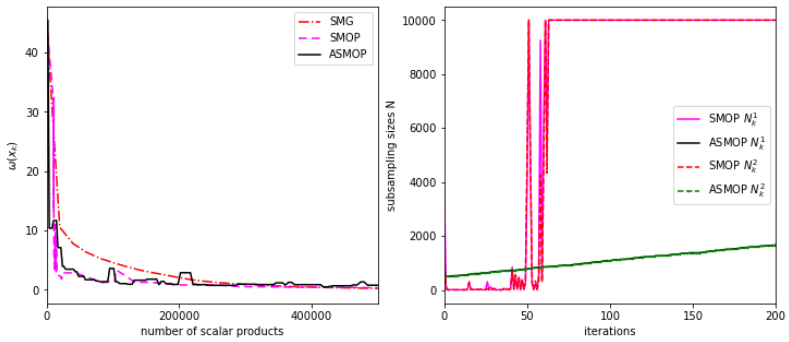

Figure 2: MNIST dataset, problem (31), $N = 10^4, n = 1024$. Optimality measure against computational cost (left) and sample sizes behavior (right). Parameters: $x_0 = (0.05, 0.05, ..., 0.05), \delta_0 = 1, \delta_{max} = 8, \gamma_1 = 0.5, \gamma_2 = 2, \nu = 10^{-4}, \eta = 0.01, \varepsilon = 10^{-5}$.

Using a standard procedure, as in [18] and [11], it is possible to locate the entire Pareto front by utilizing an algorithm which finds Pareto critical points. The procedure is based on choosing the starting approximation of the front, expanding it by adding points in its neighbourhood, applying the chosen algorithm (ASMOP in our case) for several iterations for each point and finally updating the front with the resulting points so that no point is dominated by another. The following figure shows the approximation of the convex Pareto front for the regularized logistic regression problem 31 with CIFAR10 dataset.

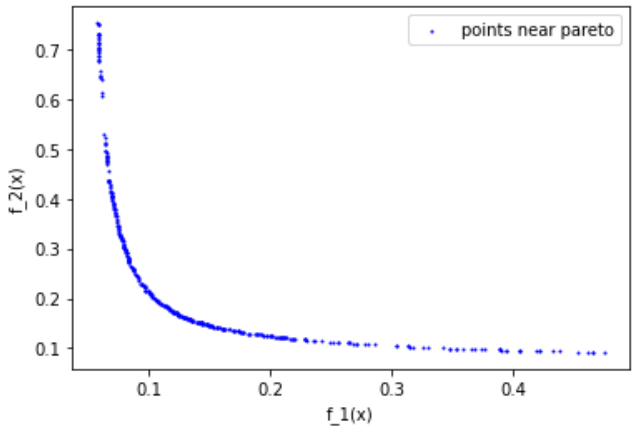

Figure 3: CIFAR10 dataset, problem (31), $N = 10^4, n = 3072$. Pareto front approximation [18] using ASMOP. Parameters: $\delta_0 = 1, \delta_{max} = 8, \gamma_1 = 0.5, \gamma_2 = 2, \nu = 10^{-4}, \eta = 0.25, \varepsilon = 10^{-5}$.

5.2 Logistic regression and Least squares

We have also tested how the ASMOP behaves when the criteria are two completely different loss functions. Once again we have used CIFAR10 and MNIST dataset for an image classification problem. The problem we are solving is (30), however this time the component functions are

$$f^1(x) = \frac{1}{N} \sum_{j \in \mathcal{N}^1} \log(1 + e^{(-y_j(x^T a_j))}) + \frac{\lambda_1}{2} \|\hat{x}\|^2$$

and

$$f^2(x) = \frac{1}{N} \sum_{j \in \mathcal{N}^2} \frac{1}{2}(x^T a_j - y_j)^2 \tag{32}$$

where N is the size of the respective sample groups, $x \in \mathbb{R}^n$ is the vector of model coefficients, $\hat{x}$ coefficient vector without the intercept, a_j the attribute vector of the sample and y_j its respective label. By minimizing this loss function, we get coefficients that are adjusted for both machine learning models. Specifically, for CIFAR10 it will use a regularized logistic regression to differentiate images of cars and planes, and weighted least squares method to differentiate birds from cats, and analogously for MNIST. In the similar manner, we set the initial subsampling sizes $N_0^i = 0.05N$, and the step size update to $\Delta N_k^i = 0.001N$. The predetermined nonmonotonicity sequences were set to $t_k = \frac{1}{(k+1)^{1.1}}$ and $\bar{t}_k = \frac{1}{(k+1)^{1.1}}$ as in the previous experiment. The following figures show $\omega(x_k)$ values in terms of number of scalar products for a fixed budget of $2 \cdot 10^6$ for CIFAR10, and $5 \cdot 10^5$ for MNIST dataset. In both Figures 4 and 5 it can be seen that for the given budget ASMOP shows efficiency and a large decrease in $\omega(x_k)$ value for a small cost.

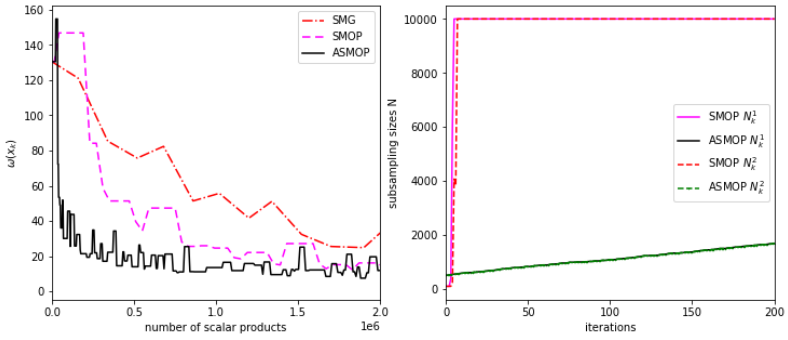

Figure 4: CIFAR10 dataset, problem (32), $N = 10^4, n = 3072$. Optimality measure against computational cost (left) and sample sizes behavior (right). Parameters: $x_0 = (0, 0, ..., 0), \delta_0 = 1, \delta_{max} = 8, \gamma_1 = 0.5, \gamma_2 = 2, \nu = 10^{-4}, \eta = 0.05, \varepsilon = 10^{-5}$.

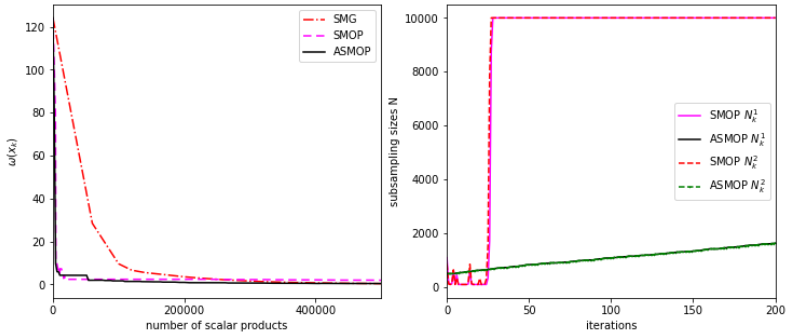

Figure 5: MNIST dataset, problem (32), $N = 10^4, n = 1024$. Optimality measure against computational cost (left) and sample sizes behavior (right). Parameters: $x_0 = (0.1, 0.1, ..., 0.1), \delta_0 = 0.1, \delta_{max} = 3, \gamma_1 = 0.5, \gamma_2 = 2, \nu = 10^{-4}, \eta = 0.1, \varepsilon = 10^{-5}$.

5.3 Nonmonotonicity parameters

In the previous experiments, we set parameters t_k and $\bar{t}_k$ to be $\frac{1}{(k+1)^{1.1}}$. By adjusting these settings it is possible to increase or decrease the tolerance of the nonmonotonicity, which leads to different algorithm behavior. We set

$$\bar{t}_k = \frac{C_2}{(k+1)^{1.1}}$$

and tested three different scenarios ($C_2 \in \{1, 100, 1000\}$) in order to see how the relaxation of the condition $\rho_{\mathcal{D}_k} > \nu$ impacts the performance. The following table shows the chosen settings of the compared algorithms for MNIST dataset, whereas the rest of the parameters were set as in the second experiment.

(MNIST)	C_2
ASMOP1	1
ASMOP2	10^2
ASMOP3	10^3

Table 1: MNIST dataset. Different nonmonotonicity settings for ASMOP versions

We compared these three algorithms similarly as in previous experiments, by criticality measure $\omega(x_k)$ in terms of number of scalar products. The problem being solved is (32), and we showcase results for both datasets.

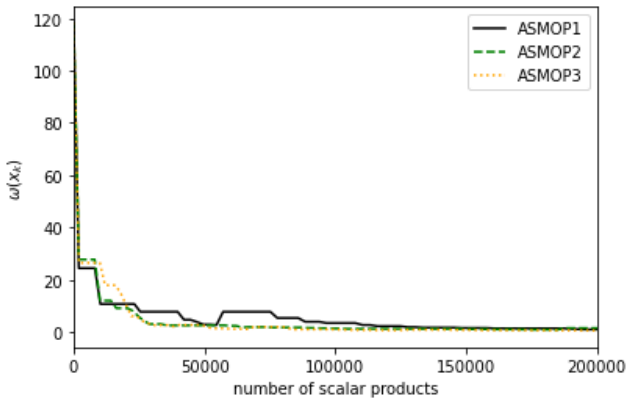

Figure 6: MNIST dataset, problem (32) different settings 5.3,$N = 10^4, n = 1024$. Optimality measure against computational cost. Parameters: $x_0 = (0.1, 0.1, ..., 0.1), \delta_0 = 0.1, \delta_{max} = 3, \gamma_1 = 0.5, \gamma_2 = 2, \nu = 10^{-4}, \eta = 0.1, \varepsilon = 10^{-5}$.

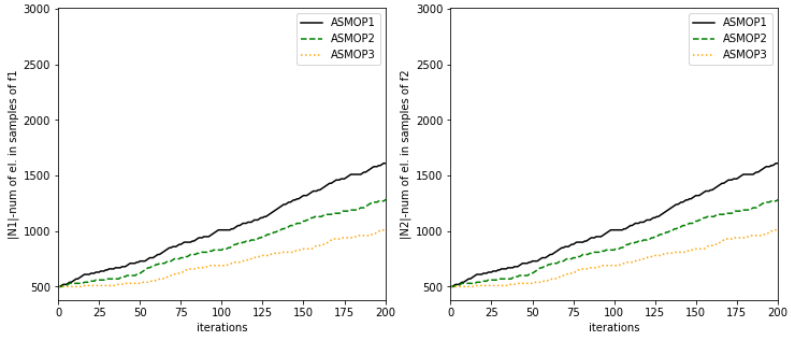

Figure 7: MNIST dataset, problem (32) different settings 5.3,$N = 10^4, n = 1024$. Sample sizes behavior. Parameters: $x_0 = (0.1, 0.1, ..., 0.1), \delta_0 = 0.1, \delta_{max} = 3, \gamma_1 = 0.5, \gamma_2 = 2, \nu = 10^{-4}, \eta = 0.1, \varepsilon = 10^{-5}$.

It is noticeable that the subsampling sizes are increased less frequently for the versions that have a more relaxed coefficient $\rho_{\mathcal{D}_k}$, which means that the higher tolerance leads to the condition $\rho_{\mathcal{D}_k} > \nu$ being satisfied more often in Step 2 of the algorithm. For the CIFAR10 dataset, we increased C_2 parameter as in the following table

(CIFAR10)	C_2
ASMOP1	1
ASMOP2	10^5
ASMOP3	10^7

Table 2: CIFAR10 dataset. Different nonmonotonicity settings for ASMOP versions

The following figures show that for the budget of $3 \cdot 10^5$, the ASMOP3 version doesn't increase the subsampling size, however it showcases similar behaviour as other algorithms.

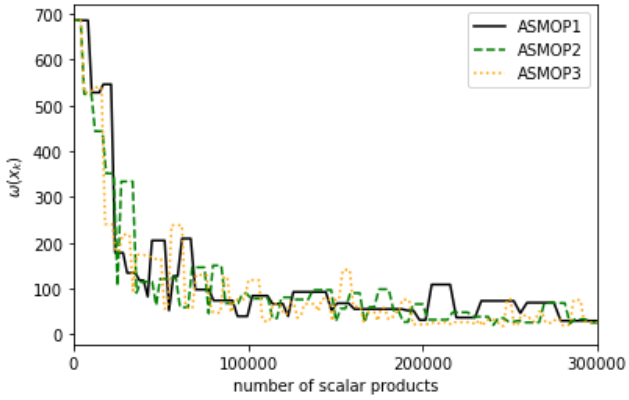

Figure 8: CIFAR10 dataset, problem (32) different settings 5.3,$N = 10^4, n = 1024$. Optimality measure against computational cost. Parameters: $x_0 = (0.1, 0.1, ..., 0.1), \delta_0 = 1, \delta_{max} = 8, \gamma_1 = 0.5, \gamma_2 = 2, \nu = 10^{-4}, \eta = 0.05, \varepsilon = 10^{-5}$.

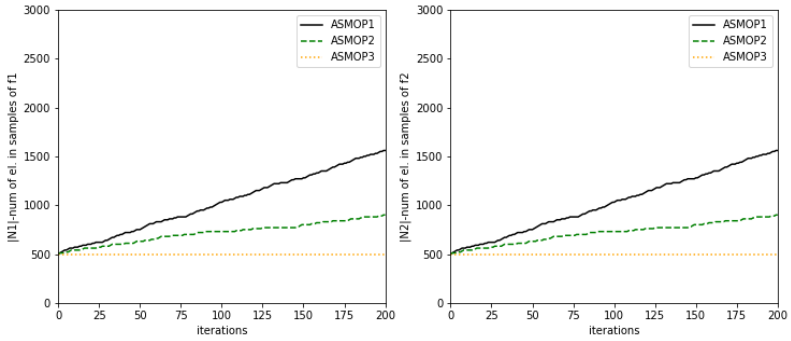

Figure 9: CIFAR10 dataset, problem (32) different settings 5.3,$N = 10^4, n = 1024$. Sample sizes behavior. Parameters: $x_0 = (0.1, 0.1, ..., 0.1), \delta_0 = 1, \delta_{max} = 8, \gamma_1 = 0.5, \gamma_2 = 2, \nu = 10^{-4}, \eta = 0.05, \varepsilon = 10^{-5}$.

6 Conclusion

We proposed an additional sampling algorithm for minimization of finite sum vector functions and thus extend additional sampling trust-region approach to multi-objective optimization problems. This way we create a stochastic MOP method that, depending on a problem at hand, behaves like a mini-batch or increasing sample scheme. The proposed method is supported by theoretical analysis which shows the almost sure convergence of a subse-

quence of iterates towards a Pareto critical point under some standard assumptions for stochastic and MOP framework. Numerical results conducted on several configurations of some representative machine learning problems show the efficiency of the proposed scheme and its competitiveness with relevant existing counterparts for multi-objective problems.

7 Funding

This research was supported by the Science Fund of the Republic of Serbia, GRANT No 7359, Project title - LASCADO.

References

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

[2] 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.*

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

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

[5] 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*

[6] 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.*

[7] 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.*

[8] Krejić, N., & Krklec, N., (2013) Line search methods with variable sample size for unconstrained optimization, *Journal of Computational and Applied Mathematics 245, pp. 213-231.*

[9] Krejić, N. & Krklec Jerinkić, N. (2015) Nonmonotone line search methods with variable sample size, *Numerical Algorithms, 68(4), 711-739.*

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

[11] Krejić, N., Krklec Jerinkić, N. & Rutešić, L. (2025) SMOP: Stochastic trust region method for multi-objective problems *https://arxiv.org/abs/2501.06350*

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

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

[14] 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.*

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

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

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

[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] Yousefi, M. & Calomardo, Á. M. (2022) A stochastic nonmonotone trust-region training algorithm for image classification *16th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS) pp. 522-529*

[20] Sun, S. & Nocedal, J. (2023) A trust region method for noisy unconstrained optimization. *Math. Program, 202, 445–472*

[21] Cao, L., Berahas, A.S. & Scheinberg, K. (2024) First- and second-order high probability complexity bounds for trust-region methods with noisy oracles. *Math. Program, 207, 55–106*

[22] 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.*

[23] Krejić, N., Lužanin, Z., Ovcin, Z., & Stojkovska, I. (2015). Descent direction method with line search for unconstrained optimization in noisy environment. *Optimization Methods and Software, 30(6), 1164–1184.*

[24] Serafino, D., Krejić, N., Krklec Jerinkić, N. & Viola, M. (2023) LSOS: Line-search Second-Order Stochastic optimization methods for nonconvex finite sums *Math. Comp, 92, 1273-1299*

[25] Iusem, A.N., Jofré, A., Oliveira R.I. & Thompson P. (2019) Variance-Based Extragradient Methods with Line Search for Stochastic Variational Inequalities *SIAM Journal on Optimization, 29, 175 - 206*

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