Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Continuous Optimization

A stochastic multiple gradient descent algorithm

Quentin Mercier[a,*], Fabrice Poirion[a], Jean-Antoine Désidéri[b]

[a] ONERA DMAS Université Paris Saclay Onera the French Aerospace Lab, 29 avenue de la Division Leclerc, Chatillon 92320, France
[b] Inria, 2004 Route des Lucioles, Valbonne 06902, France

ARTICLE INFO

Article history:
Received 2 March 2017
Accepted 23 May 2018
Available online 31 May 2018

Keywords:
Multiple objective programming
Multiobjective stochastic optimization
Stochastic gradient algorithm
Multiple gradient descent algorithm
Common descent vector

ABSTRACT

In this article, we propose a new method for multiobjective optimization problems in which the objective functions are expressed as expectations of random functions. The present method is based on an extension of the classical stochastic gradient algorithm and a deterministic multiobjective algorithm, the Multiple Gradient Descent Algorithm (*MGDA*). In *MGDA* a descent direction common to all specified objective functions is identified through a result of convex geometry. The use of this common descent vector and the Pareto stationarity definition into the stochastic gradient algorithm makes the algorithm able to solve multiobjective problems. The mean square and almost sure convergence of this new algorithm are proven considering the classical stochastic gradient algorithm hypothesis. The algorithm efficiency is illustrated on a set of benchmarks with diverse complexity and assessed in comparison with two classical algorithms (*NSGA-II*, *DMS*) coupled with a Monte Carlo expectation estimator.

1. Introduction

Manufacturers are ever looking for designing products with better performance, higher reliability at lower cost and risk. One way to address these antagonistic objectives is to use multiobjective optimization approaches. But real world problems are rarely described through a collection of fixed parameters and uncertainty has to be taken into account, may it appear in the system description itself or in the environment and operational conditions. Indeed the system behavior can be very sensitive to modifications in some parameters (Arnaud & Poirion, 2014; Matthies, Brenner, Bucher, & Soares, 1997; Papadimitriou, Katafygiotis, & Au, 1997). This is why uncertainty has to be introduced in the design process from the start. Optimization under uncertainty has known important advances since the second-half of the 20th century (Bellman & Zadeh, 1970; Dantzig, 2004; Klamroth, Köbis, Schöbel, & Tammer, 2017) and various approaches have been proposed including robust optimization, which encompasses today a rather large field of robustness concepts such as the "worst case" or the mean and variance" concepts (Gabrel, Murat, & Thiele, 2014), and stochastic optimization where uncertain parameters are modeled through random variables with a given distribution and where the probabilistic information is directly introduced in the numerical approaches. In that context the uncertain multiobjective problems are written in terms of the expectation of each objective. In our paper we shall focus on this last interpretation of the optimization problem. Considering single objective stochastic optimization problems, a large variety of numerical approaches (Roy, Hinduja, & Teti, 2008; Sahinidis, 2004) can be found in the literature. Two main distinct approaches exist, one based on stochastic approximations such as the Robbins Monro algorithm and the various stochastic gradient approaches (Ermoliev & Wets, 1988; Ermoliev & Gaivoronski, 1992; Robbins & Monro, 1951), the second one based on scenario approaches (Nemirovski & Shapiro, 2006; Shapiro, 2003), the latter being more frequently applied for chance constrained problems.

Regarding stochastic multiobjective optimization the literature is less prolific: the various approaches proposed are based on classical deterministic algorithm such as genetic algorithms coupled with a robust formulation where the random quantities appearing (such as the mean values or standard deviations) are either obtained analytically for simple objectives (Caballero, Cerdá, del Mar Muñoz, & Rey, 2004; Deb & Gupta, 2005; Fliege & Werner, 2014) or estimated using a sample averaging approach using scenarios (Bonnel & Collonge, 2014; Diaz, Handl, & Xu, 2017; Fliege & Xu, 2011; Löhndorf, 2016; Mattson & Messac, 2005; Wang, Guo, Zheng, & Wang, 2015). In this paper, we propose a new algorithm for constructing the set of Pareto stationary points of a multiobjective optimization problem written in terms of the mean objective functions. The method is based on the use of the *MGDA* algorithm (Désidéri, 2012; 2014) and more precisely on the existence of a common descent vector analogous to the steepest descent vector of Fliege and Svaiter (2000), together with a stochastic gradient algorithm. Convergences of this new algorithm will be proved

[*] Corresponding author.
E-mail addresses: quentin.mercier@onera.fr (Q. Mercier), poirion@onera.fr (F. Poirion), jean-antoine.desideri@inria.fr (J.-A. Désidéri).

https://doi.org/10.1016/j.ejor.2018.05.064
0377-2217/© 2018 Elsevier B.V. All rights reserved.

and several illustrations given. The paper is organized as follows. In Section 2 the Multiple Gradient Descent Algorithm (*MGDA*) is recalled. In Section 3, after introducing some probabilistic notations and results which will be used for the convergence proofs, we introduce the problem under consideration and introduce the Stochastic Multiple Gradient Descent Algorithm (*SMGDA*). Then we shall prove two types of convergence. In Section 4 illustrations of the *SMGDA* algorithm will be given and compared to the classical Sample Average Approximation (SAA) approach (Shapiro, 2003).

2. Multiple gradient descent algorithm (*MGDA*)

The Multiple Gradient Descent Algorithm (*MGDA*) was originally introduced in Désidéri (2009) and Désidéri (2012) to solve general multiobjective optimization problems involving differentiable cost functions. Variants were proposed in Désidéri (2014), but more recently the algorithm was slightly revised in Désidéri (2015) to apply to cases where the number m of objective functions exceeds the dimension n of the working design space.

Recently, the revised version of *MGDA* was applied in a deterministic setting, to a time periodic problem governed by the time dependent compressible Navier–Stokes equations (Désidéri and Duvigneau). There, six parameters defining pulsating jets on a flat plate have been optimized to reduce drag over a time period. The multiple gradients were the realizations at 800 time steps of the gradient of drag with respect to the six parameters. By *MGDA*, drag was reduced at every time steps of the period. Presently, following Désidéri and Duvigneau, we summarize the method used to construct the descent direction from a set made of multiple gradients, using slightly different notations being necessary in the subsequent stochastic framework. For the sake of clarity, the calculation of the common descent vector is only presented in the linearly independent gradients case where the number of objectives m is supposed inferior to the dimension of the design space n.

2.1. Multiobjective problem statement

Let m and n be two arbitrary integers and consider the multiobjective optimization problem consisting in minimizing m differentiable objective functions $\{f_j(\mathbf{x})\}$ in some open admissible domain $\mathcal{D}_a \subseteq \mathbb{R}^n$ ($j = 1, \dots, m$; $f_j \in C^1(\mathcal{D}_a)$). Given a starting point $\mathbf{x}_0 \in \mathcal{D}_a$ and a vector $\mathbf{d} \in \mathbb{R}^n$, one forms the directional derivatives

$$f'_j = \left[\nabla_{\mathbf{x}} f_j(\mathbf{x}_0)\right]^t \mathbf{d}, \tag{1}$$

where $\nabla_{\mathbf{x}}$ is the symbol for the gradient w.r.t. $\mathbf{x}$ and the superscript ^{t} stands for transposition. One seeks for a vector $\mathbf{d}$ such that the scalar product of any objective gradient $\nabla_x f_j(\mathbf{x}_0)$ with the vector $\mathbf{d}$ remains strictly positive

$$f'_j > 0. \tag{2}$$

If such a vector $\mathbf{d}$ exists, the direction of the vector $(-\mathbf{d})$ is said to be a local descent direction common to all objective functions. Then evidently, infinitely many other such directions also exist, and our algorithm permits to identify at least one.

2.2. Convex hull, two lemmas and basic MGDA

We recall the following :

Definition 1. The convex hull of a family of m vectors $\{\mathbf{u}_j\}$ ($j = 1, \dots, m$; $\mathbf{u}_j \in \mathbb{R}^n$), is the set of all their convex combinations

$$\bar{\mathcal{U}} = \left\{ \mathbf{u} \in \mathbb{R}^n \text{ such that } \mathbf{u} = \sum_{j=1}^{m} \alpha_j \mathbf{u}_j;\ \alpha_j \in \mathbb{R}^+\ (\forall j);\ \sum_{j=1}^{m} \alpha_j = 1 \right\}. \tag{3}$$

Evidently, given the m vectors $\{\mathbf{u}_j\}$ in $\mathbb{R}^n$, the convex hull $\bar{\mathcal{U}}$ is a convex, closed and bounded subset of the finite-dimensional subspace spanned by these vectors. Its image in the affine space $\mathbb{R}^n$ in which vectors are associated with representatives of same origin O, is a convex polytope with at most m vertices. Then, we have:

Lemma 1. *Given an $n \times n$ real-symmetric positive-definite matrix $\mathbf{A}_n$, the associated scalar product*

$$\left(\mathbf{u}, \mathbf{v}\right) = \mathbf{u}^t \mathbf{A}_n \mathbf{v} \quad (\mathbf{u}, \mathbf{v} \in \mathbb{R}^n), \tag{4}$$

and Euclidean norm

$$\|\mathbf{u}\| = \sqrt{\mathbf{u}^t \mathbf{A}_n \mathbf{u}}, \tag{5}$$

the convex hull $\bar{\mathcal{U}}$ admits a unique element $\xi^\star$ of minimum norm.

Proof. - Existence: $\bar{\mathcal{U}}$ is compact and $\|.\|$ is a continuous function.
- Uniqueness: suppose that ξ_1 and ξ_2 are two realizations of the minimum $\mu = \arg\min_{\mathbf{u}\in\bar{\mathcal{U}}} \|\mathbf{u}\|$ so that $\mu = \|\xi_1\| = \|\xi_2\|$ and let

$$\xi_s = \frac{1}{2}(\xi_2 + \xi_1), \quad \xi_d = \frac{1}{2}(\xi_2 - \xi_1),$$

so that:

$$\left(\xi_s, \xi_d\right) = \frac{1}{4}\left(\xi_2 + \xi_1, \xi_2 - \xi_1\right) = \frac{1}{4}\left(\|\xi_2\|^2 - \|\xi_1\|^2\right) = 0\,.$$

Hence $\xi_s \perp \xi_d$, and since $\xi_s \in \bar{\mathcal{U}}$, $\|\xi_s\| \geq \mu$, and

$$\mu^2 = \|\xi_2\|^2 = \|\xi_s + \xi_d\|^2 = \|\xi_s\|^2 + \|\xi_d\|^2 \geq \mu^2 + \|\xi_d\|^2 \Longrightarrow \xi_d = 0.$$

□

Lemma 2. *The minimum-norm element $\xi^\star$ defined in Lemma 1 satisfies the following relation for any element $\mathbf{u}$ in the convex hull $\bar{\mathcal{U}}$*

$$\left(\mathbf{u}, \xi^\star\right) \geq \|\xi^\star\|^2. \tag{6}$$

Proof. Let $\mathbf{u} \in \bar{\mathcal{U}}$, arbitrary. Let $r = \mathbf{u} - \xi^\star$; by convexity of $\bar{\mathcal{U}}$, any convex combination of $\xi^\star$ and $\mathbf{u}$ is an element of $\bar{\mathcal{U}}$

$$(1-\epsilon)\xi^\star + \epsilon\mathbf{u} = \xi^\star + \epsilon r \in \bar{\mathcal{U}}\,, \quad \epsilon \in [0,1].$$

By definition of $\xi^\star$, $\|\xi^\star + \epsilon r\| \geq \|\xi^\star\|$, that is

$$\left(\xi^\star + \epsilon r, \xi^\star + \epsilon r\right) - \left(\xi^\star, \xi^\star\right) = 2\epsilon\left(\xi^\star, r\right) + \epsilon^2\|r\|^2 \geq 0\,,$$

and this requires that the coefficient $(\xi^\star,\ r)$ of ϵ be non-negative. □

Then consider the set $\{\mathbf{u}_j\}_{j\in[\![1,m]\!]}$ where each element is the gradient of the objective function f_j at point $\mathbf{x}_0$

$$\mathbf{u}_j = \nabla_{\mathbf{x}} f_j(\mathbf{x}_0). \tag{7}$$

If the vector $\xi^\star$ defined in Lemma 1 is nonzero, the vector

$$\mathbf{d} = \mathbf{A}_n \xi^\star \tag{8}$$

is also nonzero, and is a solution to the problem stated in (1)-(2) since by virtue of Lemma 2

$$\left(\mathbf{u}_j, \xi^\star\right) = \mathbf{u}_j^t \mathbf{A}_n \xi^\star = \mathbf{u}_j^t \mathbf{d} \geq \|\xi^\star\|^2 > 0. \tag{9}$$

The situation in which $\xi^\star = 0$, or equivalently, when there exists a set $\alpha = \{\alpha_j\}$ of m positive real numbers such that

$$\sum_{j=1}^{m} \alpha_j \nabla f_j(\mathbf{x}_0) = 0 \text{ and } \sum_{j=1}^{m} \alpha_j = 1, \tag{10}$$

is said to be one of "Pareto stationarity". The relationship between Pareto optimality and Pareto stationarity was made precise by the following Désidéri (2014) and Désidéri (2015)

Theorem 1. *If the objective functions are differentiable and convex in some open ball $\mathcal{B} \subseteq \mathcal{D}_a$ about $\mathbf{x}_0$, and if $\mathbf{x}_0$ is Pareto optimal, then the Pareto stationarity condition is satisfied at $\mathbf{x}_0$.*

Proof. Without loss of generality, suppose that $f_j(\mathbf{x}_0) = 0$ for $j \in [\![1, m]\!]$. Since, by hypothesis, $\mathbf{x}_0$ is Pareto optimal, a single, arbitrary objective function cannot be improved (here diminished below 0) under the constraint of no-degradation of the others. In particular, $\mathbf{x}_0$ solves the problem

$$\min_{\mathbf{x}} f_m(\mathbf{x}) \ / \ \text{subject to} : f_j(\mathbf{x}) \leq 0, \ \ j \in [\![1, m-1]\!]. \tag{11}$$

Let $\bar{\mathcal{U}}_{m-1}$ be the convex hull of the $m-1$ gradients $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{m-1}\}$ and

$$\xi^\star_{m-1} = \underset{\mathbf{u} \in \bar{\mathcal{U}}_{m-1}}{\text{argmin}} \|\mathbf{u}\|. \tag{12}$$

Existence, uniqueness and following property of this element have already been established (Lemmas 2.1 and 2.2)

$$\left(\mathbf{u}_j, \xi^\star_{m-1}\right) \geq \left\| \xi^\star_{m-1} \right\|^2, \ \ j \in [\![1, m-1]\!]. \tag{13}$$

Two situations are then possible :

1. Either $\xi^\star_{m-1} = 0$, and the Pareto stationarity condition is satisfied at $\mathbf{x} = \mathbf{x}_0$ with $\alpha_m = 0$.
2. Or $\xi^\star_{m-1} \neq 0$. Then let $\phi_j(\epsilon) = f_j(\mathbf{x}_0 - \epsilon \xi^\star_{m-1})$ $(j = 1, \ldots, m-1)$ so that $\phi_j(0) = 0$ and $\phi_j'(0) = -\left(\mathbf{u}_j, \xi^\star_{m-1}\right) \leq -\left\| \xi^\star_{m-1} \right\|^2 < 0$, and for sufficiently-small ϵ

$$\phi_j(\epsilon) = f_j(\mathbf{x}_0 - \epsilon \xi^\star_{m-1}) < 0, \ \ j \in [\![1, m-1]\!]. \tag{14}$$

This result confirms that for the constrained minimization problem (11), Slater's constraint-qualification criterion is satisfied, and optimality requires the satisfaction of the Karush–Kuhn–Tucker (KKT) condition (Boyd & Vandenberghe, 2004), that is, the Lagrangian

$$\mathsf{L} = f_m(\mathbf{x}) + \sum_{j=1}^{m-1} \lambda_j f_j(\mathbf{x}) \tag{15}$$

must be stationary, and this gives

$$\mathbf{u}_m + \sum_{j=1}^{m-1} \lambda_j \mathbf{u}_j = 0 \tag{16}$$

in which $\lambda_j > 0$ for $j \in [\![1, m-1]\!]$ by saturation of the constraints $(f_j(\mathbf{x}_0) = 0)$ and sign convention. Finally, $\Lambda = 1 + \sum_{j=1}^{m-1} \lambda_j > 1$. Thus by dividing the above equation by $\Lambda \neq 0$, the result is achieved.

□

Note that this proof is valid for all m and n, in particular in situations in which $m \geq n$, encountered in particular in multi-point optimization when the number of points is larger than the number of variables, as well as $m < n$ more typical of multidisciplinary optimization. Hence, the Pareto stationarity condition generalizes to the multiobjective context, the classical stationarity condition expressing that an unconstrained differentiable function is extremal.

We now return to the non-trivial case of a point $\mathbf{x}_0$ that is *not Pareto stationary* and we suppose that the vectors $\xi^\star$ and $\mathbf{d}$ ($\xi^\star \neq 0$; $\mathbf{d} \neq 0$) have been identified (see next subsection). Then we define *MGDA* as the iteration which transforms $\mathbf{x}_0$ in

$$\mathbf{x}_1 = \mathbf{x}_0 - \rho \, \mathbf{d}, \tag{17}$$

where $\rho > 0$ is some appropriate step size. In many cases in engineering sciences, the step size can be adjusted after an analysis of the the physical scales involved is made (Désidéri and Duvigneau). PDE-constrainted optimization is our ultimate goal, and realistically, the number of cost function evaluations per optimization iteration should remain small. Additionally, in steady problems, a constant step size is often adequate although not optimal. This is our choice here. Nevertheless, in case where step size adaptation is really beneficial, it may be realized by constraint violation limitation, and/or accurate, or coarse one dimensional optimization in the direction of search; see for example Désidéri (2018).

Thus *MGDA* is an extension to the multiobjective context of the classical steepest descent method, in which the direction of search is taken to be the vector $\mathbf{d}$ defined above. At convergence, the limiting point is Pareto stationary. We now examine how can the vector $\mathbf{d}$ be computed in practice.

2.3. QP formulation and hierarchical Gram-Schmidt orthogonalization

By letting

$$\xi^\star = \sum_{j=1}^{m} \alpha_j \mathbf{u}_j = \mathbf{U}\alpha, \tag{18}$$

where $\mathbf{u}_j = \nabla_{\mathbf{x}} f_j(\mathbf{x}_0)$, $\mathbf{U}$ is the $n \times m$ matrix whose jth column contains the n components of vector $\mathbf{u}_j$, the identification of vector $\xi^\star$ can be made by solving the following Quadratic Programming (QP) problem for the unknown vector of coefficients $\alpha = \{\alpha_j\}$

$$\xi^\star = \arg \min_{\alpha \in \mathbb{R}^m} \frac{1}{2} \alpha^t \mathbf{H} \alpha \tag{19}$$

subject to

$$\alpha_j \geq 0, \ \ j \in [\![1, m]\!], \quad \sum_{j=1}^{m} \alpha_j = 1, \tag{20}$$

where $\mathbf{H} = \mathbf{U}^t \mathbf{A}_n \mathbf{U}$. Note that if vector $\xi^\star$ is unique, vector α may not be.

However, if the family of gradients is linearly-independent, which requires in particular that $m \leq n$, it is possible to choose the scalar product, through the definition of matrix $\mathbf{A}_n$, in such a way that the given gradients are 2 by 2 orthogonal, and of norm 1. Then the QP-formulation admits the trivial solution

$$\xi^\star = \frac{1}{m} \sum_{j=1}^{m} \mathbf{u}_j \tag{21}$$

To characterize matrix $\mathbf{A}_n$, first apply a Gram-Schmidt orthogonalization process to the vectors $\{\mathbf{u}_j\}$ and get a new family of vectors, $\{\mathbf{v}_j\}$, 2 by 2 orthogonal with respect to the standard Euclidean scalar product $(\mathbf{v}_j^t \mathbf{v}_k = 0, \ j \neq k)$; define the following diagonal matrix

$$\Delta = \mathbf{Diag}\left(\mathbf{v}_j^t \mathbf{v}_j\right). \tag{22}$$

Then:

$$\mathbf{A}_n = \mathbf{W}^t \mathbf{W} + (\mathbf{I} - \Pi)^2, \quad \mathbf{W} = \left(\mathbf{U}^t \mathbf{U}\right)^{-1} \mathbf{U}^t \mathbf{V} \Delta^{-1} \mathbf{V}^t, \tag{23}$$

where $\Pi = \mathbf{V} \Delta^{-1} \mathbf{V}^t$ is the projection matrix onto subspace spanned by the gradients (Désidéri, 2015).

Once matrix $\mathbf{A}_n$ is defined, the descent direction $(-\mathbf{d})$ is given by (8), and a descent step can be performed by *MGDA*, (17).

Remark 1. In the case of a linearly-dependent family of gradients, only a subfamily of $\{\mathbf{u}_j\}$ is used in the Gram–Schmidt process. The elements of this subfamily are selected one-by-one according to a specific hierarchical principle which tends to make the cone associated with the hull of the subfamily as large as possible. A new family of same rank and made of vectors two-by-two orthogonal is thus constructed, $\{\mathbf{v}_j\}$. Then, one usually resorts to solving the QP-problem but reformulated in this basis which permits a very stable numerical treatment (Désidéri, 2015; Désidéri and Duvigneau). The case of exception is when Pareto stationarity is detected; then, the algorithm is terminated. However, in any case, this methodological enhancement is not necessary here since the number m of objective functions is less than the dimension n of the admissible working space, and the gradients are assumed to be linearly independent.

3. The stochastic multiple gradient descent algorithm (*SMGDA*)

Using results given in the previous section, we are going to extend the *MGDA* algorithm to the stochastic context using a stochastic gradient like algorithm. Classical probabilistic results are recalled in the first subsection. Then the problem formulation and the *SMGDA* algorithm are described. In the last subsection two convergence results are given.

3.1. Probabilistic prerequisites

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be an abstract probabilistic space, and $W : \Omega \to \mathbb{R}^d,\ \omega \mapsto W(\omega)$ a given random vector. We denote μ the distribution of the random variable W and $\mathcal{W}$ its image space $W(\Omega) \subset \mathbb{R}^d$. Let $W_1, \ldots, W_p, \ldots$ be independent copies of the random variable W which will be used to generate independent random samples with distribution μ. We denote $\mathcal{F}_k = \sigma(W_1, \ldots, W_k)$ the σ-algebra generated by the k first random variables W_i. Since $\mathcal{F}_{k-1} \subset \mathcal{F}_k$ the sequence $\{\mathcal{F}_k\}_{k\geq 1}$ is a filtration denoted $\mathcal{F}$.

Definition 2. A sequence (X_n) of integrable random variables is a supermartingale relatively to the filtration $\mathcal{F}$ if X_n is $\mathcal{F}_n$ measurable and if and only if

$$\mathbb{E}(X_{n+1}|\mathcal{F}_n) \leq X_n$$

almost surely (a.s.) where $\mathbb{E}(X_{n+1}|\mathcal{F}_n)$ denotes the conditional expectation of the random variable X_{n+1} respectively to the σ-algebra $\mathcal{F}_n$.

The two following classical results will be used in the convergence proof:

Lemma 3. *Let $\mathcal{B} \subset \mathcal{A}$ be two σ-algebras from probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and X and Y be two independent random variables such that X is independent of $\mathcal{B}$ and Y is $\mathcal{B}$-measurable. We consider f, a measurable bounded function that takes its values in $\mathbb{R}$. Then*

$$\begin{cases} \mathbb{E}[f(X,Y)|\mathcal{B}] = \varphi(Y) \\ \varphi(y) = \mathbb{E}[f(X,y)] \end{cases}.$$

Theorem 2. *Let $(X_k)_{k\in\mathbb{N}}$ be a positive supermartingale. Then there exists a random variable X_∞ such that X_k converge toward X_∞ almost surely*

$$\mathbb{P}\left(\lim_{k\to\infty} X_k = X_\infty\right) = 1.$$

3.2. Problem statement

Throughout the paper the standard inner product on $\mathbb{R}^n$ will be used and denoted $\langle \cdot, \cdot \rangle$, the norm being denoted $\|\cdot\|$.

Consider m functions f_j: $\mathbb{R}^n \times \mathcal{W} \to \mathbb{R}, j = 1, \ldots, m$. The problem addressed in this paper is to solve the mean multiobjective optimization problem written

$$\min_{\mathbf{x}\in\mathbb{R}^n} \{\mathbb{E}[f_1(\mathbf{x}, W), \mathbb{E}[f_2(\mathbf{x}, W)], \ldots, \mathbb{E}[f_m(\mathbf{x}, W)]\}. \tag{24}$$

More precisely we want to construct a set of points that belongs to the associated Pareto set. As it is written, problem (24) is a deterministic problem but in general the objective function expectations are not known. A classical approach is to replace each expectancy by an estimator built using independent samples w_k of the random variable W (Bonnel & Collonge, 2014; Fliege & Xu, 2011). As for stochastic gradient algorithms, the algorithm we propose does not need to calculate the mean objective functions and is only based on the values of the stochastic functions gradients. The classical stochastic gradient algorithm is based on a descent direction given by the objective function gradient. Here the descent vector which will be used is the common descent vector constructed in the previous section for the *MGDA* algorithm. In the stochastic context this common descent vector is random and defined by the random convex combination

$$\xi^\star(\mathbf{x}, W) = \sum_{j=1}^{m} \alpha_j(\mathbf{x}, W)\nabla f_j(\mathbf{x}, W)\ a.s.,\ \ \mathbf{x} \in \mathbb{R}^n \tag{25}$$

with

$$\sum_j \alpha_j(\mathbf{x}, W) = 1\ a.s. \tag{26}$$

by construction.

The flow chart of *SMGDA* (*Stochastic Multiple Gradient Descent Algorithm*) algorithm is described below.

Algorithm 1: SMGDA.

```
input:
  • An initial point $\mathbf{X}_0$ of the design space
  • A number of iterations $N$
  • A $\sigma$-sequence $\{\epsilon_k\}_{k\in\mathbb{N}}$
begin
    $\mathbf{X} = \mathbf{X}_0$ ;
    for $k \in [\![1, N]\!]$ do
        Generate a sample $w_k$ of random variable $W_k$;
        Evaluate the objective functions and their gradients
        $(\mathbf{X}_{k-1}, w_k) \longrightarrow (f_j(\mathbf{X}_{k-1}, w_k), \nabla f_j(\mathbf{X}_{k-1}, w_k))$;
        Calculate the common descent vector $\xi^\star(\mathbf{X}_{k-1}, w_k)$;
        Update the current parameter values :
        $\mathbf{X}_k = \mathbf{X}_{k-1} - \epsilon_k \xi^\star(\mathbf{X}_{k-1}, w_k)$ .
```

Remark 2. Two parameters require user adjustment: the number of iterations for the stochastic algorithm, and the σ-sequence that define the step size.

Remark 3. No stopping criterion is proposed as it is the case for most stochastic algorithms since there exists no efficient ones.

More generally we shall consider the random sequence $(\mathbf{X}_k)$ defined by the recurrence relation

$$\mathbf{X}_k = \mathbf{X}_{k-1} - \epsilon_k \xi^\star(\mathbf{X}_{k-1}, W_k). \tag{27}$$

3.3. Convergence proofs

Two types of convergences will be proved in this section, the first one being a mean square convergence in the Hilbert space $L^2(\Omega)$, the second one being an almost sure point-wise convergence. The two proofs are extensions of the stochastic gradient convergence proofs and are based on classical assumptions which can be found for instance in Dodu, Goursat, Hertz, Quadrat, and Viot (1981) and Polyak (1976).

The notation $\mathcal{P}_D^\star$ (resp. $\mathcal{P}_O^\star$) will denote the Pareto solution set (resp. the Pareto front). For any $\mathbf{x} \in \mathbb{R}^n$ the notation $\mathbf{x}^\perp$ will denote an element of the Pareto set which minimizes the distance between the point $\mathbf{x}$ and a point of the Pareto set $\mathcal{P}_D{}^\star$

$$\mathbf{x}^\perp \in \underset{\mathbf{u}\in\mathcal{P}_D^\star}{\text{argmin}}\{\|\mathbf{x} - \mathbf{u}\|\}. \tag{28}$$

The convergence results rely upon the following hypotheses.

H1 Problem (24) admits a nonempty Pareto solution set $\mathcal{P}_D^\star$.
H2 The random variables $f_j(\mathbf{x}, W)$ are integrable for $j = 1, \ldots, m$ and for all $\mathbf{x} \in \mathbb{R}^n$.
H3 The functions $\mathbf{x} \mapsto f_j(\mathbf{x}, W) : \mathbb{R}^n \to \mathbb{R}$ are convex and their derivatives exist almost surely for $j = 1, \ldots, m$.

H4 The partial gradient of function f_j with respect to $\mathbf{x}$ is almost surely uniformly bounded by a strictly positive real number M_j

$$\|\nabla f_j(\mathbf{x},W)\| \le M_j \ a.s., \quad \mathbf{x}\in\mathbb{R}^n.$$

H5 For any objective function f_j, there exists a positive real number c_j such that for any $\mathbf{x}$ in $\mathbb{R}^n$ the following relation holds

$$f_j(\mathbf{x},W) - f_j(\mathbf{x}^\perp,W) \ge \frac{c_j}{2}\|\mathbf{x}-\mathbf{x}^\perp\|^2 \, a.s. \ ; \quad j=1,\dots,m.$$

H6 The sequence $\{\epsilon_k\}_{k\in\mathbb{N}}$ is a σ-sequence

$$\sum_{k=0}^{\infty}\epsilon_k = \infty$$

$$\sum_{k=0}^{\infty}\epsilon_k^2 < \infty.$$

Some properties of the common descent vector $\xi^\star$ will be needed further.

Proposition 1. *The norm of the common descent vector $\xi^\star$ is almost surely uniformly bounded by a positive real number $M_{\xi^\star}$*

$$\|\xi^\star(\mathbf{x},W)\| \le M_{\xi^\star} \ a.s., \quad \mathbf{x}\in\mathbb{R}^n.$$

Proof. Using the definition of the common descent vector

$$\forall(\mathbf{x},w)\in\mathbb{R}^n\times\mathcal{W}; \ \ \|\xi^\star(\mathbf{x},w)\| = \|\sum_{j=1}^{m}\alpha_j(\mathbf{x},w)\nabla f_j(\mathbf{x},w)\|$$

with $0\le\alpha_j(\mathbf{x},w)\le 1$ and $\sum_j\alpha_j(\mathbf{x},w)=1$, we can write

$$\|\xi^\star(\mathbf{x},w)\| \le \sum_{j=1}^{m}\|\alpha_j(\mathbf{x},w)\nabla f_j(\mathbf{x},w)\| \le \sum_{j=1}^{m}\|\nabla f_j(\mathbf{x},w)\|.$$

Under assumption H4

$$\|\xi^\star(\mathbf{x},w)\| \le \sum_{j=1}^{m}M_j.$$

□

The mean common descent vector is defined as:

Definition 3.

$$\begin{aligned}\Xi^\star(\mathbf{x}) = \mathbb{E}[\xi^\star(\mathbf{x},W)] &= \mathbb{E}\left[\sum_{j=1}^{m}\alpha_j(\mathbf{x},W)\nabla f_j(\mathbf{x},W)\right]\\ &= \sum_{j=1}^{m}\mathbb{E}\big[\alpha_j(\mathbf{x},W)\nabla f_j(\mathbf{x},W)\big]\end{aligned}$$

Proposition 2. *The mean common descent vector satisfies the following property for any $\mathbf{x}$ in $\mathbb{R}^n$*

$$\Xi^\star(\mathbf{x})(\mathbf{x}-\mathbf{x}^\perp) \ge \frac{C}{2}\|\mathbf{x}-\mathbf{x}^\perp\|^2$$

with $C=\min c_j$ and where the c_j are defined in H5.

Proof. From hypothesis H3, we know that any objective function f_j is almost surely convex

$$\nabla f_j(\mathbf{x},W)(\mathbf{x}-\mathbf{x}^\perp) \ge f_j(\mathbf{x},W) - f_j(\mathbf{x}^\perp,W) \ a.s., \quad \mathbf{x}\in\mathbb{R}^n.$$

Therefore, using assumption H5,

$$\nabla f_j(\mathbf{x},W)(\mathbf{x}-\mathbf{x}^\perp) \ge \frac{c_j}{2}\|\mathbf{x}-\mathbf{x}^\perp\|^2 \ a.s..$$

Introducing the coefficients that define $\xi^\star(x, W)$, we can write

$$\sum_{j=1}^{m}\alpha_j(\mathbf{x},W)\nabla f_j(\mathbf{x},W)(\mathbf{x}-\mathbf{x}^\perp) \ge \sum_{j=1}^{m}\alpha_j(\mathbf{x},W)\frac{c_j}{2}\|\mathbf{x}-\mathbf{x}^\perp\|^2 \, a.s..$$

Let $C=\min_j c_j$, we can write

$$\sum_{j=1}^{m}\alpha_j(\mathbf{x},W)\nabla f_j(\mathbf{x},W)(\mathbf{x}-\mathbf{x}^\perp) \ge \frac{C}{2}\|\mathbf{x}-\mathbf{x}^\perp\|^2\sum_{j=1}^{m}\alpha_j(\mathbf{x},W) \, a.s..$$

Since the $\{\alpha_j\}$ sum up to 1 by construction

$$\sum_{j=1}^{m}\alpha_j(\mathbf{x},W)\nabla f_j(\mathbf{x},W)(\mathbf{x}-\mathbf{x}^\perp) \ge \frac{C}{2}\|\mathbf{x}-\mathbf{x}^\perp\|^2 \, a.s..$$

The proof follows by taking the expectation

$$\begin{aligned}\Xi^\star(\mathbf{x})(\mathbf{x}-\mathbf{x}^\perp) &= \sum_{j=1}^{m}\mathbb{E}[\alpha_j(\mathbf{x},W)\nabla f_j(\mathbf{x},W)](\mathbf{x}-\mathbf{x}^\perp)\\ &\ge \frac{C}{2}\|\mathbf{x}-\mathbf{x}^\perp\|^2.\end{aligned}$$

□

Remark 4 (Weaker hypothesis H5)**.** The approach of keeping the same hypothesis as the mono-objective does not take into account the Pareto order. This makes the hypothesis very strong, because the relation

$$f_j(\mathbf{x},W) - f_j(\mathbf{x}^\perp,W) \ge \frac{c_j}{2}\|\mathbf{x}-\mathbf{x}^\perp\|^2 \, a.s.$$

is supposed true for all objectives $(j=1,\dots,m)$. Using the Pareto dominance approach, we can easily weaken this hypothesis. Considering that $\mathbf{x}^\perp$ dominates almost surely the point $\mathbf{x}$ and that the inequality of hypothesis H5 is true for at least one objective ($\ell\in[\![1,m]\!]$), it is possible to demonstrate the same property for the mean descent vector

$$\left.\begin{aligned}&\exists\ell\in[1,m], \ f_\ell(\mathbf{x},W) - f_\ell(\mathbf{x}^\perp,W) \ge \tfrac{c_\ell}{2}\|\mathbf{x}-\mathbf{x}^\perp\|^2\\ &\forall j\in[1,m]\setminus\{\ell\}, \ f_j(\mathbf{x},W) - f_j(\mathbf{x}^\perp,W) \ge 0\end{aligned}\right. a.s..$$

It follows immediately that

$$\xi^\star(\mathbf{x},W)(\mathbf{x}-\mathbf{x}^\perp) \ge \alpha_\ell(\mathbf{x},W)\frac{c_\ell}{2}\|\mathbf{x}-\mathbf{x}^\perp\|^2 \, a.s.,$$

and therefore,

$$\Xi^\star(\mathbf{x})(\mathbf{x}-\mathbf{x}^\perp) \ge \mathbb{E}[\alpha_\ell(\mathbf{x},W)]\frac{c_\ell}{2}\|\mathbf{x}-\mathbf{x}^\perp\|^2.$$

3.3.1. Mean square convergence of SMGDA

We introduce the filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$ where the σ-algebras are generated by the k first random variables of the sequence $(W_n)_{n\in\mathbb{R}^n}$

$$\mathcal{F}_k = \sigma(W_1,\dots,W_k).$$

By construction, the random variable $\mathbf{X}_k$ is $\mathcal{F}_k$-measurable for any $k\in\mathbb{N}$. From now on, we denote the common descent vector $\xi^\star(\mathbf{X}_k,W_{k+1})$ by the notation $\xi_k^\star$ and we use $\Xi_k^\star$ for the mean descent vector $\Xi^\star(\mathbf{X}_k)$.

Lemma 4.

$$\mathbb{E}\big[\langle\mathbf{X}_k-\mathbf{X}_k^\perp, \xi_k^\star-\Xi_k^\star\rangle|\mathcal{F}_k\big] = 0 \, a.s.$$

Proof. This results directly from Lemma 3 since the random variable W_{k+1} is independent from the σ-algebra $\mathcal{F}_k$

$$\mathbb{E}\big[\langle\mathbf{X}_k-\mathbf{X}_k^\perp, \xi_k^\star-\Xi_k^\star\rangle|\mathcal{F}_k\big] = \varphi(\mathbf{X}_k) \, a.s.,$$

with φ the function defined by

$$\forall\mathbf{x}\in\mathbb{R}^n, \ \varphi(\mathbf{x}) = \langle\mathbf{x}-\mathbf{x}^\perp, \mathbb{E}[\xi^\star(\mathbf{x},W_{k+1})] - \Xi^\star(\mathbf{x})\rangle.$$

The conclusion follows from the definition of the mean descent vector

$$\Xi^\star(\mathbf{x}) = \mathbb{E}_{W_{k+1}}[\xi^\star(\mathbf{x}, W_{k+1})].$$

□

Theorem 3. *The sequence of random variables* $\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_n$ *constructed using the* SMGDA *algorithm converges in mean square towards a point* $\mathbf{X}^\perp$ *of the Pareto set* $\mathcal{P}_D^\star$

$$\lim_{k\to+\infty} \mathbb{E}[\|\mathbf{X}_k - \mathbf{X}_k^\perp\|^2] = 0.$$

Proof. Let $\mathcal{L}_k^\perp$ denote the square distance between $\mathbf{X}_k$ and one of its closest point in $\mathcal{P}_D^\star$: $\mathbf{X}_k^\perp$

$$\mathcal{L}_k^\perp = \|\mathbf{X}_k - \mathbf{X}_k^\perp\|^2.$$

As $\mathbf{X}_{k+1}^\perp$ is one of the closest point of $\mathcal{P}_D^\star$ to $\mathbf{X}_{k+1}$, we have

$$\mathcal{L}_{k+1}^\perp \le \|\mathbf{X}_{k+1} - \mathbf{X}_k^\perp\|^2.$$

We now introduce the recurrence relation which describes the *SMGDA* algorithm

$$\forall k \in \mathbb{N}, \ \ \mathbf{X}_{k+1} = \mathbf{X}_k - \epsilon_k \xi_k^\star$$

into the latest relation

$$\begin{aligned}\mathcal{L}_{k+1}^\perp &\le \|\mathbf{X}_k - \epsilon_k\xi_k^\star - \mathbf{X}_k^\perp\|^2 \\ &\le \mathcal{L}_k^\perp + \epsilon_k^2\|\xi_k^\star\|^2 - 2\epsilon_k\langle \mathbf{X}_k - \mathbf{X}_k^\perp, \xi_k^\star\rangle.\end{aligned}$$

Adding the null term

$$\langle \mathbf{X}_k - \mathbf{X}_k^\perp, \Xi_k^\star - \Xi_k^\star\rangle$$

to the right hand side of the last relation

$$\mathcal{L}_{k+1}^\perp \le \mathcal{L}_k^\perp + \epsilon_k^2\|\xi_k^\star\|^2 - 2\epsilon_k\langle \mathbf{X}_k - \mathbf{X}_k^\perp, \xi_k^\star - \Xi_k^\star + \Xi_k^\star\rangle. \qquad (29)$$

Using the results of Propositions 1 and 2, we obtain

$$\mathcal{L}_{k+1}^\perp \le \mathcal{L}_k^\perp(1-\epsilon_k C) + \epsilon_k^2 M_{\xi^\star}^2 - 2\epsilon_k\langle \mathbf{X}_k - \mathbf{X}_k^\perp, \xi_k^\star - \Xi_k^\star\rangle.$$

We then take the conditional expectation of the expression with respect to the element $\mathcal{F}_k$ of the filtration $(\mathcal{F})$

$$\mathbb{E}[\mathcal{L}_{k+1}^\perp|\mathcal{F}_k] \le \mathbb{E}[\mathcal{L}_k^\perp(1-\epsilon_k C) + \epsilon_k^2 M_{\xi^\star}^2 - 2\epsilon_k\langle \mathbf{X}_k - \mathbf{X}_k^\perp, \xi_k^\star - \Xi_k^\star\rangle|\mathcal{F}_k].$$

Since the random variable $\mathbf{X}_k$ is $\mathcal{F}_k$-measurable, we can write

$$\mathbb{E}[\mathcal{L}_{k+1}^\perp|\mathcal{F}_k] \le \mathcal{L}_k^\perp(1-\epsilon_k C) + \epsilon_k^2 M_\xi^2 - 2\epsilon_k\mathbb{E}[\langle \mathbf{X}_k - \mathbf{X}_k^\perp, \xi_k^\star - \Xi_k^\star\rangle|\mathcal{F}_k].$$

Introducing the result of Lemma 4 yields the following relation

$$\mathbb{E}[\mathcal{L}_{k+1}^\perp|\mathcal{F}_k] \le \mathcal{L}_k^\perp(1-\epsilon_k C) + \epsilon_k^2 M_{\xi^\star}^2.$$

Since $\mathbb{E}[\,\mathbb{E}[\mathcal{L}_{k+1}^\perp|\mathcal{F}_k]\,] = \mathbb{E}[\mathcal{L}_{k+1}^\perp]$,

$$\mathbb{E}[\mathcal{L}_{k+1}^\perp] \le \mathbb{E}[\mathcal{L}_k^\perp](1-\epsilon_k C) + \epsilon_k^2 M_{\xi^\star}^2.$$

Considering the above relation for N consecutive terms, we have

$$\mathbb{E}[\mathcal{L}_{k+1+N}^\perp] \le \mathbb{E}[\mathcal{L}_k^\perp]\prod_{\ell=k}^{k+N}(1 - C\epsilon_\ell) + \sum_{\ell=k}^{k+N}\epsilon_\ell^2 M_{\xi^\star}^2. \qquad (30)$$

The proof of the convergence follows from the fact that the two sequences $\sum_{\ell=k}^{k+N}\epsilon_\ell^2 M_{\xi^\star}^2$ and $\prod_{\ell=k}^{k+N}(1-C\epsilon_\ell)$ converge towards 0, the first one because $(\epsilon_k)_{k\in\mathbb{N}}$ is a σ-sequence, the second one because of the convergence of its logarithm image. Finally we have proved that

$$\lim_{N\to\infty}\left(\mathbb{E}[\mathcal{L}_{k+1+N}^\perp]\right) = 0$$

which proves the mean square convergence theorem. □

A convergence speed result is also available for the mean square convergence.

Theorem 4. *Let* $\mathbf{X}_0$ *be an initial design point for the stochastic optimization problem. If the following* σ*-sequence is used in the algorithm*

$$\epsilon_k = \frac{1}{kC/2 + \dfrac{M_\xi^2}{\mathbb{E}[\mathcal{L}_0^\perp]C/2}},$$

then

$$\mathbb{E}[\mathcal{L}_k^\perp] \le \frac{1}{\dfrac{(C/2)^2}{M_\xi^2}k + \dfrac{1}{\mathbb{E}[\mathcal{L}_0^\perp]}}.$$

The proof of this convergence speed is exactly the same as the one given in Dodu et al. (1981) for mono-objective problems and will not be recalled here.

Remark 5. It can be seen that the convergence speed depends on the chosen σ-sequence.

3.3.2. *Almost sure convergence of* SMGDA

Theorem 5. *The sequence of random variables* $\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_n$ *constructed using the* SMGDA *algorithm converges almost surely towards a point* $\mathbf{X}^\perp$ *of the Pareto set* $\mathcal{P}_D^\star$

$$\mathbb{P}\left(\left\{\lim_{k\to\infty}(\mathbf{X}_k - \mathbf{X}_k^\perp) = 0\right\}\right) = 1.$$

Proof. Let $(Y_k)_{k\in\mathbb{N}}$ be the random sequence defined by

$$Y_k = \|\mathbf{X}_k - \mathbf{X}_k^\perp\|^2 + M_{\xi^\star}^2\sum_{\ell\ge k}\epsilon_\ell^2.$$

Using the inequality (29), we can write

$$\begin{aligned}Y_{k+1} &\le \|\mathbf{X}_k - \mathbf{X}_k^\perp\|^2 + \epsilon_k^2 M_{\xi^\star}^2 \\ &+ \sum_{\ell\ge k+1}\epsilon_\ell^2 M_{\xi^\star}^2 - 2\epsilon_k\langle \mathbf{X}_k - \mathbf{X}_k^\perp, \xi_k^\star - \Xi_k^\star + \Xi_k^\star\rangle.\end{aligned}$$

Proposition 2 allows us to bound from above by 0 the last term of this relation

$$-2\epsilon_k\langle \mathbf{X}_k - \mathbf{X}_k^\perp, \Xi_k^\star\rangle \le -\epsilon_k C\|\mathbf{X}_k - \mathbf{X}_k^\perp\|^2 \le 0,$$

which leads to the following inequation

$$Y_{k+1} \le \|\mathbf{X}_k - \mathbf{X}_k^\perp\|^2 + \epsilon_k^2 M_{\xi^\star}^2 + \sum_{\ell\ge k+1}\epsilon_\ell^2 M_{\xi^\star}^2 - 2\epsilon_k\langle \mathbf{X}_k - \mathbf{X}_k^\perp, \xi_k^\star - \Xi_k^\star\rangle.$$

Taking the conditional expectation of Y_{k+1} with respect to the σ-algebra $\mathcal{F}_k$, we obtain

$$\mathbb{E}[Y_{k+1}|\mathcal{F}_k] \le \mathbb{E}[Y_k|\mathcal{F}_k] - 2\epsilon_k\mathbb{E}[\langle \mathbf{X}_k - \mathbf{X}_k^\perp, \xi_k^\star - \Xi_k^\star\rangle|\mathcal{F}_k].$$

Knowing that Y_k is $\mathcal{F}_k$-mesurable and using the Lemma 4 we finally obtain the following expression

$$\mathbb{E}[Y_{k+1}|\mathcal{F}_k] \le Y_k.$$

The random process $(Y_k)_{k\in\mathbb{N}}$ is a supermartingale which is obviously positive. Therefore, using Theorem 2, the random process $(Y_k)_{k\in\mathbb{N}}$ converges almost surely toward a random variable Y_∞. Using Fatou's lemma, we can now bound the random variable Y_∞ by the following expression

$$\begin{aligned}0 \le \mathbb{E}[\lim_{k\to\infty}\left(\inf_{\ell\ge k} Y_\ell\right)] = \mathbb{E}[Y_\infty] &\le \lim_{k\to\infty}\left(\inf_{\ell\ge k}\mathbb{E}[Y_\ell]\right) \\ &\le \lim_{k\to\infty}\left(\mathbb{E}[\mathcal{L}_k^\perp] + \sum_{\ell\ge k}\epsilon_\ell^2 M_{\xi^\star}^2\right).\end{aligned}$$

The mean square convergence and the fact that the second term is the remainder of the 2*nd* order series of (ϵ_k) allow us to deduce that

$$\mathbb{E}[Y_\infty] = 0.$$

Knowing that (Y_k) is a positive random process implies that $Y_\infty = 0$ almost surely

$$\mathbb{P}\left(\lim_{k\to\infty} Y_k = 0\right) = 1.$$

□

4. Illustrations

The efficiency and the reliability of the method is assessed by comparing the solution obtained by *SMGDA* and by two other solvers: *NSGA-II* (Deb, Pratap, Agarwal, & Mayarivan, 2002), and *DMS* (Custodió, Madeira, Vaz, & Vincente, 2011) on several classical deterministic benchmark problems described in Huband, Hingston, Barone, and While (2006). The problems are chosen in order to present different situations : convex, nonconvex and discontinuous Pareto sets. Uncertainties are added to each problem by introducing random variables into the objective functions. Since the expectations appearing in the optimization problem described by Eq. (24) are not available, a sample average approximation approach is used for *NSGA-II* and *DMS* in order to evaluate them:

$$\mathbb{E}[f(\mathbf{x}, W)] \approx \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}, w_i), \tag{31}$$

where w_i are independent samples of the random variable W. The number N of samples plays a crucial role in the efficiency of the algorithm: a too small value results in a wide confidence interval and a poor estimate of the objective function, while an excessive value results in a dramatic increase of the computational cost. The test cases are conducted taking into account a budget based on the maximum number of calls to the objective functions. In order to compare the performance of the three algorithms tested in this section, two classical indicators are introduced: a performance indicator called Purity (Bandyopadhyay, Pal, & Aruna, 2004), which compares the number of non-dominated points an algorithm is able to find to a reference front built using the results of the three optimizers, and the well known Hypervolume indicator (Auger, Bader, Brockhoff, & Zitzler, 2009) which gives an indication on both the spreading of the front and its quality by calculating the sum of the hypervolumes generated between all non-dominated points and a reference point taken in the objective space. This last metric is illustrated in Fig. 4, where the reference point is represented by the symbol ■, and where the hypervolume corresponds to the green area. For both indicators the higher the score is the better the performance is. The tuning of *NSGA-II* and *DMS* parameters (including the number of samples used for the sample average approximation) is not straightforward and we used an auxiliary genetic algorithm in order to find the parameter values which maximizes the resulting Hypervolume measure for each problem. Due to the stochastic nature of the problems the fitness considered is the mean value of the resulting Hypervolume measure estimated on a sample of 5 independent runs for the same set of parameter values. The tuning optimizer is run for a population of 20 individuals and 15 generations.

In this section we shall begin by presenting in details the *MOP2* test case : the exact formulation of the problem is written, the insertion of the uncertainties and the results of each algorithm are detailed. The results of the other test cases are analyzed using performance profiles which gives an indication of performance on the overall set of problems (Dolan & Moré, 2002).

4.1. MOP2 problem

The first test case presented is a randomized version of the *MOP2* test case (Huband et al., 2006). It involves two non convex objective functions and fifteen design variables:

$$\min_{\mathbf{x}\in[-4,4]^{15}} \left\{ \begin{array}{l} \mathbb{E}[f_1(\mathbf{x}, W)] = \mathbb{E}\left[1 - \exp\left(-\sum_{i=1}^{15}(\mathbf{x}_i - \frac{1+W_{1i}}{\sqrt{15}})^2\right)\right] \\ \mathbb{E}[f_2(\mathbf{x}, W)] = \mathbb{E}\left[1 - \exp\left(-\sum_{i=1}^{15}(\mathbf{x}_i + \frac{1+W_{2i}}{\sqrt{15}})^2\right)\right] \end{array} \right\}, \tag{32}$$

where $\{W_{1i}\}_{i\in\{1,\,15\}}$ and $\{W_{2i}\}_{i\in\{1,\,15\}}$ are thirty independent random variables with uniform distribution on the interval $[-0.7, 0.7]$. We shall denote W the random vector $(W_{1,\,i};\ W_{2,\,i})\ i\in\{1,\ 15\}$.

The reference point considered for defining the Hypervolume indicator is the point $f_{ref} = (1.1, 1.1)$. As it is explained above, *NSGA-II* and *DMS* parameters have been optimized using a genetic algorithm in order to maximize the Hypervolume measure calculated with f_{ref} as the reference point. Ten independent runs of *NSGA-II* and *DMS* with the optimized parameters are generated. The run with the highest hypervolume is the only one considered in the results. A budget of 10^5 calls to the objective functions is allocated for both *NSGA-II* and *DMS* algorithms while only 10^4 calls are allocated to *SMGDA* and they are shared over 10^2 initial points. Once the three algorithms are stopped, a last estimation of the final mean performance is done using a sample average approximation with 10^5 samples. For this particular problem, *NSGA-II* parameters are set to a population of 211 individuals over 236 generations. The crossover probability is tuned to 92%, and the crossover index to 1.05. The mutation index parameter has shown to have a very low influence on the Hypervolume measure result. Thus it has not been considered in the parameter optimization and was set to 10 in all problems. The sample average approximation uses 2 samples which is very small, but it gives the best Hypervolume measure for the 10^5 allowed calls. The *NSGA-II* results are expected to have a low Purity score as accuracy will be impacted by the small number of samples used for the expectation approximations. The number of samples for *DMS* was set to 45 which allows to use more than 2000 calls to the objective functions. No step size criterion was imposed for *DMS* algorithm. The critical tuning of *SMGDA* lies in its σ-sequence that rules the step size. For the *MOP2* problem,we used the sequence

$$\epsilon_k = \frac{10^{-1}}{k+1},$$

k being the algorithm iteration index.

Looking at the Pareto front found by the three algorithm and illustrated on Fig. 1, *SMGDA* seems to give the best results: most of the solutions found by *SMGDA* are dominating the solutions provided by the other solvers. This is confirmed by looking at the Purity metric illustrated in Fig. 2(a). Due to the gradient low values of the objective functions far from the front, some points of *SMGDA* have not been able to converge for the allowed budget. Analyzing the Hypervolume metric results for this particular problem, *SMGDA* is able to give results very close to the maximum obtained by combining the solutions of the three algorithms together which define the reference Pareto front used for characterizing the Purity metric.

In order to illustrate the impact of uncertainties on the solutions, the probability distribution of the *SMGDA* solutions was also built. For each converged design point $\mathbf{x}^*$ in the Pareto set, 10^5 independent samples of the random functions $f_i(\mathbf{x}^*, W)$ have been generated, each sample yielding a Pareto front, solution of a new optimization problem. A Gaussian kernel density estimator was then used in order to build the probability distribution of the random Pareto front. It can be seen on Fig. 3 that this distribution has two peaks located at each edge of the front and that it is rather widely spread out in the middle. This indicates that the uncertain parameters have a much greater impact on the solution of

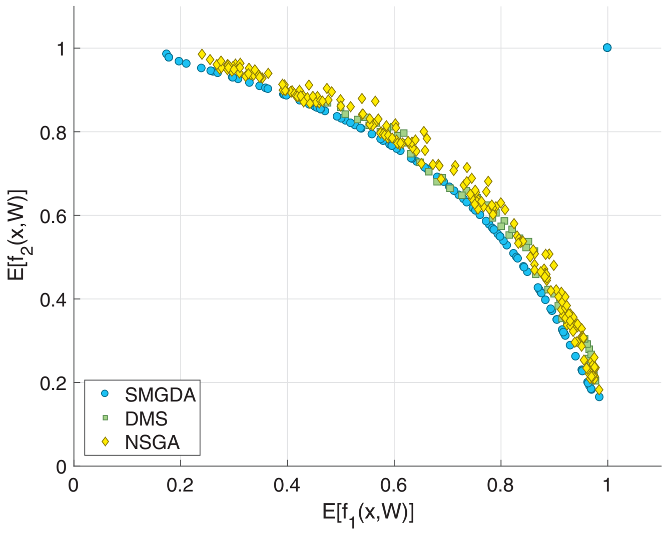


Fig. 1. Pareto fronts given by the three solvers for the *MOP2* problem.

Table 1
Benchmark problems.

Problem	Design variable number	Random variable number	*SMGDA* calls	*NSGA-II* & *DMS* calls
MOP2	15	30	10^4	10^5
MOP3	2	18	10^3	10^4
MOP6	2	3	10^3	10^4
ZDT{1,2,3}	30	32	5.10^3	5.10^4
JOS1	30	60	5.10^3	5.10^4
JOS2	30	32	5.10^3	5.10^4
SCH1	1	4	10^3	10^4
IM1	2	3	10^3	10^4

the multiobjective optimization problem than on each single objective optimization problem.

4.2. Additional numerical tests

The three algorithms are now compared using several other benchmark tests described in Table 1. The last two columns give the number of total calls budget during the optimization process for each algorithm tested. The number of calls allowed for *SMGDA* is set to be 10 times less than for *DMS* and *NSGA-II*.

In this section the Pareto front is not drawn for each problem, but performance profiles using Purity and Hypervolume metrics are used in order to compare the performance of the three algorithms. The performance profiles correspond to a cumulative distribution function that gives an indication of the percentage of problems considered solved for a certain threshold τ of the ratio

$$r_{p,s} = \frac{t_{p,s}}{\min\{t_{p,\bar{s}},\ \bar{s} \in \mathcal{S}\}},$$

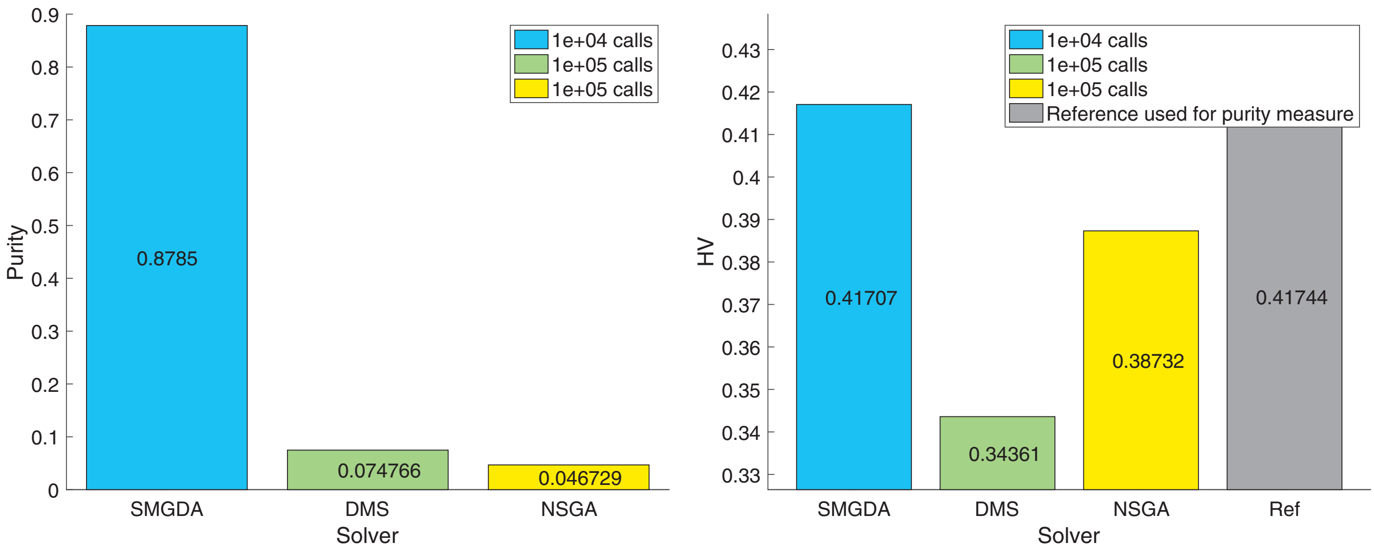
where p is an optimization problem belonging to the set $\mathcal{P}$ of benchmark problems addressed, s is the solver used in $\mathcal{S} = \{SMGDA\ , DMS\ , NSGA\text{-}II\ \}$ and t is a performance indicator for which a lower score indicates a better performance. Thus, in this section, it is actually the inverse of Purity and Hypervolume metrics which are used. This leads, for each solver, to the following expression of the cumulative distribution function $\rho(\tau)$

$$\rho_s(\tau) = \frac{1}{|\mathcal{P}|} \times |\{p \in \mathcal{P},\ r_{p,s} \le \tau\}|.$$

Thus, for each solver s, the value of $\rho_s(1)$ is the number of problems for which the performance of algorithm s is superior to the other two.

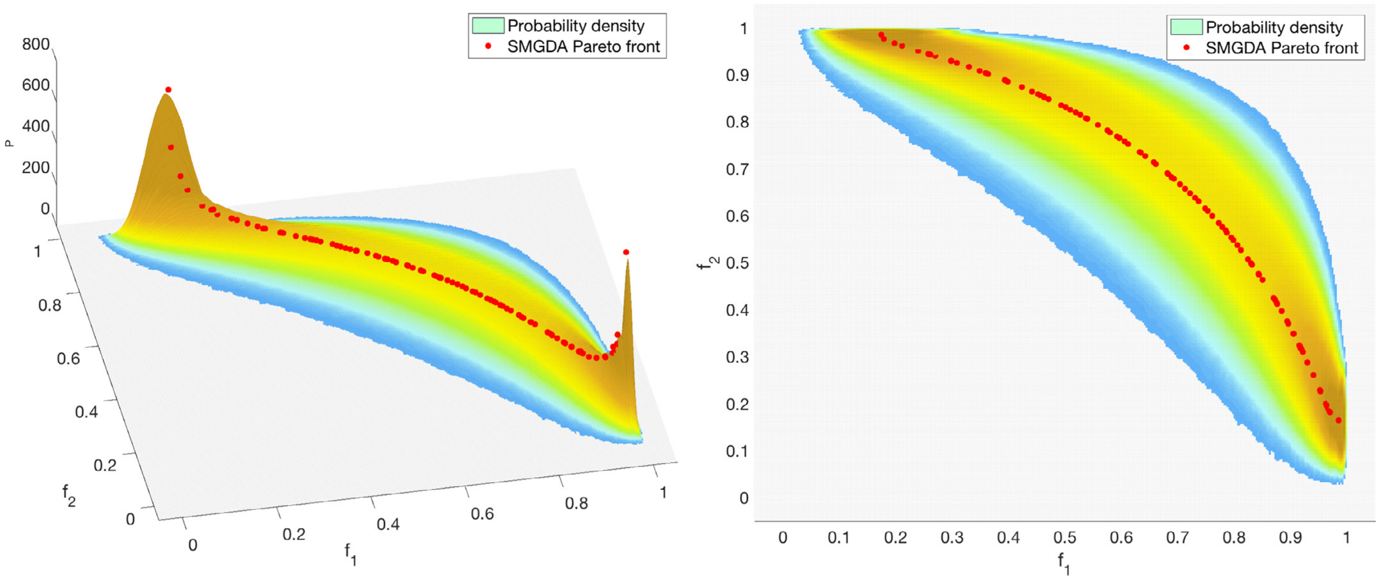
The set of problems studied covers a large range of optimization problems with convex and non convex objective functions. Two problems with a multimodal objective function are considered in this benchmark (*MOP6*, *ZDT3*). Because these two last problems have Pareto stationary points which are not Pareto optimal, *SMGDA* converges to a Pareto optimal point only when the initialization permits to avoid local minima. This can be observed on Fig. 4 and shows that *SMGDA* should be used with care when dealing with multimodal objectives.

For τ equal to 1, *SMGDA* outperforms the other two algorithms. Regarding the performance profiles based on the Purity metric represented on Fig. 5, *SMGDA* has a better performance for eight test problems and is the only algorithm able to reach $\rho_s = 1$. Which means that *NSGA-II* and *DMS* have not been able generate any non-dominated point on some of the addressed problems for the



(a) Purity (b) Hypervolume measure

Fig. 2. Performance indicators for *MOP2* problem.



(a) Side view

(a) Top view

Fig. 3. Density of probability of the Pareto front given by the *SMGDA* algorithm.

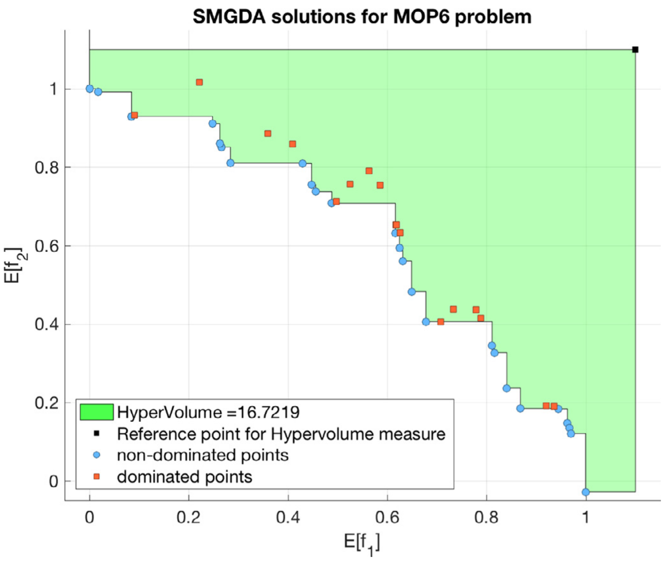


Fig. 4. *SMGDA* solutions for *MOP6* problem and its Hypervolume measure. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

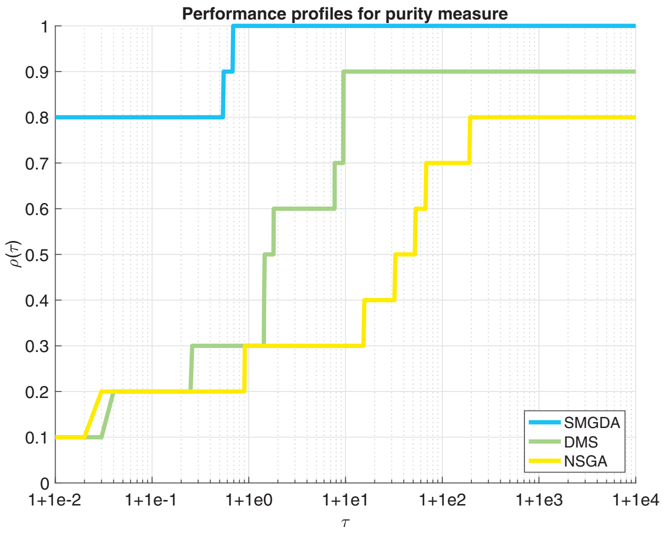


Fig. 5. Performance profile of the Purity indicator.

allocated budget. Because *SMGDA* may converge towards Pareto stationary but not optimal points, *SMGDA* obtains a good score using the Purity metric for all the problems except for the two multimodal test cases. But the algorithm reach nevertheless the value of $\rho_s = 1$ for $\tau = 2$ which means that the performance of *SMGDA* in terms of Purity measure was found inferior to the best performance by a factor of 2 only. Whereas, for certain test cases, *DMS* (resp. *NSGA-II*) can result in a Purity score up to 10 (resp. 100) times lower than the winning algorithm. This demonstrates the capability of *SMGDA* to perform well and to give good quality results for all the benchmark problems addressed.

Since the parameters of both *DMS* and *NSGA-II* have been optimized specifically for the Hypervolume metric, the two algorithms show a better performance profile than the ones constructed for the Purity metric as it is illustrated in Fig. 6. *NSGA-II* algorithm, especially, outperforms the other two algorithms for three test problems. Even if the *SMGDA* performance is lower for the Hypervolume metric, it is nevertheless rather efficient since the performance reaches $\rho_s = 1$ for a very low value of τ. For its less performing test case, *SMGDA* performance score was only 1.031 times less than the best algorithm performance.

The performance profiles presented in this section show that *SMGDA* can compete successfully with classical algorithms used for multiobjective optimization problems, at least for the two performance metric introduced. The numerical efficiency of *SMGDA* comes mainly from the fact that no estimator construction is necessary to evaluate the objective functions. Moreover the algorithm efficiency does not depend on the number of random variables introduced in the objective functions nor on the number of objective functions. The weakness of the method is the necessity to have the gradient analytic expressions, their numerical calculation would of course increase the computation time.

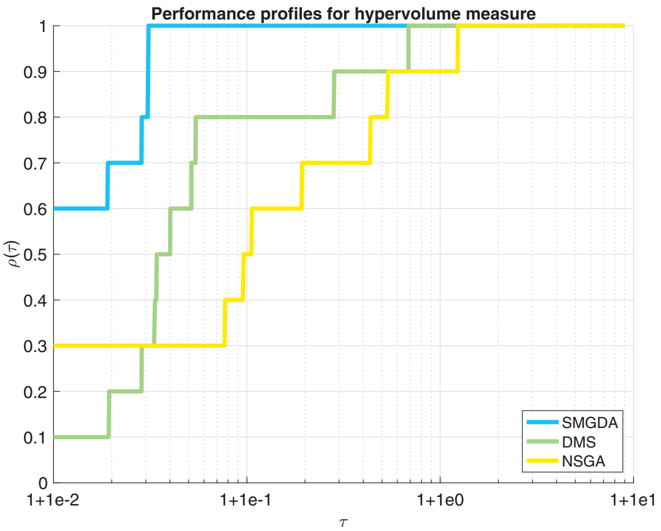


Fig. 6. Performance profile of the Hypervolume indicator.

5. Conclusions

In this article, we have proposed a novel algorithm for solving a stochastic multiobjective optimization problem. It is based on two ingredients: a common random descent vector and an extension of the classical stochastic gradient algorithm. Because the algorithm necessitates only a single iteration loop, it is less time demanding than classical approaches based on sample averaging approximation methods. Two types of convergence have been proved based on rather restrictive assumptions. As it is the case for stochastic gradient algorithms, no efficient stopping criterion exists. Comparisons with *NSGA-II* and *DMS* on a set of benchmark problems have shown the very good behaviour of the proposed method which requires much less iterations to converge than the two other solvers tested. Compared to a genetic algorithm there is no exchange of information between the initial points, which may a priori seem to yield a suboptimal decision but which renders the algorithm entirely and readily parallelizable: the computation time can be divided by the number of threads. Of course the objective functions needs to be regular but the approach should be easily extended to nonregular objective functions considering descent direction obtained by subgradients. We are actually working along this path.

References

Arnaud, R., & Poirion, F. (2014). Optimization of an uncertain aeroelastic system using stochastic gradient approaches. *Journal of Aircraft, 51*(3), 1061–1066.

Auger, A., Bader, J., Brockhoff, D., & Zitzler, E. (2009). Theory of hyervolume indicator: optimal μ-distributions and the choice of the reference point. In Proceeding of the tenth ACM SIGEVO workshop on foundations of genetic algorithm, (pp. 87–102).

Bandyopadhyay, S., Pal, S. K., & Aruna, B. (2004). Multiobjective GAs, quantitive indices and pattern classification. *IEEE Transaction on Systems, Man , and Cybernetics – Part B: Cybernetics, 34*(5), 2088–2099.

Bellman, R. E., & Zadeh, L. A. (1970). Decision-making in a fuzzy environment. *Management Science, 17*(4), B-141–B-164.

Bonnel, H., & Collonge, J. (2014). Stochastic optimization over a pareto set associated with a stochastic multi-objective optimization problem.. *Journal of Optimization Theory and Applications, 162*(2), 405–427.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Caballero, R., Cerdá, E., del Mar Muñoz, M., & Rey, L. (2004). Stochastic approach versus multiobjective approach for obtaining efficient solutions in stochastic multiobjective programming problems. *European Journal of Operational Research, 158*(3), 633–648.

Custodió, A. L., Madeira, J. F. A., Vaz, A. I. F., & Vincente, L. N. (2011). Direct multisearch for multiobjective optimization. *SIAM Journal on Optimization, 21*(3), 1109–1140.

Dantzig, G. B. (2004). Linear programming under uncertainty. *Management Science, 50*(12_supplement), 1764–1769.

Deb, K., & Gupta, H. (2005). Searching for robust Pareto-optimal solutions in multi-objective optimization. In Proceedings of the Evolutionary Multi-Criterion Optimization. EMO, 3410, 150–164.

Deb, K., Pratap, A., Agarwal, S., & Mayarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transaction on Evolutionary Computation, 6*(2), 181–197.

Désidéri, J. A. (2009). Multiple-gradient descent algorithm (MGDA. *Technical report no. 6953)*. Inria: Sophia Antipolis.

Désidéri, J. A. (2012). Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathematique, 3810*(5), 229–332.

Désidéri, J. A. (2014). Multiple-gradient descent algorithm (*MGDA*) for Pareto-front identification. In W. Fitzgibbon, W. Kuznetsov, Y. Neittaanmäki, & O. Pironneau (Eds.), *Modeling, simulation and optimization for science and technology* (pp. 41–58). Dordrecht, Netherlands:: Springer,Dordrecht.

Désidéri, J. A. (2015). Révision de l'algorithme de descente à gradients multiples (MGDA) par orthogonalisation hiérarchique. *Technical report no. 8710)*.. Inria: Sophia Antipolis.

Désidéri, J. A. (2018). Quasi-Riemannian Multiple Gradient Descent Algorithm for constrained multiobjective differential optimization. *Technical report no. 9159)*. Inria: Sophia Antipolis.

Désidéri, J. A., & Duvigneau, R., 2017. Parametric optimization of pulsating jets in unsteady flow by multiple-gradient descent algorithm (MGDA). In B.N. Chetverushkin, W. Fitzgibbon, Y.A. Kuznetsov, P. Neittaanmäki, J. Periaux, & O. Pironneau (Eds.), Contributions to partial differential equations and applications, computational methods in applied sciences. https://www.springer.com/us/book/9783319783246.

Diaz, J. E., Handl, J., & Xu, D. L. (2017). Evolutionary robust optimization in production planning: interactions between number of objectives, sample size and choice of robustness measure. *Computers & Operations Research, 79*, 266–278.

Dodu, J. C., Goursat, M., Hertz, A., Quadrat, J. P., & Viot, M. (1981). Méthodes de gradient stochastique pour l'optimisation des investissements dans un réseau électrique. *EDF, Bulletin de la Direction des études et recherches, Série C, Mathématique, Informatique, 2*, 133–167.

Dolan, E. D., & Moré, J. J. (2002). Benchmarking optimization software with performance profiles. *Mathematical Programming, 91*(2), 201–213.

Ermoliev, Y., & Wets, R. J. B. (1988). *Numerical techniques for stochastic optimization*. NY: Springer-Verlag New York.

Ermoliev, Y. M., & Gaivoronski, A. A. (1992). Stochastic quasigradient methods for optimization of discrete event systems. *Annals of Operations Research, 39*(1), 1–39.

Fliege, J., & Svaiter, B. F. (2000). Steepest descent methods for multicriteria optimization. *Mathematical Methods of Operations Research, 51*(3), 479–494.

Fliege, J., & Werner, R. (2014). Robust multiobjective optimization & applications in portfolio optimization. *European Journal of Operational Research, 234*(2), 422–433.

Fliege, J., & Xu, H. (2011). Stochastic multiobjective optimization: Sample average approximation and applications. *Journal of Optimization Theory and Applications, 151*(1), 135–162.

Gabrel, V., Murat, C., & Thiele, A. (2014). Recent advances in robust optimization: An overview. *European Journal of Operational Research, 235*(3), 471–483.

Huband, S., Hingston, P., Barone, L., & While, L. (2006). A review of multi-objective test problems and a scalable test problem toolkit. *IEEE Transaction on Evolutionary Computation, 10*(5), 477–506.

Klamroth, K., Köbis, E., Schöbel, A., & Tammer, C. (2017). A unified approach to uncertain optimization. *European Journal of Operational Research, 260*(2), 403–420.

Löhndorf, N. (2016). An empirical analysis of scenario generation methods for stochastic optimization. *European Journal of Operational Research, 255*(1), 121–132.

Matthies, H. G., Brenner, C. E., Bucher, C. G., & Soares, C. G. (1997). Uncertainties in probabilistic numerical analysis of structures and solids-stochastic finite elements. *Structural Safety, 19*(3), 283–336.

Mattson, C. A., & Messac, A. (2005). Pareto frontier based concept selection under uncertainty, with visualization. *Optimization and Engineering, 6*(1), 85–115.

Nemirovski, A., & Shapiro, A. (2006). Scenario approximations of chance constraints. In G. Calafiore, & F. Dabbene (Eds.), *Probabilistic and randomized methods for design under uncertainty* (pp. 3–47). London, UK:: Springer, London.

Papadimitriou, C., Katafygiotis, L. S., & Au, S. K. (1997). Effects of structural uncertainties on TMD design: A reliability-based approach. *Journal of Structural Control, 4*(1), 65–88.

Polyak, B. (1976). Convergence and convergence rate of iterative stochastic algorithms. *Automatica i Telemekhanika, 12*, 83–94.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics, 22*(3), 400–407.

Roy, R., Hinduja, S., & Teti, R. (2008). Recent advances in engineering design optimisation: Challenges and future trends. *CIRP Annals, 57*(2), 697–715.

Sahinidis, N. V. (2004). Optimization under uncertainty: state-of-the-art and opportunities. *Computers & Chemical Engineering, 28*(6), 971–983.

Shapiro, A. (2003). Monte carlo sampling approach to stochastic programming. *ESAIM: Proceedings, 13*, 65–73.

Wang, Z., Guo, J., Zheng, M., & Wang, Y. (2015). Uncertain multiobjective traveling salesman problem. *European Journal of Operational Research, 241*(2), 478–489.