Comput Optim Appl (2017) 68:317–331
DOI 10.1007/s10589-017-9921-x

Descent algorithm for nonsmooth stochastic multiobjective optimization

Fabrice Poirion[1] · **Quentin Mercier**[1] · **Jean-Antoine Désidéri**[2]

Received: 23 September 2016 / Published online: 28 June 2017

Abstract An algorithm for solving the expectation formulation of stochastic nonsmooth multiobjective optimization problems is proposed. The proposed method is an extension of the classical stochastic gradient algorithm to multiobjective optimization using the properties of a common descent vector defined in the deterministic context. The mean square and the almost sure convergence of the algorithm are proven. The algorithm efficiency is illustrated and assessed on an academic example.

Keywords Multiobjective optimization · Stochastic · Nonsmooth · Almost sure convergence

1 Introduction

The gradient algorithm or steepest descent method can be used to minimize any real value differentiable function defined on an Euclidean space (or more generally on an Hilbert space). When the function is convex the method converges towards a global minimum. Even if its numerical behaviour may present some drawbacks—slow convergence, oscillations, the method relies on rigorous convergence theorems and has given birth to more elaborate and efficient methods (conjugate gradient and Newton

✉ Fabrice Poirion
poirion@onera.fr

Quentin Mercier
quentin.mercier@onera.fr

Jean-Antoine Désidéri
jean-antoine.desideri@inria.fr

1 ONERA The French Aerospace Lab, Palaiseau, France

2 INRIA, Nice, France

methods). Very rapidly appeared the need to override the smoothness assumption of the objective function. Indeed many real life problems involve nonsmooth objectives: minimize the maximum constraint or vibration eigen-frequency for a structure, optimize a structure topology with contact constraints, etc. Non regular analysis has allowed to develop a theoretical framework to address the nonsmooth optimization problem: the subgradient plays the role of the gradient in the descent method. In addition real-life problems deal frequently with uncertain parameters. The gradient algorithm has been generalized to take into account uncertainties when they are modelled as random variables. Again convergence of the stochastic gradient algorithm can be shown [9]. For instance in [1] the authors use such an algorithm to solve a small industrial problem taking into account material uncertainties. Regarding smooth multiobjective optimization, Désidéri [7,8] has extended the gradient algorithm using a common descent vector built from the convex hull spanned by each objective gradient. Using the same argument Wilppu et al. [15] has generalized Désidéri's MGDA algorithm to nonsmooth objective functions deriving a common descent direction. In [6] the authors consider the same deterministic multiobjective optimization problem for quasi-convex objective functions and use a generalization of the scalar-valued subgradient method.

The purpose of the paper is to propose a method to construct the set of Pareto stationary points of a stochastic multiobjective optimization problem written in terms of the mean objective functions for nonsmooth objective functions. Convergence will be proved and an illustration given. The paper is organized as follow. In Sect. 2 various definitions and results on nonsmooth analysis are recalled. In Sect. 3 we introduce the problem under consideration and introduce the stochastic multi subgradient descent algorithm (SMSGDA). Then two types of convergence will be given. The last section is devoted to an illustration of the algorithm.

2 Preliminaries

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

Definition 1 A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\lambda \in [0, 1]$ the following inequality holds:

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}).$$

Definition 2 The directional derivative at $\mathbf{x}$ along the direction $\mathbf{v} \in \mathbb{R}^n$ of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined by the limit :

$$f'(\mathbf{x}; \mathbf{v}) = \lim_{t \downarrow 0} \frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t}.$$

Any convex function f is continuous and differentiable almost everywhere. Moreover there exists at each point $\mathbf{x}$ a lower affine function which is identical to f at $\mathbf{x}$. This affine function defines the equation of a plane called tangent plane. When the function f is differentiable at $\mathbf{x}$ there is only one tangent plane defined by the gradient

$\nabla f(\mathbf{x})$. When f is nondifferentiable at $\mathbf{x}$ there exists an infinity of tangent planes which define the subdifferential.

Definition 3 The subdifferential of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{x}$ is the set

$$\partial f(\mathbf{x}) = \{\mathbf{s} \in \mathbb{R}^n : f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{s}, \mathbf{y} - \mathbf{x} \rangle \ \forall \mathbf{y} \in \mathbb{R}^n\}. \tag{1}$$

This set is nonempty, convex, closed and reduced to $\nabla f(\mathbf{x})$ when f is differentiable.

The next result allows to use the notion of subdifferential for characterizing optimums.

Theorem 1 [2] *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. The following statements are equivalent*

1. *f is minimized at $\mathbf{x}^*$: $f(\mathbf{y}) \geq f(\mathbf{x}^*)\ \forall \mathbf{y} \in \mathbb{R}^n$,*
2. *$0 \in \partial f(\mathbf{x}^*)$,*
3. *$f'(\mathbf{x}^*, \mathbf{d}) \geq 0\ \forall \mathbf{d} \in \mathbb{R}^n$.*

A vector $\mathbf{d}$ is called a descent direction if $\exists t_0 > 0 \mid f(\mathbf{x} + t\mathbf{d}) < f(\mathbf{x})$ for all $t \in [0, t_0]$. A natural extension of the gradient algorithm would be to replace the gradient by any subgradient of the subdifferential. But not all subgradients define a descent direction. To override this difficulty one can use bundle methods [11,12]. In [15] the authors proposes an algorithm for constructing a descent vector of a convex nonsmooth function.

Multiobjective optimization is based on the notion Pareto optimality and weak Pareto optimality. Consider m convex functions $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots m$ and the unconstrained optimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \{f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})\}. \tag{2}$$

A solution $\mathbf{x}^*$ of problem (2) is Pareto optimal if no point $\mathbf{x}$ such that $f_i(\mathbf{x}) \leq f_i(\mathbf{x}^*)\ \forall i = 1, \ldots, m$ and $f_j(\mathbf{x}) < f_j(\mathbf{x}^*)$ for an index $j \in \{1, \ldots, m\}$ exists. It is weakly Pareto optimal if no point $\mathbf{x}$ such that $f_i(\mathbf{x}) < f_i(\mathbf{x}^*)\ \forall i = 1, \ldots, m$ exists. In the convex case the Pareto stationarity is equivalent to weak Pareto optimality. A complete review on multiobjective optimization can be found in [13].

3 The SMSGDA algorithm

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 first k 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

$\mathcal{F}$. We recall that a property depending on ω is said to be true almost surely (a.s.) if it is true for all values of ω except on a set of zero probability.

Definition 4 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 $\mathcal{F}_n$.

3.2 Problem statement

Consider m convex random functions $f_i : \mathbb{R}^n \times \mathcal{W} \to \mathbb{R}$, $i = 1, \dots m$. The problem addressed in this paper is to solve the mean multiobjective optimization problem:

$$\min_{x \in \mathbb{R}^n} \{\mathbb{E}[f_1(\mathbf{x}, W(\omega))], \mathbb{E}[f_2(\mathbf{x}, W(\omega))], \dots, \mathbb{E}[f_m(\mathbf{x}, W(\omega))]\}. \tag{3}$$

More precisely we want to construct the set of Pareto optimal solutions. As it is written, problem (3) 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, [3,10]. The algorithm we propose does not need to calculate the mean objective functions and is based only on the construction of a common descent vector.

In the smooth deterministic context, the existence of a common descent vector has been proved by Désidéri [8]: it is constructed as the minimum norm vector in the convex hull spanned by the gradients of the objective functions. In the nonsmooth context a natural extension is to consider the minimum norm element in the convex hull spanned by the union of each objective function subdifferential. We shall prove that this element, if it exists, is a common descent vector.

Lemma 1 *Let C be the convex hull of the union of all the subdifferentials $\partial f_i(\mathbf{x})$ of the objective functions appearing in problem (2). Then there exists a unique vector $\mathbf{p}^* = \text{argmin}_{p \in C} ||\mathbf{p}||$ such that*

$$\forall \mathbf{p} \in C \ : \ \langle \mathbf{p}, \mathbf{p}^* \rangle \geq \langle \mathbf{p}^*, \mathbf{p}^* \rangle = ||\mathbf{p}^*||^2.$$

Proof This result is a general property of closed convex sets and its proof is exactly the same as the one given in [7,15]. □

Before going on with the existence of a common descent vector we shall need to introduce the notion of Pareto stationary point.

Definition 5 A point $\mathbf{x}$ is said to be Pareto stationary if there exists a null convex combination of subgradients $\xi_i \in \partial f_i(\mathbf{x})$:

$$\exists \lambda_i \ ; \ \lambda_i \geq 0 \ ; \ \sum_{i=1}^{m} \lambda_i = 1 \ ; \ \mid \sum_{i=1}^{m} \lambda_i \xi_i = 0.$$

Theorem 2 *Let C be the convex set defined in Lemma 1 and $\mathbf{p}^*$ its minimum norm element. Then either we have*

1. $\mathbf{p}^* = 0$ *and the point* $\mathbf{x}$ *is Pareto stationary or*
2. $\mathbf{p}^* \neq 0$ *and the vector* $-\mathbf{p}^*$ *is a common descent direction for every objective function.*

Proof Since $\mathbf{p}^* \in C$, it can be written as a convex combination of elements of the union of subdifferentials $\partial f_i(\mathbf{x})$. Moreover, since a subdifferential is itself a convex set, this sum can be written as a convex combination involving a single element of each set $\partial f_i(\mathbf{x})$:

$$\mathbf{p}^* = \sum_{i=1}^{m} \lambda_i \xi_i^* \ ; \ \xi_i^* \in \partial f_i(\mathbf{x}).$$

with $\lambda_i \geq 0 \ ; \ \sum_{i=1}^{m} \lambda_i = 1$. Now if $\mathbf{p}^* = 0$ the point $\mathbf{x}$ is Pareto stationary by definition. Now if $\mathbf{p}^* \neq 0$, since it is the minimum norm element of the set C we have, using Lemma 1, $\langle \xi_i, \mathbf{p}^* \rangle \geq ||\mathbf{p}^*||^2 > 0$ for all $\xi_i \in \partial f_i(\mathbf{x})$. Therefore $\langle \xi_i, -\mathbf{p}^* \rangle < 0$ which implies that $-\mathbf{p}^*$ is a descent direction for each function $f_i, i = 1, \ldots, m$, (Theorem 4.5 in [2]). □

Corollary 1 *Let ω be given in Ω and consider the deterministic multiobjective optimization problem:*

$$\min_{\mathbf{x} \in \mathbb{R}^n} \{f_1(\mathbf{x}, W(\omega)), f_2(\mathbf{x}, W(\omega)), \ldots, f_m(\mathbf{x}, W(\omega))]\}. \tag{4}$$

Then either a point $\mathbf{x}$ *is Pareto stationary or there exists a common descent direction* $\mathbf{d}(\omega)$ *for the objective functions* $f_i(\mathbf{x}, W(\omega)), i = 1, \ldots, m$.

The descent vector depends of $\mathbf{x}$ and ω and will be considered as a random vector $d(\mathbf{x}, \omega)$ defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

3.3 The algorithm

We give now the successive steps of the algorithm that we propose.

1. Choose an initial point $\mathbf{x}_0$ in the design space, a number N of iterations and a σ-sequence $t_k : \sum t_k = \infty \ ; \ \sum t_k^2 < \infty$,
2. at each step k, draw a sample w_k of the random variable $W_k(\omega)$,
3. construct a descent vector $\mathbf{d}_i(\mathbf{x}_{k-1}, w_k)$ of the objective function $f_i(\mathbf{x}_{k-1}, w_k)$,
4. construct the common descent vector $\mathbf{d}(\mathbf{x}_{k-1}, w_k)$,
5. update the current point : $\mathbf{x}_k = \mathbf{x}_{k-1} + t_k \, \mathbf{d}(\mathbf{x}_{k-1}, w_k)$.

The last step of the algorithm defines a sequence of random variables on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ through the following relation

$$X_k(\omega) = X_{k-1}(\omega) - t_k\, d(X_{k-1}(\omega), W_k(\omega)). \tag{5}$$

Remark 1 We have not yet given any information on how the common descent direction is built. Except for some particular cases it is usually not possible to characterize the whole subdifferential of a function. We shall specify in the last section the method used to construct this direction based on subdifferential approximations.

The next section will be devoted to the convergence proofs of this algorithm.

3.4 Convergences

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 minimize the distance between the point $\mathbf{x}$ and a point of the Pareto set $\mathcal{P}_D{}^\star$.

$$\mathbf{x}^\perp \in \underset{\mathbf{u}\in\mathcal{P}_D^\star}{\operatorname{argmin}} \{\|\mathbf{x} - \mathbf{u}\|\}. \tag{6}$$

The convergence proofs are based on the following set of assumptions:

- H1 Problem (3) admits a nonempty Pareto solution set $\mathcal{P}_D^\star$.
- H2 The random variables $\omega \mapsto f_i(\mathbf{x}, W(\omega))$ are integrable for $i = 1, \ldots, m$ and $\mathbf{x} \in \mathbb{R}^n$.
- H3 Functions $\mathbf{x} \mapsto f_i(\mathbf{x}, W(\omega)) : \mathbb{R}^n \to \mathbb{R}$ are almost surely convex.
- H4 The sets $\partial f_i(\mathbf{x}, W(\omega))\ ;\ i = 1, \ldots, m$ are uniformly bounded :

$$\sup\{||s(\omega)||, s(\omega) \in \partial f_i(\mathbf{x}, W(\omega)), \mathbf{x} \in \mathbb{R}^n, \omega \in \Omega\} \leq q_i\,.$$

- H5 $\exists c_i \in \mathbb{R}^+,\ \forall(\mathbf{x}, w) \in \mathbb{R}^n \times \mathcal{W},\ f_i(\mathbf{x}, w) - f_i(\mathbf{x}^\perp, w) \geq \frac{c_i}{2}\|\mathbf{x} - \mathbf{x}^\perp\|^2\ ;\ i = 1, m$
- H6 The sequence $\{t_k\}_{k=1,m}$is a σ-sequence.

Theorem 3 *Under assumptions H1–H6 ,*

1. *the sequence of random variables $X_k(\omega)$ defined by relation (5) converges in mean square towards a point X^* of the Pareto set:*

$$\lim_{k\to+\infty} \mathbb{E}[\|X_k(\omega) - X^\star\|^2] = 0.$$

2. *The sequence converges almost surely towards $X^\star$.*

$$\mathbb{P}\left(\left\{\omega \in \Omega,\ \lim_{k\to\infty} X_k(\omega) = X^\star\right\}\right) = 1.$$

Before going on with the proof of this theorem we shall need two intermediate results.

Proposition 1 *The common descent vector d is almost surely bounded:*

$$\exists q,\ \forall x \in \mathbb{R}^n,\ \|d(\mathbf{x}, W(\omega))\| \leq q\ \ a.s.$$

Proof By construction d is a convex combination of descent vectors d_i:

$$\forall (\mathbf{x}, w) \in \mathbb{R}^n \times \mathcal{W}\ ;\ \|d(\mathbf{x}, w)\| = \left\| \sum_{i=1}^{m} \alpha_i(\mathbf{x}, w) d_i(\mathbf{x}, \xi) \right\|$$

with $0 \leq \alpha_i(\mathbf{x}, w) \leq 1$ and $\sum_i \alpha_i(\mathbf{x}, w) = 1$. Therefore

$$\|d(\mathbf{x}, w)\| \leq \sum_{i=1}^{m} \|\alpha_i(\mathbf{x}, w) d_i(\mathbf{x}, w)\| \leq \sum_{i=1}^{m} \|d_i(\mathbf{x}, w)\|.$$

Following assumption H4

$$\|d(\mathbf{x}, w)\| \leq \sum_{i=1}^{m} q_i\ ;\ \forall (\mathbf{x}, w) \in \mathbb{R}^n \times \mathcal{W}.$$

□

Proposition 2 *Let* $\bar{d}(\mathbf{x}) = \mathbb{E}[d(\mathbf{x}, W(\omega))]$ *be the mean common descent vector at point* $\mathbf{x}$. *Then the following relation holds:*

$$\forall \mathbf{x} \in \mathbb{R}^n,\ \langle \bar{d}(\mathbf{x}), (\mathbf{x} - \mathbf{x}^{\perp}) \rangle \geq \frac{c}{2} \|\mathbf{x} - \mathbf{x}^{\perp}\|^2$$

where $\mathbf{x}^{\perp}$ *denotes the one of the closest point* $\mathcal{P}_D^{\star}$ *from* $\mathbf{x}$ *and* $c = \min c_i$ *where the* c_i *are defined in H5*

Proof Following assumption H3, functions f_i are convex almost surely. From the definition of the subdifferential we can write:

$$\forall i = 1, m,\ \forall \mathbf{x} \in \mathbb{R}^n,\ \langle d_i(\mathbf{x}, W(\omega)), (\mathbf{x} - \mathbf{x}^{\perp}) \rangle \geq f_i(\mathbf{x}, W(\omega)) - f_i(\mathbf{x}^{\perp}, W(\omega))\ a.s.$$

Using assumption H5,

$$\langle d_i(\mathbf{x}, W(\omega)), (\mathbf{x} - \mathbf{x}^{\perp}) \rangle \geq \frac{c_i}{2} \|\mathbf{x} - \mathbf{x}^{\perp}\|^2\ a.s.$$

and

$$\sum_{i=1}^{m} \alpha_i(\mathbf{x}, W(\omega)) \langle d_i(\mathbf{x}, W(\omega)), (\mathbf{x} - \mathbf{x}^{\perp}) \rangle \geq \sum_{i=1}^{m} \alpha_i(\mathbf{x}, W(\omega)) \frac{c_i}{2} \|\mathbf{x} - \mathbf{x}^{\perp}\|^2.$$

Let $C = \inf_i c_i$, then

$$\sum_{i=1}^{m} \alpha_i(\mathbf{x}, W(\omega))\langle d_i(\mathbf{x}, W(\omega)), (\mathbf{x} - \mathbf{x}^{\perp})\rangle \geq \frac{C}{2}\|\mathbf{x} - \mathbf{x}^{\perp}\|^2 \sum_{i=1}^{m} \alpha_i(\mathbf{x}, W(\omega)).$$

Since $\sum \alpha_i = 1$,

$$\sum_{i=1}^{m} \alpha_i(\mathbf{x}, W(\omega))\langle d_i(\mathbf{x}, W(\omega)), (\mathbf{x} - \mathbf{x}^{\perp})\rangle \geq \frac{C}{2}\|\mathbf{x} - \mathbf{x}^{\perp}\|^2.$$

Therefore, taking the expected values of both sides:

$$\langle \bar{d}(\mathbf{x}), (\mathbf{x} - \mathbf{x}^{\perp})\rangle = \mathbb{E}\left[\sum_{i=1}^{m} \alpha_i(\mathbf{x}, W(\omega))\langle d_i(\mathbf{x}, W(\omega)), (\mathbf{x} - \mathbf{x}^{\perp})\rangle\right] \geq \frac{C}{2}\|\mathbf{x} - \mathbf{x}^{\perp}\|^2.$$

□

Remark 2 (Weaker hypothesis H5) The approach of keeping the same hypothesis as the single objective does not take into account the pre-order relation induced by the definition of Pareto optimality. This makes the hypothesis very strong, because the relation

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

is supposed true for all objectives ($j = 1, \ldots, m$). 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.

$$\begin{cases} \exists \ell \in [\![1, m]\!], \;\; f_\ell(\mathbf{x}, W) - f_\ell(\mathbf{x}, W) \geq \frac{c_\ell}{2}\|\mathbf{x} - \mathbf{x}^{\perp}\|^2 \\ \forall j \in [\![1, m]\!] \setminus \{\ell\}, \;\; f_j(\mathbf{x}, W) - f_j(\mathbf{x}, W) \geq 0 \end{cases} \quad a.s.$$

It follows immediately that

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

and therefore, taking the expectation of each side:

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

We can now prove the two types of convergence of the algorithm

Proof of the second order convergence Let $X_k(\omega)$ defined by recurrence relation (5) and $X_k^{\perp}(\omega)$ its projection over the Pareto front. We denote l_k the square distance $\|X_k - X_k^{\perp}\|^2$. We have to prove that $\lim_{k\to\infty} \mathbb{E}(l_k) = 0$.

We have trivially $l_{k+1} \leq \|X_{k+1} - X_k^{\perp}\|^2$. Using the recurrence relation yields the following relation:

$$\begin{aligned} l_{k+1} &\leq \|X_k - t_k d(X_k, W_{k+1}) - X_k^{\perp}\|^2 \\ &\leq l_k + t_k^2 \|d(X_k, W_{k+1})\|^2 - 2t_k \langle X_k - X_k^{\perp}, d(X_k, W_{k+1})\rangle, \end{aligned}$$

where $d(X_k, W_{k+1})$ is the common descent vector at point X_k for the optimization problem corresponding to the random drawing W_{k+1}. Using Proposition 1 this relation becomes:

$$l_{k+1} \leq l_k - 2t_k \langle X_k - X_k^{\perp}, d(X_k, W_{k+1})\rangle + q^2 t_k^2.$$

In order to compute $\mathbb{E}(l_k)$ we will use the classical result on conditional expectation $\mathbb{E}(l_{k+1}) = \mathbb{E}[\mathbb{E}(l_{k+1}|\mathcal{F}_k)]$. In order to compute $\mathbb{E}(l_{k+1}|\mathcal{F}_k)$ we shall use the following classical probability lemma [14].

Lemma 2 *Let $\mathcal{B} \subset \mathcal{A}$ be two σ-algebras 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:*

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

Indeed the random variable X_k is built from the random variables $W_1, \ldots, W_k$ and therefore is $\mathcal{F}_k$-measurable and the construction of the descent vector d involves the random variable W_{k+1} which is independent of the σ-algebra $\mathcal{F}_k$. Therefore

$$\mathbb{E}[\langle d(X_k, W_{k+1}), X_k - X_k^{\perp}\rangle|\mathcal{F}_k] = \langle \mathbb{E}_{W_{k+1}}[d(X_k, W_{k+1})], X_k - X_k^{\perp}\rangle,$$

where $\mathbb{E}_{W_{k+1}}$ denotes the expectation relatively to the distribution of W_{k+1}. Hence $\mathbb{E}_{W_{k+1}}[d(X_k, W_{k+1})]$ is the mean common descent vector at point X_k: $\bar{d}(X_k)$.

$$\mathbb{E}(l_{k+1}|\mathcal{F}_k) \leq l_k - 2t_k \langle X_k - X_k^{\perp}, \bar{d}(X_k)\rangle + q^2 t_k^2. \tag{7}$$

Using Proposition 2 and taking the expectation of both sides yields:

$$\mathbb{E}[l_{k+1}] \leq \mathbb{E}[l_k)(1 - ct_k) + q^2 t_k^2, \tag{8}$$

Iterating this relation m times:

$$\mathbb{E}[l_{k+m}] \leq \mathbb{E}[l_k] \prod_{i=0}^{m-1} (1 - ct_{k+i}) + q^2 \sum_{i=0}^{m-1} t_{k+i}^2.$$

The proof follows from the fact that the two sequences $\sum_{j=k}^{k+m} t_j^2 q^2$ and $\prod_{j=k}^{k+m} (1 - ct_j)$ converge towards 0, the first one because (t_j) is a σ-sequence, the second one follows from the convergence of its logarithm image. Finally we have proved that

$$\lim_{m \to \infty} (\mathbb{E}[l_{k+m}]) = 0,$$

which proves the mean square convergence theorem. □

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

$$Y_k = l_k + \sum_{i \geq k} t_i^2 q^2.$$

Taking the conditional expectation of both sides relatively to the σ-algebra $\mathcal{F}_k$

$$E[Y_{k+1} | \mathcal{F}_k] = E[l_{k+1} | \mathcal{F}_k] + \sum_{i \geq k+1} t_i^2 q^2.$$

Using the inequality (7) we can write,

$$E[Y_{k+1} | \mathcal{F}_k] \leq l_k^2 + q^2 t_k^2 \leq Y_k.$$

The random process $(Y_k)_{k \in \mathbb{N}}$ is a submartingale which is obviously positive. To conclude the proof we shall need the following result:

Theorem 4 [14] *Let $(Y_k)_{k \in \mathbb{N}}$ be a positive submartingale. Then there exists a random variable Y_∞ such that Y_k converges toward Y_∞ almost surely.*

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

Using the Fatou lemma, we can now bound the random variable Y_∞ by the following expression :

$$\begin{aligned} 0 \leq \mathbb{E}[\lim_{k \to \infty} \left(\inf_{j \geq k} Y_j \right)] = \mathbb{E}[Y_\infty] &\leq \lim_{k \to \infty} \left(\inf_{j \geq k} \mathbb{E}[Y_j] \right) \\ &\leq \lim_{k \to \infty} \left(\mathbb{E}[l_k] + \sum_{j \geq k} t_j^2 M_\xi^2 \right). \end{aligned}$$

The mean square convergence and the fact that the second term is the remainder of the 2nd order series of (t_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 Illustration

The algorithm is based on the construction of a common descent direction which necessitates to specify the subdifferentials of the objective functions. As said earlier the computation of a subdifferential is a difficult task and several approaches exist in order to approximate it, the most popular one being the bundle method [2,12]. In this application we will use an approximation proposed by Burke et al. [4,5] who, using the property that a generalized subdifferential can be related to limits of convex combinations of gradients at nearby points, have developed a gradient sampling algorithm for nonconvex single objective optimization. In practice, for the SMSGDA algorithm, the common descent direction is built at each step as the minimum norm element in the convex hull spanned by a set of N gradients for each objective function, calculated in the neighborhood of the current point.

The method will be illustrated on two simple numerical tests. The first one is constructed in order that the expectancies appearing in the mean problem (3) can be explicitly known. It will allow to assess the results of the stochastic algorithm using a deterministic algorithm such as the one proposed in [15]. We consider two objective functions f_1 and f_2 defined by:

$$f_1(x, y, \omega) = -.5 + U(\omega) \times |x + 1| + 0.5 \times (y + 1)^2 \tag{9}$$

$$f_2(x, y, \omega) = -.5 + \sqrt{(x - 1)^2} + V(\omega) \times (y - 1)^2, \tag{10}$$

where $U(\omega)$ and $V(\omega)$ are independent uniform random variables over the interval $[1. - \alpha, 1. + \alpha]$; $\alpha \geq 0$. We want to solve the following optimization problem:

$$\min_{x,y}\{\mathbb{E}[f_1(x, y, \omega)],\ \mathbb{E}[f_2(x, y, \omega)]\}. \tag{11}$$

For this first problem $\mathbb{E}[f_2(x, y, \omega)]$ and $\mathbb{E}[f_1(x, y, \omega)]$ are equal to the function f_2 and f_1 for the parameter value $\alpha = 0$. In a first step we construct the Pareto set of problem (11) using a deterministic algorithm and compare it with the Pareto set obtained by using SMSGDA. Figure 1 represents the two deterministic functions $\mathbb{E}(f_1)$ and $\mathbb{E}(f_2)$ as well as the corresponding Pareto set plotted in red (resp. green) on function f_1 (resp. $f2$) plot. Figure 2 compares in the objective space the Pareto set obtained by using the deterministic and stochastic algorithm. We can check that they are the same. The Pareto set for the perturbed problem obtained by considering three samples of the random variables U and V are also drawn on the same plot, illustrating the uncertainty effect. The σ sequence which was introduced in the algorithm is the following:

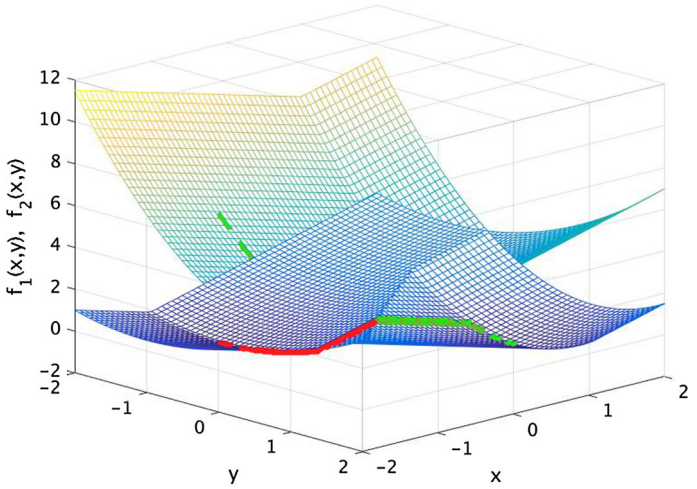

Fig. 1 Objective functions and Pareto solutions

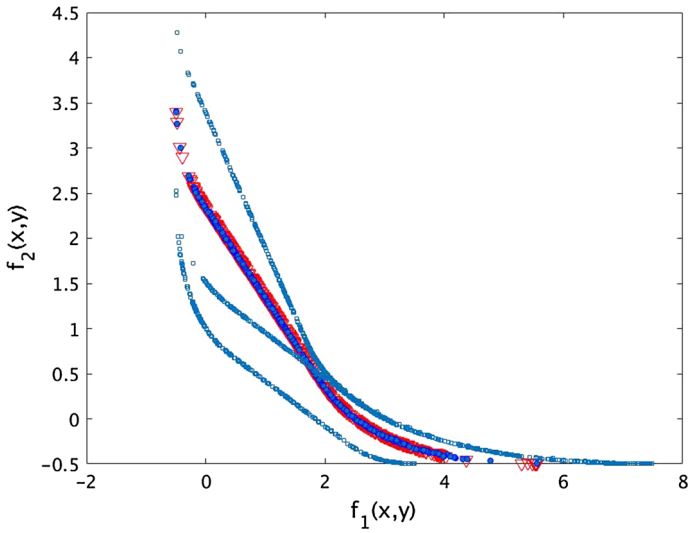

Fig. 2 Comparison of the Pareto set—*red inverted triangle* reference solution; *blue filled circle* SMSGDA; *green square* Pareto set samples (Color figure online)

$$t_k = a/(b + d \times k^c)\ ;\ a = 1, b = .1, c = .6, d = 1. \tag{12}$$

Four hundred initial points were uniformly distributed over $[-2, 2] \times [-2, 2]$ and 1000 iterations per points used with $\alpha = .3$. Each subdifferential at point $\mathbf{x}$ is approximated using eight gradients calculated in the ball $\mathbf{x} + 0.1 \times \mathcal{B}$ where $\mathcal{B} = \{\mathbf{x} \mid ||\mathbf{x}|| \leq 1\}$.

Figure 3 represents the Pareto set in the design space. In the second numerical test we add a third random variable and change the location of the random variables in order to introduce a nonlinear dependency between each function and the random parameters:

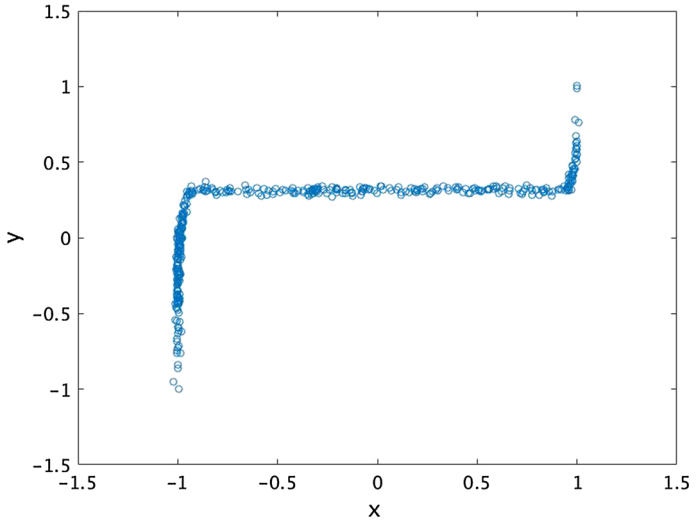

Fig. 3 Pareto set in the design space

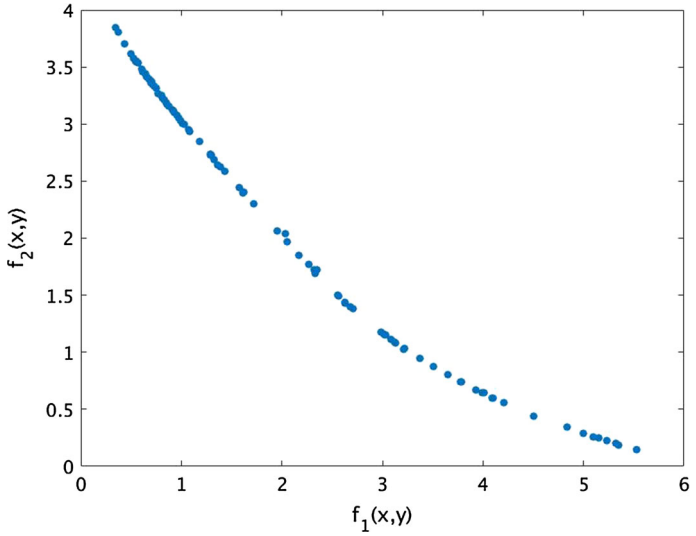

Fig. 4 Pareto set in the objective space

$$f_1(x, y, \omega) = -.5 + |x + 1 + U(\omega)| + 0.5 \times (y + 1 + V(\omega))^2 \tag{13}$$

$$f_2(x, y, \omega) = -.5 + \sqrt{(x + W(\omega))^2} \times (y - 1).^2, \tag{14}$$

where $W(\omega)$ is a $N(0, \beta)$ Gaussian variable. The same parameter values as in the previous illustration are used and $\beta = .3$. Figure 4 represents the Pareto set in the objective space. In order to evaluate the effect of uncertainty on the objective functions considered at the optimal design points (x^*, y^*) and on the corresponding Pareto front,

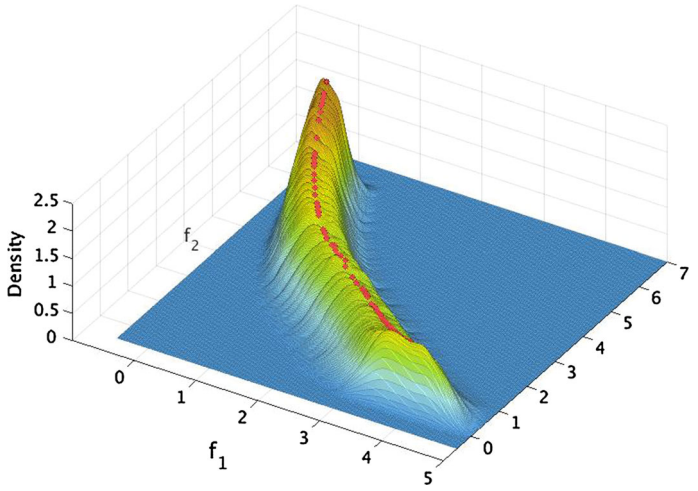

Fig. 5 Probability density of final solutions—*red filled square* Pareto set of the mean problem (Color figure online)

we have drawn on Fig. 4 the distribution of $[f_1(x^*, y^*, W), f_2(x^*, y^*, W)]$ along the Pareto front and we have superimposed the Pareto front of the mean problem (Fig. 5).

5 Conclusion

We have proposed a descent algorithm for stochastic nonsmooth unconstrained multi-objective optimization when the objective functions are convex. The descent vector is based on the common descent vector introduced in the work of Désidéri [8] and Wilppu et al. [15]. The descent step size is given by the values of a σ-sequence. Mean square and almost sure convergences were proved. A simple numerical test has shown the capability of such an algorithm. As it is the case with stochastic algorithms there exists no efficient stopping criteria for the algorithm. Because there is no exchange of information between the initial points the algorithm is entirely and readily parallelisable: one may divide the computation time by the number of cores.

References

1. Arnaud, R., Poirion, F.: Optimization of an uncertain aeroelastic system using stochastic gradient approaches. J. Aircr. **51**, 1061–1066 (2014)
2. Bagirov, A., Karmitsa, N., Mäkelä, M.: Introduction to Nonsmooth Optimization: Theory Practice and Software. Springer, Berlin (2014)
3. Bonnel, H., Collonge, J.: Stochastic optimization over a Pareto set associated with a stochastic multi-objective optimization problem. J. Optim. Theory Appl. **162**, 405–427 (2014)
4. Burke, J.V., Lewis, A.S., Overton, M.L.: Approximating subdifferentials by random sampling of gradients. Math. Oper. Res. **27**, 567–584 (2002)
5. Burke, J.V., Lewis, A.S., Overton, M.L.: A robust gradient sampling algorithm for nonsmooth, non-convex optimization. SIAM J. Optim. **15**, 751–779 (2005)

6. Da Cruz Neto, J., Da Silva, G., Ferreira, O., Lopes, J.: A subgradient method for multiobjective optimization. Comput. Optim. Appl. **54**, 461–472 (2013)
7. Désidéri, J.: Multi-gradient descent algorithm (MGDA). Technical report 6953, INRIA. (2009)
8. Désidéri, J.: Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. CRAS Paris Ser. I **350**, 313–318 (2012)
9. Duflo, M.: Random Iterative Models. Springer, Berlin (1997)
10. Fliege, J., Xu, H.: Stochastic multiobjective optimization: sample average approximation and applications. J. Optim. Theory Appl. **151**, 135–162 (2011)
11. Kiwiel, K.: Methods of Descent for Nondifferentiable Optimization, Lecture Notes in Mathematics, vol. 1133. Springer, Berlin (1985)
12. Mäkelä, M.: Survey of bundle methods for nonsmooth optimization. Optim. Methods Softw. **17**, 1–29 (2002)
13. Miettinen, K.: Nonlinear Multiobjective Optimization, International Series in Operations Research and Management Science, vol. 12. Springer, Berlin (1998)
14. Revuz, D., Yor, M.: Continuous Martingales and Brownian Motion. Springer, Berlin (1991)
15. Wilppu, O., Karmitsa, N., Mäkelä, M.: New multiple subgradient descent bundle method for nonsmooth multiobjective optimization. Technical report 1126, TUCS, University of Turku, Finland. (2014)