RANDOM PERTURBATION OF THE VARIABLE METRIC METHOD FOR UNCONSTRAINED NONSMOOTH NONCONVEX OPTIMIZATION

ABDELKRIM EL MOUATASIM*, RACHID ELLAIA*
JOSÉ E. SOUZA DE CURSI**

* Ecole Mohammadia d'Ingénieurs
LERMA, Avenue Ibn Sina BP765—Agdal, Rabat, Morocco
e-mail: {moatasim,ellaia}@emi.ac.ma

** LMR – UMR 6138 CNRS, INSA—Rouen
Avenue de l'Université BP 8, Saint-Etienne du Rouvray, France, FR-76831
e-mail: souza@insa-rouen.fr

We consider the global optimization of a nonsmooth (nondifferentiable) nonconvex real function. We introduce a variable metric descent method adapted to nonsmooth situations, which is modified by the incorporation of suitable random perturbations. Convergence to a global minimum is established and a simple method for the generation of suitable perturbations is introduced. An algorithm is proposed and numerical results are presented, showing that the method is computationally effective and stable.

Keywords: nonconvex optimization, stochastic perturbation, variable metric method, nonsmooth optimization, generalized gradient

1. Introduction

Nonsmooth optimization problems are often found in economics (Outrata, 1987; Outrata *et al.*, 1998), control (Mäkelä and Neittaanmäki, 1992; Batukhtin *et al.*, 1998) data analysis (Bagirov *et al.*, 2003), or design centering (Nguyen and Strodiot, 1992). In such problems, a typical situation is the determination of a *column vector* $\boldsymbol{x}^\star \in E = \mathbb{R}^n$ such that

$$\boldsymbol{x}^* = \arg\min_E f, \tag{1}$$

where the objective function $f : E \longrightarrow \mathbb{R}$ *does not* satisfy convexity assumptions, is continuous but *not* differentiable on a finite or countable subset of E (this is the case when, e.g., f is assumed to be only locally Lipschitz continuous, cf. (Mäkelä and Neittaanmäki, 1992)). The problem (1) is that of unconstrained optimization, since no limitation is imposed on the design variables $\boldsymbol{x}$. In practice, lower and upper bounds the variables are often introduced:

$$\ell_i \le x_i \le u_i, \quad \ell_i \in \mathbb{R} \cup \{-\infty\}, \quad u_i \in \mathbb{R} \cup \{+\infty\},$$

$i = 1, \ldots, n$.

In this case, we set $\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_n)^t$ and $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)^t$, and we look for the minimum in a box region:

$$\begin{aligned} \boldsymbol{x}^* &= \arg\min_S f, \\ S &= [\boldsymbol{\ell}, \boldsymbol{u}] = \{\boldsymbol{x} \in E \mid \boldsymbol{\ell} \le \boldsymbol{x} \le \boldsymbol{u}\}. \end{aligned} \tag{2}$$

Although some restrictions on the values of the design variables have been introduced, we shall also consider the problem (2) as a member of a family of unconstrained ones. In fact, the numerical methods for its solution are strictly analogous to those of (1). In the sequel, our model problem will be (3).

Due to the lack of differentiability, derivatives of the objective function f are not available and specific methods must be considered, such as generalized gradients or bundle methods. Both these approaches are based on the idea of *local underevaluating affine functions* (overevaluating affine functions may also be considered), i.e., the construction of a family $\{\varphi_\lambda\}_\lambda$ of affine functions such that $\varphi_\lambda \le f$ on a subset $E_\lambda \subset E$ (usually, E_λ is a neighbourhood of a given point x_λ). Different approaches may be found in the literature for the construction of such a family, but a simple approach consists in using generalized gradients such as Clarke's ones (note that, in general, subgradients are not anywhere defined for nonconvex f).

Bundle methods construct approximate underevaluations of the objective function by gathering the general-

ized gradients obtained in previous iterations into a bundle. A search direction for the objective function can be determined by solving a quadratic direction finding problem (Mākelā and Neittaanmāki, 1992) and, in convex optimization, global convergence of bundle methods with limited past information can be guaranteed by using aggregation strategies (Kiwiel, 1994) in order to accumulate the information furnished by previous iterations (Hiriart-Urruty and Lemaréchal, 1993; Lemaréchal *et al.*, 1981; Schramm and Zowe, 1992).

Generalized gradient methods construct underevaluations or overevaluations by considering affine functions supported by the generalized gradient, and extend the methods of smooth optimization to the nonsmooth situation by using a generalized gradient $\boldsymbol{p}$ instead of the gradient ∇f, which may be undefined. Consider a standard descent method starting at an initial guess $\boldsymbol{x}_0$ and generating a sequence of points $\{\boldsymbol{x}_k\}_{k\geq 0} \subset S$ as follows:

$$\boldsymbol{x}_0 \in S \text{ given, } \forall k \geq 0: \quad \boldsymbol{x}_{k+1} = \boldsymbol{Q}\left(\boldsymbol{x}_k, \boldsymbol{d}_k, \omega_k\right), \quad (3)$$

where

$$\boldsymbol{Q}\left(\boldsymbol{x}, \boldsymbol{d}, \omega\right) = \boldsymbol{x} + \omega \boldsymbol{d}, \quad (4)$$

$\boldsymbol{d}_k \in E$ is a descent direction (i.e., $f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k)$) and $\omega_k \in \mathbb{R}$ ($\omega_k \geq 0$ and) is a step length. It is expected that $\omega_k > 0$ and $\boldsymbol{d} \neq \boldsymbol{0}$ while convergence is not obtained. Usually, the descent direction is obtained from information furnished by gradients. For instance, the classical steepest descent uses $\boldsymbol{d}_k = -\boldsymbol{g}_k$, where $\boldsymbol{g}_k = \boldsymbol{\nabla} f(\boldsymbol{x}_k)$. When $\boldsymbol{\nabla} f(\boldsymbol{x}_k)$ is not defined but a generalized gradient $\boldsymbol{p}_k$ at the point $\boldsymbol{x}_k$ is available, we may use $\boldsymbol{g}_k = \boldsymbol{p}_k$ in order to get a descent direction: the method extends directly to the nonsmooth situation. In practice, subgradients or gradients at points close to $\boldsymbol{x}_k$ may also be used (Clarke, 1975).

At present, subgradient methods are regarded as the most effective and reliable methods for nonsmooth unconstrained problems (Boyd *et al.*, 2003; Ellaia, 1992; Hiriart-Urruty and Lemaréchal, 1993; Kiwiel, 1985; Larsson *et al.*, 1996). Due to their simple structure, they are widely used, but several difficulties persist in nonsmooth optimization. For instance, directions generated by using generalized gradients may lead to nondescent search directions, and arbitrarily generalized gradients do not necessarily become small in the neighbourhood of an optimal point, which makes the definition of stopping criteria difficult. In addition, the convergence of the sequence $\{\boldsymbol{x}_k\}_{k\geq 0}$ to a global minimum $\boldsymbol{x}^\star$ is not ensured under the lack of convexity.

In this work, we consider *variable metric methods* (Fletcher, 1980), which have been proven efficient and robust for nonsmooth problems (Lemaréchal, 1982; Uryasev, 1991). In these methods, the descent direction is usually given by

$$\boldsymbol{d}_k = \boldsymbol{u}_k\left(\boldsymbol{x}_k, \boldsymbol{x}_{k-1}, \boldsymbol{B}_k\right), \quad (5)$$

where $\boldsymbol{B}_k$ is an $n \times n$ matrix. In order to simplify the analysis, we assume that there is a constant $M \in \mathbb{R}^+$ (see also Remark 1) for which we have

$$\|\boldsymbol{d}_k\| \leq M. \quad (6)$$

The determination of a step $\omega_k \geq 0$ often involves a one-dimensional search and a previously established maximal step $\overline{\omega}$. For instance, the optimal step is

$$\begin{aligned} &\omega_k = \arg\min_W f\left(\boldsymbol{x}_k + \omega \boldsymbol{d}_k\right), \\ &W = \left\{\omega \mid \boldsymbol{x}_k + \omega \boldsymbol{d}_k \in S, \;\; 0 \leq \omega \leq \overline{\omega}\right\}. \end{aligned} \quad (7)$$

Accordingly, the step is given by a function $\omega : E \times E \to \mathbb{R}$, so that

$$\omega_k = \omega\left(\boldsymbol{x}_k, \boldsymbol{d}_k\right), \quad 0 \leq \omega\left(\boldsymbol{x}_k, \boldsymbol{d}_k\right) \leq \overline{\omega}. \quad (8)$$

We shall use the optimal step defined above in our numerical experiments (cf. Section 4).

In order to prevent the algorithm from convergence to a local minimum and the above-mentioned difficulties concerning the descent direction in the nonconvex framework, we introduce a controlled random search (Bouleau, 1986; Carson and Maria, 1997; Dorea, 1990; Ellaia and Elmouatasim, 2004; Pogu and Souza, 1994; Souza de Cursi, 1992a; Souza de Cursi *et al.*, 2003; 2005; Wang and Spall, 1999). In this approach, the sequences $\{\boldsymbol{x}_k\}_{k\geq 0}$, $\{\boldsymbol{d}_k\}_{k\geq 0}$, $\{\boldsymbol{\omega}_k\}_{k\geq 0}$ become those of random variables $\{\boldsymbol{X}_k\}_{k\geq 0}$, $\{\boldsymbol{D}_k\}_{k\geq 0}$, $\{\Omega_k\}_{k\geq 0}$ and the descent iterations are modified as follows:

$$\begin{aligned} &\boldsymbol{X}_0 = \boldsymbol{x}_0 \in S \text{ given,} \\ &\qquad \forall k \geq 0: \boldsymbol{X}_{k+1} = \boldsymbol{Q}\left(\boldsymbol{X}_k, \boldsymbol{D}_k, \Omega_k\right) + \boldsymbol{P}_k, \end{aligned} \quad (9)$$

$$\boldsymbol{D}_k = \boldsymbol{u}_k\left(\boldsymbol{X}_k, \boldsymbol{X}_{k-1}\right), \quad (10)$$

$$\Omega_k = \omega\left(\boldsymbol{X}_k, \boldsymbol{D}_k\right), \quad 0 \leq \omega\left(\boldsymbol{X}_k, \boldsymbol{D}_k\right) \leq \overline{\omega}, \quad (11)$$

where $\boldsymbol{P}_k$ is a suitable random vector of stochastic perturbation. A convenient choice of $\{\boldsymbol{P}_k\}_{k\geq 0}$ ensures the convergence of the sequence $\{\boldsymbol{X}_k\}_{k\geq 0}$ to a solution (see Section 4). If $\boldsymbol{x}^\star$ is not unique, the limit random variable $\boldsymbol{X} = \lim_{k\to+\infty} \boldsymbol{X}_k$ describes the set of solutions (i.e., $\boldsymbol{X}$ is supported by the set of solutions). A practical implementation of (9)–(11) involves finite samples of $\boldsymbol{P}_k$ that will be detailed in Section 4.3: only a finite number of $\boldsymbol{X}_k$ are generated and the iterations lead to an approximation of one from among all of the global minimizers.

2. Notation and Assumptions

We denote by $\mathbb{R}$ the set of real numbers $(-\infty, +\infty)$. $\mathbb{R}^+$ stands for the set of nonnegative real numbers $[0, +\infty)$ and $E = \mathbb{R}^n$ means the n-dimensional real Euclidean space. For $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^t \in E$, $\boldsymbol{x}^t$ denotes

the transpose of $\boldsymbol{x}$. We denote by $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^t\boldsymbol{x}} = (x_1^2 + \cdots + x_n^2)^{1/2}$ the Euclidean norm of $\boldsymbol{x}$, and by $(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{x}^t\boldsymbol{y}$ the scalar product on E. We shall denote by $\boldsymbol{Id}$ the $n \times n$ identity matrix.

The objective function is $f : S \longrightarrow \mathbb{R}$, and we assume that f has a lower bound in S, which is denoted by l^*:

$$\min_S f = l^* \in \mathbb{R}. \tag{12}$$

Let us introduce $S_\alpha = C_\alpha \cap S$, $C_\alpha = \{\boldsymbol{x} \in E \mid f(\boldsymbol{x}) \leq \alpha\}$. We assume that

$$f \text{ is locally Lipschitz continuous on } S, \tag{13}$$

$$\forall \alpha > l^* : \ S_\alpha \text{ is not empty, closed and bounded,} \tag{14}$$

$$\forall \alpha > l^\star : \ \mathrm{meas}\,(S_\alpha) > 0, \tag{15}$$

where $\mathrm{meas}\,(S_\alpha)$ is the measure of S_α.

Since E is a finite dimensional space, the assumption (14) is satisfied, e.g., for either S bounded or f coercive, i.e.,

$$\lim_{\|\boldsymbol{x}\| \to +\infty} f(\boldsymbol{x}) = +\infty.$$

Equation (15) is equivalent to the assumption that S contains a sequence of neighbourhoods of an optimal point $\boldsymbol{x}^\star$ having a strictly positive measure, i.e., $\boldsymbol{x}^\star$ can be approximated by a sequence of points of the interior of S. As has previously been remarked, $\boldsymbol{x}^\star$ is not assumed to be unique and the iterations generate approximations of one from among all of the global minimizers.

From (6), (8) and (14), we get

$$\beta(\alpha,\overline{\omega}) = \sup\big\{\, \|\boldsymbol{y} - \boldsymbol{Q}\,(\boldsymbol{x},\boldsymbol{d},\omega)\| : (\boldsymbol{x},\boldsymbol{y}) \in S_\alpha \times S_\alpha,\ 0 \leq \omega \leq \overline{\omega}\big\} < +\infty. \tag{16}$$

In addition, we assume that

$$f \text{ is not constant in } S. \tag{17}$$

3. Generalized Gradient and Variable Metric Descent

As has already been mentioned, the objective function $f : E \longrightarrow \mathbb{R}$ is assumed to be locally Lipschitz continuous, i.e., at any point $\boldsymbol{x} \in E$, there exists a nonnegative scalar $K\,(\boldsymbol{x})$ and $\varepsilon\,(\boldsymbol{x}) > 0$ such that

$$\|\boldsymbol{y} - \boldsymbol{x}\| \leq \varepsilon\,(\boldsymbol{x}) \text{ and } \|\boldsymbol{y}' - \boldsymbol{x}\| \leq \varepsilon\,(\boldsymbol{x}) \Longrightarrow |f\,(\boldsymbol{y}) - f\,(\boldsymbol{y}')\,| \leq K\,(\boldsymbol{x})\,\|\boldsymbol{y} - \boldsymbol{y}'\|\,. \tag{18}$$

Nonsmooth Lipschitz continuous functions may have a countable number of points of nondifferentiability. In addition, *f is not* assumed to be convex: the subdifferential may be empty and generalized gradients, such as Clarke's ones, must be used. Recall that $D_c f(\boldsymbol{x})(\boldsymbol{v})$, the Clarke directional derivative of f at $\boldsymbol{x}$ along the direction $\boldsymbol{v}$, is defined by (Clarke, 1983):

$$D_c f(\boldsymbol{x})(\boldsymbol{v}) = \limsup_{y \to x,\ \theta \to 0+} \frac{f\,(\boldsymbol{y} + \theta\boldsymbol{v}) - f(\boldsymbol{y})}{\theta} = \inf_{\alpha,\theta}\left(\sup_{\|\boldsymbol{y}-\boldsymbol{x}\| \leq \alpha,\ 0<t<\theta} \frac{f\,(\boldsymbol{y}+t\boldsymbol{v}) - f(\boldsymbol{y})}{t}\right). \tag{19}$$

The generalized gradients are elements of the set

$$\partial f\,(\boldsymbol{x}) = \{\boldsymbol{p} \in E \mid (\boldsymbol{p},\boldsymbol{v}) \leq D_c f(\boldsymbol{x})(\boldsymbol{v}), \quad \forall \boldsymbol{v} \in E\}\,.$$

If f is continuously differentiable at $\boldsymbol{x}$, then $D_c f(\boldsymbol{x})(\boldsymbol{v}) = (\nabla f(\boldsymbol{x}),\boldsymbol{v})$ and we have $\partial f\,(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$. The generalized gradients are connected with affine underevaluations of f. Consider an affine function $\varphi : E \longrightarrow \mathbb{R}$ given by

$$\varphi\,(\boldsymbol{y}) = (\boldsymbol{p},\boldsymbol{y} - \boldsymbol{x}) + f(\boldsymbol{x}),$$

which underevaluates f on a ball $B_\eta\,(\boldsymbol{x})$ with a centre $\boldsymbol{x}$ and a radius $\eta > 0$:

$$\forall \boldsymbol{y} \in B_\eta\,(\boldsymbol{x}) : \ \varphi\,(\boldsymbol{y}) \leq f\,(\boldsymbol{y})\,.$$

Let B denote the ball with a centre $\mathbf{0}$ and a radius 1, and I_η the interval $(0,\eta)$, i.e.,

$$B = \{\boldsymbol{v} \in E \mid \|\boldsymbol{v}\| \leq 1\}\,, \quad I_\eta = (0,\eta)\,.$$

Then

$$\forall\,(\boldsymbol{v},t) \in B \times I_\eta\colon t\,(\boldsymbol{p},\boldsymbol{v}) + f(\boldsymbol{x}) \leq f\,(\boldsymbol{x}+t\boldsymbol{v})\,.$$

Thus, for $\theta > \eta$,

$$(\boldsymbol{p},\boldsymbol{v}) \leq \frac{f\,(\boldsymbol{x}+t\boldsymbol{v}) - f(\boldsymbol{x})}{t} \leq \sup_{\|\boldsymbol{y}-\boldsymbol{x}\| \leq \alpha,\ 0<t<\theta} \frac{f\,(\boldsymbol{y} + t\boldsymbol{v}) - f(\boldsymbol{y})}{t}$$

and, by (19),

$$(\boldsymbol{p},\boldsymbol{v}) \leq \limsup_{y \to x,\ \theta \to 0+} \frac{f\,(\boldsymbol{y} + \theta\boldsymbol{v}) - f(\boldsymbol{y})}{\theta} = D_c f(\boldsymbol{x})(\boldsymbol{v}).$$

Since both $\boldsymbol{v} \to D_c f(\boldsymbol{x})(\boldsymbol{v})$ and $\boldsymbol{v} \to (\boldsymbol{p},\boldsymbol{v})$ are positively homogeneous, this inequality shows that $\boldsymbol{p}$ is a generalized gradient. Analogous calculations can be performed for overevaluations. If

$$\forall \boldsymbol{y} \in B_\eta\,(\boldsymbol{x}) : \ \varphi\,(\boldsymbol{y}) \geq f\,(\boldsymbol{y})\,,$$

then $\boldsymbol{p}$ is a generalized gradient:

$$(\boldsymbol{p},\boldsymbol{v}) \leq D_c\,(-f)\,(\boldsymbol{x})(-\boldsymbol{v}) = D_c f(\boldsymbol{x})(\boldsymbol{v}).$$

As has previously been observed, these deliberations provide a simple way to extend descent methods based on the gradient to the nonsmooth situation under consideration: if the objective function f is differentiable at $\boldsymbol{x}_k$, the descent direction $\boldsymbol{d}_k$ is determined by using the standard gradient $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k)$. Otherwise, we consider a local affine underevaluation or overevaluation $\varphi_k(\boldsymbol{y}) = (\boldsymbol{p}_k, \boldsymbol{y} - \boldsymbol{x}_k) + f(\boldsymbol{x}_k)$, and we use $\boldsymbol{g}_k = \boldsymbol{p}_k$ for the determination of the descent direction. In practice, φ_k may be numerically approximated using the values of f or ∇f at points close to $\boldsymbol{x}_k$. This approach is particularly suitable for the situation under consideration, since f is differentiable almost everywhere (i.e., except in a set having a zero Lebesgue measure (Mākelā and Neittaanmāki, 1992)). For instance, if f is not differentiable at $\boldsymbol{x}_k$ and fails to be convex in the neighbourhood of $\boldsymbol{x}_k$, the evaluation may be performed at a different point, close to $\boldsymbol{x}_k$, where f is differentiable.

As has previously been observed, we consider *variable metric methods*, which are regarded as efficient and robust. This class of methods includes, e.g., the popular Davidon-Fletcher-Powell method (Davidon, 1991), where the descent direction is

$$\boldsymbol{u}_k(\boldsymbol{x}_k, \boldsymbol{x}_{k-1}, \boldsymbol{B}_k) = -\boldsymbol{B}_k \boldsymbol{g}_k,$$

and $\{\boldsymbol{B}_k\}_{k\geq 0}$ is a sequence of $n \times n$ matrices such that $\boldsymbol{B}_0 = \boldsymbol{Id}$ is the $n \times n$ identity matrix, and

$$\boldsymbol{B}_k = \boldsymbol{B}_{k-1} + \frac{(\boldsymbol{x}_k - \boldsymbol{x}_{k-1})(\boldsymbol{x}_k - \boldsymbol{x}_{k-1})^t}{(\boldsymbol{x}_k - \boldsymbol{x}_{k-1})^t (\boldsymbol{g}_k - \boldsymbol{g}_{k-1})} - \frac{\boldsymbol{B}_{k-1}(\boldsymbol{g}_k - \boldsymbol{g}_{k-1})(\boldsymbol{g}_k - \boldsymbol{g}_{k-1})^t \boldsymbol{B}_{k-1}}{(\boldsymbol{g}_k - \boldsymbol{g}_{k-1})^t \boldsymbol{B}_{k-1}(\boldsymbol{g}_k - \boldsymbol{g}_{k-1})}.$$

In the sequel, we consider a general variable metric descent method satisfying Eqns. (5)–(8). The mathematical results given in Section 4 are valid for any method satisfying these assumptions.

In our numerical examples, we shall consider a particular method where the descent direction is

$$\boldsymbol{u}_k(\boldsymbol{x}_k, \boldsymbol{x}_{k-1}, \boldsymbol{B}_k) = \begin{cases} -\dfrac{\boldsymbol{B}_k \boldsymbol{g}_k}{\|\boldsymbol{B}_k \boldsymbol{g}_k\|} & \text{if } \boldsymbol{B}_k \boldsymbol{g}_k \neq \boldsymbol{0}, \\ \boldsymbol{0} & \text{if } \boldsymbol{B}_k \boldsymbol{g}_k = \boldsymbol{0}, \end{cases} \tag{20}$$

and the sequence $\{\boldsymbol{B}_k\}_{k\geq 0}$ is formed by $\boldsymbol{B}_0 = \boldsymbol{Id}$ and $\boldsymbol{B}_k$ recursively defined from $\boldsymbol{B}_{k-1}$:

$$(\boldsymbol{B}_k)_{ij} = b_{step} \sum_{s=1}^{j-1} (\boldsymbol{B}_{k-1})_{is} \times \left[\sum_{l=1}^{n} (\boldsymbol{B}_{k-1})_{li} (\boldsymbol{g}_k)_i (\boldsymbol{g}_{k-1})_l + (\boldsymbol{g}_{k-1})_i (\boldsymbol{g}_k)_l \right], \quad 1 \leq i \leq n, \quad 1 \leq j \leq n, \tag{21}$$

where b_{step} is a step size with respect to the matrix B_{k-1} and $0 < b_{step} < 1$, see, e.g., Algorithm 7.2 in (Uryasev, 1991), where a recommended value is $b_{step} = \alpha_3 = 0.55$.

This choice leads to a descent method such that $\|\boldsymbol{d}_k\| \leq 1$, and the assumption (6) is thus satisfied.

The step ω_k has to be determined by an independent rule. The classical choices are, e.g., a fixed step, Wolfe's rule or an optimal step. In our calculations, we shall use the optimal step approach.

Remark 1. This approach extends to the alternative method where

$$\boldsymbol{u}_k(\boldsymbol{x}_k, \boldsymbol{x}_{k-1}, \boldsymbol{B}_k) = -\boldsymbol{B}_k \boldsymbol{g}_k.$$

Let f be Lipschitz of rank $K(\boldsymbol{x})$ near $\boldsymbol{x}$. Then

(a) $\partial f(\boldsymbol{x})$ is a nonempty, convex, weakly-compact subset of E, and $\|\boldsymbol{p}\| \leq K(\boldsymbol{x})$ for every $\boldsymbol{p}$ in $\partial f(\boldsymbol{x})$.

(b) For every $\boldsymbol{v}$ in E, we have that $D_c f(\boldsymbol{x})(\boldsymbol{v}) = \max\{(\boldsymbol{p}, \boldsymbol{v}) : \boldsymbol{p} \in \partial f(\boldsymbol{x})\}$,

see Proposition 2.1.2, p. 27, in (Clarke, 1983).

From (13) and (14), we get

$$\gamma_1 = \sup\{\|\boldsymbol{p}\| : \boldsymbol{x} \in C_\alpha, \boldsymbol{p} \in \partial f(\boldsymbol{x})\} < +\infty. \tag{22}$$

Using (22) and (21), it follows (by recurrence) that

$$\gamma_2 = \sup\{\|\boldsymbol{B}_k\| : \boldsymbol{x}_k \in C_\alpha\} < +\infty$$

and

$$M = \sup\{\|\boldsymbol{d}\| : \boldsymbol{x} \in C_\alpha\} < +\infty. \tag{23}$$

Thus, we have that Eqn. (16) as well as Lemmas 1 and 2 (see next section) remain valid.

4. Random Perturbation of the Variable Metric (RPVM)

The main difficulty remains the lack of convexity: if f is not convex, the Kuhn-Tucker points may not correspond to the global minima. In the sequel, we shall improve this by using an appropriate random perturbation: as has previously been observed, the real quantities are replaced by random variables (cf. Eqns. (9)–(11)). Since

$$\boldsymbol{Q}(\boldsymbol{X}_k, \boldsymbol{D}_k, \Omega_k) + \boldsymbol{P}_k = \boldsymbol{Q}(\boldsymbol{X}_k, \boldsymbol{D}_k + \boldsymbol{P}_k/\Omega_k, \Omega_k),$$

the stochastic iterations may be considered as perturbations in the descent direction $\boldsymbol{d}_k$. In the sequel, we describe the general properties of these elements leading to convenient sequences and we show that sequences of Gaussian vectors may be used.

4.1. General Properties of the Random Perturbation. A simple way for the generation of a convenient sequence of perturbations $\{\boldsymbol{P}_k\}_{k\geq O}$ is

$$\boldsymbol{P}_k = \xi_k \boldsymbol{Z}_k, \tag{24}$$

where

(1) $\{\xi_k\}_{k\geq 0}$ is a nonincreasing sequence of *strictly positive* real numbers *converging to zero* and such that $\xi_0 \leq 1$.

(2) $\{\boldsymbol{Z}_k\}_{k\geq 0}$ is a sequence of random vectors taking their values in E. We assume that there exists a *decreasing* function $t \mapsto g_k(t)$, $g_k(t) > 0$ on $\mathbb{R}^+$ such that the probability density function ϕ_k of $\boldsymbol{Z}_k$ satisfies

$$\phi_k(\boldsymbol{z}) \geq g_k(\|\boldsymbol{z}\|) > 0, \quad \forall \boldsymbol{z} \in E.$$

We write $\boldsymbol{x} < \boldsymbol{y}$ if there exist $1 \leq i \leq n$ such that $x_i < y_i$.

Let us denote by Φ_k the cumulative distribution function of $\boldsymbol{Z}_k$:

$$\Phi_k(\boldsymbol{y}) = P(\boldsymbol{Z}_k < \boldsymbol{y}),$$

and by $F_{k+1}(\boldsymbol{y} \mid \boldsymbol{X}_k = \boldsymbol{x})$ the conditional cumulative distribution function of $\boldsymbol{X}_{k+1}$:

$$F_{k+1}(\boldsymbol{y} \mid \boldsymbol{X}_k = \boldsymbol{x}) = P(\boldsymbol{X}_{k+1} < \boldsymbol{y} \mid \boldsymbol{X}_k = \boldsymbol{x}).$$

The conditional probability density function of $\boldsymbol{X}_{k+1}$ is denoted by ρ_{k+1}. By (9) and (24) we have that if $\boldsymbol{X}_k = \boldsymbol{x}$, then

$$\boldsymbol{X}_{k+1} = \boldsymbol{Q}(\boldsymbol{x}, \boldsymbol{d}, \omega) + \xi_k \boldsymbol{Z}_k.$$

We also have

$$F_{k+1}(\boldsymbol{y} \mid \boldsymbol{X}_k = \boldsymbol{x}) = P\left(\boldsymbol{Z}_k < \frac{\boldsymbol{y} - \boldsymbol{Q}(\boldsymbol{x}, \boldsymbol{d}, \omega)}{\xi_k}\right)$$

and

$$F_{k+1}(\boldsymbol{y} \mid \boldsymbol{X}_k = \boldsymbol{x}) = \Phi_k\left(\frac{\boldsymbol{y} - \boldsymbol{Q}(\boldsymbol{x}, \boldsymbol{d}, \omega)}{\xi_k}\right).$$

Thus,

$$\rho_{k+1}(\boldsymbol{y} \mid \boldsymbol{X}_k = \boldsymbol{x}) = \frac{1}{\xi_k^n}\phi_k\left(\frac{\boldsymbol{y} - \boldsymbol{Q}(\boldsymbol{x}, \boldsymbol{d}, \omega)}{\xi_k}\right) : \ \boldsymbol{y} \in E, \tag{25}$$

where $n = \dim(E)$, and (16) shows that

$$\boldsymbol{y} \in S_\alpha \Rightarrow \phi_k\left(\frac{\boldsymbol{y} - \boldsymbol{Q}(\boldsymbol{x}, \boldsymbol{d}, \omega)}{\xi_k}\right) \geq g_k\left(\frac{\beta(\alpha, \overline{\omega})}{\xi_k}\right). \tag{26}$$

4.2. Convergence to a Global Minimum. Let us introduce $U_k = f(\boldsymbol{X}_k)$. Since, at each iteration number $k \geq 0$, the step ω_k is determined in a way that reduces the value of the objective function (cf. Eqn. (7)), the sequence $\{U_k\}_{k\geq O}$ is decreasing by construction. Moreover, it has a lower bound given by l^* (cf. Eqn. (12)),

$$\forall k \geq 0 : l^* \leq U_{k+1} \leq U_k. \tag{27}$$

Thus, $\{U_k\}_{k\geq 0}$ is decreasing and bounded from below by l^*, i.e., there exists $U \geq l^*$ such that $U_k \to U$ for $k \to +\infty$. The aim is to establish that $U = l^*$.

We need the following auxiliary result:

Lemma 1. *Let $\boldsymbol{P}_k = \xi_k \boldsymbol{Z}_k$ and $\gamma = f(\boldsymbol{x}_0)$, $\hat{S}_\theta = \{\boldsymbol{x} \in E \mid f(\boldsymbol{x}) < \theta\}$. Then there exists $\nu > 0$ such that*

$$\forall k \geq 0 : P(U_{k+1} < \theta \mid U_k \geq \theta) \geq \frac{\mathrm{meas}(\hat{S}_\theta)}{\xi_k^n} g_k\left(\frac{\beta(\gamma, \overline{\omega})}{\xi_k}\right) > 0, \quad \forall \theta \in (l^*, l^* + \nu],$$

where $n = \dim(E)$.

Proof.

Step 1. Let $\nu > 0$ and $\theta \in (l^*, l^* + \nu]$. Since $S_\alpha \subset \hat{S}_\theta$, for $l^* < \alpha < \theta$, from (15) it follows that $\hat{S}_\theta$ is not empty and has a strictly positive measure.

Step 2. Assume that, for any $k > 0$, there exists $\theta_k \in (l^*, l^* + 1/k]$ such that $\mathrm{meas}\left(S - \hat{S}_{\theta_k}\right) = 0$. Then

$$\begin{aligned} l^* + \frac{1}{k} \geq \theta_k &\Longrightarrow \hat{S}_{\theta_k} \subset \hat{S}_{l^* + \frac{1}{k}} \\ &\Longrightarrow S - \hat{S}_{l^* + \frac{1}{k}} \subset S - \hat{S}_{\theta_k} \\ &\Longrightarrow \mathrm{meas}\left(S - \hat{S}_{l^* + \frac{1}{k}}\right) = 0. \end{aligned}$$

Let

$$A = \{\boldsymbol{x} \in S \mid f(\boldsymbol{x}) > l^*\}.$$

We have

$$A = \bigcup_{k=1}^{+\infty} A_k,$$

$$A_k = \left\{\boldsymbol{x} \in S \mid f(\boldsymbol{x}) \geq l^* + \frac{1}{k}\right\} = S - \hat{S}_{l^* + \frac{1}{k}}.$$

Since $\mathrm{meas}(A_k) = 0$, from the Borel-Cantelli lemma it follows that $\mathrm{meas}(A) = 0$. Thus, $f(\boldsymbol{x}) = l^*$ almost everywhere on S and f is constant on S, which contradicts (17).

Step 3. Consequently, there is $\nu = 1/k > 0$ such that $\mathrm{meas}(S - \hat{S}_\theta) > 0$ for $\theta \in (l^*, l^* + \nu]$. For any such θ we have

$$\begin{aligned} P\left(\boldsymbol{X}_k \notin \hat{S}_\theta\right) &= P\left(\boldsymbol{X}_k \in S - \hat{S}_\theta\right) \\ &= P\left(\boldsymbol{X}_k \in S - \hat{S}_\theta \mid \boldsymbol{X}_{k-1} \in S\right). \end{aligned}$$

Thus,

$$P\left(\boldsymbol{X}_k \notin \hat{S}_\theta\right) = \int_S P\left(\boldsymbol{X}_k \in \mathrm{d}\boldsymbol{x}\right) \int_{S-\hat{S}_\theta} \rho_k\left(\boldsymbol{y} \mid \boldsymbol{X}_{k-1} = \boldsymbol{x}\right) \mathrm{d}\boldsymbol{y}.$$

Since the sequence $\{U_i\}_{i\geq 0}$ is decreasing, we have

$$\{\boldsymbol{X}_i\}_{i\geq 0} \subset S_\gamma \subset S. \tag{28}$$

Thus, Eqns. (25) and (26) show that

$$\rho_k\left(\boldsymbol{y} \mid \boldsymbol{X}_k = \boldsymbol{x}\right) \geq \frac{1}{\xi_{k-1}^n} g_{k-1}\left(\frac{\beta(\gamma,\overline{\omega})}{\xi_{k-1}}\right) > 0, \forall k \geq 1.$$

Consequently,

$$P\left(\boldsymbol{X}_k \notin \hat{S}_\theta\right) \geq \int_S P\left(\boldsymbol{X}_k \in \mathrm{d}\boldsymbol{x}\right) \times \int_{S-\hat{S}_\theta} \frac{1}{\xi_{k-1}^n} g_{k-1}\left(\frac{\beta(\gamma,\overline{\omega})}{\xi_{k-1}}\right) \mathrm{d}\boldsymbol{y},$$

and we have

$$P\left(\boldsymbol{X}_k \notin \hat{S}_\theta\right) \geq \frac{\mathrm{meas}(S-\hat{S}_\theta)}{\xi_{k-1}^n} g_{k-1}\left(\frac{\beta(\gamma,\overline{\omega})}{\xi_{k-1}}\right) \times \int_S P\left(\boldsymbol{X}_k \in \mathrm{d}\boldsymbol{x}\right). \tag{29}$$

Equation (28) gives

$$\int_S P\left(\boldsymbol{X}_k \in \mathrm{d}\boldsymbol{x}\right) = P\left(\boldsymbol{X}_k \in S\right) = 1.$$

Thus, (29) yields

$$P\left(\boldsymbol{X}_k \notin \hat{S}_\theta\right) \geq \frac{\mathrm{meas}(S-\hat{S}_\theta)}{\xi_{k-1}^n} g_{k-1}\left(\frac{\beta(\gamma,\overline{\omega})}{\xi_{k-1}}\right) > 0 \quad \text{for any} \quad \theta \in (l^*, l^*+\nu].$$

Step 4. We have (see, e.g., Eqn. (27)):

$$P\left(U_{k+1} < \theta \mid U_k \geq \theta\right) = P\left(\boldsymbol{X}_{k+1} \in \hat{S}_\theta \mid \boldsymbol{X}_k \notin \hat{S}_\theta\right).$$

Since $P(X_k \notin \hat{S}_\theta) > 0$, we get

$$P\left(\boldsymbol{X}_{k+1} \in \hat{S}_\theta \mid \boldsymbol{X}_k \notin \hat{S}_\theta\right) = \frac{P\left(\boldsymbol{X}_{k+1} \in \hat{S}_\theta,\ \boldsymbol{X}_k \notin \hat{S}_\theta\right)}{P\left(\boldsymbol{X}_k \notin \hat{S}_\theta\right)}.$$

Accordingly,

$$P\left(\boldsymbol{X}_{k+1} \in \hat{S}_\theta, \boldsymbol{X}_k \notin \hat{S}_\theta\right) = \int_{S-\hat{S}_\theta} P\left(\boldsymbol{X}_k \in \mathrm{d}\boldsymbol{x}\right) \int_{\hat{S}_\theta} \rho_{k+1}\left(\boldsymbol{y} \boldsymbol{X}_k = \boldsymbol{x}\right) \mathrm{d}\boldsymbol{y}.$$

From Eqns. (28) and (26) it follows that

$$P\left(\boldsymbol{X}_{k+1} \in \hat{S}_\theta, \boldsymbol{X}_k \notin \hat{S}_\theta\right) \geq \int_{S_\gamma - \hat{S}_\theta} P\left(\boldsymbol{X}_k \in \mathrm{d}\boldsymbol{x}\right) \int_{\hat{S}_\theta} \frac{1}{\xi_k^n} g_k\left(\frac{\beta(\gamma,\overline{\omega})}{\xi_k}\right) \mathrm{d}\boldsymbol{y},$$

and

$$P\left(\boldsymbol{X}_{k+1} \in \hat{S}_\theta, \boldsymbol{X}_k \notin \hat{S}_\theta\right) \geq \frac{\mathrm{meas}(\hat{S}_\theta)}{\xi_k^n} g_k\left(\frac{\beta(\gamma,\overline{\omega})}{\xi_k}\right) \int_{S_\gamma - \hat{S}_\theta} P\left(\boldsymbol{X}_k \in \mathrm{d}\boldsymbol{x}\right).$$

Consequently,

$$P\left(\boldsymbol{X}_{k+1} \in \hat{S}_\theta, \boldsymbol{X}_k \notin \hat{S}_\theta\right) \geq \frac{\mathrm{meas}(\hat{S}_\theta)}{\xi_k^n} g_k\left(\frac{\beta(\gamma,\overline{\omega})}{\xi_k}\right) P\left(\boldsymbol{X}_k \notin \hat{S}_\theta\right),$$

and

$$P\left(\boldsymbol{X}_{k+1} \in \hat{S}_\theta \mid \boldsymbol{X}_k \notin \hat{S}_\theta\right) \geq \frac{\mathrm{meas}(\hat{S}_\theta)}{\xi_k^n} g_k\left(\frac{\beta(\gamma,\overline{\omega})}{\xi_k}\right).$$

■

Global convergence follows from the Borel-Cantelli lemma:

Lemma 2. *Let $\{U_k\}_{k\geq 0}$ be a decreasing sequence, bounded from below by l^*. Then there exists U such that $U_k \to U$ for $k \to +\infty$. Assume that there exists $\nu > 0$ such that, for any $\theta \in (l^*, l^*+\nu]$, there is a sequence of strictly positive real numbers $\{c_k(\theta)\}_{k\geq 0}$ which satisfies*

$$\forall k \geq 0 : P\left(U_{k+1} < \theta \mid U_k \geq \theta\right) \geq c_k(\theta) > 0, \quad \sum_{k=0}^{+\infty} c_k(\theta) = +\infty. \tag{30}$$

Then $U = l^*$ almost surely.

Proof. See, e.g., (Pierre and Renée, 1996; Pogu and Souza, 1994). ■

Proposition 1. *Let $\gamma = f(x_0)$. Assume that the sequence $\{\xi_k\}_{k\geq 0}$ is nonincreasing and*

$$\sum_{k=0}^{+\infty} g_k\left(\frac{\beta(\gamma,\overline{\omega})}{\xi_k}\right) = +\infty. \tag{31}$$

Then $U = l^*$ almost surely.

Proof. Let

$$c_k(\theta) = \frac{\text{meas}(\hat{S}_\theta)}{\xi_k^n} g_k\left(\frac{\beta(\gamma,\overline{\omega})}{\xi_k}\right) > 0. \qquad (32)$$

Since the sequence $\{\xi_k\}_{k\geq 0}$ is nonincreasing, we have

$$c_k(\theta) \geq \frac{\text{meas}(\hat{S}_\theta)}{\xi_0^n} g_k\left(\frac{\beta(\gamma,\overline{\omega})}{\xi_k}\right) > 0.$$

From Eqn. (31) it follows that

$$\sum_{k=0}^{+\infty} c_k(\theta) \geq \frac{\text{meas}(\hat{S}_\theta)}{\xi_0^n} \sum_{k=0}^{+\infty} g_k\left(\frac{\beta(\gamma,\overline{\omega})}{\xi_k}\right) = +\infty.$$

Using Lemmas 1 and 2, we have $U = l^*$ almost surely. ■

Theorem 1. *Let $\boldsymbol{Z}_k = \boldsymbol{Z}$, where $\boldsymbol{Z}$ is a random variable following $N(\boldsymbol{0}, \sigma\boldsymbol{Id})$, $\sigma > 0$, and let*

$$\xi_k = \sqrt{\frac{a}{\ln(k+d)}}, \qquad (33)$$

where $a > 0$, $d > 0$ and k is the iteration number. Then, for a sufficiently large, $U = l^*$ almost surely.

Proof. We have

$$\phi_k(\boldsymbol{z}) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left(-\frac{1}{2}\left\|\frac{\boldsymbol{z}}{\sigma}\right\|^2\right) = g_k(\|\boldsymbol{z}\|) > 0.$$

Consequently,

$$g_k\left(\frac{\beta(\gamma,\overline{\omega})}{\xi_k}\right) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n (k+d)^{\beta(\gamma,\overline{\omega})^2/(2\sigma^2 a)}}.$$

For a such that

$$0 < \frac{\beta(\gamma,\overline{\omega})^2}{2\sigma a} < 1,$$

we have

$$\sum_{k=0}^{+\infty} g_k\left(\frac{\beta(\gamma,\overline{\omega})}{\xi_k}\right) = +\infty,$$

and, from the preceding proposition, we have $U = l^*$ almost surely. ■

4.3. Practical Implementation. The above results suggest the following numerical algorithm:

(1) An initial guess $\boldsymbol{x}_0 \in S$ is given.

(2) At the iteration number $k \geq 0$, $\boldsymbol{x}_k$ is known and $\boldsymbol{x}_{k+1}$ is determined by performing the following three substeps:

(2.1) *Unperturbed descent:* We determine the descent direction $\boldsymbol{d}_k$ and the step ω_k using the standard descent method (3)–(8). This generates the first trial point:

$$\boldsymbol{T}_{k+1}^0 = \boldsymbol{Q}(\boldsymbol{x}_k, \boldsymbol{d}_k, \omega_k).$$

(2.2) *Perturbation:* We determine a sample $(\boldsymbol{P}_k^1, \ldots, \boldsymbol{P}_k^{n_{sto}})$ of $n_{sto}(k)$ variates from $\boldsymbol{P}_k$. This generates $n_{sto}(k)$ new trial points:

$$\boldsymbol{T}_{k+1}^i = \boldsymbol{T}_{k+1}^0 + \boldsymbol{P}_k^i, \quad i = 1, \ldots, n_{sto}(k).$$

(2.3) *Dynamics:* We determine $\boldsymbol{x}_{k+1}$ by selecting it from the set of available points:

$$\mathcal{A}_k = \left\{\boldsymbol{x}_k, \boldsymbol{T}_{k+1}^0, \ldots, \boldsymbol{T}_{k+1}^{n_{sto}}\right\}.$$

In order to satisfy the assumptions leading to mathematical results, Substep (2.1) may use a descent method satisfying Eqns. (3)–(8). As was shown in Theorem 1, Substep (2.2) may use $\boldsymbol{P}_k^i = \xi_k \boldsymbol{Z}^i$, where $\mathcal{Z} = \left(\boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^{n_{sto}}\right)$ is a sample of $N(\boldsymbol{0}, \sigma\boldsymbol{Id})$ and ξ_k is given by Eqn. (33). The choice in Substep (2.3) must fulfil Lemma 2. For instance, we can consider elitistic dynamics:

$$\boldsymbol{x}_{k+1} = \arg\min\left\{f(\boldsymbol{x}) : \ \boldsymbol{x} \in \mathcal{A}_k\right\}. \qquad (34)$$

In practice, a maximum iteration number $k_{\max}$ has to be considered: the iterations are stopped when $k = k_{\max}$.

Remark 2. The dynamics of Metropolis *do not* satisfy Lemma 2. An alternative formulation, compatible with this dynamics, can be found, e.g., in (Souza de Cursi, 1992b).

5. Numerical Experiments

In this section, we present the results of some numerical experiments performed using the algorithm described in Section 4.3.

In Substep (2.1), we use the variable metric descent method given by Eqns. (3)–(8), (20) and (21). The descent directions are generated using generalized gradients (see Section 3): in the neighbourhood of a point where the objective function is nondifferentiable, a random convex combination of two elements of the generalized gradient provides a descent direction. The main parameters of the method are $\overline{\omega}$ and b_{step}. Their influence has been studied.

In Substep (2.2), we use a fixed size for all the samples, $n_{sto}(k) = n_{sto}$, independent of the iteration number k. $\mathcal{Z}$ is a sample of $N(\boldsymbol{0}, \boldsymbol{Id})$ (thus, $\sigma = 1$) and $\xi_k = \sqrt{a/\ln(k+2)}$. The main parameters are n_{sto} and $a > 0$. Their influence has been studied.

In Substep (2.3), we use the elitistic dynamics described by Eqn. (34).

The experiments were performed on a workstation IBM Intel(R) Pentium(R) 4 CPU 3.00 GHz, 504 MB RAM. The serial CPU yields the mean CPU time counted in seconds for one run.

5.1. Examples. Owing to volume limitations, we do not present here a complete set of numerical experiments performed and we focus our attention on some classical test functions.

Example 1. (Crescent, see (Mākelā and Neittaanmāki, 1992))

$$f(\boldsymbol{x}) = \max\left\{x_1^2 + (x_2-1)^2 + x_2 - 1, -x_1^2 - (x_2-1)^2 + x_2 + 1\right\}.$$

The starting point is

$$x_1 = -1.5, \quad x_2 = 2.0.$$

Example 2. (Mifflin 2, see (Mākelā and Neittaanmāki, 1992))

$$f(\boldsymbol{x}) = -x_1 + 2(x_1^2 + x_2^2 - 1) + 1.75|x_1^2 + x_2^2 - 1|.$$

The starting point is

$$x_1 = -1, \quad x_2 = -1.$$

Example 3. (Wolfe, see (Zowe, 1985))

$$\begin{cases} f(\boldsymbol{x}) = f_1(\boldsymbol{x}), & x_1 > |x_2|, \\ f(\boldsymbol{x}) = f_2(\boldsymbol{x}), & 0 < x_1 \le |x_2|, \\ f(\boldsymbol{x}) = f_3(\boldsymbol{x}), & x_1 \le 0, \\ f_1(\boldsymbol{x}) = 5\sqrt{9x_1^2 + 16x_2^2}, \\ f_2(\boldsymbol{x}) = 9x_1 + 16|x_2|, \\ f_3(\boldsymbol{x}) = 9x_1 + 16|x_2| - x_1^9. \end{cases}$$

The starting point is

$$x_1 = 3, \quad x_2 = 2.$$

Example 4. (Colville 1, see (Bihain , 1984))

$$\begin{cases} \min f(\boldsymbol{x}) = \displaystyle\sum_{i=1}^{5} d_i x_5^3 + \sum_{i=1}^{5}\sum_{j=1}^{5} c_{ij} x_j x_i \\ \qquad\qquad + \displaystyle\sum_{j=1}^{5} e_j x_j \\ \text{subject to: } \displaystyle\sum_{j=1}^{5} a_{ij} x_j \ge b_i, \quad i = 1, \ldots, 10, \\ \qquad\qquad x_j \ge 0, \quad j = 1, \ldots, 5. \end{cases}$$

$$A = \begin{pmatrix} -16 & 2 & 0 & 1 & 0 \\ 0 & -2 & 0 & 4 & 2 \\ -3.5 & 0 & 2 & 0 & 0 \\ 0 & -2 & 0 & -4 & -1 \\ 0 & -9 & -2 & 1 & -2.8 \\ 2 & 0 & -4 & 0 & 0 \\ -1 & -1 & -1 & -1 & -1 \\ -1 & -2 & -3 & -2 & -1 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix},$$

$$C = \begin{pmatrix} 30 & -20 & -10 & 32 & -10 \\ -20 & 39 & -6 & -31 & 32 \\ -10 & -6 & 10 & -6 & -10 \\ 32 & -31 & -6 & 39 & -20 \\ -10 & 32 & -10 & -20 & 30 \end{pmatrix},$$

$$b = \begin{pmatrix} -40 \\ -2 \\ -0.25 \\ -4 \\ -4 \\ -1 \\ -40 \\ -60 \\ 5 \\ 1 \end{pmatrix}, \quad d = \begin{pmatrix} 4 \\ 8 \\ 10 \\ 6 \\ 2 \end{pmatrix},$$

$$e^t = \begin{pmatrix} -15 & -27 & -36 & -18 & -12 \end{pmatrix}.$$

The starting point is

$$x_1 = 0, \quad x_2 = 0, \quad x_3 = 0, \quad x_4 = 0 \text{ and } x_5 = 1.$$

The restrictions are penalized using the term

$$p(\boldsymbol{x}) = \max\left[0, \max_{1\le i\le 10}\left(b_i - \sum_{j=1}^{5} a_{ij} x_j\right)\right] + \sum_{i=1}^{5} \max(0, -x_i).$$

Thus, we minimize $f_\lambda = f + \lambda p$. The calculations use $\lambda = 100$.

Example 5. (Gill, see (Kiwiel, 1989))

$$\begin{cases} f(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), f_3(\boldsymbol{x})\}, \\ f_1(\boldsymbol{x}) = \sum_{i=1}^{10} (x_i - 1)^2 + 10^3 \sum_{i=1}^{10} \left(x_i^2 - \frac{1}{4}\right)^2, \\ f_2(\boldsymbol{x}) = \sum_{i=2}^{30} \left[\sum_{j=2}^{10} x_j (j-1) \left(\frac{i-1}{29}\right)^{j-2} \right. \\ \qquad \left. - \left(\sum_{j=1}^{10} x_j \left(\frac{i-1}{29}\right)^{j-1}\right)^2 - 1\right]^2 \\ \qquad + x_1^2 + \left(x_2 - x_1^2 - 1\right)^2, \\ f_3(\boldsymbol{x}) = \sum_{i=2}^{10} \left[100(x_i - x_{i-1}^2)^2 + (1 - x_i)^2\right]. \end{cases}$$

The starting point is

$$x_i = -0.1, \quad 1 \le i \le 10.$$

5.2. Results. In order to analyze the influence of various parameters, we consider two basic sets of values:

Case 1: $\bar{\omega} = 100$, $k_{\max} = 100$, $n_{sto} = 500$, $b_{step} = 0.1$ and $a = 1$.

Case 2: $\bar{\omega} = 100$, $k_{\max} = 500$, $n_{sto} = 500$, $b_{step} = 0.1$ and $a = 1$.

In Tables 1 and 2, we show the observed influence of the variation of a single parameter while the others remain at their original values for Cases 1 and 2, respectively. The minimal values of the objective functions $f(\boldsymbol{x}^*) = 0,\ -1,\ -8$ and 8.24 in Examples Crescent, Mifflin 2, Wolfe and Gill were respectively found for Cases 1 and 2, but the minimal value of the objective function $f(\boldsymbol{x}^*) = -32.34$ was produced in Example Colville 1 for Case 2. The other local minimum values are shown in Tables 1 and 2. We give the iteration number $k_{\max}$ in Table 1 since it is the same in Cases 1 and 2.

Example 4 may be also solved by considering a box bounded situation where $\boldsymbol{\ell} = 0$ and $\boldsymbol{u} = +\infty$. In this case, the penalty term becomes

$$p(\boldsymbol{x}) = \max\left[0, \max_{1 \le i \le 10}\left(b_i - \sum_{j=1}^{5} a_{ij} x_j\right)\right].$$

This approach led to analogous results.

In Table 3 we give results of comparison with:

- the New Variable Metric Algorithm (NVMA), see, e.g., (Uryasev, 1991), and

Table 1. Optimal values obtained in Case 1. The most influential parameter is the number of generated points.

Example		Crescent	Mifflin 2	Wolfe	Colville 1	Gill
$\bar{\omega}$	1	0.00000	−1.00000	−8.0000	−32.2795	8.2452
	10	0.00001	−0.99998	−7.9994	−32.2714	8.2976
	100	0.00015	−0.99952	−7.9855	−32.1958	10.7980
	500	0.00379	−0.99647	−7.9521	−32.2835	189.022
	1000	0.00585	−0.98763	−7.7064	−32.2835	189.022
$k_{\max}$	10	0.00384	−0.99706	−7.9855	−30.7826	66.4160
	50	0.00015	−0.99827	−7.9855	−32.1562	15.3921
	100	0.00015	−0.99952	−7.9855	−32.1958	10.7980
	500	0.00003	−1.00000	−7.9912	−32.3344	10.7980
	1000	0.00003	−1.00000	−7.9978	−32.3365	10.0830
n_{sto}	0	4.25000	4.75000	60.208	20.0000	189.022
	10	0.01022	−0.99127	−7.7394	−32.0894	22.2780
	50	0.00699	−0.99879	−7.9888	−32.0173	31.4441
	100	0.00101	−0.99983	−7.9815	−32.1694	17.4487
	500	0.00015	−0.99952	−7.9855	−32.1958	10.7980
	1000	0.00011	−0.99997	−7.9889	−32.3253	13.8606
b_{step}	0.001	0.00015	−0.99952	−7.9855	−32.2500	14.1922
	0.01	0.00015	−0.99952	−7.9855	−32.2500	15.2089
	0.1	0.00015	−0.99952	−7.9855	−32.1958	10.7980
a	0.01	0.00000	−1.00000	−8.0000	−32.2795	8.2460
	0.1	0.00002	−0.99999	−8.0000	−32.2714	8.2940
	1	0.00015	−0.99952	−7.9855	−32.1958	10.7980
CPU	time	1.39	0.95	1.31	17.60	91.13

Table 2. Optimal values obtained in Case 2. The most influential parameter is the number of generated points.

Example		Crescent	Mifflin 2	Wolfe	Colville 1	Gill
$\bar{\omega}$	1	0.00000	−1.00000	−8.0000	−32.2860	8.2444
	10	0.00000	−1.00000	−7.9999	−32.3385	8.2876
	100	0.00002	−1.00000	−7.9912	−32.3344	10.0830
	500	0.00101	−0.99898	−7.9968	−32.3226	189.022
	1000	0.00172	−0.99939	−7.9343	−32.3226	189.022
n_{sto}	0	4.25000	4.75000	60.2079	20.0000	189.022
	10	0.00166	−0.99942	−7.8961	−32.1431	19.9605
	50	0.00058	−0.99943	−7.9933	−32.1960	14.2083
	100	0.00078	−0.99982	−7.9937	−32.2367	13.2332
	500	0.00002	−1.00000	−7.9912	−32.3344	10.0830
	1000	0.00003	−0.99997	−7.9979	−32.3362	11.7762
b_{step}	0.001	0.00002	−1.00000	−7.9912	−32.3344	10.9185
	0.01	0.00002	−1.00000	−7.9912	−32.3344	10.7423
	0.1	0.00002	−1.00000	−7.9912	−32.3344	10.0830
a	0.01	0.00000	−1.00000	−8.0000	−32.2860	8.2444
	0.1	0.00000	−1.00000	−8.0000	−32.3464	8.2904
	1	0.00002	−1.00000	−7.9912	−32.3344	10.0830
CPU	time	6.56	4.87	6.29	87.29	463.01

Table 3. Results of the estimation of the minimal value of the objective obtained by using three methods.

Example	Crescent	Mifflin 2	Wolfe	Colville 1	Gill
NVMA	0.00000	−1.00000	−8.0000	−32.3306	9.7873
SIGMA	0.00000	−1.00000	−8.0000	−32.3386	8.2722
RPVM	0.00000	−1.00000	−8.0000	−32.3464	8.2444

- the Stochastic Integration Global Minimization Algorithm (SIGMA), see, e.g., (Aluffi-Pentini *et al.*, 1988).

The RPVM method has high reliability, as shown by the presented results: convergence to global optima was obtained independently of the starting point. The method is also less sensitive to the choice of the parameters than pure stochastic methods. The presented combination of deterministic and stochastic descents increases robustness.

In order to get some information about the robustness, we studied the behavior of the method when using various samples of random vectors. We denote by $\widehat{l^*}$ and $\widehat{\boldsymbol{x}^*}$ the estimates of the optimal values of l^* and $\boldsymbol{x}^*$ provided by the method, respectively. Run i yields the variates $\widehat{l_i^*}$ and $\widehat{\boldsymbol{x}_i^*}$: a set of nr runs yields nr-samples of $\widehat{l^*}$ and $\widehat{\boldsymbol{x}^*}$, which allows us to get the estimates of their mean values and variances. In Table 4 we show the results obtained from $nr = 100$ independent runs. We denote by $E(\widehat{l^*})$ and $V(\widehat{l^*})$ the *estimates* of the mean and the variance of $\widehat{l^*}$ provided by the produced results. Analogous notation is used for $\widehat{\boldsymbol{x}^*}$. We observe that the results are stable, with small variance.

6. Concluding Remarks and Future Work

We have considered a general problem of unconstrained continuous optimization where the objective function may be nonsmooth. Standard methods for smooth functions usually generate a descent direction by using the gradient and may be extended to nonsmooth situations by using a generalized gradient instead of the standard one whenever necessary. For instance, Clarke's generalized gradients or more general affine local underevaluations (or overevaluations) can be used at the points where the objective function is not differentiable. In accordance with this observation, we considered a general variable metric descent method and introduced suitable affine local approximations to be used.

However, in the lack of convexity assumptions, convergence to a global minimum cannot be ensured. We also introduced a stochastic modification of the descent method involving the incorporation of a random perturbation $\boldsymbol{P}_k$, which may be interpreted as a perturbation of the descent direction. This approach leads to a stochastic descent method where the deterministic sequence generated by the variable metric descent method is replaced by a sequence of random variables. A mathematical result concerning convergence to a global minimum was established for a convenient class of random perturbations. We established that perturbations such that $\boldsymbol{P}_k = \xi_k \boldsymbol{Z}$ belong to this class if $\boldsymbol{Z}$ is a Gaussian random vector ($N(\boldsymbol{0}, \boldsymbol{Id})$ variate), and $\{\xi_k\}_{k\geq 0}$ is a nonincreasing sequence of strictly positive real numbers converging to zero and such that $\xi_0 \leq 1$. This provides a simple method for generation of convenient perturbations.

We proposed an algorithm for the implementation of the method and presented the results of some numerical experiments. The results show that, on the one hand, the method is effective and, on the other, the introduction of stochastic perturbations significantly improves performance. Robustness was analyzed using different samples generated at independent runs: the results are stable with small variance, as shown in Table 4.

The main difficulty in the practical use of stochastic methods is connected with the tuning of parameters. We analyzed the practical effect of variations in the main parameters of the proposed algorithm: $a, \overline{\omega}, k_{\max}, b_{step}$ and n_{sto}. As is shown in Tables 1 and 2, the influence of b_{step} is small. The parameters a and $\overline{\omega}$ have slight influence: a large set of values led to good results. A more influential

Table 4. Analysis of robustness on 100 tests.

Example	Crescent	Mifflin 2	Wolfe	Colville 1	Gill
$E\left(\widehat{l^*}\right)$	0.000	−1.000	−8.000	−32.3464	8.245
$V(\widehat{l^*})$	5E−28	2E−15	1E−13	1.8E−12	3E−07
$\widehat{\boldsymbol{x}^*}$	$\widehat{x_1^*}$ = 0.001 $\widehat{x_2^*}$ = 0.000	$\widehat{x_1^*}$ = 1.000 $\widehat{x_2^*}$ = 0.000	$\widehat{x_1^*}$ = −1.000 $\widehat{x_2^*}$ = 0.000	$\widehat{x_1^*}$ = 0.2999 $\widehat{x_2^*}$ = 0.3329 $\widehat{x_3^*}$ = 0.3999 $\widehat{x_4^*}$ = 0.4284 $\widehat{x_5^*}$ = 0.2255	$\widehat{x_1^*}$ = 0.456, $\widehat{x_6^*}$ = 0.441 $\widehat{x_2^*}$ = 0.362, $\widehat{x_7^*}$ = 0.256 $\widehat{x_3^*}$ = 0.164, $\widehat{x_8^*}$ = 0.110 $\widehat{x_4^*}$ = 0.087, $\widehat{x_9^*}$ = 0.077 $\widehat{x_5^*}$ = 0.071, $\widehat{x_{10}^*}$ = 0.060

parameter is the number of trial points generated, $k_{\max} \times n_{sto}$: for a small number of iterations $k_{\max}$, a large number of perturbations n_{sto} is needed, while small n_{sto} requires large $k_{\max}$.

Variable metric descent methods can be extended to constrained optimization. Future work will consider generalization to affine and nonlinear constraints.

References

Aluffi-Pentini F., Parisi V. and Zirilli F. (1988): *ALGORITHM 667: SIGMA – A Stochastic-Integration Global Minimization Algorithm*. — ACM Trans. Math. Softw., Vol. 14, No. 4, pp. 366–380.

Bagirov A.M., Rubinov A.M., Soukhoroukova N.V. and Yearwood J. (2003): *Unsupervised and supervised data classification via nonsmooth and global optimization*. — Trab. Invest. Oper., Vol. 11, No. 1, pp. 1–93.

Batukhtin V.D., Kirillova F.M. and Ukhobotov V.I. (1998): *Nonsmooth and discontinuous problems of control and optimization*. — Proc. IFAC Workshop NDPCO'98, Chelyabinsk, Russia, pp. 25–34.

Bihain A. (1984): *Optimization of upper semidifferentiable functions*. — J. Optim. Theory Applic., Vol. 4, No. 4, pp. 545–568.

Bouleau N. (1986): *Variables Aléatoires et Simulation*. — Paris: Editions Hermann.

Boyd S., Xiao L. and Mutapcic A. (2003): *Subgradient Methods*. — Notes for EE392o, Stanford University.

Carson Y. and Maria A. (1997): *Simulation optimization: Methods and applications*. — Proc. *Winter Simulation Conf.*, Atlanta, GA, USA, pp. 118–126.

Clarke F. (1983): *Optimization and Nonsmooth Analysis*. — New York: Wiley.

Clarke F. (1975): *Generalized gradient and applications*. — Trans. Amer. Math. Soc., Vol. 205, pp. 247–262.

Davidon W. (1991): *Variable metric method for minimization*. — SIAM J. Optim., Vol. 1, No. 1, pp. 1–17.

Dorea C.C.Y. (1990): *Stopping rules for a random optimization method*. — SIAM J. Contr. Optim., Vol. 28, No. 4, pp. 841–850.

Ellaia R. (1992): *Contributions à l'optimisation globale et à l'analyse nondifférentiable*. — Thèse d'état, Université Mohammed V, Faculté des Sciences, Rabat, Morocco.

Ellaia R. and Elmouatasim A., (2004): *Random perturbation of reduced gradient method for global optimization*. — Proc. Conf. *Modelling, Computation and Optimization, MCO'04*, Metz, France, (to appear.)

Fletcher R. (1980): *Practical Methods of Optimization, Vol.2*. — New York: Wiley.

Hiriart-Urruty J. and Lemaréchal C. (1993): *Convex Analysis and Minimization Algorithms II*. — Berlin: Springer.

Kiwiel K.C. (1985): *Method of Descent for Nondifferentiable Optimization*. — Berlin: Springer.

Kiwiel K.C. (1989): *An ellipsoid trust region bundle method for nonsmooth convex minimization*. — SIAM J. Contr. Optim., Vol. 27, No. 4, pp. 737–757.

Kiwiel K.C. (1994): *Free-steering relaxation methods for problems with strictly convex costs and linear constraints*. — Tech. Rep. IIASA, No. A-2361, Laxenburg, Austria.

Larsson T., Patrksson M. and Stromberg A.B. (1996): *Conditional subgradient optimization—Theory and applications*. — Europ. J. Oper. Res., Vol. 88, No. 2, pp. 382–403.

Lemaréchal C. (1982): *Numerical experiments in nonsmooth optimization*. — Proc. IIASA Workshop *Progress in Nondifferentiable Optimization*, Laxemburg, Austria, pp. 61–84.

Lemaréchal C., Strodiot J.J. and Bihain A. (1981): *On a bundle algorithm for nonsmooth optimization*, In: Nonlinear Programming 4 (O. Mangasarian, R. Meyer and S. Robinson, Eds.). — New York: Academic Press, pp. 245–282.

Mākelā M.M. and Neittaanmākі P. (1992): *Nonsmooth Optimization: Analysis and Algorithms with Applications to Optimal Control*. — London: World Scientific.

Nguyen V.H. and Strodiot J.J. (1992): *Computing a Global Optimal Solution to a Design Centering Problem*. — Tech. Rep., No. 88/12, Facultés Universitaires Notre-Dame de la Paix, Belgium.

Outrata J. (1987): *On numerical solution of optimal control problems with nonsmooth objectives: Application to economic models*. — Kybernetika, Vol. 23, pp. 54–66.

Outrata J., Kocvara M. and Zowe J. (1998): *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints: Theory, Applications & Numerical Results*. — Dordrecht: Kluwer.

Pierre L. and Renée T. (1996): *On the Deng-Lin random number generators and related methods*, In: Global Optimization in Action (Pinter J., Ed.). — Les cahiers du GERAD, Dordrecht: Kluwer.

Pogu M. and Souza J.E. (1994): *Global optimization by random perturbation of the gradient method with a fixed parameter*. — J. Glob. Optim., Vol. 5, No. 2, pp. 159–180.

Schramm H. and Zowe J. (1992): *A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results*. — SIAM J. Optim., Vol. 2, No. 1, pp. 121–152.

Souza de Cursi J.E. (1992a): *Minimisation stochastique de fonctionnelles non convexes en dimension finie*. — Tech. Rep. ECN, Available at `http://meca.insa-rouen.fr/~souza`

Souza de Cursi J.E. (1992b): *Recuit simulé et application à l'approximation par des sommes d'exponentielles*. — Tech. Rep. ECN, Available at `http://meca.insa-rouen.fr/~souza`

Souza de Cursi J.E., Ellaia R. and Bouhadi M. (2003): *Global optimization under nonlinear restrictions by using stochastic perturbations of the projected gradient*, In: Frontiers In Global Optimization (C.A. Floudas and Panos Pardalos, Eds.). — Boston, MA: Kluwer Academic Publishers, Nonconvex Optim. Appl., Vol. 74, pp. 541–561.

Souza de Cursi J.E., Ellaia R. and El Mouatasim A. (2005): *Stochastic perturbation of active set method for nonconvex nonsmooth optimization problem with linear constraints*. — RAIRO-Recherche Operational, (submitted).

Uryasev S.P. (1991): *New variable metric algorithms for nondifferentiable optimization problems*. — J. Optim. Theory Applic., Vol. 71, No. 2, pp. 359–388.

Wang I.J. and Spall J.C. (1999): *A constrained simultaneous perturbation stochastic approximation algorithm based on penalty functions*. — Proc. Amer. Contr. Conf. San Diego, CA, Vol. 1, pp. 393–399.

Zowe J. (1985): *Nondifferentiable optimization*, In: Computational Mathematical Programming (K. Schittkowski, Ed.). — Berlin: Springer Verlag, pp. 323–356.

Received: 19 April 2006
Revised: 28 September 2006
Re-revised: 6 November 2006