Filomat 32:19 (2018), 6799–6807
https://doi.org/10.2298/FIL1819799K

Published by Faculty of Sciences and Mathematics,
University of Niš, Serbia
Available at: http://www.pmf.ni.ac.rs/filomat

A Nonmonotone Line Search Method for Stochastic Optimization Problems

Nataša Krejić[a], Sanja Lončar[b]

[a] *Department of Mathematics and Informatics, University of Novi Sad, Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia*
[b] *Novi Sad School of Business,Vladimira Perića-Valtera 4, 21000 Novi Sad, Serbia*

Abstract. A nonmonotone line search method for solving unconstrained optimization problems with the objective function in the form of mathematical expectation is proposed and analyzed. The method works with approximate values of the objective function obtained with increasing sample sizes and improves accuracy gradually. Nonmonotone rule significantly enlarges the set of admissible search directions and prevents unnecessarily small steps at the beginning of the iterative procedure. The convergence is shown for any search direction that approaches the negative gradient in the limit. The convergence results are obtained in the sense of zero upper density. Initial numerical results confirm theoretical results and show efficiency of the proposed approach.

1. Introduction

The problem that we consider is an unconstrained problem of the form

$$\min_{x \in R^p} f(x), \tag{1}$$

where the objective function f is given as

$$f(x) = E(g(x, \omega)). \tag{2}$$

The mathematical expectation E is defined with respect to ω in the probability space $(\Omega, \mathcal{F}, P)$. It is assumed that the function $g : R^p \times \Omega \to R$ is known analytically or provided by a black box oracle with desired accuracy. But the analytical form of the function f is seldom available and needs to be approximated in some way. The most common approximation is the Sample Average Approximation defined as

$$G(x, w) = \frac{1}{n} \sum_{j=1}^{n} g(x, \omega_j), \tag{3}$$

where $\omega = \{w_1, \ldots, w_n\}$ is random sample of size n. The sample size n represents a tradeoff between precision and cost, as large sample size provides better approximation but causes higher computation costs and vice

2010 *Mathematics Subject Classification*. Primary 90C15; Secondary 90C30
Keywords. zero upper density convergence, unconstrained stochastic problem, sample average approximation
Received: 23 February 2017; Revised: 15 April 2017; Accepted: 15 April 2017
Communicated by Dragan S. Djordjević
Nataša Krejić is supported by Ministry of Education and Science, Republic of Serbia, Grant no. 174030
Email addresses: natasak@uns.ac.rs (Nataša Krejić), sanja.lonchar@gmail.com (Sanja Lončar)

versa. Problems of this type appear in many applications, for example in mathematical models obtained by simulations or whenever the set of model parameters is not known or is subject to noise. Thus, there is a great need to solve them efficiently. In general, a large sample size is needed to obtain approximations of reasonable accuracy. This fact causes large computational effort in solving (1) as the computation of the objective function, as well as its derivatives, becomes very costly. The general approach is to consider a sequence of approximations (2) with an increasing sample size, i.e., with a different sample size in each iteration and lower the cost of the overall optimization procedure. The problem (1) is closely related to the problem arising in machine learning where one has to minimize a finite, but a very large sum of functions, see [3, 4].

There are many different approaches for a choice of the sequence $\{n(i)\}$ of sample sizes at each iteration. The dominant way of sample size scheduling is an increasing sample size sequence that results in smaller computational costs than working with a large sample from the beginning. One can distinguish between two main approaches in the sample size scheduling - a predetermined sample size schedule, for example [12] or an adaptive sample size schedule, [6, 11, 13]. An overview of different sample size scheduling is presented in [7].

The classical approach in deterministic optimization for unconstrained optimization is to apply a line search method, either monotone and based on Armijo type decrease condition, or one of the well known nonmonotone line search methods. The monotone line search method for (1)-(2) with a predetermined sample size sequence is defined and considered for problems of type (1) in [12]. The method is based on a decrease determined by the Armijo rule in each iteration, for the approximate objective function defined with the current sample in the iteration. The search direction is an approximate negative gradient. It is shown that the method converges with zero upper density. However, the decrease obtained with the Armijo rule at each iteration is, in fact, a decrease of the approximate objective function at that iteration and does not necessarily imply a decrease of the true objective function of (1). On the other hand, the strict decrease condition might cause a rather small step size and thus trap the algorithm in a narrow valley of the objective function. This is specially the case when the derivatives are not available. Hence, a nonmonotone line search, which does not require a strict decrease in each iteration and allows for large step sizes, might be a better option for the overall optimization procedure, in particular for the stochastic problems. An additional property of a nonmonotone line search procedure is the step sizes are in general larger and there is more freedom with the search direction. In this paper we consider the nonmonotone line search rule due to Li, Fukushima [8] that is successfully applied in many papers, for deterministic and stochastic problems, for example see [1, 6].

The main contribution of this paper is a generalization of the results presented in [12] in the following sense. First, we define a nonmonotone line search strategy that allows us to take an arbitrary search direction, not necessarily strictly decreasing for the current approximate objective function. The search directions need to approach the negative gradient only in the limit. Furthermore, the step size rule allows us more freedom and hence generates a sequence that might approach the solution faster. We prove the convergence of the proposed algorithm in the sense of zero upper density, as in [12]. Finally, we present a set of initial testing results that confirm the theoretical results and provide empirical evidence for the proposed algorithm.

2. Preliminaries

In this section we briefly repeat the results of Wardy [12] that will allow us to propose a nonmonotone line search method and prove its convergence. Let us first state the definition of upper density convergence.

Definition 2.1. *Let K be a set of integers. The upper density of K, denoted by $ud(K)$ is the quantity*

$$ud(K) = \limsup_{i\to\infty} \frac{|K \cap [1, i]|}{i}\ , \tag{4}$$

where $|S|$ denotes cardinality of set S, and for integers i and j, $j \geq i$

$$[i, j] := \{i, i+1, \ldots, j\}.$$

The convergence in upper density is defined and proved by means of optimality function. The function $\theta : R^p \to R^+$ is an optimality function if $\theta(x) = 0$ if and only if x satisfies the optimality conditions.

Definition 2.2. *An algorithm which generates sequences $x_1, x_2, \ldots$ in $\mathbb{R}^p$ is said to converge with zero upper density (ud) on compact sets if with probability 1, if $\{x_i\}$ is a bounded sequence, then there exists a set of integers J, such that $ud(J) = 0$ and $\theta(x_i) \underset{i \notin J}{\to} 0$.*

We will prove that the nonmonotone line search method we propose here converges in upper density as in [12]. To do so, we need to assume the following.

Assumption A1. *[12]*

$$\text{If } x_i \to x,\ x_i \in R^p, i = 1, 2, 3, .. \text{ then } \theta(x) = 0 \text{ if and only if } \theta(x_i) \to 0. \tag{5}$$

The optimality function we consider is the norm of the gradient of the objective function and thus the assumption above is satisfied.

An algorithm which generates sequence $\{x_i\}_{i \in N}$, converges with *zero upper density* on a compact set if the sequence is bounded and there exists w.p.1 a set J with $ud(J) = 0$, such that the any accumulation point of subsequence $\{x_i\}_{i \in N \setminus J}$ satisfies the optimality conditions.

Let us now recall the notation needed for formulation of conditions for convergence with zero upper density on compact sets, [12]. For every compact set $\Gamma \subset R^p$, $r \geq 0$, $s \geq 0$ and integer i, let us define the following events:

- $E_i(\Gamma, r)$ is the event that $x_i \in \Gamma$ and $\theta(x_i) \geq r$.
- $G_i(\Gamma, s)$ is the event that $x_i \in \Gamma$ and $f(x_{i+1}) - f(x_i) \geq -s$.
- $H_i(\Gamma, s)$ is the event that $x_i \in \Gamma$ and $f(x_{i+1}) - f(x_i) \geq s$.

Here, $\mathcal{F}_i$ is the σ-algebra generated by all the information leading to the construction of x_i.

The following two conditions together constitute a sufficient condition for the convergence in zero upper density if f is continuous function and the iterations are generated by a line search with a random sample of predetermined size at each iteration, [12]. Let C_i be an arbitrary event from $\mathcal{F}_i$.

Condition 2.3. *[12] For every compact set $\Gamma \subset R^p$ and $r > 0$, there exists $s > 0$ such that, for every $\epsilon > 0$, there exists an integer I such that for every $i \geq I$ and event $C_i \in \mathcal{F}_i$*

$$P(G_i(\Gamma, s) | E_i(\Gamma, r), C_i) < \epsilon \tag{6}$$

Condition 2.4. *[12] For every compact set $\Gamma \subset R^p$, $s > 0$ and $\epsilon > 0$, there exists an integer I such that for every $i \geq I$ and event $C_i \in \mathcal{F}_i$*

$$P(H_i(\Gamma, s) | C_i) < \epsilon \tag{7}$$

The following two assumptions characterise the problem we consider more closely.

Assumption A2. *The objective function f has the form (2), and $g(\cdot, \omega) \in C^2(R^p)$.*

Assumption A3. *For every compact set $\Gamma \subset R^p$, there exists $K > 0$ such that, for every $x \in \Gamma$ and $\omega \in \Omega$,*

$$|g(x, \omega)| + \|\frac{\partial g}{\partial x}(x, \omega)^T\| + \|\frac{\partial^2 g}{\partial x^2}(x, \omega)\| \leq K, \tag{8}$$

where $\| \cdot \|$ denotes vector norm, or induced matrix norm, depending on context.

The consequence of A3 is that f is continuously differentiable and ∇f is Lipschitz continuous on compact sets, so

$$\nabla f(x) = E\left(\frac{\partial g}{\partial x}(x, \omega)^T\right). \tag{9}$$

This fact justifies the choice of $\|\nabla f(x)\|$ as the optimality function i.e. $\theta(x) = \|\nabla f(x)\|$. Clearly, the condition (5) holds.

3. The Nonmonotone Line Search Method

Line Search algorithm presented here is a modification of the algorithm presented in [12]. Instead of monotone Armijo-type line search with negative gradient as the search direction, we use a general search direction satisfying (12), and nonmonotone Armijo rule. The nonmonotonicity is defined by a sequence $\{\epsilon_i\}_{i \in N}$ such that

$$\epsilon_i > 0, \sum_{i=0}^{\infty} \epsilon_i < \infty. \tag{10}$$

Algorithm 3.1. *Input:* $x_0 \in R^p$, $\{n(i)\}_{i \in N}$, $\{\epsilon_i\}_{i \in N}$, $\alpha \in (0,1)$, $\beta \in (0,1)$

Step 0. *Set* $i = 0$.

Step 1. *Randomly draw* $n(i)$ *sample points* $\omega^i := \{\omega_{i,1}, \omega_{i,2}, \dots, \omega_{i,n(i)}\} \in \Omega$.

Step 2. *Choose a search direction* h_i.

Step 3. *Set* $k(i)$ *to be the smallest integer* k *satisfying*

$$G(x_i - \beta^k h_i, \omega^i) - G(x_i, \omega^i) \leq -\alpha \beta^k \|h_i\|^2 + \epsilon_i. \tag{11}$$

Set $x_{i+1} = x_i - \beta^{k(i)} h_i$, $i = i + 1$ *and go to Step 1.*

In Step 3 our goal is to find the step size that satisfies the nonmonotone Armijo condition, i.e. find the appropriate $k(i)$ that satisfies (11). Notice that Algorithm is well defined for an arbitrary search direction as $\epsilon_i > 0$ so for any h_i there exists $k(i)$ large enough such that (11) holds and Step 3 finishes with a finite $k(i)$.

Theorem 3.2. *Assume that A2-A3 hold. If the search directions* h_i *in Step 2 of Algorithm are chosen such that*

$$\lim_{i \to \infty} \frac{\|\nabla G(x_i, \omega^i) - h_i\|}{\epsilon_i} = 0, \tag{12}$$

where $G(x_i, \omega^i) := \frac{1}{n(i)} \sum_{j=1}^{n(i)} g(x_i, \omega_{i,j})$ *and* $\nabla G(x_i, \omega^i) := \frac{\partial G}{\partial x}(x_i, \omega^i)^T$, *then Algorithm converges with zero upper density on compact sets to a stationary point of (1).*

Proof. To prove the statement we need to show that Conditions 1 and 2 hold. Then the statement follows by Theorem 2.1 in [12]. Let $\Gamma \subset R^p$ be a compact set. First, we show that the sequence $\|h_i\|$ is bounded from above. Due to (12), there exists a constant K_0 such that $\|h_i - \nabla G(x_i, \omega^i)\| \leq K_0$. Also, (8) guaranties that there exists $K_1 > 0$ such that $\|\nabla G(x_i, \omega^i)\| \leq K_1$. So, for $M = 2\max\{K_0, K_1\}$, we have

$$\|h_i\| \leq \|h_i - \nabla G(x_i, \omega^i)\| + \|\nabla G(x_i, \omega^i)\| \leq M. \tag{13}$$

Therefore, $\|h_i\|$ is bounded from above. Let us prove now that

$$\lim_{i \to \infty} \frac{|h_i^T h_i - \nabla G(x_i, \omega^i)^T h_i|}{\epsilon_i} = 0. \tag{14}$$

Given that

$$0 < |h_i^T h_i - \nabla G(x_i, \omega^i)^T h_i| \leq \|h_i - \nabla G(x_i, \omega^i)\| \cdot \|h_i\| \tag{15}$$

and that $\|h_i\|$ is bounded, the limit (12) implies that (14) holds.

Let us now prove that for an arbitrary compact set $\Gamma \subset R$ there exists an integer $\bar{k}$ such that for every $x_i \in \Gamma$ we have . Let $x_i \in \Gamma$ and $\lambda \geq 0$. By the Mean value theorem we have

$$G(x_i - \lambda h_i, \omega^i) - G(x_i, \omega^i) = -\lambda \frac{\partial G}{\partial x}(x_i, \omega^i) h_i + \lambda^2 \int_0^1 (1-s) \langle \frac{\partial^2 G}{\partial x^2}(x_i - s\lambda h_i, \omega^i) h_i, h_i \rangle ds \tag{16}$$

By (14), there exists an integer i_0 such that for every $i \geq i_0$

$$-\lambda \frac{\partial G}{\partial x}(x_i, \omega^i) h_i \leq -\lambda \|h_i\|^2 + \epsilon_i. \tag{17}$$

By Cauchy - Schwarz's inequality, continuity of $\frac{\partial^2 G}{\partial x^2}(\cdot, \omega^i)$ and boundedness of $\|h_i\|$ we obtain

$$|\lambda^2 \int_0^1 (1-s) \langle \frac{\partial^2 G}{\partial x^2}(x_i - s\lambda h_i, \omega^i) h_i, h_i \rangle ds| \leq \lambda^2 K \|h_i\|^2 \tag{18}$$

Now, (16)-(18) implies

$$G(x_i - \lambda h_i, \omega^i) - G(x_i, \omega^i) \leq -\lambda(1 - \lambda K)\|h_i\|^2 + \epsilon_i \tag{19}$$

Substituting $\lambda = \beta^k$ in the above inequality, we get that (11) is satisfied if $\beta^k \leq (1-\alpha)/K$ holds, i.e, (11) holds for all $k \geq \frac{\log((1-\alpha)/K)}{\log \beta}$. Therefore, there exists $\bar{k}$ such that $k(i) \leq \bar{k}$.

Let us consider Condition 2.3. Let $\Gamma \subset R^p$ be a compact set. Take $r > 0$ and $s = \frac{1}{2}\alpha\beta^{\bar{k}} r^2$ and $\epsilon > 0$. We can choose $\delta \in (0, r)$ such that

$$\alpha\beta^{\bar{k}}(r-\delta)^2 \geq s.$$

As $\sum_{i=0}^{\infty} \epsilon_i < \infty$, there exists an integer i_1 such that for every $i \geq i_1$ we have $\epsilon_i \leq \delta$.

Let $A(i)$ be the event: $x_i \in \Gamma$, and

$$\|\nabla f(x_i) - \nabla G(x_i, \omega^i)\| < \frac{\delta}{2},\ |f(x_i) - G(x_i, \omega^i)| < \frac{\delta}{2},\ |f(x_{i+1}) - G(x_{i+1}, \omega^i)| < \frac{\delta}{2}.$$

By the Weak Law of Large Numbers there exists an integer i_2 such that for every $i \geq i_2$

$$P(A(i)|C_i, x_i \in \Gamma) \geq 1 - \epsilon.$$

With $I = \max\{i_0, i_1, i_2\}$, for all $i \geq I$, if $A(i)$ is satisfied and $\|\nabla f(x_i)\| \geq r$ then

$$\begin{aligned} \|\nabla f(x_i) - h_i)\| &= \|\nabla f(x_i) - G(x_i, \omega^i) + G(x_i, \omega^i) - h_i\| \\ &\leq \|\nabla f(x_i) - G(x_i, \omega^i)\| + \|G(x_i, \omega^i) - h_i\| \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta, \end{aligned}$$

and

$$\begin{aligned} \|h_i\| &= \|\nabla f(x_i) - (\nabla f(x_i) - h_i)| \geq \Big| \|\nabla f(x_i)\| - \|(\nabla f(x_i) - h_i)\| \Big| \\ &\geq \|\nabla f(x_i)\| - \|(\nabla f(x_i) - h_i)\| \geq r - \delta. \end{aligned}$$

Then the following holds

$$\begin{aligned} f(x_{i+1}) - f(x_i) &= f(x_{i+1}) - G(x_{i+1}, \omega^i) - \Big(f(x_i) - G(x_i, \omega^i)\Big) + G(x_{i+1}, \omega^i) - G(x_i, \omega^i) \\ &\leq \delta - \alpha\beta^{\bar{k}}\|h_i\|^2 + \epsilon_i \leq 2\delta - \alpha\beta^{\bar{k}}(r-\delta)^2 \leq -s. \end{aligned}$$

The above inequalities imply that under $E_i(\Gamma, r)$ and C_i, $A(i)$ implies $\overline{G_i(\Gamma, s)}$. Therefore, $G_i(\Gamma, s)$ implies $\overline{A(i)}$, and conditional probability of $G_i(\Gamma, s)$ is less or equal than the conditional probability of $\overline{A(i)}$. As

$$P(\overline{A(i)}|C_i, x_i \in \Gamma) \leq \epsilon,$$

we conclude that

$$P(G_i(\Gamma, s)|E_i(\Gamma, r), C_i) < \epsilon$$

i.e., Condition 2.3 is fulfilled.

To prove Condition 2.4 we consider again a compact set $\Gamma \subset R$, $s > 0$ and $\epsilon > 0$. As $\sum_{i=0}^{\infty} \epsilon_i < \infty$, we can take an integer i_0 such that for every $i \geq i_0$ there holds

$$\epsilon_i \leq \frac{s}{3}.$$

As f is Lipschitz continuous on Γ and (13) holds, for $x_{i+1} = x_i - \beta^{k(i)} h_i$ there exist constants $L > 0$, and $M > 0$ such that

$$|f(x_{i+1}) - f(x_i)| \leq LM\beta^{k(i)}.$$

Thus, there exists an integer $\bar{k}$, such that if $k(i) \geq \bar{k}$, then

$$f(x_{i+1}) - f(x_i) \leq s. \tag{20}$$

Now, we consider the case $k(i) \leq \bar{k}$. Let $B(i)$ be the event

$$x_i \in \Gamma,\ k(i) \leq \bar{k},\ |f(x_i) - G(x_i, \omega^i)| < \frac{s}{3},\ |f(x_{i+1}) - G(x_{i+1}, \omega^i)| < \frac{s}{3}.$$

If the event $B(i)$ is realized, then

$$\begin{aligned} f(x_{i+1}) - f(x_i) &= f(x_{i+1}) - G(x_{i+1}, \omega^i) - \left(f(x_i) - G(x_i, \omega^i)\right) + G(x_{i+1}, \omega^i) - G(x_i, \omega^i) \\ &\leq \frac{2s}{3} - \alpha\beta^{k(i)}||h_i||^2 + \epsilon_i \leq s. \end{aligned}$$

Again, by the Weak law of large number, there exists an integer i_1, such that for all $i \geq i_1$

$$P(B(i), k(i) \leq \bar{k}|C_i, x_i \in \Gamma) \geq 1 - \epsilon.$$

Taking $I = \max\{i_0, i_1\}$, we have that for all $i \geq I$ and $C_i \in \mathcal{F}_i$

$$P(\overline{B(i)}, k(i) \leq \bar{k}|C_i, x_i \in \Gamma) \leq \epsilon. \tag{21}$$

Now, (21), (20) and (7) imply that Condition 2.4 is fulfilled. As Conditions 2.3 - 2.4 are satisfied, the statement follows by Theorem 2.1 in [12]. □

4. Numerical Results

In this section we report some preliminary numerical results that confirm theoretical results and demonstrate efficiency of the proposed approach. We consider the following four test examples, defined as

$$g(x, \omega) = \phi(\omega x), \quad \omega : \mathcal{N}(1, \sigma^2),$$

where $\phi : R^p \to R$. The testing is done for two variance levels $\sigma^2 = 0.1$ and $\sigma^2 = 1$, using test functions ϕ taken from [2] and [9]:

AP Aluffi-Pentini's Problem, $p = 2$

$$g(x, \omega) = 0.25(\omega x_1)^4 - 0.5(\omega x_1)^2 + 0.1(\omega x_1) + 0.5(\omega x_2)^2.$$

EXP Exponential Problem $p = 10$

$$g(x, \omega) = \exp(-0.5 \sum_{i=1}^{p} (\omega x_i)^2).$$

SAL Salomon Problem $p = 10$

$$g(x, \omega) = 1 - \cos(2\pi \|\omega x\|) + 0.1\|\omega x\|, \text{ where } \|\omega x\| = \sqrt{\sum_{i=1}^{p} (\omega x_i)^2}.$$

SPH Sphere function or first function of De Jongs $p = 10$

$$g(x, \omega) = \sum_{i=1}^{p} (\omega x_i)^2.$$

Theoretical results are obtained for the case $n \to \infty$. But clearly, in actual implementation one can work only with finite sample size. Let $n_{\max}$ denote the maximal sample size allowed and we fixed $n_{\max} = 100$ for the first two problems, $n_{\max} = 1300$ for the third problem and $n_{\max} = 200$ for the last problem. The choice of $n_{\max}$ is highly non-trivial but we will not discuss it here as our aim is only to illustrate the potential advantages of nonmonotone line search rule.

The algorithm is implemented and tested against classical Armijo monotone line search rule ($\epsilon_i = 0$ in Algorithm) for two search directions, the first one being the negative gradient while the second direction is the finite difference approximation of the negative gradient $\nabla_\xi G(x_i, \omega^i)$, defined in [10]. The jth component is defined as

$$\frac{G(x_i + \xi e_j, \omega^i) - G(x_i - \xi e_j, \omega^i)}{2\xi},$$

where e_j denotes the jth coordinate vector in $\mathbb{R}^p$ and $\xi = 10^{-4}$. The sequence $\{\epsilon_i\}$ is defined as $\epsilon_i = 2^{-i}, i = 1, 2, \ldots$. The convergence condition in Theorem 1 suggests that one should improve the gradient approximation defined above as the iterates progress towards the stationary point. Therefore, it would be natural to consider diminishing ξ in the finite difference approximation, perhaps even connected with ϵ_i and the sample size in each iteration. But the finite difference approximation with small ξ tends to be unstable in the stochastic problems, [10] and diminishing ξ might in fact deteriorate the quality of approximation. Furthermore, the numerical results clearly indicate that the approximate gradient defined by $\xi = 10^{-4}$ works well. This fact might also indicate that the convergence condition might be weakened, perhaps for a different type of convergence in probability. We leave this issue for future research. Therefore, we have implemented four different methods.

- NM1 Nonmonotone line search with the negative gradient search direction, $h_i = \nabla G(x_i, \omega^i)$
- NM2 Nonmonotone line search with the finite difference approximation of the negative gradient. $h_i = \nabla_\xi G(x_i, \omega^i)$
- M1 Monotone (Armijo) line search with the the negative gradient search direction, $h_i = \nabla G(x_i, \omega^i)$
- M2 Monotone (Armijo) line search with the finite difference approximation of the negative gradient. $h_i = \nabla_\xi G(x_i, \omega^i)$

The sample size in each iteration is defined as $n(i+1) = \min\{\lceil 1.1n(i) \rceil, n_{max}\}$, with the initial value $n(0) = 3$ and a new sample of the size $n(i)$ is generated in ith iteration. The algorithmic parameters are the same for all problems, the starting point is $x_0 = 10 \cdot [1, 1, \ldots, 1]^T$, $\alpha = 10^{-4}$ and backtracking is performed with $\beta = 0.5$. We also limited the number of backtracking steps to 5. The stopping criteria is satisfied in x_i if the norm of the gradient or its approximation is smaller than 10^{-2} and $n(i) = n_{max}$. The number of function evaluations is used as the algorithm performance measure. Thus, for NM1 and M1, each gradient calculation is counted as p function evaluation, while for NM2 and M2 we used the two-sided approximation of gradient, so each gradient calculation is counted as $2p$ function evaluation. The method is stopped if the maximal allowed number of function evaluation is exhausted, with the maximal number set to 10^7.

In the testing process, we generated 5 independent samples for each variance levels and all problems are tested using the same collection of samples.

The results are shown at Figure 1, using the performance profile graph [5], where the cost function is defined as the number of function evaluations. The graph clearly indicates that the nonmonotone line search outperforms the classical Armijo line search at the considered test collection for both search directions. As expected, negative gradient performs better than the finite difference approximation of the negative gradient but nevertheless works reasonable well, which is an important property for problems where the function is calculated using a black box and the exact gradient of g is not available.

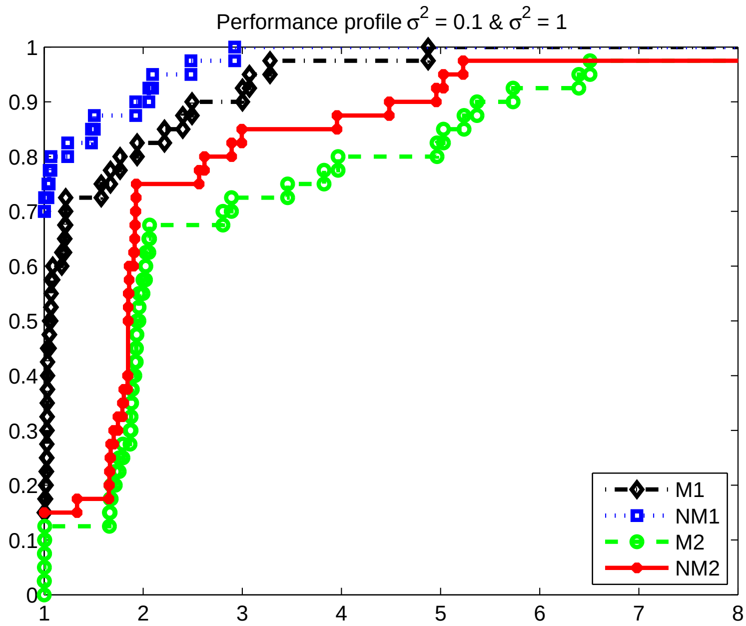

Figure 1: Performance profile for methods M1, NM1, M2, NM2 and two variance levels 0.1 and 1.

Acknowledgement We are grateful to the anonymous referee whose comments help us to improve the paper.

References

[1] M. Ahookhosh, S. Ghaderi, On efficiency of nonmonotone Armijo-type line searches, Applied Mathematical Modelling, 43,170–190.

[2] M. M. Ali, C. Khompatraporn, Z. B. Zabinsky, A numerical evaluation of several stochastic algorithms on selected continuous global optimization test problems, Journal of Global Optimization, 31(4) (2005) 635–672.

[3] R. H. Byrd, G. M. Chin, W. Neveitt, J. Nocedal, On the Use of Stochastic Hessian Information in Optimization Methods for Machine Learning, SIAM J. Optim. 21(3) (2011) 977–995.

[4] R. H. Byrd, G. M. Chin, J. Nocedal, Y. Wu, Sample size selection in optimization methods for machine learning, Mathematical Programming, 134(1) (2012) 127–155.

[5] E. D. Dolan, J. J. Moré, Benchmarking optimization software with performance profiles, Mathematical programming, 91(2) (2002) 201–213.

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

[7] N. Krejić, N. Krklec Jerinkić, Stochastic gradient methods for unconstrained optimization, Pesquisa Operacional, 34(3) (2014) 373–393.

[8] D. H. Li, M. Fukushima, A derivative-free line search and global convergence of Broyden-like method for nonlinear equations, Optimization Methods and Software, 13 (2000) 181–201.

[9] M. Molga, C. Smutnicki, Test functions for optimization needs, 2005, `http://www.vafaeijahan.com/en/wp-content/uploads/2012/02/Test-functions-for-optimization-needs.pdf`

[10] J. C. Spall, Introduction to stochastic search and optimization: estimation, simulation, and control. Wiley-Interscience Serises in Discrete Mathematics, New Jersey 2003.

[11] Y. Wardi, A stochastic steepest-descent algorithm, Journal of Optimization Theory and Applications 59(2) (1988) 307–323.

[12] Y. Wardi, Stochastic algorithms with Armijo stepsizes for minimization of functions, Journal of Optimization Theory and Applications 64(2) (1990) 399–417.

[13] D. Yan, H. Mukai, Optimization Algorithm with Probabilistic Estimation, Journal of Optimization Theory and Applications 79(2) (1993) 345–371.