Nonmonotone line search methods with variable sample size

Nataša Krejić[*] Nataša Krklec Jerinkić [*]

April 21, 2014

Abstract

Nonmonotone line search methods for unconstrained minimization with the objective functions in the form of mathematical expectation are considered. The objective function is approximated by the sample average approximation (SAA) with a large sample of fixed size. The nonmonotone line search framework is embedded with a variable sample size strategy such that different sample size at each iteration allow us to reduce the cost of the sample average approximation. The variable sample scheme we consider takes into account the decrease in the approximate objective function and the quality of the approximation of the objective function at each iteration and thus the sample size may increase or decrease at each iteration. Nonmonotonicity of the line search combines well with the variable sample size scheme as it allows more freedom in choosing the search direction and the step size while the sample size is not the maximal one and increases the chances of finding a global solution. Eventually the maximal sample size is used so the variable sample size strategy generates the solution of the same quality as the SAA method but with significantly smaller number of function evaluations. Various nonmonotone strategies are compared on a set of test problems.

Key words: nonmonotone line search, sample average approximation, variable sample size

[*] Department of Mathematics and Informatics, Faculty of Science, University of Novi Sad, Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia, e-mail: natasak@uns.ac.rs,natasa.krklec@dmi.uns.ac.rs. Research supported by Serbian Ministry of Education Science and Technological Development, grant no. 174030

1 Introduction

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

$$\min_{x \in \mathbb{R}^n} f(x) := E[F(x, \xi)]$$

where $\xi \in \Omega$ is a random vector. As $f(x)$ is rarely available analytically, one of the common approaches is to approximate the problem with

$$\min_{x \in \mathbb{R}^n} \hat{f}_N(x) = \frac{1}{N} \sum_{i=1}^{N} F(x, \xi_i). \tag{1}$$

applying the sample average approximation [1], [28]. Here N represents the sample size and $\{\xi_1, \ldots, \xi_N\}$ is a fixed sample generated at the beginning of the optimization process and kept throughout the whole process. In general, $F(x, \xi_i)$ does not have to be random. Many data fitting problems can be formulated in the same form as (1). For further references one can see [12]. The same type of problems arise in machine learning, see [5, 6]. We will treat N as the sample size in this paper.

In general, the sample size N is some large number and the evaluation of $\hat{f}_N$ is expensive. Thus applying an optimization method on the function $\hat{f}_N(x)$ can be very costly. Therefore, methods that start with a small sample and increase the sample size throughout the optimization process are developed in many papers, see [9], [12], [16], [24], [25], [27]. In these methods, at each iteration one considers an approximate objective function of the form given by (1) but with the current sample size N_k that might be smaller than N. The dominant way of dealing with the sample sizes (the schedule sequence from now on) is to consider an increasing sequence and thus ensure increasing precision during the optimization procedure regardless of the progress made, or not made, in decreasing the objective function given by (1). Intuitively it might be appealing to take the progress in the objective function as well as the error of the sample average approximation into account when designing the schedule sequence. An algorithm that allows the sample size to oscillate according to some measure of the decrease of the objective function is proposed in [1], [2], [3] for the trust region approach. A similar framework for the schedule sequence in the line search framework is developed in [17]. In this paper we are extending this schedule to the nonmonomotone line search framework. The main advantage of the proposed methods is the fact that they result in approximate solutions for the SAA problem but with significantly smaller computational costs than the classical SAA method.

A strong motivation for using nonmonotone line search is coming from problems where the search direction is not necessary descent. This happens, for instance, when derivatives are not available. This scenario is very realistic in stochastic optimization framework where only input-output information are available. When a variable schedule sample average approximation is used, the objective function at an arbitrary iteration is not necessarily equal to (1) and thus a descent direction for $\hat{f}_N$ does not need to be descent for the current approximation of the objective function $\hat{f}_{N_k}$. Furthermore, some efficient quasi-Newton methods, for example the SR1 update, do not produce descent direction at every single iteration, see [23]. It is well known that even some very efficient gradient-related methods do not posses monotonicity property at all, for example the spectral gradient methods, [19], [26], [30]. In these cases, it is useful to consider nonmonotone rules which do not require decrease of the objective function at every iteration. Moreover, when it comes to global convergence, numerical results in [32], [8], [26], [31] suggest that nonmonotone techniques have better chances of finding global optimizers than their monotone counterparts.

Evaluating optimization methods is a problem itself and it has been the main issue of some research efforts [21], [11]. Methods considered in this paper are evaluated mainly by means of the efficiency index [18] and the performance profile [11]. Both quantities are defined with respect to the number of function evaluations.

In this paper we introduce and analyze a class of algorithms that use nonmonotone line search rules which fit the variable sample size context developed in [17]. The nonmonotone line search framework for (1) as well as the schedule sequence are defined in the following section. In Section 3 we prove global convergence results for general search direction. A generalization of the results regarding the descent directions and R-linear convergence, which are obtained in [32] and [8], is presented in Section 4. A set of numerical examples that illustrate the properties of the considered methods is presented in Section 5. Two sets of examples are considered, the first one consists of academic optimization problems in noisy environment where the mathematical expectation is the true objective function that is approximated by (1). The second example comes from a research on various factors affecting the metacognition and the feeling of knowing among 746 students in Serbia. [1] and this example fits the framework of data fitting as defined in [12].

2 The algorithms

Suppose that N_{max} is some substantially large but finite positive integer and $\{\xi_1, \ldots, \xi_{N_{\max}}\}$ is a generated sample. The problem we consider is (1) with $N = N_{max}$, i.e.

$$\min_{x \in \mathbb{R}^n} \hat{f}_{N_{\max}}(x). \tag{2}$$

The algorithm we state allows us to vary the sample size $N_k \leq N_{\max}$ across iterations and therefore we are considering different functions $\hat{f}_{N_k}$ during the optimization process. Eventually $\hat{f}_{N_{\max}}$ will be considered and (2) will be solved.

The line search rule we consider seeks for a step size α_k that satisfies the condition

$$\hat{f}_{N_k}(x_k + \alpha_k p_k) \leq \tilde{C}_k + \varepsilon_k - \eta dm_k(\alpha_k), \tag{3}$$

where $\eta \in (0, 1]$, $\tilde{C}_k$ is a parameter related to the approximate objective function $\hat{f}_{N_k}(x_k)$, ε_k provides nonmonotonicity and $dm_k(\alpha_k)$ measures the decrease. The function $\hat{f}_{N_k}(x)$ is computed by (1) using the subset of the first N_k elements of $\{\xi^1, \ldots, \xi^{N_{\max}}\}$. This rather general form of the line search rule allows us to consider the stochastic counterparts of the main nonmonotone line search algorithms.

Let us first consider the measure of decrease represented by the function $dm_k(\alpha)$ defined in the following two ways. The first one is

$$dm_k(\alpha) = -\alpha p_k^T \nabla \hat{f}_{N_k}(x_k). \tag{4}$$

This definition is used only if $p_k^T \nabla \hat{f}_{N_k}(x_k) < 0$ and in that case $dm_k(\alpha) > 0$ for every $\alpha > 0$. The second option is to define

$$dm_k(\alpha) = \alpha^2 \beta_k, \tag{5}$$

where $\{\beta_k\}$ is a bounded sequence of positive number satisfying the following implication

$$\lim_{k \in K} \beta_k = 0 \Rightarrow \lim_{k \in K} \nabla \hat{f}_{N_{max}}(x_k) = 0, \tag{6}$$

[1] We are grateful to Ivana Rančić who provided the data. The research is done within the project no. 179010, financed by the Ministry of Education, Science and Technological Development, Republic of Serbia.

for every infinite subset of indices $K \subseteq \mathbb{N}$. This sequence is introduced in [10]. Besides some increasingly accurate approximation of $\|\nabla \hat{f}_{N_{max}}(x_k)\|$, a suitable choice for β_k can be even some positive constant.

The parameters $\varepsilon_k > 0$ make the line search (3) well defined for an arbitrary search direction p_k. The nonmonotone rules which contain a sequence of nonnegative parameters $\{\varepsilon_k\}_{k\in\mathbb{N}}$ were introduced in [20] for the first time and successfully used in many other algorithms, see [4] for example. The following property of the parameter sequence is assumed

$$\varepsilon_k > 0, \sum_k \varepsilon_k = \varepsilon < \infty. \tag{7}$$

Finally, let us comment the parameters $\tilde{C}_k$. Again, two different parameters are considered. The first one is defined as

$$\tilde{C}_k = \max\{C_k, \hat{f}_{N_k}(x_k)\}. \tag{8}$$

Here C_k is a convex combination of the objective function values in previous iterations as introduced in [32]. A nonmonotone generalized of the Armijo type rule with such C_k is considered in [7]. However, we are dealing with a different function at every iteration and $C_k \geq \hat{f}_{N_k}(x_k)$ might not be true. In order to make the algorithm well defined an additional definition is specified in (8) to ensure $\tilde{C}_k \geq \hat{f}_{N_k}(x_k)$. The definition of C_k is conceptually the same as in [32] except for a modification needed due to the variable sample size scheme. Therefore, we define C_k recursively with

$$C_{k+1} = \frac{\tilde{\eta}_k Q_k}{Q_{k+1}} C_k + \frac{1}{Q_{k+1}} \hat{f}_{N_{k+1}}(x_{k+1}), \quad C_0 = \hat{f}_{N_0}(x_0), \tag{9}$$

where

$$Q_{k+1} = \tilde{\eta}_k Q_k + 1, \quad Q_0 = 1, \quad \tilde{\eta}_k \in [0,1]. \tag{10}$$

The parameter $\tilde{\eta}_k$ determines the level of monotonicity regarding C_k. Notice that $\tilde{\eta}_{k-1} = 0$ yields $\tilde{C}_k = \hat{f}_{N_k}(x_k)$. On the other hand, the choice $\tilde{\eta}_k = 1$ for every k generates the average

$$C_k = \frac{1}{k+1} \sum_{i=0}^{k} \hat{f}_{N_i}(x_i). \tag{11}$$

Clearly, $1 \leq Q_k \leq k+1$ for every k. Furthermore, one can see that C_k is a convex combination of $\hat{f}_{N_0}(x_0), ..., \hat{f}_{N_k}(x_k)$. Moreover, it can be proved that the following lemma, analogous to the statements in [32] (Theorem 2.2, (2.15)), holds.

Lemma 2.1. *Suppose that $\tilde{\eta}_k \in [\eta_{min}, \eta_{max}]$ for every k where $0 \leq \eta_{min} \leq \eta_{max} \leq 1$ and Q_k is defined by (10).*

1) If $\eta_{min} = 1$, then $\quad \lim_{k \to \infty} Q_k^{-1} = 0$.

2) If $\eta_{max} < 1$, then $\quad 0 \leq Q_k \leq (1 - \eta_{max})^{-1}$ *and* $\lim_{k \to \infty} Q_k^{-1} > 0$.

The previous lemma distinguishes two cases regarding C_k or more precisely regarding $\tilde{\eta}_k$. It turns out that the average (11) is a special case in terms of the convergence analysis too and a stronger result can be obtained by excluding $\tilde{\eta}_k = 1$.

In order to cover all relevant nonmonotone line search rules, we define the second possibility for $\tilde{C}_k$,

$$\tilde{C}_k = \max\{\hat{f}_{N_k}(x_k), \ldots, \hat{f}_{N_{\max\{k-M+1,0\}}}(x_{\max\{k-M+1,0\}})\}, \tag{12}$$

where $M \in \mathbb{N}$ is arbitrary but fixed. This rule originates from [14] and can be found in many successful algorithms, see [10] and [15] for example.

The following assumption ensures that the problem is well defined.

A 1. *There exists a constant M_F such that for every ξ, x, $M_F \leq F(x, \xi)$.*

The consequence of this assumption is that every function $\hat{f}_{N_k}$ is bounded from below with the same constant M_F and therefore the sequence C_k is also bounded from below, i.e. $M_F \leq C_k$. The same holds for $\tilde{C}_k$ as well.

Each iteration of the method we consider generates a new iteration x_{k+1} and a new sample size N_{k+1}. The new iteration x_{k+1} is obtained as $x_{k+1} = x_k + \alpha_k p_k$ using (3) and the function $\hat{f}_{N_k}$. After that the sample size and the objective function are updated as follows. The new sample size N_{k+1} is determined by Algorithms 1 and 2 which are stated below. Besides the schedule sequence $\{N_k\}$, two additional sequences $N_k^{\min}$ and N_k^+ are also defined. The sequence $N_k^{\min}$ is nondecreasing and represents the lower bounds for the schedule sequence. The sequence N_k^+ is generated by Algorithm 1 and represents the candidate sample sizes which are further considered in Algorithm 2. The Algorithms 1 and 2 presented in this paper are adjustments of the corresponding algorithms analyzed in [17]. The adjustments are made to fit the nonmonotone framework and are mainly technical. A more detailed analysis can be found in [17] but we state the algorithms here for the sake of completeness.

The following algorithm yields the candidate sample size N_k^+. Notice that it is constructed to provide $N_k^{min} \leq N_k^+ \leq N_{max}$. The algorithm relies on a good balance between the progress made in decreasing the objective function (measured by dm_k) and the (lack of) precision in the current approximate objective function $\hat{f}_{N_k}$. The lack of precision is defined as

$$\varepsilon_{\delta}^{N_k}(x_k) = \hat{\sigma}_{N_k}(x_k) \frac{\alpha_{\delta}}{\sqrt{N_k}}, \tag{13}$$

which is an approximate width of the confidence interval around $f(x_k)$ with $\hat{\sigma}_{N_k}(x_k)$ being the sample standard deviation and α_{δ} the corresponding quantile of the Gaussian distribution $\mathcal{N}(0,1)$ with $\delta = 0.95$.

ALGORITHM 1.

```
S0 Input parameters: $dm_k,\ N_k^{min},\ \varepsilon_{\delta}^{N_k}(x_k),\ \nu_1 \in (0,1), d \in (0,1]$.
S1 Determine $N_k^+$
    1) $dm_k = d\ \varepsilon_{\delta}^{N_k}(x_k) \quad \rightarrow \quad N_k^+ = N_k$.
    2) $dm_k > d\ \varepsilon_{\delta}^{N_k}(x_k)$
       Starting with $N = N_k$, while $dm_k > d\ \varepsilon_{\delta}^{N}(x_k)$ and $N > N_k^{min}$,
       decrease $N$ by 1 and calculate $\varepsilon_{\delta}^{N}(x_k) \quad \rightarrow \quad N_k^+$.
    3) $dm_k < d\ \varepsilon_{\delta}^{N_k}(x_k)$
        i) $dm_k \geq \nu_1 d\ \varepsilon_{\delta}^{N_k}(x_k)$
           Starting with $N = N_k$, while $dm_k < d\ \varepsilon_{\delta}^{N}(x_k)$ and $N < N_{max}$,
           increase $N$ by 1 and calculate $\varepsilon_{\delta}^{N}(x_k) \quad \rightarrow \quad N_k^+$.
        ii) $dm_k < \nu_1 d\ \varepsilon_{\delta}^{N_k}(x_k) \quad \rightarrow \quad N_k^+ = N_{max}$.
```

Acceptance of the candidate sample size is decided within Algorithm 2. Notice that $N_{k+1} \geq N_k^+$.

ALGORITHM 2.

```
S0 Input parameters: $N_k^+,\ N_k,\ x_k,\ x_{k+1}$.
S1 Determine $N_{k+1}$
    1) If $N_k^+ \geqslant N_k$ then $N_{k+1} = N_k^+$.
    2) If $N_k^+ < N_k$ compute
```

$$\rho_k = \left| \frac{\hat{f}_{N_k^+}(x_k) - \hat{f}_{N_k^+}(x_{k+1})}{\hat{f}_{N_k}(x_k) - \hat{f}_{N_k}(x_{k+1})} - 1 \right|.$$

i) If $\rho_k < \frac{N_k - N_k^+}{N_k}$ put $N_{k+1} = N_k^+$.

ii) If $\rho_k \geq \frac{N_k - N_k^+}{N_k}$ put $N_{k+1} = N_k$.

The safeguard algorithm stated above is supposed to prohibit an unproductive decrease in the sample size. The right-hand side of the inequalities in S1 2) i)-ii) imply that if the proposed decrease $N_k - N_k^+$ is relatively large, then the chances of accepting the smaller sample size N_k^+ are larger. This reasoning is motivated by the empirical results which suggest that a large decrease in the sample size is almost always productive. On the other hand if N_k is close to N_{max} and the proposed decrease is relatively small it is far less likely that the decrease in the schedule sequence is meaningful.

The lower bound N_k^{min} is updated as follows.

- If $N_{k+1} \leq N_k$ then $N_{k+1}^{min} = N_k^{min}$.
- If $N_{k+1} > N_k$ and
  - N_{k+1} is a sample size which has not been used so far then $N_{k+1}^{min} = N_k^{min}$.
  - N_{k+1} is a sample size which has been used before and we have made a big enough decrease of the function $\hat{f}_{N_{k+1}}$ since the last time it has been used, then $N_{k+1}^{min} = N_k^{min}$.
  - N_{k+1} is a sample size which has been used before and we have not made a big enough decrease of the function $\hat{f}_{N_{k+1}}$ since the last time, then $N_{k+1}^{min} = N_{k+1}$.

The decrease of the function is not big enough if

$$\frac{\hat{f}_{N_{k+1}}(x_{h(k)}) - \hat{f}_{N_{k+1}}(x_{k+1})}{k+1-h(k)} < \frac{N_{k+1}}{N_{max}} \varepsilon_\delta^{N_{k+1}}(x_{k+1}).$$

where $h(k)$ is the iteration at which we started to use the sample size N_{k+1} for the last time. Notice that the average decrease of the function $\hat{f}_{N_{k+1}}$ after the iteration $h(k)$ is obtained on the left-hand side The average decrease is compared to the lack of precision throughout the ratio N_{k+1}/N_{max}. This means that a stronger decrease is required if the function $\hat{f}_{N_{k+1}}$ is closer to $\hat{f}_{N_{max}}$ as the real objective function is $\hat{f}_{N_{max}}$.

Now, we can state the main algorithm. The important modifications regarding the algorithm from [17] are in steps S4 and S6. The search direction does not have to be decreasing in general and the line search rule is changed. Consequently, the definition of dm_k is altered and therefore the input parameter of Algorithm 1 is modified, but the mechanism for searching N_k^+ remains the same. Another important modification in comparison with [17] is that Algorithm 3 does not have any stopping criterion. The reason is that in general the gradient of function $\hat{f}_{N_{max}}$ is not available.

ALGORITHM 3.

S0 Input parameters: $M, N_{max}, N_0^{min} \in \mathbb{N},\ x_0 \in \mathbb{R}^n,\ \delta, \beta, \nu_1 \in (0,1),\ \eta \in (0,1],\ 0 \leq \eta_{min} \leq \eta_{max} \leq 1,\ \{\varepsilon_k\}_{k \in \mathbb{N}}$ satisfying (7).

S1 Generate the sample realization: $\xi_1, \ldots, \xi_{N_{max}}$.
Set $N_0 = N_0^{min}$, $C_0 = \hat{f}_{N_0}(x_0)$, $Q_0 = 1$, $\tilde{C}_0 = C_0$, $k = 0$.

S2 Compute $\hat{f}_{N_k}(x_k)$ and $\varepsilon_\delta^{N_k}(x_k)$.

S3 Determine the search direction p_k.

S4 Find the smallest nonnegative integer j such that $\alpha_k = \beta^j$ satisfies

$$\hat{f}_{N_k}(x_k + \alpha_k p_k) \leq \tilde{C}_k + \varepsilon_k - \eta dm_k(\alpha_k).$$

S5 Set $s_k = \alpha_k p_k$ and $x_{k+1} = x_k + s_k$.

S6 Determine the candidate sample size N_k^+ using Algorithm 1 and $dm_k = dm_k(\alpha_k)$.

S7 Determine the sample size N_{k+1} using Algorithm 2.

S8 Determine the lower bound of sample size N_{k+1}^{min}.

S9 Determine $\tilde{C}_{k+1}$ using (8) or (12).

S10 Set $k = k + 1$ and go to step S2.

3 General search direction

In this section, we analyze the case where the search direction might be nondescent. The convergence analysis is conducted in two main stages. First, we prove that Algorithm 3 eventually ends up with $N_k = N_{max}$ for all k large enough and thus (2) is eventually solved. The second part of the analysis deals with the function $\hat{f}_{N_{max}}$. In order to prove that the schedule sequence becomes stationary with $N_k = N_{\max}$ for k large enough, we need to prove that a subsequence of $\{dm_k(\alpha_k)\}_{k\in\mathbb{N}}$ tends to zero. This is done by considering separately two definitions of $\tilde{C}_k$. Result for the line search with $\tilde{C}_k = \max\{C_k, \hat{f}_{N_k}(x_k)\}$ is stated in the next lemma. An additional assumption stated below is needed.

A 2. *For every ξ, $F(\cdot, \xi) \in C^1(\mathbb{R}^n)$.*

Lemma 3.1. *Suppose that assumptions A1 - A2 are satisfied and there exists $\tilde{n} \in \mathbb{N}$ such that $N_k = N$ for every $k \geq \tilde{n}$. Then Algorithm 3 with $\tilde{C}_k$ defined by (8) satisfies*

$$\liminf_{k\to\infty} dm_k(\alpha_k) = 0.$$

Moreover, if $\eta_{max} < 1$ it follows that

$$\lim_{k\to\infty} dm_k(\alpha_k) = 0.$$

Proof. First of all, recall that the line search is such that for every $k \geq \tilde{n}$ we have

$$\hat{f}_N(x_{k+1}) \leq \tilde{C}_k + \varepsilon_k - \eta dm_k \tag{14}$$

where $dm_k = dm_k(\alpha_k)$. Furthermore, $C_k \leq \max\{C_k, \hat{f}_{N_k}(x_k)\} = \tilde{C}_k$. Therefore, the following is true for every $k \geq \tilde{n}$

$$\begin{aligned} C_{k+1} &= \frac{\tilde{\eta}_k Q_k}{Q_{k+1}} C_k + \frac{1}{Q_{k+1}} \hat{f}_N(x_{k+1}) \\ &\leq \frac{\tilde{\eta}_k Q_k}{Q_{k+1}} \tilde{C}_k + \frac{1}{Q_{k+1}} (\tilde{C}_k + \varepsilon_k - \eta dm_k) \\ &= \tilde{C}_k + \frac{\varepsilon_k}{Q_{k+1}} - \eta \frac{dm_k}{Q_{k+1}} \end{aligned}$$

The previous equality follows from the equality $Q_{k+1} = \tilde{\eta}_k Q_k + 1$. Moreover, using $Q_{k+1} \geq 1$ we obtain

$$C_{k+1} \leq \tilde{C}_k + \varepsilon_k - \eta \frac{dm_k}{Q_{k+1}}. \tag{15}$$

Now, from (14) and (15) there follows

$$\begin{aligned}\tilde{C}_{k+1} &= \max\{C_{k+1}, \hat{f}_N(x_{k+1})\} \\ &\leq \max\{\tilde{C}_k + \varepsilon_k - \eta\frac{dm_k}{Q_{k+1}}, \tilde{C}_k + \varepsilon_k - \eta dm_k\} \\ &= \tilde{C}_k + \varepsilon_k - \eta\frac{dm_k}{Q_{k+1}}\end{aligned} \tag{16}$$

and for every $s \in \mathbb{N}$

$$\tilde{C}_{\tilde{n}+s} \leq \tilde{C}_{\tilde{n}} + \sum_{j=0}^{s-1} \varepsilon_{\tilde{n}+j} - \eta \sum_{j=0}^{s-1} \frac{dm_{\tilde{n}+j}}{Q_{\tilde{n}+j+1}}. \tag{17}$$

The sequence $\{\varepsilon_k\}_{k\in\mathbb{N}}$ satisfies (7). Furthermore, assumption A1 implies $C_k \geq M_F$ and we obtain

$$0 \leq \eta \sum_{j=0}^{s-1} \frac{dm_{\tilde{n}+j}}{Q_{\tilde{n}+j+1}} \leq \tilde{C}_{\tilde{n}} - M_F + \varepsilon := C.$$

Now, letting $s \to \infty$ we obtain

$$0 \leq \sum_{j=0}^{\infty} \frac{dm_{\tilde{n}+j}}{Q_{\tilde{n}+j+1}} \leq \frac{C}{\eta} < \infty. \tag{18}$$

Suppose that $dm_k \geq \bar{d} > 0$ for all k sufficiently large. Then $Q_k \leq k+1$ implies

$$\sum_{j=0}^{\infty} \frac{dm_{\tilde{n}+j}}{Q_{\tilde{n}+j+1}} \geq \sum_{j=0}^{\infty} \frac{\bar{d}}{\tilde{n}+j+2} = \infty$$

which is the contradiction with (18). Therefore, there must exist a subset of iterations K such that $\lim_{k\in K} dm_k = 0$ and the statement of this lemma follows.

Finally, assume that $\eta_{max} < 1$. From (18) we conclude that $\lim_{k\to\infty} Q_k^{-1} dm_k = 0$. Since Lemma 2.1 implies that $\lim_{k\to\infty} Q_k^{-1} > 0$ it follows that $\lim_{k\to\infty} dm_k = 0$. This completes the proof. ■

The analogous statement can be proved for

$$\tilde{C}_k = \max\{\hat{f}_{N_k}(x_k), \ldots, \hat{f}_{N_{\max\{k-M+1,0\}}}(x_{\max\{k-M+1,0\}})\}.$$

The existence of a subsequence of $\{dm_k(\alpha_k)\}_{k\in\mathbb{N}}$ that vanishes will be enough to prove the first stage result in the convergence analysis. In order to specify the subsequence that tends to zero, suppose that $\tilde{n} \geq M$ is the iteration such that for every $k \geq \tilde{n}$ the sample size is fixed. Furthermore, if we define $s(k) = \tilde{n} + kM$ then by the definition of $\tilde{C}_k$ we have $\tilde{C}_{s(k)} = \max\{\hat{f}_N(x_{s(k)}), \ldots, \hat{f}_N(x_{s(k)-M+1})\}$. Let $v(k)$ be the index such that $\tilde{C}_{s(k)} = \hat{f}_N(x_{v(k)})$ and notice that $v(k) \in \{s(k-1)+1, \ldots, s(k-1)+M\}$. Finally, define

$$K = \{v(k) - 1\}_{k\in\mathbb{N}}.$$

The proof of the following lemma is essentially the same as the proof of Proposition 1 from [10], applied on the function $\hat{f}_N$ and thus we omit it here.

Lemma 3.2. *Suppose that assumptions A1 - A2 are satisfied and there exists $\tilde{n} \in \mathbb{N}$ such that $N_k = N$ for every $k \geq \tilde{n}$. Then Algorithm 3 with $\tilde{C}_k$ defined by (12) satisfies:*

1) $\tilde{C}_{s(k+1)} \leq \tilde{C}_{s(k)} + \sum_{i=0}^{M-1} \varepsilon_{s(k)+i} - \eta dm_{v(k+1)-1}, \quad k \in \mathbb{N},$

2) $\tilde{C}_{s(m+1)} \leq \tilde{C}_{s(1)} + \sum_{k=1}^{m}\sum_{i=0}^{M-1} \varepsilon_{s(k)+i} - \eta \sum_{k=1}^{m} dm_{v(k+1)-1}, \quad m \in \mathbb{N},$

3) $\lim_{k\in K} dm_k(\alpha_k) = 0.$

The result stated in Lemma 3.1 concerning the case $\eta_{max} < 1$ is attainable in this case as well, but under stronger assumptions on the search directions and the objective function as will be shown in Section 4.

The previous two lemmas imply that

$$\liminf_{k\to\infty} dm_k(\alpha_k) = 0. \tag{19}$$

Now we are able to state the conditions which ensure that the schedule sequence eventually becomes stationary with $N_k = N_{max}$ for k large enough. The conditions are essentially the same as for the monotone line search rule from [17], Lemma 4.1. The main difference with respect to the monotone case lies in Lemma 3.2 and Lemma 3.1. Thus we give a short version of the proof here. The following assumption is needed.

A 3. *There exist $\kappa > 0$ and $n_1 \in \mathbb{N}$ such that $\varepsilon_{\delta}^{N_k}(x_k) \geq \kappa$ for every $k \geq n_1$.*

Lemma 3.3. *Suppose that the assumptions A1 - A3 are satisfied. Then there exists $q \in \mathbb{N}$ such that for every $k \geq q$ the sample size used by Algorithm 3 is maximal, i.e. $N_k = N_{max}$.*

Proof. First of all, recall that Algorithm 3 does not have any stopping criterion and the number of iterations is infinite by default. Notice that Algorithm 2 implies that $N_{k+1} \geq N_k^+$ is true for every k. Now, let us prove that sample size can not be stacked at a size that is lower than the maximal one.

Suppose that there exists $\tilde{n} > n_1$ such that $N_k = N^1 < N_{max}$ for every $k > \tilde{n}$ and define $dm_k = dm_k(\alpha_k)$. In that case (19) is valid, i.e. $\liminf_{k \to \infty} dm_k = 0$. On the other hand, we have that $\varepsilon_\delta^{N^1}(x_k) \geq \kappa > 0$ for every $k \geq \tilde{n}$ which means that $\nu_1 d\, \varepsilon_\delta^{N_k}(x_k)$ is bounded from below for every k sufficiently large. Therefore, there exists at least one $p \geq \tilde{n}$ such that $dm_p < \nu_1 d\, \varepsilon_\delta^{N^1}(x_p)$. However, the construction of Algorithm 1 would then imply $N_p^+ = N_{max}$ and we would have $N_{p+1} = N_{max} > N^1$ which is in contradiction with the current assumption that sample size stays at N^1.

We have just proved that sample size can not stay on $N^1 < N_{max}$. The rest of the proof is completely analogous to the proof of Lemma 4.1 in [17]. ■

Now, we prove that after a finite number of iterations, all the remaining iterates of the algorithm belong to the level set defined in the next lemma no matter which of the two definitions of dm_k is used. The level set does not depend on the starting point x_0 as it is usual in deterministic framework, but on the point at which the schedule sequence becomes stationary with $N_k = N_{max}$.

Lemma 3.4. *Suppose that A1 - A3 are satisfied. Then there exists a finite $q \in \mathbb{N}$ such that for every $k \geq q$ the iterates x_k belong to the level set*

$$\mathcal{L} = \{x \in \mathbb{R}^n \mid \hat{f}_{N_{max}}(x) \leq \tilde{C}_q + \varepsilon\}. \tag{20}$$

Proof. Lemma 3.3 implies the existence of a finite number $\tilde{n}$ such that $N_k = N_{max}$ for every $k \geq \tilde{n}$. If $\tilde{C}_k$ is defined by (8), for every $s \in \mathbb{N}$ inequality (17) is true. Therefore, we conclude that for every $s \in \mathbb{N}$

$$\tilde{C}_{\tilde{n}+s} \leq \tilde{C}_{\tilde{n}} + \sum_{j=0}^{s-1} \varepsilon_{\tilde{n}+j} - \eta \sum_{j=0}^{s-1} \frac{dm_{\tilde{n}+j}}{Q_{\tilde{n}+j+1}} \leq \tilde{C}_{\tilde{n}} + \varepsilon.$$

Since $\hat{f}_{N_{max}}(x_{\tilde{n}+s}) \leq \tilde{C}_{\tilde{n}+s}$ by definition, we obtain that

$$\hat{f}_{N_{max}}(x_k) \leq \tilde{C}_{\tilde{n}} + \varepsilon$$

holds for every $k \geq \tilde{n}$ which proves the statement with $q = \tilde{n}$. On the other hand, if (12) is used for $\tilde{C}_k$, Lemma 3.2 implies

$$\tilde{C}_{s(m+1)} \leq \tilde{C}_{s(1)} + \sum_{k=1}^{m} \sum_{i=0}^{M-1} \varepsilon_{s(k)+i} - \eta \sum_{k=1}^{m} dm_{v(k+1)-1} \leq \tilde{C}_{s(1)} + \varepsilon$$

where $s(m) = \tilde{n} + mM$ and $\hat{f}_{N_{max}}(x_{v(m)}) = \tilde{C}_{s(m)}$. In fact, $\tilde{C}_{s(k)} \leq \tilde{C}_{s(1)} + \varepsilon$ for every $k \in \mathbb{N}$. Moreover, since $\tilde{C}_{s(k)} = \max\{\hat{f}_{N_{max}}(x_{s(k-1)+1}), \ldots, \hat{f}_{N_{max}}(x_{s(k-1)+M})\}$ we have that $\hat{f}_{N_{max}}(x_{s(k-1)+j}) \leq \tilde{C}_{s(k)}$ for every $j \in \{1, \ldots, M\}$ and every $k \in \mathbb{N}$. Notice that $\tilde{C}_{s(1)} = \max\{\hat{f}_{N_{max}}(x_{\tilde{n}+1}), \ldots, \hat{f}_{N_{max}}(x_{\tilde{n}+M})\}$. Therefore, for every $k > \tilde{n}$

$$\hat{f}_{N_{max}}(x_k) \leq \tilde{C}_{s(1)} + \varepsilon = \tilde{C}_{\tilde{n}+M} + \varepsilon$$

which yields the result with $q = \tilde{n} + M$. ■

The rest of this section is devoted to general search directions. Therefore, the decrease measure is defined by (5) and the line search rule is

$$\hat{f}_{N_k}(x_k + \alpha_k p_k) \leq \tilde{C}_k + \varepsilon_k - \alpha_k^2 \beta_k. \tag{21}$$

The convergence results are stated in the following theorems and the statements are generalizations of the corresponding results from [32], [7] and [10] given that we consider different objective functions, $\tilde{C}_k$ and the sequence ε_k.

Theorem 3.1. *Suppose that A1 - A3 hold together with (6) and that the sequences of search directions $\{p_k\}_{k\in\mathbb{N}}$ and iterates $\{x_k\}_{k\in\mathbb{N}}$ of Algorithm 3 with the line search rule (21) are bounded. Then there exists an accumulation point (x^*, p^*) of the sequence $\{(x_k, p_k)\}_{k\in\mathbb{N}}$ that satisfies the following inequality*

$$p^{*T} \nabla \hat{f}_{N_{max}}(x^*) \geq 0.$$

If, in addition, $\tilde{C}_k$ is defined by (8) with $\eta_{max} < 1$, then the previous inequality holds for every accumulation point (x^*, p^*).

Proof. Notice that under these assumptions, Lemma 3.3 implies the existence of $\tilde{n} \in \mathbb{N}$ such that $N_k = N_{max}$ for every $k \geq \tilde{n}$. Moreover, there exists a subset $K_0 \subseteq \mathbb{N}$ such that $\lim_{k \in K_0} dm_k(\alpha_k) = \lim_{k \in K_0} \alpha_k^2 \beta_k = 0$. Furthermore, since $\{(x_k, p_k)\}_{k \in \mathbb{N}}$ is bounded, there exists at least one subset $K \subseteq K_0$ and points x^* and p^* such that $\lim_{k \in K} x_k = x^*$ and $\lim_{k \in K} p_k = p^*$. Therefore it follows that $\lim_{k \in K} \alpha_k^2 \beta_k = 0$. Moreover, if $\tilde{C}_k$ is defined by (8) with $\eta_{max} < 1$, then Lemma 3.1 implies that the whole sequence $\alpha_k^2 \beta_k$ converges to zero and thus any subsequence also converges to zero. The rest of the proof is completely analogous to the proof of Theorem 1 in [10]. ■

Roughly speaking, the previous theorem give the conditions under which the algorithm generates limit points such that no further descent direction is attainable. An additional assumption concerning the search directions p_k is needed to ensure that these limit points are stationary for $\hat{f}_{N_{max}}$.

A 4. *The sequence of search directions p_k is bounded and satisfies the following implication for any subset of iterations K*

$$\lim_{k \in K} p_k^T \nabla \hat{f}_{N_k}(x_k) = 0 \;\Rightarrow\; \lim_{k \in K} \nabla \hat{f}_{N_k}(x_k) = 0.$$

A 5. *Search directions p_k satisfy the condition* $\lim_{k \to \infty} p_k^T \nabla \hat{f}_{N_{max}}(x_k) \leq 0$.

Notice that the previous assumption is satisfied if we are eventually able to produce a descent search directions for $\hat{f}_{N_{max}}$. One possibility would be to use increasingly accurate finite differences to approximate the gradient.

Theorem 3.2. *Suppose that A1 - A5 and (6) hold and that the sequence $\{x_k\}_{k \in \mathbb{N}}$ of Algorithm 3 with the line search rule (21) is bounded. Then there exists an accumulation point of $\{x_k\}_{k \in \mathbb{N}}$ which is stationary for function $\hat{f}_{N_{max}}$. If, in addition, $\tilde{C}_k$ is defined by (8) with $\eta_{max} < 1$, then every accumulation point of $\{x_k\}_{k \in \mathbb{N}}$ is stationary for $\hat{f}_{N_{max}}$.*

Proof. Theorem 3.1 implies the existence of an accumulation point (x^*, p^*) of the sequence $\{(x_k, p_k)\}_{k \in \mathbb{N}}$ that satisfies

$$p^{*T} \nabla \hat{f}_{N_{max}}(x^*) \geq 0. \tag{22}$$

If $\tilde{C}_k$ is defined by (8) with $\eta_{max} < 1$, x^* can be considered as an arbitrary accumulation point. Let $K \subseteq \mathbb{N}$ be the subset of indices such that $\lim_{k \in K}(x_k, p_k) = (x^*, p^*)$. Since the search directions are bounded by assumption A4 and $\nabla \hat{f}_{N_{max}}$ is continuous as a consequence of assumption A2, assumption A5 implies that

$$p^{*T} \nabla \hat{f}_{N_{max}}(x^*) = \lim_{k \in K} p_k^T \nabla \hat{f}_{N_{max}}(x_k) \leq 0$$

which together with (22) implies $p^{*T}\nabla \hat{f}_{N_{max}}(x^*) = 0$. Finally, assumption A4 implies that $\nabla \hat{f}_{N_{max}}(x^*) = 0$. ■

Notice that the nonmonotone line search rules proposed in this section yielded the same result regarding achievement of the maximal sample size N_{max} as in the case of the monotone rule presented in [17]. The convergence results rely on the analysis applied on the function $\hat{f}_{N_{max}}$. The main result is the existence of an accumulation point which is stationary for $\hat{f}_{N_{max}}$ without imposing the assumption of descent search directions. Moreover, if the parameter $\tilde{C}_k$ is defined by (8) with $\eta_{max} < 1$, every accumulation point is stationary under the same assumptions.

4 Descent search direction

This section is devoted to the case where the gradient of $\hat{f}_{N_k}$ is available and a descent search direction is used at every iteration. Therefore, throughout this section we consider Algorithm 3 with the line search

$$\hat{f}_{N_k}(x_k + \alpha_k p_k) \leq \tilde{C}_k + \varepsilon_k + \eta \alpha_k p_k^T \nabla \hat{f}_{N_k}(x_k),\ \eta \in (0,1). \tag{23}$$

The parameters ε_k allow an additional degree of freedom for the step length and thus increase the chances of larger step sizes. This framework yields the possibility of obtaining convergence result where every accumulation point is stationary for the relevant objective function. Moreover, the R-linear rate of convergence is attainable if the sequence $\{\varepsilon_k\}$ is chosen such that it converges to zero R-linearly. Let us first introduce two additional assumptions and state an important technical results.

A 6. *For every ξ the gradient function $\nabla_x F(\cdot, \xi)$ is Lipschitz continuous on any bounded set.*

A 7. *There exist positive constants c_1 and c_2 such that for all k sufficiently large search directions p_k satisfy*

$$p_k^T \nabla \hat{f}_{N_k}(x_k) \leq -c_1 \|\nabla \hat{f}_{N_k}(x_k)\|^2,$$

$$\|p_k\| \leq c_2 \|\nabla \hat{f}_{N_k}(x_k)\|.$$

Lemma 4.1. *Suppose that assumptions A1 - A3 and A6 - A7 are satisfied. Then there exist positive constants $\bar{\beta}_0$ and c_3 such that for every k sufficiently large the following two inequalities hold*

$$dm_k(\alpha_k) \geq \bar{\beta}_0 \|\nabla \hat{f}_{N_{max}}(x_k)\|^2, \tag{24}$$

$$\|\nabla \hat{f}_{N_{max}}(x_{k+1})\| \leq c_3 \|\nabla \hat{f}_{N_{max}}(x_k)\| \tag{25}$$

Proof. Lemma 3.3 implies the existence of $\tilde{n}$ such that for every $k \geq \tilde{n}$ the sample size is $N_k = N_{max}$. Let us distinguish two types of iterations for $k \geq \tilde{n}$. The first type is when the full step is accepted, i.e. when $\alpha_k = 1$. In that case, A7 directly implies that $dm_k(\alpha_k) \geq c_1 \|\nabla \hat{f}_{N_{max}}(x_k)\|^2$. The second type is when $\alpha_k < 1$, i.e. there exists $\alpha_k' = \alpha_k/\beta$ such that

$$\hat{f}_{N_{max}}(x_k + \alpha_k' p_k) > \hat{f}_{N_{max}}(x_k) + \eta \alpha_k' p_k^T \nabla \hat{f}_{N_{max}}(x_k).$$

Furthermore, assumption A6 implies the Lipschitz continuity of the gradient $\nabla \hat{f}_{N_{max}}$ on $\{x \in \mathbb{R}^n | x = x_k + tp_k, t \in [0,1], k \geq \tilde{n}\}$. Therefore, there exists a constant $L > 0$ such that

$$\hat{f}_{N_{max}}(x_k + \alpha_k' p_k) \leq \frac{L}{2}(\alpha_k')^2 \|p_k\|^2 + \hat{f}_{N_{max}}(x_k) + \alpha_k' (\nabla \hat{f}_{N_{max}}(x_k))^T p_k$$

Combining the previous two inequalities and using the assumption A7 we obtain $\alpha_k \geq c_1 2\beta(1-\eta)/Lc_2^2$ and $dm_k(\alpha_k) \geq \|\nabla \hat{f}_{N_{max}}(x_k)\|^2 \; c_1^2 2\beta(1-\eta)/Lc_2^2$. Therefore, we conclude that for every $k \geq \tilde{n}$ inequality (24) holds with $\bar{\beta}_0 = \min\{c_1, \frac{c_1^2 2\beta(1-\eta)}{c_2^2 L}\}$. Moreover, A6 and A7 imply that (25) holds with $c_3 = 1 + Lc_2$. ■

The conditions for the global convergence are stated in the following two theorems. In the case of $\tilde{C}_k$ being defined by (8), the convergence results is stated in the following theorem. The proof is completely analogous to the corresponding one in [7], although the line search rule is stated with $\tilde{C}_k$. The same technique is used in the proof of Theorem 4.1 in [17]. Thus we omit the proof here.

Theorem 4.1. *Suppose that A1 - A4 hold and that the level set (20) is bounded. If $\tilde{C}_k$ is defined with (8), then there exists a subsequence of iterates $\{x_k\}_{k \in \mathbb{N}}$ that converges to a stationary point of $\hat{f}_{N_{max}}$. Moreover, if $\eta_{max} < 1$ then every accumulation point of $\{x_k\}_{k \in \mathbb{N}}$ is a stationary for $\hat{f}_{N_{max}}$.*

If $\tilde{C}_k$ is defined by (12), the statement and the proof are conceptually the same as Theorem 2.1 in [8] but with some technical differences caused by $\tilde{C}_k$ and $\varepsilon_k > 0$. For the sake of completeness we state the proof here.

Theorem 4.2. *Suppose that assumptions A1 - A3 and A6 - A7 are satisfied and that the level set (20) is bounded. If $\tilde{C}_k$ is defined by (12), then every accumulation point of the sequence $\{x_k\}_{k\in\mathbb{N}}$ generated by Algorithm 3 is stationary for $\hat{f}_{N_{max}}$.*

Proof. Under the assumptions of this theorem, Lemma 3.3 implies the existence of $\tilde{n}$ such that for every $k \geq \tilde{n}$ the sample size is $N_k = N_{max}$. Then, Lemma 3.2 implies that $\liminf_{k\to\infty} dm_k(\alpha_k) = 0$ and the subset K such that

$$\lim_{k\in K} dm_k(\alpha_k) = 0 \tag{26}$$

is defined as $K = \{v(k)-1\}_{k\in\mathbb{N}}$, where $v(k)$ is such that $\hat{f}_{N_{max}}(x_{v(k)}) = \tilde{C}_{s(k)}$, $\tilde{C}_{s(k)} = \max\{\hat{f}_{N_{max}}(x_{s(k)}), \ldots, \hat{f}_{N_{max}}(x_{s(k)-M+1})\}$ and $s(k) = \tilde{n} + kM$. Notice that $v(k) \in \{\tilde{n} + (k-1)M + 1, \ldots, \tilde{n} + kM\}$ and $v(k+1) \in \{\tilde{n} + kM + 1, \ldots, \tilde{n} + (k+1)M\}$. Therefore $v(k+1) - v(k) \leq 2M - 1$. This implies that for every $k \in \mathbb{N}$, $k \geq \tilde{n}$ there exists $\tilde{k} \geq k$, $\tilde{k} \in K$ such that

$$\tilde{k} - k \leq 2M - 2. \tag{27}$$

Let us define $k \geq q = \tilde{n} + M$. Lemma 4.1 implies that for every $k \geq q$ the inequality (24) holds which together with (26) implies

$$\lim_{k\in K} \|\nabla \hat{f}_{N_{max}}(x_k)\| = 0. \tag{28}$$

Also, (25) holds which together with (27) implies that for every $k \in \mathbb{N}$, $k > \tilde{n}$ there exists $\tilde{k} \geq k$, $\tilde{k} \in K$ such that $\|\nabla \hat{f}_{N_{max}}(x_k)\| \leq c_3^{2M-2} \|\nabla \hat{f}_{N_{max}}(x_{\tilde{k}})\|$. The previous inequality together with (28) yields $\lim_{k\to\infty} \|\nabla \hat{f}_{N_{max}}(x_k)\| = 0$. ■

After proving the global convergence result, we will analyze the convergence rate. Following the ideas from [32] and [8], we will prove that R-linear convergence for strongly convex functions can be obtained. Notice that the results presented in [7] do not include R-linear convergence rate and that the rule considered in [8] assumes $\varepsilon_k = 0$. Thus the results presented here generalize the existing theory. An additional assumption is needed.

A 8. *For every ξ, $F(\cdot, \xi)$ is a strongly convex function.*

A consequence of assumption A8 is that for every sample size N, $\hat{f}_N$ is a strongly convex function as well. Therefore, there exists $\gamma > 0$ such that for every N and every $x, y \in \mathbb{R}^n$

$$\hat{f}_N(x) \geq \hat{f}_N(y) + (\nabla \hat{f}_N(y))^T (x-y) + \frac{1}{2\gamma}\|x-y\|^2. \tag{29}$$

Furthermore, if x^* is the unique minimizer of $\hat{f}_N$ then

$$\frac{1}{2\gamma}\|x-x^*\|^2 \leq \hat{f}_N(x) - \hat{f}_N(x^*) \leq \gamma \|\nabla \hat{f}_N(x)\|^2 \tag{30}$$

As the objective function is convex we have that $\mathcal{L}$ is bounded and the iterative sequence is bounded. In order to prove R-linear convergence we impose an additional assumption on the sequence $\{\varepsilon_k\}_{k\in\mathbb{N}}$ which yields the technical result presented in Lemma 4.2.

A 9. *The sequence $\{\varepsilon_k\}_{k\in\mathbb{N}}$ is positive and converges to zero R-linearly.*

Lemma 4.2. *If the assumption A9 is satisfied, then for every $\theta \in (0,1)$ and $q \in \mathbb{N}$*

$$s_k = \sum_{j=1}^{k} \theta^{j-1} \varepsilon_{q+k-j}$$

converges to zero R-linearly.

Proof. Assumption A9 implies the existence of a constant $\rho \in (0,1)$ and a constant $C > 0$ such that $\varepsilon_k \leq C\rho^k$ for every $k \in \mathbb{N}$. Now, since $\rho, \theta \in (0,1)$, we can define $\gamma = \max\{\rho, \theta\} < 1$ such that for every $k \in \mathbb{N}$

$$s_k \leq \sum_{j=1}^{k} \theta^{j-1} C \rho^{q+k-j} \leq \sum_{j=1}^{k} C\gamma^{q+k-1} = C_1 a_k$$

where $C_1 = C\gamma^{q-1}$ and $a_k = k\gamma^k$. We will show that the sequence $\{a_k\}_{k\in\mathbb{N}}$ converges to zero R-linearly. Define a constant $s = (1+\gamma)/2\gamma$. Clearly $s > 1$. Furthermore, we define an additional sequence $\{c_k\}_{k\in\mathbb{N}}$ as follows

$$c_1 = \left(s^{(\ln s)^{-1}-1} \ln s\right)^{-1},$$

$$c_{k+1} = c_k \frac{ks}{k+1},\ k = 1, 2, \ldots,$$

and obviously $c_k = c_1 s^{k-1}/k$. In order to prove that $c_k \geq 1$ for every $k \in \mathbb{N}$, we define

$$f(x) = \frac{s^{x-1}}{x}$$

and search for its minimum on the interval $(0, \infty)$. As $f'(x) = s^{x-1}(x \ln s - 1)/x^2$ and $f'(x) = 0$ for $x^* = (\ln s)^{-1} > 0$, i.e. $x^* \ln s = 1$, and $f''(x^*) = \frac{s^{x^*} \ln s}{s x^{*2}} > 0$, x^* is the minimizer and for every $k \in \mathbb{N}$

$$\frac{s^{k-1}}{k} = f(k) \geq f(x^*) = s^{(\ln s)^{-1}-1} \ln s.$$

Therefore,

$$c_k = c_1 \frac{s^{k-1}}{k} \geq \left(s^{(\ln s)^{-1}-1} \ln s\right)^{-1} \left(s^{(\ln s)^{-1}-1} \ln s\right) = 1.$$

Now, let us define $b_k = a_k c_k$. Notice that $a_k \leq b_k$. Moreover,

$$b_{k+1} = a_{k+1} c_{k+1} = (k+1)\gamma^{k+1} c_k s \frac{k}{k+1} = s\gamma k \gamma^k c_k = t b_k$$

where $t = s\gamma = \frac{1+\gamma}{2} < 1$. So there exists a constant $B > 0$ such that $b_k \leq B t^{k-1}$. Finally, we obtain $s_k \leq C_1 a_k \leq C_1 b_k \leq C_2 B t^{k-1}$, and thus $\{s_k\}_{k\in\mathbb{N}}$ converges to zero R-linearly. ■

Next, we prove the R-linear convergence result for the sequence of iterates.

Theorem 4.3. *Suppose that the assumptions A1 - A4 and A6 - A9 are satisfied and that $\tilde{C}_k$ is defined by (12) or by (8) with $\eta_{max} < 1$. Then the sequence of iterates $\{x_k\}_{k\in\mathbb{N}}$ generated by Algorithm 3 with the line search (23) converges R-linearly to the unique minimizer x^* of function $\hat{f}_{N_{max}}$.*

Proof. First, notice that the assumptions of this theorem imply the existence of a finite number $\tilde{n}$ such that $N_k = N_{max}$ for every $k \geq \tilde{n}$. Moreover, it follows that there exists a finite integer $q \geq \tilde{n}$ such that for every $k \geq q$ iterate x_k belongs to the level set (20). Furthermore, strong convexity of the function $\hat{f}_{N_{max}}$ implies the boundedness and convexity of that level set. Therefore, there exists at least one accumulation point of the sequence $\{x_k\}_{k\in\mathbb{N}}$. Moreover, Theorems 4.1 and 4.2 imply that every accumulation point of that sequence is stationary for the function $\hat{f}_{N_{max}}$. On the other hand, strong convexity of the objective function implies that there is only one minimizer. Therefore, we conclude that $\lim_{k\to\infty} x_k = x^*$. Furthermore, according to Lemma 4.1, there are constants $c_3 = 1 + c_2 L$ and $\bar{\beta}_0 = \min\{c_1, \frac{c_1^2 2\beta(1-\eta)}{c_2^2 L}\}$ such that (24) and (25) hold for every $k > q$. Since (30) holds for $N = N_{max}$, it is sufficient to prove that $\hat{f}_{N_{max}}(x_k) - \hat{f}_{N_{max}}(x^*)$ converges to zero R-linearly.

Suppose that $\tilde{C}_k$ is defined by (8) with $\eta_{max} < 1$. Then, (17) is valid for every $k \geq q$ with $dm_k(\alpha_k) = -\alpha_k p_k^T \nabla \hat{f}_{N_{max}}(x_k)$. Moreover, Lemma 2.1 implies that $0 \leq Q_k \leq (1-\eta_{max})^{-1}$ for every k and therefore for every $k \geq q$

$$\tilde{C}_{k+1} \leq \tilde{C}_k + \varepsilon_k - \eta(1-\eta_{max}) dm_k(\alpha_k). \tag{31}$$

Subtracting $\hat{f}_{N_{max}}(x^*)$ from both sides and using (24) we obtain

$$\tilde{C}_{k+1} - \hat{f}_{N_{max}}(x^*) \leq \tilde{C}_k - \hat{f}_{N_{max}}(x^*) + \varepsilon_k - \bar{\beta}_1 \|\nabla \hat{f}_{N_{max}}(x_k)\|^2 \tag{32}$$

where $\bar{\beta}_1 = \eta(1-\eta_{max})\bar{\beta}_0$. Now, define $b = (\bar{\beta}_1 + \gamma(Lc_2+1)^2)^{-1}$. We distinguish two types of iterations for $k \geq q$.

If $\|\nabla \hat{f}_{N_{max}}(x_k)\|^2 < b(\tilde{C}_k - \hat{f}_{N_{max}}(x^*))$, inequalities (25) and (30) imply

$$\hat{f}_{N_{max}}(x_{k+1}) - \hat{f}_{N_{max}}(x^*) < \gamma(1+Lc_2)^2 b(\tilde{C}_k - \hat{f}_{N_{max}}(x^*)).$$

Setting $\theta_1 = \gamma(1+Lc_2)^2 b$ we obtain

$$\hat{f}_{N_{max}}(x_{k+1}) - \hat{f}_{N_{max}}(x^*) < \theta_1(\tilde{C}_k - \hat{f}_{N_{max}}(x^*)) \tag{33}$$

where $\theta_1 \in (0,1)$. If $\tilde{C}_{k+1} = \hat{f}_{N_{max}}(x_{k+1})$, then (33) obviously implies

$$\tilde{C}_{k+1} - \hat{f}_{N_{max}}(x^*) < \theta_1(\tilde{C}_k - \hat{f}_{N_{max}}(x^*)).$$

If $\tilde{C}_{k+1} = C_{k+1}$, then

$$\begin{aligned}
\tilde{C}_{k+1} - \hat{f}_{N_{max}}(x^*) &= C_{k+1} - \hat{f}_{N_{max}}(x^*) \\
&= \frac{\tilde{\eta}_k Q_k}{Q_{k+1}} C_k + \frac{\hat{f}_{N_{max}}(x_{k+1})}{Q_{k+1}} - \frac{\tilde{\eta}_k Q_k + 1}{Q_{k+1}} \hat{f}_{N_{max}}(x^*) \\
&\leq \frac{\tilde{\eta}_k Q_k}{Q_{k+1}} (\tilde{C}_k - \hat{f}_{N_{max}}(x^*)) + \frac{\hat{f}_{N_{max}}(x_{k+1}) - \hat{f}_{N_{max}}(x^*)}{Q_{k+1}} \\
&\leq (1 - \frac{1}{Q_{k+1}})(\tilde{C}_k - \hat{f}_{N_{max}}(x^*)) + \frac{\theta_1(\tilde{C}_k - \hat{f}_{N_{max}}(x^*))}{Q_{k+1}} \\
&= (1 - \frac{1-\theta_1}{Q_{k+1}})(\tilde{C}_k - \hat{f}_{N_{max}}(x^*)) \\
&\leq (1 - (1-\eta_{max})(1-\theta_1))(\tilde{C}_k - \hat{f}_{N_{max}}(x^*)).
\end{aligned}$$

In the last inequality, we used the fact that $Q_{k+1} \leq (1-\eta_{max})^{-1}$. Therefore, we conclude that

$$\tilde{C}_{k+1} - \hat{f}_{N_{max}}(x^*) \leq \bar{\theta}_1(\tilde{C}_k - \hat{f}_{N_{max}}(x^*)) \tag{34}$$

where $\bar{\theta}_1 = \max\{\theta_1, 1-(1-\eta_{max})(1-\theta_1)\} \in (0,1)$.

On the other hand, if $\|\nabla \hat{f}_{N_{max}}(x_k)\|^2 \geq b(\tilde{C}_k - \hat{f}_{N_{max}}(x^*))$, inequality (32) implies $\tilde{C}_{k+1} - \hat{f}_{N_{max}}(x^*) \leq \bar{\theta}_2(\tilde{C}_k - \hat{f}_{N_{max}}(x^*)) + \varepsilon_k$ where $\bar{\theta}_2 = 1 - b\bar{\beta}_1$ and therefore $\bar{\theta}_2 \in (0,1)$. So, for every $k \in \mathbb{N}_0$

$$\tilde{C}_{q+k+1} - \hat{f}_{N_{max}}(x^*) \leq \theta(\tilde{C}_{q+k} - \hat{f}_{N_{max}}(x^*)) + \varepsilon_{q+k}$$

where $\theta = \max\{\bar{\theta}_1, \bar{\theta}_2\} \in (0,1)$. By the induction argument we obtain that for every $k \in \mathbb{N}$

$$\tilde{C}_{q+k} - \hat{f}_{N_{max}}(x^*) \leq \theta^k(\tilde{C}_q - \hat{f}_{N_{max}}(x^*)) + \sum_{j=1}^{k} \theta^{j-1}\varepsilon_{q+k-j}.$$

Finally, recalling that $\hat{f}_{N_k}(x_k) \leq \tilde{C}_k$, we obtain

$$\hat{f}_{N_{max}}(x_{q+k}) - \hat{f}_{N_{max}}(x^*) \leq \theta^k(\tilde{C}_q - \hat{f}_{N_{max}}(x^*)) + \sum_{j=1}^{k} \theta^{j-1}\varepsilon_{q+k-j}.$$

which together with Lemma 4.2 implies the existence of $\theta_3 \in (0,1)$ and $M_q > 0$ such that for every $k \in \mathbb{N}$

$$\|x_{q+k} - x^*\| \leq \theta_3^k M_q.$$

Now, suppose that $\tilde{C}_k$ is defined by (12). Then, for every $k \in \mathbb{N}$

$$\tilde{C}_{s(k+1)} \leq \tilde{C}_{s(k)} + \sum_{i=0}^{M-1} \varepsilon_{s(k)+i} - \eta d m_{v(k+1)-1} \tag{35}$$

where $s(k) = \tilde{n} + kM$ and $\hat{f}_{N_{max}}(x_{v(k)}) = \tilde{C}_{s(k)}$. Together with (24), the previous inequality implies

$$\begin{aligned} \tilde{C}_{s(k+1)} - \hat{f}_{N_{max}}(x^*) &\leq \tilde{C}_{s(k)} - \hat{f}_{N_{max}}(x^*) - \eta\bar{\beta}_0\|\nabla \hat{f}_{N_{max}}(x_{v(k+1)-1})\|^2 \\ &+ \sum_{i=0}^{M-1} \varepsilon_{s(k)+i}. \end{aligned}$$

Define $b = (\bar{\beta}_0 + \gamma c_3^2)^{-1}$. If $\|\nabla \hat{f}_{N_{max}}(x_{v(k+1)-1})\|^2 \geq b(\tilde{C}_{s(k)} - \hat{f}_{N_{max}}(x^*))$, for $\theta_1 = (1 - \eta\bar{\beta}_0 b) \in (0,1)$ we obtain

$$\tilde{C}_{s(k+1)} - \hat{f}_{N_{max}}(x^*) \leq \theta_1(\tilde{C}_{s(k)} - \hat{f}_{N_{max}}(x^*)) + \sum_{i=0}^{M-1} \varepsilon_{s(k)+i}.$$

On the other hand, if $\|\nabla \hat{f}_{N_{max}}(x_{v(k+1)-1})\|^2 < b(\tilde{C}_{s(k)} - \hat{f}_{N_{max}}(x^*))$ then (30) and $\tilde{C}_{s(k+1)} = \hat{f}_{N_{max}}(x_{v(k+1)})$ imply

$$\tilde{C}_{s(k+1)} - \hat{f}_{N_{max}}(x^*) \leq \gamma c_3^2 \|\nabla \hat{f}_{N_{max}}(x_{v(k+1)-1})\|^2 < \theta_2(\tilde{C}_{s(k)} - \hat{f}_{N_{max}}(x^*))$$

where $\theta_2 = \gamma c_3^2 b \in (0,1)$. Therefore, for every $k \in \mathbb{N}$

$$\tilde{C}_{s(k+1)} - \hat{f}_{N_{max}}(x^*) \leq \theta(\tilde{C}_{s(k)} - \hat{f}_{N_{max}}(x^*)) + \sum_{i=0}^{M-1} \varepsilon_{s(k)+i}$$

where $\theta = \max\{\theta_1, \theta_2\} \in (0,1)$. Using the induction argument, we obtain

$$\tilde{C}_{s(k+1)} - \hat{f}_{N_{max}}(x^*) \leq \theta^k(\tilde{C}_{s(1)} - \hat{f}_{N_{max}}(x^*)) + \sum_{j=1}^{k} \sum_{i=0}^{M-1} \theta^{j-1} \varepsilon_{s(k+1-j)+i}$$

Moreover, $\hat{f}_{N_{max}}(x_{s(k)+j}) \leq \tilde{C}_{s(k+1)}$ holds for every $j \in \{1, \ldots, M\}$ and every $k \in \mathbb{N}$ and therefore

$$\hat{f}_{N_{max}}(x_{s(k)+j}) - \hat{f}_{N_{max}}(x^*) \leq \theta^k V + r_k \tag{36}$$

where $V = \tilde{C}_{s(1)} - \hat{f}_{N_{max}}(x^*) \geq 0$ and $r_k = \sum_{j=1}^{k} \sum_{i=0}^{M-1} \theta^{j-1} \varepsilon_{s(k+1-j)+i}$. Now, assumption A9 implies the existence of $\rho \in (0,1)$ and $C > 0$ such that $\varepsilon_k \leq C\rho^k$ for every k. Defining $C_1 = MC\rho^{\tilde{n}}$ and $\gamma_1 = \max\{\rho^M, \theta\}$ we obtain $\gamma_1 < 1$ and

$$\begin{aligned} r_k &\leq \sum_{j=1}^{k} \sum_{i=0}^{M-1} \theta^{j-1} C \rho^{s(k+1-j)+i} \leq MC \sum_{j=1}^{k} \theta^{j-1} \left(\rho^M\right)^{(k+1-j)} \rho^{\tilde{n}} \\ &\leq MC\rho^{\tilde{n}} \sum_{j=1}^{k} \gamma_1^{j-1} \gamma_1^{k+1-j} = C_1 \sum_{j=1}^{k} \gamma_1^k = C_1 k \gamma_1^k. \end{aligned}$$

Following the ideas from the proof of Lemma 4.2, we conclude that r_k converges R-linearly and therefore there exist $D > 0$ and $\bar{\theta} = \max\{\theta, t\} \in (0,1)$ such that

$$\hat{f}_{N_{max}}(x_{s(k)+j}) - \hat{f}_{N_{max}}(x^*) \leq \bar{\theta}^k D.$$

The previous inequality and (30) imply the existence of $\theta_3 \in (0,1)$ and $M_h > 0$ such that for every $k \in \mathbb{N}$ and $j \in \{1, \ldots, M\}$

$$\|x_{\tilde{n}+kM+j} - x^*\| \leq \theta_3^k M_h$$

or equivalently for every $j \in \{1, \ldots, M\}$ and every $s \in \mathbb{N}, s \geq M$

$$\|x_{\tilde{n}+s} - x^*\| \leq \theta_3^{\frac{s-j}{M}} M_h \leq \theta_3^{\frac{s}{M}-1} M_h = \theta_4^s M_m.$$

where $\theta_4 = \theta_3^{\frac{1}{M}} \in (0,1)$ and $M_m = \frac{M_h}{\theta_3} > 0$ which completes the proof. ■

If the rate of convergence in the above theorem is compared with the precise rates known for gradient methods, like those presented in [22], several differences can be commented. First of all, the rate of descent gradient methods with monotone line search is q-linear. The R-linear obtained here is a natural consequence of the nonmonotonicity and variable sample scheme. All estimates in Theorem 4.3 are true only for the iterates with $N_k = N_{\max}$. The nonmonotonicity implies R-linear convergence even with $\varepsilon_k = 0$, see [8, 32]. Thus the additional freedom in the step length selection obtained with adding ε_k is not causing a decrease in the convergence rate providing that $\{\varepsilon_k\}$ converges R-linearly. However, the upper bounds obtained in [22] for the monotone gradient methods are more precise than the bounds obtained here. For example, Theorem 2.1.15 in [22] states that

$$\|x_k - x^*\| \leq \left(\frac{Q_f - l}{Q_f + l}\right)^k \|x_0 - x^*\|$$

for the monotone gradient method with constant step size $\alpha_k = h \in (0, 2/(\gamma + L))$, where $Q_f = L/\gamma$ and $l = 1/2\gamma$. The bounds in Theorem 4.3 are expressed in terms of constants that depend on γ, L as well, but also on the rate of convergence of $\{\varepsilon_k\}$, the back-tracking parameter β, $\eta_{\max}$ or M, and the iteration in which the scheduling sequence becomes stationary with $N_k = N_{\max}$. Thus, a more precise estimation like the one cited above is not likely to be obtained for the nonmonotone methods with variable sample scheme.

5 Numerical results

In this section we apply Algorithm 3 with the safeguard proposed in Algorithm 2 and compare six different line search methods with different search directions. The advantages of the variable sample scheme with respect to the classical SAA methods as well as with respect to some heuristic schedule updates are demonstrated in [17]. Thus our main interest here is to compare different line search rules, in particular monotone versus nonmonotone ones, for solving the SAA problem. In the first subsection, we consider a set of deterministic problems which are transformed to include the noise. The second subsection is devoted to a problem with real data. The data is collected from a survey that examines the influence of the various factors on the metacognition and the feeling of knowing of the students. The total sample size is 746. Linear regression is used as the model and the least squares problem is considered. This is the form of the objective function which is considered in [12] and therefore we compare Algorithm 3 with the scheme proposed in that paper.

Algorithm 3 is implemented with the stopping criterion $\|g_k^{N_{max}}\| \leq 0.1$ where $g_k^{N_{max}}$ is an approximation or the true gradient of the function $\hat{f}_{N_{max}}$. The maximal sample size is $N_{max} = 100$ for the first set of problems and the initial sample size is $N_0 = 3$. Alternatively, the algorithm terminates if 10^7 function evaluations is exceeded. When the gradient is used each of its components is counted as one function evaluation. In the first subsection, the results are obtained from eight replications of each algorithm and the average values are reported. All the algorithms use the backtracking technique with $\beta = 0.5$. The parameters from Algorithm 1 are $\nu_1 = 0.1$ and $d = 0.5$. The confidence level is $\delta = 0.95$ which yields the lack of precision parameter $\alpha_\delta = 1.96$.

We list the line search rules as follows. The rules where the parameter $\tilde{\eta}_k = 0.85$ is given refer to $\tilde{C}_k$ defined by (8), while $M = 10$ determines the rule with $\tilde{C}_k$ defined by (12). The choice for this parameters is motivated by the results in [32] and [8]. We denote the approximation of the gradient $\nabla \hat{f}_{N_k}(x_k)$ by g_k. When the gradient is available, $g_k = \nabla \hat{f}_{N_k}(x_k)$.

(B1) $\hat{f}_{N_k}(x_k + \alpha_k p_k) \leq \hat{f}_{N_k}(x_k) + \eta \alpha_k p_k^T g_k$

(B2) $\hat{f}_{N_k}(x_k + \alpha_k p_k) \leq \hat{f}_{N_k}(x_k) + \varepsilon_k - \alpha_k^2 \beta_k$

(B3) $\hat{f}_{N_k}(x_k + \alpha_k p_k) \leq \tilde{C}_k + \varepsilon_k - \alpha_k^2 \beta_k, \quad \tilde{\eta}_k = 0.85$

(B4) $\hat{f}_{N_k}(x_k + \alpha_k p_k) \leq \tilde{C}_k + \eta \alpha_k p_k^T g_k, \quad M = 10$

(B5) $\hat{f}_{N_k}(x_k + \alpha_k p_k) \leq \tilde{C}_k + \varepsilon_k - \alpha_k^2 \beta_k, \quad M = 10$

(B6) $\hat{f}_{N_k}(x_k + \alpha_k p_k) \leq \tilde{C}_k + \eta \alpha_k p_k^T g_k, \quad \tilde{\eta}_k = 0.85$

The rules B1, B4 and B6 assume the descent search directions and the parameter η is set to 10^{-4}. The initial member of the sequence which makes the nondescent directions acceptable is defined by $\varepsilon_0 = \max\{1, |\hat{f}_{N_0}(x_0)|\}$ while the rest of it is updated by $\varepsilon_k = \varepsilon_0 k^{-1.1}$ but only if the sample size does not change, i.e. if $N_{k-1} = N_k$. Otherwise, $\varepsilon_k = \varepsilon_{k-1}$.

The search directions are of the form

$$p_k = -H_k g_k.$$

We make 4 different choices for the matrix H_k and obtain the following directions, [29].

(NG) The negative gradient direction is obtained by setting $H_k = I$ where I represents the identity matrix.

(BFGS) This direction is obtained by using the BFGS formula for updating the inverse Hessian

$$H_{k+1} = (I - \frac{1}{y_k^T s_k} s_k y_k^T) H_k (I - \frac{1}{y_k^T s_k} y_k s_k^T) + \frac{1}{y_k^T s_k} s_k s_k^T$$

where $y_k = g_{k+1} - g_k$, $s_k = x_{k+1} - x_k$ and $H_0 = I$.

(SG) The spectral gradient direction is defined by setting $H_k = \gamma_k I$ where

$$\gamma_k = \frac{\|s_{k-1}\|^2}{s_{k-1}^T y_{k-1}}$$

(SR1) The symmetric rank-one direction is defined by $H_0 = I$ and

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)(s_k - H_k y_k)^T}{(s_k - H_k y_k)^T y_k}.$$

If the gradient is available, the negative gradient is descent direction. Moreover, BFGS and SG implementations also ensure the descent search direction. Furthermore, we define $\beta_k = |g_k^T H_k g_k|$ where H_k is one of the matrices defined above.

We also tested the algorithm with the following gradient approximations. FD stands for the centered finite difference estimator while FuN represents the simultaneous perturbations approximation that allows the standard normal distribution for the perturbation sequence [13].

(FD) For $i = 1, 2, \ldots, n$

$$(g_k)_i = \frac{\hat{f}_{N_k}(x_k + he_i) - \hat{f}_{N_k}(x_k - he_i)}{2h},$$

where e_i is the ith column of the identity matrix and $h = 10^{-4}$.

(FuN) For $i = 1, 2, \ldots, n$

$$(g_k)_i = \frac{\hat{f}_{N_k}(x_k + h\Delta_k) - \hat{f}_{N_k}(x_k - h\Delta_k)}{2h} \Delta_{k,i},$$

where $h = 10^{-4}$ and random vector $\Delta_k = (\Delta_{k,1}, ..., \Delta_{k,n})^T$ follows the multivariate standard normal distribution.

The criterion for comparing the algorithms is the number of function evaluations.

5.1 Noisy problems

We use 7 test functions from the Moré test collection available at the web page [33]: Freudenstein and Roth, Jennrich and Sampson, Biggs EXP6, Osborne II, Trigonometric, Broyden Tridiagonal and Broyden Banded. They are converted into noisy problems in two ways. The first one is by adding the noise, and the second one involves multiplication by random vector which then affects the gradient as well. The noise is represented by the random vector ξ with the Normal distribution $\mathcal{N}(0,1)$. If we denote the deterministic test function by $h(x)$, we obtain the objective functions $f(x) = E(F(x,\xi))$ by the following modifications:

(N1) $F(x,\xi) = h(x) + \xi$

(N2) $F(x, \xi) = h(x) + \|\xi x\|^2$.

These modifications yield 14 test problems. The average number of function evaluations in 8 replications is used as the main criterion for comparison. Let us denote the average by ϕ_i^j where i represents the method (determined by the line search and the search direction) and j represents the problem. We define the efficiency index as in [18], i.e. for the method i the efficiency index is

$$\omega_i = \frac{1}{14} \sum_{j=1}^{14} \frac{\min_i \phi_i^j}{\phi_i^j}.$$

We also report the level of nonmonotonicity. If the number of iterations is k and s is the number of iterations at which the accepted step size would not be accepted if the line search rule was B1, then we define the nonmonotonicity index by

$$\mu = \frac{s}{k}.$$

The numbers in the following two tables refer to the average values of 8 independent runs. Table 1 represents the results obtained by applying the methods with the gradient, while the subsequent table refers to the gradient approximation approach. The SR1 method is not tested with the line search rules which assume the descent search directions and therefore the efficiency index is omitted in that cases. The same is true for the nonmonotonicity. For the same reason we omit the line search rules B1, B4 and B6 in Table 2.

	Efficiency index (ω)				Nonmonotonicity index (μ)			
	NG	SG	BFGS	SR1	NG	SG	BFGS	SR1
B1	0.2471	0.3975	0.5705	/	0.0000	0.0000	0.0000	/
B2	0.0774	0.4780	0.5474	0.4750	0.4835	0.2081	0.1616	0.2541
B3	0.0783	0.4927	0.5306	0.4401	0.4426	0.2083	0.1708	0.2810
B4	0.0620	0.6468	0.4200	/	0.4070	0.1049	0.0998	/
B5	0.0798	0.5157	0.5043	0.4725	0.4060	0.1998	0.1722	0.2593
B6	0.1064	0.6461	0.4690	/	0.3430	0.1050	0.0944	/

Table 1: The gradient-based methods

	Efficiency index (ω)			
	SG-FD	SG-FuN	BFGS-FD	SR1-FD
B2	0.6832	0.4536	0.7316	0.6995
B3	0.6957	0.4164	0.7149	0.6576
B5	0.7255	0.4286	0.6808	0.7156
	Nonmonotonicity index (μ)			
	SG-FD	SG-FuN	BFGS-FD	SR1-FD
B2	0.1693	0.1008	0.1349	0.2277
B3	0.1682	0.1166	0.1449	0.2516
B5	0.1712	0.1248	0.1453	0.2410

Table 2: The gradient-free methods

Among the 21 tested methods presented in Table 1, the efficiency index suggests that the best one is the spectral gradient method combined with the line search rule B4. However, we can see that the results also suggest that the negative gradient and the BFGS search direction should be combined with the monotone line search rule B1. The SR1 method works slightly better with the line search B2 than with B5 and we can say that it is more efficient with lower levels of nonmonotonicity. Looking at the SG method, we can conclude that large nonmonotonicity was not beneficial for that method either. In fact, B4 has the lowest nonmonotonicity if we exclude B1.

The results considering the spectral gradient method are consistent with the deterministic case because it is known that the monotone line search can inhibit the benefits of scaling the negative gradient direction. However, these testings suggest that allowing too much nonmonotonicity can deteriorate the performance of the algorithms.

The results from Table 2 imply that B5 is the best choice if we consider the spectral gradient or SR1 method with the finite difference gradient approximation. Furthermore, the finite difference approximation for the BFGS direction achieves the best performance when combined with B2. This line search is the best choice for simultaneous perturbation approach as well. However, the simultaneous perturbation approximation of the gradient provided the least preferable results in general. The reason is probably the fact that the simultaneous perturbation provided rather poor approximations of the gradient in our test examples as the number of iterations is not very large and the asymptotic features of that approach could not develop.

The number of iterations where a decrease of sample size is proposed in Algorithm 1 varies across the methods and problems and it goes up to 46% of iterations. The safeguard rule defined in Algorithm 2 prevents some of these decreases so the average number of iterations with decrease of the sample size for BFGS and SR1 is approximately 6% and for NG it is around 7%. The corresponding number for SG method is 9% and it goes down to 4% when FuN approximation of the gradient is used. The following example of the sample size sequence obtained by a single run of the SG with the line search B4 for Biggs EXP6 problem is obtained: $(N_0, \ldots, N_{17}) = (3, 100, 7, 3, 100, 15, 3, 3, 100, 100, 100, 100, 100, 100, 56, 100, 100, 100)$.

The modification N1 is supposed to be suitable for examining the convergence towards local versus global optimizers. However, the numerical results we obtained are not conclusive in that sense and we list here only the results regarding particular cases which are contrary to the common belief that nonmonotonicity implies more frequent convergence to global optimizers. In the Freudenstein and Roth problem for example, B1 converges to the global minimum in all 8 replications, B6 converges to the global minimum only once while the other methods are trapped at the local solutions. Furthermore, in the Broyden Banded problem, B4 and B6 are carried away from the global solution, while the other methods converge towards it. The case where the noise affects the gradient is harder for tracking global optimizers. However, the SG method with the line searches that allow only the descent directions (B1, B4 and B6) converges to the point with the lower function value when the Broyden Tridiagonal problem is concerned. Furthermore, in the Osborne II problem the SG with the Armijo line search B1 provided the lowest function value.

The efficiency index yields similar conclusions as the performance profile analysis [11]. At the end of this subsection, we show the performance profiles for the best methods on this particular test collection: SG in the gradient-based case (Figure 1) and BFGS-FD in the gradient-free case (Figure 2). The first graphic in both figures provides the results when the problems of the form (N1) are considered, the second one refers to the problems (N2) while the third one gathers all 14 problems together.

Figure 1 shows that B4 clearly outperforms all the other line search rules in (N1) case, while in (N2) case B6 is highly competitive. If we take a look at all of the considered problems together, B4 is clearly the best choice. In the BFGS-FD case, B2 and B3 seem to work better than B5 with B2 being the better one in the cases where the noise affects the search direction, i.e. when (N2) formulation is considered. Clearly, all the conclusions we have presented here are influenced by the test examples we consider.

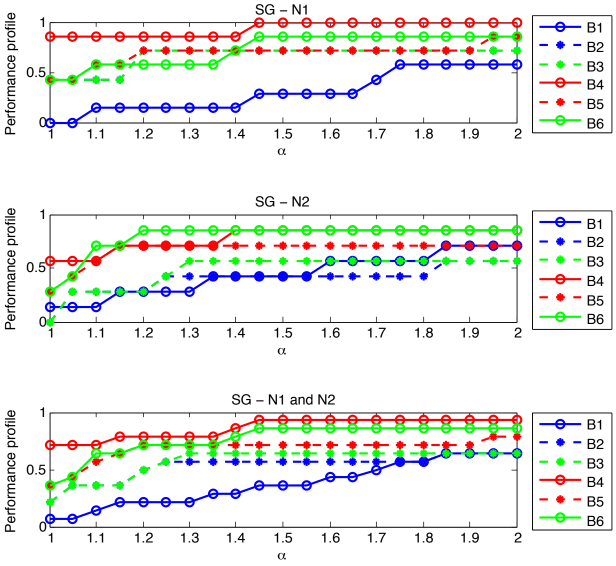

Figure 1: The SG methods in noisy environment

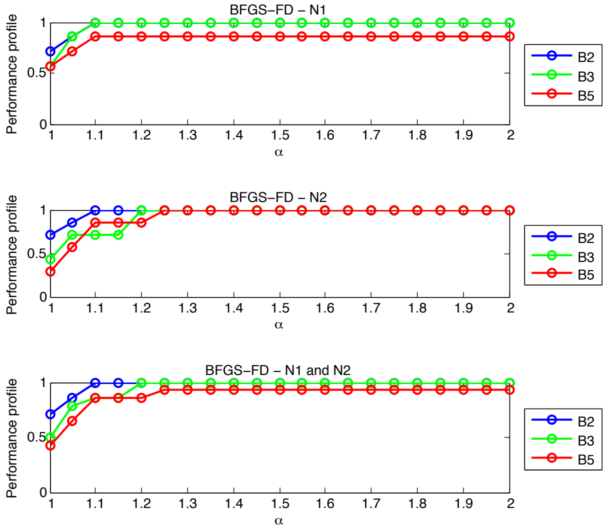

Figure 2: The BFGS-FD methods in noisy environment

5.2 Application to the least squares problems

As we already mentioned, this subsection is devoted to the real data problem. The data comes from a survey that was conducted among 746 students in Serbia. The goal of this survey was to determine how different factors affect the feeling of knowing (FOK) and metacognition (META) of the students. We will not go into further details of this survey since our aim is only to compare different algorithms. Therefore, we only present the number of function evaluations (ϕ) and nonmonotonicity index (μ) defined above.

Linear regression is used as the model and the parameters are searched for throughout the least squares problem. Therefore, we obtain two problems of the form $\min_{x \in \mathbb{R}^n} \hat{f}_N(x)$ where

$$\hat{f}_N(x) = \frac{1}{N} \sum_{i=1}^{N} (x^T a_i - y_i)^2.$$

The sample size is $N = N_{max} = 746$ and the number of factors examined is $n = 4$. Vectors a_i, $i = 1, 2, \ldots, 746$ represent the factors and y_i, $i = 1, 2, \ldots, 746$ represent the FOK or the META results obtained from the survey.

The same type of problem is considered in [12]. Therefore, the variable sample size scheme proposed in this paper is compared with the dynamics of increasing the sample size that is proposed in [12] (Heuristic). We state the results in Table 3 and Table 4. Heuristic assumes that the sample size increases by the rule $N_{k+1} = \lceil \min\{1.1 N_k, N_{max}\} \rceil$. Since the gradients are easy to obtain the gradient-based approach is a good choice and we use the spectral gradient method with the different line search rules. Algorithm 3 is used with the same parameters like in the previous subsection and the stopping criterion $\|g_k^{N_{max}}\| \leq 10^{-2}$.

	Algorithm 3		Heuristic	
SG	ϕ	μ	ϕ	μ
B1	9.4802E+04	0.0000	1.2525E+05	0.0000
B2	5.3009E+04	0.2105	6.0545E+04	0.2105
B3	5.3009E+04	0.2105	6.0545E+04	0.2105
B4	4.4841E+04	0.1765	9.4310E+04	0.2121
B5	5.3009E+04	0.2105	7.1844E+04	0.1967
B6	4.5587E+04	0.1176	1.1178E+05	0.1343

Table 3: The FOK analysis results, ϕ - efficiency index, μ - nonmonotonicity index

	Algorithm 3		Heuristic	
SG	ϕ	μ	ϕ	μ
B1	1.6716E+05	0.0000	2.1777E+05	0.0000
B2	3.3606E+04	0.0909	6.2159E+04	0.2632
B3	3.3606E+04	0.0909	6.1408E+04	0.1897
B4	3.8852E+04	0.1538	6.6021E+04	0.1607
B5	3.3606E+04	0.0909	6.1408E+04	0.1897
B6	3.8852E+04	0.1538	1.4953E+05	0.1053

Table 4: The META analysis results, ϕ - efficiency index, μ - nonmonotonicity index

First of all notice that the Algorithm 3 performs better than Heuristic in all cases. Also, the monotone line search B1 performs the worst in both problems and both presented algorithms. When the FOK problem is considered, the best results are obtained with the line search B4 applied within Algorithm 3, although B6 is highly competitive in that case. Both of the mentioned line search rules have modest but strictly positive nonmonotonicity coefficients. However, when the Heuristic is applied, the additional term ε_k turns out to be quite useful since the best performance is obtained by B2 and B3.

While the analysis of FOK provides the results similar to the ones in the previous subsection, the META yields rather different conclusions. In that case, the lowest number of function evaluations is achieved by the line search rules B2, B3 and B5. However, the results are not that different because the level of nonmonotonicity for those methods is not too large. Similar results are obtained for Heuristic where B3 and B5 were the best with the medium level of nonmonotonicity.

6 Conclusions

The methods presented in the paper combine nonmonotone line search rules with a variable sample size strategy and aim to cost efficient algorithms for solving the SAA problem. At each iteration a new sample size is chosen in accordance with the progress made in decreasing the objective function and the precision measured by an approximate width of the confidence interval. The intuitive motivation is the fact that nonmonomotonicity appears naturally in stochastic environment and nonmonotone rules provide more freedom in choosing the search direction as well as the step size.

The convergence results that are obtained include global convergence statements for the two dominant nonmonotone line search strategies as well as an analysis of the rate of convergence. It is shown that R-linear convergence could be achieved with nonmonotone rules if the search directions are descent.

The influence of nonmontonicity is tested numerically using four different search directions and six different line search rules. The obtained results clearly favor the nonmonotone rules. One important conjecture is that the level of nonmonotonicity should not be too hight i.e. the best results are obtained with the nonmonotone rules that allow relatively modest number of iterations with large step sizes or nondescent directions. On the other hand it appears that the strictly decreasing directions, like BFGS, combine slightly better with the Armijo type monotone rule. The SG direction appears to be more efficient with some level of nonmonotonicity, which is the same behavior as in deterministic problems. The best results we obtained are with the SG direction embedded in B4 if the gradients are available. The possibility of nonmonotone strategy appears to be particularly important if the gradients are not available and one is using finite difference gradient approximations. The gradient approximations by finite differences seems to work equally well regardless of the the choice of second order direction. All our conclusions are clearly influenced by the test collection. The academic set of test examples is chosen such that the objective functions posses local and global minimizers. The ideas was to investigate if the nonmonotone strategies would force the convergence towards global minimizers, following the common belief for deterministic problems. Contrary to our expectations, the experiments did not confirm this belief and further research is needed in this direction. In fact, as one of the referees pointed out, a broad comparison of nonmonotone rules for either deterministic or stochastic problems has never been done and it would be an interesting and useful result. Future research direction we plan to pursue is an extension of the variable sample size strategy presented here to constrained problems.

Acknowledgment We are grateful to the anonymous referees whose constructive remarks helped us to improve this paper.

References

[1] F. Bastin, Trust-Region Algorithms for Nonlinear Stochastic Programming and Mixed Logit Models, *PhD thesis, University of Namur, Belgium, 2004.*

[2] F. Bastin, C. Cirillo, P. L. Toint, An adaptive Monte Carlo algorithm for computing mixed logit estimators, *Computational Management Science 3(1), (2006), pp. 55-79.*

[3] F. Bastin, C. Cirillo, P. L. Toint, Convergence theory for nonconvex stochastic programming with an application to mixed logit, *Math. Program., Ser. B 108 (2006) pp. 207-234.*

[4] E. G. Birgin, N. Krejić, J. M. Martínez, Globaly convergent inexact quasi-Newton methods for solving nonlinear systems, *Numer. Algorithms 32 (2003) pp. 249-260.*

[5] R. Byrd, G. Chin, W. Neveitt, J. Nocedal On the Use of Stochastic Hessian Information in Optimization Methods for Machine Learning, *SIAM J. on Optimization, vol 21, issue 3 (2011), pp. 977-995.*

[6] R. Byrd, G. Chin, J. Nocedal, Y. Wu, Sample Size Selection in Optimization Methods for Machine Learning, *Mathematical Programming Vol. 134, Issue 1 (2012), pp. 127-155.*

[7] W. Cheng, D.H. Li, A derivative-free nonmonotone line search and its applications to the spectral residual method, *IMA Journal of Numerical Analysis 29 (2008) pp. 814-825*

[8] Y.H. Dai, On the nonmonotone line search, *J. Optim. Theory Appl., 112 (2002), pp. 315-330.*

[9] G. Deng, M. C. Ferris, Variable-Number Sample Path Optimization, *Mathematical Programming Vol. 117, No. 1-2 (2009) pp. 81-109.*

[10] M.A. Diniz-Ehrhardt, J. M. Martínez, M. Raydan, A derivative-free nonmonotone line-search technique for unconstrained optimization, *Journal of Computational and Applied Mathematics Vol. 219, Issue 2 (2008) pp. 383-397.*

[11] E. D. Dolan, J. J. Moré, Benchmarking optimization software with performance profiles, *Math. Program., Ser. A 91: 201-213 (2002), pp. 201-213*

[12] M. P. Friedlander, M. Schmidt, Hybrid deterministic-stochastic methods for data fitting, *SIAM J. Scientific Computing 34 No. 3 (2012), pp. 1380-1405.*

[13] M. C. Fu, Gradient Estimation, *S.G. Henderson and B.L. Nelson (Eds.), Handbook in OR & MS Vol. 13 (2006), pp. 575-616.*

[14] L. Grippo, F. Lampariello, S. Lucidi, A nononotone line search technique for Newton's method, *SIAM J. Numerical Analysis Vol. 23, No. 4 (1986), pp. 707-716.*

[15] L. Grippo, F. Lampariello, S. Lucidi, A class of nonmonotone stabilization methods in unconstrained optimization, *Numer. Math. 59 (1991), pp. 779-805.*

[16] T. Homem-de-Mello, Variable-Sample Methods for Stochastic Optimization, *ACM Transactions on Modeling and Computer Simulation Vol. 13, Issue 2 (2003), pp. 108-133.*

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

[18] N. Krejić, S. Rapajić, Globally convergent Jacobian smoothing inexact Newton methods for NCP, *Computational Optimization and Applications Vol. 41, Issue 2 (2008), pp. 243-261.*

[19] W. La Cruz, J. M. Martínez, M. Raydan, Spectral residual method without gradient information for solving large-scale nonlinear systems of equations, *Math. Comput. 75 (2006), pp. 1429-1448.*

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

[21] D. J. Lizotte, R. Greiner, D. Schuurmans, An experimental methodology for response surface optimization methods, *Journal of Global Optimization Vol. 53, Issue 4 (2012), pp. 699-736.*

[22] Y. Nesterov, Introductory Lectures on Convex Optimization, Kluwer Academic Publishers, 2004.

[23] J. Nocedal, S. J. Wright, Numerical Optimization, *Springer, 1999.*

[24] R. Pasupathy, On Choosing Parameters in Retrospective-Approximation Algorithms for Stochastic Root Finding and Simulation Optimization, *Operations Research Vol. 58, No. 4 (2010), pp. 889-901.*

[25] E. Polak, J. O. Royset, Eficient sample sizes in stochastic nonlinear programing, *Journal of Computational and Applied Mathematics Vol. 217, Issue 2 (2008), pp. 301-310.*

[26] M. Raydan, The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem, *SIAM J. Optimization 7 (1997), pp. 26-33.*

[27] J. O. Royset, Optimality functions in stochastic programming, *Math. Program. Vol. 135, Issue 1-2 (2012), pp. 293-321.*

[28] A. Shapiro, A. Ruszczynski, Stochastic Programming, *Vol. 10 of Handbooks in Operational Research and Management Science. Elsevier, 2003, pp. 353-425.*

[29] J. C. Spall, Introduction to Stochastic Search and Optimization, *Wiley-Interscience serises in discrete mathematics, New Jersey, 2003.*

[30] R. Tavakoli, H. Zhang, A nonmonotone spectral projected gradient method for large-scale topology optimization problems, *Numerical Algebra, Control and Optimization Vol. 2, No. 2 (2012), pp. 395-412.*

[31] P. L. Toint, An assessment of nonmonotone line search techniques for unconstrained optimization, *SIAM J. Scientific Computing Vol. 17, No. 3 (1996), pp. 725-739.*

[32] H. Zhang, W. W. Hager, A nonmonotone line search technique and its application to unconstrained optimization *SIAM J. Optim. 4 (2004), pp. 1043-1056.*

[33] http://www.uni-graz.at/imawww/kuntsevich/solvopt/results/moreset.html.