Spectral projected gradient method for stochastic optimization

Nataša Krejić[*] Nataša Krklec Jerinkić [*]

June 15, 2018

Abstract

We consider the Spectral Projected Gradient method for solving constrained optimization problems with the objective function in the form of mathematical expectation. It is assumed that the feasible set is convex, closed and easy to project on. The objective function is approximated by a sequence of different Sample Average Approximation functions with different sample sizes. The sample size update is based on two error estimates - SAA error and approximate solution error. The Spectral Projected Gradient method combined with a nonmonotone line search is used. The almost sure convergence results are achieved without imposing explicit sample growth condition. Preliminary numerical results show the efficiency of the proposed method.

Key words: spectral projected gradient, constrained stochastic problems, sample average approximation, variable sample size

1 Introduction

The problem that we consider is a constrained optimization problem of the form

$$\min f(x) = E[F(x, \xi)] \text{ subject to } x \in \Omega, \tag{1}$$

[*] 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

where $\Omega \subset \mathbb{R}^n$ is a convex and compact set, $\xi : \mathcal{A} \to \mathbb{R}^m$ is a random vector from a probability space $(\mathcal{A}, \mathcal{F}, \mathcal{P})$ and $F(\cdot, \xi) \in C^2(\Omega_e)$ with $\Omega \subset \Omega_e \subset \mathbb{R}^n$. The mathematical expectation that defines the objective function makes this problem difficult as$f$ is rarely available analytically and, even when it is available, it usually includes multiple integrals. Thus, the common approach is to approximate the objective function with Sample Average Approximation, SAA. The quality of approximation depends on the sample size and taking a large sample ensures good matching between the original problem (1) and the approximate problem. There are many possibilities for the sample choice, depending on properties of the underlying random variable ξ and availability of data, see Byrd et al. [10], Fu [14], Polak and Royset [28], Shapiro et al. [30], Spall [32].

The Sample Average Approximation is defined as

$$f_{\mathcal{N}}(x) = \frac{1}{N} \sum_{i=1}^{N} F(x, \xi^i)$$

for a given sample set $\mathcal{N} := \{\xi^1, \ldots, \xi^N\} \subset \mathcal{A}$. In general the quality of approximation depends heavily on the sample size N and taking N as large as computationally feasible is desirable in applications. Ultimately, letting the sample size $N \to \infty$ should result in some kind of asymptotic convergence, for example almost sure convergence or convergence in probability. On the other hand large N makes the evaluation of $f_{\mathcal{N}}$ and its derivatives expensive. Assuming that the sample size N is chosen appropriately, one can derive bounds on the difference between the solution of (1) and the approximate problem

$$\min f_{\mathcal{N}}(x) \text{ subject to } x \in \Omega. \tag{2}$$

Several results of this kind are available in Shapiro et al. [30], Spall [32]. The problem (2) is a reasonable approximation of the original one under a set of standard assumptions that we will state precisely later on. In general, one can be interested in solving the SAA problem for some finite, possibly very large N, as well as obtaining asymptotic results, that is, the results that cover the case $N \to \infty$, even if in practical applications one deals with a finite value of N. A naive application of an optimization solver to (2) is very often prohibitively costly if N is large due to the cost of calculating $f_{\mathcal{N}}(x)$ and its gradient. Thus, there is a vast literature dealing with variable sample scheme.

Two main approaches can be distinguished. In the first approach the objective function $f_{\mathcal{N}}$ is replaced with $f_{\mathcal{N}_k}(x)$ at each iteration k and the iterative procedure is essentially a two step procedure of the following form. Given the current approximation x_k and the sample size N_k, one has to find s_k such that the value of $f_{\mathcal{N}_k}(x_k + s_k)$ is decreased. After that we set $x_{k+1} = x_k + s_k$ and choose a new sample size N_{k+1}. The key ingredient of this procedure is the choice of N_{k+1}. The schedule of sample sizes $\{N_k\}$ should be defined in such way that either $N_k = N$ for k large enough or $N_k \to \infty$ if one is interested in asymptotic properties. Keeping in mind that $\min f_{\mathcal{N}_k}$ is just an approximation of the original problem and that the cost of each iteration depends on N_k, it is rather intuitive to start the optimization procedure with smaller samples and gradually increase the sample size N_k as the solution is approached. Thus, the most common schedule sequence would be an increasing sequence $N_0, N_1, \ldots$. In the case of solving the approximate problem for a finite N one can also consider a (possibly oscillating) scheduling sequence that takes into account the cost of each iteration and the progress made in function decrease. Such procedure results in a more efficient method than the corresponding procedure with strictly increasing schedule sequence, Bastin [2], Bastin et al. [3], Krejić and Krklec [20], Krejić and Krklec Jerinkić [21], Krejić and Martínez [22]. The results presented in Friedlander and Schmidt [13] are also closely related.

Regarding the almost sure convergence and considering the case $N \to \infty$, a strictly increasing scheduling sequence that goes to infinity is first considered in Wardi [33]. An Armijo type line search method is combined with SAA approach. The convergence is proved with upper zero density. An extension of [33] for the unconstrained case is presented in Yan and Mukai [35], where the adaptive precision is proposed, that is, the sequence $\{N_k\}_{k\in\mathbb{N}}$ is not determined in advance as in [33] but it is adapted during the iterative procedure. Nevertheless the sample size has to satisfy $N_k \to \infty$. The convergence result, again for the unconstrained case, is slightly stronger as the convergence with probability 1 is proved under the set of appropriate assumptions. The more general result that applies to both gradient and subgradient methods is obtained in Shapiro and Wardi [31]. The convergence with probability 1 is proved for the SAA gradient and subgradient methods assuming that the sample size tends to infinity and that the iterative sequence posses some additional properties. In this paper we are proposing a sample scheduling that ensures almost sure convergence of the SPG method for solving (1).

The second approach, often called the Surface Response Method, is again a two step procedure. It consists of a sequence of different SAA problems with different sample sizes that are approximately solved. After solving one SAA problem, the sample size is increased and the following SAA problem is again approximately solved. The main questions which crucially determine the efficiency of this procedure is the number of stages (that is, the number of different SAA problems to be solved) and the precision of each approximate solution. For further details one can see Pasupathy [27], Polak and Royset [28], Royset [29].

In this paper we are considering the constrained problem (1) assuming that the feasible set Ω is easy to project on. Typical case would be a box or polyhedron. The Spectral Projected Gradient method (Birgin et al. [7, 8]) is a well known for its efficiency and simplicity. The SPG method is applied to the sequence of different SAA approximate problems (2) coupled with a suitable sample scheduling scheme that yields almost sure convergence. Thus, the principal aims of this paper are: a) to derive an efficient sample scheduling that will yield an iterative sequence with almost sure convergence to a stationary point of the original problem through SAA approximation with $N \to \infty$, through efficient and computationally feasible optimization procedure for solving (2), and b) to prove the almost sure convergence of the SPG method with an appropriate sample scheduling. It is important to notice that although the logic behind the sample scheduling is similar to the corresponding one in Krejić and Krklec [20], Krejić and Krklec Jerinkić [21], the conceptual change from a finite sample size in the SAA approximation in [20, 21] to an infinite sample introduces nontrivial changes in the sample scheduling. The same is true for theoretical consideration. Additional changes are due to the difference between unconstrained and constrained problems.

The paper is organized as follows. Some preliminaries are given in Section 2. The SPG algorithm and the appropriate scheduling algorithm are stated in Section 3, while the convergence results are presented in Section 4. Section 5 contains numerical experiments that confirm the theoretical results. Some conclusions are drawn in Section 6. The almost sure convergence will be abbreviated with a.s. in the rest of the paper.

2 Preliminaries

We will assume that the samples used for calculating the Sample Average Approximation at each iteration are available and taken cumulatively. So, for two positive integers N and M with $N < M$ we define the sample sets

$$\mathcal{N} := \{\xi^1, \ldots, \xi^N\} \subset \{\xi^1, \ldots, \xi^N, \ldots, \xi^M\} =: \mathcal{M}$$

and the corresponding approximation of the objective function is

$$f_{\mathcal{N}}(x) = \frac{1}{N}\sum_{i=1}^{N} F(x, \xi^i).$$

Clearly, $N \in \mathbb{N}$ defines the sample set $\mathcal{N}$ and vice versa. We will use both notations in the sequel.

The following set of assumptions makes the problem well defined and allows us to work with the function $f_{\mathcal{N}}$ and its gradient.

Assumption A1. The set $\Omega \subset \mathbb{R}^n$ is compact and convex.

Assumption A2. For any ξ from $(\mathcal{A}, \mathcal{F}, \mathcal{P})$ there holds $F(\cdot, \xi) \in C^2(\Omega_e)$, where $\Omega \subset \Omega_e \subset \mathbb{R}^n$ and Ω_e is open and bounded. Furthermore, F and ∇F are dominated by an integrable functions and $E[\nabla F(x, \xi)] = \nabla E[F(x, \xi)]$ holds.

Assumption A3. The sample set $\{\xi^1, \xi^2, \ldots\}$ is i.i.d.

The assumptions A1-A2 imply that $F(x, \xi)$ and $\nabla F(x, \xi)$ are bounded on Ω since they are continuous and the feasible set is bounded. The assumption A2 states that F and ∇F are dominated by integrable functions. In other words we assume that there exist functions $K_1(\xi)$ and $K_2(\xi)$ such that for $x \in \Omega$

$$F(x, \xi) \leq K_1(\xi), \ \nabla F(x, \xi) \leq K_2(\xi),$$

and such that $E(K_1) < \infty, \ E(K_2) < \infty$. Moreover, the second part of the assumption A2 together with A3 implies that f is finite valued and continuous on any compact set (Shapiro et al. [30], Theorem 7.48) which implies that f is bounded on Ω. The same is true for the gradient, (Shapiro et al. [30], Theorem 7.52)

$$\nabla f_{\mathcal{N}}(x) = \frac{1}{N}\sum_{i=1}^{N} \nabla F(x, \xi^i).$$

Both $f_{\mathcal{N}}(x)$ and $\nabla f_{\mathcal{N}}(x)$ are bounded and uniformly continuous on the compact set Ω under the stated assumptions. The assumptions A1-A3 together also imply that the SAA gradients uniformly converge to the true gradient value with probability 1, that is,

$$\lim_{N\to\infty} \sup_{x\in\Omega} \|\nabla f_{\mathcal{N}}(x) - \nabla f(x)\| = 0 \; a.s. \tag{3}$$

Let us now define the following two functions that will be used later on to define the sample update. Roughly speaking, the first one measures the quality of approximation $f_{\mathcal{N}}(x) \approx f(x)$. This function is used in the Algorithms and does not necessarily represent the true (theoretically sound) error of the approximation. For example, it can be defined trivially as $\nu(x, N) = 1/N$.

Assumption A4. Assume that $\nu : \mathbb{R}^n \times \mathbb{N} \to \mathbb{R}_+$ is such that for any finite valued N there exists ν_N such that

$$\nu(x, N) \geq \nu_N > 0 \quad \text{for every} \quad x \in \Omega. \tag{4}$$

The error function $\nu(x_k, N_k)$ is used to define the new sample size N_{k+1} such that the approximation error is well balanced with the decrease of $f_{\mathcal{N}_k}$. The decrease is denoted by dm_k and it approximates the difference $f_{\mathcal{N}_k}(x_k) - f_{\mathcal{N}_k}(x_{k+1})$.

Assumption A5. Assume that $\gamma : \mathbb{N} \to (0, 1)$ is such that γ is increasing function of N and

$$\lim_{N\to\infty} \gamma(N) = 1. \tag{5}$$

One obvious possibility is to define $\gamma(N) = \exp(-1/N)$.

Notation: For a given sample set $\mathcal{N}_k$ and $x_k \in \mathbb{R}^n$ we denote $g_k = \nabla f_{\mathcal{N}_k}(x_k)$. The orthogonal projection on Ω is denoted by $P_\Omega(\cdot)$, that is,

$$P_\Omega(x) = arg\ min_{z\in\Omega}\|z - x\|,$$

where the norm is assumed to be Euclidean.

3 Algorithms

The iterative method we consider is defined by Algorithm 1-2 below. The main algorithm is Algorithm 1 which defines a new iteration using the spectral projected gradient method with nonmonotone line search, for a given sample size N_k. The sample size is updated through Algorithm 2. Two sequences, $\{N_k\}$ and $\{N_k^{\min}\}$ are defined, with N_k being the actual sample size and $N_k^{\min}$ the lower bound of the sample size. The nonmonotone line search is defined by a sequence $\{\varepsilon_k\}$ such that

$$\sum_{k=0}^{\infty} \varepsilon_k \leq \varepsilon < \infty \quad \text{and} \quad \varepsilon_k > 0 \quad \text{for every } k.$$

Algorithm 1

Given $N_0 = N_0^{\min}, x_0 \in \Omega, 0 < \alpha_{\min} < \alpha_{\max}, \alpha_0 \in [\alpha_{\min}, \alpha_{\max}], \beta, \eta \in (0,1), \{\varepsilon_k\}$. Set $k = 0$.

Step 1. Test for stationarity. If

$$\|P_\Omega(x_k - g_k) - x_k\| = 0$$

increase $N_k = N_k + 1$ until $\|P_\Omega(x_k - g_k) - x_k\| \neq 0$. Set $N_k^{\min} = N_k$.

Step 2. Compute the search direction

$$p_k = P_\Omega(x_k - \alpha_k g_k) - x_k.$$

Step 3. Find the smallest nonnegative integer j such that $\lambda_k = \beta^j$ satisfies

$$f_{\mathcal{N}_k}(x_k + \lambda_k p_k) \leq f_{\mathcal{N}_k}(x_k) + \eta \lambda_k p_k^T g_k + \varepsilon_k.$$

Define $s_k = \lambda_k p_k$ and set $x_{k+1} = x_k + s_k$.

Step 4. Update $\mathcal{N}_{k+1}$ within Algorithm 2. Set $\mathcal{I}_k = \mathcal{N}_{k+1} \cap \mathcal{N}_k$ and $y_k = \nabla f_{\mathcal{I}_k}(x_{k+1}) - \nabla f_{\mathcal{I}_k}(x_k)$.

Step 5. Compute $b_k = s_k^T y_k$ and $a_k = s_k^T s_k$ and set

$$\alpha_{k+1} = \min\{\alpha_{\max}, \max\{\alpha_{\min}, a_k/b_k\}\}.$$

Step 6. Set $k = k + 1$ and go to Step 1.

Before we state Algorithm 2, let us comment on the algorithm above. First, notice that the sequence of iterates $\{x_k\}_{k\in\mathbb{N}}$ remains in the feasible set Ω. This is a consequence of Ω being convex and compact and of definition of a search direction - the projection is performed only once at each iteration and the line search is performed within Ω.

As we already mentioned, the progress made in each iteration is measured by the decrease measure dm_k. Let us define

$$dm_k = -\lambda_k p_k^T g_k.$$

Notice that Step 1 of the previously stated algorithm ensures that dm_k used in Step 4 of Algorithm 2, is strictly positive assuming that there is no infinite loop in Step 1. The possibility of an infinite loop is discussed below Lemma 3.1. In all other cases the strictly positive dm_k comes from the fact that $p_k^T g_k = 0$ implies $\|P_\Omega(x_k - g_k) - x_k\| = 0$ according to Lemma 3.1 from Birgin et al. [7]. For the sake of completeness we state this important result below.

Lemma 3.1. *[7] Define $g_t(x) = P_\Omega(x - t\nabla f(x)) - x$. For all $x \in \Omega$, $t \in (0, \alpha_{max}]$,*

(i) $\nabla^T f(x) g_t(x) \leq -\frac{1}{t}\|g_t(x)\|^2 \leq -\frac{1}{\alpha_{max}}\|g_t(x)\|^2$

(ii) The vector $g_t(x^*)$ vanishes if and only if x^* is a stationary point for (1).

Notice that the parameter α_{max} stated in the previous lemma coincides with the upper bound of the spectral coefficient. Parameters α_{min} and α_{max} represent safeguard parameters which ensure that the spectral coefficient remains uniformly bounded away from zero as well as from infinity. One common choice is to set α_{min} to $10^{-4}, 10^{-8}$ or even 10^{-30}. On the other hand, the upper bound is usually set on $10^4, 10^8$ or even 10^{30}. Now, if we assume $\alpha_{max} \geq 1$ (which is usually true), this lemma implies that x^* is a stationary point for (1) if

$$P_\Omega(x^* - \nabla f(x^*)) - x^* = 0.$$

Algorithm 1 implies that either there exists x_k such that

$$\|P_\Omega(x_k - \nabla f_{\mathcal{N}}(x_k)) - x_k\| = 0, \ N \geq N_k, \tag{6}$$

i.e., Step 1 is performed infinitely many times, or the algorithm generates an infinite sequence $\{x_k\}$. If x_k is such that (6) holds, then (3) implies

$$\lim_{N\to\infty} \nabla f_{\mathcal{N}}(x_k) = \nabla f(x_k)$$

and therefore, $\|P(x_k - \nabla f(x_k)) - x_k\| = 0$, that is, x_k is a stationary point for (1). Given that in actual implementation the algorithm will eventually stop at some finite N, the stationary point x_k which satisfies (6) will be detected. So, from now on we assume that (6) does not occur.

As $\varepsilon_k > 0$, Step 3 necessarily terminates with a finite j for any search direction. So this step is well defined. Nevertheless, the search direction p_k calculated at Step 3 is descent direction for $f_{\mathcal{N}_k}$ at x_k, as stated in Lemma 3.1 above. The additional term ε_k allows more freedom in the choice of step length and allows the nonmonotonicity that ensures that the good properties of spectral projected gradient method are preserved, see Birgin et al. [7, 8], Li and Fukushima [26] . It is important to state here that any other nonmonotone rule like the rules considered in [7], Grippo et al. [15, 16], Krejić and Krklec Jerinkić [21], Zhang and Hager [34] could be applied here, but the analysis would be a bit more cumbersome technically than with the rule we employ at Step 3.

The spectral coefficient α_k is calculated using the intersection of two consecutive samples. It is easy to notice that $I_k = \min\{N_k, N_{k+1}\}$ so both gradient values, $\nabla f_{\mathcal{I}_k}(x_{k+1})$ and $\nabla f_{\mathcal{I}_k}(x_k)$, are available and no additional gradient values are needed for the calculation of the spectral coefficient. One could easily state

$$y_k = \nabla f_{\mathcal{N}_{k+1}}(x_{k+1}) - \nabla f_{\mathcal{N}_k}(x_k)$$

instead of y_k defined in Step 4 of the algorithm. In fact, the question of the best sample for calculation of y_k is still unsolved and there are many discussions in the literature, see Mokhtari and Ribeiro [24], Byrd et al. [9, 10, 11], Krejić et al. [19]. In the deterministic case y_k satisfies

$$y_k = \left(\int_0^1 \nabla^2 f(x_k + ts_k) dt \right) s_k.$$

As we are dealing with the expectation with respect to ξ, the variance of ξ and the corresponding variances of f and its derivatives, play an important role in the last equation. Furthermore, y_k defines the spectral coefficient a_k/b_k, which is the Rayleigh quotient relative to the average Hessian matrix $\int_0^1 \nabla^2 f(x_k + ts_k)dt$, see Birgin et al. [8]. The eigenset of $\nabla^2 f_{\mathcal{N}}$ is clearly influenced by the sample set $\mathcal{N}$, so the definition of y_k should reflect this fact as well. The choice we made here is a consequence of empirical experience that yielded a strong preference towards the definition of y_k stated in Step 4.

Recall that $dm_k = -\lambda_k p_k^T g_k$ and $dm_k > 0$. The algorithm for the sample size update is as follows.

Algorithm 2

```
         Given $dm_k, N_k, N_k^{min}, x_k, x_{k+1}$.

Step 1.  Candidate $N_k^+$.
         Set $N = \max\{N_k, N_k^{\min}\}$

  Step 1.1  If $dm_k = \nu(x_k, N_k)$ set $N_k^+ = N$.
  Step 1.2  If $dm_k > \nu(x_k, N_k)$
            While $dm_k > \nu(x_k, N)$ and $N > N_k^{\min}$ set
                        $N = N - 1.$
            End(While).
            Set $N_k^+ = N$.
  Step 1.3  If $dm_k < \nu(x_k, N_k)$
            While $dm_k < \nu(x_k, N)$ set
                        $N = N + 1.$
            End(While).
            Set $N_k^+ = N$.

Step 2.  Update of $N_{k+1}$.
         If $N_k^+ < N_k$ and
            $\rho_k = \left| \frac{f_{\mathcal{N}_k^+}(x_k) - f_{\mathcal{N}_k^+}(x_{k+1})}{f_{\mathcal{N}_k}(x_k) - f_{\mathcal{N}_k}(x_{k+1})} - 1 \right| \geq \frac{N_k - N_k^+}{N_k},$     (7)
         set $N_{k+1} = N_k$. Otherwise set $N_{k+1} = N_k^+$.

Step 3.  Update of $N_k^{\min}$.

    3.1  If $N_{k+1} = N_k$ set $N_{k+1}^{\min} = N_k^{\min}$.
```

3.2 If $N_{k+1} \neq N_k$ update $N_k^{\min}$ by the following rule.
If N_{k+1} has been used in some of the previous iterations and

$$\frac{f_{\mathcal{N}_{k+1}}(x_{h(k)}) - f_{\mathcal{N}_{k+1}}(x_{k+1})}{k+1-h(k)} \leq \gamma(N_{k+1})\nu(x_{k+1}, N_{k+1}), \tag{8}$$

where $h(k)$ is the iteration in which we started to use N_{k+1} for the last time, set $N_{k+1}^{\min} > N_k^{\min}$.
Otherwise set $N_{k+1}^{\min} = N_k^{\min}$.

A few words on Algorithm 2 are due as well. The main objective of the variable sample scheme is to ensure some balance between the computational costs and precision of the Sample Average Approximation. The principal idea is to increase or decrease the sample size according to the progress made in decreasing the objective function. Such approach ensures that we work with low precision whenever possible, saving the computational effort if possible. At the same time the presented scheme ensures that the sample size increases to infinity and allows us to prove the almost sure convergence, as we will demonstrate later on.

The main ingredients of Algorithm 2 are the decrease in the objective function measured by dm_k and the precision of the sample average approximation measured by $\nu(x_k, N_k)$. The sample size is increased or decreased in such a way that these two measures trail each other. To achieve a balance between dm_k and $\nu(x_k, N)$ we are in fact constructing two sample size sequences, N_k and $N_k^{\min}$. The sample size is defined within Step 1-2. In Step 1 the candidate N_k^+ is determined to preserve the balance between dm_k and $\nu(x_k, N)$. Notice that the candidate N_k^+ is constructed in such way that provides that the lower bound restriction, $N_k^+ \geq N_k^{min}$ holds. This follows from the fact that the starting trial sample size is $N = \max\{N_k, N_k^{\min}\}$ and, according to Step 1.2, it is decreased only if the lower bound allows the decrease. If $N_k^+ < N_k$, that is, if a decrease of the sample size is proposed, we perform an additional check stated in Step 2 to avoid possibly unproductive decreases. The parameter ρ_k is calculated only if we have a possible sample decrease as it measures similarity between the two models, represented by $f_{\mathcal{N}_k^+}$ and $f_{\mathcal{N}_k^+}$, through the decrease obtained at the current iteration. If these two model functions are close to each other, the parameter ρ_k is relatively small (close to zero) and we chose to work with the cheaper model function, that is, we set $N_{k+1} = N_k^+$. Otherwise, we do not allow the decrease in the sample size since the model functions are too different and taking the smaller sample size may lead us in a wrong direction. In that case, we chose the larger sample size, that is, we chose the objective function estimate which we believe is closer to the final objective f. The quantity on the right hand side of (7) is motivated by our previous numerical testings. Basically, it encourages larger decreases of the sample size since they were usually beneficial for the overall performance of the algorithm. Moreover, notice that $N_{k+1} \geq N_k^{min}$. If $N_{k+1} = N_k^+$, this is obvious since we saw that $N_k^+ \geq N_k^{min}$. On the other hand, $N_{k+1} = N_k$, but this may happen only if $N_k > N_k^+$. Therefore, the lower bound condition is satisfied.

The second sequence $N_k^{\min}$ is updated in Step 3 and it is clearly nondecreasing. It represents the smallest precision allowed at each stage of the optimization process and its role is to eventually push N_k towards infinity, even with the oscillations of N_k that are permitted by the algorithm. Algorithm 2 is essentially inspired by the variable sample scheme for solving the so called SAA problem with finite N, as presented in Krejić and Krklec Jerinkić [20, 21] for unconstrained problems. The first idea of this kind is developed in Bastin [2], Bastin et al. [3, 4] for the trust region approach and SAA methods. Although this Step is crucial for the convergence analysis, the lower bound increase is rarely activated according to our experience. The main idea in Step 3 is to track the progress related to various precision levels, that is, to various sample sizes in our case. The left hand side in the inequality (8) represents the average decrease between the last iteration at which we used the sample size and the new iteration with the same sample size. If this decrease is relatively modest, we assume that the oscillations were not that beneficial and increase the lower bound for the sample size, that is, the minimal allowed precision. The right hand side of the inequality (8) depends on the sample size as we do not want to treat all precision levels equally - we want to be more rigorous when the final objective function is approached. This Algorithm is conceptually based on [20] but, as we are now interested in pushing the sample size to the infinity, the steps are adjusted to allow an infinite increase of the sample size.

4 Convergence theory

The convergence results developed in this section show that the proposed method generates an iterative sequence that converges a.s. towards a solution of (1). Under the assumptions stated in Section 2 the first theorem states that the sample size generated by Algorithm 2 goes to infinity. After that, we prove a.s. convergence with an additional, but standard assumption regarding the error of the SAA approximation.

Remark 1. The sequence generated by Algorithm 1 is obviously random and thus all the relevant quantities such as N_k depend on the particular sample realization ω. However, the subsequent result shows that the sample size tends to infinity for any given ω. Thus, the result holds surely and not just almost surely. In order to avoid cumbersome notation, we will omit ω in the sequel and write, for example, $\{N_k\}_{k\in\mathbb{N}}$ instead of $\{N_k(\omega)\}_{k\in\mathbb{N}}$ and subsequence K_1 instead of subsequence $K_1(\omega)$.

Theorem 4.1. *Assume that A1-A5 hold. Then* $\lim_{k\to\infty} N_k = \infty$.

Proof. First, let us show that the sequence $\{N_k\}_{k\in\mathbb{N}}$ can not become stationary. Assume that there are $\bar{N}_0$ and $\bar{k}_0$ such that

$$N_k = \bar{N}_0 \quad \text{for every} \quad k \geq \bar{k}_0. \tag{9}$$

Then, for each $k \geq \bar{k}_0$ we have

$$f_{\bar{N}_0}(x_{k+1}) \leq f_{\bar{N}_0}(x_k) - \eta dm_k + \varepsilon_k$$

by Step 3 of Algorithm 1. Thus

$$f_{\bar{N}_0}(x_{\bar{k}_0+m}) \leq f_{\bar{N}_0}(x_{\bar{k}_0}) - \eta \sum_{j=0}^{m-1} dm_{\bar{k}_0+j} + \sum_{j=1}^{m-1} \varepsilon_{\bar{k}_0+j}$$

for arbitrary $m \in \mathbb{N}$. Recall that the sequence of iterates remains in Ω and therefore $f_{\bar{N}_0}(x_{\bar{k}_0+m})$ is bounded from below for every m. Moreover, as $0 < \sum_{k=0}^{\infty} \varepsilon_k < \infty$ and by rearranging the previous inequality we conclude that $\sum_{j=0}^{\infty} dm_{\bar{k}_0+j}$ is finite. Therefore, using the fact that dm_k is nonnegative, we obtain

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

On the other hand, as a consequence of the Assumption A4, for each $k \geq \bar{k}_0$ we have

$$\nu(x_k, N_k) = \nu(x_k, \bar{N}_0) \geq \nu_{\bar{N}_0} > 0,$$

so there exists $\bar{k}_1 > \bar{k}_0$ such that

$$\nu(x_{\bar{k}_1}, N_{\bar{k}_1}) > dm_{\bar{k}_1}.$$

However, Step 1.3 of Algorithm 2 implies that $N_{\bar{k}_1+1} > N_{\bar{k}_1} = \bar{N}_0$, which is a contradiction with (9).

Algorithm 2 ensures that $N_{k+1} \geq N_k^{\min}$. So if $\lim_{k\to\infty} N_k^{\min} = \infty$ we obviously have the statement. Thus, in the sequel we consider the opposite case and show that the statement is true in that case as well.

Let us now assume that $N_k^{\min} = N_{\max}$ for $k \geq \bar{k}_2$. Notice that the lower bound $N_k^{\min}$ is nondecreasing and it can be increased only throughout Step 1 of Algorithm 1 or in Step 3.2 of Algorithm 2. Since $N_k^{\min}$ is assumed to be bounded, i.e., $N_k^{\min} \leq N_{\max}$, Step 1 of Algorithm 1 can happen only a finitely many times. Therefore, without loss of generality we may exclude this scenario. On the other hand, in general, there are two possible outcomes of Step 3 of Algorithm 2, $N_{k+1}^{\min} = N_k^{\min}$ or $N_{k+1}^{\min} > N_k^{\min}$. The second outcome is obviously not possible for $k \geq \bar{k}_2$, so we must have $N_{k+1}^{\min} = N_k^{\min}, k \geq \bar{k}_2$. This further implies that we have one of the following three possibilities for each $k \geq \bar{k}_2$.

M1 $N_{k+1} = N_k$

M2 $N_{k+1} \neq N_k$ and N_{k+1} has not been used before

M3 $N_{k+1} \neq N_k$, N_{k+1} has been used before and

$$\frac{f_{\mathcal{N}_{k+1}}(x_{h(k)}) - f_{\mathcal{N}_{k+1}}(x_{k+1})}{k+1-h(k)} \geq \gamma(N_{k+1})\nu(x_{k+1}, N_{k+1})$$

Assume that the statement of this theorem is not true so there exists an infinite subsequence of $\{N_k\}_{k\in\mathbb{N}}$ such that its elements are bounded. Then there must exist a finite sample size $\bar{N}_1$ which is visited infinitely many times, that is, there exists an infinite subsequence $K_0 = \{k \geq \bar{k}_2 : N_{k+1} = \bar{N}_1\}$, for some $\bar{N}_1 < \infty$. Since we proved that the sequence $\{N_k\}$ can not be stationary, there exists an infinite subsequence $K_1 \subset K_0$ such that M1 does not hold for $k \in K_1$. More precisely, there must exist an infinite subsequence of iterations in which the sample size is changed on $\bar{N}_1$, that is, there exists $K_1 = \{k \geq \bar{k}_2 : N_k \neq N_{k+1} = \bar{N}_1\}$. Moreover, by excluding the first member of the sequence K_1 we obtain an infinite subsequence $K_2 \subset K_1$ that makes the scenario M2 impossible as well since the sample size $\bar{N}_1$ is obviously used before during the optimization process. Therefore, for every $k \in K_2$

$$\frac{f_{\bar{\mathcal{N}}_1}(x_{h(k)}) - f_{\bar{\mathcal{N}}_1}(x_{k+1})}{k+1-h(k)} \geq \gamma(\bar{N}_1)\nu(x_{k+1}, \bar{N}_1).$$

Let $K_3 = \{k+1, k \in K_2\}$. Notice that this corresponds to the subsequence of iterations in which the sample size $\bar{N}_1$ is revisited. Therefore, iterates $x_{h(k)}$ and x_{k+1} from the previous inequality must be a neighboring iterates of the sequence $\{x_k\}_{k \in K_3}$. Also, notice that $k+1-h(k) > 1$. Denoting $\{x_k\}_{k \in K_3} := \{x_{k_j}\}_{j \in \mathbb{N}}$ we obtain that the following holds for every $j \in \mathbb{N}$

$$f_{\bar{N}_1}(x_{k_j}) \geq f_{\bar{N}_1}(x_{k_{j+1}}) + \gamma(\bar{N}_1) e(x_{k_{j+1}}, \bar{N}_1) \tag{10}$$

Given that, according to assumptions A4-A5,

$$\gamma(\bar{N}_1)\nu(x_{k_{j+1}}, \bar{N}_1) \geq \gamma(\bar{N}_1)\nu_{\bar{N}_1} = c > 0,$$

(10) implies that $f_{\bar{N}_1}$ is unbounded on Ω which is clearly wrong. Thus the statement is proved. □

Let us now proceed to prove the almost sure convergence results for Spectral Projected Gradient method defined in Algorithm 1. Recall that $g_k = \nabla f_{\mathcal{N}_k}(x_k)$ and p_k is a search direction.

Lemma 4.1. *Assume that A1-A5 hold and let $K \subset \mathbb{N}$ be a subset of iterations such that*

$$\lim_{k \in K} x_k = x^*, \; \lim_{k \in K} p_k = 0.$$

Then x^* is a stationary point for (1) a.s.

Proof. Given that the search directions p_k converge to zero through K, we have

$$\begin{aligned} 0 &= \lim_{k \in K} p_k = \lim_{k \in K} [P_\Omega(x_k - \alpha_k g_k) - x_k] \\ &= \lim_{k \in K} P_\Omega(x_k - \alpha_k g_k) - \lim_{k \in K} x_k \\ &= P_\Omega(x^* - \lim_{k \in K} \alpha_k g_k) - x^*. \end{aligned}$$

The sequence of spectral coefficients α_k is bounded and thus there exists $K_1 \subset K$ such that $\lim_{k \in K_1} \alpha_k = \alpha^* \in [\alpha_{\min}, \alpha_{\max}]$.

As $N_k \to \infty$ we have

$$\lim_{k \in K_1} g_k = \lim_{k \in K_1} \nabla f_{\mathcal{N}_k}(x_k) = \nabla f(x^*) \quad \text{a.s.}$$

Thus

$$0 = P_\Omega(x^* - \lim_{k \in K_1} \alpha_k g_k) - x^* = P_\Omega(x^* - \alpha^* \nabla f(x^*)) - x^*.$$

Now, the statement follows by Lemma 3.1. □

In order to prove the main results, we make an additional assumption.

Assumption A6. Let ω be an arbitrary sample path generated by Algorithm 1. Assume that $e : \mathbb{R}^n \times \mathbb{N} \to \mathbb{R}_+$ is a function with the following properties

$$\lim_{N\to\infty} \sup_{x\in\Omega} e(x,N) = 0, \tag{11}$$

$$|f(x) - f_{\mathcal{N}}(x)| \leq e(x,N) \; a.s. \; x \in \Omega, N \in \mathbb{N}, N \geq N^0(w). \tag{12}$$

Notice that (12) is assumed with probability 1 while (11) is without any random elements and thus holds surely. The function e is deterministic and the only property that holds with probability (one) is that it bounds the error that depends on stochastic quantity $f_{\mathcal{N}}(x)$. Although this assumption seems restrictive, the results presented in Homem-de-Mello [17] show that such functions exist provided, for example, that the variance of $F(x,\xi)$ is uniformly bounded on Ω. This is true for the class of problems where the Gaussian noise is added to some smooth deterministic function to obtain $F(x,\xi)$, or, more generally, if $F(x,\xi) = h_1(x) + h_2(x)\xi$ where h_1 and h_2 are smooth functions and the variance of ξ is bounded. In that case, one possible choice is

$$e(x,N) = C(x)\sqrt{\frac{\ln(\ln(N))}{N}}, \tag{13}$$

where $C(x)$ is related to the variance of $F(x,\xi)$. It is important to notice that in general, as pointed out by one of the Reviewers, $C(x)$ might be infinite and thus the condition (11) might be violated.

Given that Theorem 4.1 states that N_k grows to infinity, it follows that the error bound (13) derived in Proposition 3.5 [17] for cumulative samples remains a relevant choice, at least for theoretical purposes. However, the log bound is considered as too conservative from the practical point of view and it is often approximated with a bound of sample variance type (Bastin [2]). More precisely, the bound that is frequently used is $z\frac{\hat{\sigma}(x_k,N_k)}{\sqrt{N_k}}$ where z is a suitable quantile and $\hat{\sigma}^2(x_k,N_k)$ is a sample variance of F, that is,

$$\hat{\sigma}^2(x_k,N_k) = \frac{1}{N_k - 1}\sum_{i\in\mathcal{N}_k} \left(F(x_k,\xi^i) - f_{\mathcal{N}_k}(x_k)\right)^2.$$

We use this bound in implementation in order to get numerical results. We also set $\nu(x,N) = e(x,N)$.

The main convergence results are given in Theorem 4.2 and Theorem 4.3 below. The first theorem claims that there exists an accumulation point which is stationary. In the second theorem we prove a stronger result, that each strictly strong accumulation point is stationary. Both of the results hold a.s. To the best of our knowledge, a stronger result (that each accumulation point is stationary a.s.) can not be attained under this setup. However, according to the analysis presented in this paper, one can achieve the result of every accumulation point being stationary a.s. if the sample size is increased fast enough. We state this result in Theorem 4.4 for completeness.

Theorem 4.2. *Assume that A1-A6 hold. For almost every sequence* $\{x_k\}_{k\in\mathbb{N}}$ *generated by Algorithm 1, there exists an accumulation point of* $\{x_k\}_{k\in\mathbb{N}}$ *which is stationary for (1).*

Proof. Let us demonstrate that there exists at least one subsequence of $\{p_k\}$ which converges to zero.

Suppose that there exists $p > 0$ such that for every $k \in \mathbb{N}$

$$\|p_k\|^2 \geq p > 0. \tag{14}$$

Then, Lemma 3.1 implies the following inequalities

$$p_k^T \nabla f_{\mathcal{N}_k}(x_k) \leq -\frac{1}{\alpha_{\max}} \|p_k\|^2 \leq -\frac{1}{\alpha_{\max}} p := -\bar{p} < 0, \; k \in \mathbb{N} \tag{15}$$

and

$$dm_k = -\lambda_k p_k^T \nabla f_{\mathcal{N}_k}(x_k) \geq \lambda_k \bar{p}, \quad k \in \mathbb{N}. \tag{16}$$

Suppose that $\lambda_k \geq \bar{\lambda} > 0, \; k \in \mathbb{N}$. In that case

$$dm_k \geq \bar{\lambda}\bar{p} := \bar{d} > 0, \quad k \in \mathbb{N}. \tag{17}$$

Define

$$\tilde{e}_N = \sup_{x\in\Omega} e(x, N).$$

Moreover, given that $\{x_k\}_{k\in\mathbb{N}} \subset \Omega$ and $N_k \to \infty$, according to assumption A6 we have that for all k large enough (more precisely, for $k \geq \bar{k}_1 = \bar{k}_1(\omega)$ where $\bar{k}_1$ is such that $N_k \geq N^0(\omega)$ for every $k \geq k_1$) there holds

$$|f(x_k) - f_{\mathcal{N}_k}(x_k)| \leq e(x_k, N_k) \leq \tilde{e}_{N_k} \text{ a.s.}$$

and (11) implies

$$\lim_{k\to\infty} \tilde{e}_{N_k} = 0. \tag{18}$$

Thus, for every $k \geq \bar{k}_1$ we have a.s.

$$\begin{aligned} f(x_{k+1}) &\leq f_{\mathcal{N}_k}(x_{k+1}) + \tilde{e}_{N_k} \\ &\leq f_{\mathcal{N}_k}(x_k) + \varepsilon_k - \eta d m_k + \tilde{e}_{N_k} \\ &\leq f(x_k) + 2\tilde{e}_{N_k} + \varepsilon_k - \eta \bar{d}. \end{aligned}$$

Let $q \in (0, \eta\bar{d})$ be an arbitrary constant. Then, (18) implies that $\tilde{e}_{N_k} < (\eta\bar{d} - q)/2$ for every k large enough, that is, there exists $\bar{k}_2 = \bar{k}_2(\omega)$ such that

$$2\tilde{e}_{N_k} < \eta\bar{d} - q \quad \text{for every} \quad k \in \mathbb{N},\ k \geq \bar{k}_2(\omega) \tag{19}$$

and for every $k \geq \bar{k}(\omega) = \max\{\bar{k}_1, \bar{k}_2\}$

$$f(x_{k+1}) \leq f(x_k) + \varepsilon_k - q \quad \text{a.s,}$$

which furthermore implies that a.s.

$$f(x_{\bar{k}+s}) \leq f(x_{\bar{k}}) + \sum_{j=0}^{s-1} \varepsilon_j - sq \leq f(x_{\bar{k}}) + \varepsilon - sq,\ s \in \mathbb{N}.$$

Letting s tend to infinity we obtain that f is unbounded which is not possible.

Now, suppose that there is a subsequence $K_1(\omega) \subseteq \{\bar{k}+1, \bar{k}+2, ...\}$ such that

$$\lim_{k\in K_1} \lambda_k = 0.$$

Since the initial step size in the line search is 1 at every iteration and the backtracking with parameter β is employed, for every $k \in K_1$ large enough there must be at least one unsuccessful trial in Step 3 of Algorithm 1. Let us denote by λ'_k the trial preceding the successful trial λ_k, that is, let $\lambda_k = \beta\lambda'_k$. In that case the line search rule implies that for every $k \in K_1$ there exists $\lambda'_k = \lambda_k/\beta$ such that $\lim_{k\in K_1} \lambda'_k = 0$ and

$$f_{\mathcal{N}_k}(x_k + \lambda'_k p_k) > f_{\mathcal{N}_k}(x_k) + \eta\lambda'_k p_k^T g_k + \varepsilon_k.$$

As $\varepsilon_k > 0$ we have

$$f_{\mathcal{N}_k}(x_k + \lambda'_k p_k) > f_{\mathcal{N}_k}(x_k) + \eta\lambda'_k p_k^T g_k.$$

The Mean Value Theorem implies the existence of $t_k \in (0,1)$ such that

$$p_k^T \nabla f_{\mathcal{N}_k}(x_k + t_k \lambda'_k p_k) \geq \eta p_k^T \nabla f_{\mathcal{N}_k}(x_k). \tag{20}$$

Given that $\{p_k\}$ and $\{x_k\}$ are bounded, there exists $K_2(\omega) \subseteq K_1$ such that $\lim_{k \in K_2}(x_k, p_k) = (x^*, p^*)$ and

$$\lim_{k \in K_2} x_k + t_k \lambda'_k p_k = x^*.$$

Therefore, taking limits on both sides of (20) we get

$$(p^*)^T \nabla f(x^*) \geq \eta (p^*)^T \nabla f(x^*) \text{ a.s.} \tag{21}$$

The condition $\eta \in (0,1)$ and (21) together yield

$$(p^*)^T \nabla f(x^*) \geq 0 \text{ a.s.} \tag{22}$$

On the other hand, taking limit for $k \in K_2$ in (15) we obtain

$$0 \leq (p^*)^T \nabla f(x^*) \leq -\bar{p} < 0 \text{ a.s.} \tag{23}$$

which is clearly in contradiction with (22). Therefore, we conclude that (14) is wrong and there exists a subsequence of $\{p_k\}_{k \in \mathbb{N}}$ that converges to zero, that is, there exists $K_3(\omega) \subseteq \mathbb{N}$ such that $\lim_{k \in K_3} p_k = 0$. Again, $\{x_k\}$ is bounded and there must exist $K_4(\omega) \subseteq K_3$ such that

$$\lim_{k \in K_4} p_k = 0 \quad \text{and} \quad \lim_{k \in K_4} x_k = \tilde{x}.$$

Finally, Lemma 4.1 implies that $\tilde{x}$ is a stationary point for (1) a.s. and the statement follows.□

In order to show the stronger result, we state the following definition of strictly strong accumulation point, Yan and Mukai [35].

Definition 4.1. *[35] A point x^* is called strictly strong accumulation point of the sequence $\{x_k\}_{k \in \mathbb{N}}$ if there exists a subsequence $K \subseteq \mathbb{N}$ and a constant $b \in \mathbb{N}$ such that $\lim_{k_i \in K} x_{k_i} = x^*$ and $k_{i+1} - k_i \leq b$ for any two consecutive elements $k_i, k_{i+1} \in K$.*

Theorem 4.3. *Assume that A1-A6 hold. For almost every sequence $\{x_k\}_{k \in \mathbb{N}}$ generated by Algorithm 1, every strictly strong accumulation point of the sequence $\{x_k\}_{k \in \mathbb{N}}$ is stationary for (1).*

Proof. Let x^* be an arbitrary strictly strong accumulation point of the sequence $\{x_k\}_{k\in\mathbb{N}}$. That means that there is a subsequence $K \subseteq \mathbb{N}$ and a positive constant b such that $\lim_{k\in K} x_k = x^*$ and $k_{i+1} - k_i \leq b$ for every $i \in \mathbb{N}$ where $\{x_k\}_{k\in K} := \{x_{k_i}\}_{i\in\mathbb{N}}$.

Suppose that $dm_k \geq \bar{d} > 0$ for every $k \in K$. As shown in the proof of Theorem 4.2, for every $k \geq \bar{k}(\omega)$ we have

$$f(x_{k+1}) \leq f(x_k) + 2\tilde{e}_{N_k} + \varepsilon_k - \eta dm_k \text{ a.s.} \tag{24}$$

Without loss of generality, we can assume that $K \subseteq \{\bar{k}+1, \bar{k}+2, ...\}$. Define $s_i := k_{i+1} - k_i$. Then it holds for every $i \in \mathbb{N}$ that

$$f(x_{k_{i+1}}) \leq f(x_{k_i}) + \sum_{j=0}^{s_i-1}(2\tilde{e}_{N_{k_i+j}} + \varepsilon_{k_i+j}) - \eta \sum_{j=0}^{s_i-1} dm_{k_i+j} \quad \text{a.s.}$$

Since $dm_k \geq 0$ for every k, $dm_{k_i} \geq \bar{d}$ and $s_i \leq b$, we obtain that that for every $i \in \mathbb{N}$

$$f(x_{k_{i+1}}) \leq f(x_{k_i}) + \sum_{j=0}^{b-1}(2\tilde{e}_{N_{k_i+j}} + \varepsilon_{k_i+j}) - \eta\bar{d} \quad \text{a.s.}$$

Now, letting i tend to infinity and using the fact that

$$\lim_{i\to\infty} \sum_{j=0}^{b-1}(2\tilde{e}_{N_{k_i+j}} + \varepsilon_{k_i+j}) = 0,$$

we obtain

$$f(x^*) \leq f(x^*) - \eta\bar{d} < f(x^*) \quad \text{a.s.}$$

So, we conclude that a.s. there exists $K_1 \subseteq K$ such that

$$0 = \lim_{k\in K_1} dm_k = \lim_{k\in K_1} \lambda_k \nabla f_{\mathcal{N}_k}^T(x_k) p_k.$$

Moreover, Lemma 3.1 implies the descent property of p_k

$$p_k^T \nabla f_{\mathcal{N}_k}(x_k) \leq -\frac{1}{\alpha_{\max}} \|p_k\|^2 < 0, \quad k \in \mathbb{N}.$$

Therefore, if $\lambda_k \geq \bar{\lambda} > 0$ for all $k \in K_1$ we obtain $\lim_{k\in K_1} p_k = 0$ and the statement follows by Lemma 4.1. Else, suppose that there exists $K_2 \subseteq K_1$ such that

$$\lim_{k\in K_2} \lambda_k = 0.$$

Using the Mean Value Theorem together with the descent property of the search direction and following the proof of Theorem 4.2, we obtain the existence of $K_3 \subseteq K_2$ such that $\lim_{k \in K_3} p_k = 0$, which completes the proof. □

Remark 2. A few words are due here in order to relate the result of the above theorem with the existing ones. Clearly, the definition of strictly strong accumulation point is not common in the deterministic optimization. However it appears to be necessary in the context of almost sure convergence if one wants to avoid conditions on growth of the sample sizes. The result we obtained here is analogous to the results for unconstrained case presented in Yan and Mukai [35]. Comparing with the results presented in Wardi [33] there is a clear trade-off, either one obtains weaker convergence, in upper mid density, or imposes the assumption of strictly strong accumulation points. Proving the convergence in upper density for the method we consider here seems to be possible although technically demanding. Another possibility would be to assume that $\lim_{k \to \infty} \lambda_k p_k = 0$ as in Shapiro et al. [30]. In that case the corresponding result, convergence w.p.1, follows along the same ideas as in [33]. Given that imposing the growth condition might cause very rapid increase in N_k and thus make the optimization procedure more expensive we believe that the conditions of Theorem 4.3 represent a good balance between theoretical and practical issues. Nevertheless, we state the following analysis for completeness.

Theorem 4.4. *Assume that A1-A6 hold. For almost every sequence $\{x_k\}_{k \in \mathbb{N}}$ generated by Algorithm 1, every accumulation point of the sequence $\{x_k\}_{k \in \mathbb{N}}$ is stationary for (1) if the following holds*

$$\sum_{k=0}^{\infty} \tilde{e}_{N_k} < \infty, \tag{25}$$

where $\tilde{e}_{N_k} = \sup_{x \in \Omega} e(x, N_k)$.

Proof. As in Theorem 4.3 we obtain that (24) holds and thus the following holds for every $k \geq \bar{k}(\omega)$

$$f(x_k) \leq f(x_{\bar{k}}) + \sum_{j=\bar{k}}^{k-1} (2\tilde{e}_{N_j} + \varepsilon_j) - \eta \sum_{j=\bar{k}}^{k-1} dm_j \quad \text{a.s.}$$

As f is bounded and $2\tilde{e}_{N_j} + \varepsilon_j$ is summable, we obtain that

$$\lim_{k\to\infty} dm_k = 0 \quad \text{a.s.}$$

Let x^* be an arbitrary accumulation point of the sequence $\{x_k\}_{k\in\mathbb{N}}$, that is let K_0 be a set of positive integers such that $\lim_{k\in K_0} x_k = x^*$. Following the steps from the proof of Theorem 4.2 we conclude that the existence of $K_1 \subseteq K_0$ such that $p_k \geq p > 0$ for every $k \in K_1$ is impossible. Therefore, $\lim_{k\in K_0} p_k = 0$ a.s. which together with Lemma 4.1 implies that x^* is a stationary point a.s. This completes the proof. □

If we consider the bound (13), the assumption (25) is true if we take $N_k \geq e^k$ for instance. So, in order to obtain the stronger convergence result, one can use Algorithm 1 for finitely many iterations and then switch to exponential growth of the sample size. This kind of hybrid approach would combine adaptive sample update and the conservative exponential update which provides better theoretical results. It would be interesting to find out an optimal switching point i.e. an iteration k_0 or the sample size N_{k_0} after which the sample size should grow exponentially.

5 Numerical results

In this section we report preliminary numerical results. The test collection consists of 5 academic problems and the M/M/1 problem. The experiments are designed to investigate the efficiency of the variable sample size (VSS) scheme proposed in Algorithm 2 as well as the properties of Spectral Projected Gradient method in stochastic environment. Thus the VSS is compared with four other sample size schemes combined with the SPG method.

The setup for testing is defined as follows.

$$\|P_\Omega(x_k - g_k) - x_k\| \leq \epsilon_1 \quad \text{and} \quad \frac{e(x_k, N_k)}{\max\{|f_{\mathcal{N}_k}(x_k)|, 1\}} \leq \epsilon_2 \tag{26}$$

are satisfied for some k within at most 10^7 function evaluations. In other words, x_k is an approximate stationary point of $\min_{x\in\Omega} f_{\mathcal{N}_k}(x)$ with the tolerance ϵ_1 and the relative/absolute error estimate of an approximation $f_{\mathcal{N}_k}(x_k) \approx f(x_k)$ is at most ϵ_2. The counting of function evaluations includes counting each gradient ∇F evaluation as n evaluations of f. Since the error bound (13) is considered as too conservative for practical implementations (Bastin [2]), we employ the sample variance

$$\hat{\sigma}^2(x_k, N_k) = \frac{1}{N_k - 1} \sum_{i \in \mathcal{N}_k} \left(F(x_k, \xi^i) - f_{\mathcal{N}_k}(x_k) \right)^2$$

and set

$$\nu(x_k, N_k) = e(x_k, N_k) = 1.96 \frac{\hat{\sigma}(x_k, N_k)}{\sqrt{N_k}}$$

as the precision measure. Function γ defined in Assumption A5 is set to $\gamma(N_k) = \exp(-1/N_k)$. The sequence $\{\varepsilon_k\}_{k \in \mathbb{N}}$ is defined as

$$\varepsilon_0 = \max\{1, |f_{\mathcal{N}_0}(x_0)|\}, \ \varepsilon_{k+1} = \varepsilon_0 k^{-1.1}.$$

The rest of the parameters needed in Algorithm 1 are $\beta = 0.5$, $\eta = 10^{-4}$, $\alpha_{min} = 10^{-8}$ and $\alpha_{max} = 10^8$.

The proposed method terminates either because the number of function evaluations reaches 10^7 or because (26) is satisfied for some finite N_k and x_k. Either way, it terminates with some finite sample size. Let us denote this sample size by N_{max}. VSS method is compared with four other sample size update schemes referred to as HEUR1, HEUR2, HEUR3 and SAA. HEUR1 uses update $N_{k+1} = \min\{\lceil 1.1 N_k \rceil, N_{max}\}$, HEUR2 takes $N_{k+1} = \min\{\lceil e^k \rceil, N_{max}\}$, HEUR3 approximately solves the sequence of problems (2) where the sample size $|\mathcal{N}|$ is increased by 10% until it reaches N_{max} and SAA takes $N_k = N_{max}$ for all k. The initial sample size is $N_0 = 3$ for all tested problems and N_{max} heavily depends on the problem and the variance level. In order to make the comparison fair, the sample generated within VSS runs is used for the other tested schemes as well. The same is true for the starting point. We performed 10 independent runs for each tested problem. The question of suitable $N_{\max}$ is far from simple and there is a number of paper that deals with this issues, for example Bayraksan and Morton [5], Homem-de-Mello and Bayraksan [6]. Given that the principal goal of this paper is not the quality of solution for a finite $N_{\max}$ but the almost sure convergence for unbounded sample, we adopted this simple rule for $N_{\max}$ in our tests.

The tested schemes HEUR1 and SAA have been used in our previous work where the sample size was finite and unconstrained optimization problems were considered. HEUR1 was motivated by Friedlander and Schmidt [13], although this is not the only paper that suggests this kind of scheduling sequence. Since the SPG method is linearly convergent, Pasupathy [27] also suggests the 10% growth of the sample size. However, [27] uses the so called diagonalization where a sequence of optimization problems is solved with the precision $q_k = K/\sqrt{Nk}$. We implemented this scheme (HEUR3) as well, with $K = \epsilon_1 \sqrt{N_{max}}$ to attain the same final precision, but the results were not good since it did not converge within 10^7 function evaluations in a vast majority of the tested problems. The remaining scheme HEUR2 was motivated by the result given in Theorem 4.4 and some other related results ([6] for instance) where the minimal growth rate for the sample size is given in order to provide stronger convergence result. Moreover, notice that VSS may exceed the final sample size within the optimization process which gives an advantage to the other tested schemes. Finally, one can always find a heuristic which will perform better on a particular problem than the scheme proposed in this paper. But the main advantage of VSS is its adaptive nature and inner mechanism which works for an arbitrary problem. Tuning the parameters is clearly problem dependent and a generic approach is tested here.

The test examples are defined as

$$F(x,\xi) = h(\xi x), \quad \xi : \mathcal{N}(1,\sigma^2),$$

where $h : \mathbb{R}^n \to \mathbb{R}$. Two levels of variance are tested, $\sigma^2 = 0.1$ and $\sigma^2 = 1$. Regarding the constraints, two cases are considered with respect to the solution - if constraints are active at the solution then we denote the feasible set is denoted by Ω_a, and if the constraints are inactive, the feasible set is Ω_{ia}. The dimension of all examples is $n = 10$ and the feasible set is n-dimensional box of the form $[l,u]^n$. Starting points are chosen randomly within a feasible set and the stopping criterion parameters are $\epsilon_1 = 10^{-2}$ and $\epsilon_2 = 0.05$.

Functions h are originally taken from Montaz Ali et al. [25]. We list the problems together with active/inactive case constraints, the range of the sample size N_{max} and the mean values $\bar{N}_{\max}$ in all the runs.

P1 Exponential problem

$$F(x,\xi) = e^{-0.5\|\xi x\|^2},$$

$$\Omega_a = [0.3, 0.5]^{10},\ \Omega_{ia} = [-1,1]^{10}, \quad N_{max} \in [3, 208],\ \bar{N}_{\max} = 97.$$

P2 Griewank problem

$$F(x,\xi) = 1 + \frac{1}{4000}\|\xi x\|^2 - \prod_{i=1}^{10} \cos\left(\frac{\xi x_i}{\sqrt{i}}\right),$$

$$\Omega_a = [100, 200]^{10},\ \Omega_{ia} = [-600, 600]^{10}, \quad N_{max} \in [3, 25731],\ \bar{N}_{\max} = 2800.$$

P3 Neumaier 3 problem

$$F(x, \xi) = \sum_{i=1}^{10} (\xi x_i - 1)^2 - \sum_{i=2}^{10} \xi x_i \xi x_{i-1},$$

$$\Omega_a = [0, 10]^{10},\ \Omega_{ia} = [-100, 100]^{10}, \quad N_{max} \in [3, 3274],\ \bar{N}_{\max} = 1301.$$

P4 Salomon problem

$$F(x, \xi) = 1 - \cos(2\pi \|\xi x\|^2) + 0.1\|\xi x\|^2,$$

$$\Omega_a = [10, 50]^{10},\ \Omega_{ia} = [-100, 100]^{10}, \quad N_{max} \in [3, 3651],\ \bar{N}_{\max} = 1312.$$

P5 Sinusoidal problem

$$F(x, \xi) = -2.5 \prod_{i=1}^{10} \sin(\xi x_i - 30) - \prod_{i=1}^{10} \sin(5(\xi x_i - 30)),$$

$$\Omega_a = [0, 2]^{10},\ \Omega_{ia} = [0, 180]^{10}, \quad N_{max} \in [3, 1440],\ \bar{N}_{\max} = 205.$$

M/M/1 queueing problem is often used in stochastic optimization for illustration of real world problems, [18, 1]. This is a parameter estimation problem. It aims to estimate a distribution parameter that minimizes the objective function which takes into account the expected number of customers in a queue and the expected service time given by the mean value of the Exponential distribution. The goal is to minimize the expected time that a client spends in a queue (waiting time and service time) taking into account the (nonzero) costs associated to (nonzero) waiting and service times The solution mainly serves for finding the optimal service time in a sense that it gives an information weather the service time should be improved. In this approach it is important to notice that some balance between the server cost with the above stated objective should exist. This problem can be solved analytically and for that particular reason it makes a good example for evaluating the stochastic procedures. The expected steady-state number of customers in a queue with arrival rate 1 and mean service time x is denoted by $L(x)$. More detailed description of queueing-type problems, as well as other possible approaches can be found in Andradottir [1] or Law [23] for example.

We consider a special two dimensional case (two queues) where the cost induced by the service improvement is addressed within the following objective function

$$\min_{x\in(0,1)\times(0,1)} f(x) = \frac{1}{x_1} + \frac{1}{x_2} + \frac{10}{x_1 x_2} + L(x_1) + L(x_2),$$

where $L(x_i)$ is mathematical expectation of a random variable that follows Geometrical distribution with parameter $1 - x_i$, that is,

$$L(x_i) = E\left(X(x_i)\right), \quad \mathcal{P}\left(X(x_i) = k\right) = x_i^k(1 - x_i),\ k = 0, 1, 2, \ldots.$$

Therefore, $L(x_i) = x_i/(1 - x_i)$ and the analytical solution of the problem is known: $x^* = (0.787, 0.787)^T$ with $f(x^*) = 26.0764$. Geometrically distributed random variable can be generated with Uniform distribution as

$$X(x_i) = \left\lceil \left| \frac{\ln \xi}{\ln x_i} \right| - 1 \right\rceil, \quad \xi : \mathcal{U}(0, 1).$$

We define function F from (1) by

$$F(x, \xi) = \frac{1}{x_1} + \frac{1}{x_2} + \frac{10}{x_1 x_2} + \left\lceil \left| \frac{\ln \xi}{\ln x_1} \right| - 1 \right\rceil + \left\lceil \left| \frac{\ln \xi}{\ln x_2} \right| - 1 \right\rceil.$$

As suggested in Kao et al. [18], finite differences (Fu [14]) are employed to approximate the gradient. More precisely $g(x, \xi) \approx \nabla F(x, \xi)$ with $g(x, \xi) = (g_1(x, \xi), g_2(x, \xi))^T$ and

$$g_1(x, \xi) = -\frac{1}{x_1^2} - \frac{10}{x_1^2 x_2} + \frac{1}{h}\left(\left\lceil \left| \frac{\ln(\xi)}{\ln(x_1 + h)} \right| - 1 \right\rceil - \left\lceil \left| \frac{\ln \xi}{\ln x_1} \right| - 1 \right\rceil \right),$$

$$g_2(x, \xi) = -\frac{1}{x_2^2} - \frac{10}{x_2^2 x_1} + \frac{1}{h}\left(\left\lceil \left| \frac{\ln(\xi)}{\ln(x_2 + h)} \right| - 1 \right\rceil - \left\lceil \left| \frac{\ln \xi}{\ln x_2} \right| - 1 \right\rceil \right).$$

In order to avoid singularities and obtain closed feasible set, we define $\Omega = [0 + \epsilon_3, 1 - \epsilon_3]^2$ and use $h = 10^{-2}$, $\epsilon_3 = 0.05$. The starting point is $x_0 = (0.1, 0.1)^T$ and the parameters from (26) are $\epsilon_1 = 10^{-1}$ and $\epsilon_2 = 10^{-2}$.

The results are mainly presented through performance profile graphs (Dolan and Moré [12]) using the number of function evaluations as the cost function. Roughly speaking, performance profile states the probability that the considered method close enough to the best method among all tested methods, where closeness is measured by α presented on x-axes. Specifically, $\alpha = 1$ gives us information about the percentage of the tested problems in which the considered method performance was the best one in the considered metrics.

The results for the problems P1-P5 are presented in Figures 1 and 2. Only the runs with at least one successful method are included, i.e. 82% of all the runs performed on these problems. The rest of runs are considered unsuccessful since none of the tested methods converged within 10^7 function evaluations. However, the resulting average optimality measure is of order 10^{-1} and the failure is mainly due to budget constraint (10^7 FEVs).

Figure 1 shows that VSS outperforms all other methods if one considers the overall results. It exhibits particularly good behavior for problems with solutions on the boundary of feasible set. The worst performance of VSS is observed for higher variance and inactive constraints as in these cases it is outperformed by the exponential growth sample scheme. Clearly, the behavior of the tested methods heavily depends on the problem - on its structure as well as on the objective function form which determines the variance of F. Heuristic approaches performed well when the solution is inside the feasible set, but the overall results (Figure 2) show that the usage of VSS is more efficient in terms of number of function evaluations than the other tested schemes.

In Figure 3 we present the sample scheduling for a randomly chosen run, showing the adaptive nature of the scheduling as the sequence of sample sizes is clearly nonmonotone but eventually growing.

The analytical solution of the M/M/1 queueing problem is known. Figure 4 plots the relative error of the form $(f(x_k) - f(x^*))/f(x^*)$ against the number of function evaluations FEV. The average values of all 10 runs are shown. The figure shows that VSS is highly competitive with the other tested schemes in the a budget framework. SAA and HEUR3 were too expensive and therefore their graphics are not visible on this figure. All the tested runs were successful for M/M/1 queueing problem. Average sample size $\bar{N}_{\text{max}}$ for the queuing problem is 3917 and $\bar{N}_{max} \in [3782, 4108]$. The mean value of the objective function $f(x)$ at the final iteration among these 10 runs is approximately 26.108 for all tested methods and the centered sample variance of $f(x^*)$ is of order 10^{-4}. Therefore, all the methods yield solutions of practically the same quality.

Finally, Figure 5 presents common performance profile for P1-P5 problems and M/M/1 queueing problem and confirms that VSS remains the most effective.

6 Conclusions

The method we propose and analyze in this paper consists of two components. An efficient sample scheduling update based on the progress achieved in the current iteration with the current SAA approximate function and the precision of SAA approximation, is coupled with the SPG method. A nonmonotone line search is considered as the SPG behaves much better if some nonmonotonicity is allowed. It is assumed that the feasible set is easy to project on and therefore the principal advantages of the SPG method, efficiency and simplicity, yielded a fast and reliable method for solving the constrained problems with the objective function in the form of mathematical expectation. The sample size is pushed to infinity and the almost sure convergence is proved under a set of appropriate assumptions. No growth condition on the sample sizes is assumed what is particularly important from the practical point of view, as the fast increase in the sample size very often yields an expensive method. The set of assumptions is compatible with the corresponding results for the unconstrained case. The assumption of strictly strong accumulation point yields almost sure convergence. Apparently this assumption is necessary if one wants to avoid the growth condition.

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

References

[1] S. Andradottir, A scaled stochastic approximation algorithm, *Management Science 42(4), (1996), pp. 475-498.*

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

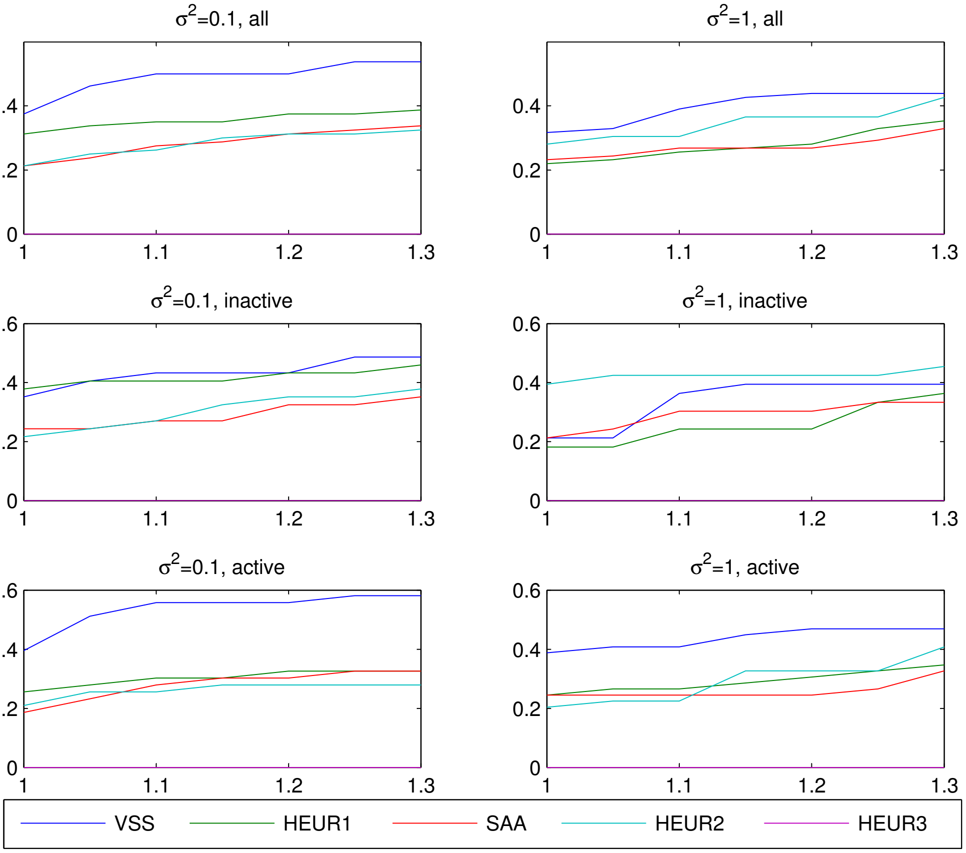

Figure 1: Examples P1-P5. The graphs on the left present profiles for $\sigma^2 = 0.1$ while the graphs on the right present profiles for $\sigma^2 = 1$. The pair of graphs on the top represents the profiles for both constraints activity (active and inactive at the solution), the pair in the middle are the profiles for inactive constraints while the bottom pair contains the profiles for the case of active constraints

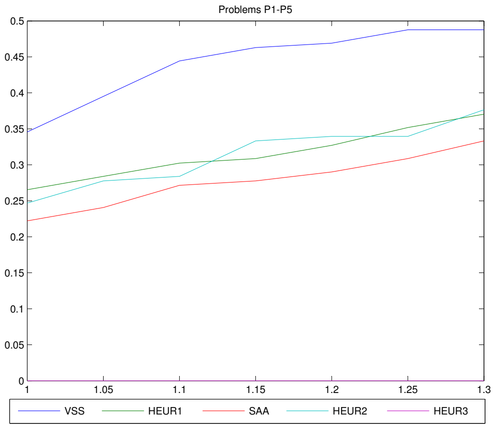

Figure 2: Problems P1-P5, all noise levels, active and inactive constraints

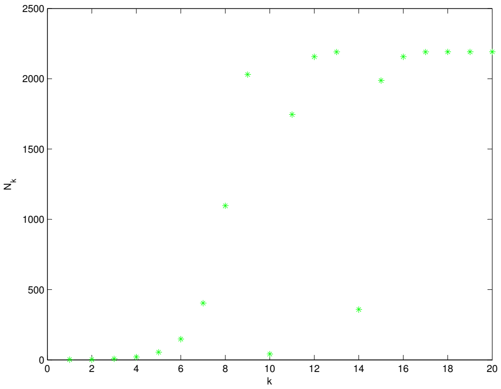

Figure 3: Sample sizes for one run of VSS method

[3] 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.*

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

[5] G. Bayraksan, D.P. Morton, A sequential sampling procedure for stochastic programming, *Operational Research 59 (2011) pp. 898-913.*

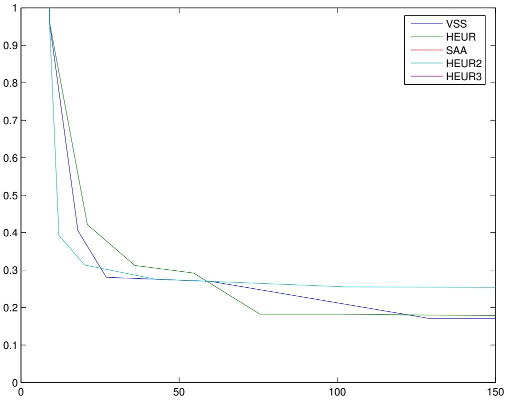

Figure 4: M/M/1 problem, all methods

[6] T. Homem-de-Mello, G. Bayraksan, Monte Carlo sampling-based methods for stochastic optimization, *Surveys in Operations Research and Management Science 19 (2014), pp. 56-85.*

[7] E.G., Birgin, J.M. Martínez, M. Raydan, Nonmonotone Spectral Projected Gradients on Convex Sets, *SIAM Journal on Optimization. 10 (2000) pp. 1196-1211.*

[8] E.G., Birgin, J.M. Martínez, M. Raydan, Spectral Projected Gradient methods: Review and Perspectives, *Journal of Statistical Software 60 (3), (2014).*

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

[10] R. Byrd, G. Chin, W. Neveitt, J. Nocedal, Sample size selection in Optimization Methods for Machine Learning, *Mathematical Programming 134(1) (2012) pp. 127-155.*

[11] R.H. Byrd, S.L. Hansen, J. Nocedal, Y.Singer, A stochastic Quasi-Newton method for large scale optimization, *arxiv.org/abs/1401.7020).*

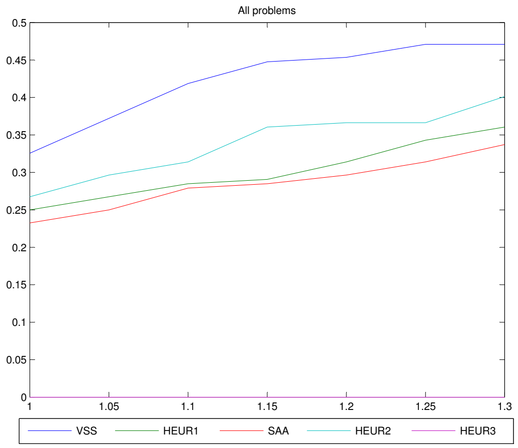

Figure 5: All problems, all methods

[12] E. D. Dolan, J. J. Moré, Benchmarking optimization software with performance profiles, *Mathematical Programming Ser. A 91 (2002), pp. 201-213.*

[13] M. P. Friedlander, M. Schmidt, Hybrid deterministic-stochastic methods for data fitting, *SIAM Journal on Scientific Computing 34(3) (2012), pp. 1380-1405.*

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

[15] L. Grippo, F. Lampariello, S. Lucidi, A nonmonotone line search technique for Newton's method, *SIAM Journal on Numerical Analysis 23(4) (1986), pp. 707-716.*

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

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

[18] C. Kao, W. T. Song, S. Chen, A modified Quasi-Newton Method for Optimization in Simulation, *International Transactions on Operations Research 4(3) (1997), pp. 223-233.*

[19] N. Krejić, Z. Lužanin, Z. Ovcin, I. Stojkovska, Descent direction method with line search for unconstrained optimization in noisy environment, *Optimization Methods and Software 30(6) (2015), pp. 1164-1184.*

[20] 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.*

[21] N. Krejić, N. Krklec Jerinkić, Nonmonotone line search methods with variable sample size, *Numerical Algorithms 68 (2015), pp. 711-739.*

[22] N. Krejić and J. M. Martínez, Inexact Restoration approach for minimization with inexact evaluation of the objective function, *Mathematics of Computation 85, 300 (2016), 1775-1791.*

[23] A.M. Law, Simulation Modeling and Analysis, *McGraw-Hill Education, 2014.*

[24] A. Mokhtary, A. Ribeiro, Global Convergence of Online Limited Memory BFGS, *Journal of Machine Learning Research 16 (2015) 3151-3181.*

[25] M. Montaz Ali, C. Khompatraporn, Z. B. Zabinsky, A Numerical Evaluation of Several Stochastic Algorithms on Selected Continous Global Optimization Test Problems, *Journal of Global Optimization 31(4) (2005), pp.635-672 .*

[26] 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), pp. 181-201.*

[27] R. Pasupathy, On Choosing Parameters in Retrospective-Approximation Algorithms for Stochastic Root Finding and Simulation Optimization, *Operations Research 58(4) (2010), pp. 889-901.*

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

[29] J. O. Royset, Optimality functions in stochastic programming, *Mathematical Programming 135(1-2) (2012), pp. 293-321.*

[30] A. Shapiro, D. Dentcheva, A. Ruszczynski, Lectures on stochastic programming: Modeling and theory, *MPS/SIAM Series on Optimization 9, 2009.*

[31] A. Shapiro, Y. Wardi, Convergence analysis of gradient descent stochastic algorithms, *Journal of Optimization Theory and Applications, 91(2) (1996), pp. 439-454.*

[32] J. C. Spall, Introduction to Stochastic Search and Optimization, *Wiley-Interscience Serises in Discrete Mathematics, New Jersey, 2003.*

[33] Y. Wardi, Stochastic Algorithms with Armijo Stepsizes for Minimization of Functions, *Journal of Optimization Theory and Applications 64 (1990), 399-417.*

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

[35] D. Yan, H. Mukai, Optimization Algorithm with Probabilistic Estimation *Journal of Optimization Theory and Applications 64, 79(2) (1993), pp. 345-371.*