Spectral projected gradient method for stochastic optimization

Nataša Krejić* Nataša Krklec Jerinkić *

July 1, 2015

Abstract

We consider the Spectral Projected Gradient method for solving constrained optimization porblems 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 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. 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 and Science, 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 analytical expression of f is rarely available 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, but makes the approximate problem more expensive as function evaluations depend on the sample size. There are many possibilities for the sample choice, depending on properties of the underlying random variable ξ and availability of data, see [8, 12, 24, 26, 28].

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. 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 [26, 28]. 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 i.e. 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 and such procedure results in a more efficient method than the corresponding procedure with strictly increasing schedule sequence, [2, 3, 19, 20, 21]. The results presented in [11] 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 [29]. An Armijo type line search method is combined with SAA approach. The convergence is proved with upper zero density. An extension of [29] for the unconstrained case is presented in [31], where the adaptive precision is proposed i.e. the sequence $\{N_k\}_{k\in\mathbb{N}}$ is not determined in advance as in [29] 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 [27]. 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.

The second approach, often called the Surface Response Method, is again a two step procedure. It consists of a sequence of 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 procedures is the number of stages (i.e. the number of different SAA problems to be solved) and the precision of each approximate solution. For

further details one can see [23, 24, 25].

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 [5, 6] is a well known for its efficiency and simplicity. In this paper we consider application of the SPG method to the case of SAA approximate problems (2) coupled with a suitable sample scheduling scheme. Thus the principal aims of this paper are: a) to derive an efficient sample scheduling that will yield an 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.

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.

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 closed 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.

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

The assumptions A1-A2 imply that $F(x,\xi), \nabla F(x,\xi), f(x)$ and $\nabla f(x)$ are bounded for $x \in \Omega$ and ξ from $(\mathcal{A}, \mathcal{F}, \mathcal{P})$. Furthermore, the equality

$$E[\nabla F(x,\xi)] = \nabla E[F(x,\xi)]$$

holds and thus the sample average approximation of $\nabla f(x)$ is given by

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

Both $f_{\mathcal{N}}(x)$ and $\nabla f_{\mathcal{N}}(x)$ are bounded and uniformly continuous on Ω.

The assumptions A1-A3 together imply that the SAA gradients uniformly converge to the true gradient value with probability 1, i.e.

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

Let us now define the following two functions that will be used later on to define the sample update.

Assumption A4. Assume that $e : \mathbb{R}^n \times \mathbb{N} \to \mathbb{R}_+$ is a function with the following properties

$$\lim_{N\to\infty} \sup_{x\in\Omega} e(x,N) = 0, \tag{4}$$

$$|f(x) - f_{\mathcal{N}}(x)| \leq e(x,N) \ a.s. \ x \in \Omega, N \in \mathbb{N}, \tag{5}$$

and for any finite valued N there exists e_N such that

$$e(x,N) \geq e_N > 0 \quad \text{for every} \quad x \in \Omega.$$

One possible choice is

$$e(x,N) = C\sqrt{\frac{\ln(\ln(N))}{N}}, \tag{6}$$

where C is a positive constant. This error bound is derived in [15] (Proposition 3.5) for cumulative samples. However, it is considered as too conservative from the practical point of view and it is often approximated with a bound of sample variance type. The error function $e(x_k, N_k)$ 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

$$\lim_{N\to\infty} \gamma(N) = 1. \tag{7}$$

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 $P_\Omega(\cdot)$, i.e.

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

where the norm is assumed to be Euclidian.

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 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

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

Algorithm 1

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

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

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

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 $dm_k = e(x_k, N_k)$ set $N_k^+ = N$.

Step 1.2 If $dm_k > e(x_k, N_k)$

While $dm_k > e(x_k, N)$ and $N > N_k^{\min}$ set

$$N = N - 1.$$

End(While).

Set $N_k^+ = N$.

Step 1.3 If $dm_k < e(x_k, N_k)$

While $dm_k < e(x_k, N)$ set

$$N = N + 1.$$

End(While).

Set $N_k^+ = N$.

Step 2. Update of N_{k+1}.

If $N_k^+ < N_k$ and

$$\rho_k = \left| \frac{f_{\mathcal{N}_k^+}(x_k) - f_{\mathcal{N}_k^+}(x_{k+1})}{f_{\mathcal{N}_k}(x_k) - f_{\mathcal{N}_k}(x_{k+1})} - 1 \right| \geq \frac{N_k - N_k^+}{N_k}$$

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

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

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 $N_{k+1}^{\min} = N_k^{\min}$.

In order to comment Algorithm 1, we state the following important result from [5].

Lemma 3.1. *[5] 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).

If we assume $\alpha_{max} \geq 1$, 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

$$\|P_\Omega(x_k - \nabla f_{\mathcal{N}}(x_k)) - x_k\| = 0, \ N \geq N_k \tag{8}$$

or the algorithm generates an infinite sequence $\{x_k\}$. If x_k is such that (8) holds, then (3) implies

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

and therefore, $\|P(x_k - \nabla f(x_k)) - x_k\| = 0$, i.e. x_k is a stationary point for (1). So, from now on we assume that (8) 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, [22, 5, 6]. It is important to state here that any other nonmonotone rule like the rules considered in [5, 13, 14, 20, 30] 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 [7, 8, 9, 17]. In the deterministic case y_k satisfies

$$y_k = \left(\int_0^1 \nabla^2 f(x_{k+1} + ts_k)dt\right) s_k.$$

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 $\int_0^1 \nabla^2 f(x_k + ts_k)dt$, [6]. 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.

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 allow 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 $e(x_k, N_k)$. 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 $e(x_k, N)$ 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 $e(x_k, N)$. If $N_k^+ < N_k$ i.e. if a decrease of the sample size is proposed, we perform an additional check stated in Step 2 to avoid possibly unproductive decreases. The second sequence $N_k^{\min}$ is updated in Step 3 and it is clearly nondecreasing. It represent 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 SAA problem with finite N, as presented in [19, 20] for unconstrained problems. The first idea of this kind is developed in [2, 3, 4] for the trust region approach and SAA methods.

4 Convergence theory

The fact that the sample size N_k goes to infinity is not obvious in Algorithm 2 and it is proved in the next theorem.

Theorem 4.1. *Assume that A1-A3 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

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

by Step 3 of Algorithm 1. Thus

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

for arbitrary $m \in \mathbb{N}$. Given that $f_{\bar{\mathcal{N}}_0}$ is bounded from below and $0 < \sum_{k=0}^{\infty} \varepsilon_k < \infty$, the last inequality yields

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

On the other hand, for each $k \geq \bar{k}_0$ we have

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

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

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

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 have the statement.

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, Step 1 of Algorithm 1 can happen at most 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

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

Assume that the statement of this lemma is not true so there exists an infinite sequence of $\{N_k\}_{k\in\mathbb{N}}$ such that its elements are bounded. Then there must exist an infinite sequence $K_0 = \{k \geq \bar{k}_2 : N_{k+1} = \bar{N}_1\}$, for some $\bar{N}_1$. Since the sequence $\{N_k\}$ is not 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 sequence $K_1 = \{k \geq \bar{k}_2 : N_k \neq N_{k+1} = \bar{N}_1\}$. 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. Therefore, for every $k \in K_2$

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

where $h(k) \in K_2$, except for the first element in K_2. Notice that $k+1-h(k) > 1$. So for $k_j \in K_2$, the inequality

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

holds with possible exception for $k_j = 1$. Given that

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

(10) implies that $f_{\bar{\mathcal{N}}_1}$ is unbounded on K_2 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.

Lemma 4.1. *Assume that A1-A3 hold and that $K \subset \mathbb{N}$ is such that*

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

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

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. □

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 almost surely.

Theorem 4.2. *Assume that A1-A5 hold and let $\{x_k\}_{k\in\mathbb{N}}$ be a sequence generated by Algorithm 1. Then there exist an accumulation point of $\{x_k\}_{k\in\mathbb{N}}$ which is stationary for (1) almost surely.*

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 $k \in \mathbb{N}$

$$\|p_k\|^2 \geq p > 0. \tag{11}$$

Then, Lemma 3.1 implies the following inequalities

$$p_k^T \nabla f_{\mathcal{N}_k}(x_k) \leq -\frac{1}{\alpha_{\max}} \|p_k\|^2 \leq -\frac{1}{\alpha_{\max}} p := -\bar{p} < 0, \ k \in \mathbb{N} \tag{12}$$

and

$$dm_k = -\lambda_k p_k^T \nabla f_{\mathcal{N}_k}(x_k) \geq \lambda_k \bar{p}, \quad k \in \mathbb{N}. \tag{13}$$

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

$$dm_k \geq \bar{\lambda}\bar{p} := \bar{d} > 0, \quad k \in \mathbb{N}. \tag{14}$$

Define

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

Such $\tilde{e}_N$ exists due to (4) and the definition of $e(x, N)$. Given that $\{x_k\}_{k\in\mathbb{N}} \subset \Omega$, $N_k \to \infty$ and A4 holds, we have

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

and

$$\lim_{k\to\infty} \tilde{e}_{N_k} = 0. \tag{15}$$

Thus, for every $k \in \mathbb{N}$ 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, (15) implies that $\tilde{e}_{N_k} < (\eta\bar{d} - q)/2$ for every k large enough, i.e. there exists $\bar{k}$ such that

$$2\tilde{e}_{N_k} < \eta\bar{d} - q \quad \text{for every} \quad k \in \mathbb{N},\ k \geq \bar{k}. \tag{16}$$

Thus, for every $k \geq \bar{k}$

$$f(x_{k+1}) \leq f(x_k) + \varepsilon_k - q,$$

which furthermore implies that

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

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

Now, suppose that there is a subsequence $K_1 \subseteq \mathbb{N}$ such that

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

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

$$p_k^T \nabla f_{\mathcal{N}_k}(x_k + t_k \lambda'_k p_k) \geq \eta p_k^T \nabla f_{\mathcal{N}_k}(x_k). \tag{17}$$

Given that $\{p_k\}$ and $\{x_k\}$ are bounded, there exists $K_2 \subseteq K_1$ 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 (17) we get

$$(p^*)^T \nabla f(x^*) \geq \eta (p^*)^T \nabla f(x^*) \text{ a.s.} \tag{18}$$

The condition $\eta \in (0,1)$ and (18) together yield

$$(p^*)^T \nabla f(x^*) \geq 0 \text{ a.s.} \tag{19}$$

On the other hand, taking limit for $k \in K_2$ in (12) we obtain

$$0 \leq (p^*)^T \nabla f(x^*) \leq -\bar{p} < 0 \text{ a.s.} \tag{20}$$

which is clearly in contradiction with (19). Therefore, we conclude that (11) is wrong and there exists a subsequence of $\{p_k\}_{k \in \mathbb{N}}$ that converges to zero, i.e. there exists $K_3 \subseteq \mathbb{N}$ such that $\lim_{k \in K_3} p_k = 0$. Again, $\{x_k\}$ is bounded and there must exist $K_4 \subseteq K_3$ 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 [31].

Definition 4.1. *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. *Suppose that A1-A5 hold and let $\{x_k\}_{k \in \mathbb{N}}$ be a sequence generated by Algorithm 1. Then every strictly strong accumulation point of the sequence $\{x_k\}_{k \in \mathbb{N}}$ is stationary for (1) almost surely.*

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, in that case the following holds for every $k \in K$

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

This implies that for every $i \in \mathbb{N}$

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

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

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

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 1. 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 [31]. Comparing with the results presented in [29] 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 [26]. In that case the corresponding result, convergence w.p.1, follows along the same ideas as in [29]. 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.

5 Numerical results

In this section we report the results obtained on a set of examples from [19] and on a stochastic optimization problem from queuing theory, the well known M/M/1 problem, see [16]. 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 two other sample size schemes combined with the SPG method.

We assume that the method (approximately) solves the problem if the inequalities

$$\|P_\Omega(x_k - g_k) - x_k\| \leq \epsilon_1 \quad \text{and} \quad \frac{e(x_k, N_k)}{\max\{|f_{\mathcal{N}_k}(x_k)|, 1\}} \leq \epsilon_2 \tag{21}$$

are satisfied for some k within at most 10^7 evaluations of f. 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 (6) is

considered as too conservative for practical implementations [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

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

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 (21) 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 two other sample size update schemes referred to as HEUR and SAA. HEUR uses update $N_{k+1} = \min\{[1.1 N_k], N_{max}\}$ while 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 HEUR and SAA runs as well. The same is true for the starting point. We performed 10 independent runs for each tested problem.

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 $\sigma^2 = 1$. Also, we consider two cases regarding the constraints - active, where the feasible set is denoted by Ω_a, and inactive with the feasible set 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 [18]. 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

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

$$\Omega_a = [0.3, 0.5]^{10},\ \Omega_{ia} = [-1, 1]^{10}, \quad N_{max} \in [3, 210],\ \bar{N}_{\max} = 100.$$

P2 Griewank problem

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

$$\Omega_a = [100, 200]^{10},\ \Omega_{ia} = [-600, 600]^{10}, \quad N_{max} \in [3, 25731],\ \bar{N}_{\max} = 3675.$$

P3 Neumaier 3 problem

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

$$\Omega_a = [0, 10]^{10},\ \Omega_{ia} = [-100, 100]^{10}, \quad N_{max} \in [3, 3274],\ \bar{N}_{\max} = 1257.$$

P4 Salomon problem

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

$$\Omega_a = [10, 50]^{10},\ \Omega_{ia} = [-100, 100]^{10}, \quad N_{max} \in [479, 3012],\ \bar{N}_{\max} = 1520.$$

P5 Sinusoidal problem

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

$$\Omega_a = [0, 2]^{10},\ \Omega_{ia} = [0, 180]^{10}, \quad N_{max} \in [4, 1352],\ \bar{N}_{\max} = 287.$$

M/M/1 queueing problem is often used in stochastic optimization for illustration of real world problems. More detailed description of queueing-type problems can be found in [1]. We consider two-dimensional case of the form

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

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

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

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

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

We define function F from (1) by

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

As suggested in [16], finite differences [12] 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

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

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

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 (21) are $\epsilon_1 = 10^{-1}$ and $\epsilon_2 = 10^{-2}$.

The results are presented through performance profile graphs [10] where the main criterion is the number of function evaluations. The results for the problems P1-P5 are presented in Figure 1, the results for M/M/1 queueing problem are presented in Figure 2 and all the tested problems together are presented in Figure 3. These graphics clearly indicate the efficiency of the proposed VSS method.

Figure 1 shows that VSS clearly outperforms the the remaining two methods in all cases. The behaviour of HEUR and SAA depend on the variance level and the type of a solution. If the constraints are active, smaller variance gives a small advantage to the SAA method but the increased variance

clearly favours the HEUR update. However, if the solution is in an interior of the box, the situation is quite opposite. In general, graphs in a first row indicate that it is advantageous to use HEUR update if a variance is relatively large, while in the case of smaller variance SAA and HEUR perform practically the same. VSS seems to be highly efficient and increases its advantage even more if the variances is smaller. In that case, inactive constraints favor HEUR over SAA but this ordering changes if the solution is on the boundary. If $\sigma^2 = 1$ and the solution is in the interior of Ω SAA and HEUR perform almost the same, however moving solution to the boundary moves SAA below HEUR and gives even more advantage to VSS. Notice that the later conclusion holds regardless of the tested variance level. VSS clearly yields the best performance.

Average sample size $\bar{N}_{\max}$ for the queuing problem is 4294 and $\bar{N}_{max} \in [3651, 6945]$. The mean value of the objective function $f(x)$ at the final iteration among these 10 runs is 26.081 for VSS and 26.082 for the remaining two methods. The centered sample variance of $f(x^*)$ is 2.22 10^{-5}, 3.11 10^{-5} and 2.49 10^{-5} for VSS, HEUR and SAA, respectively. Therefore, all three methods yield solutions of practically the same quality. However, applying VSS scheme seems to be beneficial since it decreases the necessary number of function evaluations significantly.

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 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 uncon-

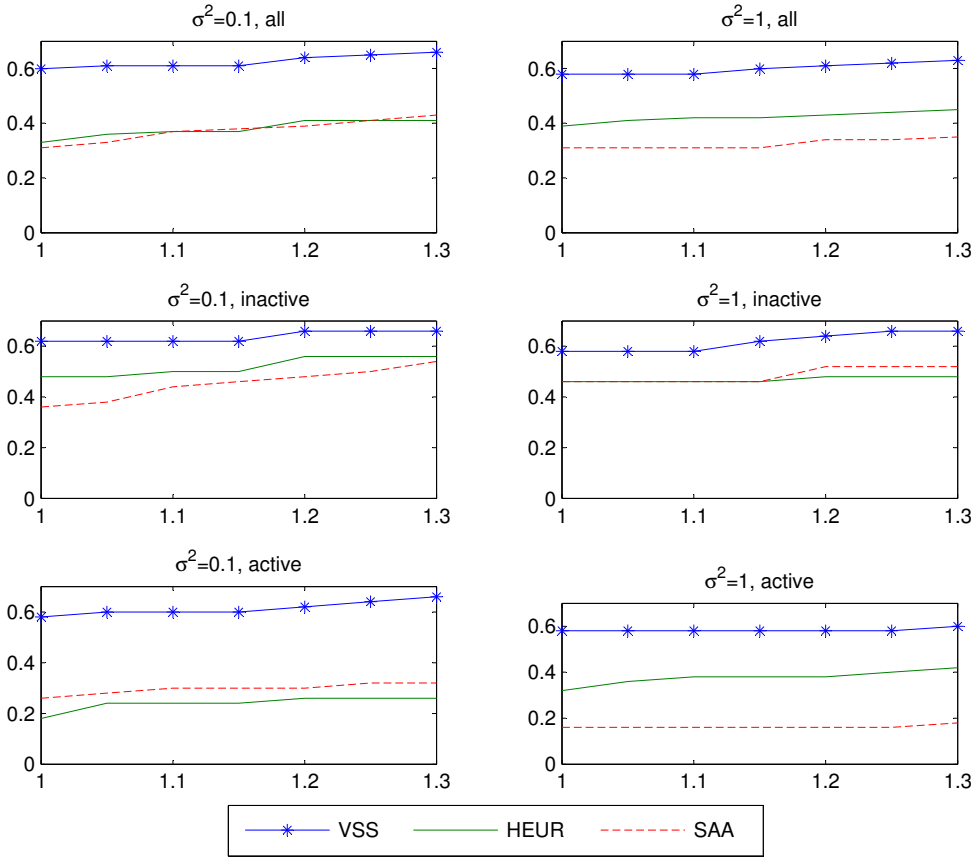

Figure 1: Examples P1-P5

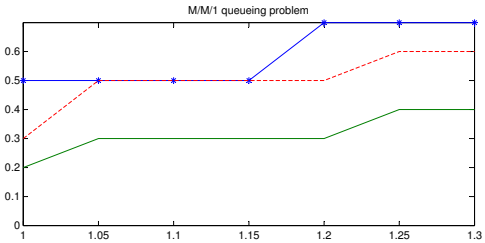

Figure 2: M/M/1 problem

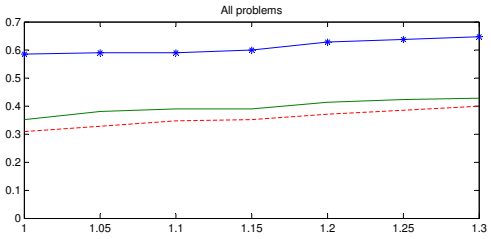

Figure 3: All problems

strained case. The assumption of strictly strong accumulation point yields almost sure convergence. Apparently this assumptions is necessary if one wants to avoid the growth condition.

Numerical results demonstrate the efficiency of this approach. SPG is considered with the sample scheduling proposed in this paper as well as with the sample scheduling generated by a simple heuristics that increases the sample size independently of the SPG method. Furthermore, for large but fixed sample size we also tested the full SAA approach with the SPG. The test examples are both academic and real life problems. It is shown that VSS scheduling generates approximate solutions with less computational effort. It is also clear that VSS is less dependent on the noise level than the other tested methods. An important property of the VSS SPG method is the efficiency in the case of active as well as in the case of inactive constraints.

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.*

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

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

[7] 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.*

[8] 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.*

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

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

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

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

[13] 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.*

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

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

[16] 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.*

[17] N. Krejić, Z. Lužanin, Z. Ovcin, I. Stojkovska, Descent direction method with line search for unconstrained optimization in noisy environment, *Optimization Methods and Software, DOI:10.1080/10556788.2015.1025403* .

[18] 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* .

[19] 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.*

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

[21] N. Krejić and J. M. Martínez, Inexact Restoration approach for minimization with inexact evaluation of the objective function, Mathematics of Computation (to appear).

[22] 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.*

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

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

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

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

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

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

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

[30] 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.*

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