# Stochastic gradient methods for unconstrained optimization

Nataša Krejić* Nataša Krklec Jerinkić *

March 12, 2014

**Abstract**

This papers presents an overview of gradient based methods for minimization of noisy functions. It is assumed that the objective functions is either given with error terms of stochastic nature or given as the mathematical expectation. Such problems arise in the context of simulation based optimization. The focus of this presentation is on the gradient based Stochastic Approximation and Sample Average Approximation methods. The concept of stochastic gradient approximation of the true gradient can be successfully extended to deterministic problems. Methods of this kind are presented for the data fitting and machine learning problems.

**Key words:** unconstrained optimization, stochastic gradient, stochastic approximation, sample average approximation

## 1 Introduction

Stochastic optimization problems appear in all areas or engineering, physical and social sciences. Typical applications are model fitting, parameter estimation, experimental design, performance evaluation etc. The models we are considering here can be written in the form

$$\min_{x\in\Omega} f(x) \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is either observed with noise or is defined as the mathematical expectation. In fact the objective function depends on a vector of random variables $\xi$ from some probability space that might be know or unknown, depending on application. Thus the exact evaluation of $f(x)$ is impossible to evaluate and it is necessary to use simulation to estimate the objective function value.

The feasible set $\Omega$ can be defined by constraints of different types - simple box constraints, deterministic constraints, chance constraints, constraints in the

*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

form of mathematical expectation etc. In this paper we consider only the case $\Omega = \mathbb{R}^n$. More precisely, we consider only two types of stochastic problems. The first type are problems with random objective function,

$$\min_x F(x, \xi),$$

where $\xi$ represents the noise (or randomness) and $x$ is the decision variable. Such models appear when the decision has to be taken before the full information about the problem parameters is known. The lack of information in that case is represented by random vector $\xi$. The second type of problems are the problems with the mathematical expectation as the objective function

$$\min_x E(F(x, \xi)). \tag{2}$$

Although the noise is technically removed in the above problem, it is rather hard to solve it as the expectation is hard or even impossible to state analytically even if the distribution of $\xi$ is known.

Methods for solving stochastic optimization problems are combination of ideas from numerical optimization and statistics. Thus the class of popular methods include simulation-based methods, direct methods for stochastic search, annealing type algorithms, genetic algorithms, methods of reinforced learning, statistical methods and many others, [34], [7]. Among all of them we restrict our attention here on two methods typically used in simulation based optimization: Stochastic Approximation, SA, and Sample Average Approximation, SAA.

Stochastic Approximation methods are introduced in the seminal paper of Robbins and Monro, [30] and remain a popular choice for solving stochastic optimization problems. They relay mainly on noisy gradient evaluations and depend heavily on the choice of steplength sequence. The choice of this sequence is the subject of many research efforts as well as other techniques for accelerating the convergence of SA methods. Sample average Approximation methods can be seen as an alternative to SA methods. In this approach a sample from the underlying distribution is used to construct a deterministic sample average problem which can be solved by optimization methods. However the sample used for the SAA approximation very often needs to be large and a naive application of standard nonlinear optimization techniques is not feasible. Therefore there has been extensive research in variable sample size methods that reduce the cost of SAA.

Both SA and SAA methods are considered here in the framework of gradient-related optimization (gradient methods, subgradient methods, second order quasi Newton methods) as well as in the derivative-free framework. This survey will largely deal with gradient methods for stochastic optimization and an interested reader can look at Spall, [34] for an overview of other methods. This paper is organized as follows. In Section 2 we discuss the SA method and its modifications. Section 3 contains results for the SAA methods and unconstrained problems with the mathematical expectation objective function. Two important

deterministic problems that are rather similar to SAA problem are discussed in Section 4 as well as methods for obtaining their solution that rely on stochastic gradients. Finally, some conclusion and research perspectives are presented in Section 5.

## 2 Stochastic Approximation Methods

There have been countless applications of Stochastic Approximation (SA) method since the work of Robbins and Monro, [30]. In this section we give an overview of its main properties and some of its generalizations. The problem we consider is

$$\min_{x\in\mathbb{R}^n} f(x) \tag{3}$$

assuming that only the noisy measurements $\hat{f}(x)$ of the function and its gradient $\hat{g}(x)$ are available. Let us start by considering the SA algorithm for solving systems of nonlinear equations as it was defined originally in [30]. The convergence theory presented here relays on imposition of statistical conditions on the objective function and the noise. Convergence analysis can be conducted throughout differential equations as well, see [34] and [25] for further references.

Consider the system of nonlinear equations

$$g(x) = 0,\ g : \mathbb{R}^n \to \mathbb{R}^n, \tag{4}$$

with $g(x)$ being the gradient of $f(x)$. Suppose that only the measurements with noise that depends on the iteration as well as on the decision variable $x$

$$\hat{g}_k(x) = g(x) + \xi_k(x) \tag{5}$$

are available. Then the SA is defined by

$$x_{k+1} = x_k - a_k\hat{g}_k(x_k). \tag{6}$$

The sequence of step sizes $\{a_k\}_{k\in\mathbb{N}}$ is also called the gain sequence and it has dominant influence on the convergence.

Let $\{x_k\}$ be a sequence generated by an SA method. Denote by $\mathcal{F}_k$ the $\sigma$-algebra generated by $x_0, x_1, \ldots, x_k$. If the problem has an unique solution $x^*$ the set of assumptions that ensures the convergence of an SA method is the following.

**S 1.** *The gain sequence satisfies:*

$$a_k > 0,\ \lim_{k\to\infty} a_k = 0,\ \sum_{k=0}^{\infty} a_k = \infty \quad \textit{and} \quad \sum_{k=0}^{\infty} a_k^2 < \infty$$

.

**S 2.** *For some symmetric, positive definite matrix $B$ and for every $\eta \in (0, 1)$,*

$$\inf_{\eta<\|x-x^*\|<\frac{1}{\eta}} (x - x^*)^T Bg(x) > 0.$$

**S 3.** *For all $x$ and $k$, $E[\xi_k(x)|\mathcal{F}_k] = 0,\ E[\|\xi_k(x)\|^2] < \infty$.*

**S 4.** *There exists a constant $c > 0$ such that for all $x$ and $k$,*

$$\|g(x)\|^2 + E\left[\|\xi_k(x)\|^2|\mathcal{F}_k\right] \leq c(1 + \|x\|^2).$$

The first assumption which implies that the step sizes converge to zero is standard in stochastic algorithms, see [34]. The second condition, $\sum_{k=0}^{\infty} a_k = \infty$ is imposed in order to avoid inefficiently small step sizes. On the other hand, the summability condition on $a_k^2$ ensures stability. Its role is to decrease the influence of the noise when the iterates come into a region around the solution. An example of a sequence that satisfies the first assumption is

$$a_k = \frac{a}{(k+1)^\alpha} \tag{7}$$

where $\alpha \in (0.5, 1]$ and $a$ is some positive constant. The condition of zero mean is also standard in stochastic optimization. Its implication is that $\hat{g}_k(x)$ is an unbiased estimator of $g(x)$. Notice that under assumption S3, the condition in S4 is equal to

$$E\left[\|\hat{g}_k(x)\|^2\right] \leq c(1 + \|x\|^2).$$

Therefore, the mean of $\|\hat{g}_k(x)\|^2$ can not grow faster than a quadratic function of $x$. Under these assumptions, the almost sure convergence of the SA algorithm can be established. The convergence in mean square i.e. $E[|x_k - x^*|]^2 \to 0$ as $k \to \infty$ was proved in [30] and the theorem bellow states a stronger result, the almost sure convergence.

**Theorem 2.1.** *[34] Consider the SA algorithm defined by (6). Suppose that assumptions S1 - S4 hold and that $x^*$ is a unique solution of the system (4). Then $x_k$ converges almost surely to $x^*$.*

Closely related and more general result is proved in Bertsekas, Tsitsiklis [6] where the gradient-related method of the form

$$x_{k+1} = x_k + a_k(s_k + \xi_k) \tag{8}$$

is considered. Here $\xi_k$ is either stochastic or deterministic error, $a_k$ is a sequence of diminishing step sizes that satisfy assumption S1 and $s_k$ is a descent direction. In this context, the direction $s_k$ is not necessarily the gradient but it is gradient-related. The convergence is stated in the following theorem.

**Theorem 2.2.** *[6] Let $\{x_k\}$ be a sequence generated by (8), where $s_k$ is a descent direction. Assume that S1 and S3 - S4 hold and that there exist positive scalars $c_1$ and $c_2$ such that*

$$c_1\|\nabla f(x_k)\|^2 \leq -\nabla f(x_k)^T s_k,\ \|s_k\| \leq c_2(1 + \|\nabla f(x_k)\|).$$

*Then, either $f(x_k) \to -\infty$ or else $f(x_k)$ converges to a finite value and $\lim_{k\to\infty} \nabla f(x_k) = 0$. Furthermore, every limit of $\{x_k\}$ is a stationary point of $f$.*

The gain sequence is the key element of the SA method. It has impact on stability as well as on convergence rate. Under some regularity conditions, Fabian [13], the asymptotic normality of $x_k$ is obtained. More precisely,

$$k^{\frac{\alpha}{2}}(x_k - x^*) \to^d \mathcal{N}(0, \Sigma),\ k \to \infty$$

where $\to^d$ denotes the convergence in distribution, $\alpha$ refers to (7) and $\Sigma$ is the covariance matrix that depends on the gain sequence and on the Hessian of $f$. Therefore, the iterate $x_k$ approximately has normal distribution $\mathcal{N}(x^*, k^{-\alpha}\Sigma)$ for large $k$. Due to assumption S1, the maximal convergence rate is obtained for $\alpha = 1$. However, this reasoning is based on the asymptotic result. Since the algorithms are finite in practice, it is often desirable to set $\alpha < 1$ because $\alpha = 1$ yields smaller steps. Moreover, if we want to minimize $\|\Sigma\|$, the ideal sequence would be

$$a_k = \frac{1}{k+1} H(x^*)^{-1}$$

where $H(x)$ denotes the Hessian matrix of $f$, Benveniste et al. [5]. Even though this result is purely theoretical, sometimes the Hessian at $x^*$ can be approximated by $H(x_k)$ and that way one can enhance the rate of convergence.

Two main drawbacks of the SA method are slow convergence and the fact that the convergence theory applies only if the solution of (4) is unique i.e. only if $f$ has an unique minimizer. Thus several generalizations are developed to address these two issues. One can easily see from (7) that the gain coefficients increase with the increase of $a$. On the other hand a large $a$ might have negative influence on stability. Therefore several generalizations of the gain coefficients are considered in the literature. One possibility, Spall [35], is to introduce the so called stability constant $A > 0$, and obtain

$$a_k = \frac{a}{(k+1+A)^\alpha}.$$

Now, the values of $a$ and $A$ can be chosen together to ensure effective practical performance of the algorithm, allowing for larger $a$ and thus producing larger step sizes in latter iterations, when the effect of $A$ is small, and avoiding instability in early iterations. The empirically recommended value of $A$ is at most 10% of the iterations allowed or expected during the optimization process, for more details see [35].

Several generalizations of the SA method are based on adaptive step sizes that try to adjust the step size at each iteration to the progress achieved in the previous iterations. The first attempt of this kind has been made in Kesten [20] for one dimensional problems. The main idea is to monitor the changes in the sign of $x_{k+1} - x_k$. If the sign of the difference between two consecutive iterations starts to change frequently we are probably in the domain of noise and therefore small steeps are needed to avoid oscillations. This idea is generalized in Delyon, Juditsky [11] for multidimensional problems. The gain coefficients are defined as

$$a_k = \frac{a}{s_k + 1},\ \ s_{k+1} = s_k + I(\hat{g}_{k+1}^T \hat{g}_k),$$

where $I$ is the identification function defined as $I(t) = 1$ if $t < 0$ and $I(t) = 0$ if $t \geq 0$. The method is accelerating the convergence of SA and its almost sure convergence is proved under the standard assumptions.

The idea of sign changes is further developed in Xu, Dai [38]. It is shown that the sequence $\{s_k/k\}$ stands for the change frequency of the sign of $\hat{g}_{k+1}^T \hat{g}_k$ in some sense. The assumption in [38] is that the noise $\xi_k$ is state independent. The theoretical analysis shows that in that case $s_k/k$ converges to $P(\xi_1^T \xi_2 < 0)$ in the mean square sense. Based on that result, a switching algorithm that uses the switching parameter $t_k$, defined as

$$t_k = \left| \frac{s_k}{k} - P(\xi_1^T \xi_2 < 0) \right| \tag{9}$$

is proposed. Then the gain coefficients are defined as

$$a_k = \begin{cases} \frac{a}{(k+1+A)^\alpha}, & \text{if } t_k \geq v \\ \frac{a}{(k+1+A)^\beta}, & \text{if } t_k < v \end{cases}, \tag{10}$$

where $0.5 \leq \alpha < \beta \leq 1$, $v$ is a small positive constant and $a, A$ are the constants from assumption S1. To prove the convergence of the switching algorithm (9)-(10) one additional assumption is introduced in [38].

**S 5.** *$G(x) = g(x) - x$ is a weighted maximum norm pseudo-contraction operator. That is for all $x \in \mathbb{R}^n$ there exists a positive vector $w$ and some $x^* \in \mathbb{R}^n$ such that*

$$\|G(x) - x^*\|_w \leq \beta \|x - x^*\|_w,$$

*where $\beta \in [0, 1)$ and $\|\cdot\|_w$ is defined as $\|x\|_w = \max\{x(i)w(i)^{-1},\ i = 1, \ldots, n\}$ where $x(i)$ and $w(i)$ are the $i$th components of $x$ and $w$ respectively.*

**Theorem 2.3.** *[38] Suppose that assumptions S1 - S2 and S5 hold. Then for $\{x_k\}$ generated through (9)-(10) we have $x_k \to x^*$ as $k \to \infty$ with probability 1.*

If the objective function is given in the form of mathematical expectation the adaptive step length sequence can be determined as proposed in Yousefian et al. [36]. For the problem

$$\min_x f(x) := E[F(x, \xi)] \tag{11}$$

the following assumptions are stated.

**S 6.** *The function $F(\cdot, \xi)$ is convex on a closed and convex set $D \subset \mathbb{R}^n$ for every $\xi \in \Omega$, and the expected value $E[F(x, \xi)]$ is finite for every $x \in D$.*

**S 7.** *The errors $\xi_k$ in the noisy gradient $\hat{g}_k$ are such that for some $\mu > 0$,*

$$E[\|\xi_k\|^2 | \mathcal{F}_k] \leq \mu^2 \ \text{ a.s. } \ \text{for all } k \geq 0.$$

A self-adaptive scheme is based on the error minimization and the convergence result is as follows.

**Theorem 2.4.** *[36] Let assumptions S6 and S7 hold. Let the function $f$ be differentiable over the set $D$ with Lipschitz gradient and assume that the optimal set of problem (11) is nonempty. Assume that the step size sequence $\{a_k\}$ is generated through the following self-adaptive scheme*

$$a_k = a_{k-1}(1 - ca_{k-1}) \text{ for all } k \geq 1,$$

*where $c > 0$ is a scalar and the initial step size is such that $0 < a_0 < 1/c$. Then the sequence $\{x_k\}$ converges almost surely to a random point that belongs to the optimal set.*

An important choice for the gain sequence is a constant sequence. Although such sequences do not satisfy assumption S1 and almost sure convergence to solution can not be obtained, it can be shown that a constant step size can conduct the iterations to a region that contains the solution. This result initiated development of a cascading steplength SA scheme in [36] where a fixed step size is used until some neighborhood of the solution is reached. After that, in order to come closer to the solution, the step size is decreased and again the fixed step size is used until the ring around the solution is sufficiently tighten up. That way, the sequence of iterates is guided towards the solution.

A hybrid method which combines the SA gain coefficients and the step sizes obtained from the inexact Armijo line search under the assumptions valid for the SA method is considered in Krejić et al. [24]. The method takes the advantages of both approaches, safe convergence of SA method and fast progress obtained by line search if the current iterate if far away from the solution (where the SA steps would be unnecessarily small). The step size is defined according to the following rule. For a given $C > 0$ and a sequence $\{a_k\}$ that satisfies assumption S1 we define

$$\alpha_k = \begin{cases} a_k & \text{if } \|\hat{g}(x_k)\| \leq C \\ \beta_k & \text{if } \|\hat{g}(x_k)\| > C, \end{cases} \tag{12}$$

where $\beta_k$ is obtained from the Armijo inequality

$$\hat{f}(x_k - \beta_k \hat{g}(x_k)) \leq \hat{f}(x_k) - c\beta_k \|\hat{g}(x_k)\|^2.$$

After that the new iteration is obtained as

$$x_{k+1} = x_k - \beta_k \hat{g}(x_k). \tag{13}$$

The existence of $C$ such that the gain coefficient (12) is well defined as well as the convergence of the sequence generated by (13) is proved in [24] under one additional assumption.

**S 8.** *Observation noise is bounded and there exists a positive constant $M$ such that $\|\xi_k(x)\| \leq M$ a.s. for all $k$ and $x$.*

**Theorem 2.5.** *[24] Assume that Assumptions S1-S4 and S8 hold, the gradient $g$ is Lipschitz continuous with the constant $L$, and the Hessian matrix $H(x^*)$*

*exists and is nonsingular. Let*

$$C \geq \max\left\{\frac{4(1-c_1)}{\underline{\alpha}c_1}, \frac{M+2\sqrt{2ML}+1}{1-c_1}\right\}$$

*where*

$$\underline{\alpha} = \frac{(1-c_1)(2\sqrt{2ML}+1)}{2L(M+2\sqrt{2ML}+1)}.$$

*Let $\{x_k\}$ be an infinite sequence generated by (13). Then $x_k \to x^*$ a.s.*

Many important issues regarding the convergence of the SA methods are not mentioned so far. One effective possibility to speed up the convergence is to apply the averaging to the sequence of gradient estimations $\hat{g}(x_k)$ as suggested in Andradottir [1]. It is shown that the rate of convergence could be significantly better than the rate of SA if two conditionally independent gradient estimations are generated and the new iteration is obtained using a scaled linear combination of the two gradients with the gain coefficient. More details on this procedure are available in [1]. Let us also mention a robust SA scheme that determines an optimal constant step length based on minimization of the theoretical error for a pre-specified number of steps [27].

We have assumed in the above discussion that the noisy gradient values are available. This is the case for example if the analytical expression of $F$ in (11) is available. In this case, under certain assumption we can interchange the expectation and derivative and thus a sample average approximation of the gradient can be calculated. It is important to be able to use a sample gradient estimation with relatively modest sample size as calculation of the sample gradient is in general expensive for large samples. However it is safe to claim that the analytical expression for the gradient calculation is not available in many cases and thus the only input data we have are (possibly noisy) function values. Thus the gradient approximation with finite differences appears to be a natural choice in many applications. The first method of this kind is due to Keifer, Wolfowitz, [21]. Many generalizations and extensions are later considered in the literature, see Fu [15] for example. Among many methods of this kind the Stochastic Perturbation method is particularly efficient as it uses one two function values to obtain a good gradient approximation, see [35] for implementation details.

The questions of stopping criteria, global convergence, search directions which are not gradient related, and other important questions are beyond the scope of this paper. An interested reader might look at Spall [34], Shapiro et al. [32] for guidance on these issues and relevant literature.

# 3 Sample Average Approximation

Sample Average Approximation (SAA) is a widely used technique for approaching the problems of the form

$$\min_x f(x) = E\left[F(x,\xi)\right]. \tag{14}$$

The basic idea is to approximate the objective function $f(x)$ with the sample mean

$$\hat{f}_N(x) = \frac{1}{N}\sum_{i=1}^{N} F(x, \xi_i) \tag{15}$$

where $N$ is the size of a sample represented by i.i.d. random vectors $\xi_1, \ldots, \xi_N$. Under the standard assumption such as finite variance of $F(x, \xi)$ the (strong) Law of Large Numbers implies that $\hat{f}_N(x)$ converges to $f(x)$ almost surely. Moreover, if $F(x, \xi)$ is dominated by an integrable function, then the uniform almost sure convergence of $\hat{f}_N(x)$ on the compact subsets of $\mathbb{R}^n$ is obtained.

Within the SAA framework the original problem (14) is replaced by the approximate problem

$$\min_x \hat{f}_N(x), \tag{16}$$

and thus the key question is the relationship between their respective solutions as $N$ tends to infinity. Denote by $X^*$ the set of optimal solutions of the problem (14) and let $f^*$ be the optimal value of the objective function. Furthermore, denote by $\hat{X}_N^*$ and $\hat{f}_N^*$ the set of optimal solutions and the corresponding optimal values, respectively, of the problem (16). Then, the following result holds.

**Theorem 3.1.** *[32] Suppose that there exists a compact set $C \subset \mathbb{R}^n$ such that $X^*$ is nonempty and $X^* \subset C$. Assume that the function $f$ is finite valued and continuous on $C$ and that $\hat{f}_N$ converges to $f$ almost surely, uniformly on $C$. Also, suppose that for $N$ large enough the set $\hat{X}_N^*$ is nonempty and $\hat{X}_N^* \subset C$. Then $\hat{f}_N^* \to f^*$ and the distance between sets $\hat{X}_N^*$ and $X^*$ tends to zero almost surely as $N \to \infty$.*

Let $\bar{x}_N$ be an approximate solution of the problem (14). Clearly $\hat{f}_N(\bar{x}_N)$ can be calculated for a given sample. The Central Limit Theorem can be used to obtain an approximation of the error bound $c_N(\bar{x}_N)$ such that the inequality $|\hat{f}_N(\bar{x}_N) - f(\bar{x}_N)| \leq c_N(\bar{x}_N)$ holds with some high probability $\delta \in (0, 1)$. For example, using the sample variance $\hat{\sigma}_N^2(\bar{x}_N) = \frac{1}{N-1}\sum_{i=1}^{N}(F(\bar{x}_N, \xi_i) - \hat{f}_N(\bar{x}_N))^2$ the following error bound is obtained

$$\varepsilon_\delta^N(\bar{x}_N) = \frac{\hat{\sigma}_N(\bar{x}_N)}{\sqrt{N}} z_{\frac{1+\delta}{2}}, \tag{17}$$

with $z$ being the quantile of the standard normal distribution. The error bound is directly proportional to the variance of the estimator $Var(\hat{f}_N(\bar{x}_N))$. Therefore, in order to provide a tight bound one can consider some techniques for reducing variance such as the quasi-Monte Carlo or Latin hypercube sampling [32]. However, these techniques tend to deteriorate the i.i.d. assumption. This issue is addressed further on in this section.

The gap $g(\bar{x}_N) = f(\bar{x}_N) - f(x^*)$ where $x^*$ is a solution of the original problem can be estimated as well. Clearly $g(\bar{x}_N) \geq 0$. To obtain an upper bound suppose that $M$ independent samples of size $N$ are available, i.e.

we have i.i.d. sample $\xi_1^m, \ldots, \xi_N^m$, $m = 1, \ldots, M$. Denote by $\hat{f}_N^{m*}$ the relevant (nearly) optimal values and define $\hat{f}_{N,M}^* = \frac{1}{M}\sum_{m=1}^M \hat{f}_N^{m*}$ and $\hat{\sigma}_{N,M}^2 = \frac{1}{M}\left(\frac{1}{M-1}\sum_{m=1}^M \left(\hat{f}_N^{m*} - \hat{f}_{N,M}^*\right)^2\right)$. Then, the upper bound estimator for the gap $g(\bar{x}_N)$ is

$$\hat{f}_{N'}(\bar{x}_N) + z_\delta \frac{\hat{\sigma}_{N'}(\bar{x}_N)}{\sqrt{N'}} - \hat{f}_{N,M}^* + t_{M-1,\delta}\hat{\sigma}_{N,M}$$

where $N'$ is some large enough sample size and $t_{M-1,\delta}$ is the quantile of Student's distribution with $M-1$ degrees of freedom. It should be mentioned that the sample size bounds such that the solutions of an approximate problem are nearly optimal for the true problem with some high probability are mainly too conservative for practical applications in general. For further references on this topic, see [32] for instance.

Recall that almost sure convergence $\hat{f}_N(x)$ towards $f(x)$ is achieved if the sample is i.i.d. under standard assumptions. However, if the sample is not i.i.d. the almost sure convergence of $\hat{f}_N$ is achievable only if the sample size $N$ which defines the SAA problem grows at the certain rate, Homem-de-Mello [19]. The analysis presented in [19] allows for biased estimators $\hat{f}_N(x)$ if $\hat{f}_N(x)$ is at least asymptotically unbiased. Let us first assume that the sample $\xi_1^k, \ldots, \xi_{N_k}^k$ generated at the iteration $k$ is independent of the sample at the previous iteration for every $k$. The following assumptions are needed.

**R 1.** *For each $x$, there exists $M(x) > 0$ such that $\sup_{i,k} F(x, \xi_i^k) \leq M(x)$ with probability 1.*

**R 2.** *For each $x$, we have that $\lim_{k\to\infty} E(\hat{f}_N(x)) = f(x)$.*

**Theorem 3.2.** *[19] Suppose that assumptions R1-R2 hold and that the sample size sequence $\{N_k\}$ satisfies $\sum_{k=1}^\infty \alpha^{N_k} < \infty$ for all $\alpha \in (0,1)$. Then $\hat{f}_{N_k}(x)$ converges to $f(x)$ almost surely.*

For example, $N_k \geq \sqrt{k}$ satisfies the previously stated summability condition.

The rate of convergence is also addressed in [19], i.e. the error bounds for $|\hat{f}_{N_k}(x) - f(x)|$ are developed. In the case where $N_k \geq c_1 k^\rho$ for $c_1 > 0$ and $\rho > 2$ it can be proved, under some additional assumptions, that for every $k$ sufficiently large the following inequality holds almost surely

$$|\hat{f}_{N_k}(x) - f(x)| \leq \sqrt{Var(F(x, \xi_1^k))}\sqrt{\frac{\ln N_k}{N_k}} + |E(\hat{f}_{N_k}(x)) - f(x)|.$$

Moreover, if the sample is cumulative, the corresponding error bound is

$$|\hat{f}_{N_k}(x) - f(x)| \leq C\sqrt{\frac{\ln(\ln(N_1 + \ldots + N_k))}{N_1 + \ldots + N_k}} \tag{18}$$

where $C$ is some positive constant.

The above analysis provides a justification for the SAA approximation as well as a guidance for choosing $N$ in (16). Thus from now on we concentrate on gradient methods for solving (16). Several papers exploit ideas from deterministic optimization. Generally speaking we are 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 (16) is very often prohibitively costly if $N$ is large due to the cost of calculating $\hat{f}_N(x)$ and its gradient. Thus there is a vast literature dealing with variable sample scheme for solving (16).

Two main approaches can be distinguished. In the first approach the objective function $\hat{f}_N$ is replaced with $\hat{f}_{N_k}(x)$ 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 $\hat{f}_{N_k}(x_k + s_k)$ 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.

The second approach, often called the diagonalization scheme or 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. So for the current $x_k$ and $N_k$ the problem (16) with $N = N_k$ is approximately solved (within an inner loop) for $\tilde{x}_{N_k}$ starting with $x_k$ as the initial approximation. After that we set $x_{k+1} = \tilde{x}_{N_k}$ and choose the new sample size $N_{k+1}$. Two important points in this procedure are the choice of $N_{k+1}$ and the precision in solving each of the optimization problems $\min \hat{f}_{N_k}$.

Let us now look into algorithms of the first kind. Keeping in mind that $\min \hat{f}_{N_k}$ 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 $N_0, N_1, \ldots$. The convergence theory for this kind of reasoning is introduced in Wardi, [40] where an Armijo type line search method is combined with SAA approach. In order to solve the problem of type (14), the iterative sequence is generated as

$$x_{k+1} = x_k - \alpha_k \nabla \hat{f}_{N_k}(x_k) \tag{19}$$

where $N_k$ is the sample size used at iteration $k$ and $\alpha_k$ is the largest number in $(0, 1]$ satisfying the inequality

$$\hat{f}_{N_k}(x_k - \alpha_k \nabla \hat{f}_{N_k}(x_k)) \leq \hat{f}_{N_k}(x_k) - \eta \alpha_k \|\nabla \hat{f}_{N_k}(x_k)\|^2. \tag{20}$$

The method is convergent with zero upper density [40], assuming that $N_k \to \infty$. More precisely, the following statement is proved.

**Theorem 3.3.** *[40] Assume that the function f is given by (14) and that F is twice continuously differentiable on $\mathbb{R}^n$ for every $\xi$. Furthermore assume that for every compact set $D \subset \mathbb{R}^n$, there exists $K > 0$ such that for every $x \in D$ and every $\xi$*

$$|F(x,\xi)| + \|\frac{\partial F}{\partial x}(x,\xi)\| + \|\frac{\partial^2 F}{(\partial x)^2}(x,\xi)\| \leq K.$$

*If $N_k \to \infty$ then the sequence $\{x_k\}$ given by (19) converges with zero upper density on compact sets.*

An extension of the above work is presented in Yan, Mukai [37] where the adaptive precision is proposed i.e. the sequence $\{N_k\}_{k\in\mathbb{N}}$ is not determined in advance as in [40] but it is adapted during the iterative procedure. Nevertheless the sample size has to satisfy $N_k \to \infty$. The convergence result 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, Wardi [33] where the convergence with probability 1 is proved for sample average gradient and subgradient methods assuming that the sample size tends to infinity.

In practical applications, the sample size is finite. So, let us now suppose that $N_{max}$ is the sample size which makes $\hat{f}_{N_{max}}$ good approximation of the original objective function. Very often we assume that the sample is generated at the beginning of the process which justifies considering the SAA objective function as deterministic. In this case one wishes again to decrease the cost of the optimization process by decreasing the number of function and gradient evaluations. Let us now look closer at the possible schedule of $N_k$. Clearly the sample size should be equal to $N_{\max}$ at the final stages of the optimization procedure to ensure that the problem (16) with $N = N_{\max}$ is solved. Thus one can consider even some heuristic schedule [14] to generate a non-decreasing sequence $\{N_k\}$ which eventually becomes stationary with $N_k = N_{\max}$ for $k$ large enough. For example, a simple way to define such sequence could be to increase $N_k$ by a fixed number every $K$ iterations.

The problem of scheduling can be approached from a different perspective in the following manner. Instead of constantly increasing the sample size one could monitor the progress in decreasing the (approximate) objective function and choose the next sample size according to that progress. One algorithm of this kind is presented in Deng, Ferris [12] where the Bayes risk is used to decide the scheduling sequence within a trust region method. Another class of results in the framework of trust region methods is presented in Bastin [2] and Bastin et al. [3], [4]. The key point of the approach considered in [2, 3, 4] is that the sample sizes might oscillate during the iterative process, i.e. $\{N_k\}$ is not necessarily non-decreasing at the initial stages of the iterative process. Eventually $N_k = N_{\max}$ is reached and (16) is solved, but very often with smaller costs if compared with an increasing scheduling. The efficiency of this approach comes from the fact that there is a balance between the precision of the objective function approximation $\hat{f}_{N_k}$ and the progress towards the solution. The same

idea is further developed for the line search methods in Krejić,Krklec [22] as follows.

Let us assume that the gradient of $\nabla F$ is available and that the search direction $p_k$ satisfies $p_k^T \nabla \hat{f}_{N_k}(x_k) < 0$. The Armijo rule with $\eta \in (0,1)$ is applied to find $\alpha_k$ such that $x_{k+1} = x_k + \alpha_k p_k$ where

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

The sample size is updated as follows. First, the candidate sample size $N_k^+$ is determined by comparing the measure of decrease in the objective function $dm_k = -\alpha_k (\nabla \hat{f}_{N_k}(x_k))^T p_k$ and the so called lack of precision $\varepsilon_\delta^N(x)$ defined by (17). The main idea is to find the sample size $N_k^+$ such that $dm_k \approx \varepsilon_\delta^{N_k^+}(x^k)$. The reasoning behind this idea is the following. If the decrease measure $dm_k$ is greater than the lack of precision $\varepsilon_\delta^{N_k^+}(x^k)$, the current approximation is probably far away from the solution. In that case, there is no need to impose high precision and therefore the sample size is decreased if possible. The candidate sample size does not exceed $N_{max}$, but there is also the lower bound, i.e. $N_k^{\min} \leq N_k^+ \leq N_{\max}$. This lower bound is increased only if $N_{k+1} > N_k$ and there is not enough progress concerning the function $\hat{f}_{N_{k+1}}$. After finding the candidate sample size, a safeguard check is performed in order to prohibit the decrease of the sample size which might be unproductive. More precisely, if $N_k^+ < N_k$ the following parameter is calculated

$$\rho_k = \frac{\hat{f}_{N_k^+}(x^k) - \hat{f}_{N_k^+}(x^{k+1})}{\hat{f}_{N_k}(x^k) - \hat{f}_{N_k}(x^{k+1})}.$$

If $\rho_k$ is relatively small, then it is presumed that these two model functions are too different and thus there is no gain in decreasing the sample size. So, $N_{k+1} = N_k$. In all the other cases, the decrease is accepted and $N_{k+1} = N_k^+$. The convergence analysis relays on the following important result which states that after some finite number of iterations, the objective function becomes $\hat{f}_{N_{max}}$ and (16) is eventually solved.

**Theorem 3.4.** *[22] Suppose that $F(\cdot, \xi)$ is continuously differentiable and bounded from below for every $\xi$. Furthermore, suppose that there exist a positive constant $\kappa$ and number $n_1 \in \mathbb{N}$ such that $\varepsilon_\delta^{N_k}(x_k) \geq \kappa$ for every $k \geq n_1$. Then there exists $q \in \mathbb{N}$ such that $N_k = N_{max}$ for every $k \geq q$.*

Let us now present some results for the second type of methods, the so called diagonalization methods described above. One possibility to determine the sample sizes in the sequence of optimization problems to be solved is presented in Royset [31] where an optimality function is used to determine when to switch on to a larger sample size. The optimality function is defined by mapping $\theta : \mathbb{R}^n \to (-\infty, 0]$ which, under standard conditions, satisfies $\theta(x) = 0$ if and only if $x$ is a solution in some sense. For unconstrained problem $\theta(x) = -\frac{1}{2}\|\nabla f(x)\|^2$ and its SAA approximation is given by $\theta_N(x) = -\frac{1}{2}\|\nabla \hat{f}_N(x)\|^2$. Under the set of

standard assumptions, almost sure convergence of $\theta_N(x)$ towards $\theta(x)$ is stated together with asymptotic normality.

Denote by $\tilde{x}_{N_k}$ the iterate obtained after a finite number of iterations of an algorithm applied on the SAA problem with sample size $N_k$ where $\tilde{x}_{N_{k-1}}$ is the initial point. The point $\tilde{x}_{N_k}$ is an approximate solution of (16) with $N = N_k$ and it is assumed that the optimization algorithm used to determine that point is successful in the following sense.

**R 3.** *For any $N_k$ every accumulation point $\tilde{x}_{N_k}$ of the sequence generated by the optimization method for solving*

$$\min \hat{f}_{N_k}(x)$$

*satisfies $\theta_{N_k}(\tilde{x}_{N_k}) = 0$ almost surely.*

The algorithm proposed in [31] increases the sample size when

$$\theta_{N_k}(x_k) \geq -\delta_1 \Delta(N_k),$$

where $\delta_1$ is some positive constant and $\Delta$ is a function that maps $\mathbb{N}$ into $(0, \infty)$ and satisfies $\lim_{N\to\infty} \Delta(N) = 0$. The sample size is assumed to be strictly increasing and unbounded, but the exact dynamics of increasing is not specified. The convergence of the algorithm is proved under one additional assumption.

**R 4.** *On any given set $S \subset \mathbb{R}^n$, the function $F(\cdot, \xi)$ is continuously differentiable and $F(\cdot, \xi)$ and $\|\nabla_x F(\cdot, \xi)\|$ are dominated by an integrable function.*

**Theorem 3.5.** *[31] Suppose that the assumptions R3-R4 are satisfied and that the sequence of iterates generated by the algorithm proposed in [31] is bounded. Then, every accumulation point $\hat{x}$ of that sequence satisfies $\theta(\hat{x}) = 0$ almost surely.*

The relation between the sample size and the error tolerance for each of the optimization problems solved within the diagonalization methods is considered in Pasupathy [28]. The error tolerance here is a small number $\varepsilon_k$ which almost surely satisfies $\|\tilde{x}_{N_k} - x^*_{N_k}\| \leq \varepsilon_k$ where $\tilde{x}_{N_k}$ and $x^*_{N_k}$ represent the approximate and the true (unique) solution of the corresponding SAA problem, respectively. A measure of effectiveness is defined as $q_k = \|\tilde{x}_{N_k} - x^*_{N_k}\|^2 W_k$ where $W_k$ represents the number of simulation calls needed to obtain the approximate solution $\tilde{x}_{N_k}$. Since the almost sure convergence is analyzed, it is assumed that $N_k \to \infty$ and $\varepsilon_k \to 0$. It is proved that the measure of effectiveness is bounded in a stochastic sense if the following three conditions hold.

**R 5.** *If the numerical procedure used to solve SAA problems exhibits linear convergence, we assume that* $\liminf_{k\to\infty} \varepsilon_k \sqrt{N_{k-1}} > 0$.
*If the numerical procedure used to solve SAA problems exhibits polynomial convergence of order $p > 1$, we assume* $\liminf_{k\to\infty} (\ln(1/\sqrt{N_{k-1}})/\ln(\varepsilon_k)) > 0$.

**R 6.** $\limsup_{k\to\infty} (\sum_{j=1}^{k} N_j)\varepsilon_k^2 < \infty$.

**R 7.** $\limsup_{k\to\infty}(\sum_{j=1}^{k} N_j)N_k^{-1} < \infty$.

If any of the above conditions is violated, then $q_k$ tends to infinity in probability. The key point of analysis in [28] is that the error tolerance should not be decreased faster than the sample size is increased. The dynamics of change depends on the convergence rate of numerical procedures used to solve the SAA problems. Moreover, the mean squared error analysis implies the choice of $\varepsilon_k$ and $N_k$ such that

$$0 < \limsup_{k\to\infty} \varepsilon_k \sqrt{N_k} < \infty. \tag{21}$$

In order to further specify the choice of the optimal sequence of sample sizes, the following theorem is stated.

**Theorem 3.6.** *[28] Suppose that (21) holds together with the assumptions R5-R7. If the numerical procedure used to solve SAA problems exhibits linear convergence, then* $\limsup_{k\to\infty} N_k/N_{k-1} < \infty$. *If the numerical procedure used to solve SAA problems exhibits polynomial convergence of order* $p > 1$, *then* $\limsup_{k\to\infty} N_k/N_{k-1}^p < \infty$.

More specific recommendations are given for linear, sublinear and polynomial rates in [28]. For example, if the applied algorithm is linearly convergent, then the linear growth of a sample size is recommended, i.e. it can be set $N_{k+1} = \lceil 1.1N_k \rceil$ for example. Also, in that case, exponential or polynomial growth of order $p > 1$ are not recommended. However, if the polynomial rate of convergence of order $p > 1$ is achieved, then we can set $N_{k+1} = \lceil N_k^{1.1} \rceil$ or $N_{k+1} = \lceil e^{N_k^{1.1}} \rceil$ for instance. Furthermore, it is implied that the error tolerance sequence should be of the form $K/\sqrt{N_k}$ where $K$ is some positive constant.

The diagonalization methods are defined for a finite $N$ as well. One possibility is presented in Polak, Royset [29] where the focus is on finite sample size $N$ although the almost sure convergence is addressed. The idea is to approximately solve the sequence of SAA problems with $N = N_k$, $k = 1, \ldots, s$ applying $n_k$ iterations at every stage $k$. The sample size is nondecreasing and the sample is assumed to be cumulative. The method consists of three phases. The first phase provides the estimates of relevant parameters such as the sample variance. In the second phase, the scheduling sequence is obtained. Finally, the sequence of the SAA problems is solved in the last phase.

An additional optimization problem is formulated and solved in the second phase in order to find the number $s$ of the SAA problem to be solved, the sample sizes $N_k, k = 1, \ldots, s$ and the number of iterations $n_k$ that are applied to solve the corresponding SAA problem. The objective function of this additional problem is the overall cost $\sum_{k=1}^{s} n_k w(N_k)$, where $w(N)$ is the estimated cost of one iteration of the algorithm applied on the function $\hat{f}_N$. For example $w(N) = N$. The constraint for this problem is motivated by the stopping criterion $f(x) - f^* \leq \varepsilon(f(x_0) - f^*)$ where $f^*$ is the optimal value of the objective function. More precisely, the cost-to-go is defined as $e_k = f(x_{n_k}^k) - f^*$ where $x_{n_k}^k$ is the last iteration at the stage $k$. Furthermore, the upper bound estimate

for $e_s$ is determined as follows. Let $\Delta(N)$ be the function defined as in Royset [31].

**R 8.** *There exists strictly decreasing function* $\Delta(N) : \mathbb{N} \to (0, \infty)$ *such that* $\lim_{N\to\infty} \Delta(N) = 0$ *and* $|\hat{f}_{N_k}(x) - f(x)| \leq \Delta(N_k)$ *holds almost surely.*

One may use a bound like (18), but it is usually too conservative for practical implementations. Therefore, $\Delta(N)$ is estimated with the confidence interval bound of the form (17) where the variance is estimated in the first stage. The following bound is derived

$$e_s \leq e_0\theta^{l_0(s)} + 4\sum_{k=1}^{s} \theta^{l_k(s)}\Delta(N_k),$$

where $l_k(s)$ represents the remaining number of iterations after the stage $k$ and $\theta$ defines the rate of convergence of the deterministic method applied on SAA. The initial cost-to-go from $e_0 = f(x_0^1) - f^*$ is also estimated in the first phase. Finally, the efficient strategy is obtained as the solution of the following problem

$$\min_{s\in\mathbb{N}} \min_{n_k, N_k} \sum_{k=1}^{s} n_k w(N_k) \quad \text{s.t.} \quad e_0\theta^{l_0(s)} + 4\sum_{k=1}^{s} \theta^{l_k(s)}\Delta(N_k) \leq \varepsilon e_0,\ N_k \geq N_{k-1}.$$

In order to prove the asymptotic result, the following assumption regarding the optimization method used at each stage is imposed.

**R 9.** *The numerical procedure used to solve SAA problems almost surely exhibits linear rate of convergence with parameter* $\theta \in (0, 1)$.

**Theorem 3.7.** *[29] Suppose that the assumptions R8-R9 hold and that the sample size sequence tends to infinity. Then* $\lim_{s\to\infty} e_s = 0$ *almost surely.*

# 4 Applications to deterministic problems

A number of important deterministic problems can be written in the form of

$$\min_x \quad \hat{f}_N(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \tag{22}$$

where $f_i(x)$ are given functions and $N$ is a large integer. For example, least squares and maximum likelihood problems are of this form. The objective function in (22) and its gradient are generally expensive to compute if $N$ is large. On the other hand for a given sample realization $\xi_1, ..., \xi_N$ and $f_i(x) = F(x, \xi_i)$ the SAA problems discussed in Section 3 are the same as (22). Therefore the SAA methods that deal with finite $N$ can be used for solving the deterministic problems specified in (22). The main idea of this approach is to use the same reasoning as in the variable sample schemes to decrease the cost of calculating the objective function and its gradient i.e. to approximate the function and the

gradient with $\hat{f}_{N_k}$ and $\nabla \hat{f}_{N_k}$. One application of a variable sample method to the data fitting problem is presented in Krejić, Krklec Jerinkić, [23]. In this section we consider two important problems, data fitting and machine learning, and methods for their solutions that use stochastic gradient approximation in the sense of approximate gradient as explained above.

The data fitting problem of the form (22) is considered in Friedlander, Schmidt [14]. The problem is solved a by quasi Newton method but the gradient approximation of the SAA type is used. Let the gradient estimation at the current iteration $k$ be given as $g_k = \nabla f(x_k) + \epsilon_k$ where $\epsilon$ is the error term. The following assumptions are stated.

**P 1.** *The functions $f_1, \dots, f_N$ are continuously differentiable and the function $f$ is strongly convex with parameter $\mu$. Also, the gradient $\nabla f$ is Lipschitz continuous with parameter $L$.*

**P 2.** *There are constants $\beta_1 \geq 0$ and $\beta_2 \geq 1$ such that $\|\nabla f_i(x)\|^2 \leq \beta_1 + \beta_2 \|\nabla f(x)\|^2$ for all $x$ and $i = 1, \dots, N$.*

The algorithm can be considered as an increasing sample size method where the sample size is bounded with $N$. The main issue in [14] is the rate of convergence and the convergence analysis is done with the assumption of a constant step size. More precisely

$$x_{k+1} = x_k - \frac{1}{L} g_k.$$

Two approaches are considered: deterministic and stochastic sampling. The deterministic sampling assumes that if the sample size is $N_k$ then the gradients to be evaluated $\nabla f_i(x_k),\ i = 1, \dots, N_k$, are determined in advance. For example, the first $N_k$ functions are used to obtain the gradient approximation $g_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \nabla f_i(x_k)$. On the other hand, stochastic sampling assumes that the gradients $\nabla f_i(x), i = 1, \dots, N_k$, to be evaluated are chosen randomly. We state the relevant results considering R-linear rate of convergence. In the case of deterministic gradient, q-linear convergence is also attained but under stronger conditions on the increase of the sample size.

**Theorem 4.1.** *[14] Suppose that the assumptions P1-P2 hold and that $(N - N_k)/N = \mathcal{O}(\gamma^{k/2})$ for some $\gamma \in (0, 1)$. Then for any $\varepsilon > 0$, $\sigma = \max\{\gamma, 1 - \mu/L\} + \varepsilon$ and every $k$, in deterministic case we obtain*

$$f(x_k) - f(x^*) = (f(x_0) - f(x^*))\mathcal{O}((1 - \mu/L + \varepsilon)^k) + \mathcal{O}(\sigma^k).$$

*Moreover, if $(N - N_k)/(N_k N) = \mathcal{O}(\gamma^k)$, in stochastic case we obtain*

$$E\left[f(x_k) - f(x^*)\right] = (f(x_0) - f(x^*))\mathcal{O}((1 - \mu/L + \varepsilon)^k) + \mathcal{O}(\sigma^k).$$

Machine learning applications which usually assume large number of training points can also be viewed as problems of the form (22). Methods for solving such problems are the subject of Byrd et al. [8] and Byrd et al. [9]. The main idea in [8] is to create methods which use the second order derivative

information but with cost comparable to the steepest descent method. The focus is on using a cheep Hessian approximations $\nabla^2 \hat{f}_{S_k}(x_k)$ where $S_k$ is a number of training points, i.e. the Hessian-related sample size at iteration $k$. More precisely, matrix-free conjugate gradient method is applied in order to obtain the search direction $p_k$ as an approximate solution of the system

$$\nabla^2 \hat{f}_{S_k}(x_k)p = -\nabla \hat{f}_{N_k}(x_k).$$

Here, $N_k$ is a sample size related to the gradient and the function approximation. This can be considered as an inexact Newton method. In the relevant examples, $p_k$ is guaranteed to be a descent search direction and therefore the Armijo line search is applied. The proposed method (named S-Newton) does not specify the dynamic of changing the sample sizes. It only requires that the variable sample strategy is used and $S_k < N_k$. The analysis is conducted for the full gradient case.

**Theorem 4.2.** *[8] Suppose that the function $\hat{f}_N$ is twice continuously differentiable and uniformly convex and that there exists a constant $\gamma > 0$ such that $x^T \nabla^2 \hat{f}_{S_k}(x_k)x \geq \gamma \|x\|$ for every $k$ and $x$. Then the sequence generated by the S-Newton method with $N_k = N$ satisfies $\lim_{k\to\infty} \|\nabla \hat{f}_N(x_k)\| = 0$.*

The same result can be obtained for the so called SLM method which uses matrix-free limited memory BFGS method. In that case, the conjugate gradient method is used to obtain the search direction where the sub-sampled Hessian approximation $\nabla^2 \hat{f}_{S_k}(x_k)$ is used for the initial matrix-vector product at every iteration. The line search uses Wolfe conditions for choosing the suitable step size.

The dynamic of increasing the sample size in the machine learning problem is addressed in [9]. The main idea is to estimate the sample size which makes the search direction $p_k$ descent for the objective function $\hat{f}_N$ without evaluating the true gradient $\nabla f_N(x)$. The approximation of the negative gradient $p_k = -\nabla \hat{f}_{N_k}(x_k)$ is a descend direction if for some $\theta \in [0, 1]$, the following inequality holds

$$\|\nabla \hat{f}_N(x_k) - \nabla \hat{f}_{N_k}(x_k)\| \leq \theta \|\nabla \hat{f}_{N_k}(x_k)\|. \tag{23}$$

Since $E[\|\nabla \hat{f}_N(x_k) - \nabla \hat{f}_{N_k}(x_k)\|^2] = \|Var(\nabla \hat{f}_{N_k}(x_k))\|_1$ and $N$ is large, inequality (23) is approximated by

$$\frac{\|\hat{\sigma}^2_{N_k}(\nabla f_i(x_k))\|_1}{N_k} \leq \theta^2 \|\nabla \hat{f}_{N_k}(x_k)\|^2 \tag{24}$$

where $\hat{\sigma}^2_{N_k}$ is a sample variance related to the chosen sample of the size $N_k$. The algorithm for the sample size schedule proposed in [9] can be described as follows. After finding the step size $\alpha_k$ such that $\hat{f}_{N_k}(x_k + \alpha_k p_k) < \hat{f}_{N_k}(x_k)$ and setting $x_{k+1} = x_k + \alpha_k p_k$, a new sample of the same size $N_k$ is chosen. If inequality (24) holds for the new sample, the sample size remains unchanged,

i.e. $N_{k+1} = N_k$. Otherwise, the sample is augmented and the new sample size is determined by

$$N_{k+1} = \left\lceil \frac{\|\hat{\sigma}^2_{N_k}(x_k)\|_1}{\theta^2 \|\nabla \hat{f}_{N_k}(x_k)\|^2} \right\rceil .$$

In order to conduct the complexity analysis, the constant step size is considered and the q-linear convergence rate is analyzed.

**Theorem 4.3.** *[9] Suppose that the function $\hat{f}_N$ is twice continuously differentiable and $x^*$ is a solution of the problem (22) with $\hat{f}_N(x^*) = 0$. Furthermore, assume that there are constants $0 < \lambda < L$ such that $\lambda\|h\|^2 \leq h^T \nabla^2 \hat{f}_N(x) h \leq L\|h\|^2$ for all $x$ and $h$. Let the sequence of iterates be generated by $x_{k+1} = x_k - \alpha \nabla \hat{f}_{N_k}(x_k)$ where $\alpha = (1-\theta)/L$ and $\theta \in (0,1)$. If the condition (23) is satisfied at iteration $k$, then*

$$\hat{f}_N(x_{k+1}) \leq (1 - \frac{\beta\lambda}{L}) \hat{f}_N(x_k)$$

*where $\beta = \frac{(1-\theta)^2}{2(1+\theta)^2}$. Moreover, if (23) holds for every $k$, then*

$$\lim_{k \to \infty} x_k = x^*.$$

The schedule $\{N_k\}$ for the gradient estimations is extended to the second order approximations to define a Newton-type method, the S-Newton method defined in [8]. This method uses the updating of $N_k$ described above while the Hessian-related sample size $S_k$ follows the dynamic of $N_k$. More precisely, $S_k = RN_k$ where $R$ is some positive number substantially smaller than 1. Also, the sample used for the Hessian approximation is assumed to be a subset of the sample used for the gradient and the function approximations. The stopping criterion for the conjugate gradient method used for obtaining the search direction is more complex than in [8] since it is related to the sample size. Wolfe conditions are imposed to obtain a suitable step size.

# 5 Conclusions

In this survey we considered unconstrained problems with the stochastic or expectation objective function. We focused our attention on two specific classes of gradient-related methods: Stohastic Approximation and Sample Average Approximation and many other important approaches are left out for the sake of brevity. An interested reader should consults [32, 34] for the initial guidance into stochastic optimization problems. Several natural extensions are easily incorporated in the framework considered in this paper, for example search directions with second order information which usually yield faster convergence, but also require additional cost, [8, 9, 10]. In order to decrease the linear algebra costs, one can consider preconditioners as their construction might be a nontrivial issue due to the presence of random variable. On the other hand,

given that there are a lot of problems yielding only input-output information, an interesting approach within SA and SAA frameworks is based on zero order information, [17]. Constrained problems are always of great interest and some recent research of penalty methods within SA methodology is presented in [39]. Projection and filter methods with variable sample size might be a valuable topic of future research. Among the others, chance constrained problems are especially challenging. In all the considered methods, deriving some complexity bounds can be of great interest from practical point of view.

# References


[1] S. Andradottir, A Scaled Stochastic Approximation Algorithm, *Management Science 42(4) (1996), 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, *Math. Program., Ser. B 108 (2006) pp. 207-234.*

[5] A. Benveniste, M. Metivier, P. Priouret, Adaptive Algorithms and Stochastic Approximations, *Springer-Verlag, New York, Vol. 22, (1990).*

[6] D.P. Bertsekas, J.N. Tsitsiklis Gradient Convergence in Gradient Methods with Errors, *SIAM J. Optimization 10(2) (2000), 627-642.*

[7] J.R. Birge, F. Louveaux, Introduction to Stochastic Programming, Springer 1997.

[8] R. H. Byrd, G. M. Chin, W. Neveitt, J. Nocedal, On the Use of Stochastic Hessian Information in Optimization Methods for Machine Learning, *SIAM J. Optim., 21 (3), (2011) pp. 977995.*

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

[10] R. H. Byrd, S. L. Hansen, J. Nocedal, Y. Singer, A Stochastic Quasi-Newton Method for Large-Scale Optimization, *Technical report, arXiv:1401.7020 [math.OC].*

[11] B. Delyon, A. Juditsky, Accelerated stochastic approximation, *SIAM J. Optimization, Vol.3, No.4, 1993, pp. 868-881.*

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

[13] V. Fabian, On Asymptotic Normality in Stochastic Optimization, *The Annals of Mathematical Statistics, Vol. 39, 1968, pp. 1327-1332.*

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

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

[16] S. Ghadimi, G. Lan, H. Zhang, Mini-batch Stochastic Approximation Methods for Nonconvex Stochastic Composite Optimization, *Technical report, arXiv:1308.6594 [math.OC].*

[17] S. Ghadimi, G. Lan, Stochastic first and zeroth-order methods for nonconvex stochastic programming, *SIAM Journal on Optimization 23(4) (2013), pp. 2341-2368.*

[18] T. Homem-de-Mello, On rates of convergence for stochastic optimization problem under non-independent and identocally distributed sampling, *SIAM J. Optim. Vol. 19, No. 2 (2008), pp. 524-551.*

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

[20] H. Kesten, Accelerated stochastic approximation, *Ann. Math. Statist., 29, 1958, pp. 41-59.*

[21] J. Kiefer, J. Wolfowitz, Stochastic Estimation of the Maximum of a Regression Function, *The Annals of Mathematical Statistics 23(3) (1952), 462-466.*

[22] 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.*

[23] N. Krejić, N. Krklec Jerinkić, Nonmonotone line search methods with variable sample size, *T*echnical report, `http://www.optimization-online.org/DB_HTML/2013/05/3902.html`

[24] N. Krejić, Z. Lužanin, I. Stojkovska, A gradient method for unconstrained optimization in noisy environment, *Applied Numerical Mathematcs 70 (2013), 1-21.*

[25] H. Kushner, Stochastic Approximation: A survey, Computational Statistics, 2(1) (2010), 87-96.

[26] K. Marti Stochastic Optimization Methods, Springer 2005.

[27] A. Nemirovski, A. Juditsky, G. Lan, A. Shapiro, Robust stochastic approximation approach to stochastic programming, *SIAM Journal on Optimization* Vol. 19, No. 4 (2009), pp. 1574-1609.

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

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

[30] H. Robbins, S. Monro A Stochastic Approximation Method, *The Annals of Mathematical Statistics, 22(3) (1951), 400-407.*

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

[32] A. Shapiro, D. Dentcheva, A. Ruszczynski Lectures on Stochastic Programming, *SIAM 2009.*

[33] A. Shapiro, Y. Wardi, Convergence Analysis of Stochastic Algorithms, *Mathematics of Operations Research 21(3) (1996), 615-628.*

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

[35] J. C. Spall, Implementation of the Simultaneous Perturbation Algorithm for Stochastic Optimization, *IEEE Transactions on Aerospace and Electronic Systems 34 (3) (1998), 817-823.*

[36] F. Yousefian, A. Nedic, U.V. Shanbhag, On stochastic gradient and subgradient methods with adaptive steplength sequences, *Automatica 48 (1),2012, pp. 56-67.*

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

[38] Z. Xu, Y.H.Dai New stochastic approximation algorithms with adaptive step sizes, *Optimization Letters 6 (2012), 1831-1846.*

[39] X. Wang, S. Ma, Y. Yuan, Penalty Methods with Stochastic Approximation for Stochastic Nonlinear Programming, *Technical report, arXiv:1312.2690 [math.OC].*

[40] Y. Wardi, Stochastic Algorithms with Armijo Stepsizes for Minimization of Functions, *Journal of Optimization Theory and Applications 64(2) (1990), 399-417.*