ON THE EFFICIENCY OF A RANDOM SEARCH METHOD

G. Pérez-Lechuga\textsuperscript{1}, J.C.S. Tuoh-Mora \textsuperscript{1}, E. Morales-Sánchez \textsuperscript{1}, M. M. Álvarez Suárez\textsuperscript{1},

\textsuperscript{1} Centro de Investigación Avanzada en Ingeniería Industrial
Universidad Autónoma del Estado de Hidalgo
ICBI, CIAII, Carretera Pachuca-Tulancingo Km.4.5, Pachuca Hgo, México.

Abstract:- This paper studies the efficiency of the random search reported by Rubinstein [1] and widely studied by Pérez-Lechuga [2]. We proof that the efficiency of the selected random search algorithm is a linear function both of the step size and the direction of the descent movement. We report the theoretical results.

Key Words:- Stochastic optimization, random search, stochastic approximation, gradient approximation, quasigradient method.

1. Introduction

Stochastic approximation algorithms can be used in system optimization problems for which only noisy measurements of the system are available without knowing the gradient of the objective function. This type of problem can be found in adaptative control, neural network training, experimental design, stochastic optimization and many other areas.

The main idea of the stochastic quasigradient methods is to solve a wide class of optimization problems with a complex nature of objective functions and constraints. These methods are stochastic algorithmic procedures for solving general constrained problems with nondifferentiable, nonconvex functions.

For stochastic programming problems, these techniques generalize the well known stochastic approximation method for unconstrained optimization of the expectation of a random function to problems involving general constraints.

Consider the general stochastic programming problem

\[
\text{Minimize } F_0(x) = \mathbb{E}[f_0(x, \omega)], \quad (1)
\]

Subject to \( x \in S \subset \mathbb{R}^n \), where

\[
S = \{ x \mid f_i(x, \omega) \leq 0, \ i = 1, \ldots, m \}, \quad (2)
\]

\( \mathbb{E} \) is the operation of mathematical expectation with respect of some probability –
space $(\Omega, \mathcal{A}, P)$, and $\omega \in \Omega$.

The more trouble on solving the problem (1) to (2) is that, it is only feasible to calculate the exact values of the functions
\[ F_i(x) = E[f_i(x, \omega)] = \int f_i(x, \omega) P(d\omega), \]
\[ i = 0, \ldots, m \]
in exceptional cases for special types of probability measures $P(\omega)$.

If the functions $F_i(x)$ have uniformly bounded second derivatives at $x \in \{x_s\}_{s=0}^{\infty}$ then for the random vectors $\xi_i(s)$ defined as (see [3])
\[ \sum_{j=1}^{n} \frac{f_i(x_s + \Delta_s e_j, \omega_s) - f_i(x_s, \omega_s)}{\Delta_s}, \quad (3) \]
we would have $E[\xi_i(s) \mid x_s] = F_i(x_s) + b_i(s), \| b_i(s) \| \leq \Delta_s$. Where $e_j$ is the unit vector on the $j$th axis and $\Delta_s$ is a positive constant.

The random sequence $x_{s+1} = x_s - \rho_s \xi_s, s = 0, 1, \ldots$ converges with probability 1 to the solution of (1) if the following conditions are satisfied with probability 1. For the step size: $\rho_s \geq 0$, $\sum_s \rho_s = \infty$, $\sum_j E[\rho_s \| \Delta_s \| + \rho_s^2] < \infty$. For the quasi-gradient $\xi_s$, $E[\xi_s \mid x^*] = \nabla F_0(x^*) + o(\Delta_s)$.  

2. The random search algorithms

In this section we introduce the random search algorithms for optimization problems, in which the computing cost of the random search at a point increases as the point tends to satisfy appropriate optimality conditions. The algorithms progress faster beginning from initial conditions far away from an optimal or suboptimal point, and they gain precision with expense of efficiency as such a point is approached. The algorithms start at some $x_0 \in \mathcal{S}$, and they generate a sequence $x_0, x_1, \ldots \in \mathcal{S}$. These are descent algorithms, in the sense that the sequences $F(x_0), F(x_1), \ldots$, are monotone decreasing. They generate $x_{i+1}$ from $x_i$ by random search techniques, using the sufficient descent principle. This principle enables that the algorithms do not adopt the first point $y \in \mathbb{R}^n$ found by random search satisfying $F(y) < F(x_i)$ as $x_i$, but they rather wait to find a point $y$ for which $F(x_i) - F(y)$ is large by some criteria [4]. The amount of descent $F(x_{i+1}) - F(x_i)$, like the amount of time the algorithm spent for the random search at $x_i$, depends on the desired extent for $x_i$; the less desirable $x_i$, the larger the descent will tend to be.

Random search techniques have been an object of research for quite some time. The concept has been initially introduced by Anderson [5] and then developed by Rastgrin [6]. The idea is to determine a descent direction at random, by using a distribution on the unit sphere around $x_i$, and then, to find a suitable step size. The step size is used by minimizing $F$ along the descent direction. This is determined adaptively, based on the ratio of successful to failed attempts (by random searches) to reduce $F$. The basic property of this algorithms is that $x_i$ reaches a solution of (1) with probability 1 using a prescribed tolerance as $i \to \infty$. Thus in descent algorithms, one random search is conducted at $x_i$ to generate $x_{i+1}$. For the convergence states, the probability of $x_i$ satisfying $f(x_i) \leq \inf \{ f(x) \mid x \in \mathcal{S} \} + \epsilon$ (for a given $\epsilon > 0$) approaches 1 as $i \to \infty$. 

2
2.1. Stochastic Quasigradient Methods

Stochastic quasigradient methods are a set of techniques useful to solve optimization problems with objective functions and constraints of such a complex nature which make impossible to calculate the precise values of these functions (let alone of their derivatives). The basic idea is to use statistical estimates for the values of the functions rather than precise values. For stochastic programming problems these methods generalize the well-known stochastic approximation method for unconstrained optimization of the expectation of a random function. This stochastic problem can be defined as follows.

$$\min \{ \mathbb{E}_\omega f(x, \omega) : x \in S \},$$

(4)

where $x$ represents the variable to be chosen optimally, $S$ is a set of constraints, and $\omega$ is a random variable belonging to some probabilistic space $(\Omega, \mathcal{B}, P)$. Here $\mathcal{B}$ is a Borel field and $P$ is a probabilistic measure. In this problem we assume that $S$ is defined in such a way that the projection operation $x \rightarrow \Pi_S(x)$ is comparatively inexpensive from a computational point of view, where

$$\Pi_S(x) = \arg\min_{Z \in S} \| x - Z \|$$

In this case it is possible to implement a stochastic quasigradient algorithm of the following type

$$x_{i+1} = \Pi_S(x_i - \rho_i \varphi_i),$$

(5)

Here $x_i$ is the current approximation of the optimal solution, $\rho_i$ is the steep size, and $\varphi_i$ is a random step direction. This step direction may, for instance, be a statistical estimate of the gradient (or subgradient in the nondifferentiable case) of $f(x)$, then $\varphi_i \equiv \xi_i$, such that

$$\mathbb{E}(\xi_i | x_1, \ldots, x_i) = \nabla F_i(x_i) + a_i,$$

(6)

$$i = 0, \ldots, m$$

where $a_i$ decreases for an increasing number of iterations, the vector $\xi_i$ is called a stochastic quasigradient of functions $F_i(x)$, and $\nabla F(x)$ is the subgradient of $F(x)$ in each point $x_i$. Usually $\rho_i \rightarrow 0$ as $i \rightarrow \infty$ and therefore $\| x_{i+1} - x_i \| \rightarrow 0$.

Algorithm (5) can also resolve with problems with more general constraints formulated in terms of mathematical expectations $\mathbb{E}_\omega [f_i(x, \omega) \leq 0], i = 1, \ldots, m$, by making use of penalty functions or Lagrangians.

3. Problem definition

The idea of efficiency was introduced by Rubinstein et al. [1] as follows. Let $x_{i+1}$ be the point reached after one single iteration, and $\Delta F_i = F_{i+1} - F_i$ the increment of the value of $F$. The efficiency of the random search algorithms is defined as

$$C = -\frac{\mathbb{E}(\Delta F_i)}{\mathbb{E}(N_i)},$$

(7)

$$D = C [\mathbb{Var} \Delta f_i]^{-1/2},$$

(8)

where $N_i$ is the number of observations (measurements) of the convex function $F(x)$ required for the algorithm at the $i$th step.

In this paper we are interested in evaluate the efficiency of the following algorithm [2]:
\[ x_{i+1} = x_i - \alpha_i \gamma_i \xi_i, \quad (9) \]

where \( \alpha_i \) is the step size, \( \gamma_i \) is a normalization factor proposed by [1], \( \xi_i = Y_{in}^0 B_{il}^0 \), and \( Y_{il}^0 = \min\{Y_{il}, \ldots, Y_{il}^d\} \) denotes the difference \( \Psi \), \( f \) and \( \nabla f(x_i) \) are independent and uniformly distributed vectors on the surface of such sphere. We assume that \( f(x, W) = f(x) + W \), where \( W \sim N(0, \sigma^2) \). From the convexity of \( f \) we have

\[ f(x_i + \Delta x_i) - f(x_i) \geq \langle \Delta x_i, \nabla f(x_i) \rangle \]

or equivalently

\[ f(x_{i+1}) = f(x_i + \Delta x_i) = f(x_i) + \langle \Delta x_i, \nabla f(x_i) \rangle + \delta(\Delta x_i) = f(x_i) + \| \Delta x_i \| \| \nabla f(x_i) \| \cos \theta + \delta(\Delta x_i), \]

where \( \cos \theta \) is the angle between the unit vectors \( \Delta x_i \) and \( \nabla f(x_i) \), and \( \delta(\Delta x_i) \to 0 \) as \( \| \Delta x_i \| \to 0 \). We analyze two cases. In the first, we consider the noise \( W = 0 \), and in the second, \( W \sim N(0, \sigma^2) \).

**First case:** \( W = 0 \). From (9), note that

\[ \Delta x_i = x_{i+1} - x_i = -\alpha_i \gamma_i Y_{il}^0 B_{il}^0, \quad (11) \]

Substituting (10) in (9) we obtain

\[ f(x_{i+1}) = f(x_i) + \alpha_i \gamma_i Y_{il}^0 \left\| B_{il}^0 \right\| \| \nabla f(x_i) \| \cos \theta + \delta(\Delta x_i) \]

Taking into account that \( \left\| B_{il}^0 \right\| = 1 \), and for \( \Delta x_i \) sufficiently small, then

\[ f(x_{i+1}) = f(x_i) + \alpha_i \gamma_i Y_{il}^0 \cos \theta, \]

therefore

\[ \Delta f_i = \alpha_i \gamma_i Y_{il}^0 \cos \theta. \]

Thus, by (7)

\[ C_i = \frac{E[\langle f(x_i) - f(x_{i+1}) \rangle]}{\mathcal{H}} = \frac{E[\Delta f_i]}{\mathcal{H}} \]

\[ = \frac{1}{\mathcal{H}} \alpha_i \gamma_i Y_{il}^0 E[\cos \theta] \quad (12) \]

where the probability density function of the random angle is given by (see [7])

\[ \zeta_n(\theta) = \frac{\sin^{n-2}(\theta)}{\int_0^{\pi/2} \sin^{n-2}(\theta) d\theta} \]

\[ = \varphi_n \sin^{n-2}(\theta), \quad -\pi/2 \leq \theta \leq \pi/2, \quad (13) \]

where

\[ \varphi = \frac{\Gamma(n/2)}{\sqrt{\pi} \Gamma((n-1)/2)}, \]

and \( \Gamma \) denotes the gamma function.

If \( h_i = \alpha_i \gamma_i Y_{il}^0 \), then substituting (13) in (12) we have that

\[ E[\Delta f_i] = h_i \int_{-\pi/2}^{\pi/2} \cos(\theta) \zeta_n(\theta) d\theta = 2\varphi h_i \int_0^{\pi/2} \cos \theta \sin^{n-2}(\theta) d\theta = \frac{2\varphi h_i}{n-1}, \]

therefore (7) takes the form

\[ -\frac{E(\Delta f_i)}{E(N_i)} = \frac{2\varphi h_i}{\mathcal{H}(n-1)}, \quad (14) \]
Second case: \( W \sim N(0, \sigma^2) \). Consider the event
\[
x_{i+1} = \begin{cases} x_i - h_i B_{il}^0, & \text{if } \Upsilon = \min \{ \Upsilon_{il} \} \\ x_i, & \text{in other case} \end{cases}
\]

(15)

note that for any \( i \) iteration, \( \Upsilon_{il} \) is such that
\[
\Upsilon_{il} = \psi(x_i + B_{il}^0, W_{il}) - \psi(x_i, W_{il}) = f(x_i + B_{il}^0) + W_{il} - f(x_i) - W_{il} = f(x_i + B_{il}) - f(x_i)
\]

(16)

where \( W_{il} \) and \( W_{i0} \) are the realizations of the noise observed at points \( x_i + B_{il} \) and \( x_i \). The probability of this event is (see [2])
\[
P = \frac{1}{2} \left( 1 + \phi \left( \frac{|\Delta f_i|}{2\sigma} \right) \right).
\]

where \( \phi(y) = \int_0^y 2\pi^{-1/2}e^{-t^2}dt \). Thus
\[
P = \frac{1}{2} \left( 1 + \phi \left( \frac{|h_i \cos \theta|}{2\sigma} \right) \right).
\]

As in [1], let \( Q \) be the random variable defined by
\[
Q = \begin{cases} \cos \theta, & \text{with probability } \theta \\ -\cos \theta, & \text{with probability } (1-P) \end{cases}
\]

for \(-\pi/2 \leq \theta \leq \pi/2\). Then, taking the mathematical expectation in \( Q \) we obtain
\[
\mathbb{E}[Q] = \int_{-\pi/2}^{\pi/2} \theta P[\cos \theta \sin^{-2}(\theta)]d\theta - \int_{-\pi/2}^{\pi/2} \theta (1-P)[\cos \theta \sin^{-2}(\theta)]d\theta = 2\theta \int_{0}^{\pi/2} P[\cos \theta \sin^{-2}(\theta)]d\theta - 2\theta \int_{0}^{\pi/2}(1-P)[\cos \theta \sin^{-2}(\theta)]d\theta.
\]

After some algebraic manipulations, we have
\[
\frac{\mathbb{E}[Q]}{2\varrho} = \int_{0}^{\pi/2} \cos \theta \sin^{-2}(\theta) \phi \left( \frac{|\Delta f_i|}{2\sigma} \right) d\theta = \int_{0}^{\pi/2} \cos \theta \sin^{-2}(\theta) \phi \left( \frac{|h_i \cos \theta|}{2\sigma} \right) d\theta.
\]

(17)

Then, \( C_i \) takes the form
\[
\frac{2\varrho}{H} \int_{0}^{\pi/2} \cos \theta \sin^{-2}(\theta) \phi \left( \frac{|h_i \cos \theta|}{2\sigma} \right) d\theta.
\]

(18)

In the final analysis we let us estimate the variance of \( \Delta f_i \) from \( \mathbb{E}[Q^2] - \mathbb{E}[Q]^2 \). Note that
\[
\mathbb{E}[Q^2] = 2\varrho \int_{0}^{\pi/2} P \cos^2 \theta \sin^{-2}(\theta)d\theta - 2\varrho \int_{0}^{\pi/2} (1-P) \cos^2 \theta \sin^{-2}(\theta)d\theta = 2\varrho \int_{0}^{\pi/2} \phi \left( \frac{|h_i \cos \theta|}{2\sigma} \right) \cos^2 \theta \sin^{-2}(\theta)d\theta
\]

Since, for \( x \) small,
\[
\phi(x) \approx \frac{1}{2} + \frac{1}{\sqrt{2\pi}}(x - \frac{x^3}{3} + \ldots),
\]

(19)

then \( \mathbb{E}[Q] \) can be written as
\[
\varrho \left[ \frac{1}{n-1} + \frac{h_i}{\sigma\sqrt{2\pi}} \int_{0}^{\pi/2} \cos^2 \theta \sin^{-2}(\theta)d\theta \right]
\]

and \( \mathbb{E}[Q^2] \) is defined by
\[
\varrho \left[ \int_{0}^{\pi/2} \sin^{n+1}(\theta)d\theta \right. + \frac{h_i}{\sigma\sqrt{2\pi}} \int_{0}^{\pi/2} \cos^3 \theta \sin^{-1}(\theta)d\theta \right],
\]

5
thus (7) takes the form
\[ C_n = \frac{\varrho_n}{\mathcal{H}(n-1)} + \frac{h_i \varrho_n}{\mathcal{H}(n-1) \sqrt{2\pi}} \int_0^{\pi/2} \cos^2 \theta \sin^{n-2}(\theta) d\theta. \] (20)

Finally, and using (19), (7) can be defined as
\[ D_n = \frac{\varrho}{\mathcal{H}(n-1) \text{Var}[Q]} + \frac{h_i \varrho_n}{\mathcal{H}(n-1) \text{Var}[Q]} \int_0^{\pi/2} \cos^2 \theta \sin^{n-2}(\theta) d\theta. \] (21)

Where (20) and (21) can be written in the linear form
\[ C_n = a_n + b_n h_i, \quad D_n = a_n' + b_n' h_i, \] (22)

where \( a_n = \frac{\varrho}{\mathcal{H}(n-1)} \), and \( a_n' = \frac{\varrho}{\mathcal{H}(n-1) \text{Var}[Q]} \).

\[ b_n = \frac{h_i \varrho_n}{\mathcal{H}(n-1) \sqrt{2\pi}} \int_0^{\pi/2} \cos^2 \theta \sin^{n-2}(\theta) d\theta, \] and

\[ b_n' = \frac{h_i \varrho_n}{\mathcal{H}(n-1) \sqrt{2\pi}} \int_0^{\pi/2} \cos^2 \theta \sin^{n-2}(\theta) d\theta. \]

Table 1 shows some values of \( C_n \). Here, 
\[ \varrho_n = \int_0^{\pi/2} \cos^2 \theta \sin^{n-2}(\theta) d\theta. \]

<table>
<thead>
<tr>
<th>( n )</th>
<th>( \varrho_n )</th>
<th>( \vartheta_n )</th>
<th>( C_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.2837</td>
<td>0.7853</td>
<td>0.1283 + 0.0402 h_i</td>
</tr>
<tr>
<td>3</td>
<td>0.5641</td>
<td>0.3331</td>
<td>0.0282 + 0.0749 h_i</td>
</tr>
<tr>
<td>4</td>
<td>0.5641</td>
<td>0.1963</td>
<td>0.0188 + 0.0441 h_i</td>
</tr>
<tr>
<td>5</td>
<td>0.1880</td>
<td>0.1333</td>
<td>4.7 E-3 + (1E-3) h_i</td>
</tr>
<tr>
<td>6</td>
<td>0.0940</td>
<td>0.0981</td>
<td>1.8 E-2 + (3.6E-4) h_i</td>
</tr>
</tbody>
</table>

Table 1: Efficiency of \( C_n \)

4. Conclusions
For the algorithm presented, the efficiency of the search can be viewed as a linear function of the \( h_i \) parameter, and of the evaluated points on the surface of the unit sphere. Where \( h_i = \alpha_i \gamma_i T^0_{im}. \)

5. References


