Journal of Statistical Planning and Inference 142 (2012) 633–644

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A generalization of the Solis–Wets method

Miguel de Carvalho

Institute of Mathematics, Analysis, and Applications, Ecole Polytechnique Fédérale de Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 13 April 2011
Received in revised form
4 May 2011
Accepted 31 August 2011
Available online 1 October 2011

Keywords:
Extremum estimators
Improving hit-and-run algorithm
Solis–Wets method
Stochastic optimization
Zigzag algorithm

ABSTRACT

In this paper we focus on the application of global stochastic optimization methods to extremum estimators. We propose a general stochastic method—the master method—which includes several stochastic optimization algorithms as a particular case. The proposed method is sufficiently general to include the Solis–Wets method, the improving hit-and-run algorithm, and a stochastic version of the zigzag algorithm. A matrix formulation of the master method is presented and some specific results are given for the stochastic zigzag algorithm. Convergence of the proposed method is established under a mild set of conditions, and a simple regression model is used to illustrate the method.

1. Introduction

Extremum estimators are one of the most extensive classes of methods for estimating the parameters of a statistical model of interest (Amemiya, 1985; Newey and McFadden, 1994; Andrews, 1999; Romano and Shaikh, 2010). The ordinary least squares, the generalized method of moments, and maximum likelihood methods are defined by the solution of an optimization problem of interest, and thus are instances of extremum estimators. One advantage of this general class of estimators is its elegant asymptotic theory, which reduces to a set of general results. Despite their appealing features, in many instances of interest these estimators are analytically intractable, and we frequently lack a closed-form solution for computing estimates based on extremum estimators. An approach to overcome this problem is to rely on optimization algorithms and obtain such estimates computationally. Two questions then arise. First, is there a method which outperforms all others? Second, what type of algorithm should we use to perform the optimization? An answer to the first question is provided by the No Free Lunch theorem—an impossibility result which precludes the existence of a general purpose strategy, robust *a priori* to any type of optimization problem (Wolpert and Macready, 1997). In the second question, a major concern is related with convergence features of the method used to iterate towards the optimal solution. If the method converges to a local solution, consistency of the extremum estimator is no longer ensured (cf. Newey and McFadden, 1994; Gan and Jiang, 1999). Hence, one should avoid to rely on optimization methods which may converge to a local solution, since it is the global solution that has noteworthy asymptotic features. Two types of algorithm are typically adopted to tackle such problems, namely deterministic and stochastic optimization methods. The former includes the Newton–Raphson algorithm, the steepest descent method, among many others (Nocedal and Wright, 1999). In this paper we focus on stochastic optimization algorithms, and these include random search methods (Spall, 2003; Zabinsky, 2003), the simulated annealing technique (Bohachevsky et al., 1986), the improving hit-and-run algorithm (Zabinsky et al., 1993; Andersen and Diaconis, 2007), the conditional Gaussian martingale algorithm (Esquível, 2006), among many others.

E-mail address: Miguel.Carvalho@epfl.ch

0378-3758/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2011.08.016

Stochastic search and optimization algorithms are applied in many fields. The topic includes applications ranging from game theory (Pakes and McGuire, 2001) to the clustering of multivariate data (Booth et al., 2008); an overview of stochastic search and optimization methods can be found, for instance, in Duflo (1996), Spall (2003), and Zabinsky (2003).

A major contribution of this paper is given by a master method from which several stochastic optimization algorithms are a special instance. The generality of the proposed method is enough to include the conceptual algorithm of Solis and Wets (1981) as a particular case, and we establish the convergence of our master method under a set of fairly mild assumptions. Just as a master key opens several locks, each of which also having its own key, the establishment of convergence of the master method allows us to reach the convergence of all the algorithms that it includes. An important instance of this method, to which we devote some attention, is the stochastic zigzag method—an algorithm largely inspired in the works of Mexia et al. (1999) and Pereira and Mexia (2010). We apply this instance of our method to obtain maximum likelihood estimates in a classical problem of statistical risk modeling related to NASA's (National Aeronautics and Space Administration) first shuttle tragedy (Dalal et al., 1989).

This paper is structured as follows. In the next section some preliminary concepts are introduced, and in Section 3 we revisit the meta-approach of Solis and Wets (1981). The master method is introduced in Section 4, and its convergence is established in Section 5. Concluding remarks are given in Section 6. Proofs are in Appendix.

2. Preliminaries

2.1. Problem formulation

We start with the definition of extremum estimator.

Definition 2.1. An estimator $\widehat{\boldsymbol{\theta}}_n$ is an extremum estimator if there is a parameter objective function $\mathcal{T}_n$, such that

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{T}_n(\boldsymbol{\theta}). \tag{1}$$

Some remarks on notation: $\boldsymbol{\Theta}\subseteq\mathbb{R}^k$ denotes a parameter space; n is the sample size; the true parameter will be denoted by $\boldsymbol{\theta}_o$. As a consequence of (1), the optimization problem at the center of our attention is

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{T}_n(\boldsymbol{\theta}). \tag{2}$$

For concreteness observe that the ordinary least squares estimator is obtained by setting $\mathcal{T}_n(\boldsymbol{\beta}) = -(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$, where $\mathbf{y}$, $\mathbf{X}$, and $\boldsymbol{\beta}$ denote the response vector, the design matrix, and the regression parameter, respectively.

Even though our main interest lies over the optimization problem (2), the procedures developed in this paper carry over *mutatis mutandis* to other unconstrained optimization problems of interest.

An alternative definition of extremum estimator is given by the following condition:

$$\mathcal{T}_n(\widehat{\boldsymbol{\theta}}_n) = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{T}_n(\boldsymbol{\theta}) + \mathrm{o_p}(1). \tag{3}$$

Here and below, we say that a random variable X_n is $\mathrm{o_p}(1)$, if for any $\varepsilon>0$ it holds that $\mathbb{P}[|X_n|>\varepsilon]\longrightarrow 0$, as $n\to\infty$. From the conceptual standpoint, the definition provided in (3) is preferable since it only requires that $\mathcal{T}_n(\widehat{\boldsymbol{\theta}}_n)$ is within $\mathrm{o_p}(1)$ from the global maximum of $\mathcal{T}_n(\widehat{\boldsymbol{\theta}}_n)$ is within $\mathrm{o_p}(1)$. This overcomes the question of existence and it is also more suitable for computational purposes (Andrews, 1999). To reduce the burden of notation, below we omit the subscript n in the extremum estimator and in the parameter objective function.

2.2. An overview of random search techniques

Suppose we have a random sample of size n from a population of interest, and we aim to estimate $\boldsymbol{\theta}_o$ by solving the optimization problem (2). From the conceptual standpoint, for a fixed n, we can think of the graph of $\mathcal{T}$, $\mathrm{gr}(\mathcal{T}) = \{(\boldsymbol{\theta},\mathcal{T}(\boldsymbol{\theta})) : \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$, as another population of interest from which we intend to consistently estimate the parameters:

$$\left(\arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{T}(\boldsymbol{\theta}),\ \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{T}(\boldsymbol{\theta})\right).$$

To do so, suppose that we collect a random sample $\{(\boldsymbol{\theta}_i,\mathcal{T}(\boldsymbol{\theta}_i)\}_{i=1}^{p}$ from such population. Hence for each sampled value $\boldsymbol{\theta}_i$, we also inquire its corresponding image value $\mathcal{T}(\boldsymbol{\theta}_i)$. Assume that such a sample is collected sequentially and that during each extraction period we compute

$$\tilde{\boldsymbol{\theta}}_{i+1} = \begin{cases} \boldsymbol{\theta}_0, & \Leftarrow i=0,\\ \tilde{\boldsymbol{\theta}}_i \mathbb{I}\{\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)\geq\mathcal{T}(\boldsymbol{\theta}_{i+1})\} + \boldsymbol{\theta}_{i+1}\mathbb{I}\{\mathcal{T}(\boldsymbol{\theta}_{i+1})>\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)\}, & \Leftarrow i\in\mathbb{N}, \end{cases} \tag{4}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. As we shall see below, the procedure described above contains the quintessence of the classical pure random search algorithm (Zabinsky, 2003, Section 2).

Classical random search algorithm

1. Choose an initial value of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, either randomly or deterministically. Set $i=0$ and $\tilde{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0$.
2. Generate a new independent value $\boldsymbol{\theta}_{i+1}$ from a probability distribution f, with support over $\boldsymbol{\Theta}$. If $\mathcal{T}(\boldsymbol{\theta}_{i+1}) > \mathcal{T}(\tilde{\boldsymbol{\theta}}_i)$, set $\tilde{\boldsymbol{\theta}}_{i+1} = \boldsymbol{\theta}_{i+1}$; else set $\tilde{\boldsymbol{\theta}}_{i+1} = \tilde{\boldsymbol{\theta}}_i$. Increment i.

The convergence of the algorithm stated above was established in the seminal work of Solis and Wets (1981). The crux of their work lies in the introduction of a conceptual algorithm which includes among others the above-mentioned algorithm. To shed some light on some standard variants of the classical random search algorithm, observe that

- other types of processes can be used in lieu of (4);
- independence in the choice of the values of $\boldsymbol{\theta}_i$ can be dropped;
- the probability distribution f can be allowed to have a support defined over $\mathbb{R}^k \supseteq \boldsymbol{\Theta}$.[1]

3. Revisiting the Solis–Wets framework

3.1. Preliminaries and notation

In the sequel we present some definitions and notation. We start by introducing the essential supremum, a concept which is related with the maximum. It turns out however that the essential supremum is more suited for computational purposes than the maximum itself; the concept of optimality region is also presented below. The following shorthand notation will be useful:

$$\nabla_t = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \mathcal{T}(\boldsymbol{\theta}) < t\}, \quad \overline{\nabla}_t = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \mathcal{T}(\boldsymbol{\theta}) \le t\},$$

$$\Delta_t = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \mathcal{T}(\boldsymbol{\theta}) > t\}, \quad \underline{\Delta}_t = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \mathcal{T}(\boldsymbol{\theta}) \ge t\}. \tag{5}$$

Definition 3.1. Let $\mathcal{T} : \boldsymbol{\Theta} \to \overline{\mathbb{R}}$ be a measurable function. The essential supremum is defined as

$$\operatorname*{ess\,sup}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{T}(\boldsymbol{\theta}) \equiv \inf\{t : \mathcal{T} \le t, \text{ a.e.}\}.$$

Remark 3.1. Observe that the essential supremum is tantamount to

$$\sup\{t : \lambda(\Delta_t) > 0\},$$

where $\lambda(\cdot)$ denotes the Lebesgue measure. This type of representation is actually preferred by Solis and Wets (1981).

To ease notation, below we use τ to denote the essential supremum. Let us recollect two properties of the essential supremum which will be necessary in latter developments. First, if the maximizer of $\mathcal{T}$ is unique, and $\mathcal{T}$ is continuous, the essential supremum τ coincides with the maximum, i.e.

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{T}(\boldsymbol{\theta}) \Rightarrow \tau = \mathcal{T}(\widehat{\boldsymbol{\theta}}). \tag{6}$$

Second, the measurable function $\mathcal{T}$ cannot take values above its essential supremum, except on a measure-zero set:

$$\mathcal{T} \le \tau, \text{ a.e.} \tag{7}$$

These properties can be found for instance in Capiński and Kopp (1998, pp. 66 and 289). We now formally define the concept of optimality zone.

Definition 3.2. Let τ denote the essential supremum of $\mathcal{T}$. The optimality zone for the argument of the maximum of $\mathcal{T}$ is given by the set-valued function $\mathcal{O} : \mathbb{R}^2_+ \rightrightarrows \boldsymbol{\Theta}$, defined as

$$\mathcal{O}_{\varepsilon,M} = \begin{cases} \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \mathcal{T}(\boldsymbol{\theta}) > \tau - \varepsilon\} & \Leftarrow \tau \in \mathbb{R}, \\ \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \mathcal{T}(\boldsymbol{\theta}) > M\} & \Leftarrow \tau = +\infty. \end{cases}$$

The arguments of the optimality zone $\mathcal{O}_{\varepsilon,M}$ can be interpreted as a tolerance and a threshold, respectively. Using the notation introduced in (5), we can restate the optimality zones as

$$\mathcal{O}_{\varepsilon,M} = \begin{cases} \Delta(\tau-\varepsilon) & \Leftarrow \tau \in \mathbb{R}, \\ \Delta(M) & \Leftarrow \tau = +\infty. \end{cases}$$

The next definition closes our conceptual framework.

[1] Obviously, due adaptations are necessary; otherwise some problems can arise in Step 2; the lapse of such modifications may preclude the computation of the image for certain values of $\boldsymbol{\theta}$ which are not included in the domain of $\mathcal{T}$.

Definition 3.3. A function $\mathrm{c}:\boldsymbol{\Theta}\times\mathbb{R}^k\to\boldsymbol{\Theta}$ is a compass function if

$$\begin{cases}(\mathcal{T}\circ\mathrm{c})(\boldsymbol{\theta}_a,\boldsymbol{\theta}_b)\geq\mathcal{T}(\boldsymbol{\theta}_a), & \forall(\boldsymbol{\theta}_a,\boldsymbol{\theta}_b)\in\boldsymbol{\Theta}\times\mathbb{R}^k,\\ (\mathcal{T}\circ\mathrm{c})(\boldsymbol{\theta}_a,\boldsymbol{\theta}_b)\geq\mathcal{T}(\boldsymbol{\theta}_b), & \forall(\boldsymbol{\theta}_a,\boldsymbol{\theta}_b)\in\boldsymbol{\Theta}\times\boldsymbol{\Theta}.\end{cases}$$

Example 3.1. A simple example of compass function is the mapping:

$$\mathrm{c}(\boldsymbol{\theta}_a,\boldsymbol{\theta}_b)=\boldsymbol{\theta}_a\mathbb{I}_{\{\boldsymbol{\theta}_a\in\underline{\Delta}(\mathcal{T}(\boldsymbol{\theta}_b))\}}(\boldsymbol{\theta}_a,\boldsymbol{\theta}_b)+\boldsymbol{\theta}_b\mathbb{I}_{\{\boldsymbol{\theta}_a\in\nabla(\mathcal{T}(\boldsymbol{\theta}_b))\}}(\boldsymbol{\theta}_a,\boldsymbol{\theta}_b).$$

If we define

$$\tilde{\boldsymbol{\theta}}_{i+1}=\mathrm{c}(\tilde{\boldsymbol{\theta}}_i,\boldsymbol{\theta}_{i+1}),$$

then it holds that

$$\tilde{\boldsymbol{\theta}}_{i+1}=\tilde{\boldsymbol{\theta}}_i\mathbb{I}\{\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)\geq\mathcal{T}(\boldsymbol{\theta}_{i+1})\}+\boldsymbol{\theta}_{i+1}\mathbb{I}\{\mathcal{T}(\boldsymbol{\theta}_{i+1})>\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)\},$$

and so we recover the above-mentioned probabilistic recursive translation of the pure random search algorithms.

We refer to this function as a 'compass', because this is the mapping that guides the process of selection of the maxima.

4. The master method

4.1. Introducing the master method

Before we present the *modus operandi* of our master method some notation is necessary: $\mathbf{Z}$ will be used to denote the iterates of the algorithm; the parameter c controls the length of each run of the algorithm.

Modus operandi of the master method ($c\in\mathbb{N}$)

0. Set $i,j=1$. Find $\mathbf{a},\mathbf{b}\in\boldsymbol{\Theta}$, and set $\tilde{\boldsymbol{\theta}}_0$ and $\boldsymbol{\theta}_0$ equal to $\arg\max_{\mathbf{x}\in\{\mathbf{a},\mathbf{b}\}}\mathcal{T}(\mathbf{x})$. Further, set $\mathfrak{z}_1$ and $\mathbf{Z}_{1,1}$ equal to $\arg\min_{\mathbf{x}\in\{\mathbf{a},\mathbf{b}\}}\mathcal{T}(\mathbf{x})$.
1. If $c>1$, generate $\mathfrak{Z}_{i,j}$ from the probability space $(\mathbb{R}^k,\mathbb{B}(\mathbb{R}^k),\mathbb{P}_{i,j})$, and set $\mathbf{Z}_{i,j+1}=\mathfrak{Z}_{i,j}$. Else, go to Step 2.
2. If $j<c-1$, increment j, and return to Step 1. Otherwise, set $\tilde{\boldsymbol{\theta}}_i=\mathrm{c}(\tilde{\boldsymbol{\theta}}_{i-1},\boldsymbol{\theta}_i)$, where $\boldsymbol{\theta}_i=\arg\max_{q\in\{1,\ldots,c\}}\mathcal{T}(\mathbf{Z}_{i,q})$, and set $j=1$.
3. Generate $\mathfrak{z}_i$ from the probability space $(\mathbb{R}^k,\mathbb{B}(\mathbb{R}^k),\mathbb{P}_i)$, set $\mathbf{Z}_{i,1}=\mathfrak{z}_i$, increment i and j, and return to Step 1.

Some comments on this general algorithm are in order:

- The parameter c can be defined *a priori* by the user, and it can take any positive integer value. As a rule of thumb, we suggest taking c as random (e.g. drawn from discrete uniform distribution $U\{1,\ldots,k\}$).
- Observe that Step 0 simply initiates the algorithm. If we repeat Step 1 for a fixed i, we construct the iterates $\mathfrak{Z}_{i,1},\ldots,\mathfrak{Z}_{i,c-1}$. In Step 2 we update the compass and obtain the next 'candidate' for argument of the maximum, proposed by the algorithm. The repetition of Step 3 yields $\mathfrak{z}_2,\mathfrak{z}_3,\ldots$
- Here and below, we refer to each $\mathfrak{z}_i$ as a seed. If the seeds are independent and identically distributed, we refer to the master method as pure. If the probability measure $\mathbb{P}_p$ depends on some probability measure(s) $\mathbb{P}_q$, with $q<p$, then the master method will be called adaptive. Further, we refer to each $\mathbf{Z}_{i,j}$ as an iterate, and to each sequence $\mathbf{Z}_{i,1},\ldots,\mathbf{Z}_{i,c}$ as a course. Using this terminology, we can say that the consecutive repetition of Step 1 builds a course. Similarly, if we rerun serially Step 3 we obtain a sequence of seeds.
- The mechanics of the algorithm is perhaps better understood through the law of movement of the iterates:

$$\mathbf{Z}_{i,j}=\mathfrak{z}_i\mathbb{I}(j=1\vee c=1)+\mathfrak{Z}_{i,j-1}\mathbb{I}(j\in\{2,\ldots,c\}\wedge c>1). \tag{8}$$

To gain some insight on the mechanics of the method, consider the case wherein $c=1$. Throughout this section, this benchmark case will be invoked frequently; in this case we have that

$$\boldsymbol{\theta}_i=\arg\max_{q\in\{1\}}\mathcal{T}(\mathbf{Z}_{i,q})=\mathbf{Z}_{i,1}=\mathfrak{z}_i$$

and hence $\boldsymbol{\theta}_i=\mathfrak{z}_i=Z_{i,1}$. Additionally, j becomes inactive in the algorithm, since under these circumstances, Step 1 is never triggered. Consequently, for $c=1$ the algorithm can be equivalently rewritten as follows.

Modus operandi of the master method ($c=1$)

0. Set $i=1$. Find $\boldsymbol{\theta}_1\in\boldsymbol{\Theta}$, and set $\tilde{\boldsymbol{\theta}}_0=\boldsymbol{\theta}_1$.
1. Set $\tilde{\boldsymbol{\theta}}_i=\mathrm{c}(\tilde{\boldsymbol{\theta}}_{i-1},\boldsymbol{\theta}_i)$, and increment i.
2. Generate $\boldsymbol{\theta}_i$ from the probability space $(\mathbb{R}^k,\mathbb{B}(\mathbb{R}^k),\mathbb{P}_i)$, return to Step 1.

It turns out that this is precisely the Solis and Wets (1981, p. 19) method. Hence, the classical Solis–Wets conceptual algorithm is a particular case of our master method, when $c=1$. The master method is simply a generalization of this method which follows a course between any two seeds. These and other features will become more clear after the introduction of a matrix formulation of the master method in Section 4.3.

4.2. Stochastic zigzag methods

We will be particularly interested in the instance of the proposed method where Step 1 takes the form:

1. If $c>1$, generate $\alpha_{i,j}$ from the probability space $(\mathbb{R},\mathbb{B}(\mathbb{R}),\mathbb{P}_{i,j})$, and set $\mathbf{Z}_{i,j+1}=\alpha_{i,j}\boldsymbol{\theta}_{i-1}+(1-\alpha_{i,j})\eth_i$. Else, go to Step 2.

Essentially the layout of the stochastic zigzag algorithm is the following. In Step 0 we initialize the algorithm, and sample from the line which passes through the points $\boldsymbol{\theta}_1$ and $\eth_1$. The consecutive application of Step 1, simply collects a random sample of c points from such line. In Step 2, we refresh the compass function C, obtaining the next 'candidate', for the argument of the maximum, yield by the algorithm. We then move to Step 3, where a new seed is generated. Again, we sample from the line which passes through the estimated argument of the maximum of the previous line and the new generated seed, and repeat the procedure described above (eventually *ad infinitum*).

In Figs. 1 and 2, we exemplify the stochastic zigzag method using the Styblinski–Tang function (Spall, 2003, p. 46):

$$L(\theta_1,\theta_2)=\tfrac{1}{2}[\theta_1^4-16\theta_1^2+5\theta_1+\theta_2^4-16\theta_2^2+5\theta_2]. \tag{9}$$

Other variants of the stochastic zigzag algorithm are also included in the general method; for example, if $c=2$, we get the improving hit-and-run algorithm (Zabinsky et al., 1993; Andersen and Diaconis, 2007). We can also consider an alternative shape for the line, and the generality of the master method is such that we may even take a different shape per course. The specification given above was chosen because of its simplicity and ease of implementation (see Theorem 4.1 in Section 4.3).

4.3. Matrix formulation of the master method

In this subsection, we present a matrix representation of the master method. This conceptual framework will help us to clarify some features of the method, and, as we shall see latter, it reduces the burden of computational implementation. To be able to present this formulation, we need to consider a stopping time for the method, which we denote by r. From the theoretical standpoint, we can consider for instance the time of entry in the optimal zone. This can be defined for every $\varepsilon,M>0$, as

$$r_{\varepsilon,M}=\inf\{i\in\overline{\mathbb{N}}:\boldsymbol{\theta}_i\in\mathcal{O}_{\varepsilon,M}\},$$

where $\overline{\mathbb{N}}=\mathbb{N}\cup\{\infty\}$. It is simple to show that this is a stopping time with respect to the natural filtration $\mathbb{F}_i=\sigma(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_i)$; analogous stopping times can be found in introductory textbooks (Williams, 1991, Section 10.8), so we skip the details. The crux of the proof is given by observing that for every $i\in\overline{\mathbb{N}}$

$$\{r_{\varepsilon,M}\le i\}=\bigcup_{p=1}^{i}\{\boldsymbol{\theta}_p\in\mathcal{O}_{\varepsilon,M}\}\in\mathbb{F}_i.$$

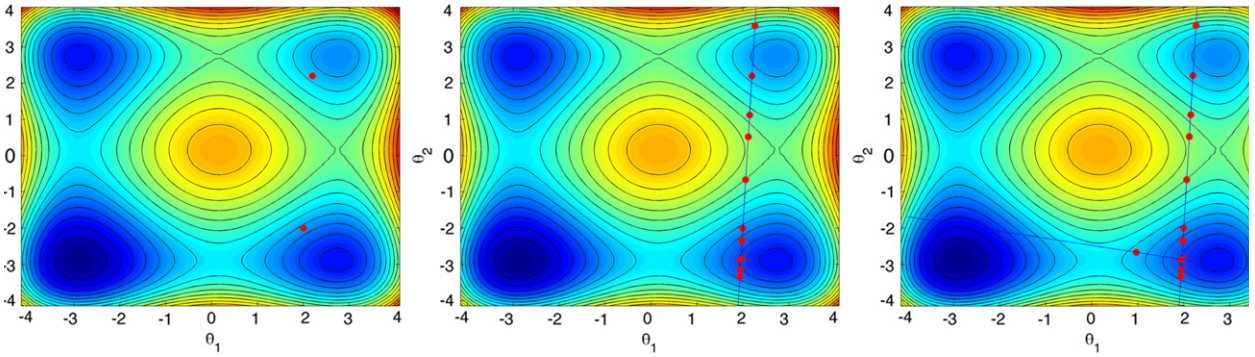

Fig. 1. The figure represents the initialization of the stochastic zigzag method. In the picture in the left we start with points **a** and **b** to initialize the search. The picture in the center illustrates that in Step 1 we collect a random sample ($c=10$) from the line which passes through **a** and **b**. The remaining picture depicts Steps 2 and 3 wherein after estimating the argument of the maximum of the first line, we generate another seed and extract a sample from the new line which passes through such points.

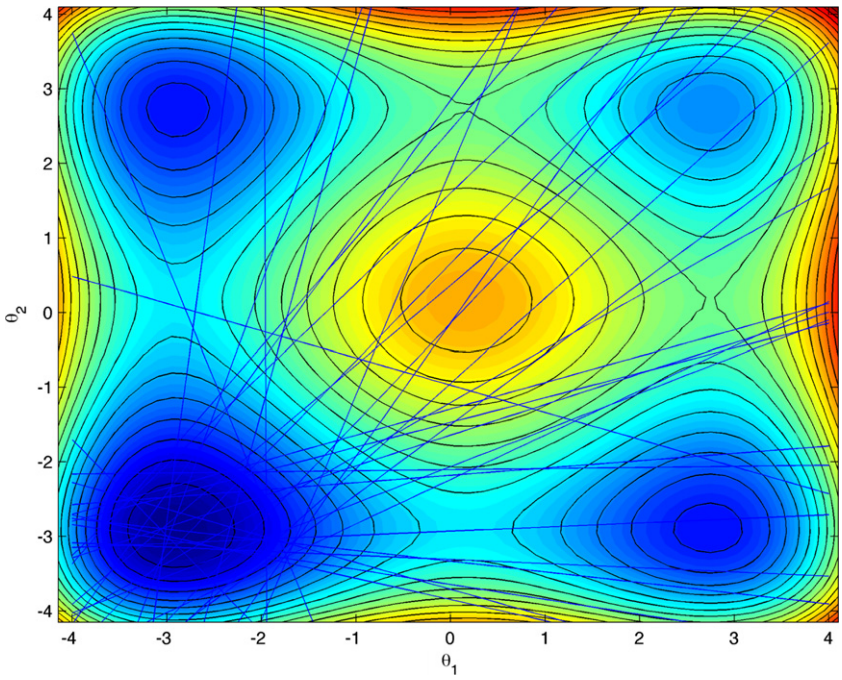

Fig. 2. The application of the stochastic zigzag method to the Styblinski–Tang function ($r=30$; $c=10$). This function pertains to a class which is typically used to assess the performance of an optimization algorithm (see e.g. Spall, 2003). The functional form of this function is given in formula (9).

The law of movement of the iterates (8) allows us to describe the mechanics of the method in a matrix form, by defining the iterative matrix $\mathbf{Z}$ as the $(r\times kc)$-matrix

$$\mathbf{Z}\equiv\begin{bmatrix}\mathbf{Z}_{1,1} & \mathbf{Z}_{1,2} & \cdots & \mathbf{Z}_{1,c}\\ \mathbf{Z}_{2,1} & \mathbf{Z}_{2,2} & \cdots & \mathbf{Z}_{2,c}\\ \vdots & \vdots & \vdots & \vdots\\ \mathbf{Z}_{r,1} & \mathbf{Z}_{r,2} & \cdots & \mathbf{Z}_{r,c}\end{bmatrix}=\begin{bmatrix}\vartheta_1 & \mathfrak{Z}_{1,1} & \cdots & \mathfrak{Z}_{1,c-1}\\ \vartheta_2 & \mathfrak{Z}_{2,1} & \cdots & \mathfrak{Z}_{2,c-1}\\ \vdots & \vdots & \vdots & \vdots\\ \vartheta_r & \mathfrak{Z}_{r,1} & \cdots & \mathfrak{Z}_{r,c-1}\end{bmatrix}=\begin{bmatrix}\mathbf{z}_1\\ \mathbf{z}_2\\ \vdots\\ \mathbf{z}_r\end{bmatrix}$$

and the map-iterative matrix $\mathbf{T}_{\mathbf{Z}}$ as the $(r\times c)$-matrix

$$\mathbf{T}_{\mathbf{Z}}\equiv\begin{bmatrix}\mathcal{T}(\mathbf{Z}_{1,1}) & \mathcal{T}(\mathbf{Z}_{1,2}) & \cdots & \mathcal{T}(\mathbf{Z}_{1,c})\\ \mathcal{T}(\mathbf{Z}_{2,1}) & \mathcal{T}(\mathbf{Z}_{2,2}) & \cdots & \mathcal{T}(\mathbf{Z}_{2,c})\\ \vdots & \vdots & \vdots & \vdots\\ \mathcal{T}(\mathbf{Z}_{r,1}) & \mathcal{T}(\mathbf{Z}_{r,2}) & \cdots & \mathcal{T}(\mathbf{Z}_{r,c})\end{bmatrix}=\begin{bmatrix}\mathbf{T}_{\mathbf{z}_1}\\ \mathbf{T}_{\mathbf{z}_2}\\ \vdots\\ \mathbf{T}_{\mathbf{z}_r}\end{bmatrix}.$$

For illustration let us rethink the case wherein $c=1$. Then the iterative matrix $\mathbf{Z}$ and the map-iterative matrix $\mathbf{T}_{\mathbf{Z}}$ become

$$\mathbf{Z}=\begin{bmatrix}\vartheta_1\\ \vartheta_2\\ \vdots\\ \vartheta_r\end{bmatrix},\quad \mathbf{T}_{\mathbf{Z}}=\begin{bmatrix}\mathcal{T}(\vartheta_1)\\ \mathcal{T}(\vartheta_2)\\ \vdots\\ \mathcal{T}(\vartheta_r)\end{bmatrix}.$$

The affinity between the Solis–Wets conceptual algorithm and the master method now becomes more clear. In the particular case $c=1$ the iterative matrix degenerates into a matrix composed uniquely by seeds, i.e. random draws generated from the probability space $(\mathbb{R}^k,\mathbb{B}(\mathbb{R}^k),\mathbb{P}_i)$.

In the stochastic zigzag method, the following matrix is also used

$$\boldsymbol{\alpha}\equiv\begin{bmatrix}\alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,c-1}\\ \alpha_{2,1} & \alpha_{2,2} & \cdots & \alpha_{2,c-1}\\ \vdots & \vdots & \vdots & \vdots\\ \alpha_{r,1} & \alpha_{r,2} & \cdots & \alpha_{r,c-1}\end{bmatrix}=\begin{bmatrix}\boldsymbol{\alpha}_1\\ \boldsymbol{\alpha}_2\\ \vdots\\ \boldsymbol{\alpha}_r\end{bmatrix}.$$

In the next theorem we show how the matrix representation of the stochastic zigzag method can ease its implementation.

Theorem 4.1 (*Kronecker–zigzag decomposition*). *The ith zigzag course can be rewritten as*

$$\mathbf{z}_i = [\eth_i \ \vdots \ \boldsymbol{\alpha}_i \otimes \boldsymbol{\theta}_{i-1} + (\mathbf{1}_{c-1}^{\mathrm{T}} - \boldsymbol{\alpha}_i) \otimes \eth_i], \quad i = 1, \ldots, r,$$

where $\boldsymbol{\theta}_{i-1}$ *is defined accordingly to the formulation of the stochastic zigzag method given above.*

The latter result warrants some comments. Roughly speaking, it states that the law of movement of each iterate, can be readily extended to describe the whole law of movement of a zigzag course by replacing the scalar product with the Kronecker product and performing the necessary scalar to vector adaptations. To put this differently, using the binary operation $\otimes$, we are able to build in a step each line of the iterative matrix $\mathbf{Z}$. The latter result thus allows us to easily implement computationally the algorithm by a 'loop' which is stated below in pseudocode.

Pseudocode implementation of the stochastic zigzag method

```
randomize:
   seeds;
   alpha.
for i=1 to r,
   compute θ_{i-1};
   compute z_i;
   increment i.
```

We give below an illustration of the Kronecker–zigzag decomposition and the stochastic zigzag method.

Example 4.1 (*Minimizing the Styblinski–Tang function*). Suppose that the following matrices were randomly generated

$$\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \boldsymbol{\alpha}_3 \end{bmatrix} = \begin{bmatrix} 2/3 & 1/3 \\ -1 & -1/3 \\ -2 & -1 \end{bmatrix}, \quad \eth = \begin{bmatrix} -4 & 1 \\ 0 & 0 \\ 2 & 0 \end{bmatrix}, \quad \tilde{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0 = [-1 \ \ 4].$$

By Theorem 4.1, it follows that

$$\mathbf{z}_1 = [\eth_1 \ \vdots \ \boldsymbol{\alpha}_1 \otimes \boldsymbol{\theta}_0 + (\mathbf{1}_2^{\mathrm{T}} - \boldsymbol{\alpha}_1) \otimes \eth_1] = [-4 \ \ 1 \ \ -2 \ \ 3 \ \ -3 \ \ 2].$$

Hence

$$\mathbf{T}_{\mathbf{z}_1} = [-15 \ \ -53 \ \ -58]$$

and so $\boldsymbol{\theta}_1 = [-3 \ \ 2]$. Similarly, we build second and third lines of the iterative matrix $\mathbf{Z}$:

$$\mathbf{z}_2 = [\eth_2 \ \vdots \ \boldsymbol{\alpha}_2 \otimes \boldsymbol{\theta}_1 + (\mathbf{1}_2^{\mathrm{T}} - \boldsymbol{\alpha}_2) \otimes \eth_2] = [0 \ \ 0 \ \ 3 \ \ -2 \ \ 1 \ \ -2/3].$$

This yields

$$\mathbf{T}_{\mathbf{z}_2} = [0 \ \ -53 \ \ -10.12],$$

so that $\boldsymbol{\theta}_2 = [3 \ \ -2]$. Finally, we have

$$\mathbf{z}_3 = [\eth_3 \ \vdots \ \boldsymbol{\alpha}_3 \otimes \boldsymbol{\theta}_2 + (\mathbf{1}_2^{\mathrm{T}} - \boldsymbol{\alpha}_3) \otimes \eth_3] = [2 \ \ 0 \ \ 0 \ \ 4 \ \ 1 \ \ 2].$$

Thus

$$\mathbf{T}_{\mathbf{z}_3} = [-19 \ \ 10 \ \ -24]$$

and so $\boldsymbol{\theta}_3 = [1 \ \ 2]$. Thus, we have the following iterative matrix $\mathbf{Z}$ and corresponding map-iterative matrix $\mathbf{T}_{\mathbf{Z}}$:

$$\mathbf{Z} = \begin{bmatrix} -4 & 1 & -2 & 3 & -3 & 2 \\ 0 & 0 & 3 & -2 & 1 & -2/3 \\ 2 & 0 & 0 & 4 & 1 & 2 \end{bmatrix}, \quad \mathbf{T}_{\mathbf{Z}} = \begin{bmatrix} -15 & -53 & -58 \\ 0 & -53 & -10.12 \\ -19 & 10 & -24 \end{bmatrix}.$$

Example 4.2 (*Maximum likelihood estimation*). We now consider an example of maximum likelihood estimation in a logistic regression model. The data are from 23 flights of the space shuttle Challenger previous to the accident of 1986, wherein the shuttle blew up during takeoff. On the morning of this catastrophic accident, the O-rings were 22 °F below the minimum temperature recorded in all the previous flights. There has been a large discussion in the literature about how to conduct a scientific risk analysis which allows one to predict the O-rings failure/success from its temperature (see, for instance, Dalal et al., 1989; Maranzano and Krzysztofowicz, 2008). The simplest possibility is by a logistic regression model where the variables of interest are the temperature of the primary O-rings of the space shuttle and an indicator of failure/success of the O-rings during takeoff; the data can be found in Christensen (1990, Section 2.6). The interest is thus in modeling the probability p_i that at least one O-ring fails, by taking temperature t_i as a covariate. This can be accomplished

by considering the model:

$$\log\left\{\frac{p_i}{1-p_i}\right\} = \theta_\alpha + \theta_\beta t_i, \tag{10}$$

where θ_α and θ_β, respectively, denote an intercept and a slope parameter. By (10), we can rewrite the probability that at least one O-ring fails, in case i, as

$$p_i = \frac{\exp\{\theta_\alpha + \theta_\beta t_i\}}{1+\exp\{\theta_\alpha + \theta_\beta t_i\}},$$

so that the log-likelihood can be written as

$$\ell = \sum_{i=1}^{n} \log\{p_i \times \mathbb{I}(\mathit{failure}_i)\} + \sum_{i=1}^{n} \log\{(1-p_i) \times \mathbb{I}(\mathit{success}_i)\}. \tag{11}$$

The estimation objects of interest are the intercept and slope parameters of model (10). Table 1 summarizes the estimates obtained by the application of the stochastic zigzag method to the log-likelihood (11). The results are from the averaging of a Monte Carlo simulation study with 500 trials; per each value of c we considered $r=1000$. As it can be observed from Table 1, the estimates are close to the ones presented by Christensen (1990, p. 56), namely $(\widehat{\theta}_\alpha,\widehat{\theta}_\beta)=(15.04,-0.2321)$. As pointed out by one of the reviewers, the comparison of the log-likelihood of our estimates and the one of $(\widehat{\theta}_\alpha,\widehat{\theta}_\beta)$ is critical for our comparison. Apart from the case $c=2$, where we obtain a slightly lower value, the log-likelihood of our average estimates gives -10.1576, which is the same value we would obtain with $(\widehat{\theta}_\alpha,\widehat{\theta}_\beta)$. To avoid unfair comparisons with the estimates reported in Christensen's book, we used the same number of decimal places in computing the log-likelihoods reported in Table 1.

4.4. A short note on the construction of confidence intervals

This subsection is devoted to the construction of confidence intervals for the maximum of $\mathcal{T}$, through the use of the image of the first column of the map-iterative matrix $\mathbf{T_Z}$. As we shall see below, if the master method is pure, and the seeds are uniformly distributed over $\boldsymbol{\Theta}$, then it is possible to take advantage of a result on extreme value theory due to de Haan (1981). In the sequel, let $\mathcal{T}_{\vartheta(1)} \leq \cdots \leq \mathcal{T}_{\vartheta(r)}$ denote the order statistics of the sequence of the image of the seeds, where r denotes a finite (possibly degenerated) stopping time.

Theorem 4.2 (*Confidence intervals for the maximum—de Haan, 1981*). *Consider the sequence of independent and identically distributed* ϑ_i *with uniform distribution over* $\boldsymbol{\Theta}$. *Further, consider the auxiliary set-valued function* $\Xi : \mathbb{N} \times [0;1] \rightrightarrows \mathbb{R}$, *defined as*

$$\Xi(i,p) = \left[\mathcal{T}_{\vartheta(i)}; \mathcal{T}_{\vartheta(i)} + \frac{\mathcal{T}_{\vartheta(i)} - \mathcal{T}_{\vartheta(i-1)}}{(1-p)^{-2/k}-1}\right].$$

The following large sample result holds

$$\mathbb{P}[\mathcal{T}_i(\boldsymbol{\theta}_o) \in \Xi(i,p)] - (1-p) = o(1).$$

Proof. See de Haan (1981, pp. 467–469). □

Remark 4.1. This result has also been used to conduct inference. For example, Veall (1990) relied on Theorem 4.2 to develop a statistical procedure for testing if a certain solution is global. Hence, if the method is pure and the seeds are uniformly distributed, Veall's test can be implemented here by using the first column of the iterative matrix $\mathbf{Z}$.

The method is extremely easy to apply using the following inputs: two order statistics $(\mathcal{T}_{\vartheta(r)},\mathcal{T}_{\vartheta(r-1)})$, level of significance p, and the dimension of the optimization problem k; further details on how to construct confidence intervals with Theorem 4.2, can be found in de Carvalho (2011).

Table 1
Estimates of the intercept and slope parameters $(\theta_\alpha,\theta_\beta)$, for the logistic regression model (11), obtained by the stochastic zigzag method.

Outputs from the Monte Carlo simulation study	c		
	2	3	4
Average estimate	(14.9121, −0.2302)	(14.9932,0.5871)	(14.9945, −0.2315)
Standard deviation	(−0.2302,0.0145)	(−0.2314,0.0086)	(−0.2315,0.0067)
Log-likelihood	−10.1578	−10.1576	−10.1576

5. Convergence

In this section we study the convergence of the general algorithm introduced above. We start with the introduction of some preliminary considerations. As a consequence of the compass updating rule of the master algorithm, $\tilde{\boldsymbol{\theta}}_i = \mathrm{c}(\tilde{\boldsymbol{\theta}}_{i-1}, \boldsymbol{\theta}_i)$, it holds that the sequence $\{\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)\}_{i\in\mathbb{N}}$ is increasing:

$$\mathcal{T}(\tilde{\boldsymbol{\theta}}_i) = (\mathcal{T}\circ\mathrm{C})(\tilde{\boldsymbol{\theta}}_{i-1}, \boldsymbol{\theta}_i) \geq \mathcal{T}(\tilde{\boldsymbol{\theta}}_{i-1}).$$

This reasoning can be extended by induction, being valid that for every positive integer κ

$$\mathcal{T}(\tilde{\boldsymbol{\theta}}_{i+\kappa}) \geq \mathcal{T}(\tilde{\boldsymbol{\theta}}_i).$$

This simple fact plays an important role in the establishment of the following trinity of elementary results.

Proposition 5.1. *For every positive integer κ, we have that:*

1. *If $\boldsymbol{\theta}_i \in \mathcal{O}_{\varepsilon,M}$, then $\tilde{\boldsymbol{\theta}}_{i+\kappa} \in \mathcal{O}_{\varepsilon,M}$.*
2. *If $\tilde{\boldsymbol{\theta}}_i \in \mathcal{O}_{\varepsilon,M}$, then $\tilde{\boldsymbol{\theta}}_{i+\kappa} \in \mathcal{O}_{\varepsilon,M}$.*
3. $\{\tilde{\boldsymbol{\theta}}_\kappa \in \mathcal{O}^c_{\varepsilon,M}\} \subseteq \{\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_{\kappa-1} \in \mathcal{O}^c_{\varepsilon,M}\} \cap \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{\kappa-1} \in \mathcal{O}^c_{\varepsilon,M}\}$.

Claims 1 and 2 of the foregoing theorem, translate the idea that if an iterate of the algorithm falls in the optimal zone, then it remains there forever. Claim 3 will be particularly useful in the proof of convergence of the general algorithm stated above.

Theorem 5.1 (*Convergence of the pure master method, Part I*).

1. *Suppose that $\mathcal{T}$ is bounded from above. Further, suppose that the master method is pure, and that* $\forall B \in \mathbb{B}(\boldsymbol{\Theta})\ \lambda(B) > 0 \Rightarrow \mathbb{P}[\mathfrak{d}_1 \in B] > 0$. *Then* $\mathbb{P}[\tilde{\boldsymbol{\theta}}_i \in \mathcal{O}^c_{\varepsilon,M}] = \mathrm{o}(1)$.
2. *Suppose that $\mathcal{T}$ is bounded from above. Then $\mathcal{T}(\tilde{\boldsymbol{\theta}}_i) - T = \mathrm{o}(1)$, a.s., where T is a random variable such that $\mathbb{P}[T = \tau] = 1$.*

The latter result warrants some remarks. Claim 1 states that the probability of failing to sample the optimality region, approaches 0 as the number of iterates increases, whereas Claim 2 ensures that the sequence $\{\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)\}_{i\in\mathbb{N}}$ converges a.s. to a random variable T which is indistinguishable from the essential supremum. The proof of Claim 2 is entirely robust to both the pure stochastic method and the adaptive master method, and so the second result also holds for the adaptive master method. It then arises the question. Is the first claim of the previous theorem also extendable to the adaptive master method? This question is addressed in the next theorem.

Theorem 5.2 (*Convergence of the adaptive master method, Part I*). *Suppose that $\mathcal{T}$ is bounded from above. Further, suppose that the master method is adaptive, and that* $\inf_{1\leq p\leq i-1} \mathbb{P}[\mathfrak{d}_p \in \mathcal{O}^c_{\varepsilon,M}] = \mathrm{o}(1)$. *Then* $\mathbb{P}[\tilde{\boldsymbol{\theta}}_i \in \mathcal{O}^c_{\varepsilon,M}] = \mathrm{o}(1)$.

It is important to underscore that the hypothesis considered here to establish the convergence of the adaptive master method is known in the literature, and is tantamount to the one used by Esquível (2006). Note however that whereas Esquível used this condition to establish the convergence of the adaptive random search, here it is used in the more general context of the adaptive master method.

Theorem 5.3 (*Convergence of the pure master method, Part II*). *Suppose that $\mathcal{T}$ is bounded from above. Further, suppose that master method is pure, and that* $\forall B \in \mathbb{B}(\boldsymbol{\Theta})\ \lambda(B) > 0 \Rightarrow \mathbb{P}[\mathfrak{d}_1 \in B] > 0$. *Further, suppose that $\mathcal{T}(\boldsymbol{\theta})$ is continuous and that $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{T}(\boldsymbol{\theta})$. Then, it holds that $\mathcal{T}(\tilde{\boldsymbol{\theta}}_i) - \mathcal{T}(\widehat{\boldsymbol{\theta}}) = \mathrm{o}(1)$, a.s. If furthermore $\boldsymbol{\Theta} \subset \mathbb{R}^k$ is compact, then $\tilde{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}} = \mathrm{o}(1)$, a.s.*

A similar result can be established if the master method is adaptive. Again, the framework need to be suitably accommodated by using Esquível's (2006) condition.

Theorem 5.4 (*Convergence of the adaptive master method, Part II*). *Suppose that $\mathcal{T}$ is bounded from above. Further, suppose that master method is adaptive, and that* $\inf_{1\leq p\leq i-1} \mathbb{P}[\mathfrak{d}_p \in \mathcal{O}^c_{\varepsilon,M}] = \mathrm{o}(1)$. *Suppose in addition that $\mathcal{T}(\boldsymbol{\theta})$ is continuous and that $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{T}(\boldsymbol{\theta})$. Then, it holds that $\mathcal{T}(\tilde{\boldsymbol{\theta}}_i) - \mathcal{T}(\widehat{\boldsymbol{\theta}}) = \mathrm{o}(1)$ a.s., as $i \to \infty$. If furthermore $\boldsymbol{\Theta} \subset \mathbb{R}^k$ is compact, then $\tilde{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}} = \mathrm{o}(1)$ a.s., as $i \to \infty$.*

6. Summary

This paper introduces the master method—a general algorithm which comprises several stochastic optimization algorithms as a particular case. The generality of the master method is considerable including, for instance, the conceptual algorithm of Solis and Wets (1981) and the improving hit-and-run algorithm (Zabinsky et al., 1993). Another specific

embodiment of the master method is provided by the stochastic zigzag method—an optimization algorithm which is based on the works of Mexia et al. (1999) and Pereira and Mexia (2010). We introduce a matrix formulation of the algorithm which brings new insights into the general method and diminishes the burden of implementation. The stochastic convergence of the master method is here achieved under a fairly mild set of conditions, and we illustrate the method by revisiting a classical problem in statistical risk modeling.

Acknowledgments

I am grateful to Tiago Mexia, Manuel Esquível, Viatcheslav Melas, Vanda Inácio, Anthony Davison, Narayanaswamy Balakrishnan, Miguel Fonseca, Feridun Turkman, and to five anonymous referees for helpful suggestions and recommendations that led to a significant improvement of this paper. Financial support from *Centro de Matemática e Aplicações*, *Universidade Nova de Lisboa* and *Fundação para a Ciência e Tecnologia* is greatly acknowledged.

Appendix

Proof of Theorem 4.1. Just note that

$$\begin{aligned}\mathbf{z}_i &= [\eth_i \ \vdots \ \alpha_{i,1}\boldsymbol{\theta}_{i-1}+(1-\alpha_{i,1})\eth_i \ \cdots \ \alpha_{i,c-1}\boldsymbol{\theta}_{i-1}+(1-\alpha_{i,c-1})\eth_i]\\ &= [\eth_i \ \vdots \ \alpha_{i,1}\boldsymbol{\theta}_{i-1} \ \cdots \ \alpha_{i,c-1}\boldsymbol{\theta}_{i-1}]+[\mathbf{0} \ \vdots \ (1-\alpha_{i,1})\eth_i \ \cdots \ (1-\alpha_{i,c-1})\eth_i]\\ &= [\eth_i \ \vdots \ \boldsymbol{\alpha}_i \otimes \boldsymbol{\theta}_{i-1}+(\mathbf{1}_{c-1}^{\mathrm{T}}-\boldsymbol{\alpha}_i)\otimes \eth_i]. \qquad \square\end{aligned}$$

Proof of Proposition 5.1.

1. We just deal with the case where the essential supremum is finite, because the case where $\tau=\infty$ is similar. Given that the sequence $\{\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)\}_{i\in\mathbb{N}}$ is increasing, we have that for every positive integer κ
$$\mathcal{T}(\tilde{\boldsymbol{\theta}}_{i+\kappa}) \geq \mathcal{T}(\tilde{\boldsymbol{\theta}}_i) = \mathcal{T}(\mathrm{c}(\tilde{\boldsymbol{\theta}}_{i-1},\boldsymbol{\theta}_i)) \geq \mathcal{T}(\boldsymbol{\theta}_i). \tag{12}$$
Further, since by assumption $\boldsymbol{\theta}_i \in \mathcal{O}_{\varepsilon,M}$, it holds that
$$\mathcal{T}(\boldsymbol{\theta}_i) > \tau-\varepsilon. \tag{13}$$
The final result now follows by combining inequalities (12) and (13).
2. We only consider the case in which $\tau \in \mathbb{R}$, given that the case where $\tau=\infty$ is similar. Since by assumption we have that $\tilde{\boldsymbol{\theta}}_i \in \mathcal{O}_{\varepsilon,M}$, then it holds that
$$\mathcal{T}(\tilde{\boldsymbol{\theta}}_i) > \tau-\varepsilon.$$
The final result follows directly since $\{\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)\}_{i\in\mathbb{N}}$ is increasing.
3. By Claims 1 and 2, we have that for every positive integer κ
$$(\boldsymbol{\theta}_{\kappa-1} \in \mathcal{O}_{\varepsilon,M} \vee \tilde{\boldsymbol{\theta}}_{\kappa-1} \in \mathcal{O}_{\varepsilon,M}) \Rightarrow \tilde{\boldsymbol{\theta}}_{\kappa} \in \mathcal{O}_{\varepsilon,M}. \tag{14}$$
Applying the contrapositive law to (14), yields
$$\tilde{\boldsymbol{\theta}}_{\kappa} \in \mathcal{O}^c_{\varepsilon,M} \Rightarrow \begin{cases}\boldsymbol{\theta}_{\kappa-1} \in \mathcal{O}^c_{\varepsilon,M}\\ \tilde{\boldsymbol{\theta}}_{\kappa-1} \in \mathcal{O}^c_{\varepsilon,M}\end{cases} \Rightarrow \begin{cases}\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{\kappa-1} \in \mathcal{O}^c_{\varepsilon,M},\\ \tilde{\boldsymbol{\theta}}_1,\ldots,\tilde{\boldsymbol{\theta}}_{\kappa-1} \in \mathcal{O}^c_{\varepsilon,M},\end{cases}$$
where the last implication follows directly from Claims 1 and 2. $\square$

Proof of Theorem 5.1. The proof is as follows.

1. As a consequence of Proposition 5.1, it holds that
$$\mathbb{P}[\tilde{\boldsymbol{\theta}}_i \in \mathcal{O}^c_{\varepsilon,M}] \leq \mathbb{P}\left[\bigcap_{1\leq p\leq i-1}\{\tilde{\boldsymbol{\theta}}_p \in \mathcal{O}^c_{\varepsilon,M}\} \cap \{\boldsymbol{\theta}_p \in \mathcal{O}^c_{\varepsilon,M}\}\right] \leq \mathbb{P}\left[\bigcap_{1\leq p\leq i-1}\{\boldsymbol{\theta}_p \in \mathcal{O}^c_{\varepsilon,M}\}\right]. \tag{15}$$
Since by definition $\boldsymbol{\theta}_p = \arg\max_{q\in\{1,\ldots,c\}} \mathcal{T}(\mathbf{Z}_{p,q})$, then it holds that $\{\boldsymbol{\theta}_p \in \mathcal{O}^c_{\varepsilon,M}\} \subseteq \{\eth_p \in \mathcal{O}^c_{\varepsilon,M}\}$, for any positive integer p. This latter observation combined with (15) yields
$$\mathbb{P}[\tilde{\boldsymbol{\theta}}_i \in \mathcal{O}^c_{\varepsilon,M}] \leq \mathbb{P}\left[\bigcap_{1\leq p\leq i-1}\{\eth_p \in \mathcal{O}^c_{\varepsilon,M}\}\right] = \mathbb{P}[\eth_1 \in \mathcal{O}^c_{\varepsilon,M}]^{i-1}.$$
The final result now holds since by assumption $\mathbb{P}[\eth_1 \in \mathcal{O}^c_{\varepsilon,M}] < 1$.

2. Start by noting that $\{\mathcal{T}(\tilde{\boldsymbol{\theta}}_i),\mathbb{F}_i\}_{i\in\mathbb{N}}$ is a submartingale, where $\mathbb{F}_i=\sigma(\tilde{\boldsymbol{\theta}}_1,\ldots,\tilde{\boldsymbol{\theta}}_i)$ denotes the natural filtration, i.e.

$$\mathbb{E}[\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)|\mathbb{F}_i]=\mathbb{E}[(\mathcal{T}\circ c)(\tilde{\boldsymbol{\theta}}_{i-1},\boldsymbol{\theta}_i)|\mathbb{F}_i]\geq\mathbb{E}[\mathcal{T}(\tilde{\boldsymbol{\theta}}_{i-1})|\mathbb{F}_i]=\mathcal{T}(\tilde{\boldsymbol{\theta}}_{i-1}),\ \text{a.s.}$$

Since this submartingale is bounded from above, it is a.s. convergent to a random variable T.[2] Moreover, since ε is arbitrary, the preceding claim implies that

$$\mathbb{P}[\mathcal{T}(\boldsymbol{\theta}_i)<\tau]=o(1). \tag{16}$$

Fatou's lemma yields

$$\mathbb{P}[T<\tau]=\mathbb{P}[\liminf_{i\to\infty}\{\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)<\tau\}]\leq\limsup_{i\to\infty}\mathbb{P}[\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)<\tau]=0,$$

where the last equality follows by (16). Furthermore, recall that (7) holds, i.e.

$$\mathcal{T}\leq\tau,\ \text{a.e.}$$

In particular, this implies that for every positive integer i we have $\mathbb{P}[\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)>\tau]=0$. Consequently it holds that

$$\mathbb{P}[\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)>\tau]=o(1). \tag{17}$$

Therefore, again by Fatou's lemma

$$\mathbb{P}[T>\tau]=\mathbb{P}[\liminf_{i\to\infty}\{\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)>\tau\}]\leq\limsup_{i\to\infty}\mathbb{P}[\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)>\tau]=0,$$

where the last equality is a consequence of (17). □

Proof of Theorem 5.2. Our approach is similar to the one used in the previous proof. By a similar reasoning, it holds that

$$\mathbb{P}[\tilde{\boldsymbol{\theta}}_i\in\mathcal{O}^c_{\varepsilon,M}]\leq\mathbb{P}\left[\bigcap_{1\leq p\leq i-1}\{\vartheta_p\in\mathcal{O}^c_{\varepsilon,M}\}\right]\leq\inf_{1\leq p\leq i-1}\mathbb{P}[\vartheta_p\in\mathcal{O}^c_{\varepsilon,M}],$$

from where the final result follows. □

Proof of Theorem 5.3. The proof is split into two claims. The first claim establishes that $\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)-\mathcal{T}(\widehat{\boldsymbol{\theta}})=o(1)$, a.s. and the second claim shows that $\tilde{\boldsymbol{\theta}}_i-\widehat{\boldsymbol{\theta}}=o(1)$, a.s.

1. We first show that the sequence $\{\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)\}_{i\in\mathbb{N}}$ converges in probability to $\mathcal{T}(\tilde{\boldsymbol{\theta}}_0)$. Consider an arbitrary $\varepsilon>0$, and start by noting that

$$\mathbb{P}[|\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)-\mathcal{T}(\widehat{\boldsymbol{\theta}})|\geq\varepsilon]=\mathbb{P}[\{\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)\leq\mathcal{T}(\widehat{\boldsymbol{\theta}})-\varepsilon\}\cup\{\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)\geq\mathcal{T}(\widehat{\boldsymbol{\theta}})+\varepsilon\}]. \tag{18}$$

By (6) it holds that the essential supremum and the maximum coincide, and so by definition of essential supremum it holds that $\mathbb{P}[\{\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)\geq\mathcal{T}(\widehat{\boldsymbol{\theta}})+\varepsilon\}]=0$. This implies that (18) can be rewritten as

$$\mathbb{P}[|\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)-\mathcal{T}(\widehat{\boldsymbol{\theta}})|\geq\varepsilon]=\mathbb{P}[\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)\leq\mathcal{T}(\widehat{\boldsymbol{\theta}})-\varepsilon]=\mathbb{P}[\tilde{\boldsymbol{\theta}}_i\in\mathcal{O}^c_{\varepsilon,M}].$$

As a consequence of Proposition 5.1, it holds that

$$\mathbb{P}[\tilde{\boldsymbol{\theta}}_i\in\mathcal{O}^c_{\varepsilon,M}]\leq\mathbb{P}[\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{i-1}\in\mathcal{O}^c_{\varepsilon,M}]\leq\mathbb{P}[\vartheta_1,\ldots,\vartheta_{i-1}\in\mathcal{O}^c_{\varepsilon,M}]=(\mathbb{P}[\vartheta_1\in\mathcal{O}^c_{\varepsilon,M}])^{i-1}.$$

Since by assumption $\mathbb{P}[\vartheta_1\in\mathcal{O}^c_{\varepsilon,M}]<1$, the last inequality establishes that $\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)-\mathcal{T}(\widehat{\boldsymbol{\theta}})=o_p(1)$. The remaining part of the proof follows by a standard argument, given that the sequence $\{\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)\}_{i\in\mathbb{N}}$ is increasing, as this implies that the sequence of events $E_{i,\varepsilon}=\{|\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)-\mathcal{T}(\widehat{\boldsymbol{\theta}})|\leq\varepsilon\}$ is contractive, i.e. $E_{i+1,\varepsilon}\subseteq E_{i,\varepsilon}$, for every $i\in\mathbb{N}$ and $\varepsilon>0$. Thus by a standard argument,[3] convergence in probability implies that for every $\varepsilon>0$

$$\mathbb{P}[\lim_{i\to\infty}|\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)-\mathcal{T}(\boldsymbol{\theta}_o)|\leq\varepsilon]=1.$$

Given that ε is arbitrary, we get

$$\mathbb{P}[\lim_{i\to\infty}|\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)-\mathcal{T}(\widehat{\boldsymbol{\theta}})|=0]=1,$$

from where the final result follows.

2. We now assume that $\boldsymbol{\Theta}$ is compact, and suppose towards a contradiction that (Theorem 5.3) does not hold. Then for every ω on a set of positive probability $\Omega\subset\mathbb{R}^k$

$$\exists\varepsilon>0,\quad\forall p\in\mathbb{N}\quad\exists N_i>p\quad|\tilde{\boldsymbol{\theta}}_i(\omega)-\widehat{\boldsymbol{\theta}}|>\varepsilon. \tag{19}$$

[2] Since by assumption $\mathcal{T}$ is bounded from above, it holds that $\sup_i\mathbb{E}[\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)]<\infty$. Consequently, Doob's martingale convergence theorem can be applied, hence establishing the a.s. convergence to T.

[3] Recall that when a sequence of events E_i is either expansive or contracting it holds that $\lim_{i\to\infty}\mathbb{P}[E_i]=\mathbb{P}[\lim_{i\to\infty}E_i]$; see Ross (1996, p. 2).

Now for all $\omega \in \Omega$ the sequence $\{\tilde{\boldsymbol{\theta}}_i(\omega)\}_{i\in\mathbb{N}}$ is a sequence of points in a compact set $\boldsymbol{\Theta}$ and by Bolzano–Weierstrass theorem there is a convergent subsequence $\{\tilde{\boldsymbol{\theta}}_{i_\kappa}(\omega)\}_{\kappa\in\mathbb{N}}$ of $\{\tilde{\boldsymbol{\theta}}_i(\omega)\}_{i\in\mathbb{N}}$. This subsequence must converge to $\widehat{\boldsymbol{\theta}}$ because if the limit were $\boldsymbol{\theta}_a$ then, by the continuity of $\mathcal{T}$ we would have the sequence $\{\mathcal{T}(\tilde{\boldsymbol{\theta}}_{i_\kappa})(\omega)\}_{\kappa\in\mathbb{N}}$ converging to $\mathcal{T}(\boldsymbol{\theta}_a)=\mathcal{T}(\widehat{\boldsymbol{\theta}})$. Since $\widehat{\boldsymbol{\theta}}$ is the unique maximizer of $\mathcal{T}$ in $\boldsymbol{\Theta}$ we have $\boldsymbol{\theta}_a=\widehat{\boldsymbol{\theta}}$. Finally, observe that the subsequence $\{\tilde{\boldsymbol{\theta}}_{i_\kappa}(\omega)\}_{\kappa\in\mathbb{N}}$ also verifies the condition expressed in (19) for κ large enough, which yields the desired contradiction. □

Proof of Theorem 5.4. Using an argument similar to the proof of Theorem 5.1 we get that

$$\mathbb{P}[\tilde{\boldsymbol{\theta}}_i \in \mathcal{O}^c_{\varepsilon,M}] \le \mathbb{P}[\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{i-1} \in \mathcal{O}^c_{\varepsilon,M}] \le \mathbb{P}[\mathfrak{z}_1,\ldots,\mathfrak{z}_{i-1} \in \mathcal{O}^c_{\varepsilon,M}] \le \inf_{1\le p\le i-1} \mathbb{P}[\mathfrak{z}_p \in \mathcal{O}^c_{\varepsilon,M}].$$

This establishes that $\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)-\mathcal{T}(\widehat{\boldsymbol{\theta}})=\mathrm{o_p}(1)$, and the a.s. convergence can be achieved by the same argument used in the proof of Theorem 5.3. The remaining part of the proof is the same as above. □

References

Amemiya, T., 1985. Advanced Econometrics. Harvard University Press, Cambridge.

Andersen, H.C., Diaconis, P., 2007. Hit and run as a unifying device. Journal de la Société Française de Statistique 148, 5–28.

Andrews, D., 1999. Estimation when a parameter is on a boundary. Econometrica 67, 1341–1383.

Bohachevsky, I.O., Johnson, M.E., Stein, M.L., 1986. Generalized simulated annealing for a function optimization. Technometrics 28, 209–217.

Booth, J., Casella, G., Hobert, J., 2008. Clustering using objective functions and stochastic search. Journal of the Royal Statistical Society B 70, 119–139.

Capiński, M., Kopp, E., 1998. Measure, Integral and Probability. Springer, London.

Christensen, R., 1990. Log-Linear Models. Springer, New York.

Dalal, S.R., Fowlkes, E.B., Hoadley, B., 1989. Risk analysis of the space shuttle: pre-Challenger prediction of failure. Journal of the American Statistical Association 84, 945–957.

de Carvalho, M., 2011. Confidence intervals for the minimum of a function using extreme value statistics. International Journal of Mathematical Modelling & Numerical Optimisation 2, 288–296.

de Haan, L., 1981. Estimation of the minimum of a function using order statistics. Journal of the American Statistical Association 76, 467–469.

Duflo, M., 1996. Algorithmes Stochastiques. Springer, Berlin.

Esquível, M.L., 2006. A conditional Gaussian martingale algorithm for global optimization. In: Gavrilova, M., et al. (Eds.), Proceedings of International Conference on Computational Science and its Applications, CSA 2006, Glasgow, 2006, Lecture Notes in Computer Science, vol. 3982. , Springer, Berlin, pp. 813–823.

Gan, L., Jiang, J., 1999. A test for a global optimum. Journal of the American Statistical Association 94, 847–854.

Maranzano, C.J., Krzysztofowicz, R., 2008. Bayesian reanalysis of the Challenger O-ring data. Risk Analysis 28, 1053–1067.

Mexia, J.T., Pereira, D.G., Baeta, J., 1999. L2 environmental indexes. Biometrical Letters 36, 137–143.

Newey, W., McFadden, D., 1994. Large sample estimation and hypothesis testing. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. 4. , Elsevier Science, Amsterdam, pp. 2112–2245.

Nocedal, J., Wright, S., 1999. Numerical Optimization. Springer, New York.

Pakes, A., McGuire, P., 2001. Stochastic algorithms, symmetric Markov perfect equilibrium and the curse of dimensionality. Econometrica 69, 1261–1282.

Pereira, D.G., Mexia, J.T., 2010. Comparing double minimization and zigzag algorithms in joint regression analysis: the complete case. The Journal of Statistical Computation and Simulation 80, 133–141.

Romano, J.P., Shaikh, A.M., 2010. Inference for the identified set in partially identified econometric models. Econometrica 78, 169–211.

Ross, S., 1996. Stochastic Processes. Wiley, New York.

Solis, F.J., Wets, R.J.-B., 1981. Minimization by random search techniques. Mathematics of Operations Research 6, 19–30.

Spall, J.C., 2003. Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control. Wiley, Hoboken.

Veall, M.R., 1990. Testing for a global minimum in an econometric context. Econometrica 58, 1459–1465.

Williams, D., 1991. Probability with Martingales. Cambridge University Press, Cambridge.

Wolpert, D.H., Macready, W.G., 1997. No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation 1, 67–82.

Zabinsky, Z.B., 2003. Stochastic Adaptive Search in Global Optimization. Springer, New York.

Zabinsky, Z.B., Smith, R.L., McDonald, J.F., Romeijn, H.E., Kaufman, D.E., 1993. Improving hit-and-run for global optimization. Journal of Global Optimization 3, 171–192.