

# A generalization of the Solis–Wets method

Miguel de Carvalho<sup>\*†</sup>  
Institute of Mathematics  
Ecole Polytechnique Fédérale de Lausanne

## Abstract

This paper is concerned with the application of global stochastic optimization methods to extremum estimators. We propose a general stochastic method — the master method —, which includes several other 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.

**Key words:** Extremum estimators; Improving hit-and-run algorithm; Solis–Wets method; Stochastic optimization; Zigzag algorithm.

**MSC[2010]:** 62F10; 68W01.

## 1 Introduction

Extremum estimators are one of the most extensive classes of methods for obtaining estimates 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. Stated differently, in several key cases we 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. The questions then arise. First, is there a method which outperforms all others? Second, what type of algorithm should one 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

---

<sup>\*</sup>I am grateful to J. Tiago Mexia, Manuel L. Esquivel, Viatcheslav Melas and Vanda Inácio for helpful suggestions and recommendations. Financial support from *Centro de Matemática e Aplicações, Universidade Nova de Lisboa* and *Fundação para a Ciência e Tecnologia* is greatly acknowledged.

<sup>†</sup>*Address for correspondence:* M. de Carvalho, Institute of Mathematics, Ecole Polytechnique Fédérale de Lausanne, Station 8, 1015 Lausanne, Switzerland. E-mail: [miguel.carvalho@epfl.ch](mailto:miguel.carvalho@epfl.ch).

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 the focus will, however, be on stochastic optimization algorithms. 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 (Esquivel, 2006), *etc.* 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. In addition, we establish the convergence of the master method under a set of fairly mild assumptions. Just as a master key can be used to open several doors, the establishment of convergence of the master method, allows us to reach to the convergence of all the algorithms that it includes. An important instance of this method, to which we also devote some time, is the stochastic zigzag method — an algorithm largely inspired in the works of Mexia et al. (1999) and Pereira and Mexia (2010).

The remainder of this paper is structured as follows. In the next section some preliminary concepts are introduced. Section 3 revisits the meta-approach of Solis and Wets (1981). Section 4 is devoted to the introduction of the master method. The convergence of the proposed method is established in Section 5. Concluding remarks are provided in Section 6.

## 2 Preliminaries

### 2.1 Problem formulation

The exposition starts with the definition of extremum estimator.

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

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

Some remarks regarding notation:  $\Theta \subseteq \mathbb{R}^k$  denotes a parameter space;  $n$  is the sample size; the true parameter will be denoted by  $\boldsymbol{\theta}_o$  (the “o” stands for optimal).

As a consequence of (1), the optimization problem at the center of our attention is

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

For the sake of concreteness observe that the ordinary least squares estimator is obtained by setting  $\mathcal{T}_n(\boldsymbol{\beta}) = -(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ , 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 relies 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 means of the following condition

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

Here and below, we say that a random variable  $X_n$  is  $o_p(1)$ , if for any  $\varepsilon > 0$  the following condition is verified

$$\mathbb{P}[|X_n| > \varepsilon] \xrightarrow[n \rightarrow \infty]{} 0.$$

From the conceptual standpoint, the definition provided in (3) is preferable since it only requires that  $\mathcal{T}_n(\hat{\boldsymbol{\theta}}_n)$  is within  $o_p(1)$  from the global maximum of  $\mathcal{T}_n(\boldsymbol{\theta}_n)$ . This overcomes the question of existence and it is also more suitable for computational purposes (Andrews, 1999).

In order to reduce the burden of notation, hereafter we omit the subscript  $n$  in the extremum estimator and in the parameter objective function.

## 2.2 An overview of random search techniques

Suppose one has available a random sample of size  $n$ , from a population of interest. With such a sample at hand, we aim to obtain estimates of  $\boldsymbol{\theta}_o$  by solving the optimization problem (2). From the conceptual standpoint, for a fixed  $n$ , we can also think of the graph of  $\mathcal{T}$  as another population of interest from which we intend to consistently estimate the parameters<sup>1</sup>

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

In order 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

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

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).

### Classical random search algorithm

1. Choose an initial value of  $\boldsymbol{\theta}$ , say  $\boldsymbol{\theta}_0 \in \Theta$ , 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  $\Theta$ . 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 relies in the introduction of a conceptual algorithm which includes among others the algorithm stated above. To shed some light on some standard variants of the classical random search algorithm, observe that

<sup>1</sup>Recall that the graph of  $\mathcal{T}$  is defined  $Gr(\mathcal{T}) = \{(\boldsymbol{\theta}, \mathcal{T}(\boldsymbol{\theta})) : \boldsymbol{\theta} \in \Theta\}$ .

- other types of processes can be used in lieu of (4);
- independence in the choice of the values of  $\theta_i$  can be dropped;
- the probability distribution  $f$  can be allowed to have a support defined over  $\mathbb{R}^k \supseteq \Theta$ .<sup>2</sup>

### 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 deeply 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

$$\begin{aligned} \nabla_t &= \{\theta \in \Theta : \mathcal{T}(\theta) < t\}, & \bar{\nabla}_t &= \{\theta \in \Theta : \mathcal{T}(\theta) \leq t\}, \\ \Delta_t &= \{\theta \in \Theta : \mathcal{T}(\theta) > t\}, & \underline{\Delta}_t &= \{\theta \in \Theta : \mathcal{T}(\theta) \geq t\}. \end{aligned} \quad (5)$$

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

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

**Remark 3.1.** Observe that the essential supremum is tantamount to

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

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

In order to ease notation, below we use  $\tau$  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  $\tau$  coincides with the maximum, i.e.

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \mathcal{T}(\theta) \Rightarrow \tau = \mathcal{T}(\hat{\theta}). \quad (6)$$

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

$$\mathcal{T} \leq \tau, \text{ a.e.} \quad (7)$$

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  $\tau$  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  $\mathcal{O} : \mathbb{R}_+^2 \rightrightarrows \Theta$ , defined as

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

---

<sup>2</sup>Obviously, due adaptations are necessary; otherwise some problems can arise in Step 2. In effect, the lapse of such modifications may preclude the computation of the image for certain values of  $\theta$  which are not included in the domain of  $\mathcal{T}$ .

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

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

The next definition closes our conceptual framework.

**Definition 3.3.** A function  $C : \Theta \times \mathbb{R}^k \rightarrow \Theta$  is a compass function if

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

**Example 3.1.** A simple example of compass function is given by the mapping

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

If we define

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

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.

It is worth mentioning that we refer to this function as a ‘compass’, because this mapping guides the process of selection of the maxima. In the next section, we introduce the master method — a broad algorithm which includes several other optimization algorithms.

## 4 The master method

### 4.1 Introducing the master method

We open this subsection with the introduction of a general method from which several other algorithms are particular cases. Before we present the *modus operandi* of this master method some comments on notation are 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. Further discussion on notation is also given below.

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

0. Set  $i, j = 1$ . Find  $\mathbf{a}, \mathbf{b} \in \Theta$ , 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

$$\mathbf{Z}_{1,1} \text{ equal to } \arg \min_{\mathbf{x} \in \{\mathbf{a}, \mathbf{b}\}} \mathcal{T}(\mathbf{x}).$$

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  $\tilde{\boldsymbol{\theta}}_i = C(\tilde{\boldsymbol{\theta}}_{i-1}, \boldsymbol{\theta}_i)$ , where  $\boldsymbol{\theta}_i = \arg \max_{q \in \{1, \dots, c\}} \mathcal{T}(\mathbf{Z}_{i,q})$ , 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  $U\{1, \dots, k\}$ ).
- Observe that Step 0 simply initiates the algorithm. If we repeat Step 1 for a fixed  $i$ , we construct the iterates  $\mathfrak{z}_{i,1}, \mathfrak{z}_{i,2}, \dots, \mathfrak{z}_{i,c-1}$ . 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  $\mathfrak{z}_2, \mathfrak{z}_3, \text{etc.}$
- 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. For each  $i$  we will refer to sequence  $\mathbf{Z}_{i,1}, \dots, \mathbf{Z}_{i,c}$  as a course. At the light of 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, which can be written as

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

In order 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 such 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  $\boldsymbol{\theta}_i = \mathfrak{z}_i = \mathbf{Z}_{i,1}$ . Additionally,  $j$  becomes inactive in the the algorithm, since under these circumstances, Step 1 is never activated. 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  $\boldsymbol{\theta}_1 \in \Theta$ , and set  $\tilde{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_1$ .
1. Set  $\tilde{\boldsymbol{\theta}}_i = C(\tilde{\boldsymbol{\theta}}_{i-1}, \boldsymbol{\theta}_i)$ , 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–Wets method (cf. Solis and Wets, 1981, pg. 19). 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 which we present in the next subsection.

## 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  $\mathbf{Z}_{i,j+1} = \alpha_{i,j}\boldsymbol{\theta}_{i-1} + (1 - \alpha_{i,j})\mathfrak{z}_i$ . 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  $\mathfrak{z}_1$ . 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*).

For the sake of illustration in Figures 1 and 2, we exemplify the stochastic zigzag method to the Styblinski–Tang function (Spall, 2003).

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

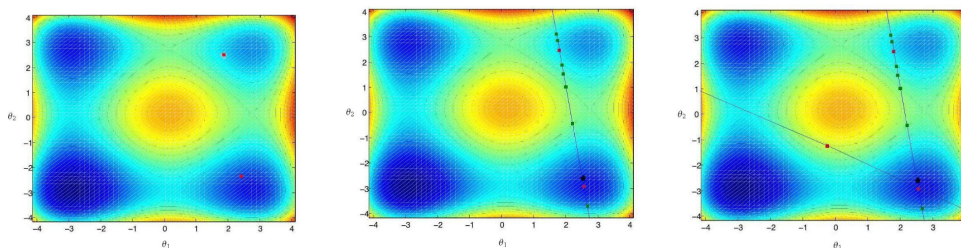


Figure 1: The figure represents the initialization of the stochastic zigzag method. In the picture in the left we start by finding points  $\mathbf{a}$  and  $\mathbf{b}$  which initialize the algorithm. The second picture illustrates that in Step 1 we collect a random sample ( $c = 10$ ) from line which passes through  $\mathbf{a}$  and  $\mathbf{b}$ . The remaining picture depicts steps 2 and 3 where 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 by such points.

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 could also have considered an alternative shape for the line. In fact, given the generality of the master method, we are even able to consider a different shape per each disparate course. The specification given above was chosen because of its simplicity and ease of implementation (see Theorem 4.1 of the next subsection).

## 4.3 Matrix formulation of the master method

In this subsection, we shed some light on the matrix representation of the master method. This conceptual framework will help us clarify some features of the method. Additionally, as we shall see latter, this representation reduces the burden of computational implementation. In order to

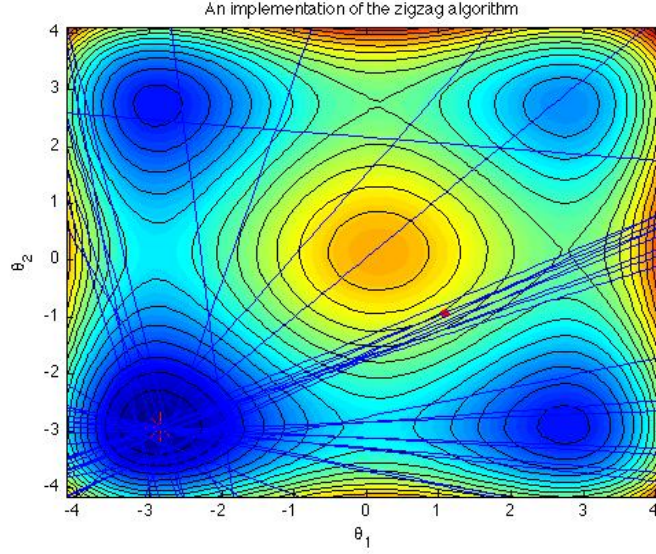


Figure 2: The application of the stochastic zigzag method to the Styblinski-Tang function ( $r = 30$ ;  $c = 10$ ). This function pertains to a class which is typically used to assess the performance of an optimization algorithm (see e.g. Spall, 2003). The functional form of this function is given in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

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, one can consider for instance the time of entry in the optimal zone. In fact, this can be defined for every  $\epsilon, M > 0$ , as

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

It can be easily shown that this is a stopping time, with respect to the natural filtration  $\mathbb{F}_i = \sigma(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_i)$ . In effect, analogous stopping times can be found even in introductory textbooks (e.g. Williams, 1991), so we skip the details. The crux of the proof is given by observing that for every  $i \in \bar{\mathbb{N}}$

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

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

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



Further, we will refer to the map-iterative matrix  $\mathbf{T}_{\mathbf{Z}}$  as the  $(r \times c)$ -matrix

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

For the sake of 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

$$\mathbf{Z} = \begin{bmatrix} \mathfrak{z}_1 \\ \mathfrak{z}_2 \\ \vdots \\ \mathfrak{z}_r \end{bmatrix}, \quad \mathbf{T}_{\mathbf{Z}} = \begin{bmatrix} \mathcal{T}(\mathfrak{z}_1) \\ \mathcal{T}(\mathfrak{z}_2) \\ \vdots \\ \mathcal{T}(\mathfrak{z}_r) \end{bmatrix}.$$

The affinity between the Solis–Wets conceptual algorithm and the master method introduced above, now becomes more clear. In fact, in the particular case wherein  $c = 1$ , 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 particular case of the stochastic zigzag method, the following matrix also finds application

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

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  $i$ -th zigzag course can be rewritten as*

$$\mathbf{z}_i = \begin{bmatrix} \mathfrak{z}_i \\ \vdots \\ \boldsymbol{\alpha}_i \otimes \boldsymbol{\theta}_{i-1} + (\mathbf{1}_{c-1}^T - \boldsymbol{\alpha}_i) \otimes \mathfrak{z}_i \end{bmatrix}, \quad i = 1, \dots, r$$

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

*Proof.* Just note that

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

□

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 simply replacing the scalar product with the Kronecker product; this can only be done

after performing the necessary scalar to vector adaptations. To put this differently, making use of 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

```

• rand:
  - seeds;
  - alpha.

for i=1 to r,
  - compute theta_i-1;
  - compute z_i;
  - increment i.

```

In the following we introduce an illustration of the Kronecker–zigzag decomposition and of the stochastic zigzag method.

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

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} 2/3 & 1/3 \\ -1 & -1/3 \\ -2 & -1 \end{bmatrix}, \quad \mathfrak{z} = \begin{bmatrix} -4 & 1 \\ 0 & 0 \\ 2 & 0 \end{bmatrix}, \quad \tilde{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0 = [ -1 \quad 4 ].$$

By Theorem 4.1, it follows that

$$\mathbf{z}_1 = \left[ \mathfrak{z}_1 \dot{:} \boldsymbol{\alpha}_1 \otimes \boldsymbol{\theta}_0 + (\mathbf{1}_2^T - \boldsymbol{\alpha}_1) \otimes \mathfrak{z}_1 \right] = [ -4 \quad 1 \quad -2 \quad 3 \quad -3 \quad 2 ].$$

Hence

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

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

$$\mathbf{z}_2 = \left[ \mathfrak{z}_2 \dot{:} \boldsymbol{\alpha}_2 \otimes \boldsymbol{\theta}_1 + (\mathbf{1}_2^T - \boldsymbol{\alpha}_2) \otimes \mathfrak{z}_2 \right] = [ 0 \quad 0 \quad 3 \quad -2 \quad 1 \quad -\frac{2}{3} ].$$

This yields

$$\mathbf{T}_{\mathbf{z}_2} = [ 0 \quad -53 \quad -10, 12 ],$$

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

$$\mathbf{z}_3 = \left[ \mathfrak{z}_3 \dot{:} \boldsymbol{\alpha}_3 \otimes \boldsymbol{\theta}_2 + (\mathbf{1}_2^T - \boldsymbol{\alpha}_3) \otimes \mathfrak{z}_3 \right] = [ 2 \quad 0 \quad 0 \quad 4 \quad 1 \quad 2 ].$$

Thus

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

and so  $\boldsymbol{\theta}_3 = [ 1 \quad 2 ]$ . 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, where the shuttle blew up during takeoff. On the morning of this catastrophic accident, the O-rings were 22°F below the temperature recorded in all prior launches. 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). The interest is thus to model 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

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

where  $\theta_\alpha$  and  $\theta_\beta$  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 loglikelihood can be written as

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

The estimation objects of interest are the intercept and slope parameters of model (10). Table (4.2) summarizes the estimates obtained by the application of the stochastic zigzag method to the loglikelihood (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 (4.2), the estimates are close to the ones presented by Christensen (1990), namely  $(\hat{\theta}_\alpha, \hat{\theta}_\beta) = (15.04, -0.2321)$ .

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)

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.

#### 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  $\Theta$ , then it is possible to take advantage of a result on extreme value theory due to de Haan (1981). In the sequel, let  $\mathcal{T}_{\mathfrak{z}(1)} \leq \mathcal{T}_{\mathfrak{z}(2)} \leq \dots \leq \mathcal{T}_{\mathfrak{z}(r)}$  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  $\mathfrak{z}_i$  with uniform distribution over  $\Theta$ . Further, consider the auxiliary correspondence  $\Xi : \mathbb{N} \times [0; 1] \rightrightarrows \mathbb{R}$ , defined as follows*

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

The following large sample result holds

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

*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, in the same spirit, if the method is pure and the seeds are uniformly distributed, Veall’s test can be implemented here by making use of the first column of the iterative matrix  $\mathbf{Z}$ .*

It is worth noting that the method is extremely easy to apply making use of the following inputs: two order statistics  $(\mathcal{T}_{\mathfrak{z}(r)}, \mathcal{T}_{\mathfrak{z}(r-1)})$ , 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 (2010).

## 5 Convergence of the master method

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

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

This reasoning can be extended by induction, being valid that for every positive integer  $\kappa$

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

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

**Proposition 5.1.** *For every positive integer  $\kappa$ , we have that*

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

*Proof.*

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, it holds that for every positive integer  $\kappa$

$$\mathcal{T}(\tilde{\boldsymbol{\theta}}_{i+\kappa}) \geq \mathcal{T}(\tilde{\boldsymbol{\theta}}_i) = \mathcal{T}(C(\tilde{\boldsymbol{\theta}}_{i-1}, \boldsymbol{\theta}_i)) \geq \mathcal{T}(\boldsymbol{\theta}_i). \quad (12)$$

Further, since by assumption  $\boldsymbol{\theta}_i \in \mathcal{O}_{\epsilon, M}$ , it holds that

$$\mathcal{T}(\boldsymbol{\theta}_i) > \tau - \epsilon. \quad (13)$$

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  $\tilde{\boldsymbol{\theta}}_i \in \mathcal{O}_{\epsilon, M}$ , then it holds that

$$\mathcal{T}(\tilde{\boldsymbol{\theta}}_i) > \tau - \epsilon.$$

The final result follows directly as a consequence of the sequence  $\{\mathcal{T}(\tilde{\boldsymbol{\theta}}_i)\}_{i \in \mathbb{N}}$  being increasing.

3. As a consequence of claims 1 and 2, we have that for every positive integer  $\kappa$

$$\left( \boldsymbol{\theta}_{\kappa-1} \in \mathcal{O}_{\epsilon, M} \vee \tilde{\boldsymbol{\theta}}_{\kappa-1} \in \mathcal{O}_{\epsilon, M} \right) \Rightarrow \tilde{\boldsymbol{\theta}}_{\kappa} \in \mathcal{O}_{\epsilon, M}. \quad (14)$$

A counter-reciprocal argument of (14), yields

$$\tilde{\boldsymbol{\theta}}_{\kappa} \in \mathcal{O}_{\epsilon, M}^c \Rightarrow \begin{cases} \boldsymbol{\theta}_{\kappa-1} \in \mathcal{O}_{\epsilon, M}^c \\ \tilde{\boldsymbol{\theta}}_{\kappa-1} \in \mathcal{O}_{\epsilon, M}^c \end{cases} \Rightarrow \begin{cases} \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{\kappa-1} \in \mathcal{O}_{\epsilon, M}^c \\ \tilde{\boldsymbol{\theta}}_1, \dots, \tilde{\boldsymbol{\theta}}_{\kappa-1} \in \mathcal{O}_{\epsilon, M}^c \end{cases},$$

where the last implication follows directly from claims 1 and 2.

□

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. The next theorem starts the analysis of convergence of the master method.

**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 the following condition holds

$$\forall B \in \mathbb{B}(\boldsymbol{\Theta}) \quad \lambda(B) > 0 \Rightarrow \mathbb{P}[\mathfrak{z}_1 \in B] > 0. \quad (15)$$

Then

$$\mathbb{P}[\tilde{\boldsymbol{\theta}}_i \in \mathcal{O}_{\epsilon, M}^c] = o(1).$$

2. Suppose that  $\mathcal{T}$  is bounded from above. Then

$$\mathcal{T}(\tilde{\boldsymbol{\theta}}_i) - T = o(1), \text{ a.s.},$$

where  $T$  is a random variable such that  $\mathbb{P}[T = \tau] = 1$ .

*Proof.* The proof is as follows.

1. As a consequence of Proposition 5.1, it holds that

$$\begin{aligned} \mathbb{P}[\tilde{\boldsymbol{\theta}}_i \in \mathcal{O}_{\epsilon, M}^c] &\leq \mathbb{P} \left[ \bigcap_{1 \leq p \leq i-1} \{ \tilde{\boldsymbol{\theta}}_p \in \mathcal{O}_{\epsilon, M}^c \} \cap \{ \boldsymbol{\theta}_p \in \mathcal{O}_{\epsilon, M}^c \} \right] \\ &\leq \mathbb{P} \left[ \bigcap_{1 \leq p \leq i-1} \{ \boldsymbol{\theta}_p \in \mathcal{O}_{\epsilon, M}^c \} \right]. \end{aligned} \quad (16)$$

Observe now that since by definition  $\boldsymbol{\theta}_p = \arg \max_{q \in \{1, \dots, c\}} \mathcal{T}(\mathbf{Z}_{p,q})$ , then it holds that  $\{ \boldsymbol{\theta}_p \in \mathcal{O}_{\epsilon, M}^c \} \subseteq \{ \mathfrak{z}_p \in \mathcal{O}_{\epsilon, M}^c \}$ , for any positive integer  $p$ . This latter observation combined with (16) yields

$$\mathbb{P}[\tilde{\boldsymbol{\theta}}_i \in \mathcal{O}_{\epsilon, M}^c] \leq \mathbb{P} \left[ \bigcap_{1 \leq p \leq i-1} \{ \mathfrak{z}_p \in \mathcal{O}_{\epsilon, M}^c \} \right] = \mathbb{P}[\mathfrak{z}_1 \in \mathcal{O}_{\epsilon, M}^c]^{i-1}.$$

The final result now holds since by assumption  $\mathbb{P}[\mathfrak{z}_1 \in \mathcal{O}_{\epsilon, M}^c] < 1$ .

2. Start by noting that  $\{\mathcal{T}(\tilde{\boldsymbol{\theta}}_i), \mathbb{F}_i\}_{i \in \mathbb{N}}$  is a submartingale, where  $\mathbb{F}_i = \sigma(\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \dots, \tilde{\boldsymbol{\theta}}_i)$ , denotes the natural filtration. In effect, just observe that

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

Since this submartingale is bounded from above, it is *a.s.* convergent to a random variable  $T$ .<sup>3</sup> Moreover, since  $\epsilon$  is arbitrary, the preceding claim implies that

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

Fatou's lemma yields

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

where the last equality follows by (17). 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

$$\mathbb{P}[\mathcal{T}(\tilde{\boldsymbol{\theta}}_i) > \tau] = o(1). \quad (18)$$

Therefore, again by Fatou's lemma, it holds that

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

where the last equality holds as a consequence of (18).

---

<sup>3</sup>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  $T$ .

□

The latter result warrants some general remarks. Roughly speaking, claim 1 states that the probability of the algorithm failing to sample the optimality region, approaches 0 as the number of iterates increases. Further, claim 2 ensures that the sequence  $\{\mathcal{T}(\tilde{\theta}_i)\}_{i \in \mathbb{N}}$  converges *a.s.* to a random variable  $T$  which is indistinguishable from the essential supremum. The proof of the second claim is entirely robust to both the pure stochastic method and the adaptive master method. Hence, the second result also holds in what concerns 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 the following condition holds*

$$\inf_{1 \leq p \leq i-1} \mathbb{P}[\mathfrak{z}_p \in \mathcal{O}_{\epsilon, M}^c] = o(1). \quad (19)$$

Then

$$\mathbb{P}[\tilde{\theta}_i \in \mathcal{O}_{\epsilon, M}^c] = o(1).$$

*Proof.* Our line of attack is similar to the previous proof. By a similar reasoning, it holds that

$$\mathbb{P}[\tilde{\theta}_i \in \mathcal{O}_{\epsilon, M}^c] \leq \mathbb{P} \left[ \bigcap_{1 \leq p \leq i-1} \{\mathfrak{z}_p \in \mathcal{O}_{\epsilon, M}^c\} \right] \leq \inf_{1 \leq p \leq i-1} \mathbb{P}[\mathfrak{z}_p \in \mathcal{O}_{\epsilon, M}^c],$$

from where the final result follows directly. □

It is important to underscore that the hypothesis considered here in order to establish the convergence of the adaptive stochastic method is known in the literature. In effect, condition (19) is tantamount to the one recommended by Esquivel (2006). Note however, that whereas Esquivel used this condition as a means 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 the following condition holds*

$$\forall B \in \mathbb{B}(\Theta) \quad \lambda(B) > 0 \Rightarrow \mathbb{P}[\mathfrak{z}_1 \in B] > 0.$$

Further, suppose that  $\mathcal{T}(\theta)$  is continuous and that  $\hat{\theta} = \arg \max_{\theta \in \Theta} \mathcal{T}(\theta)$ . Then, it holds that

$$\mathcal{T}(\tilde{\theta}_i) - \mathcal{T}(\hat{\theta}) = o(1) \text{ a.s., as } i \rightarrow \infty.$$

If furthermore  $\Theta \subset \mathbb{R}^k$  is compact, then

$$\tilde{\theta}_i - \hat{\theta} = o(1) \text{ a.s., as } i \rightarrow \infty. \quad (20)$$

*Proof.* The proof is split in two claims. The first claim establishes that  $\mathcal{T}(\tilde{\theta}_i) - \mathcal{T}(\hat{\theta}) = o(1)$ , *a.s.* The second claim shows that  $\tilde{\theta}_i - \hat{\theta} = o(1)$ , *a.s.*

1. Let us first show that the sequence  $\{\mathcal{T}(\tilde{\theta}_i)\}_{i \in \mathbb{N}}$  converges in probability to  $\mathcal{T}(\tilde{\theta}_0)$ . Consider  $\epsilon > 0$ . Start by noting that

$$\mathbb{P}[|\mathcal{T}(\tilde{\theta}_i) - \mathcal{T}(\hat{\theta})| \geq \epsilon] = \mathbb{P}[\{\mathcal{T}(\tilde{\theta}_i) \leq \mathcal{T}(\hat{\theta}) - \epsilon\} \cup \{\mathcal{T}(\tilde{\theta}_i) \geq \mathcal{T}(\hat{\theta}) + \epsilon\}]. \quad (21)$$

Observe now that by (6) it holds that the essential supremum and the maximum coincide. Hence, by definition of essential supremum it holds that  $\mathbb{P}[\{\mathcal{T}(\tilde{\boldsymbol{\theta}}_i) \geq \mathcal{T}(\hat{\boldsymbol{\theta}}) + \epsilon\}] = 0$ . This implies that (21) can be rewritten as

$$\mathbb{P}[|\mathcal{T}(\tilde{\boldsymbol{\theta}}_i) - \mathcal{T}(\hat{\boldsymbol{\theta}})| \geq \epsilon] = \mathbb{P}[\mathcal{T}(\tilde{\boldsymbol{\theta}}_i) \leq \mathcal{T}(\hat{\boldsymbol{\theta}}) - \epsilon] = \mathbb{P}[\tilde{\boldsymbol{\theta}}_i \in \mathcal{O}_{\epsilon, M}^c].$$

Now, observe that as a consequence of Proposition 5.1, it holds

$$\mathbb{P}[\tilde{\boldsymbol{\theta}}_i \in \mathcal{O}_{\epsilon, M}^c] \leq \mathbb{P}[\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{i-1} \in \mathcal{O}_{\epsilon, M}^c] \leq \mathbb{P}[\mathfrak{z}_1, \dots, \mathfrak{z}_{i-1} \in \mathcal{O}_{\epsilon, M}^c] = (\mathbb{P}[\mathfrak{z}_1 \in \mathcal{O}_{\epsilon, M}^c])^{i-1}.$$

Since by assumption  $\mathbb{P}[\mathfrak{z}_1 \in \mathcal{O}_{\epsilon, M}^c] < 1$ , the last inequality establishes that  $\mathcal{T}(\tilde{\boldsymbol{\theta}}_i) - \mathcal{T}(\hat{\boldsymbol{\theta}}) = o_p(1)$ . 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. In effect, this implies that the sequence of events  $E_{i, \epsilon} = \{|\mathcal{T}(\tilde{\boldsymbol{\theta}}_i) - \mathcal{T}(\hat{\boldsymbol{\theta}})| \leq \epsilon\}$  is contractive, i.e., it is such that  $E_{i+1, \epsilon} \subseteq E_{i, \epsilon}$ , for every  $i \in \mathbb{N}$  and  $\epsilon > 0$ . Consequently by a standard argument,<sup>4</sup> convergence in probability implies that for every  $\epsilon > 0$

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

Given that  $\epsilon$  is arbitrary, we get

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

from where the final result follows.

- Let us now suppose that  $\Theta$  is compact, and suppose by contradiction that (20) does not hold. Then for every  $\omega$  on a set of positive probability  $\Omega \subset \mathbb{R}^k$

$$\exists \epsilon > 0 \quad \forall p \in \mathbb{N} \quad \exists N_i > p \quad |\tilde{\boldsymbol{\theta}}_i(\omega) - \hat{\boldsymbol{\theta}}| > \epsilon. \quad (22)$$

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  $\Theta$  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  $\hat{\boldsymbol{\theta}}$  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  $\mathcal{T}(\boldsymbol{\theta}_a) = \mathcal{T}(\hat{\boldsymbol{\theta}})$ . Since  $\hat{\boldsymbol{\theta}}$  is the unique maximizer of  $\mathcal{T}$  in  $\Theta$  we have  $\boldsymbol{\theta}_a = \hat{\boldsymbol{\theta}}$ . Finally, observe that the subsequence  $\{\tilde{\boldsymbol{\theta}}_{i_\kappa}(\omega)\}_{\kappa \in \mathbb{N}}$  also verifies the condition expressed in (22) for  $\kappa$  large enough, which yields the desired contradiction. □

A similar result can be established if the master method is adaptive. Again, the framework has to be suitably accommodated making use Esquivel's condition (Esquivel, 2006).

**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 the following condition holds*

$$\inf_{1 \leq p \leq i-1} \mathbb{P}[\mathfrak{z}_p \in \mathcal{O}_{\epsilon, M}^c] = o(1). \quad (23)$$

---

<sup>4</sup>Recall that when a sequence of events  $E_i$  is either expansive or contracting it holds that  $\lim_{i \rightarrow \infty} \mathbb{P}[E_i] = \mathbb{P}\left[\lim_{i \rightarrow \infty} E_i\right]$ . Thus, under such circumstances one can interchange the limit with the measure. See e.g. Proposition 1.1.1 in Ross (1996).



Suppose in addition that  $\mathcal{T}(\boldsymbol{\theta})$  is continuous and that  $\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta} \in \Theta} \mathcal{T}(\boldsymbol{\theta})$ . Then, it holds that

$$\mathcal{T}(\tilde{\boldsymbol{\theta}}_i) - \mathcal{T}(\hat{\boldsymbol{\theta}}) = o(1) \text{ a.s., as } i \rightarrow \infty.$$

If furthermore  $\Theta \subset \mathbb{R}^k$  is compact, then

$$\tilde{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}} = o(1) \text{ a.s., as } i \rightarrow \infty.$$

*Proof.* By a similar reasoning to the proof of Theorem 5.4 we get that

$$\mathbb{P}[\tilde{\boldsymbol{\theta}}_i \in \mathcal{O}_{\epsilon, M}^c] \leq \mathbb{P}[\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{i-1} \in \mathcal{O}_{\epsilon, M}^c] \leq \mathbb{P}[\boldsymbol{z}_1, \dots, \boldsymbol{z}_{i-1} \in \mathcal{O}_{\epsilon, M}^c] \leq \inf_{1 \leq p \leq i-1} \mathbb{P}[\boldsymbol{z}_p \in \mathcal{O}_{\epsilon, M}^c].$$

This establishes that  $\mathcal{T}(\tilde{\boldsymbol{\theta}}_i) - \mathcal{T}(\hat{\boldsymbol{\theta}}) = o_p(1)$ . The a.s. convergence can be now achieved by the same argument used in the proof of Theorem 5.4. Hence, the remaining part of the proof is the same as above.  $\square$

## 6 Summary

This paper introduces the master method — a general algorithm which comprises several other 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.

## References

- AMEMIYA, T., 1985. *Advanced Econometrics*. Harvard University Press, Cambridge.
- ANDERSEN, H.C., DIACONIS, P., 2007. Hit and run as a unifying device. *J. Soc. Fr. Stat.* 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. *J. Roy. Statist. Soc. 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. *J. Amer. Statist. Assoc.* 84, 945–957.

- DE CARVALHO, M., 2009. Confidence intervals for the minimum of a function using extreme value statistics. submitted.
- DE HAAN, L., 1981. Estimation of the minimum of a function using order statistics. *J. Amer. Statist. Assoc.* 76, 467–469.
- DUFLO, M., 1996. *Algorithmes Stochastiques*. Springer, Berlin.
- ESQUÍVEL, M.L., 2006. A conditional Gaussian martingale algorithm for global optimization. In: M. Gavrilova et al. (Eds.), *Computational Science and its Applications, Proc. Int. Conf. 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. *J. Amer. Statist. Assoc.* 94, 847–854.
- GOLDBERG, D., 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading.
- MARANZANO, C.J., KRZYSZTOFOWICZ, R. 2008. Bayesian Reanalysis of the Challenger O-Ring Data. *Risk Anal.* 28, 1053–1067.
- MEXIA, J.T., CORTE REAL, P., 2002. Strong law of large numbers for additive extremum estimators. *Disc. Math. Prob. Statist.* 21, 81–88.
- MEXIA, J.T., CORTE REAL, P., 2003. Compact hypothesis and extremal set estimators. *Disc. Math. Prob. Statist.* 23, 103–121.
- MEXIA, J.T., PEREIRA, D.G., BAETA, J., 1999. L2 environmental indexes. *Biometrical Lett.* 36, 137–143.
- NEWBY, 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. *J. Stat. Comput. Sim.* 80, 133–141.
- PFLUG, G., 1996. *Optimization of Stochastic Models: The Interface Between Simulation and Optimization*. Kluwer Academic, Boston.
- 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. *Math. Oper. Res.* 6, 19–30.
- SPALL, J. C., 2003. *Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control*. Wiley, New Jersey.

- 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 Trans. Evol. Comput.* 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., ROMELJN, H.E., KAUFMAN, D.E., 1993. Improving hit-and-run for global optimization. *Journal of Global Optimization* 3, 171–192.