On the Information Content of Some Stochastic Algorithms

Manuel L. Esquível[1](✉), Nélio Machado[2], Nadezhda P. Krasii[3], and Pedro P. Mota[1]

[1] FCT Nova and CMA, UNL, 2829-516 Caparica, Portugal
{mle,pjpm}@fct.unl.pt
[2] FCT Nova, UNL, 2829-516 Caparica, Portugal
[3] Don State Technical University,
Gagarin Square 1, Rostov-on-Don 344000, Russian Federation

Abstract. We formulate an optimization stochastic algorithm convergence theorem, of Solis and Wets type, and we show several instances of its application to concrete algorithms. In this convergence theorem the algorithm is a sequence of random variables and, in order to describe the increasing flow of information associated to this sequence we define a filtration – or flow of σ-algebras – on the probability space, depending on the sequence of random variables and on the function being optimized. We compare the flow of information of two convergent algorithms by comparing the associated filtrations by means of the Cotter distance of σ-algebras. The main result is that two convergent optimization algorithms have the same information content if both their limit minimization functions generate the full σ-algebra of the probability space.

Keywords: Stochastic algorithms · Global optimization · Convergence of information σ-fields

1 Introduction

There are three main roots we can consider to the present work. The first is a quite general formulation of a stochastic optimization algorithm given in [SW81], studied under a different perspective in [SP99] and [PS00], and then corrected and slightly generalized in [Esq06] and having further developments and extensions in [dC12]. The subject of stochastic optimization – in the perspective adopted in this work – become stabilized with the books [Spa03, Zab03] and the work [Spa04]. The interest on the development of stochastic optimization methods continued, for instance in the work [RS03]. We refer a very effective and general approach to a substantial variety of stochastic optimization problems that takes the denomination of *the cross entropy method* proposed in a unified way in [RK04], further explained in [dBKMR05], with the convergence proved in [CJK07] and further extended in [RK08].

A. N. Shiryaev et al. (Eds.): ICSM-5 2020, PROMS 371, pp. 57–75, 2021.
https://doi.org/10.1007/978-3-030-83266-7_5

The second root originated in [SB98], is detailed in Sect. 4 for the reader's convenience, and may be broadly described as a form of conditioning of the results of any algorithm for global optimization; conditioning in the sense that the algorithm must gather enough information in order to get significant results.

This leads to the third root, namely the formalization of the concept *information*, conveyed by a random variable, as the σ-algebra generated by this random variable. This formalization encompasses many extensions and uses (see [Vid18]) for a recent and thorough account). We may initially refer with introduction of a convergence definition for σ-algebras – the so called *strong convergence* – related to the conditional expectation by Neveu in [Nev65] (or the French version in [Nev64]), with in [Kud74] a very deep study and, with developments, in [Pic98] and [Art01]. Then [Boy71] introducing a different convergence – the *Hausdorff convergence* – with an important observation in [Nev72] and further analysis in [Rog74] and [VZ93]. In the study of convergence of σ-algebras (or fields) there were many noticeable advances – and useful in our perspective – with Cotter in [Cot86] and [Cot87] extended in [ALR03] and detailed in [Bar04] and further extensions in [Kom08].

In the perspective of further developments, we mention [Wan01] and [Yin99], two works that concern the determination of the rates of convergence of stochastic algorithms allowing for the determination of adequate and most effective stopping rules for the algorithm and also [dC11] – and references therein – for a method to obtain confidence intervals for stochastic optimums.

2 Some Random Search Algorithms

We will now develop the following general idea: a convergent stochastic search algorithm for global optimization of a real valued function f defined on a domain $\mathcal{D}$ may be seen simply as a sequence of random variables $\mathbb{Y} = (Y_n)_{n\geq 1}$ such that the sequence $(f(Y_n))_{n\geq 1}$ converges (almost surely or in probability) to a random variable which gives a good estimate of $\min_{x\in\mathcal{D}} f(x)$. This sequence of random variables gives information about f on $\mathcal{D}$. A natural question is how to compare quantitatively the information brought by two different algorithms.

We now describe three algorithms which we will discuss in the following. Important issues for discussion are the convergence of the algorithm and, in case of convergence, the rate of convergence of the algorithm. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space.

2.1 The Pure Random Search Algorithm

For the general problem of minimizing $f : \mathcal{D} \subseteq \mathbb{R}^n \mapsto \mathbb{R}$ over $\mathcal{D}$, a bounded Borel set of $\mathbb{R}^n$, we consider the following natural algorithm.

S.1 Select a point x_1 at random in $\mathcal{D}$. Do $y_1 := x_1$.
S.2 Choose a point x_2 at random in $\mathcal{D}$. Do:

$$y_2 := y_1 \mathbb{I}\{f(y_1) < f(x_2)\} + x_2 \mathbb{I}\{f(y_1) \geq f(x_2)\}.$$

S.3 Repeat S.2.

To this algorithm it corresponds a probabilistic translation given in the following.

Sp.1 Let $X_1, X_2, \dots, X_n, \dots$ be independent random variables with common distribution over $\mathcal{D}$ verifying furthermore, with $\mathcal{B}(\mathcal{D})$ the Borel σ-algebra of $\mathcal{D}$:

$$\forall B \in \mathcal{B}(\mathcal{D}) \quad \lambda(B) > 0 \Rightarrow \mathbb{P}[X_1 \in B] > 0. \tag{1}$$

Sp.2 $Y_1 := X_1$
Sp.3 $Y_{n+1} = Y_n \mathbb{1}\{f(Y_n) < f(X_{n+1})\} + X_{n+1} \mathbb{1}\{f(Y_n) \geq f(X_{n+1})\}$

Having no prior information on the minimum set location and for random variables having common distribution on a bounded Borel set, a natural choice for the distribution of the random variables X_j is the uniform distribution. A non uniform distribution will distribute more mass on some particular sub domain. This may entail a loss of efficiency if the minimizer set is not contained in the more charged domain.

Remark 1 (The laws of the random variables of pure random search algorithm). We observe that for $n > 1$ we have:

$$Y_n = \sum_{k=1}^{n} X_k \mathbb{1}_{\bigcap_{j<k}\{f(X_k) \leq f(X_j)\} \cap \bigcap_{j>k}\{f(X_k) < f(X_j)\}},$$

an alternative expression that will allow us to describe the law of Y_n. Let D in the Borel σ-algebra of $\mathcal{D}$ and suppose that the random variables $(X_n)_{n\geq 1}$ are uniformly distributed in $\mathcal{D}$. We have the following disjoint union:

$$\{Y_n \in D\} \\ = \bigcup_{k=1}^{n} \left(\{X_k \in D\} \cap \bigcap_{1\leq j<k} \{f(X_k) \leq f(X_j)\} \cap \bigcap_{k<j\leq n} \{f(X_k) < f(X_j)\} \right),$$

which entails, representing by λ the Lebesgue measure over $\mathcal{D}$, that (see the Appendix, page 17, for the complete deduction):

$$\mathbb{P}[Y_n \in D] \\ = \sum_{k=1}^{n} \left(\frac{1}{\lambda(\mathcal{D})^n} \int_D \lambda(f^{-1}([f(x_k), +\infty[))^{k-1} \lambda(f^{-1}(]f(x_k), +\infty[))^{n-k} d\lambda(x_k) \right), \tag{2}$$

by Fubini theorem and by the fact that $(X_n)_{n\geq 1}$ is a sequence of independent uniformly distributed random variables on $\mathcal{D}$. Suppose furthermore that for every $x \in \mathcal{D}$ we have $\lambda(f^{-1}(\{f(x)\})) = 0$, we then have:

$$\mathbb{P}[Y_n \in D] = \frac{n}{\lambda(\mathcal{D})^n} \int_D \lambda(f^{-1}([f(x), +\infty[))^{n-1} d\lambda(x),$$

which gives us the density of Y_n with respect to the Lebesgue measure.

2.2 The Random Search Algorithm on (Nearly) Unbounded Domains

In the context of simple random search we may ask what is the natural substitute of the uniform distribution on an unbounded domain? A variant of the algorithm we now present was introduced in [Esq06] having in mind performing global optimization in unbounded domains. For bounded but large domains one may consider an algorithm using, for instance, a Gaussian distribution.

S.1 Select a point x at random in $\mathcal{D} \subset \mathbb{R}$. Do $z := x$.
S.2 Choose a point x at random in $\mathcal{D}$. Choose a point y with distribution $\mathcal{N}(x, \sigma)$ where for instance $\sigma := \mathrm{diam}(\mathcal{D})/10$. Do:

$$z := z\mathbb{I}\{f(z) < f(y)\} + y\mathbb{I}\{f(z) \geq f(y)\}.$$

S.3 Repeat S.2.

The probabilistic recursive translation of this algorithm is the following.

pS.1 Let $X_1, X_2, \ldots, X_n, \ldots$ independent random variables with common uniform distribution over $\mathcal{D}$
pS.2 $Z_1 := X_1$
pS.3 Let $Y_1, Y_2, \ldots, Y_n, \ldots$ be a sequence of independent random variables such that $Y_n \frown \mathcal{N}(X_n, \sigma)$.
pS.4 $Z_{n+1} := Z_n\mathbb{I}\{f(Z_n) < f(Y_{n+1})\} + Y_{n+1}\mathbb{I}\{f(Z_n) \geq f(Y_{n+1})\}$.

2.3 The Zig-Zag Algorithm

The zig-zag algorithm was introduced in [MPB99] (see also for other references and a convergence proof [PM10]) in order to optimize a quadratic function in two sets of multidimensional variables. The main idea of this algorithm may be simply described. In the first step we optimize in one of the sets of variables leaving the variables of the other set unchanged. On the second step, the first set of variables remains unchanged in the optimum value determined in the first step and an optimization is performed in the second set of variables. On the third step, it is now the second set of variables that remains unchanged in the optimum determined in the second step while an optimization is executed in the first set of variables. For the general case, the convergence and – if applicable – the rate of convergence issues were open problems, as far as we know.

One of the possibilities opened by this algorithm is to perform the optimization in sets of strictly smaller linear dimension than the dimension of $\mathcal{D}$. Suppose that $\mathcal{D} \subseteq \mathbb{R}^2$ is bounded.

S.1 Select a point x at random in $\mathcal{D}$. Do $z := x$.
S.2 (Optimization along a lower dimensional subset of the domain)
  S.2.1 Choose a point y at random in $\mathcal{D}$.

S.2.2 Choose, at random, points $\lambda_1, \ldots, \lambda_N \in \mathbb{R}$ such that $\lambda_j z + (1-\lambda_j) y \in \mathcal{D}$ and define x to be such that $f(x) = \min_{1 \leq j \leq N} f(\lambda_j z + (1-\lambda_j)y)$. Do:

$$z := z\mathbb{1}\{f(z) < f(x)\} + x\mathbb{1}\{f(z) \geq f(x)\}.$$

S.3 Repeat S.2

For this algorithm, one probabilistic recursive translation may be the following.

pS.1 Let $Y_1, Y_2, \ldots, Y_n, \ldots$ be a sequence of independent random variables with common uniform distribution over $\mathcal{D}$.
pS.2 $Z_1 := Y_1$
pS.3 For each $n \geq 1$, let $\lambda_1^n, \ldots, \lambda_N^n$ be independent sequences of independent random variables with uniform distribution in $[a, b]$ an interval such that:

$$\forall \lambda \in [a,b] \ \ \forall x, y \in \mathcal{D} \quad \lambda x + (1-\lambda) y \in \mathcal{D} \ .$$

which is possible as $\mathcal{D}$ is bounded.
pS.4 Define the random variable $X_n^{j_0}$ such that:

$$f(X_n^{j_0}) = \min_{1 \leq j \leq N} f(\lambda_j^n Z_n + (1-\lambda_j^n) Y_{n+1})$$

pS.5 $Z_{n+1} := Z_n \mathbb{1}\{f(Z_n) < f(X_n^{j_0})\} + X_n^{j_0} \mathbb{1}\{f(Z_n) \geq f(X_n^{j_0})\}$.

The main idea of the zig-zag algorithm may, of course, be exploited in several other ways.

3 The Solis and Wets Approach of Random Algorithms

We present this approach following the presentation in [Esq06] – which follows the context formalism of [SW81] – and observe that this approach may be applied to the algorithms described above.

3.1 The Convergence Results

We introduce some definitions which are necessary for the presentation of the convergence results. Let $f : \mathcal{D} \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ be a measurable function defined on a domain $\mathcal{D}$ that can be unbounded. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space. In order to deal with discontinuous functions, such as $\mathbb{1}_{[0,1] \setminus \{1/2\}}$, or unbounded functions, such as $\ln(|x| \, \mathbb{1}_{\mathbb{R} \setminus \{0\}} + (\infty) \mathbb{1}_{\{0\}}$, we need the following notion.

Definition 1 (Essential infimum of f in $\mathcal{D}$).

$$\alpha := \inf\{t \in \mathbb{R} : \lambda(\{x \in \mathcal{D} : f(x) < t\}) > 0\} \tag{3}$$

with λ being the Lebesgue measure on $\mathbb{R}^n$.

The formulation of general hypothesis on the function f in order to obtain the algorithm convergence requires the definition of the sets for $\epsilon > 0$ and $M < 0$.

Definition 2 (Level set of f of height ϵ over α).

$$E_{\alpha+\epsilon,M} := \begin{cases} \{x \in \mathcal{D} : f(x) < \alpha + \epsilon\} & \textit{if } \alpha \in \mathbb{R} \\ \{x \in \mathcal{D} : f(x) < M\} & \textit{if } \alpha = -\infty \end{cases} \tag{4}$$

The general form of the algorithm may be decomposed into the nuclear part which is a function verifying some condition and the procedure.

Definition 3 (The algorithm).

- *A function $\psi : \mathcal{D} \times \mathbb{R}^n \mapsto \mathcal{D} \subset \mathbb{R}^n$ such that the following hypothesis $[H1]$ is verified.*

$$[H1] : \begin{cases} \forall t, x \;\; f(\psi(t,x)) \leq f(t) \\ \forall x \in \mathcal{D} \;\; f(\psi(t,x)) \leq f(x) \end{cases} \tag{5}$$

- *A sequence of random variables given by:*

$$\begin{cases} Y_1 = X_1 \\ Y_{n+1} = \psi(Y_n, X_n) & \textit{for } n \geq 1 \end{cases} \tag{6}$$

where $X_n \frown \mathbb{P}_n$ satisfying hypothesis in Formula (1), and $\mathbb{P}_n$ being a probability measure – the law of X_n – that may depend on $\mathbb{P}_1, \ldots, \mathbb{P}_{n-1}$ in case of adaptive random search.

Remark 2 (Examples of stochastic algorithms for global optimization). The pure random search algorithm in Sect. 2.1, the random search on nearly unbounded domains in Sect. 2.2 and the zig-zag algorithm in Sect. 2.3 may be considered as instances of Solis and Wets approach. As presented, the following function obviously describes the algorithms and verifies the hypothesis $H1$ in Formula (5).

$$\psi(t,x) = t\mathbb{1}_{\{f(t)<f(x)\}}(t,x) + x\mathbb{1}_{\{f(t)\geq f(x)\}}(t,x)$$

The following result ensures the convergence of the algorithm under very general hypothesis.

Theorem 1 (A Solis and Wets' type theorem for random search algorithm convergence). *Suppose that f is measurable and bounded from below. Let α be the essential infimum of f in $\mathcal{D}$.*

$H2(\epsilon)$ For pure random search this hypothesis is defined for every $\epsilon > 0$ as:

$$\lim_{k\to+\infty} \prod_{1\leq j\leq k} \mathbb{P}[X_j \in E^c_{\alpha+\epsilon,M}] = \lim_{k\to+\infty} \prod_{1\leq j\leq k} \mathbb{P}_j[E^c_{\alpha+\epsilon,M}] = 0 \,.$$

$H'2(\epsilon)$ *For adaptive search this hypothesis is defined for every $\epsilon > 0$ as:*

$$\lim_{k \to +\infty} \inf_{1 \leq j \leq k} \mathbb{P}[X_j \in E^c_{\alpha+\epsilon,M}] = \lim_{k \to +\infty} \inf_{1 \leq j \leq k} \mathbb{P}_j[E^c_{\alpha+\epsilon,M}] = 0 \ . \tag{7}$$

If for some $\epsilon > 0$ hypothesis $H2(\epsilon)$ (in case of pure random search) or $H'2(\epsilon)$ (in case of adaptive search) are verified, then:

$$\lim_{n \to +\infty} \mathbb{P}[Y_n \in E_{\alpha+\epsilon,M}] = 1 \ . \tag{8}$$

If for every $\epsilon > 0$ hypothesis $H2(\epsilon)$ (in case of pure random search) or $H'2(\epsilon)$ (in case of adaptive search) are verified, then the sequence $(f(Y_n))_{n \geq 1}$ converges almost surely to a random variable $Min_{f,\mathbb{Y}}$ such that $\mathbb{P}[Min_{f,\mathbb{Y}} \leq \alpha] = 1$.

Proof. A first fundamental observation is that if $Y_n \in E_{\alpha+\epsilon,M}$ or $X_n \in E_{\alpha+\epsilon,M}$, then by hypothesis $H1$ we have that $Y_{n+1} \in E_{\alpha+\epsilon,M}$ and so as $(f(Y_n))_{n \geq 1}$ is decreasing, $Y_{n+k} \in E_{\alpha+\epsilon,M}$ for every $k \geq 1$. As a consequence, for $k > 1$:

$$\{Y_k \in E^c_{\alpha+\epsilon,M}\} \subseteq \{Y_1, \ldots, Y_{k-1} \in E^c_{\alpha+\epsilon,M}\} \cap \{X_1, \ldots, X_{k-1} \in E^c_{\alpha+\epsilon,M}\}.$$

as if it was otherwise we would contradict our first observation. Now, it is clear that:

$$\begin{aligned} \mathbb{P}[Y_k \in E^c_{\alpha+\epsilon,M}] &\leq \mathbb{P}\left[\bigcap_{1 \leq j \leq k-1} \{Y_j \in E^c_{\alpha+\epsilon,M}\} \cap \{X_j \in E^c_{\alpha+\epsilon,M}\}\right] \\ &\leq \mathbb{P}\left[\bigcap_{1 \leq j \leq k-1} \{X_j \in E^c_{\alpha+\epsilon,M}\}\right] . \end{aligned} \tag{9}$$

On the pure random search scenario we have that $(X_n)_{n \geq 1}$ is a sequence of iid random variables and so:

$$\mathbb{P}\left[\bigcap_{1 \leq j \leq k-1} \{X_j \in E^c_{\alpha+\epsilon,M}\}\right] = \prod_{1 \leq j \leq k-1} \mathbb{P}\left[X_j \in E^c_{\alpha+\epsilon,M}\right] = \mathbb{P}\left[X_1 \in E^c_{\alpha+\epsilon,M}\right]^{k-1} , \tag{10}$$

implying that

$$1 \geq \mathbb{P}[Y_k \in E_{\alpha+\epsilon,M}] = 1 - \mathbb{P}[Y_k \in E^c_{\alpha+\epsilon,M}] \geq 1 - \mathbb{P}\left[X_1 \in E^c_{\alpha+\epsilon,M}\right]^{k-1} .$$

Now by hypothesis in Formula (1) we have that $\mathbb{P}\left[X_1 \in E^c_{\alpha+\epsilon,M}\right] < 1$ and so conclusion in Formula (8) of the theorem now follows. On the alternative scenario of adaptive random search we still have the same conclusion but now based on the estimate:

$$\mathbb{P}\left[\bigcap_{1 \leq j \leq k-1} \{X_j \in E^c_{\alpha+\epsilon,M}\}\right] \leq \inf_{1 \leq j \leq k-1} \mathbb{P}\left[X_j \in E^c_{\alpha+\epsilon,M}\right] ,$$

instead of estimate in Formula (10) used in the pure random search case. For the second conclusion of the proof, observe that the sequence $(f(Y_n))_{n\geq 1}$ being almost surely non increasing, as a consequence of hypothesis $H1$, and bounded below is almost surely convergent to a random variable that we denote by $\mathrm{Min}_{\mathbb{Y}}$. Now, observing that for all $\epsilon > 0$:

$$\lim_{k\to+\infty} \mathbb{P}[Y_k \in E_{\alpha+\epsilon,M}] = \lim_{k\to+\infty} \mathbb{P}[f(Y_k) < \alpha + \epsilon] = 1,$$

in either pure random search or adaptive search, the conclusion follows by a standard argument (see Corollary 1. and Lemma 1. in [Esq06, p. 844]).

Remark 3. Having in mind a characterization of the speed of convergence of the algorithm it may be useful to observe that the following condition $H''2(\epsilon)$ also entails the conclusion of the theorem, although being more stringent than $H'2(\epsilon)$.

$$\lim_{k\to+\infty} \max_{1\leq j\leq k} \mathbb{P}[X_j \in E^c_{\alpha+\epsilon,M}] = \lim_{k\to+\infty} \max_{1\leq j\leq k} \mathbb{P}_j[E^c_{\alpha+\epsilon,M}] = 0. \qquad (11)$$

In order to improve Theorem 1 some additional hypothesis are needed. For instance, if the minimizer is not unique then the sequence $(Y_n)_{n\geq 1}$ may not converge. First, let us observe that if the minimizer of f is unique and f is continuous, then the essential minimum of f coincides with the minimum of f.

Proposition 1. *Let f be continuous and admiting an unique minimizer $z \in \mathcal{D}$ that is such that $f(z) = \min_{x\in\mathcal{D}} f(x)$. Then $\alpha = \min_{x\in\mathcal{D}} f(x) =: m$.*

Proof. Let $\epsilon > 0$ be given. There exists $x_\epsilon \in \mathcal{D}$ such that $m = f(z) < f(x_\epsilon) < m + \epsilon$. By the continuity we have an open neighborhood V of x_ϵ such that for all $x \in V$ we still have $m < f(x) < m + \epsilon$. As a consequence:

$$\lambda(\{x : f(x) < m + \epsilon\}) \geq \lambda(V) > 0,$$

and so $\alpha \leq m + \epsilon$ and, as ϵ is arbitrary, we have $\alpha \leq m$. Consider again a given $\epsilon > 0$. There exists $\alpha < t_\epsilon < \alpha + \epsilon$. As a consequence:

$$\lambda\left(\{x \in \mathcal{D} : f(x) < t_\epsilon\}\right) > 0,$$

and $m = \min_{x\in\mathcal{D}} f(x) < \alpha + \epsilon$. As ϵ is arbitrary we have $m \leq \alpha$ and finally the conclusion stated.

Theorem 2. *Suppose the same notations and the same set of hypothesis of Theorem 1, namely that for every $\epsilon > 0$ hypothesis $H2(\epsilon)$ (in case of pure random search) or $H'2(\epsilon)$ (in case of adaptive search) are verified. Suppose, furthermore, that f is continuous and that admits an unique minimizer $z \in \mathcal{D}$. Then we have almost surely that $\lim_{n\to+\infty} f(Y_n) = f(z)$. If, furthermore, $\mathcal{D}$ is compact then $\lim_{n\to+\infty} Y_n = z$.*

Proof. Let us first show that the sequence $(f(Y_n))_{n\geq 1}$ converges in probability to $f(z)$. Consider $\epsilon > 0$. As $f(z)$ is now the essential minimum of f in $\mathcal{D}$ we have that:

$$\mid f(Y_n) - f(z) \mid \geq \epsilon \Leftrightarrow \begin{cases} f(Y_n) \leq f(z) - \epsilon & \text{impossible} \\ f(Y_n) \geq f(z) + \epsilon, \end{cases}$$

the possible case meaning that $Y_n \in E^c_{f(z)+\epsilon}$. Now, by a similar argument as the one used in the proof of Theorem 1, we have that $X_1, \ldots, X_{n-1} \in E^c_{f(z)+\epsilon}$ and so, under each one of the alternative hypothesis, we have:

$$\mathbb{P}\left[\mid f(Y_n) - f(z) \mid \geq \epsilon\right] \leq \begin{cases} \mathbb{P}[\{X{\in}E^c_{f(z)+\epsilon}\}]^{n-1} & \text{under } H2(\epsilon) \\ \inf_{1\leq j\leq n-1} \mathbb{P}[X_j \in E^c_{f(z)+\epsilon}] & \text{under } H'2(\epsilon), \end{cases} \quad (12)$$

thus ensuring that $\lim_{n\to+\infty} \mathbb{P}[\mid f(Y_n) - f(z) \mid \geq \epsilon] = 0$. If $H2\epsilon$ (or $H'2\epsilon$) are verified for all $\epsilon > 0$ the convergence in probability follows immediately. Finally, by a standard argument, the convergence almost surely of the sequence $(f(Y_n))_{n\geq 1}$ follows because this sequence is non increasing and convergent in probability. Let us suppose now that $\mathcal{D}$ is compact and that the sequence $(Y_n)_{n\geq 1}$ does not converge to z almost surely. Then for every ω on a set of positive probability $\Omega' \subset \Omega$:

$$\exists \epsilon > 0 \;\; \forall n \in \mathbb{N} \;\; \exists N_n > n \quad \mid Y_{N_n}(\omega) - z \mid > \epsilon. \quad (13)$$

Now for all $\omega \in \Omega'$ the sequence $(Y_n)(\omega)_{n\geq 1}$ is a sequence of points in a compact set $\mathcal{D}$ and by Bolzano-Weierstrass theorem there is a convergent subsequence $(Y_{n_k})(\omega)_{k\geq 1}$ of $(Y_n)(\omega)_{n\geq 1}$. This subsequence must converge to z because if the limit were y then, by the continuity of f we would have the sequence $(f(Y_{n_k}))(\omega)_{k\geq 1}$ converging to $f(y) = f(z)$. Now as z is an unique minimizer of f in $\mathcal{D}$ we certainly have $y = z$. Finally observe that the subsequence $(Y_{n_k})(\omega)_{k\geq 1}$ also verifies the condition expressed in Formula (13) for k large enough, which yields the desired contradiction.

3.2 A Preliminary Observation on the Rate of Convergence

Results on the rate of convergence may be used to determine a stopping criterium for the algorithm. As a proxy for the speed of convergence of the algorithms in the context of the proof Theorems 1 and 2, namely for instance Formula (12), we may consider the quantity $\mathbb{P}[X_j \in E^c_{\alpha+\epsilon,M}]$ for various choices of distributions. In case of pure random search we have obviously:

$$\mathbb{P}\left[X_j \in E^c_{\alpha+\epsilon,M}\right] = \frac{\lambda(E^c_{\alpha+\epsilon,M})}{\lambda(\mathcal{D})}\ .$$

In case of random search on a nearly unbounded domain we have, (with the notations of Sect. 2.2), that:

$$\mathbb{P}[Y_j \in E^c_{\alpha+\epsilon,M}] = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{Y_j\in E^c_{\alpha+\epsilon,M}\}}\right] \mid X_j\right]\right]\,.$$

Now as we have that:

$$\mathbb{P}\left[Y_j \in E^c_{\alpha+\epsilon,M} \mid X_j = x\right] = \int_{E^c_{\alpha+\epsilon,M}} \frac{e^{-\frac{(x-u)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}} du,$$

it follows that,

$$\mathbb{E}\left[\mathbb{1}_{\{Y_j \in E^c_{\alpha+\epsilon,M}\}} \mid X_j\right] = \int_{E^c_{\alpha+\epsilon,M}} \frac{e^{-\frac{(X_j-u)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}} du,$$

which, in turn, implies that,

$$\mathbb{P}\left[Y_j \in E^c_{\alpha+\epsilon,M}\right] = \mathbb{E}\left[\int_{E^c_{\alpha+\epsilon,M}} \frac{e^{-\frac{(X_j-u)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}} du\right] = \int_{\mathcal{D}} \int_{E^c_{\alpha+\epsilon,M}} \frac{e^{-\frac{(x-u)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}} du \frac{dx}{\lambda(\mathcal{D})}$$

where the integral on the right doesn't seem easily estimable, in general. Suppose for simplification that $\mathcal{D} = [-A, +A]$ and that $E^c_{\alpha+\epsilon,M} \subseteq [-a, +a]$ where $0 < a \ll 1 \ll A$. Then, by Fubini theorem,

$$\begin{aligned}\int_{\mathcal{D}} \int_{E^c_{\alpha+\epsilon,M}} \frac{e^{-\frac{(x-u)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}} du \frac{dx}{\lambda(\mathcal{D})} &\approx \int_{-\infty}^{+\infty} \int_{E^c_{\alpha+\epsilon,M}} \frac{e^{-\frac{(x-u)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}} du \frac{dx}{\lambda(\mathcal{D})} \\ &= \frac{\lambda(E^c_{\alpha+\epsilon,M})}{\lambda(\mathcal{D})},\end{aligned}$$

allowing the conclusion that also $\mathbb{P}[Y_j \in E^c_{\alpha+\epsilon,M}] \approx \lambda(E^c_{\alpha+\epsilon,M})/\lambda(\mathcal{D})$ thus showing that the two algorithms, in the special situation assumed for simplification, are comparable in a first approximation.

4 On the Information Content of a Stochastic Algorithm

It is natural to conceive that in order for an algorithm to achieve global stochastic optimization of a function over a domain the algorithm has to collect complete – in some sense – information on the function over the domain. In [SB98] there are some very striking precise results on this idea. Let us detail Stephens and Baritompa's result. Consider a random algorithm described by a sequence of random variable $X^f_1, \ldots, X^f_n, \ldots$ for some function f on a domain $\mathcal{D}$. The closure $\overline{\mathbf{X}^f}$ of the set $\{X^f_1, \ldots, X^f_n, \ldots\}$ is a random set in $\mathcal{D}$.

Theorem 3 (*Global optimization requires global information*). *For any $r \in]0, 1[$, the following are equivalent:*

1. *The probability that the algorithm locate the global minimizers for f as points of $\overline{\mathbf{X}^f}$ is greater or equal than r, for any f in a sufficiently rich class of functions.*

2. *The probability that $x \in \overline{\mathbf{X}^f}$ is greater or equal than r, for any $x \in \mathcal{D}$ and f in a sufficiently rich class of functions.*

That is, roughly speaking, an algorithm works on any rich class of functions if and only if we have $\mathbb{P}[\overline{\mathbf{X}^f} = \mathcal{D}] = 1$. In the case of deterministic search the result is as expected, namely that the algorithm *sees* – in an intuitive yet precise sense – the global optimum for a class of functions in a domain if and only if the closure of the set of finite testing sequences, for any function, is dense in the domain. The extension of this result to the stochastic case gives the necessary and sufficient condition, in Theorem 3, that the lower bound of the probability of a stochastic algorithm *seing* the global optimum is the same as the lower bound of the probability of an arbitrary point of the domain belonging to the closure of the (random) set of finite testing sequences.

Having in mind the study of the limitations of an effective global optimization stochastic algorithm we address the problem of studying the information content of an algorithm. We recall that – as in Theorem 1 – a random algorithm may be identified with a sequence of random variables. The flow of information gained through a sequential observation of the sequence of random variables is usually described by the natural filtration associated with the sequence. In order to compare, in the information sense, two sequences of random variables we need to compare the associated natural filtrations.

In Theorem 5 below, by resorting to a natural defined notion of the information content of a stochastic algorithm, we obtain the result that two convergent algorithms have the same information content if the information generated by their respective minimizing functions is the whole available information in the probability space. So, the connection between the function and the stochastic set-up to generate stochastic algorithms for its global optimization - namely, probability space, probability laws of the algorithm – deserves to be further investigated.

In the following Sect. 4.1 we briefly recall results from [Cot86, Cot87, ALR03, Kud74, Bar04] on the set of complete sub σ-algebras of $\mathcal{F}$ as a topological metric space.

4.1 The Cotter Metric Space of the Complete σ-algebras

Recall that all random variables are defined on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We now consider $\mathfrak{F}^\star$, the set of all σ-algebras $\mathcal{G} \subseteq \mathcal{F}$ which are complete with respect to $\mathbb{P}$.

Remark 4. We may define an equivalence relation $\mathcal{R}$ on $\mathfrak{F}^\star$ by considering an equivalence relation $\sim$ for sets in $\mathcal{F}$ defined for all $G, H \in \mathcal{F}$ by:

$$G \sim H \Leftrightarrow \mathbb{P}[G \setminus H \cup H \setminus G] = 0. \tag{14}$$

As so, the quotient class $\mathfrak{F} := \mathfrak{F}^\star/\mathcal{R}$ is the class of all sub-σ-algebras of $\mathcal{F}$ with elements identified up to sets of probability zero.

Strong convergence in $L^1(\Omega, \mathcal{F}, \mathbb{P})$ – and also in $L^p(\Omega, \mathcal{F}, \mathbb{P})$ for $p \in [1, +\infty[$ – of a sequence $(\mathcal{G}_n)_{n\geq 1}$ of σ-algebras to $\mathcal{G}_\infty$ was introduced by Neveu in 1970 (see [Nev64, pp. 117–118]) with the condition that:

$$\forall X \in L^1(\Omega, \mathcal{F}, \mathbb{P}) \quad \lim_{n\to+\infty} \|\mathbb{E}[X \mid \mathcal{G}_n] - \mathbb{E}[X \mid \mathcal{G}_\infty]\|_{L^1(\Omega,\mathcal{F},\mathbb{P})} = 0, \tag{15}$$

noticing that for the sequence $(\mathcal{G}_n)_{n\geq 1}$ to converge it suffices that for all $F \in \mathcal{F}$ the sequence $(\mathbb{E}[\mathbb{1}_F \mid \mathcal{G}_n])_{n\geq 1}$ converges in probability. In 1985, Cotter showed that this notion of convergence defines a topology which is metrizable (see [Cot87]). The Cotter distance d_c is defined on $\mathfrak{F} \times \mathfrak{F}$ by:

$$\begin{aligned} d_c(\mathcal{H}, \mathcal{G}) &= \sum_{i=1}^{+\infty} \frac{1}{2^i} \min\left(\mathbb{E}\left[|\mathbb{E}[X_i \mid \mathcal{H}] - \mathbb{E}[X_i \mid \mathcal{G}]|\right], 1\right) \\ &= \sum_{i=1}^{+\infty} \frac{1}{2^i} \min\left(\|\mathbb{E}[X_i \mid \mathcal{H}] - \mathbb{E}[X_i \mid \mathcal{G}]\|_1, 1\right). \end{aligned} \tag{16}$$

with $\mathcal{H}, \mathcal{G} \in \mathfrak{F}$, $\|X\|_1$ the $L^1(\Omega, \mathcal{F}, \mathbb{P})$ norm of X, with $(X_i)_{i\in\mathbb{N}}$ a dense denumerable set in $L^1(\Omega, \mathcal{F}, \mathbb{P})$. We have that $(\mathfrak{F}, d_c)$ is a complete metric space.

We will need a consequence of the definition of the Cotter distance (see Corollary III.35, in [Bar04, p. 36]) that we quote for the reader's convenience.

Proposition 2. *Consider $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathcal{G}_3$ in $\mathfrak{F}$, Then we have that:*

$$d_c\left(\mathcal{G}_2, \mathcal{G}_3\right) \leq 2 d_c\left(\mathcal{G}_1, \mathcal{G}_3\right).$$

We will also need a remarkable result of Cotter (see Corollary 2.2 and Corollary 2.4 in [Cot87, p. 42]) that we formulate next.

Theorem 4. *Let $\mathcal{L}_\mathrm{P}$ be the metric space of the real valued random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with the metric of the convergence in probability. Let $\sigma : \mathcal{L}_\mathrm{P} \mapsto \mathfrak{F}$ that to each random variable X associates $\sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}$ the sigma-algebra generated by X. Then, considering the metric space $(\mathfrak{F}, d_c)$ with d_c the Cotter distance defined in Formulas (16), we have that σ is continuous at $X \in \mathcal{L}_\mathrm{P}$ if and only if $\sigma(X) = \mathcal{F}$.*

This result on the continuity of the map σ between metric spaces $\mathcal{L}_\mathrm{P}$ and $(\mathfrak{F}, d_c)$ will be applied to convergent sequences.

4.2 The Information Content of a Random Algorithm

Let $\mathbb{Y} = (Y_n)_{n\geq 1}$ be a stochastic algorithm for the minimization of f on a domain $\mathcal{D}$. According to Theorem 1 we may define a convergent algorithm for the minimization problem of f on the domain $\mathcal{D}$.

Definition 4. *Let α be the essential infimum of f on $\mathcal{D}$ defined in Formula (3). Following Theorem 1, the algorithm $\mathbb{Y}$* ***converges*** *on $\mathcal{D}$ if the sequence $(f(Y_n))_{n\geq 1}$ converges almost surely to a random variable $Min_{f,\mathbb{Y}}$ such that:*

$$\mathbb{P}\left[Min_{f,\mathbb{Y}} \leq \alpha\right] = 1.$$

Now given a random algorithm $\mathbb{Y} = (Y_n)_{n\geq 1}$ we define the flow of information associated to this algorithm.

Definition 5. *The* ***flow of information*** *associated to the algorithm* $\mathbb{Y} = (Y_n)_{n\geq 1}$ *for the global minimization of the function* f *is given by the natural filtration of* $(f(Y_n))_{n\geq 1}$*, which is the increasing sequence of* σ*-algebras defined by:*

$$\mathcal{F}_n^{\mathbb{Y}} := \sigma\left(f(Y_1), \ldots, f(Y_n)\right).$$

The terminal σ-algebra associated to this algorithm, $\mathcal{F}_\infty^{\mathbb{Y}}$, is naturally defined as (in the two usual notations):

$$\mathcal{F}_\infty^{\mathbb{Y}} := \sigma\left(\bigcup_{n=1}^{+\infty} \mathcal{F}_n^{\mathbb{Y}}\right) = \bigvee_{n=1}^{+\infty} \mathcal{F}_n^{\mathbb{Y}}.$$

As an immediate result we have that the filtration converges in the Cotter distance to the terminal σ-algebra.

Proposition 3. *For every stochastic algorithm* $\mathbb{Y} = (Y_n)_{n\geq 1}$*,*

$$\lim_{n\to+\infty} d_c\left(\mathcal{F}_n^{\mathbb{Y}}, \mathcal{F}_\infty^{\mathbb{Y}}\right) = 0.$$

Proof. Let's first observe that by Proposition 2.2 of Cotter (see again [Cot86]) any increasing sequence of σ-algebras converges in the Cotter distance. In fact, by a standard argument we have that:

$$\bigcap_{n=1}^{+\infty} \bigvee_{m=n}^{+\infty} \mathcal{F}_m^{\mathbb{Y}} = \mathcal{F}_\infty^{\mathbb{Y}} = \bigvee_{n=1}^{+\infty} \bigcap_{m=n}^{+\infty} \mathcal{F}_m^{\mathbb{Y}}$$

and by the result quoted this suffices to ensure that the filtration associated to the algorithm converges. Now, it is a well known fact (see [Bil95, p. 470]) that, by the definitions above, we have that almost surely:

$$\forall Z \in L^1(\Omega, \mathcal{F}, \mathbb{P}) \quad \lim_{n\to+\infty} \mathbb{E}\left[Z \mid \mathcal{F}_n^{\mathbb{Y}}\right] = \mathbb{E}\left[Z \mid \mathcal{F}_\infty^{\mathbb{Y}}\right] \tag{17}$$

as the sequence $(\mathbb{E}\left[Z \mid \mathcal{F}_n^{\mathbb{Y}}\right])_{n\geq 1}$ is uniformly integrable, (17) implies that

$$\forall Z \in L^1(\Omega, \mathcal{F}, \mathbb{P}) \quad \lim_{n\to+\infty} \left\|\mathbb{E}\left[Z \mid \mathcal{F}_n^{\mathbb{Y}}\right] - \mathbb{E}\left[Z \mid \mathcal{F}_\infty^{\mathbb{Y}}\right]\right\|_{L^1(\Omega,\mathcal{F},\mathbb{P})} = 0,$$

and this is just definition given by Formula (15).

We now compare the information content of two stochastic algorithms by comparing their information induced filtrations.

Definition 6. *Two algorithms* $\mathbb{Y}^1$ *and* $\mathbb{Y}^2$ *are* ***informationally asymptotically equivalent*** *(IAE) if and only if:*

$$\lim_{n\to+\infty} d_c\left(\mathcal{F}_n^{\mathbb{Y}^1}, \mathcal{F}_n^{\mathbb{Y}^2}\right) = 0.$$

As an easy first observation we have that two algorithms are informationally asymptotically equivalent if and only if the Cotter distance of their terminal σ-algebras is zero, that is:

Proposition 4. *Let $\mathbb{Y}^1$ and $\mathbb{Y}^2$ be two algorithms, Then:*

$$\mathbb{Y}^1 IAE\, \mathbb{Y}^2 \Leftrightarrow d_c\left(\mathcal{F}_\infty^{\mathbb{Y}^1}, \mathcal{F}_\infty^{\mathbb{Y}^2}\right) = 0. \tag{18}$$

Proof. If the two algorithms are informationally asymptotically equivalent then the condition about the terminal σ-algebras is verified as an immediate consequence of Proposition 3. In fact,

$$d_c\left(\mathcal{F}_\infty^{\mathbb{Y}^1}, \mathcal{F}_\infty^{\mathbb{Y}^2}\right) \leq d_c\left(\mathcal{F}_\infty^{\mathbb{Y}^1}, \mathcal{F}_n^{\mathbb{Y}^1}\right) + d_c\left(\mathcal{F}_n^{\mathbb{Y}^1}, \mathcal{F}_n^{\mathbb{Y}^2}\right) + d_c\left(\mathcal{F}_n^{\mathbb{Y}^2}, \mathcal{F}_\infty^{\mathbb{Y}^2}\right).$$

Now, for the converse suppose that $d_c\left(\mathcal{F}_\infty^{\mathbb{Y}^1}, \mathcal{F}_\infty^{\mathbb{Y}^2}\right) = 0$ and that the algorithms are not IAE. Then, for some $\epsilon > 0$ there exists an increasing integer sequence $(n_k^\epsilon)_{k\in\mathbb{N}}$ such that

$$\forall k \in \mathbb{N},\ \ d_c\left(\mathcal{F}_{n_k^\epsilon}^{\mathbb{Y}^1}, \mathcal{F}_{n_k^\epsilon}^{\mathbb{Y}^2}\right) \geq \epsilon.$$

We then have that for all $k \geq 1$,

$$\begin{aligned}\epsilon &\leq d_c\left(\mathcal{F}_{n_k^\epsilon}^{\mathbb{Y}^1}, \mathcal{F}_{n_k^\epsilon}^{\mathbb{Y}^2}\right) \leq d_c\left(\mathcal{F}_{n_k^\epsilon}^{\mathbb{Y}^1}, \mathcal{F}_\infty^{\mathbb{Y}^1}\right) + d_c\left(\mathcal{F}_\infty^{\mathbb{Y}^1}, \mathcal{F}_\infty^{\mathbb{Y}^2}\right) + d_c\left(\mathcal{F}_\infty^{\mathbb{Y}^1}, \mathcal{F}_{n_k^\epsilon}^{\mathbb{Y}^2}\right)\\ &= d_c\left(\mathcal{F}_{n_k^\epsilon}^{\mathbb{Y}^1}, \mathcal{F}_\infty^{\mathbb{Y}^1}\right) + d_c\left(\mathcal{F}_\infty^{\mathbb{Y}^1}, \mathcal{F}_{n_k^\epsilon}^{\mathbb{Y}^2}\right)\\ &\leq \limsup_{n\to+\infty}\left(d_c\left(\mathcal{F}_{n_k^\epsilon}^{\mathbb{Y}^1}, \mathcal{F}_\infty^{\mathbb{Y}^1}\right) + d_c\left(\mathcal{F}_\infty^{\mathbb{Y}^1}, \mathcal{F}_{n_k^\epsilon}^{\mathbb{Y}^2}\right)\right) = 0,\end{aligned}$$

again, by Proposition 3, which is a contradiction.

Our purpose now is to illustrate the intuitive idea that a convergent algorithm for minimizing a function must recover all available information about the function. For the first result we require that the algorithm exhausts all the available information in the probability space. We will suppose that the two algorithms $\mathbb{Y}^1$ and $\mathbb{Y}^2$ both converge. We will show next that, if we suppose,

$$\sigma\left(\mathrm{Min}_{f,\mathbb{Y}^1}\right) = \mathcal{F} = \sigma\left(\mathrm{Min}_{f,\mathbb{Y}^2}\right), \tag{19}$$

then, these algorithms, $\mathbb{Y}^1$ and $\mathbb{Y}^2$, are informationally asymptotic equivalent.

Theorem 5. *With the notations of Definition 4, let $\mathbb{Y}^1$ and $\mathbb{Y}^2$ be two algorithms that converge. We have that:*

$$\sigma\left(Min_{f,\mathbb{Y}^1}\right) = \mathcal{F} = \sigma\left(Min_{f,\mathbb{Y}^2}\right) \Rightarrow \mathbb{Y}^1 IAE\, \mathbb{Y}^2.$$

Proof. The proof is a consequence of the continuity of the operator that maps each random variable to the σ-algebra it generates formulated in Cotter's Theorem 4. We have that the sequences,

$$\left(\sigma\left(f(Y_n^1)\right)\right)_{n\geq 1}, \left(\sigma\left(f(Y_n^2)\right)\right)_{n\geq 1}\ ,$$

both converge in the Cotter distance to $\mathcal{F}$ by reason of the hypothesis. Now, by definition, as we have that for all $n \geq 1$,

$$\sigma\left(f(Y_n^1)\right) \subset \mathcal{F}_n^{\mathbb{Y}^1} \subset \mathcal{F}\ ,\ \sigma\left(f(Y_n^2)\right) \subset \mathcal{F}_n^{\mathbb{Y}^2} \subset \mathcal{F},$$

by Proposition 2, we have:

$$d_c\left(\mathcal{F}_n^{\mathbb{Y}^1}, \mathcal{F}\right) \leq 2d_c\left(\sigma\left(f(Y_n^1)\right), \mathcal{F}\right)\ ,\ d_c\left(\mathcal{F}_n^{\mathbb{Y}^2}, \mathcal{F}\right) \leq 2d_c\left(\sigma\left(f(Y_n^2)\right), \mathcal{F}\right)$$

and so we also have that the sequences,

$$\left(\mathcal{F}_n^{\mathbb{Y}^1}\right)_{n\geq 1}, \left(\mathcal{F}_n^{\mathbb{Y}^2}\right)_{n\geq 1},$$

converge in the Cotter distance to $\mathcal{F}$. Finally, as we have:

$$d_c\left(\mathcal{F}_n^{\mathbb{Y}^1}, \mathcal{F}_n^{\mathbb{Y}^2}\right) \leq d_c\left(\mathcal{F}_n^{\mathbb{Y}^1}, \mathcal{F}\right) + d_c\left(\mathcal{F}, \mathcal{F}_n^{\mathbb{Y}^2}\right),$$

we have the condition of Formula (19) appearing in Theorem 5.

Remark 5. If condition in Formula (19), essential in Theorem 5, is not verified – then by Cotter's theorem quoted in Theorem 4 – the map σ is not continuous at $\sigma\left(\mathrm{Min}_{f,\mathbb{Y}^1}\right)$ and $\sigma\left(\mathrm{Min}_{f,\mathbb{Y}^2}\right)$ and so – it is in general not true that the sequences $\left(\sigma\left(f(Y_n^1)\right)\right)_{n\geq 1}$ and $\left(\sigma\left(f(Y_n^2)\right)\right)_{n\geq 1}$ converge. As a consequence, despite the fact that, by Proposition 3, the sequences $\left(\mathcal{F}_n^{\mathbb{Y}^1}\right)_{n\geq 1}$ and $\left(\mathcal{F}_n^{\mathbb{Y}^2}\right)_{n\geq 1}$ both converge – to $\mathcal{F}_\infty^{\mathbb{Y}^1}$ and $\mathcal{F}_\infty^{\mathbb{Y}^2}$, respectively – we can not ensure that the condition given by Formula (18) in Proposition 4 is verified and so, we can not conclude that the two algorithms are IAE.

If moreover the algorithms are informationally asymptotic equivalent, and their associated limit minimum functions take a denumerable set of values, then their associated limit minimum functions will coincide almost surely thus saying, essentially, that two IAE convergent algorithms carry the same information content with respect to the minimization function.

Theorem 6. *With the notations of Definition 4, let $\mathbb{Y}^1$ and $\mathbb{Y}^2$ be two algorithms that converge. Let us suppose that the set $Min_{f,\mathbb{Y}^1}(\Omega) \cup Min_{f,\mathbb{Y}^2}(\Omega)$ is denumerable. We then have that:*

$$\mathbb{Y}^1 IAE\, \mathbb{Y}^2 \Rightarrow Min_{f,\mathbb{Y}^1} = Min_{f,\mathbb{Y}^2}\ \ a.\ s. \tag{20}$$

Proof. The announced result is a consequence of Proposition 4. In fact, if $\mathbb{Y}^1$ and $\mathbb{Y}^2$ are IAE then this means that:

$$\mathcal{F}_\infty^{\mathbb{Y}^1} \sim \mathcal{F}_\infty^{\mathbb{Y}^2},$$

and so by (14), for every B in the Borel σ-algebra of the reals $\mathcal{B}(\mathbb{R})$,

$$\mathbb{P}\left[\mathrm{Min}_{f,\mathbb{Y}^1}^{-1}(B) \setminus \mathrm{Min}_{f,\mathbb{Y}^2}^{-1}(B)\right] = 0 = \mathbb{P}\left[\mathrm{Min}_{f,\mathbb{Y}^2}^{-1}(B) \setminus \mathrm{Min}_{f,\mathbb{Y}^1}^{-1}(B)\right] \tag{21}$$

Now, consider $B = \{x\} \in \mathcal{B}(\mathbb{R})$. Formulas (21) imply that:

$$\mathbb{P}\left[\left\{\omega \in \Omega \mid \mathrm{Min}_{f,\mathbb{Y}^1}(\omega) \neq x\right\} \cup \left\{\omega \in \Omega \mid \mathrm{Min}_{f,\mathbb{Y}^2}(\omega) = x\right\}\right] = 1,$$

and also

$$\mathbb{P}\left[\left\{\omega \in \Omega \mid \mathrm{Min}_{f,\mathbb{Y}^2}(\omega) \neq x\right\} \cup \left\{\omega \in \Omega \mid \mathrm{Min}_{f,\mathbb{Y}^1}(\omega) = x\right\}\right] = 1.$$

Now, by considering the intersection

$$\left(\left\{\mathrm{Min}_{f,\mathbb{Y}^1} \neq x\right\} \cup \left\{\mathrm{Min}_{f,\mathbb{Y}^2} = x\right\}\right) \cap \left(\left\{\mathrm{Min}_{f,\mathbb{Y}^2} \neq x\right\} \cup \left\{\mathrm{Min}_{f,\mathbb{Y}^1} = x\right\}\right),$$

which is is a set of probability one, we get by expanding that for every $x \in \mathbb{R}$:

$$\mathbb{P}\left[\left\{\mathrm{Min}_{f,\mathbb{Y}^1} \neq x \wedge \mathrm{Min}_{f,\mathbb{Y}^2} \neq x\right\} \cup \left\{\mathrm{Min}_{f,\mathbb{Y}^1} = x \wedge \mathrm{Min}_{f,\mathbb{Y}^2} = x\right\}\right] = 1.$$

And so by considering the denumerable set $\mathrm{Im} = \mathrm{Min}_{f,\mathbb{Y}^1}(\Omega) \cup \mathrm{Min}_{f,\mathbb{Y}^2}(\Omega)$, as

$$\begin{aligned}
&\left\{\mathrm{Min}_{f,\mathbb{Y}^1} \neq \mathrm{Min}_{f,\mathbb{Y}^2}\right\} \\
&\subseteq \bigcup_{x \in \mathrm{Im}} \left\{\mathrm{Min}_{f,\mathbb{Y}^1} = x \wedge \mathrm{Min}_{f,\mathbb{Y}^2} \neq x\right\} \cup \left\{\mathrm{Min}_{f,\mathbb{Y}^1} \neq x \wedge \mathrm{Min}_{f,\mathbb{Y}^2} = x\right\} \\
&= \bigcup_{x \in \mathrm{Im}} \left(\left\{\mathrm{Min}_{f,\mathbb{Y}^1} \neq x \wedge \mathrm{Min}_{f,\mathbb{Y}^2} \neq x\right\} \cup \left\{\mathrm{Min}_{f,\mathbb{Y}^1} = x \wedge \mathrm{Min}_{f,\mathbb{Y}^2} = x\right\}\right)^c
\end{aligned}$$

we will have that $\mathbb{P}\left[\mathrm{Min}_{f,\mathbb{Y}^1} \neq \mathrm{Min}_{f,\mathbb{Y}^2}\right] = 0$, as wanted.

The particular case of an unique minimizer of a continuous function on a compact domain deserves special mention as a case where two algorithms having IAE lead to the same minimizing function almost surely.

Proposition 5. *With the notations of definition 4, let $\mathbb{Y}^1$ and $\mathbb{Y}^2$ be two algorithms that converge. Suppose, furthermore, that f is continuous, that f admits an unique minimizer z and $\mathcal{D}$ is compact. Then we have that:*

$$\mathbb{Y}^1 \mathit{IAE}\, \mathbb{Y}^2 \Rightarrow \begin{cases} \lim_{n\to+\infty} f(Y_n^1) = f(z) = \lim_{n\to+\infty} f(Y_n^2) \ \ a.\ s. \\ \lim_{n\to+\infty} Y_n^1 = z = \lim_{n\to+\infty} Y_n^2 \ \ a.\ s. \end{cases} .$$

Proof. As we have $\lim_{n\to+\infty} f(Y_n^1) = f(z) = \lim_{n\to+\infty} f(Y_n^2)$ and $\lim_{n\to+\infty} Y_n^1 = z = \lim_{n\to+\infty} Y_n^2$, by Theorem 2 and Proposition 1, we also have that $\mathrm{Min}_{f,\mathbb{Y}^1} = f(z) = \mathrm{Min}_{f,\mathbb{Y}^2}$ almost surely and so, by Theorem 6, we have the announced result.

Remark 6. Let us observe, with respect to Proposition 5, that under the hypothesis stated in that proposition, that is, if we have almost surely,

$$\mathrm{Min}_{f,\mathbb{Y}^1} = \mathrm{Min}_{f,\mathbb{Y}^2} = f(z),$$

then, by modifying $\mathrm{Min}_{f,\mathbb{Y}^1}$ and $\mathrm{Min}_{f,\mathbb{Y}^2}$ on sets of probability zero we would have that:

$$\sigma\left(\mathrm{Min}_{f,\mathbb{Y}^1}\right) = \{\emptyset, \varOmega\} = \sigma\left(\mathrm{Min}_{f,\mathbb{Y}^2}\right).$$

By Remark 4, in general, under the hypothesis of Proposition 5, the two σ-algebras $\sigma\left(\mathrm{Min}_{f,\mathbb{Y}^1}\right)$ and $\sigma\left(\mathrm{Min}_{f,\mathbb{Y}^1}\right)$ are equal to $\{\emptyset, \varOmega\}$ in $\mathfrak{F} := \mathfrak{F}^\star / \sim$ – the class of all sub-σ-algebras of $\mathcal{F}$ identified up to sets of probability zero – and so the condition in Formula (19) – which is essential in Theorem 5 – may be verified only for deterministic algorithms (as in this case all random variables are constant).

Acknowledgements. This work was partially supported, for the first and fourth authors, through the project UID/MAT/00297/2020 of the Centro de Matemática e Aplicações, financed by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) and, for the third author, through RFBR (Grant n. 19-01-00451).

A Appendix

Deduction of Formula (2). Let λ_x denote the Lebesgue measure over $\mathcal{D}$ applied to the set defined by the variable x.

$$\begin{aligned}
&\mathbb{P}[Y_n \in D] \\
&= \sum_{k=1}^{n} \mathbb{P}\left[\{X_k \in D\} \cap \bigcap_{1\le j<k} \{f(X_k) \le f(X_j\} \cap \bigcap_{k<j\le n} \{f(X_k) < f(X_j)\}\right] \\
&= \sum_{k=1}^{n} \left(\frac{1}{\lambda(\mathcal{D})^n} \int_{\mathcal{D}^n} \mathbb{1}_{\{x_k \in D\}} \prod_{1\le j<k} \mathbb{1}_{\{f(x_k)\le f(x_j\}} \right. \\
&\left. \times \prod_{k<j\le n} \mathbb{1}_{\{f(x_k)<f(x_j)\}} d\lambda(x_1) \dots d\lambda(x_n) \right) \\
&= \sum_{k=1}^{n} \left(\frac{1}{\lambda(\mathcal{D})} \int_{\mathcal{D}} d\lambda(x_k) \frac{1}{\lambda(\mathcal{D})^{k-1}} \prod_{1\le j<k} \int_{\mathcal{D}^{k-1}} \mathbb{1}_{\{f(x_k)\le f(x_j\}} d\lambda(x_j) \right. \\
&\left. \times \frac{1}{\lambda(\mathcal{D})^{n-k}} \prod_{k<j\le n} \int_{\mathcal{D}^{n-k}} \mathbb{1}_{\{f(x_k)<f(x_j)\}} d\lambda(x_j) \right) \\
&= \sum_{k=1}^{n} \left(\frac{1}{\lambda(\mathcal{D})} \int_{\mathcal{D}} \mathbb{1}_{\{x_k \in D\}} \frac{\lambda_x(\{f(x_k) \le f(x\})^{k-1}}{\lambda(\mathcal{D})^{k-1}} \right. \\
&\left. \times \frac{\lambda_x(\{f(x_k) < f(x\})^{n-k}}{\lambda(\mathcal{D})^{n-k}} d\lambda(x_k) \right) \\
&= \sum_{k=1}^{n} \left(\frac{1}{\lambda(\mathcal{D})^n} \int_{D} \lambda(f^{-1}([f(x_k), +\infty[)^{k-1} \lambda(f^{-1}(]f(x_k), +\infty[)^{n-k} d\lambda(x_k) \right).
\end{aligned}$$

References

[ALR03] Appel, M.J., LaBarre, R., Radulović, D.: On accelerated random search. SIAM J. Optim. **14**(3), 708–731 (2003). (Electronic)

[Art01] Artstein, Z.: Compact convergence of σ-fields and relaxed conditional expectation. Probab. Theory Relat. Fields **120**(3), 369–394 (2001)

[Bar04] Barty, K.: Contributions à la discrétisation des contraintes de mesurabilité pour les problèmes d'optimisation stochastique. Ph.D. thesis, École Nationale des Ponts et Chaussées, Paris, France, June 2004

[Bil95] Billingsley, P.: Probability and Measure. Wiley Series in Probability and Mathematical Statistics, 3rd edn. Wiley, New York (1995)

[Boy71] Boylan, E.S.: Equiconvergence of martingales. Ann. Math. Stat. **42**, 552–559 (1971)

[CJK07] Costa, A., Jones, O.D., Kroese, D.: Convergence properties of the cross-entropy method for discrete optimization. Oper. Res. Lett. **35**(5), 573–580 (2007)

[Cot86] Cotter, K.D.: Similarity of information and behavior with a pointwise convergence topology. J. Math. Econ. **15**(1), 25–38 (1986)

[Cot87] Cotter, K.D.: Convergence of information, random variables and noise. J. Math. Econ. **16**(1), 39–51 (1987)

[dBKMR05] de Boer, P.-T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. Ann. Oper. Res. **134**, 19–67 (2005)

[dC11] de Carvalho, M.: Confidence intervals for the minimum of a function using extreme value statistics. Int. J. Math. Modell. Numer. Optim. **2**(9), 288–296 (2011)

[dC12] de Carvalho, M.: A generalization of the Solis and Wets method. J. Stat. Plann. Inference **142**(3), 633–644 (2012)

[Esq06] Esquível, M.L.: A conditional Gaussian martingale algorithm for global optimization. In: Gavrilova, M., et al. (eds.) ICCSA 2006, Part III. LNCS, vol. 3982, pp. 841–851. Springer, Heidelberg (2006). https://doi.org/10.1007/11751595_89

[Kom08] Komisarski, A.: Distances between σ-fields on a probability space. J. Theoret. Probab. **21**(4), 812–823 (2008)

[Kud74] Kudō, H.: A note on the strong convergence of σ-algebras. Ann. Probab. **2**(1), 76–83 (1974)

[MPB99] Mexia, J.T., Pereira, D., Baeta, J.: L_2 environmental indexes. Listy Biom. **36**(2), 137–143 (1999)

[Nev64] Neveu, J.: Bases mathématiques du calcul des probabilités. Masson et Cie, Éditeurs, Paris (1964)

[Nev65] Neveu, J.: Mathematical Foundations of the Calculus of Probability. Translated by Amiel Feinstein. Holden-Day Inc., San Francisco (1965)

[Nev72] Neveu, J.: Note on the tightness of the metric on the set of complete sub σ-algebras of a probability space. Ann. Math. Stat. **43**, 1369–1371 (1972)

[Pic98] Piccinini, L.: Convergence of nonmonotone sequence of sub-σ-fields and convergence of associated subspaces $L^p(\mathcal{B}_n)(p \in [1, +\infty])$. J. Math. Anal. Appl. **225**(1), 73–90 (1998)

[PM10] Pereira, D.G., Mexia, J.T.: Comparing double minimization and zigzag algorithms in joint regression analysis: the complete case. J. Stat. Comput. Simul. **80**(1–2), 133–141 (2010)

[PS00] Peng, J., Shi, D.: Improvement of pure random search in global optimization. J. Shanghai Univ. **4**(2), 92–95 (2000)

[RK04] Rubinstein, R.Y., Kroese, D.P.: The Cross-Entropy Method. Information Science and Statistics, Springer, New York (2004). https://doi.org/10.1007/978-1-4757-4321-0. A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning

[RK08] Rubinstein, R.Y., Kroese, D.P.: Simulation and the Monte Carlo Method. Wiley Series in Probability and Statistics, 2nd edn. Wiley-Interscience, Hoboken (2008)

[Rog74] Rogge, L.: Uniform inequalities for conditional expectations. Ann. Probab. **2**, 486–489 (1974)

[RS03] Raphael, B., Smith, I.F.C.: A direct stochastic algorithm for global search. Appl. Math. Comput. **146**(2–3), 729–758 (2003)

[SB98] Stephens, C.P., Baritompa, W.: Global optimization requires global information. J. Optim. Theory Appl. **96**(3), 575–588 (1998)

[SP99] Shi, D., Peng, J.: A new theoretical framework for analyzing stochastic global optimization algorithms. J. Shanghai Univ. **3**(3), 175–180 (1999)

[Spa03] Spall, J.C.: Introduction to Stochastic Search and Optimization. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, Hoboken (2003). Estimation, Simulation, and Control

[Spa04] Spall, J.C.: Stochastic optimization. In: Handbook of Computational Statistics, pp. 169–197. Springer, Berlin (2004)

[SW81] Solis, F.J., Wets, R.J.-B.: Minimization by random search techniques. Math. Oper. Res. **6**(1), 19–30 (1981)

[Vid18] Vidmar, M.: A couple of remarks on the convergence of σ-fields on probability spaces. Statist. Probab. Lett. **134**, 86–92 (2018)

[VZ93] Van Zandt, T.: The Hausdorff metric of σ-fields and the value of information. Ann. Probab. **21**(1), 161–167 (1993)

[Wan01] Wang, X.: Convergence rate of conditional expectations. Sci. Math. Jpn. **53**(1), 83–87 (2001)

[Yin99] Yin, G.: Rates of convergence for a class of global stochastic optimization algorithms. SIAM J. Optim. **10**(1), 99–120 (1999). (Electronic)

[Zab03] Zabinsky, Z.B.: Stochastic Adaptive Search for Global Optimization. Nonconvex Optimization and its Applications, vol. 72. Kluwer Academic Publishers, Boston (2003)