On Some Stochastic Algorithms for Global Optimization

Preliminary report

Manuel L. Esquível, Miguel Carvalho, João Tiago Mexia

International Conference on Advances in Interdisciplinary Statistics and Combinatorics

Greensboro, October 13, 2007
Outline

1. Optimization, Random Algorithms and Information
   - Introduction
   - Some driving general ideas

2. Examples of Random Algorithms and Convergence Results
   - Three Random Algorithms
   - The Solis and Wets approach

3. Information and algorithms
   - Information for algorithms
   - An information distance
Introduction

Outline

1. Optimization, Random Algorithms and Information
   - Introduction
   - Some driving general ideas

2. Examples of Random Algorithms and Convergence Results
   - Three Random Algorithms
   - The Solis and Wets approach

3. Information and algorithms
   - Information for algorithms
   - An information distance
Introduction

The problem

**Global Optimization**
- $f : D \subseteq \mathbb{R}^P \mapsto \mathbb{R}$ measurable
- $\min_{x \in D} f(x) ?$
- $x_0$ such that $f(x_0) = \min_{x \in D} f(x) ?$

**Random Algorithms**
- For $n = 1, \ldots, N$, $X_n : D$ valued random variables
- $\{f(X_1), f(X_2), \ldots f(X_N)\}$ rough description of $f$ over $D$. 

**Information problem**
How to study the flow of information gained by the sequential observation of a random algorithm?
Outline

1. Optimization, Random Algorithms and Information
   - Introduction
   - Some driving general ideas

2. Examples of Random Algorithms and Convergence Results
   - Three Random Algorithms
   - The Solis and Wets approach

3. Information and algorithms
   - Information for algorithms
   - An information distance
Some driving general ideas

Stephens and Baritompa result (1998)

*Global optimization requires global information*

Consider a random algorithm $X_1^f, \ldots, X_n^f, \ldots$ for some function $f$. The closure $\overline{X^f} := \{X_1^f, \ldots, X_n^f, \ldots\}$ is a random set in $\mathcal{D}$.

**Theorem**

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

Roughly, the algorithm works on any rich class of functions iff $\mathbb{P}[\overline{X^f} = \mathcal{D}] = 1$. 
Some driving general ideas

Two possible approaches

Entropy and Information

1. Relative entropy of the sequences of random variables defining the algorithms (not yet developed).
2. The information flow generated by the sequence defining the algorithm (partially developed).
Three Random Algorithms

Outline

1. Optimization, Random Algorithms and Information
   - Introduction
   - Some driving general ideas

2. Examples of Random Algorithms and Convergence Results
   - Three Random Algorithms
   - The Solis and Wets approach

3. Information and algorithms
   - Information for algorithms
   - An information distance
Three Random Algorithms

Pure Random Search
Outline and Probabilistic Description

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:

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

S.3 Repeat S.2.

- Let $X_1, X_2, \ldots, X_n, \ldots$ be independent random variables with common distribution over $\mathcal{D}$ verifying furthermore:

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

- $Y_1 := X_1$

- $Y_{n+1} = Y_n \mathbb{I}\{f(Y_n) < f(X_{n+1})\} + X_{n+1} \mathbb{I}\{f(Y_n) \geq f(X_{n+1})\}$
Pure Random Search

The density of the algorithm random variables

Remark

$((X_n))_{n \geq 1}$ uniformly distributed in $D$. Then $\{Y_n \in D\}$ is

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

If for every $x \in D$ we have $\lambda(f^{-1}(\{x\}) = 0$, then

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

May be used for relative entropy calculation purposes
Three Random Algorithms

Random search on a nearly unbounded domain

Outline and Probabilistic Description

S.1 Select a point $x$ at random in $\mathcal{D}$. Do $z := x$.

S.2 Choose a point $x$ at random in $\mathcal{D}$. Choose a point $y$ with distribution $\mathcal{N}(x, \sigma)$ with, for instance, $\sigma := \text{diam}(\mathcal{D})/10$. Do:

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

S.3 Repeat S.2.

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

The Zig-Zag algorithm I

Outline

S.1 Select a point $x$ at random in $\mathcal{D}$. Do $z := x$.

S.2 (Optimization along an one 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
Three Random Algorithms

The Zig-Zag algorithm II
Probabilistic Description

- Let $Y_1, Y_2, \ldots, Y_n, \ldots$ be a sequence of independent random variables with common uniform distribution over $\mathcal{D}$.
- $Z_1 := Y_1$
- For each $n \geq 2$, 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:
  \[
  \forall \lambda \in [a, b] \; \forall x, y \in \mathcal{D} \; \lambda x + (1 - \lambda)y \in \mathcal{D}.
  \]
  which is possible as $\mathcal{D}$ is bounded.
- Define the random variable $X_{n0}$ such that:
  \[
  f(X_{n0}^j) = \min_{1 \leq j \leq N} f(\lambda_j^n Z_n + (1 - \lambda_j^n)Y_n)
  \]
- $Z_{n+1} := Z_n \mathbb{I}\{f(Z_n) < f(X_{n0}^j)\} + X_{n0}^j \mathbb{I}\{f(Z_n) \geq f(X_{n0}^j)\}$. 
Outline

1. Optimization, Random Algorithms and Information
   - Introduction
   - Some driving general ideas

2. Examples of Random Algorithms and Convergence Results
   - Three Random Algorithms
   - The Solis and Wets approach

3. Information and algorithms
   - Information for algorithms
   - An information distance
Some definitions
Essential infimum and optimality region

Essential infimum of $f$ in $\mathcal{D}$

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

Optimality region of $f$ of height $\epsilon$ over $\alpha$

$$E_{\alpha+\epsilon, M} := \begin{cases} 
\{x \in \mathcal{D} : f(x) < \alpha + \epsilon\} & \text{if } \alpha \in \mathbb{R} \\
\{x \in \mathcal{D} : f(x) < M\} & \text{if } \alpha = -\infty
\end{cases}$$  \hspace{1cm} (2)
The Solis and Wets approach

The structure function and the algorithm

The structure function

\[ \psi : \mathcal{D} \times \mathbb{R}^n \mapsto \mathbb{R} \text{ such that } [H1] \text{ is verified.} \]

\[ [H1] : \begin{cases} 
\forall t, x \quad f(\psi(t, x)) \leq f(t) \\
\forall x \in \mathcal{D} \quad f(\psi(t, x)) \leq \min(f(x), f(t)) 
\end{cases} \]

The algorithm

A sequence of random variables given by

\[ \begin{cases} 
Y_1 = X_1 \\
Y_{n+1} = \psi(Y_n, X_n) \quad \text{for } n \geq 1
\end{cases} \]

\[ X_n \sim \mathbb{P}_n \text{ and } \mathbb{P}_n \text{ function of } \mathbb{P}_1, \ldots, \mathbb{P}_{n-1} \text{ (adaptive r. s.)} \]
Examples of Solis and Wets approach

Remark

Pure random search, random search on nearly unbounded domains and the zig-zag algorithm are instances of Solis and Wets approach.

\[ \psi(t, x) = t \mathbb{I}_{\{f(t) < f(x)\}}(t, x) + x \mathbb{I}_{\{f(t) \geq f(x)\}}(t, x) \]

is the structure function of the algorithms and verifies the hypothesis \([H1]\).
The Solis and Wets approach

**General convergence result I**

**Hypothesis**

- $f$ bounded below.
- For pure random search, $H2(\epsilon)$:
  \[
  \lim_{k \to +\infty} \prod_{1 \leq j \leq k} \mathbb{P}[X_j \in E^c_{\alpha+\epsilon,M}] = \lim_{k \to +\infty} \prod_{1 \leq j \leq k} \mathbb{P}_j[E^c_{\alpha+\epsilon,M}] = 0. 
  \]
- For adaptive search, $H'2(\epsilon)$:
  \[
  \lim_{k \to +\infty} \inf_{1 \leq j \leq k} \mathbb{P}[X_j \in E^c_{\alpha+\epsilon,M}] = \lim_{k \to +\infty} \inf_{1 \leq j \leq k} \mathbb{P}_j[E^c_{\alpha+\epsilon,M}] = 0. 
  \]
The Solis and Wets approach

General convergence result II

Theorem

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

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

- 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 $Y_\star$ such that $\mathbb{P}[Y_\star \leq \alpha] = 1.$
General convergence result III

Remarks
- If the minimizer is not unique then the sequence \((Y_n)_{n \geq 1}\) may not converge.
- If the minimizer of \(f\) is unique and \(f\) is continuous, then the essential minimum of \(f\) coincides with the minimum of \(f\).

Theorem
- If \(f\) is continuous and admits an unique minimizer \(z \in \mathcal{D}\) then \(\lim_{n \to +\infty} f(Y_n) = f(z)\).
- If, furthermore, \(\mathcal{D}\) is compact then \(\lim_{n \to +\infty} Y_n = z\).
# Outline

1. **Optimization, Random Algorithms and Information**
   - Introduction
   - Some driving general ideas

2. **Examples of Random Algorithms and Convergence Results**
   - Three Random Algorithms
   - The Solis and Wets approach

3. **Information and algorithms**
   - Information for algorithms
   - An information distance
The flow of information of an algorithm

For an algorithm $X_1, \ldots, X_n, \ldots$ the natural filtration $G_n := \sigma(X_1, \ldots, X_n)$ describe the flow of information gained by the sequential observation of $X_1, \ldots, X_n, \ldots$.

Main driving ideas

1. To compare algorithm $X_1, \ldots, X_n, \ldots$ and algorithm $Y_1, \ldots, Y_n, \ldots$ compare the natural filtrations $\sigma(X_1, \ldots, X_n)$ and $\sigma(Y_1, \ldots, Y_n)$.

2. To compare filtrations use some notion of distance.
Outline

1. Optimization, Random Algorithms and Information
   - Introduction
   - Some driving general ideas

2. Examples of Random Algorithms and Convergence Results
   - Three Random Algorithms
   - The Solis and Wets approach

3. Information and algorithms
   - Information for algorithms
   - An information distance
An information distance

The metric space of the complete $\sigma$-algebras

Definition

$\mathcal{G}$, the set of all $\sigma$-algebras $G \subseteq \mathcal{F}$ complete with respect to $\mathbb{P}$.

A distance for complete $\sigma$-algebras

- $d$ defined on $\mathcal{G} \times \mathcal{G}$

$$d(\mathcal{H}, \mathcal{G}) = \sup_{\|X\|_2 \leq 1} \mathbb{E} \left[ \| \mathbb{E}[X|\mathcal{H}] - \mathbb{E}[X|\mathcal{G}] \| \right].$$

$\mathcal{H}, \mathcal{G} \in \mathcal{G}$ and $\|X\|_2$ the $L^2$ norm of $X$.

- $(\mathcal{G}, d)$ is a metric space.
The metric space of the complete $\sigma$-algebras

Problem: How to compute $d$ in particular cases? (difficult)

Theorem

$(O_n)_{n \geq 1}$ complete orthonormal set in $L^2(\Omega, \mathcal{F}, \mathbb{P})$. For $\mathcal{H}, \mathcal{G} \in \mathcal{F}$:

$$d(\mathcal{H}, \mathcal{G}) = \sup_{\|X\|_2 \leq 1} \mathbb{E} \left[ \sum_{n=1}^{+\infty} < X, O_n > \right] \mathbb{E}[O_n | \mathcal{H}] - \mathbb{E}[O_n | \mathcal{G}]$$

It follows that

$$\forall n \geq 1 \quad \mathbb{E}[O_n | \mathcal{H}] = \mathbb{E}[O_n | \mathcal{G}] \Rightarrow d(\mathcal{H}, \mathcal{G}) = 0,$$
**Definition: Information for algorithms**

\[ X_1^f, \ldots X_n^f, \ldots \text{ and } Y_1^f, \ldots Y_n^f, \ldots \] b two random algorithms on a function \( f \), \( H_n^f := \sigma(X_1^f, \ldots X_n^f) \) and \( G_n^f := \sigma(Y_1^f, \ldots Y_n^f) \) the natural filtrations.

- The algorithms **generate the same information** iff for every \( n \geq 1 \) we have \( d(H_n^f, G_n^f) = 0 \) for any \( f \) in a sufficiently rich class of functions.

- The algorithms **generate the same information asymptotically** iff \( \lim_{n \to +\infty} d(H_n^f, G_n^f) = 0 \) for any \( f \) in a sufficiently rich class of functions.

**Conjecture**

All convergent algorithms in a sufficiently rich class of functions generate the same information asymptotically.
Main References