2010 Mathematics Subject Classification: 62F10, 68W01.

Confidence intervals for the minimum of a function using extreme value statistics

Miguel de Carvalho

Ecole Polytechnique Fédérale de Lausanne,
Swiss Federal Institute of Technology,
CH-1015, Lausanne, Switzerland
E-mail: miguel.carvalho@epfl.ch

Abstract: Stochastic search algorithms are becoming an increasingly popular tool in the optimization community. The random structure of these methods allows us to sample from the range of a function and to obtain estimates of its global minimum. However, a major advantage of stochastic search algorithms over deterministic algorithms, which is frequently unexplored, is that they also allow us to obtain interval estimates. In this paper, we put forward such advantage by providing guidance on how to combine stochastic search and optimization methods with extreme value theory. To illustrate this approach we use several well-known objective functions. The obtained results are encouraging, suggesting that the interval estimates yield by this approach, can be helpful for supplementing point estimates produced by other sophisticated optimization methods.

Keywords: extreme value theory; stochastic optimization; unconstrained optimization.

Reference to this paper should be made as follows: Carvalho, M. de (2010) 'Confidence intervals for the minimum of a function using extreme value statistics', *Int. J. Mathematical Modelling and Numerical Optimisation*, Vol. 1, No. 1, pp.xxx–xxx.

Biographical notes: Miguel de Carvalho is post-doctoral researcher at the Ecole Polytechnique Fédérale de Lausanne, Swiss Federal Institute of Technology.

1 Introduction

The search for an input value which fulfills a well-defined output target is a problem of interest in a wide variety of scenarios. In optimization theory, the input value of interest is the one which maximizes/minimizes some profit/cost criteria. Typically, these searches are made either through deterministic or stochastic search algorithms. In this paper we are particularly concerned with optimization methods where the search direction is randomly dictated. For an inventory of deterministic

optimization methods see, for example, Nocedal and Wright (1999). Stochastic search and optimization algorithms have been applied in a wide variety of scenarios. The scope of the topic is broad enough to comprehend applications ranging from game theory (Pakes and McGuire, 2001), to the clustering of multivariate data (Booth et al., 2008). An introductory overview of stochastic search and optimization methods can be found, for instance, in Duflo (1996), or in Spall (2003).

Although the No Free Lunch theorems (Wolpert and Macready, 1997) preclude the existence of an *a priori* ideal optimization search strategy, a major advantage of stochastic search algorithms over deterministic algorithms, which is frequently unexplored, is that they also allow its user to obtain interval estimates. In this paper, we exploit and put forward some consequences of such advantage providing guidance on how to combine random optimization methods with extreme value theory (Haan and Ferreira, 2006; Resnick, 2007). This connection arises naturally since the main concern of extreme value theory lies precisely in the extremes of a sequence of random variables. Applications of extreme value theory arise in the fields of banking (Vries, 2005), finance (Poon et al., 2005), industrial production (Borbot et al., 2007), wildfire analysis (Turkman et al., 2010), sport statistics (Einmahl, 2008), among others. A broad share of this statistical paradigm is founded on Karamata's regular variation, which places the methods at an elegant mathematical support (Bingham, 1987).

2 Interval estimates for the minimum of a function

The asymptotic results derived in this section will act as a bridge linking stochastic optimization methods with extreme value theory. Let L denote a continuous cost function of interest defined over $\boldsymbol{\Theta}$ a subset of $\mathbb{R}^k$. From the conceptual stance, we can model the set $\{(\boldsymbol{\theta}, \mathrm{L}(\boldsymbol{\theta})) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, as a population of interest from which we intend to consistently estimate the parameters

$$(\boldsymbol{\theta}^\star, l^\star) := \left(\arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{L}\left(\boldsymbol{\theta}\right), \min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{L}\left(\boldsymbol{\theta}\right) \right).$$

To put this differently, we intend to use data from the graph of L to estimate its minimum. In order to do so, suppose we collect a random sample $\{(\boldsymbol{\theta}_i, \mathrm{L}(\boldsymbol{\theta}_i)\}_{i=1}^n$ from such population. Hence for each sampled value $\boldsymbol{\theta}_i$, we also inquire its corresponding image value $\mathrm{L}(\boldsymbol{\theta}_i)$. Assume that the sampling scheme collects sequentially the $\boldsymbol{\theta}_i$. This leads us to take $\boldsymbol{\theta}_1$ as our first guess for the minimum, and so we set $\widehat{\boldsymbol{\theta}}_1 := \boldsymbol{\theta}_1$. During the next extraction periods we keep computing

$$\widehat{\boldsymbol{\theta}}_{i+1} = \widehat{\boldsymbol{\theta}}_i \, \mathbb{I}\{L_{1:i} \leq \mathrm{L}(\boldsymbol{\theta}_{i+1})\} + \boldsymbol{\theta}_{i+1} \, \mathbb{I}\{L_{1:i} > \mathrm{L}(\boldsymbol{\theta}_{i+1})\}, \; i = 1, \ldots, n-1.$$

Here we use the notation $L_{1:i} \leq L_{2:i} \leq \ldots \leq L_{i:i}$ to represent the order statistics of $\mathrm{L}(\widehat{\boldsymbol{\theta}}_1), \mathrm{L}(\widehat{\boldsymbol{\theta}}_2), \ldots, \mathrm{L}(\widehat{\boldsymbol{\theta}}_i)$. Below we state Renyi's decomposition—an important result for modelling the order statistics of the image values of L (Haan and Ferreira, 2006).

Lemma 2.1. *Let E be a random variable with standard exponential distribution. Let $E_{1:n},\ldots,E_{n:n}$ denote the n-th order statistics from a standard exponential distribution. The following decomposition holds,*

$$E_{i:n} \overset{d}{=} \sum_{j=1}^{i} \frac{E_j^*}{n-j+1}, \tag{1}$$

where E_j^* are independent and identically distributed with E.

A major advantage of Renyi's decomposition, is that it allows us to write the order statistics of the image values $L_{1:n},\ldots,L_{n:n}$, as a function of the order statistics of a standard exponential distribution

$$(L_{1:n},\ldots,L_{n:n}) = (\psi(E_{1:n}),\ldots,\psi(E_{n:n})) \overset{d}{=} \left(\psi\left(\frac{E_1^*}{n}\right),\ldots,\psi\left(\sum_{j=1}^{n}\frac{E_j^*}{n-j+1}\right)\right).$$

Here ψ is an inverse function of $-\log(1-H)$, where H denotes the distribution function of the order statistics $L_{1:n},\ldots,L_{n:n}$. The archetypal type of assumption made in extreme value theory is also made here; we assume regular variation on $\psi(x)-l^\star$ at 0+, with index $\alpha>0$, i.e.:

$$\frac{\psi(tx)-l^\star}{\psi(t)-l^\star} \underset{t\to 0^+}{\longrightarrow} x^{1/\alpha}, \quad \forall x>0. \tag{2}$$

This is a minor assumption that is made in the literature in the same context (Haan, 1981). For a unified overview on regular variation, see Bingham (1987); Haan and Ferreira (2006). Renyi's decomposition is put at work below where we state a set results which will be useful for constructing interval estimates for $l^\star$.

Lemma 2.2. *The following asymptotic results hold.*

1. $\frac{L_{1:n}-l^\star}{\psi(n^{-1})-l^\star} \xrightarrow{\mathcal{D}} E_1^{*1/\alpha}$.
2. $\frac{L_{2:n}-l^\star}{\psi(n^{-1})-l^\star} \xrightarrow{\mathcal{D}} \left(E_1^*+E_2^*\right)^{1/\alpha}$.
3. $\frac{L_{1:n}-l^\star}{L_{2:n}-L_{1:n}} \xrightarrow{\mathcal{D}} \frac{E_1^{*1/\alpha}}{\left(E_1^*+E_2^*\right)^{1/\alpha}-E_1^{*1/\alpha}}$.

The next result is based on asymptotic arguments put forward by Haan (1981). This is a cornerstone result linking stochastic search and optimization methods with extreme value theory. A sketch of the proof is given in the appendix.

Theorem 2.3. *Consider the auxiliary set-valued function $\Psi:\mathbb{N}\times[0;1]\rightrightarrows\mathbb{R}$*

$$\Psi_n(p) = \left]L_{1:n} - \frac{L_{2:n}-L_{1:n}}{(1-p)^{-1/\alpha}-1}; L_{1,n}\right[. \tag{3}$$

Then, as $n\to\infty$

$$\operatorname{pr}\{l^\star\in\Psi_n(p)\}-(1-p)=o(1).$$

Theoretical applications of Theorem 2.3 can be found, for instance, in Romeijn and Smith (1994) and Carvalho (2010). We follow de Haan's general recommendation to set the parameter $\alpha = k/2$, where k denotes the number of dimensions of the problem of interest. Technical details regarding this choice are beyond the scope of this work and can be found in Section 3 of Haan (1981). Thus, in practice we use the following modified version of (3)

$$\Psi_n(p) = \left] L_{1:n} - \frac{L_{2:n} - L_{1:n}}{(1-p)^{-\frac{2}{k}} - 1}; L_{1,n} \right[. \tag{4}$$

Table 1 Search domains of the test functions and their global minimum values.

Test function	Search domain	$l^\star$
Beale	$[-4.5, 4.5]^2$	0
Easom	$[-100, 100]^2$	-1
Griewank	$[-600, 600]^2$	0
Rastrigin	$[-5.12, 5.12]^2$	0
Rosembrock	$[-5, 10]^2$	0
Styblinski–Tang	$[-8, 8]^2$	-78.33

3 Computational experiment

The method suggested by Theorem 2.3 is extremely easy to apply; it uses the following inputs: two order statistics of the image values $(L_{1:n}, L_{2:n})$; level of significance (p); dimension of the optimization problem of interest (k). Although for the sake of illustration we focus on two-dimensional test functions, the theory discussed above works for any subset of $\mathbb{R}^k$. From the practical stance, we suggest using the method discussed here as a supplement for the point estimates produced by further sophisticated methods (e.g. gradient methods, simulated annealing, evolutionary algorithms, among others).

This computational experiment intends to fill two purposes. Firstly, to provide an illustration of the confidence zones suggested by Theorem 1. Secondly, we intend to assess the degree of variation to which such interval estimates are subject, if a new computation is made, and how does this evolves when the number of sampled image values increases.

3.1 Design of the experiment

Monte Carlo simulations were considered for several (degenerated) stopping times, $r = 10.000$, 20.000, 100.000 and 500.000. Given that we run a battery of trials and Monte Carlo simulations, we use the notation $L_{i:n,j}$ to denote the i-th order statistic

from the j-th trial. The ouputs of our computational exercise are formally defined in the sequel. Consider the set-valued function $\overline{\Psi} : \mathbb{N} \times [0;1] \rightrightarrows \mathbb{R}$

$$\overline{\Psi}_n(p) = \left] n^{-1}\sum_{j=1}^{n}\left(L_{1:n,j} - \frac{L_{2:n,j} - L_{1,n,j}}{(1-p)^{-\frac{2}{k}} - 1}\right); n^{-1}\sum_{j=1}^{n} L_{1:n,j}\right[, \tag{5}$$

and define $\widehat{\sigma}^2_{LB} : [0,1] \to \mathbb{R}^+_0$, as the following function of p

$$(n-1)^{-1}\sum_{j=1}^{n}\left(\left(L_{1:n,j} - \frac{L_{2:n,j}-L_{1:n,j}}{(1-p)^{-\frac{2}{k}}-1}\right) - n^{-1}\sum_{j=1}^{n}\left(L_{1:n,j} - \frac{L_{2:n,j}-L_{1:n,j}}{(1-p)^{-\frac{2}{k}}-1}\right)\right)^2. \tag{6}$$

To measure the dispersion of the upper limit of the interval estimate, we also define

$$\hat{\sigma}^2_{UB} = (n-1)^{-1}\sum_{j=1}^{n}\left(L_{1:n,j} - n^{-1}\sum_{j=1}^{n} L_{1:n,j}\right)^2. \tag{7}$$

We considered 1.000 trials and used some well-known test functions; the selected test functions are frequently applied for assessing the performance of optimization methods (Spall, 2003; Esquível, 2007).

3.2 An account of the results

This section reports computational experience with the method presented in the foregoing section. In Table 1, we summarize useful information regarding the the search domains used, as well as their global minimums in the respective domains. Given the large number of outputs reported in Table 2, some guidance is requisite. From Table 2 we can ascertain that with the exception of the Easom and the Rastrigin functions, the interval estimate created through the application of (4) is very thin even with low computational effort. To explain the entries in Table 2, we focus on its first line. It contains the following information regarding the Beale function

$$\begin{aligned}
\overline{\Psi}_{10.000}(0.10) &=]-0,0409; 0,0048[,\\
\overline{\Psi}_{10.000}(0.05) &=]-0,0918; 0,0048[,\\
\overline{\Psi}_{10.000}(0.01) &=]-0,4985; 0,0048[.
\end{aligned}$$

These respectively correspond to the 90%, 95% and 99% interval estimates for the minimum of the Beale function build with 10.000 observations. To be more precise the approximate confidence zone are built with 10.000 observations $\times$ 1.000 trials. As it can be observed from Table 2 the values of the lower and upper bounds remain approximately constant from trial to trial. In the construction of the approximate confidence intervals there is a tradeoff between the number of trials and the number of observations to collect. Our computational experiments suggest that it is preferable to consider a larger number of observations than a larger number of replications. To state this differently, we recommend a one shot

Table 2 $(1-p)$ interval estimates for $p = 0.10/0.05/0.01$; LB and UB respectively denote the lower and upper bound of the confidence zone constructed according with (4). The values in parentheses denote sample variances defined according to (6) and (7).

10.000 observations	LB $[p = 0.10]$	LB$[p = 0.05]$	LB$[p = 0.01]$	UB
Beale	-0.0409	-0.0918	-0.4985	0.0048
	(0.0022)	(0.0096)	(0.2580)	(3e-05)
Easom	-2.3266	-4.5481	-22.3196	-0.3273
	(5.9262)	(23.5227)	(585.4046)	(0.0908)
Griewank	-0.7654	-1.8815	-10.8104	0.2391
	(0.8137)	(3.4069)	(88.7540)	(0.0142)
Rastrigin	-2.6359	-6.1850	-34.5774	0.5582
	(8.8619)	(36.4579)	(936.5940)	(0.1605)
Rosembrock	-0.5828	-1.3078	-7.1151	0.0704
	(0.4335)	(1.9305)	(52.4967)	(0.0049)
Styblinski–Tang	-79.4701	-80.8843	-92.1976	-78.1974
	(1.5938)	(7.0524)	(191.2889)	(0.0196)
20.000 observations	LB$[p = 0.10]$	LB$[p = 0.05]$	LB$[p = 0.01]$	UB
Beale	-0.0195	-0.0440	-0.2403	0.0026
	(5e-04)	(0.0023)	(0.0608)	(6e-06)
Easom	-2.8362	-5.4301	-26.1817	-0.5016
	(4.8064)	(19.4017)	(490.4135)	(0.0841)
Griewank	-0.5536	-1.3559	-7.7745	0.1685
	(0.3970)	(1.6416)	(42.3840)	(0.0080)
Rastrigin	-2.1453	-4.8858	-26.8100	0.3212
	(0.4265)	(21.9313)	(577.5952)	(0.0855)
Rosembrock	-0.2898	-0.6521	-3.5507	0.0363
	(0.1135)	(0.5022)	(13.6278)	(0.0015)
Styblinski–Tang	-78.8997	-79.6046	-85.2518	-78.2634
	(0.4113)	(1.8156)	(49.1538)	(0.0049)
100.000 observations	LB$[p = 0.10]$	LB$[p = 0.05]$	LB$[p = 0.01]$	UB
Beale	-0.0041	-0.0092	-0.0499	5e-04
	(2e-06)	(1e-04)	(0.0026)	(2e-07)
Easom	-2.0988	-3.4918	-14.6353	-0.8452
	(1.3030)	(5.6331)	(150.2689)	(0.0183)
Griewank	-0.2619	-0.6373	-3.6409	0.0760
	(0.0864)	(0.3614)	(9.4083)	(0.0015)
Rastrigin	-0.5316	-1.1991	-6.5390	0.0691
	(0.4265)	(1.8842)	(51.0218)	(0.0046)
Rosembrock	-0.0584	-0.1310	-0.7124	0.007
	(0.0043)	(0.0192)	(0.5227)	(5e-05)
Styblinski–Tang	-78.4469	-78.5898	-79.7329	-78.3183
	(0.0167)	(0.0735)	(1.9868)	(2e-04)
500.000 observations	LB$[p = 0.10]$	LB$[p = 0.05]$	LB$[p = 0.01]$	UB
Beale	-8e-04	-0.0018	-0.0097	1e-04
	(8e-07)	(4e-06)	(1e-04)	(8e-09)
Easom	-1.2893	-1.6514	-4.5486	-0.9633
	(0.1045)	(0.4600)	(12.4314)	(0.014)
Griewank	-0.1191	-0.2883	-1.6418	0.0332
	(0.0183)	(0.0769)	(2.0045)	(3e-04)
Rastrigin	-0.0118	-0.0265	-0.1440	0.0014
	(2e-04)	(7e-04)	(0.0193)	(2e-06)
Rosembrock	-0.0110	-0.0248	-0.1356	0.0015
	(2e-04)	(2e-04)	(8e-04)	(2e-06)
Styblinski–Tang	-78,3537	-78.3807	-78.5960	-78.3295
	(5e-04)	(0.0024)	(0.0643)	(8e-06)

run with a larger number of observations than averaging over several trials with a smaller number of observations. The sample variances of the lower and upper bounds of the interval estimates, respectively defined according to (6) and (7), are also reported in Table 2. As expected, the results evidence an overall tendency for approaching zero as the number of observations increases; this occurs at a different rate per test function.

Acknowledgements

I am grateful to Cláudia Neves, Vanda Inácio, António Rua, Miguel Fonseca and an anonymous referee 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.

Appendix

Proof of Lemma 2.2.
The proofs are as follows:

1. As a consequence of Lemma 2.1, it holds that
$$\frac{L_{1:n}-l^\star}{\psi(n^{-1})-l^\star}=\frac{\psi(E_{1:n})-l^\star}{\psi(n^{-1})-l^\star}\stackrel{d}{=}\frac{\psi(n^{-1}E_1^*)-l^\star}{\psi(n^{-1})-l^\star}.$$
The final result is a consequence of (2).
2. The line of attack is similar to the proof of the previous claim. As a consequence of Lemma 2.1, we have
$$\frac{L_{2:n}-l^\star}{\psi(n^{-1})-l^\star}=\frac{\psi(E_{2:n})-l^\star}{\psi(n^{-1})-l^\star}\stackrel{d}{=}\frac{\psi\left(n^{-1}\left(E_1^*+\frac{n}{n-1}E_2^*\right)\right)-l^\star}{\psi(n^{-1})-l^\star}.$$
Evoke (2) and the final result follows.
3. The ratio of interest can be conveniently rewritten as
$$\frac{L_{1:n}-l^\star}{L_{2:n}-L_{1:n}}=\frac{(L_{1:n}-l^\star)/(\psi(n^{-1})-l^\star)}{\{(L_{2:n}-l^\star)-(L_{1:n}-l^\star)\}/(\psi(n^{-1})-l^\star)}.$$
Claims 1 and 2 can now be applied to yield the final result. □

Proof of Theorem 2.3.
Start by noting that
$$\mathrm{pr}\left\{l^\star\in\Psi_n(p)\right\}=\mathrm{pr}\left\{L_{1:n}-\frac{L_{2:n}-L_{1:n}}{(1-p)^{-1/\alpha}-1}\leq l^\star\leq L_{1:n}\right\},$$
which implies that
$$\mathrm{pr}\left\{l^\star\in\Psi_n(p)\right\}=\mathrm{pr}\left\{\frac{L_{1:n}-l^\star}{L_{2:n}-L_{1:n}}\leq\frac{1}{(1-p)^{-1/\alpha}-1}\right\}.$$
As a consequence of claim 3 of Lemma 2.2, it can be shown that (Haan, 1981)
$$\mathrm{pr}\left\{\frac{L_{1:n}-l^\star}{L_{2:n}-L_{1:n}}\leq x\right\}\underset{n\to\infty}{\longrightarrow}\left(\frac{x}{1+x}\right)^\alpha,$$
implying that for n large, $\mathrm{pr}\left\{l^\star\in\Psi_n(p)\right\}$ approaches
$$\left(\frac{1}{(1-p)^{-1/\alpha}-1}\right)^\alpha\times\left(1+\frac{1}{(1-p)^{-1/\alpha}-1}\right)^{-\alpha}=1-p.$$
□

References and Notes

1 Bortot, P., Coles, S.G. and Sisson, S.A. (2007) 'Inference for steorological extremes', *Journal of the American Statistical Association*, Vol. 102, pp.84–92.

2 Bingham, N.H., Goldie, C.M. and Teugels, J.L. (1987) *Regular Variation*, Encyclopedia of Mathematics and its Applications, Cambridge University Press.

3 Booth, J., Casella, G. and Hobert, J. (2008) 'Clustering using objective functions and stochastic search', *Journal of the Royal Statistical Society - Ser. B*, Vol. 70, pp.119–139.

4 Carvalho, M. de (2010) 'A generalization of the Solis–Wets method', *Journal of Statistical Planning and Inferece*, to appear.

5 Duflo, M. (1996) *Algorithmes Stochastiques*, Springer, Mathématiques & Applications.

6 Einmahl, J.H.J. and Magnus, J.R. (2008) 'Records in athletics through extreme-value theory', *Journal of the American Statistical Association*, Vol. 103, pp.1382–1391.

7 Esquível, M.L. (2006) 'A conditional gaussian martingale algorithm for global optimization', *Lecture Notes in Computer Science*, Vol. 3982, pp.813–823.

8 Haan, L. de (1981) 'Estimation of the minimum of a function using order statistics', *Journal of the American Statistical Association*, Vol. 76, pp.467–469.

9 Haan, L. de and Ferreira, A. (2006) *Extreme Value Theory: An Introduction*, Springer, New York.

10 Nocedal, J. and Wright, S. (1999) *Numerical Optimization*, Springer-Verlag New York.

11 Pakes, A. and McGuire, P. (2001) 'Stochastic algorithms, symmetric markov perfect equilibrium and the curse of dimensionality', *Econometrica*, Vol. 69, pp.1261–1282.

12 Poon, S.-H., Rockinger, M. and Tawn, J.A. (2004) 'Extreme value dependence in financial markets: diagnostics, models, and financial implications', *The Review of Financial Studies*, Vol. 17, pp.581–610.

13 Romeijn, H.E. and Smith, R.L. (1994) 'Simulated annealing for constrained global optimization', *Journal of Global Optimization*, Vol. 5, pp.101–126.

14 Resnick, S. (2007) *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling,* Springer Series in Operations Research and Financial Engineering.

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

16 Turkman, K.F., Turkman, M.A. and Pereira, J.M. (2010) 'Asymptotic models and inference for extremes of spatio-temporal data extremes', *Extremes*, Vol. 13, pp.375–397

17 Vries, C.G. de (2005) 'The simple economics of bank fragility', *Journal of Banking and Finance*, Vol. 29, pp.803–825.

18 Wolpert, D.H. and Macready, W.G. (1997) 'No free lunch theorems for optimization', *IEEE Transactions on Evolutionary Computation*, Vol. 1, pp.67–82.