Strategy based on information entropy for optimizing stochastic functions

Tobias Christian Schmidt*
Philipps-University Marburg, Renthof 5, 35032 Marburg, Germany

Harald Ries
Philipps-University Marburg, Renthof 5, 35032 Marburg, Germany and OEC AG, Lindwurmstr. 41, 80337 Munich, Germany

Wolfgang Spirkl
Infineon Technologies A.G., Am Campeon 10, 85579 Munich, Germany

(Received 7 June 2006; revised manuscript received 6 October 2006; published 9 February 2007)

We propose a method for the global optimization of stochastic functions. During the course of the optimization, a probability distribution is built up for the location and the value of the global optimum. The concept of information entropy is used to make the optimization as efficient as possible. The entropy measures the information content of a probability distribution, and thus gives a criterion for decisions: From several possibilities we choose the one which yields the most information concerning location and value of the global maximum sought.

DOI: 10.1103/PhysRevE.75.021108 PACS number(s): 02.50.−r, 02.60.Pn, 42.15.Eq, 65.40.Gr

I. INTRODUCTION

The mathematical task of optimization is linked to thermodynamics and statistical physics in more than one way. The issue of global versus local optima is addressed by simulated annealing [1,2]. The entire optimization algorithm can be viewed as a finite time thermodynamic process in which numerical efficiency can be expressed as thermodynamical optimality [3,4]. In this contribution we use an information entropy approach to quantify the information gained in optimization.

We propose a method to optimize stochastic functions which is based on information entropy. By stochastic function we mean a function which cannot be evaluated precisely, but to which the algorithm has only indirect access, e.g., via a Monte Carlo type experiment. Thus one can only derive a probability distribution for the stochastic function, the error of which decreases with computational effort.

The stochastic function can be described by a scheme for how to get an approximation of the merit function value from the results of the random experiments and a set of pairs (a_i,b_i), where the a_i are the configurations and where every b_i is an instruction about how to conduct a random experiment.

As an example, we choose stochastic functions whose b_i are Bernoulli experiments and whose domains contain a finite number of elements. The probability of the Bernoulli experiment b_i yielding a positive result is given by the value of the stochastic merit function g_i for configuration a_i. The task is to find the maximum of the g_i with respect to value and location: $g_{\text{opt}}=g(a_{\text{opt}})$. The search for the optimum should only proceed via Bernoulli experiments.

II. INFORMATION ENTROPY STRATEGY

A. The concept

The key task of any optimization algorithm is to decide at which location to evaluate the objective function next, based on past evaluations. For a stochastic function the algorithm should additionally specify the computational effort to be invested (or alternatively the precision sought). The strategy we propose in this contribution is based on maximizing the expected information gained in each step. For this we use the term "information entropy strategy." The information entropy strategy is specified so as to optimize as efficiently as possible. What is meant by efficiency in this context? Efficiency is the ratio of gain to invested effort. We measure effort by the number of Bernoulli trials performed. The measure of gain is defined as follows. What we aspire to know are the location and the function value of the maximum. We do not seek to know the function value at other locations. Consequently we introduce the probability density function for the optimum $p_{\text{opt}}(g,i)$, which expresses the probability density that the optimum occurs with configuration i and has the value g. We refer to p_{opt} as probability distribution for the optimum. We measure the information we gain concerning the optimum of the stochastic function by the decrease in information entropy of $p_{\text{opt}}(g,i)$. Now the information entropy strategy can be outlined. Imagine the next Bernoulli trial is to be done for configuration j. Then we can calculate the expectation value of the entropy change which results from this Bernoulli trial. In order to decide the configuration for which the next Bernoulli trials should be performed, the expected entropy change following an additional trial at this configuration is calculated for all configurations. We choose that configuration with the largest expected entropy drop and perform the next Bernoulli trial there. Because this entropy drop is a measure of information gain and the number of Bernoulli trials is the measure of effort, we expect the maximum possible efficiency.

How the expected entropy changes and the probability distribution for the optimum are calculated is detailed in

*Electronic address: tobias.schmidt@physik.uni-marburg.de

Secs. II B–II E. Because the probability distribution for the optimum depends on all Bernoulli trials completed so far, the calculation of expectation values is tedious. In Secs. II B–II E we derive simplified expressions for these expectation values, which can be evaluated with moderate effort.

B. Definitions

The probability density p for the value of the objective for a given configuration i to be g, given that of n_i Bernoulli trials at that configuration k_i were successful is

$$p(g,i) = p(g,n_i,k_i) = \frac{(n_i+1)!}{k_i!(n_i-k_i)!} g^{k_i}(1-g)^{n_i-k_i}. \qquad (1)$$

The merit g is the probability that a Bernoulli trial yields a positive result. It is restricted to the interval $0 \leqslant g \leqslant 1$.

The binomial distribution is

$$P_{\text{bin}}(g,n,k) = \frac{n!}{k!(n-k)!} g^k (1-g)^{n-k}.$$

For given g and n, $P_{\text{bin}}(g,n,k)$ is the probability to get a certain value k. The normalization condition for $P_{\text{bin}}(g,n,k)$ is

$$\sum_{k=0}^{n} P_{\text{bin}}(g,n,k) = 1.$$

In our case, n and k are given. $p(g,i)$ is the probability density for g. The normalization condition is

$$\int_0^1 p(g,i)dg = 1.$$

The additional factor (n_i+1) is necessary to satisfy this condition.

The probability P_b for the value of the ith configuration to be lower than a certain value g is

$$P_b(g,i) = P_b(g,n_i,k_i) = \int_0^g p(x,i)dx. \qquad (2)$$

The index b signifies "below."

If a total of m configurations were tested then the probability $P_a(g)$ for all values to be below g is

$$P_a(g) = \prod_{i=1}^{m} P_b(g,i). \qquad (3)$$

The index a signifies "all below."

Consequently the probability distribution for the optimum $p_{\text{opt}}(g,h)$ of configuration h being the best and having an objective equal to g is

$$p_{\text{opt}}(g,h) = p(g,h)\prod_{i\neq h} P_b(g,i). \qquad (4)$$

Note that the product extends over all configurations, except configuration h.

C. Entropy and information

The total information entropy S of the probability distribution for the optimum as given in Eq. (4) is according to Shannon [5]:

$$S = -\sum_{i=1}^{i=m}\int_{g=0}^{g=1} p_{\text{opt}}(g,i)\ln[p_{\text{opt}}(g,i)]dg. \qquad (5)$$

We base the information entropy on the probability distribution for the optimum as given in Eq. (4) and *not* on the probability distribution of the value of the objective as given in Eq. (1), because we aspire to gain information about location and value of the maximum, and not about the entire function.

We choose to re-examine that configuration for which the expected information gain is largest, i.e., for which the expectation value of the entropy after performing an additional evaluation is lowest.

Calculating the total information gain in order to evaluate which configuration yields the largest gain, i.e., which i is the most "interesting" configuration is numerically demanding, in particular if many configurations have been extensively examined. The total information entropy is

$$\begin{aligned} S &= -\int_{g=0}^{g=1}\sum_{i=1}^{i=m} p_{\text{opt}}(g,i)\ln\left(P_a(g)\frac{p(g,i)}{P_b(g,i)}\right)dg \\ &= -\int_{g=0}^{g=1} \ln[P_a(g)]\sum_{i=1}^{i=m} p_{\text{opt}}(g,i)dg \\ &\quad -\int_{g=0}^{g=1}\sum_{i=1}^{i=m} p_{\text{opt}}(g,i)\ln\left(\frac{p(g,i)}{P_b(g,i)}\right)dg. \end{aligned} \qquad (6)$$

The first integral can be evaluated explicitly because the sum is a total differential,

$$\sum_{i=1}^{i=m} p_{\text{opt}}(g,i) = \frac{\mathrm{d}P_a}{\mathrm{d}g}, \qquad (7)$$

as can be seen from Eqs. (3) and (4). This reduces the first term to

$$\begin{aligned} S_I &= -\int_{g=0}^{g=1} \ln[P_a(g)]\sum_{i=1}^{i=m} p_{\text{opt}}(g,i)dg \\ &= -\int_{P_a=0}^{P_a=1} \ln(P_a)dP_a = 1. \end{aligned} \qquad (8)$$

The second term is

$$\begin{aligned} S_{II} &= -\int_{g=0}^{g=1}\sum_{i=1}^{i=m} p_{\text{opt}}(g,i)\ln\left(\frac{p(g,i)}{P_b(g,i)}\right)dg \\ &= -\sum_{i=1}^{i=m}\int_{g=0}^{g=1}\left(\prod_{j\neq i} P_b(g,j)\right)p(g,i)\ln\left(\frac{p(g,i)}{P_b(g,i)}\right)dg. \end{aligned} \qquad (9)$$

D. Expectation value of the entropy

When we decide to perform one Bernoulli trial for configuration j, we expect the system to have a certain entropy $\langle S\rangle^j$ afterwards. The entropy change depends on the outcome of the Bernoulli trial.

The expectation value $\langle S\rangle^j$ after one additional event for configuration j is

$$\langle S\rangle^j = \alpha_j S^{j+} + (1-\alpha_j) S^{j-}. \tag{10}$$

Here $\alpha_j=(k_j+1)/(n_j+2)$ is the probability of getting a positive result when re-examining the configuration j which has a record of k_j positive results out of n_j, and S^{j+} is the total entropy following a successful Bernoulli trial, where n_j and k_j would both be increased by one. Consequently $1-\alpha_j=(n_j+1-k_j)/(n_j+2)$ is the probability for a negative result and S^{j-} the entropy after a negative result if only n_j is increased by one.

E. Calculating the expected entropy change

With Eqs. (9) and (10), the expected change in the total entropy due to one additional Bernoulli trial for configuration j can be calculated. Equation (10) yields

$$\langle \Delta S\rangle^j = \alpha_j S^{j+} + (1-\alpha_j) S^{j-} - S. \tag{11}$$

Later, we will use the important fact that

$$\alpha_j \frac{P_b^{j+}(g,j)}{P_b^{(0)}(g,j)} + (1-\alpha_j)\frac{P_b^{j-}(g,j)}{P_b^{(0)}(g,j)} = 1. \tag{12}$$

This is a consequence of the fact that *a priori* the expected probability distribution after a measurement is equal to the distribution before the measurement. This is a property of all probability distributions.

Now S, S^{j+}, S^{j-} are calculated using Eq. (9) and substituted into Eq. (11):

$$\begin{aligned} S &= 1 - \sum_{i=1}^{i=m} \int_{g=0}^{g=1} \left(\prod_{j\neq i} P_b(g,j)\right) p(g,i) \ln\left(\frac{p(g,i)}{P_b(g,i)}\right) dg \\ &= 1 - \sum_{i=1}^{i=m} \int_{g=0}^{g=1} \left(\frac{P_a(g)}{P_b(g,i)}\right) p(g,i) \ln\left(\frac{p(g,i)}{P_b(g,i)}\right) dg \\ &= 1 - \int_{g=0}^{g=1} P_a(g) \sum_{i=1}^{i=m} \left(\frac{p(g,i)}{P_b(g,i)}\right) \ln\left(\frac{p(g,i)}{P_b(g,i)}\right) dg. \end{aligned} \tag{13}$$

In the following, the superscript $^{(0)}$ refers to values calculated before a new Bernoulli trial is carried out, whereas the superscript $^{j+}$ refers to a value calculated assuming a new Bernoulli trial was successful, and $^{j-}$ refers to a value assuming it was unsuccessful.

Now, how does S change when one additional measurement (Bernoulli trial) is successfully performed for configuration j? The integrand in Eq. (13) consists of a product and a sum. After a new measurement, one of the factors of the product changes and one of the summands of the sum. Thus, we can replace the initial terms by the new ones, which yields

$$S^{j+} = 1 - \int_{g=0}^{g=1} P_a^{(0)}(g) V^{j+} (A + T^{j+} - T^{(0)}) dg. \tag{14}$$

With the following abbreviations:

$$V^{j+} = \frac{P_b^{j+}(g,j)}{P_b^{(0)}(g,j)},$$

$$V^{j-} = \frac{P_b^{j-}(g,j)}{P_b^{(0)}(g,j)},$$

$$A = \sum_{i=1}^{i=m} \frac{p^{(0)}(g,i)}{P_b^{(0)}(g,i)} \ln \frac{p^{(0)}(g,i)}{P_b^{(0)}(g,i)},$$

$$T^{(0)} = \frac{p^{(0)}(g,j)}{P_b^{(0)}(g,j)} \ln \frac{p^{(0)}(g,j)}{P_b^{(0)}(g,j)},$$

$$T^{j+} = \frac{p^{j+}(g,j)}{P_b^{j+}(g,j)} \ln \frac{p^{j+}(g,j)}{P_b^{j+}(g,j)},$$

$$T^{j-} = \frac{p^{j-}(g,j)}{P_b^{j-}(g,j)} \ln \frac{p^{j-}(g,j)}{P_b^{j-}(g,j)}. \tag{15}$$

Similarly,

$$S^{j-} = 1 - \int_{g=0}^{g=1} P_a^{(0)}(g) V^{j-} (A + T^{j-} - T^{(0)}) dg. \tag{16}$$

And, of course,

$$S = 1 - \int_{g=0}^{g=1} P_a^{(0)}(g) A dg. \tag{17}$$

Equations (14), (16), and (17) are substituted into Eq. (11). Then, we use the fact that

$$[\alpha_j V^{j+} + (1-\alpha_j) V^{j-}] = 1. \tag{18}$$

See Eq. (12).

This yields

$$\begin{aligned} \langle \Delta S\rangle^j = &- \alpha_j \left(\int_{g=0}^{g=1} P_a^{(0)}(g) V^{j+} (T^{j+} - T^{(0)}) dg\right) - (1-\alpha_j) \\ &\times \left(\int_{g=0}^{g=1} P_a^{(0)}(g) V^{j-} (T^{j-} - T^{(0)}) dg\right). \end{aligned} \tag{19}$$

The term A no longer shows up in the equation. The result can further be simplified using Eq. (18):

$$\langle \Delta S\rangle^j = \int_{g=0}^{g=1} P_a^{(0)}(g) [T^{(0)} - \alpha_j V^{j+} T^{j+} - (1-\alpha_j) V^{j-} T^{j-}] dg. \tag{20}$$

This is the main equation for our strategy. Every time we want to decide the configuration for which the next Bernoulli trial should be made, we evaluate $\langle \Delta S\rangle^j$ for all configurations j and conduct the Bernoulli trial where $-\langle \Delta S\rangle^j$ is largest.

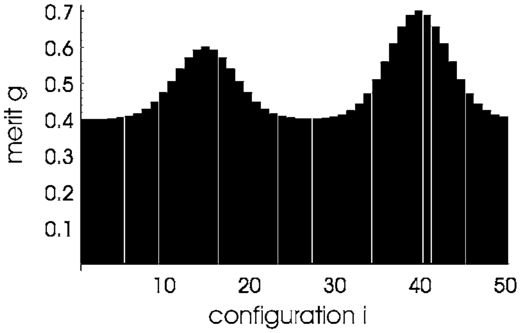

FIG. 1. The example function.

(The minus is because the smaller the entropy, the more knowledge one has.) In practice it is not necessary to calculate the $\langle\Delta S\rangle^j$ before every Bernoulli trial. Rather, we assume that the change in $\langle\Delta S\rangle^j$ is small when a small number of Bernoulli trials are made for a certain configuration. By "a small number" we mean small compared with the total number of Bernoulli trials. Hence, we proceed as follows: we calculate the $\langle\Delta S\rangle^j$, then we make a small number of Bernoulli trials for the configuration for which $-\langle\Delta S\rangle^j$ is largest, then we recalculate the $\langle\Delta S\rangle^j$, and so on.

III. TEST RESULTS

A. Applications

The application we have in mind is to choose the best from a set of virtual optical systems for illumination via Monte Carlo ray tracing. This is a standard procedure in optical design. Sending a randomly chosen ray through a virtual illumination optic is a Bernoulli trial. If the ray strikes the target surface, the outcome is "true," otherwise it is "false." Hence, every one of these illumination optic systems is an instruction on how to do a Bernoulli trial, and hence can be a b_i. If the illumination systems are a discrete subset of a parametrized set, the a_i is the parameter vector which specifies the illumination system b_i. Otherwise, one can think of the a_i simply as names of the illumination systems. By stochastic optimization we aspire to find the illumination system which directs more radiation onto the target than any of the others.

B. A naive strategy used for comparison

We use a naive and simple strategy for solving the introduced optimization problem as a benchmark for the information entropy strategy. The simple strategy carries out the same number of Bernoulli trials at all configurations. From the basic theorem of Monte Carlo integration, the necessary number of Bernoulli trials per configuration is calculated. [6]. Finding out the probability g_i of a "true" result for a Bernoulli trial by repeatedly performing Bernoulli trials is equivalent to integrating a function $f(x)$ with $x\in[0,1]$ and $f(x)=1|x\leq g_i$ and $f(x)=0|x>g_i$ with the Monte Carlo method and determining g_i from the result.

An error estimate for the integral is

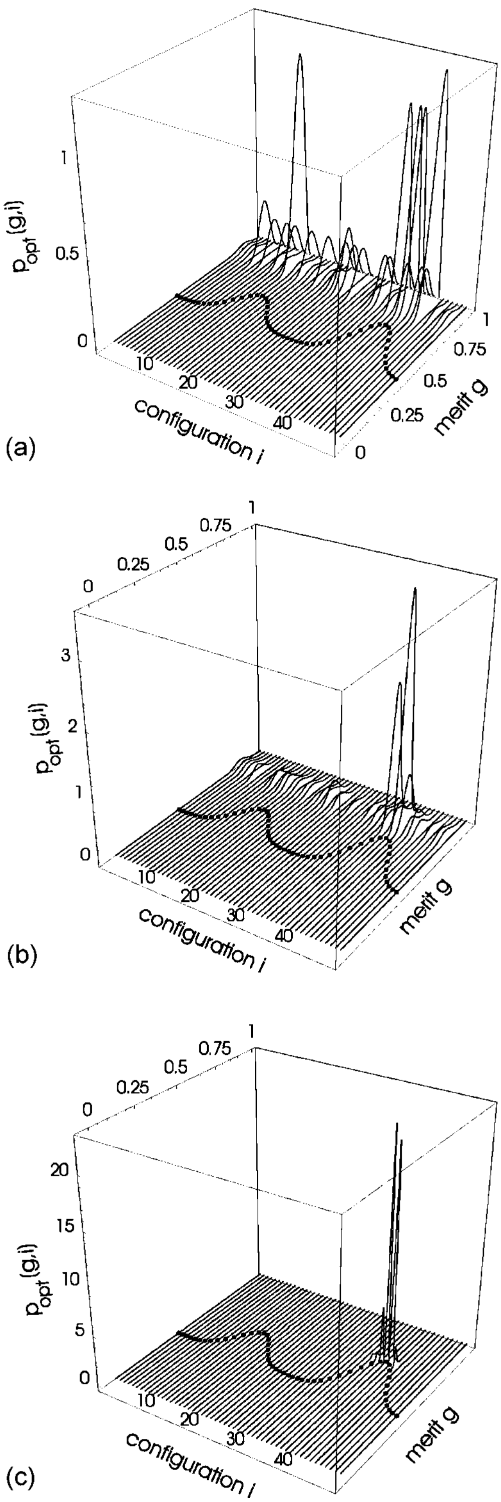

FIG. 2. Optimization with information entropy strategy. The three graphs plot the probability distribution for the optimum $p_{\text{opt}}(g,i)$ for different stages of the optimization process. In each graph, the example function is shown with dots in the horizontal plane and the probability distribution for the optimum $p_{\text{opt}}(g,i)$ is plotted in the vertical direction for each of the 50 configurations as a function of the merit function value. The plot in (a) is based on 300 Bernoulli trials, the plot in (b) on 500, and the plot in (c) on 10^4. The information entropy of the probability distribution for the optimum shown in (c) is −2.12.

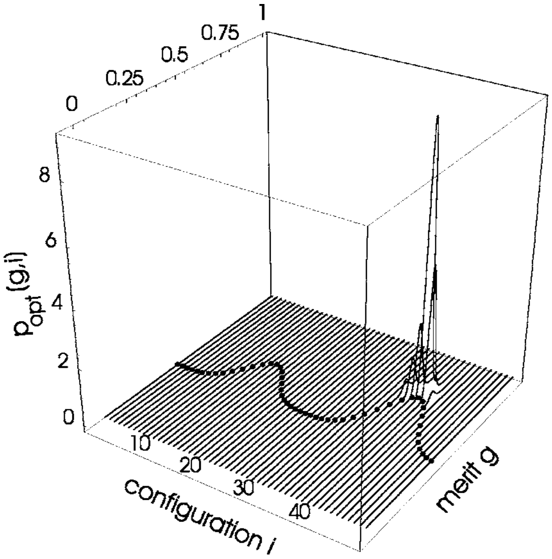

FIG. 3. This graph shows the probability distribution for the optimum calculated from 10^4 Bernoulli trials, distributed according to the naive strategy among the configurations of the example function. The probability distribution for the optimum $p_{\text{opt}}(g,i)$ is plotted for each of the 50 configurations as a function of the merit function value. In the horizontal plane the example function is shown with dots. The information entropy of the probability distribution for the optimum shown is −1.15.

$$\epsilon = V\sqrt{\frac{\langle f^2\rangle - \langle f\rangle^2}{N}}.$$

Here, N is the number of randomly chosen points and V is the volume, over which the integration extends.

The brackets indicate averaging:

$$\langle f\rangle \equiv \frac{1}{N}\sum_{i=0}^{N-1} f(x_i) \quad \text{and} \quad \langle f^2\rangle \equiv \frac{1}{N}\sum_{i=0}^{N-1} f^2(x_i).$$

For this error estimation see Ref. [6], chapter 7.

Since $x\in[0,1]$, $V=1$ and because of $f\in\{0,1\}$, $f(x_i) = f^2(x_i)$, and $\langle f^2\rangle = \langle f\rangle$, this yields

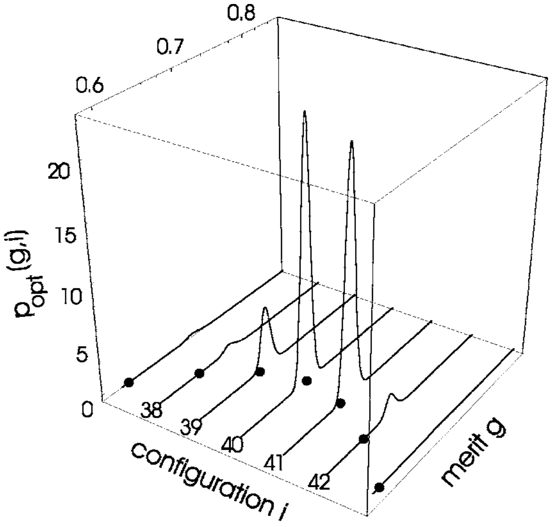

FIG. 4. This graph is a magnified section of Fig. 2(c). It shows the probability distribution for the optimum in the vicinity of its maximum for the information entropy strategy after 10^4 Bernoulli trials. The information entropy of the probability distribution for the optimum shown is −2.12.

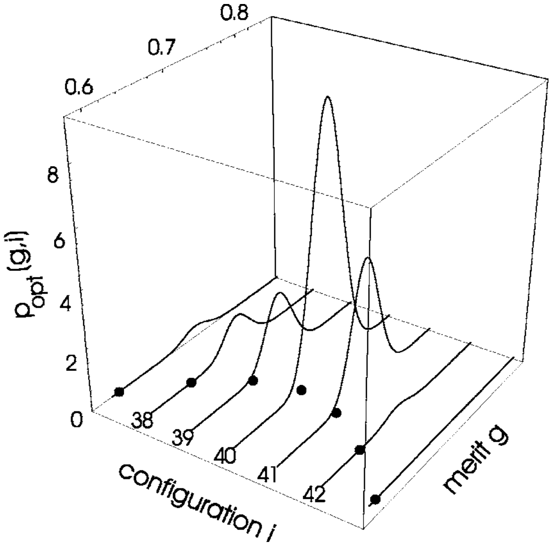

FIG. 5. This graph is a magnified section of Fig. 3. The information entropy of the probability distribution for the optimum shown is −1.15.

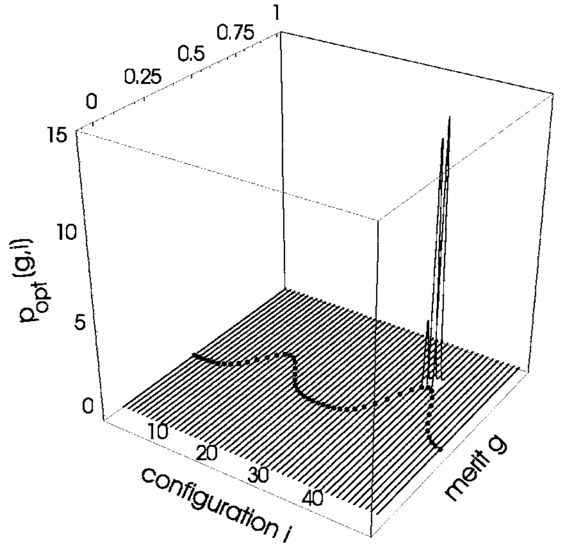

FIG. 6. This graph shows the probability distribution for the optimum calculated from 4×10^4 Bernoulli trials, distributed according to the naive strategy among the configurations of the example function. The probability distribution for the optimum $p_{\text{opt}}(g,i)$ is plotted for each of the 50 configurations as a function of the merit function value. In the horizontal plane the example function is shown with dots. The information entropy of the probability distribution for the optimum shown is −2.02.

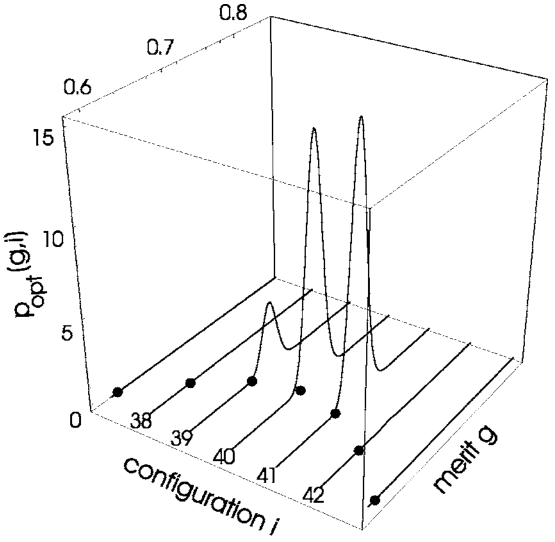

FIG. 7. This graph is a magnified section of Fig. 6. It shows the details of the peak. The information entropy of the probability distribution for the optimum shown is −2.02.

$$\epsilon = \sqrt{\langle f\rangle - \langle f\rangle^2}\sqrt{\frac{1}{N}}.$$

Because $\langle f\rangle \in [0,1]$, the maximum of $\sqrt{\langle f\rangle - \langle f\rangle^2}$ is 0.5.

If $\langle f\rangle$ is not known in advance, the error estimate is

$$\epsilon_{\max} = \frac{1}{2\sqrt{N}}.$$

To get an error smaller than $\epsilon_{\max}$, the necessary number of Bernoulli trials per configuration is

$$N = \frac{1}{4\epsilon_{\max}^2}.$$

C. Test function

We tested the information entropy strategy by applying it to an example function defined for a discrete set of 50 configurations distinguished by one parameter: $a_1 = 0.02; a_2 = 0.04; \ldots; a_{50} = 1$. The corresponding 'true' merit function values are chosen to express two peaks of different heights.

$$g_i = 0.2 \exp\left[-\left(\frac{a_i - 0.3}{0.1}\right)^2\right] + 0.3 \exp\left[-\left(\frac{a_i - 0.8}{0.1}\right)^2\right] + 0.4,$$

with $i \in \{1;\ldots;50\}$. This is called the example function, see Fig. 1. A Bernoulli trial for configuration a_i is made like this: a random number between 0 and 1 is generated. If it is smaller than g_i, the result is true, otherwise the result is false. These instructions are called b_i. A total of 10 000 Bernoulli trials were distributed among the configurations according to the information entropy strategy. We chose to make five Bernoulli trials every time the $\langle \Delta S\rangle^j$ were calculated. The graphs in Fig. 2 show the probability distribution for the optimum in different stages of the optimization process. In the beginning, one cannot see from the probability distribution for the optimum where the maximum lies, or what value it has, but after 10 000 Bernoulli trials, the location and value of the maximum are found with good precision.

D. Performance comparison

Figure 2 illustrates the evolution of the probability distribution for location and value of the optimum in the course of an optimization following the information entropy strategy. Figure 2(a) refers to the result after a total of 300 Bernoulli trials were completed, Fig. 2(b) after 500 experiments, and finally Fig. 2(c) after 10^4 Bernoulli trials were carried out. The information entropy of the probability distribution for the optimum at this point was −2.12. Note that the probability distribution for the optimum at the beginning [Fig. 2(a)] shows two peaks after which it settles at the higher peak.

We have compared the information entropy strategy to the naive strategy.

Figure 3 shows the probability distribution for the optimum calculated from 10^4 Bernoulli trials, which were distributed according to the naive strategy among the configurations of the example function. The information entropy of this probability distribution for the optimum is −1.15.

Note that after an equal number of evaluations the naive strategy correctly identifies the global maximum of the test function, however, the distribution is much broader, i.e., the

TABLE I. Results of the obtimization of the implicit example function with 50 configuarations. (a) Naive strategy, (b) information entropy strategy.

(a)			(b)		
Number of Bernouil trials in 10^3	Entropy	Computing time in minutes	Number of Bernoulli trials in 10^3	Entropy	Computing time in minutes
0	0.98	0	0	0.98	0
50	−3.586	5	25	−0.771	4
100	−3.939	11	50	−3.700	9
150	−4.120	16	75	−5.198	16
200	−4.251	22	100	−5.516	26
250	−4.382	28			

TABLE II. Results of the optimization of the implicit example function with 200 configurations. (a) Naive strategy, (b) information entropy strategy.

(a)			(b)		
Number of Bernoulli trials in 10^3	Entropy	Computing time in minutes	Number of Bernoulli trails in 10^3	Entropy	Computing time in minutes
0	0.995	0	0	0.995	0
200	−3.053	22	25	0.994	7
400	−3.930	45	50	0.993	14
600	−4.087	67	75	0.992	20
800	−4.186	89	100	0.957	28
1000	−4.321	112	125	0.880	36
			150	0.621	45
			175	−0.271	56
			200	−3.440	67
			225	−5.138	87

maximum is identified with less precision. This is illustrated in more detail in Figs. 4 and 5, which enlarge the relevant range close to the optimum of the Figs. 2(c) and 3.

For a better comparison, we allowed the naive strategy to continue until the probability distribution for the optimum roughly matched the results of Fig. 2(c). See Figs. 6 and 7. At this point the information entropy was −2.02. We found that a total of 4×10^4 Bernoulli trials were necessary. This illustrates the superiority of the information entropy strategy.

TABLE III. Results of the optimization of the implicit example function with 800 configurations. (a) Naive strategy, (b) information entropy strategy.

(a)			(b)		
Number of Bernoulli trials in 10^3	Entropy	Computing time in minutes	Number of Bernoulli trails in 10^3	Entropy	Computing time in minutes
0	0.999	0	0	0.999	0
800	−2.355	105	100	0.999	67
1600	−2.576	205	200	0.998	144
2400	−2.778	303	300	0.998	144
3200	−2.710	402	400	0.008	321
4000	−2.878	500	500	0.997	427
4800	−3.101	599	600	0.994	544
5600	−3.503	697	700	0.928	675
6400	−3.708	795	800	−2.608	816
7200	−3.807	893	900	−3.973	1092
8000	−3.944	992	1000	−4.394	1525
8800	−3.825	1091	1100	−4.724	2141
9600	−3.960	1189			
10 400	−3.972	1289			
11 200	−4.159	1389			
12 000	−4.201	1487			
12 800	−4.222	1586			
13 600	−4.124	1686			
14 400	−4.061	1787			
15 200	−3.984	1887			
16 000	−4.197	1991			

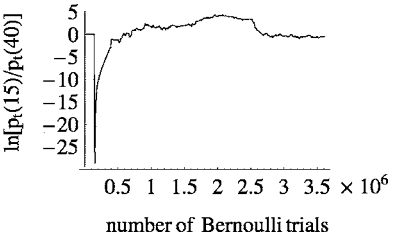

FIG. 8. Optimization of the degenerate test function with information entropy strategy. $p_t(15)$ is the total probability that the global maximum is at 0.3, whereas $p_t(40)$ is the total probability that the global maximum is at 0.8. This graph shows $\ln[p_t(15)/p_t(40)]$ as a function of the number of Bernoulli trials. The total number of Bernoulli trials is 3.61×10^6.

E. Computation time

The computation time used by the information entropy strategy is split between the time needed for carrying out the Bernoulli trials and the overhead needed to evaluate the expected entropy gain in order to decide which configuration to examine next. For this decision Eq. (20) needs to be evaluated for each configuration. Thus the time needed is roughly proportional to the number of configurations, i.e., the size of the system.

In the examples presented in Sec. III D the stochastic function used allowed a very fast evaluation of Bernoulli trials. Furthermore, the expected entropy gain was evaluated very frequently (every five Bernoulli trials). Consequently, the overhead dominated the computation time in these examples. However, this is not to be expected in practical applications, for several reasons:

(a) Additional Bernoulli trials change the expected entropy less if many trials have been previously performed. Therefore the number of Bernoulli trials carried out between consecutive evaluations of the expected entropy gain should increase in the course of the optimization, eventually rendering the computation time spent for Bernoulli trials dominant.

(b) For practical applications the computation involved in carrying out Bernoulli trials is probably more time consuming than in the simple tests used here. In particular we envision using information entropy strategy for the design of optical illumination systems, where performance is assessed via Monte Carlo ray tracing. In this field a Bernoulli trial would be equivalent to tracing a ray through an optical system which involves finding intersections at each optical surface. The duration for complex systems which may involve freeform surfaces may well be over 1 ms/ray.

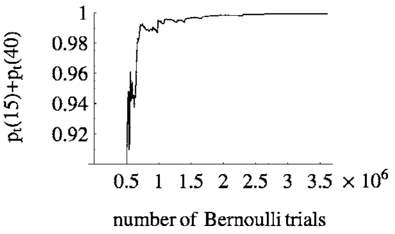

FIG. 9. Optimization of the degenerate test function with information entropy strategy. $p_t(15)$ is the total probability that the global maximum is at 0.3, whereas $p_t(40)$ is the total probability that the global maximum is at 0.8. This graph shows the sum $p_t(15)+p_t(40)$ as a function of the number of Bernoulli trials. The total number of Bernoulli trials is 3.61×10^6.

TABLE IV. Optimization of the degenerate test function with information entropy strategy. The table shows the total probability for the two maximal configurations for different numbers of Bernoulli trials.

Number of Bernoulli trials	$p_t(15)$	$p_t(40)$
1×10^6	0.830	0.165
2×10^6	0.982	0.016
3×10^6	0.439	0.560
3.61×10^6	0.389	0.610

(c) We did not code the evaluation of the expected entropy gain in the most efficient way yet. For example the term $P_a^{(0)}$ in Eq. (20) may be evaluated recursively much faster that directly via Eq. (3) as currently done. It is also possible that in the course of optimization, some configurations are recognized to be so uninteresting that they need not be considered at each evaluation. We plan to investigate these issues in future work.

In order to compare the information entropy strategy with the naive strategy in terms of computation time in a remotely realistic way with our present code, we simply used a stochastic function, which was implicitly defined via numerical root finding, such that the time needed for the Bernoulli trials was much longer. We used this implicit test function with systems of 50, 200, and 800 configurations. The results are summarized in Tables I–III. After an initial phase, during

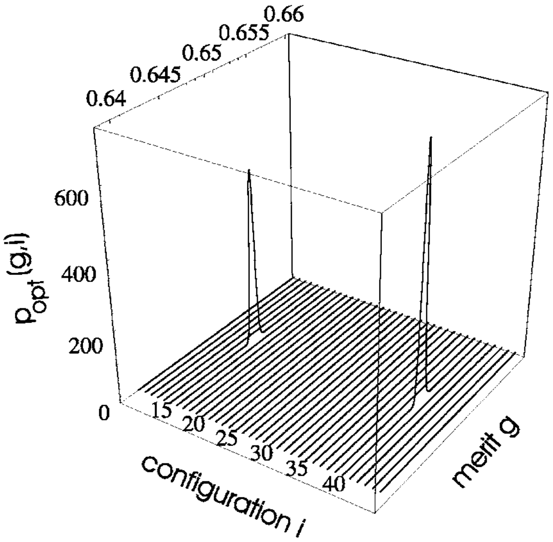

FIG. 10. Optimization with the information entropy strategy. The graph shows the probability distribution for the optimum for the degenerate test function after 3.61×10^6 Bernoulli trials. The probability distribution for the optimum is negligible outside of the section shown in the graph.

which the naive strategy is faster, the information entropy Strategy is faster in reducing the entropy of the probability distribution for the maximum. The duration of this initial phase increases with system size.

This finding is easily explained: Initially, as *a priori* all configurations are equal, the information entropy strategy coincides with the naive strategy in the choice of where to evaluate the stochastic function. Therefore the naive strategy is superior because it has no overhead. After a rough localization of the maximum, the better choice made by the information entropy strategy offsets the overhead. For larger systems a rough localization of the maximum takes longer.

F. The special case of two equally high maxima

Up to now, we were only concerned with stochastic functions which have exactly one global maximum. Since in practical applications the number of maxima is not known beforehand, it is important to know how the algorithm proceeds in the case of several equally high local maxima. Consequently, we tested the information entropy strategy on a test function with two equally high maxima, which we call the degenerate test function.

The degenerate test function is

$$g_i = 0.25\exp\left[-\left(\frac{a_i-0.3}{0.1}\right)^2\right] + 0.25\exp\left[-\left(\frac{a_i-0.8}{0.1}\right)^2\right] + 0.4, \tag{21}$$

for the parameter values $a_1=0.02; a_2=0.04; \ldots ; a_{50}=1$.

We want to ensure that the information entropy strategy identifies both maxima, and not only one of them. For each of the maxima we integrate the probability distribution for the optimum $p_{\text{opt}}(g,i)$ over the merit function value g, thus getting the total probability $p_t(i)$ that this configuration is better than all others. $p_t(i)$ is a function of the number of Bernoulli trials, since the probability distribution for the optimum is a function of the number of Bernoulli trials. Both of the maxima are found if $p_t(i)$ is of the same order of magnitude for both of the maxima and small for all the other configurations. The result is shown in Figs. 8 and 9 and in Table IV.

From Figs. 8 and 9 two things can be learned. First, one can see that for sufficiently large numbers of Bernoulli trials $p_t(15)$ and $p_t(40)$ are of the same order of magnitude and all other $p_t(i)$ are small compared to $p_t(15)$ and $p_t(40)$, since the sum of all $p_t(i)$ is one. That means that both maxima were found by the information entropy strategy (see Fig. 10). Second, the number of Bernoulli trials should not be too small. If the calculation had been stopped after 2×10^6 Bernoulli trials, the maximum a_{40} perhaps had been overlooked (see Table IV).

IV. CONCLUSIONS

Information entropy appears to be a useful criterion for the optimization of stochastic functions. However, it is important in the context of optimization to base the information entropy on the probability distribution for the optimum rather then the probability distribution of the stochastic function itself. The strategy presented in this work appears to hold good promise as a key ingredient for a global optimization algorithm for stochastic functions. In future work we plan to extend and test this concept for continuous parameter spaces of higher dimensions.

ACKNOWLEDGMENTS

We wish to thank Christopher Essex from the University of Western Ontario, Canada, Karl Heinz Hoffmann from the Technical University Chemnitz, and Bjarne Andresen from the University of Copenhagen for helpful discussions on this subject.

[1] G. Ruppeiner *et al.*, J. Phys. I **1**, 455 (1991).
[2] B. Andresen and J. M. Gordon, Phys. Rev. E **50**, 4346 (1994).
[3] W. Spirkl and H. Ries, Phys. Rev. E **52**, 3485 (1995).
[4] G. Ruppeiner, Rev. Mod. Phys. **67**, 605 LP (1995).
[5] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948); **27**, 623 (1948).
[6] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery, *Numerical Recipes in C++* (Cambridge University Press, New York, 2002), 2nd ed.