Introducing Nested Sampling

Noel van Erp & Pieter van Gelder
Structural Hydraulic Engineering and Probabilistic Design, TU Delft
Delft, The Netherlands

1 Introduction

Integration of k-variate functions is needed for assessing failure probabilities of systems where multiple load – and resistance variables are present. By reducing any k-variate function \( f \) to a corresponding monotonic descending univariate function \( g \), and by using order statistics, the integral of any k-variate function \( f \) may be evaluated using a Monte Carlo sampling scheme called Nested Sampling [1]. This paper is organised as follows. First an example of a 2-variate function will be presented to introduce the problem, followed by a presentation of the general philosophy of nest sampling with a stepwise algorithm. Technical details concerning implementation issues are treated in the appendices of this paper.

2 Case study of integrating a bivariate function \( f \) with nested sampling

Say we wish to evaluate a bivariate distribution \( f(x, y) \) numerically, where

\[
f(x, y) = \frac{\sqrt{1 - (0.7)^2}}{2\pi} \exp\left[-\frac{1}{2} \left(x^2 + 1.4xy + y^2\right)\right]
\]

(1)

where \(-5 \leq x \leq 5, -5 < y < 5\).
The total volume under the curve \( f(x, y) \) is given by the integral

\[
\int_{-5}^{5} \int_{-5}^{5} \sqrt{1 - (0.7)^2} \exp \left[ -\frac{1}{2} \left( x^2 + 1.4xy + y^2 \right) \right] \, dx \, dy = 0.9993. \tag{2}
\]

We may evaluate the integral (2) through brute force. We partition the \( x, y \)-plane in little squares with area \( dx_j \, dy_k \), \( j = 1, \ldots, 20 \), \( k = 1, \ldots, 20 \), then define the center of these areas as \( (\tilde{x}_j, \tilde{y}_k) \), and compute the strips of volume \( V_{jk} \) as

\[
V_{jk} = f(\tilde{x}_j, \tilde{y}_k) \, dx_j \, dy_k. \tag{3}
\]

In Figure 2 we give all the volume elements \( V_{jk} \) together:
Now, we may map these 3-dimensional volume elements $V_{jk}$ to corresponding 2-dimensional area elements $A_i$. This is easily done by introducing the following notation:

$$dw_i = dx_j dy_k, \quad f([\tilde{x}, \tilde{y}]) = f([\tilde{x}_j, \tilde{y}_k]),$$

where index $i$ is a function of the indices $j$ and $k$:

$$i \equiv (j - 1)20 + k$$

and $i = 1, \ldots, 400$. Using (5), we may rewrite (3) as

$$A_i = f([\tilde{x}, \tilde{y}]) dw_i.$$  \hfill (7)

In Figure 3 we give all the 400 are elements $A_i$ together:

\begin{figure}
\centering
\includegraphics[width=\textwidth]{area_elements.png}
\caption{Area elements of function $f$}
\end{figure}

Since (7) is equivalent to (3), we have that the mapping of the 3-dimensional volume elements $V_{jk}$ to their corresponding 2-dimensional area elements $A_i$ has not led to any loss of information; that is,

$$\text{area} = \sum_{i=1}^{400} A_i = \sum_{j=1}^{20} \sum_{k=1}^{20} V_{jk} = \text{volume}.$$  \hfill (9)

We now may, trivially, rearrange the elements $A_i$ in Figure 2 in descending order:
Figure 4: Ordered area elements of function $f$

Note that the horizontal axis of Figure 4 is non-dimensional. This is because we are looking at an collection of rectangular area elements ordered in one of many possible configurations. Now all these rectangular elements have a base of $dw = dx \, dy = 0.25$, being that there are 400 area elements we might view Figure 4 as a representation of some monotonic descending function $g(w)$, where $0 \leq w \leq 100$.

Figure 5: Plot of function $g$

What we now have accomplished is that we have mapped 3-dimensional volume elements (Figure 2) to 2-dimensional area elements (Figure 3) after which we have rearranged the
area elements (Figure 4) so as to get a monotonic descending ‘function’ \( g(w) \) (Figure 5). We now may integrate the univariate function \( g(w) \) and, again, get the volume we are looking for. Note that in this manner we may reduce any \( k \)-variate function to a corresponding monotonic descending univariate function \( g(w) \). We will see that the procedure of Nested Sampling is based upon the equivalence between any \( k \)-variate function and its corresponding \( g(w) \).

### 3 Sampling Abcissa’s

Now say, we have a value of \( g(w) \), without knowing the value of \( w \). Then the only thing we know about \( w \) is that it must lie somewhere in the region \( 0 \leq w \leq W \), where \( W \) is the area for which the \( k \)-variate function is defined (see for example, the area for which (1) is defined and the \( w \)-axis of Figure 5). So, we have that \( w \) is univariately uniformly distributed, \( w \sim U(0,W) \), with mean and standard deviation of

\[
E(w) = \frac{W}{2}, \quad \text{std}(w) = \frac{W}{2\sqrt{3}}. \tag{10}
\]

Now, suppose that we sample \( N \) values of \( g(w) \), that is, we have sampled \( g(w_1), \ldots, g(w_N) \), and though we still do not know the values of \( w_1, \ldots, w_N \), the one thing we now do know is that the smallest realisation of \( g(w) \) must correspond with the largest value of \( w \). This is because function \( g(w) \) is a monotonic descending function. It follows that we may use an order distribution for the unknown value \( w_{\text{max}} \):

\[
p(w_{\text{max}}) = N\left(\frac{w}{W}\right)^{N-1} \frac{1}{W} \tag{11}
\]

with mean and standard deviation of

\[
E(w_{\text{max}}) = W - \frac{1}{N+1}W, \quad \text{std}(w_{\text{max}}) = W\sqrt{\frac{N}{(N+1)^2(N+2)}}. \tag{12}
\]

and where both the values of \( N \) and \( W \) are known to us. We have that the standard deviation, that is, our uncertainty regarding the unknown value of \( w_{\text{max}} \), falls of with a factor \( N \). We will see that (12) forms the backbone of the Nested Sampling algorithm.

### 4 The Basic Nested Sampling Algorithm

In this first version of the Nested Sampling algorithm we will not protect against under- and overflow. We will just focus here on the basic philosophy which underlies Nested Sampling.
Step 1

Find \( N \) random values \( f(x, y) \) in the \( x, y \)-plane. Since we may perform the steps as shown in Figure 1 through 5, it holds trivially that these \( N \) values of \( f(x, y) \) must correspond with \( N \) values of \( g(w) \). In the absence of an explicit sorting of the volume/area elements we cannot map the \((x, y)\)-coordinates to the corresponding \( w \)-coordinate explicitly. But the thing we can do is use (12) to statistically approximate this corresponding \( w \)-coordinate for the smallest sampled value of \( f(x, y) \) and thus get our first coordinate \((w_1, g_1)\) of the unknown function \( g(w) \), where

\[
w_1 = W - \frac{1}{N+1} W, \quad g_1 = f(x^{(l)}, y^{(l)})_{\text{min}}, \tag{Alg.1}
\]

where the error of our estimated \( w_1 \) will fall of with a factor \( N \), as can be seen in (12). We now approximate the integral right of \( w_1 = W - W/(N+1) \) as

\[
A_1 = \int_{w_1}^{w} g(w)dw \approx \frac{1}{N+1} W \cdot g_1, \tag{Alg.2}
\]

and as a first approximation to the total volume:

\[
Z_1 = A_1. \tag{Alg.3}
\]

Step 2

We again find \( N \) random values \( f(x, y) \) in the \( x, y \)-plane, but now we constrain these random values to be equal or greater than the value of the minimum of the last iterate, that is, we sample \( f(x, y) \) under the constraint

\[
f(x, y) \geq g_1 \tag{Alg.4}
\]

Now since it holds trivially that the found \( N \) values of \( f(x, y) \) must correspond with \( N \) values of \( g(w) \), we may rewrite (Alg.2) as

\[
g(w) \geq g_1. \tag{Alg.5}
\]

Now, since \( g(w) \) is a monotonic descending function and since \( w_1 \) is the coordinate that corresponds with the lowerbound \( g_1 \), we may set \( 0 \leq w \leq w_1 \). Using (12) again, but replacing \( W \) with \( w_1 \), the second coordinate \((w_2, g_2)\) of the unknown function \( g(w) \), where

\[
w_2 = w_1 - \frac{1}{N+1} w_1, \quad g_2 = f(x^{(2)}, y^{(2)})_{\text{min}}, \tag{Alg.6}
\]

We now approximate the integral right of \( w_2 = w_1 - w_1/(N+1) \) as
\[ A_2 = \int_{w_2}^{w_1} g(w) dw = \frac{1}{N+1} w_1 \cdot g_2, \]  
(Alg.7)

and as a second approximation to the total volume:
\[ Z_2 = A_1 + A_2 \]  
(Alg.8)

**Step \( t \)**

For the \( t \)th iterate we find
\[ w_t = w_{t-1} - \frac{1}{N+1} w_{t-1}, \quad g_t = [f(x^{(i)}, y^{(j)})]_{\text{lin}}, \]  
(Alg.9)

and
\[ A_t = \int_{w_t}^{w_{t-1}} g(w) dw = \frac{1}{N+1} w_{t-1} \cdot g_t, \]  
(Alg.10)

and the estimate of the total volume in the \( t \)th iteration becomes:
\[ Z_t = \sum_{i=1}^{t} A_i \]  
(Alg.11)

**Termination Step.**

We have that \( \lim_{t \to \infty} w_t = 0 \), because of the identity:
\[ w_t = \left(1 - \frac{1}{N+1}\right)^t W, \]  
(Alg.12)

and at the same time we have that \( g_t \) cannot be larger than the maximum value of \( f(x, y) \). It then follows that \( \lim_{t \to \infty} A_t = 0 \). We will take as a stopping criterion the point where \( A_t \) does not contribute more than \( 1/N^2 \) part to \( Z_t \), that is,
\[ \frac{A_t}{Z_t} > \frac{1}{N^2}. \]  
(Alg.13)

However, there is of yet no rigorous criterion to ensure the validity of the above terminating condition. Termination remains a matter of user judgement about the problem in hand [1].

Since we are sampling our abcissas some sampling error will occur in our estimate of \( Z_t \), even though this error falls of a factor \( N \). We may record the end points of the likelihood intervals of the areas \( A_t \):
\[ A_t^{\text{left}} = \left[ \frac{1}{N+1} - \sqrt{\frac{N}{(N+1)^2(N+2)}} \right] w_{t-1} \cdot g_t, \]  
(Alg.14a)
in order to compute the end points of the likelihood intervals of total volume \( Z_t \):

\[
Z_{t \text{Left}} = \sum_{i=1}^{t} A_{i \text{Left}}, \quad Z_{t \text{Right}} = \sum_{i=1}^{t} A_{i \text{Right}}. \quad (\text{Alg. 15})
\]

This finishes our exposition of the basic idea behind Nested Sampling.

Now if one is to implement the above algorithm one would get an algorithm very differently form the one given by Skilling (2006). These differences are mainly differences of implementation, where the implementations of Skilling are the most efficient ones.

The reason for us to give here the naive algorithm was to point to reader to the elegantly simple idea behind Nested Sampling, without losing ourselves too much in the technicalities of optimal implementations. However, with the core idea behind Nested Sampling demonstrated, we refer the reader to appendices A, B, and C, wherein these points of optimal implementation are treated.

5 Discussion

By reducing any \( k \)-variate function \( f \) to a corresponding monotonic descending univariate function \( g \), and by using order statistics, the integral of the \( k \)-variate function \( f \) may be evaluated using Monte Carlo sampling. The procedure by which this done is called Nested Sampling. A special application of Nested Sampling is the case in which the \( k \)-variate function \( f \) is the product of a prior distribution and a likelihood function. The resulting evaluated integral then equals the evidence value of the corresponding posterior distribution. For the pseudo code of Nested Sampling we refer the interested reader to Skilling (2006).

6 Reference

Appendix A: Reduction of Computational Cost

In our algorithm we draw at each iteration $N$ new samples from $f(x,y)$ under the constraint

$$f(x^{(t+1)}, y^{(t+1)}) \geq g_t,$$

where

$$g_t \equiv \left[f(x^{(t)}, y^{(t)})\right]_{\min}.$$  \hfill (A.1)

Now, we may do this more efficiently by realizing that in iteration $t-1$ we already had $N-1$ objects that satisfied the above constraint. If we keep these $N-1$ objects, then we only need to sample one additional object in order to obtain our needed sample of $N$ objects.

In the words of Skilling (2006): ‘After each iteration $t$ we discard one object from our sample of $N$ objects. This discarded object becomes the lower bound $g_t$. The $N-1$ surviving objects are taken to the next iteration and an additional object is sampled under the constraint $f(x^{(t+1)}, y^{(t+1)}) \geq g_t$. This implementation reduces the computational costs with an order of magnitude of $N$.’

Appendix B: Guarding Against Under- and Overflow

In many problems $Z_t$ may become so large that computational overflow may occur, that is, that $Z_t$ becomes so large that there is no longer a floating number representation possible for its value. To remedy this problem we will have to work with the $\log Z_t$. We also have that certain functions $f(x,y)$ have values so large that computational overflow may also occur. So, in what follows we will sample $\log[f(x,y)]$ instead of $f(x,y)$. Likewise we have that, because of the fact that $\lim_{t \to \infty} w_t = 0$, for sufficiently large $t$ computational underflow may occur, that is, that $w_t$ becomes so small that there is no longer a floating number representation possible for its value. To remedy the latter situation we will go from the $w$ scale to the $\log w$ scale.

To go to the $u = \log w$ scale we will have to make a proper change of variable for the order distribution (11) in order to still be able to sample abcissa values. We have that $du = dw/w$ and $w = \exp u$, so (11) transforms to

$$p(u_{\max}) = \frac{N}{W^N} \exp(Nu), \quad -\infty < u \leq \log W,$$ \hfill (B1)

with mean and standard deviation

$$E(u_{\max}) = -\frac{1}{N} \log W, \quad \text{std}(u_{\max}) = \frac{1}{N}.$$ \hfill (B2)
With a repeated application of (B2) we may find the limits of the $u$ scale after the $t$th iteration to be

$$u_t = -\frac{t}{N} + \log W$$  \hspace{1cm} (B3)

It follows that the width of the $t$th interval on the original $w$ scale, $dw_t$, may be written as

$$dw_t = \exp(u_{t-1}) - \exp(u_t)$$

$$= \exp\left(-\frac{t-1}{N} + \log W\right) - \exp\left(-\frac{t}{N} + \log W\right)$$

$$= W\left[\exp\left(-\frac{t-1}{N}\right) - \exp\left(-\frac{t}{N}\right)\right]$$

$$= W\exp\left(-\frac{t-1}{N}\right)\left[1 - \exp\left(-\frac{1}{N}\right)\right]$$  \hspace{1cm} (B4)

It follows that

$$\log(dw_t) = -\frac{t-1}{N} + \log W + \log\left[1 - \exp\left(-\frac{1}{N}\right)\right]$$  \hspace{1cm} (B5)

where only the first part is dependent upon the number of iterations. We have that the log of the area element at the $t$th iteration, $\log A_t$, is the sum

$$\log A_t = \log(dw_t) + \log\left[f(x^{(t)}, y^{(t)})\right]_{\min}$$

$$= -\frac{t-1}{N} + \log W + \log\left[1 - \exp\left(-\frac{1}{N}\right)\right] + \log(g_t)$$  \hspace{1cm} (B6)

Note that (B6) is now sufficiently protected from under- and overflow. We may update $\log Z_{t-1}$ to $\log Z_t$ with $\log A_t$ using the formula for logarithmic addition [1]

$$\log[\exp(x) + \exp(y)] = \log\left\{\exp(x)\left[1 + \exp(y-x)\right]\right\}$$

$$= x + \log\left[1 + \exp(y-x)\right]$$

$$= x + \log\left[1 + \exp(g_t)\right]$$

where we set $x \equiv \log Z_{t-1}$ and $y \equiv A_t$, so as to get

$$\log Z_t = \log Z_{t-1} + \log\left[1 + \exp(\log A_t - \log Z_t)\right]$$  \hspace{1cm} (B8)

To summarize, in order to protect the algorithm from under- and overflow, which may easily occur in actual problems, the original algorithm remains unchanged, except that each iteration (B6) is used, instead of (Alg.2), (Alg.7), and (Alg.10), to compute $\log A_t$, while
$Z_t$ is updated using (B8), instead of (Alg.3), (Alg.8), and (Alg.11). Note that the termination condition (Alg.13), may be rewritten as

$$\log \mathcal{A} - \log Z_t > -2 \log N$$

(B9)

Though we remind the reader that of yet there is no rigorous criterion to ensure the validity of the above terminating condition. Termination remains a matter of user judgement about the problem in hand [1].

Appendix C: Restricting Nested Sampling to Posterior Distributions

The log area (18) elements are also used by Skilling in his algorithm. However, Skilling sets $W = 1$, because he is concerning himself only with the evaluation of posterior probability distributions; that is,

$$f(x^{(i)}, y^{(i)}) = p(x^{(i)}, y^{(i)}|D)$$

(C1)

$$= p(x^{(i)}, y^{(i)})L(x^{(i)}, y^{(i)})$$

where $p(x^{(i)}, y^{(i)})$ and $L(x^{(i)}, y^{(i)})$ are the prior and likelihood, respectively. We will now explain.

Say that the area spanned by the likelihood is $W$, then we know that the uninformative uniform prior $p(x^{(i)}, y^{(i)})$ is equal to $W^{-1}$, so

$$f(x^{(i)}, y^{(i)}) = W^{-1}L(x^{(i)}, y^{(i)}).$$

(C2)

It follows that

$$\log \left\{ f(x^{(i)}, y^{(i)}) \right\}_{\text{min}} = -\log W + \log \left\{ L(x^{(i)}, y^{(i)}) \right\}_{\text{min}}$$

(C3)

Substituting (C3) into (B6) we get

$$\log \mathcal{A}_t = -\frac{t-1}{N} + \log W + \log \left[ 1 - \exp \left( -\frac{1}{N} \right) \right] - \log W + \log (L_t)$$

(C4)

$$= -\frac{t-1}{N} + \log \left[ 1 - \exp \left( -\frac{1}{N} \right) \right] + \log (L_t)$$

where $L_t$ is the minimal sampled likelihood at the $t$th iteration:

$$L_t \equiv \left[ L(x^{(i)}, y^{(i)}) \right]_{\text{min}}.$$

(C5)

Note that $\log W$ has disappeared, or, equivalently, we may set $W = 1$, in the case we are evaluating posterior distributions with uniform priors over the region $W$. \[\]
Appendix D: Generating Monte Carlo Samples

In the case that \( f(x, y) \) is a posterior probability distribution, then we may also produce Monte Carlo samples. Every minimum realization \( [f(x^{(i)}, y^{(i)})]_{\min} \) corresponds with an area \( A_i \) as well with an sample \( (x^{(i)}, y^{(i)}) \). Now, if we record every \( (x^{(i)}, y^{(i)}) \) together with the area \( A_i \) , then we may assign a weight, that is, probability, to this sample

\[
p_i = \frac{A_i}{Z_T},
\]

where \( Z_T \) is the final estimated evidence at the last iteration \( T \). We then may estimate the expectation value of the function \( Q(x, y) \) as

\[
E(Q) = \sum_{i=1}^{T} p_i \cdot Q(x^{(i)}, y^{(i)}),
\]

and the a variation of

\[
\text{var}(Q) = \sum_{i=1}^{T} p_i \cdot [Q(x^{(i)}, y^{(i)}) - E(Q)]^2.
\]

Appendix E: Why \( N^{th} \) Order Statistics Are Used

Now we may ask ourselves why we take the \( N^{th} \) order statistic instead of, say, the 1st order statistic. The reason for this is that, as can be seen in Figure 6 where we have constructed the function \( g(w) \) by taking in (1) the limits to be \(-50 \leq x \leq 50 \) and \(-50 \leq y \leq 50 \).

Figure 6: Plot of function \( g \) for broader limits of function \( f \)

Now we must realise that for limits \(-50 \leq x \leq 50 \) and \(-50 \leq y \leq 50 \) the graph for \( 0 \leq W \leq 100 \) in Figure 6 is roughly the same the graph for \( 0 \leq W \leq 100 \) in Figure 5. The reason being that \( f(x, y) \) for \(-5 > x > 5 \) and \(-5 > y > 5 \) will generally be (much) smaller
than $f(x,y)$ for $-5 \leq x \leq 5$ and $-5 \leq y \leq 5$, combined with the fact that $g(w)$ is constructed by ordering area elements in descending order. We also have that $W = 10000$ which for $N = 100$ leaves us with a mean of about

$$E(w_{\min}) = \frac{1}{N+1} W \approx 100,$$

(E1)

and a standard deviation of about

$$std(w_{\min}) = W \sqrt{\frac{N}{(N+1)^2(N+2)}} \approx 100$$

(E2)

However this means that we leave the informative part $g(w)$, that is, its left-hand side in just one iteration step. However, if we approach the informative part from the right, using the $N$th order statistic, we have that our intervals fall off with a factor $[N/(N+1)]^t$, because of the identity

$$dw^{(t)} = \frac{w_t}{N+1}$$

(E3)

where the limit in the $t$th iteration $w_t$ is given as:

$$w_t = \left(1 - \frac{1}{N+1}\right)^t W.$$  

(E4)

So when the informative part of function $g(w)$ is reached, the intervals will in general have shrunk sufficiently to capture all the information in the compressed left-hand part of $g(w)$. 