*Proceedings of the 2010 Winter Simulation Conference*
*B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds.*

# RANDOM SEARCH IN HIGH DIMENSIONAL STOCHASTIC OPTIMIZATION

Russell Cheng

University of Southampton
Highfield
Southampton, SO17 1BJ. UNITED KINGDOM

## ABSTRACT

We consider the use of random search for high dimensional optimization problems where the objective function to be optimized can only be computed with error . Random search is easy to carry out, but extraction of information concerning the objective function is not so straightforward. We propose fitting a statistical model to the objective function values obtained in such a search, and show how the fitted model can be used to estimate the best value obtained when the search effort is limited and how this value compares with the unknown true optimum value. A possible use of this approach is in combinatorial optimization problems. The dimension in such a problem is not usually considered, but if a dimension can be associated with it, then it is likely to be high. We illustrate our method with a numerical example involving a travelling salesman problem.

## 1 INTRODUCTION

This paper considers the use of random search optimization (RSO) in simulation to minimize, an objective function $X(\theta)$, typically an expected system performance measure, that is a continuous function of a vector $\theta$ of $d$ continuous decision variables, where $\theta$ can be selected from a compact region $\Theta$ of $R^d$, and where the minimum expected performance

$$x_{\min} = \min_{\theta \in \Theta} X(\theta) \tag{1}$$

is obtained at an interior point $\theta_{\min}$ of $\Theta$. We conduct the random search in the following way. We first sample $m$ mutually independent values of $\theta$:

$$\theta_1, \theta_2, ..., \theta_m, \tag{2}$$

which we shall call *search points*, from some continuous distribution with density

$$g(\theta),\ \theta \in \Theta. \tag{3}$$

We allow for a general density rather than sampling from a uniform distribution to enable sampling to be focussed in the most promising regions of $\Theta$ based on prior information.

Then for each $\theta_i$ we make $n$ independent simulation runs, each of some predetermined and fixed standard length $t$. We shall not discuss how individual runs are conducted. For example, if a warm up period

is needed in each run, we assume that this has already been considered and dealt with. If the total time available allows a maximum of *c* simulation runs, we have that

$$c = nm. \tag{4}$$

Writing $X_i$ for $X(\theta_i)$, the *observed* performance indices are:

$$Y_{ij} = X_i + \varepsilon_{ij}, \qquad i = 1,2,...,m; j = 1,2,...,n. \tag{5}$$

where $\varepsilon_{ij}$ are random 'error' quantities with mean zero and variance $\sigma^2$. We consider only the case where the $\varepsilon_{ij}$ are mutually independent and where $\sigma^2$ is the same at all individual observations.

The averages of the observations at each $\theta_i$ are

$$W_i = n^{-1}\sum_{j=1}^{n} Y_{ij} = X_i + n^{-1}\sum_{j=1}^{n} \varepsilon_{ij} = X_i + \zeta_i, \text{ say, for } i = 1,2,...,m, \tag{6}$$

where the averaged errors $\zeta_i$ have mean zero and variance:

$$\mathrm{Var}[\zeta_i] = \sigma^2 / n = \sigma_n^2, \text{ say.} \tag{7}$$

We include the suffix *n* as a reminder that $\mathrm{Var}[\zeta_i]$ depends on *n*. The important point to note is that because the $\theta_i$ are randomly sampled, the $X_i$ are *also* random variables which are independent of the $\zeta_i$. There are thus two sources of variation: the *search induced variability* of the $X_i$ and the *simulation induced variability* of the $\zeta_i$. In all that follows we shall use the notation $X(\theta)$ when we are regarding *X* as a deterministic function of $\theta$, and the notation $X(\theta_i)$, or $X_i$, when regarding *X* as a random variable resulting from the random sampling of the $\theta_i$ in (2). We shall write $F_X(\cdot)$, and $F_\zeta(\cdot)$ to denote the cumulative distribution functions (CDF) of $X_i$ and $\zeta$ and $f_X(\cdot)$ and $f_\zeta(\cdot)$ for their probability density functions (PDF).

To date published theoretical work has focussed on the situation where *d* is known and where the minimum observed value of the $W_i$ is used to estimate $x_{\min}$. We shall write this smallest value as $W_{(m,n)}$ to indicate that its distribution depends on both *m* and *n*. Chia (2005) and Chia and Glynn (2007) have studied the behavior of $W_{(m,n)}$ under the classical assumption that $X(\theta)$ is a quadratic function of $\theta$ near $\theta_{\min}$. The main result is that $W_{(m,n)}$ has minimum variance when

$$m \sim rc^{d/(d+4)}, \; n \sim r^{-1}c^{4/(d+4)} \tag{8}$$

as $c \to \infty$, with *r* an arbitrary but fixed positive constant.

Essentially the same problem has also been considered by Yakowitz et al. (2000) who however employ a search on a low dispersion set of points, but yielding a result very similar to that of (8).

We shall also consider this problem under the same quadratic assumption for $X(\theta)$, but for the case where the dimension *d* is to be regarded as being large, and possibly not even precisely known.

Our approach is to fit a statistical model to the observations (5). To completely specify a statistical model for the observations (5) we note that $Y_{ij}$ is simply the sum of $X_i$ and $\varepsilon_{ij}$. Thus all we need to do is specify $F_X(\cdot)$ the distribution of $X$ and $F_\varepsilon(\cdot)$ the distribution of $\varepsilon$. The distribution of $Y$ is then the convolution of these two distributions. For the situation where $d$ is large we show that we do not need to consider any definite value for $d$ but can assume that a central limit theorem applies to the distribution of $X$ so that we can assume it to be normally distributed. Moreover we shall only consider the situation where the errors are normally distributed. The distribution of $Y$ is then also normally distributed.

One of the advantages of using a statistical model is that it allows one to study quantities of interest other than $W_{(m,n)}$. We shall therefore study not simply the distribution of $W_{(m,n)}$, but estimation of the unknown optimum value $x_{\min}$, and also the distribution of the value of the performance measure actually obtained when the search point $\theta_i$ corresponding to $W_{(m,n)}$ is selected. This last is perhaps the quantity of most interest in a random search. We also consider the calculation of confidence intervals (CI) for these quantities using resampling. We also suggest a modified Anderson-Darling goodness of fit test, again using bootstrap resampling, to test the adequacy of the fitted statistical model.

In our numerical example the precise form of the distribution of $X_i$ is completely know so that the optimum solution is known. We are therefore able to make definitive comparisons of the results of the RSO with this solution.

In the next section we discuss in more detail our normal model of the observations (5). The practical approach of fitting this model to data is described in Section 3. In Section 3 we also discuss the use of bootstrapping to calculate CIs and a goodness of fit test of such a model. A numerical example is given in Section 4 and a brief summary is given in Section 5.

## 2 THE NORMAL MODEL

Consider the use of RSO, as described in Section 1, to estimate the optimum expected performance $x_{\min}$ as given in (1), with the individual observations of the form (5).

We consider first the situation where, for a given search point $\theta$, $X(\theta)$ can be observed without error. Consider the following assumption.

**Assumption A** The optimum value $x_{\min}$ is obtained at an interior point $\theta_{\min} \in \Theta$ and, throughout $\Theta$, (i) $X(\theta)$ is twice continuously differentiable and (ii) $X(\theta)$ has a positive definite Hessian, $H(\theta_{\min})$, of second derivatives. Moreover let the first derivative be zero at $\theta_{\min}$.

We have the following result.

**Lemma 2.1** Under Assumption A

$$\Pr[X_i \le x] = K(x - x_{\min})^{d/2}[1 + R(x - x_{\min})], \quad \text{for } x > x_{\min}, \tag{9}$$

for some $K > 0$, where $R(x) = \mathrm{o}(1)$ as $x \to 0$.

**Proof** With no loss of generality assume the components of $\theta$ are taken in standardized units so that the contour where $X(\theta) = x$, with $x$ constant, is the boundary of the hypersphere $S(r, \theta_{\max}) = \{\theta \mid \sum_{i=1}^{m} (\theta_i - \theta_{\max i})^2 \le r^2\}$ of radius $r$ and centre $\theta_{\min}$, where $r^2 = x - x_{\min}$ and whose volume is therefore $Cr^d$ where $C$ depends on $d$ and $H(\theta_{\min})$ but is independent of $r$. Moreover if the densi-

ty (3) is strictly positive and continuous at $\theta_{\min}$ then the density in $S(r,\theta_{\min})$ is a constant to first order in *r*. Thus the proportion of points falling within $S(r,\theta_{\min})$ will be proportional to its volume $Cr^d$ also to first order in *r*. This is exactly equivalent to saying that the tail behavior of *X* near its minimum $X(\theta_{\min})$ is of the form (9). y

Lemma 2.1 shows that the left tail of the distribution of $X_i$ is closely represented by a power-law dependent on the dimension *d*. Indeed if $X(\theta)$ is exactly quadratic and the density is exactly uniform for all $\theta \in S(r,\theta_{\min})$ for some *r*, then (9) clearly holds exactly with $R(x)=0$, for all *x* sufficiently close to $x_{\min}$.

In view of Lemma 2.1 a simple model for $F_X(\cdot)$ with the correct form of left-tail behavior would be the gamma distribution with CDF

$$F_X(x \mid \alpha,\beta,\gamma) = \Gamma^{-1}(\alpha)\beta^{-\alpha}\int_{\gamma}^{w} u^{\alpha-1}\exp\{-[(u-\gamma)/\beta]\}du, \quad \text{for } x>\gamma, \tag{10}$$

where we have written $\alpha = d/2$ and $\gamma = x_{\min}$.

We are actually interested in the case where *d* is large. In this situation the gamma distribution actually converges to a normal form that is not apparent if we retain the parametrization of (10). Cheng and Iles (1990) call such a model an *embedded model* obtainable at an infinite limit. They show that if

$$\lambda = \alpha^{-1/2},\ \mu = \gamma + \alpha\beta,\ \omega = \alpha^{1/2}\beta \tag{11}$$

and let $\alpha \to \infty,\ \beta \to 0,\ \gamma \to -\infty$ in such a way that $\mu$ and $\omega$ remain fixed (with $\lambda \to 0$), then the gamma distribution converges to the normal distribution with mean $\mu$ and standard deviation $\omega$.

There are two implications of this result. Firstly, when $d$ is large, we can approximate the distribution of $X_i$ in its left tail by the normal model

$$F_X(x \mid \mu,\omega) = \Phi((x-\mu)/\omega), \tag{12}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. Secondly, it might appear that this representation does not then include an explicit parameter for the minimum $\gamma$ that we are interested in estimating. This is a consequence of the form of (11) which indicates that $\gamma$ is likely to be large and negative. The behavior of any estimate of $\gamma$ is likely therefore to be unstable and unreliable. We are therefore better off in using (12) for the distribution of $X_i$, and rather than trying to estimate $\gamma$ we should estimate a low quantile of (12) instead. We will adopt this approach and estimate

$$\delta_q = \mu + z_q\omega, \tag{13}$$

where $z_q$ is the *q*th quantile of the standard normal distribution and with $q = 0.05$ say. Thus $\delta_q$ in (13) is estimated simply by estimating $\mu$ and $\omega$.

Our statistical model of the observations (5) is completed by specifying the distribution of $\varepsilon$, the random simulation induced error arising from the within-run stochastic variation. We shall simply assume that this is normal, i.e.

$$\varepsilon \sim N(0,\sigma^2) \tag{14}$$

where $\sigma^2$ is unknown but constant. Normally distributed errors would seem a reasonable assumption in many contexts. For example in the simulation of a continuous production process the output of interest is typically an average output rate or cost.

The discussion just given is appropriate where this left-tail of the distribution of $X_i$ is well approximated by the power-law form given in Lemma 2.1. In view of our discussion, this indicates that a proportion of the smaller $W_i$ values might reasonably be assumed to be the sum of two independent and approximately normally distributed random variables. We therefore make the following assumption.

**Assumption B** Let $W_1 < W_2 < ...W_m$ be the observed averaged observations given in (6), only now taken in ranked order. Given *m*, $\rho$ can be found with $0 < \rho < 1$ for which each of the averaged observations in the subsample

$$W_1 < W_2 < ...W_\nu \tag{15}$$

where

$$\nu = \lfloor m\rho \rfloor \tag{16}$$

takes the form

$$W_i = X_i + \zeta_i \tag{17}$$

with $X_i \sim N(\mu,\omega^2)$ and $\zeta_i \sim N(0,\sigma^2/n)$ are mutually independent so that

$$W_i \sim N(\mu,\psi^2) \tag{18}$$

where

$$\psi^2 = \omega^2 + n^{-1}\sigma^2 . \tag{19}$$

In the next Section we show how to estimate the parameters $\mu, \omega$ and $\sigma$ under Assumption B.

## 3 FITTING THE NORMAL MODEL

We consider estimation of the parameters $\mu, \omega$ and $\sigma$ of the normal model given in Assumption B. It is easiest to estimate $\sigma$ first separately from $\mu$ and $\omega$.

### 3.1 Estimation of $\sigma$

As observations are replicated at each search point, an easy immediate, quite efficient, estimate of $\sigma^2$ is obtainable using

$$\hat{\sigma}^2 = m^{-1}\sum_{i=1}^{m}(n-1)^{-1}\sum_{j=1}^{n}(X_{ij} - \overline{X}_i)^2 \tag{20}$$

where

$$\overline{X}_i = n^{-1}\sum_{j=1}^{n} X_{ij} . \tag{21}$$

Here $\hat{\sigma}^2$ has the same distribution as $\tau^{-1}\chi_\tau^2$ where $\chi_\tau^2$ is a chi-squared variable with $\tau = m(n-1)$ degrees of freedom. Thus, provided $m > n$, $\hat{\sigma}$ has variance that is $O(c^{-1})$ as $c \to \infty$.

### 3.2 Estimation of $\mu$ and $\omega$

We assume that $\hat{\sigma}$ is as given in (20). Using Assumption B we shall apply the method of maximum likelihoo to a *subsample* to estimate the other two parameters $\mu$ and $\omega$. This type of method is used in estimating the value at risk of a financial portfolio (see Pickands 1975, for example). From now on we shall assume that the sample (6) has been ordered so that $W_1 < W_2 < ... < W_m$ and set $\nu = \lfloor m\rho \rfloor$ as in (16).. Clearly, as $m \to \infty$, $W_\nu$ tends in probability to the $\rho$ th quantile of the distribution (18). Thus we have to order $m^{-1/2}$, when *m* is large, that

$$\Phi\big((W_\nu - \mu)/\psi\big) = \rho,$$

i.e.

$$\mu = W_\nu - z_\rho \psi \tag{22}$$

in probability, where $z_\rho$ is the $\rho$ th quantile of the standard normal distribution and $\psi = \sqrt{\omega^2 + n^{-1}\hat{\sigma}^2}$ . Using this condition the loglikelihood is easily shown to be:

$$L(\mu, \omega \mid \hat{\sigma}) = -\frac{\nu}{2}\log(2\pi) - \nu \log(\psi) - \frac{1}{2\psi^2}\sum_{i=1}^{\nu}(W_i - W_\nu - \psi z_\rho)^2, \quad \psi \geq \hat{\sigma}. \tag{23}$$

We can obtain the maximum likelihood (ML) estimate of $\psi$ explicitly in this case:

$$\hat{\psi} = \max\left\{2^{-1}(W_\nu - \overline{W})\left\{\left[z_\rho^2 + 4\left(1 + s_W^2/(W_\nu - \overline{W})^2\right)\right]^{1/2} - z_\rho\right\},\ \ n^{-1/2}\hat{\sigma}\right\}, \tag{24}$$

with the ML estimate of $\mu$ given by

$$\hat{\mu} = W_\nu - z_p\hat{\psi}\,. \tag{25}$$

The distribution of *X* thus has estimated CDF:

$$F_X(x \mid \hat{\mu}, \hat{\omega}) = \Phi[(x - W_\nu + z_\rho\hat{\psi})/\hat{\omega}]\,, \tag{26}$$

where $\hat{\omega} = \sqrt{(\hat{\psi}^2 - n^{-1}\hat{\sigma}^2)}$ provided $\hat{\psi} > \hat{\sigma}$. The case $\hat{\psi} \leq n^{-1/2}\hat{\sigma}$ is an indication that the variance of the error term in (5) is so large that the effect of the performance index variance is lost. In this case the results of the entire search are probably suspect and cannot be relied on.

There is an issue concerning the choice of the value of $\rho$. Making it too large might result in a poor fit because normality of the $X_i$ cannot be guaranteed over the entire range of its distribution, but too small a choice would result in unnecessary loss of estimator efficiency and accuracy. We dealt with this simply by fitting the model over a range of $\rho$ values, $\rho = 0.1,0.2,...,0.9$ in our case, and then selecting a value where the estimates were reasonably stable.

### 3.3 Confidence Intervals for Quantities of Interest

In the numerical example to be presented in Section 4 we examine the distributional properties of three quantities of particular interest The first two are:

(i) $\hat{\delta}_q = \hat{\mu} + z_q\hat{\omega}$, the estimate of $\delta_q$ given in (13).

(ii) $W_1$, the smallest observed average performance measure appearing in (6).

We shall assess the effectiveness of $\hat{\delta}_q$ and $W_1$ as estimators of $\delta_q$ and $x_{\min}$ by examining their distributional properties using the *parametric bootstrap* . The method is described for example by Cheng (2006). The underlying idea is that if the assumptions concerning the distributions of $X_i$ and $\varepsilon_{ij}$ in the original observations (5) are correct then they will be well approximated by $F_X(\cdot\,|\,\hat{\mu},\hat{\omega})$ and $N(0,\hat{\sigma}^2)$ respectively. We therefore take these estimated distributions as being the true distributions and sample from them to generate *B* parametric bootstrap (BS) replicates

$$\{Y_{ij}^*(k) = X_i^*(k) + \varepsilon_{ij}^*(k),\ i = 1,2,...,m,\ j = 1,2..,n\},\ k = 1,2,...,B \tag{27}$$

with $X_i^*(k) \sim F_X(\cdot\,|\,\hat{\mu},\hat{\omega})$ and $\varepsilon_{ij}^*(k) \sim N(0,\hat{\sigma}^2)$, where an asterisk denotes a bootstrapped quantity. Each of the *B* BS replicates in (27) has exactly the same form as (5). Thus *t*, where $t = \hat{\delta}_q$ or $t = W_1$, which were calculated from the original observations (5) can also be calculated from each bootstrap replicate in (27), giving a sample of *B* bootstrap values of $t$: $\{t^*(k),\ k = 1,2,...,B\}$. Under fairly general smoothness conditions, see Bickel and Freedman (1981), the empirical distribution function (EDF) formed from this BS sample is a consistent estimate of the CDF of $t$. We can therefore use the BS sample to construct a CI for the unknown true $\delta_q$ or $x_{\min}$ value being estimated.

In the numerical examples described in Section 4 the $\{t^*(k),\ k = 1,2,...,B\}$ sample is both skewed and biased. We used the following CI based on the conventional normal approximation CI as described by Davison and Hinkley (1997, Section 5.2.1), but modified in a simple way to allow for this asymmetry in the $t^*(k)$ sample. Let $\{t^*(k), k = 1,2,...,B\}$ represent the BS sample and let $\bar{t}^*$ be the BS sample mean obtained from the BS sample.. Now assume that the sample is ordered and define $B_1$ as the subscript for which $t^*(k) \le \bar{t}^*$ for $k = 1,2,...,B_1$ and $\bar{t}^* < t^*(k)$, for $k = B_1, B_1+1,...,B$. Let

$$s_1^2 = \sum_{i=1}^{B_1}(t^*(k) - \bar{t}^*)^2/(B_1 - 1),\ s_2^2 = \sum_{i=B_1+1}^{B}(t^*(k) - \bar{t}^*)^2/(B - B_1 - 1).$$

The suggested $100(1-p)\%$ CI then has lower and upper limits

$$(t^{*L}(p),\ t^{*U}(p)) = \bar{t}^* \pm z_{p/2}s, \tag{28}$$

where $z_{p/2}$ is the upper *p/2* quantile of the $N(0,1)$ distribution and *s* is the larger of $s_1$ and $s_2$.

We consider one other quantity of particular practical interest:

(iii) $X_{(1)}$, the performance measure actually achieved when the search point corresponding to $W_1$ is selected as being the best. We write this as

$$W_1 = X_{(1)} + \zeta_1. \tag{29}$$

We have used a bracketed subscript in $X_{(1)}$ as a reminder that the actual performance measure corresponding to $W_1$ may not be the lowest actual performance measure that has been obtained amongst all the search points examined in the RSO.

The quantity $X_{(1)}$ is not observable in a real RSO. It can be estimated from the model but this is quite complicated. Instead we estimated it using the BS sample mean

$$\overline{X}^*_{(1)} = B^{-1}\sum_{k=1}^{B} X^*_{(1)}(k)\,, \tag{30}$$

We can in this case also directly calculate a CI of the form (28), but with $t^*$ now representing $X^*_{(1)}$, in exactly the same way as described for the cases $t^* = \hat{\gamma}^*$ and $t^* = W_1^*$. This is possible even in a real RSO where $X_{(1)}$ is not known, as the BS process generates all the individual $X_i^*(k)$ and $\varepsilon_{ij}^*(k)$ in (27) which are therefore all known.

The interesting issue arises here as to what is the unknown quantity that the interval (28) is the CIs *for*, when $t^* = X^*_{(1)}$. It can of course be regarded as a CI for E($X_{(1)}$), but it would be much more interesting to regard it, unconventionally, as a CI for the actual unknown $X_{(1)}$. the quantity of real interest, *even though this is random*.Though we do not give a justification here in detail it turns out that this is reasonable because the distribution of the BS observations in (27) tend to those of the original observations. We find that under suitable regularity conditions

$$\Pr(s(p,\hat{\phi}_c) \le X_{(1)} \le t(p,\hat{\phi}_c)) \to 1-p \quad \text{with probability1 as } c \to \infty\,, \tag{31}$$

where $\hat{\phi}_c = (\hat{\mu}_c, \hat{\omega}_c, \hat{\sigma}_c)$ are the estimates obtained from a RSO using $c$ observations, and $s(p,\phi)$ satisfies $F_{X_{(1)}}(s(p,\phi)\,|\,\phi) = p/2$ and $t(p,\phi)$ satisfies $F_{X_{(1)}}(t(p,\phi)\,|\,\phi) = (1-p/2)$. These latter are estimated from the BS CI (28) for $X_{(1)}$. Inversion of (31) in the usual way then gives a CI for $X_{(1)}$.

### 3.4 Goodness of Fit Test

An obvious concern is whether the fitted distribution of $X_i$, is correct or not. The Anderson-Darling (AD) statistic, $A^2$, goodness of fit test with its critical values calculated by parametric resampling can be used for this as discussed by Cheng (2006). The only change needed is that the AD statistic has to be modified to allow it to be applied to the subsample (15). In analogy to the standard AD statistic (see Anderson and Darling, 1952) we define this as

$$A^2 = \nu\int_0^{\rho} \frac{(\hat{F}_W - \tilde{F}_W)^2}{\hat{F}_W(\rho - \hat{F}_W)}\,d\hat{F}_W \tag{32}$$

where $\hat{F}_W = F_W(\cdot\,|\,\hat{\mu},\hat{\omega},\hat{\sigma})$ is the fitted distribution and $\tilde{F}_W$ is the empirical distribution function (EDF) of the subsample (15). Written out explicitly we find that (32) is equivalent to

$$A^2 = -\rho\{\nu - 1 + (\nu-1)^{-1}\sum_{i=1}^{\nu-1}(2i-1)[\log(\hat{F}_i) + \log(\rho - \hat{F}_i)]\}\,, \tag{33}$$

where $\hat{F}_i = F_W(W_i \cdot | \hat{\mu}, \hat{\omega}, \hat{\sigma})$. This form is easier to use for numerical calculations than (32). Note that the summation in (33) does not include $\hat{F}_v$ as the condition (22) requires $\hat{F}_v = \rho$.

## 4 AN EXAMPLE OF A COMBINATORIAL OPTIMIZATION PROBLEM

A situation where our model of Section 3 might be appropriate occurs in combinatorial optimization. In such problems the set of solutions is usually discrete but typically has very large cardinality. Thus it seems reasonable to assume that the distribution of the objective function values obtained in a random search of points drawn from the solution set can adequately be approximated by a continuous distribution. The dimensionality in combinatorial problems is not usually considered. We argue however that if the dimension can be defined at all, it will typically be large. These considerations suggest that our statistical model given in Assumption B might be appropriate in such problems. We demonstrate this in this Section by applying the method of the previous Section to a travelling salesman problem (TSP) containing a stochastic element.

Table 1 gives the x,y coordinates of nine randomly generated points in the unit square. We define a tour as a path that starts and ends at a given point, visiting each other point just once. The TSP problem is to find the tour with the shortest total length, *X*, assuming the Euclidean distance is used for the distance from one point to the next on a tour. The length of each tour is the sum of nine distances, so that though a dimension is not usually associated with such a problem, it seems not unreasonable to regard each tour as associated with a point in 8-dimensional space, assuming the starting point to be fixed. The search space therefore comprises the 8!/2=20160 distinct points in this space corresponding to the distinct tours of the nine original points. The problem thus has a reasonable number of points in the search space, but sufficiently small to enable all tour lengths to be evaluated.

For simplicity of presentation we reduced all tour lengths by the amount of the shortest tour length, so that the minimum transformed tour length corresponds to $x_{\min} = 0$. If we select tours at random, each being equally likely, then we have the RSO problem where the distribution of the tour lengths has CDF depicted in Figure 1. It will be seen that even in this relatively small example the left tail does have a shape one might associate with a normal tail, even if only superficially.

In our problem we assume tour lengths cannot be evaluated accurately and have added a standard normal error. Thus in our problem observed tour lengths are as in (5) with $\sigma = 1$.

Table 1: The x, y Coordinates of the Nine Points of the Travelling Salesman Problem of Section 4.

| x, y | x, y | x, y |
|---|---|---|
| 0.685, 0.991 | 0.195, 0.462 | 0.656, 0.664 |
| 0.083, 0.964 | 0.540, 0.360 | 0.054, 0.831 |
| 0.287, 0.111 | 0.673, 0.600 | 0.095, 0.206 |

Three metaexperiments were carried out, each with a different value of *c*: 100, 1,000 and 10,000, in order to encompass a representative range of *c* values that might be used in practice. Each metaexperiment was made up of 100 independent but otherwise identical experiments. In each experiment an RSO as described in Section 3 was carried out using one of the *c* values. From (8) it is clear that with *d* large we should take *m* much larger *n*. We thus took $m = c/4$ and $n = 4$ to provide $3c/4$ degrees of freedom to estimate $\sigma$ as described in Subsection 3.1.

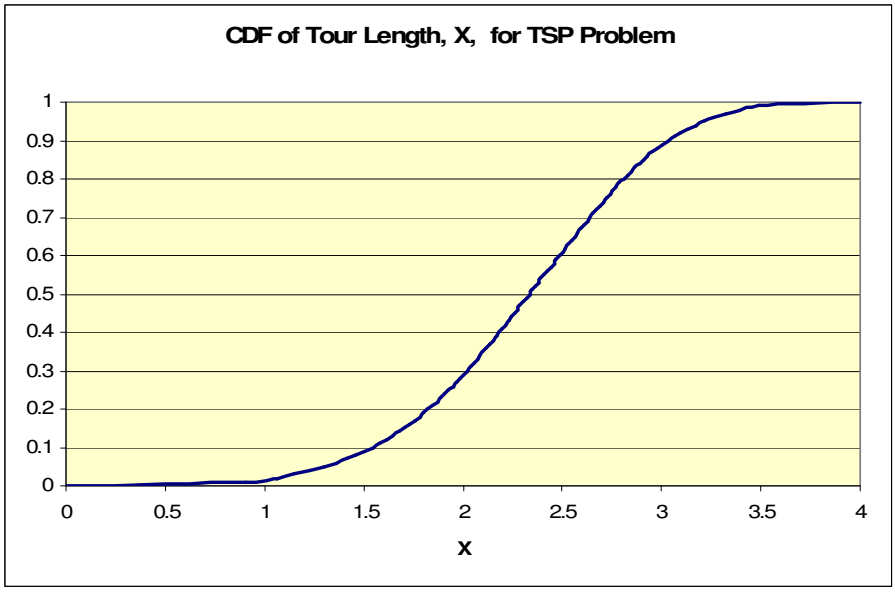


Figure 1: The CDF of $X_i$ the 20160 tour lengths in the travelling salesman problem assuming all tours are equally likely to be selected. All tours are shifted by the amount of the shortest tour length, so that the shortest tour length corresponds to $x_{\min} = 0$

The RSO produced a set of observations of the form (5), with, in each observation, $X_i$ randomlyselected from the 20160 tours, all tours being equally likely to be selected, and with the $\varepsilon_{ij}$ all standard normal variables. We then fitted the model of Assumption B. The three quantities of interest described in Subsection 3.3, namely. $\hat{\delta}$, $W_1$ and $X_{(1)}$ were then calculated.

A set of parametric bootstrap versions of the RSO was then obtained, exactly as described in Subsection 3.3 to produce the three $100(1-p)\%$ CIs: $(\hat{\delta}^{*L}(p),\ \hat{\delta}^{*U}(p))$, $(W_1^{*L}(p),\ W_1^{*U}(p))$, and $(X_{(1)}^{*L}(p),\ X_{(1)}^{*U}(p))$. These were used as CIs for $\delta_{0.05}$, so that $q = 0.05$ in Equation (13), $x_{\min}$= 0 and $X_{(1)}$ respectively. We used 90% CI's so that $p = 0.05$.

In a real RSO study none of these quantities are known. However in our experiments all are known, with $\gamma = 0$, $\delta_{0.05}$ = 1.303, and with $X_{(1)}$ easily obtained from the observations (5). Thus in all three cases we can check if the CI covers the true value or not. The metaexperiment simply replicates the entire experiment just described a number of times, $N$ say, to allow estimation of the true coverage by seeing how often each CI covers its true value.

We also carried out the goodness of fit test of Subsection 3.4 in each experiment of each metaexperiment, with the test returning the result either as 'reject' or 'not reject' the fitted statistical model, and with probability of not rejecting a correct model set at 90%.

In all experiments the number of bootstraps, $B$ was set at 100. The number of replicates, $N$, in a metaexperiment was also set at 100. Even though these are relatively low settings the results obtained provide a reasonably clear picture of the overall performance of the method over a range of conditions.

Table II summarizes the results obtained for the 3 metaexperiments carried out. The Table gives, for each metaexperiment, $\rho$, the proportion of the main RSO sample used in fitting the

model, and $p_{GoF}$, the proportion of the experiments for which the model was *not* rejected as a bad fit by the goodness of fit test. The rows M and SD are the sample mean and sample standard deviation of these samples, which therefore estimate the true behavior of each quantity. Each experiment also produced 100 BS samples of each quantity of interest giving a BS sample mean for each. The mean of these BS sample means, taken over the 100 experiments for each quantity, i.e. the grand mean for each quantity taken over all BS samples and all experiments in the metaexperiment, is given in row BM. The 100 BS values of each quantity in each experiment are also used to calculate a 90% CI. The half-width was recorded in each experiment giving a sample of 100 half-widths. The sample mean of these half widths is recorded in the CI/2 row. The final row, PCI gives the proportion of the 100 experiments that the true value of the quantity of interest is covered by its CI.

The $p_{GoF}$ values in Table 2 indicate that the model fit was satisfactory in the experiments.

The sample means and sample standard deviations in rows M and SD behave generally as one would expect, showing improved accuracy as $c$ increases.

Comparison of the bootstrap means of row BM with those in row M gives an indication of the general reliability of the bootstrap process. The BM values for $\hat{\delta}$ and $W_1$ are not otherwise of particular interest, as the true values of these latter quantities, as given in the M row, are always observable in an RSO. The BM value, (30), for the quantity $X_{(1)}$ is however of more interest, as it is a point estimate of the true $X_{(1)}$ value, which will be unknown in a practical RSO. So it is of interest see how its mean value compares with its known true value in our experiments.

For each quantity of interest, the CI half width in the CI/2 row is a direct measure of the spread in its BS sample. One would expect this to be roughly twice the sample SD value, which in most cases they are.

Perhaps of most interest in Table 2 are the PCI values, giving the observed coverages of the true values of each quantity. The nominal confidence level used in calculating the CIs is 90%. The PCI values for $\hat{\delta}_{0.05}$ and $X_{(1)}$ seem satisfactory. The PCI values for $W_1$ are erratic suggesting that $W_1$ is not a particular good estimator of the true minimum $x_{\min}$ which is as we expected, indicating that our strategy of estimating $\delta_{0.05}$ a low quantile for $X_{(1)}$ is preferable.

Table 2. Results of 3 RSO Metaexperiments for each of Three Values of $c$ in the TSP Example. The Table Entries are as explained in the Text.

| | $c$ = 100 | | | $c$ = 1000 | | | $c$ = 10000 | | |
|---|---|---|---|---|---|---|---|---|---|
| | $m$ | $\rho$ | $p_{GoF}$ | $m$ | $\rho$ | $p_{GoF}$ | $m$ | $\rho$ | $p_{GoF}$ |
| | 25 | 0.4 | 0.90 | 250 | 0.4 | 0.91 | 2500 | 0.2 | 0.92 |
| | $\hat{\delta}_{0.05}$ | $W_1$ | $X_{(1)}$ | $\hat{\delta}_{0.05}$ | $W_1$ | $X_{(1)}$ | $\hat{\delta}_{0.05}$ | $W_1$ | $X_{(1)}$ |
| M | 1.140 | 0.726 | 1.286 | 1.330 | 0.178 | 1.022 | 1.318 | -0.398 | 0.837 |
| SD | 0.422 | 0.396 | 0.417 | 0.120 | 0.306 | 0.404 | 0.036 | 0.251 | 0.388 |
| BM | 0.994 | 0.559 | 1.155 | 1.301 | 0.101 | 1.011 | 1.312 | -0.446 | 0.660 |
| CI/2 | 0.777 | 0.826 | 0.856 | 0.198 | 0.590 | 0.729 | 0.068 | 0.505 | 0.722 |
| PCI | 0.87 | 0.61 | 0.9 | 0.92 | 0.99 | 0.89 | 0.92 | 0.74 | 0.91 |

## 5 SUMMARY

We have discussed a statistical model that may be of use in analysing the results of an RSO. The use of bootstrapping enables the quality of estimates of quantities of interest to be gauged. In particular the model enables estimation not only of $X_{(1)}$, the performance actually achieved when the search point corresponding to the best observed search point is selected as being the best, but also of quantile points, $\delta_q$, as given in equation (13). The proportion of all possible solution points $\theta$ with objective function value less than (i.e. better than) $\delta_q$ is $q$. Thus comparison of the estimates of $X_{(1)}$ and $\delta_q$ gives an indication not only of the best value of the objective function obtained in the search, but also of the quality of this value.

## REFERENCES


Anderson, T. W., and D. A. Darling. 1952. Asymptotic theory of certain 'goodness of fit' criteria based on stochastic processes. *Annals of Mathematical Statistics 23*, 193-212.

Bickel, P. J., and D. A. Freedman. 1981 Some asymptotic theory for the bootstrap. *Annals of Statistics. 9*, 1196-1217.

Cheng, R. C. H., and T. C. Iles. 1990. Embedded models in three-parameter distributions and their estimation. *J. Roy. Statist. Soc. B 52*, 135-149.

Cheng, R. C. H. 2006. Validating and comparing simulation models using resampling. *J. of Simulation. 1*, 53-63.

Chia, Y. L. 2005 *Simulation-based estimation with application in biomedical research.* PhD Thesis, Stanford University.

Chia, Y. L., and P. W. Glynn. 2007. Optimal convergence rate for random search. In *Proceedings of the 2007 INFORMS Simulation Society Workshop Eds C-H Chen and S.G. Henderson* http://www.insead.edu/v1/issrw/documents/11.pdf

Davison, A. C. and D. V. Hinkley. 1997. *Bootstrap Methods and their Application.* Cambridge: Cambridge University Press.

Pickands, J. 1975. Statistical inference using extreme order statistics. *Annals of Statistics 3***,** 119-131.

Yakowitz, S., P. L'Ecuyer, and F. Vásquez-Abad. 2000. Global stochastic optimization with low-dispersion point sets. *Operations Research 48*, 939-950.


## AUTHOR BIOGRAPHY


**RUSSELL C. H. CHENG** is Emeritus Professor of Operational Research at the University of Southampton. He has an M.A. and the Diploma in Mathematical Statistics from Cambridge University, England. He obtained his Ph.D. from Bath University. He is a former Chairman of the U.K. Simulation Society, a Fellow of the Royal Statistical Society and a Member of the Operational Research Society. His research interests include: design and analysis of simulation experiments and parametric estimation methods. He was a Joint Editor of the *IMA Journal of Management Mathematics*. His email and web addresses are <R.C.H.Cheng@soton.ac.uk> and <www.personal.soton.ac.uk/rchc>.