

Geng Deng · Michael C. Ferris

# Variable-Number Sample-Path Optimization

the date of receipt and acceptance should be inserted later

**Abstract** The sample-path method is one of the most important tools in simulation-based optimization. The basic idea of the method is to approximate the expected simulation output by the average of sample observations with a common random number sequence. In this paper, we describe a new variant of Powell’s UOBYQA (*Unconstrained Optimization BY Quadratic Approximation*) method, which integrates a Bayesian Variable-Number Sample-Path (VNSP) scheme to choose appropriate number of samples at each iteration. The statistically accurate scheme determines the number of simulation runs, and guarantees the global convergence of the algorithm. The VNSP scheme saves a significant amount of simulation operations compared to general purpose ‘fixed-number’ sample-path methods. We present numerical results based on the new algorithm.

**Keywords** sample-path method, simulation-based optimization, Bayesian analysis, trust region method

**Dedication** This paper is dedicated to Stephen Robinson on the occasion of his 65th birthday. The authors are grateful for his encouragement and guidance over the past two decades, and the inspirational work he has done in the topic of this paper.

## 1 Introduction

Computer simulations are used extensively as models of real systems to evaluate output responses. The choice of optimal simulation parameters can lead to improved operation, but configuring them well remains a challenging problem. Historically, the parameters are chosen by selecting the best from a set of candidate parameter settings. *Simulation-based optimization* [12, 13, 20] is an emerging field which integrates optimization techniques into simulation analysis. The corresponding objective function is an associated measurement of an experimental simulation. Due to the complexity of the simulation, the objective function may be difficult and expensive to evaluate. Moreover, the inaccuracy of the objective function often complicates the optimization process. Indeed, derivative information is typically unavailable, so many derivative-dependent methods are not applicable to these problems.

---

This material is based on research partially supported by the National Science Foundation Grants DMI-0521953, DMS-0427689 and IIS-0511905 and the Air Force Office of Scientific Research Grant FA9550-04-1-0192

---

G. Deng  
Department of Mathematics, University of Wisconsin, 480 Lincoln Drive, Madison, WI 53706, USA, E-mail: geng@cs.wisc.edu

M. C. Ferris  
Computer Sciences Department, University of Wisconsin, 1210 West Dayton Street, Madison, WI 53706, USA, E-mail: ferris@cs.wisc.edu

Although real world problems have many forms, in this paper we consider the following unconstrained stochastic formulation:

$$\min_{x \in \mathbb{R}^n} f(x) = \mathbb{E}[F(x, \xi(\omega))]. \quad (1.1)$$

Here,  $\xi(\omega)$  is a random vector defined on a probability space  $(\Omega, \mathcal{F}, P)$ . The sample response function  $F: \mathbb{R}^n \times \mathbb{R}^d \rightarrow \mathbb{R}$  takes two inputs, the simulation parameters  $x \in \mathbb{R}^n$  and a random sample of  $\xi(\omega)$  in  $\mathbb{R}^d$ . Given a random realization  $\xi_i$  of  $\xi(\omega)$ ,  $F(x, \xi_i)$  can be evaluated via a single simulation run. The underlying objective function  $f(x)$  is computed by taking an expectation over the sample response function and has no explicit form. A basic assumption requires that the expectation function  $f(x)$  is *well defined* (for any  $x \in \mathbb{R}^n$  the function  $F(x, \cdot)$  is measurable, and either  $\mathbb{E}[F(x, \xi(\omega))_+]$  or  $\mathbb{E}[F(x, \xi(\omega))_-]$  is finite, see page 57 of [31]).

The *sample-path method* is a well-recognized technique in simulation-based optimization [11, 14, 15, 25, 26, 30]. It is sometimes called the *Monte Carlo sampling approach* [34] or the *sample average approximation method* [16, 17, 19, 33, 35, 36]. The sample-path method has been applied in many settings, including buffer allocation, tandem queue servers, network design, etc. The basic idea of the method is to approximate the expected value function  $f(x)$  in (1.1) by averaging sample response functions

$$f(x) \approx \hat{f}^N(x) := \frac{1}{N} \sum_{i=1}^N F(x, \xi_i), \quad (1.2)$$

where  $N$  is an integer representing the number of samples. Note that by fixing a sequence of i.i.d. samples  $\xi_i, i = 1, 2, \dots, N$  in (1.2), the approximate function  $\hat{f}^N$  is a deterministic function. This advantageous property allows the application of deterministic techniques to the averaged sample-path problem

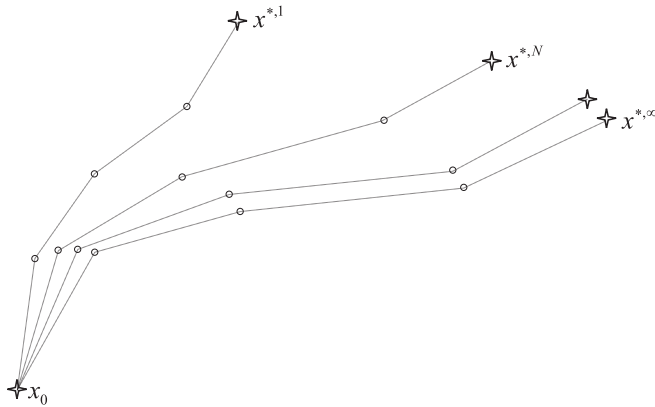
$$\min_{x \in \mathbb{R}^n} \hat{f}^N(x), \quad (1.3)$$

which serves as a substitute for (1.1). An optimal solution  $x^{*,N}$  to the problem (1.3) is then treated as an approximation of  $x^*$ , the solution of (1.1). Note that the method is not restricted to unconstrained problems as in our paper, but it requires appropriate deterministic tools (i.e., constrained optimization methods) to be used.

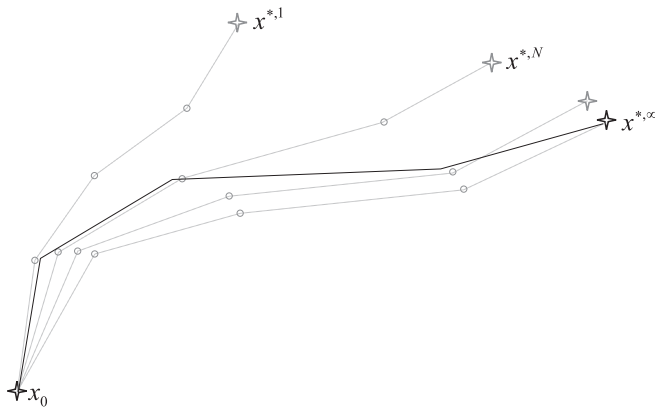
Convergence proofs of the sample-path method are given in [30, 32]. Suppose there is a unique solution  $x^*$  to the problem (1.1), then under assumptions such as the sequence of functions  $\{\hat{f}^N\}$  epiconverges to the function  $f$ , the optimal solution sequence  $\{x^{*,N}\}$  converges to  $x^*$  almost surely for all sample paths. Note that a sample path corresponds to a sequence of realized samples  $\{\xi_1, \xi_2, \dots\}$ . The almost sure statement is defined with respect to the generated probability measure  $\bar{P}$  of the sample path space  $\bar{\Omega} = \Omega \times \Omega \times \dots$ . See Figure 1 for the illustration of the sample-path optimization method.

Our purpose in this paper is to introduce a *Variable-Number Sample-Path* (VNSP) scheme, an extension of sample-path optimization. The classical sample-path method is criticized for its excessive simulation evaluations: in order to obtain a solution point  $x^{*,N}$ , one has to solve an individual optimization problem (1.3) and at each iterate  $x_k$  of the algorithm  $\hat{f}^N(x_k)$  is required (with  $N$  large). The new VNSP scheme is designed to generate different numbers of samples ( $N$ ) at each iteration. Denoting  $N_k$  as the number of samples at iteration  $k$ , the VNSP scheme integrates Bayesian techniques to determine a satisfactory  $N_k$ , which accordingly ensures the accuracy of the approximation of  $\hat{f}^N(x)$  to  $f(x)$ . The numbers  $\{N_k\}$  form a non-decreasing sequence within the algorithm, with possible convergence to infinity. The new approach is briefly described in Figure 2. Significant computational savings accrue when  $k$  is small. There is an extensive literature on using Bayesian methods in simulation output analysis. For example, Chick and Inoue [3, 4] has implemented Bayesian estimation in ordering discrete simulation systems (ranking and selection [1, 18]). Deng and Ferris [8] propose a similar Bayesian analysis to evaluate the stability of surrogate models.

Another ‘variable-sample’ scheme for sample-path optimization is proposed by Homem-de-Mello in [16]. The work proposes a framework for iterative algorithms that use, at iteration  $k$ , an estimator  $f^{N_k}$  of the true function  $f$  constructed via the sample average of  $N_k$  samples. It is shown in [16] that, if the convergence of such an algorithm requires that  $f^{N_k}(x) \rightarrow f(x)$  almost surely for all sample



**Fig. 1** Mechanism of the sample-path optimization method. Starting from  $x_0$ , for a given  $N$ , a deterministic algorithm is applied to solve the sample-path problem. The sequence of solutions  $\{x^{*,N}\}$  converges to the true solution  $x^{*,\infty} = x^*$  almost surely.



**Fig. 2** Mechanism of the new sample-path method with the VNSP scheme. Starting from  $x_0$ , the algorithm generates its iterates across different averaged sample functions. In an intermediate iteration  $k$ , it first computes a satisfactory  $N_k$  which guarantees certain level of accuracy, then an optimization step is taken exactly the same as in problem (1.3), with  $N = N_k$ . The algorithm has a globally convergent solution  $x^{*,N_\infty}$ , where  $N_\infty := \lim_{k \rightarrow \infty} N_k$ . The convergence is almost sure for all the sample paths, which correspond to different runs of the algorithm. The solution, we will prove later, matches the solution  $x^{*,\infty}$ .

paths, then it is necessary that  $N_k \rightarrow \infty$  at a certain rate. Our VNSP scheme is significantly different:  $N_k$  in our scheme is validated based on the uncertainty of the iterate  $x_k$ . We require  $x_k \rightarrow x^*$  almost surely, but we do not impose the convergence condition  $\hat{f}^{N_k} \rightarrow f$ . As a consequence,  $\{N_k\}$  is a non-decreasing sequence with the limit value  $N_\infty$  being either finite or infinite. Here is a toy example showing that the limit sample number  $N_\infty$  in our algorithm can be finite. Consider a simulation system with only ‘white noise’:

$$F(x, \xi(\omega)) = \phi(x) + \xi(\omega),$$

where  $\phi(x)$  is a deterministic function and  $\xi(\omega) \sim N(0, \sigma^2)$ . As a result, the minimizer of each piece  $F(x, \xi_i) = \phi(x) + \xi_i$  coincides with the minimizer of  $f(x) = \phi(x)$  (thus the solutions of  $\hat{f}^k$  are:  $x^{*,1} = x^{*,2} = \dots = x^{*,\infty}$ ). In this case, our VNSP scheme turns out to use a constant sequence of sample numbers  $N_k : N_1 = N_2 = \dots = N_\infty < +\infty$ . We obtain  $\lim_{k \rightarrow \infty} x_k = x^{*,N_1} = \dots = x^{*,N_\infty} = x^*$ , but obviously  $\lim_{k \rightarrow \infty} \hat{f}^{N_k} \neq f$ . However, the ‘variable-sample’ scheme in [16] still requires  $\lim_{k \rightarrow \infty} N_k = \infty$  on this example. More details about this toy example can be found in the numerical example section.

Sections of the paper are arranged as follows. In Section 2.1 we detail the underlying quadratic models that we will use and outline properties of the model construction that are relevant to the

sequel. In Section 2.2 we will provide the outline of the new algorithm, with a realization of the VNSP scheme. In Section 2.3, we describe the Bayesian VNSP scheme to determine the suitable value of  $N_k$  at iteration  $k$ . Section 3 provides an analysis of the global convergence properties of the algorithm. Finally, in Section 4, we discuss several numerical results on test functions.

## 2 The Extended UOBYQA Algorithm

We apply Powell's UOBYQA (*Unconstrained Optimization BY Quadratic Approximation*) algorithm [27] as our base sample-path optimization solver. The algorithm is a derivative-free approach and thus is a good fit for the optimization problem (1.3). It is designed to solve nonlinear problems with a moderate number of dimensions. The general structure of UOBYQA follows a *model-based approach* [5,6], which constructs a chain of local quadratic models that approximate the objective function. The method is an iterative algorithm in a trust region framework [24], but it differs from a classical trust region method in that it creates quadratic models by interpolating a set of sample points instead of using the gradient and Hessian values of the objective function (thus making it a derivative-free tool). Besides UOBYQA, other model-based software include WEDGE [21] and NEWUOA [28].

A general framework for the model-based approach is given by Conn and Toint [6], and convergence analysis is presented in [5]. In our extension of UOBYQA, we inherit several basic assumptions regarding the nature of the objective function from [5].

**Assumption 1** *For a fixed  $y \in \mathbb{R}^d$  the function  $F(\cdot, y)$  is twice continuously differentiable and its gradient and Hessian are uniformly bounded on  $\mathbb{R}^n \times \mathbb{R}^d$ . There exist constants  $\kappa_{Fg} > 0$  and  $\kappa_{Fh} > 0$ , such that the following inequalities hold:*

$$\sup_{x \in \mathbb{R}^n, y \in \mathbb{R}^d} \left\| \frac{\partial F(x, y)}{\partial x} \right\| \leq \kappa_{Fg} \quad \text{and} \quad \sup_{x \in \mathbb{R}^n, y \in \mathbb{R}^d} \left\| \frac{\partial^2 F(x, y)}{\partial^2 x} \right\| \leq \kappa_{Fh}.$$

**Assumption 2** *For a given  $y \in \mathbb{R}^d$ , the function  $F(\cdot, y)$  and the underlying function  $f(\cdot)$  are bounded below on  $\mathbb{R}^n$ .*

### 2.1 Interpolating quadratic model properties

At every iteration of the algorithm, a quadratic model

$$Q_k^N(x) = c_k^N + (g_k^N)^T (x - x_k) + \frac{1}{2}(x - x_k)^T G_k^N (x - x_k), \quad (2.1)$$

is constructed by interpolating a set of adequate points (see explanation below)  $\mathcal{I}_k = \{y^1, y^2, \dots, y^L\}$ ,

$$Q_k^N(y^i) = \hat{f}^N(y^i), \quad i = 1, 2, \dots, L. \quad (2.2)$$

We will indicate how to generate the number of samples  $N$  in Section 2.3 using a Bayesian VNSP scheme.

The point  $x_k$  acts as the center of a trust region, the coefficient  $c_k^N$  is a scalar,  $g_k^N$  is a vector in  $\mathbb{R}^n$ , and  $G_k^N$  is an  $n \times n$  real symmetric matrix. The interpolation model is expected to approximate  $\hat{f}^N$  well around the base point  $x_k$ , such that the parameters  $c_k^N, g_k^N$  and  $G_k^N$  approximate the Taylor series expansion coefficients of  $\hat{f}^N$  around  $x_k$ . Thus,  $g_k^N$  is used as a derivative estimate for  $\hat{f}^N$ . To ensure a unique quadratic interpolator, the number of interpolating points should satisfy

$$L = \frac{1}{2}(n+1)(n+2). \quad (2.3)$$

Note that the model construction step (2.1) does not require evaluations of the gradient or the Hessian. However, for each quadratic interpolation model, we require that the Hessian matrix is uniformly bounded.

**Assumption 3** *The Hessian of the quadratic function  $Q_k^N$  is uniformly bounded for all  $x$  in the trust region, i.e., there exists a constant  $\kappa_{Qh} > 0$  such that*

$$\|G_k^N\| \leq \kappa_{Qh}, \text{ for all } x \in \{x \in \mathbb{R}^n \mid \|x - x_k\| \leq \Delta_k\}.$$

The notion of adequacy of the interpolation points in a ball

$$\mathcal{B}_k(d) := \{x \in \mathbb{R}^n \mid \|x - x_k\| \leq d\}$$

is defined in [5]. As a key component of the analysis, Conn, Scheinberg, and Toint address the difference of using the classical Taylor expansion model

$$\hat{Q}_k^N(x) = \hat{f}^N(x_k) + \nabla \hat{f}^N(x_k)^T(x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 \hat{f}^N(x_k)(x - x_k)$$

and the interpolative quadratic model  $Q_k^N$ . The model  $\hat{Q}_k^N$  shares the same gradient  $\nabla \hat{f}^N(x_k)$  at  $x_k$  with the underlying function, while for the interpolative model  $Q_k^N$ , its gradient  $g_k^N$  is merely an approximation. The error in this approximation is shown in the following lemma to decrease quadratically with the trust region radius. As an implication of the lemma, within a small trust region, the model  $Q_k^N$  is also a decent approximation model.

**Lemma 1** *(Theorem 4 in [5]) Assume Assumptions 1-3 hold and  $\mathcal{I}_k$  is adequate in the trust region  $\mathcal{B}_k(\Delta_k)$ . Suppose at iteration  $k$ ,  $Q_k^N$  is the interpolative approximation model for the function  $\hat{f}^N$ , then the bias of the function value and the gradient are bounded within the trust region. There exist constants  $\kappa_{em}$  and  $\kappa_{eg}$ , for each  $x \in \mathcal{B}_k(\Delta_k)$ , the following inequalities hold*

$$|\hat{f}^N(x) - Q_k^N(x)| \leq \kappa_{em} \max[\Delta_k^2, \Delta_k^3] \quad (2.4)$$

and

$$\|\nabla \hat{f}^N(x) - g_k^N\| \leq \kappa_{eg} \max[\Delta_k, \Delta_k^2]. \quad (2.5)$$

In fact, the proof of Lemma 1 is associated with manipulating Newton polynomials instead of the Lagrange functions that UOBYQA uses. Since the quadratic model is unique via interpolation (by choice of  $L$ ), the results are valid regardless of how the model is constructed.

Implicitly, adequacy relates to good conditioning of an underlying matrix, which enables the interpolation model to work well. Improving the adequacy of the point set involves replacing a subset of points with new ones. The paper [5] shows a mechanism that will generate adequate interpolation points after a finite number of operations. UOBYQA applies a heuristic procedure, which may not guarantee these properties, but is very effective in practice. Since this point is unrelated to the issues we address here, we state the theory in terms of adequacy to be rigorous, but use the UOBYQA scheme for our practical implementation.

We have seen that  $Q_k^N$  interpolates the function  $\hat{f}^N$  at the points in  $\mathcal{I}_k$ . Let  $Q_k^\infty$  be the ‘expected’ quadratic model interpolating the function  $f$  at the same points. The following lemma provides convergence of  $Q_k^N$  to  $Q_k^\infty$ .

**Lemma 2**  $Q_k^N(x)$  converges pointwise to  $Q_k^\infty(x)$  with probability 1 (w.p.1) as  $N \rightarrow \infty$ .

*Proof* The Law of Large Numbers (LLN) guarantees the pointwise convergence of  $\hat{f}^N(x)$  to  $f(x)$  w.p.1 [31]. By solving the system of linear equations (2.2), each component of the coefficients of  $Q_k^N$ ,  $c_k^N$ ,  $g_k^N(i)$ ,  $G_k^N(i, j)$ ,  $i, j = 1, 2, \dots, n$ , is uniquely expressed as a linear combination of  $\hat{f}^N(y^i)$ ,  $\hat{f}^N(y^i)\hat{f}^N(y^j)$ ,  $i, j = 1, 2, \dots, L$ . (The uniqueness of solution requires the adequacy of the interpolation points.) Therefore, as  $N \rightarrow \infty$  the coefficients  $c_k^N$ ,  $g_k^N$ ,  $G_k^N$  converge to  $c_k^\infty$ ,  $g_k^\infty$ ,  $G_k^\infty$  w.p.1 because the values  $\hat{f}^N(y^i)$  converge to  $f(y^i)$ ,  $i = 1, 2, \dots, L$ , w.p.1. Finally, for a fixed value  $x \in \mathbb{R}^n$ ,  $Q_k^N(x)$  converges to  $Q_k^\infty(x)$  w.p.1.  $\square$

In the remainder of the section, we focus on deriving the posterior distributions of  $Q_k^\infty$  and computing the Bayes risk. These distributions will be used in Section 2.3; they are summarized in the penultimate paragraph of this subsection for a reader who wishes to skip the technical details. Assume the simulation output at points of  $\mathcal{I}_k$

$$\mathbf{F} = (F(y^1, \xi(\omega)), F(y^2, \xi(\omega)), \dots, F(y^L, \xi(\omega)))$$

is a multivariate normal variable, with mean  $\boldsymbol{\mu} = (\mu(y^1), \dots, \mu(y^L))$  and covariance matrix  $\boldsymbol{\Sigma}$ :

$$\mathbf{F} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (2.6)$$

Since the simulation outcomes are correlated, the covariance matrix is typically not a diagonal matrix. The existing data  $X^N$  can be accumulated as an  $N \times L$  matrix, with

$$X_{i,j}^N = f(y^j, \xi_i), i = 1, \dots, N, j = 1, \dots, L,$$

and  $L$  is the cardinality of the set  $\mathcal{I}_k$  defined in (2.3). The data is available before the construction of the model  $Q_k^N$ . Let  $\bar{\boldsymbol{\mu}}$  and  $\hat{\boldsymbol{\Sigma}}$  denote the sample mean and sample covariance matrix of the data. For simplicity, we introduce the notation  $\mathbf{s}_i = (F(y^1, \xi_i), \dots, F(y^L, \xi_i)), i = 1, \dots, N$ , so that

$$X^N = \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_N \end{bmatrix}. \text{ The sample mean and sample covariance matrix are calculated as}$$

$$\begin{aligned} \bar{\boldsymbol{\mu}} &= \frac{1}{N} \sum_{i=1}^N \mathbf{s}_i \\ &= (\hat{f}^N(y^1), \dots, \hat{f}^N(y^L)), \end{aligned} \quad (2.7)$$

and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{s}_i - \bar{\boldsymbol{\mu}})^T (\mathbf{s}_i - \bar{\boldsymbol{\mu}}). \quad (2.8)$$

We delve into the detailed steps of quadratic model construction in the UOBYQA algorithm. The quadratic model  $Q_k^\infty$  is expressed as a linear combination of Lagrange functions  $l_j(x)$ ,

$$Q_k^\infty(x) = \sum_{j=1}^L f(y^j) l_j(x) = \sum_{j=1}^L \mu(y^j) l_j(x), \quad x \in \mathbb{R}^n. \quad (2.9)$$

Each piece of  $l_j(x)$  is a quadratic polynomial from  $\mathbb{R}^n$  to  $\mathbb{R}$

$$l_j(x_k + s) = c_j + g_j^T s + \frac{1}{2} s^T G_j s, \quad j = 1, 2, \dots, L,$$

that has the property

$$l_j(y^i) = \delta_{ij}, \quad i = 1, 2, \dots, L,$$

where  $\delta_{ij}$  is 1 if  $i = j$  and 0 otherwise. It follows from (2.1) and (2.9) that the parameters of  $Q_k^\infty$  are derived as

$$\begin{aligned} c_k^\infty &= \mathbf{c}\boldsymbol{\mu}^T, \quad g_k^\infty = \mathbf{g}\boldsymbol{\mu}^T, \\ \text{and } G_k^\infty &= \sum_{j=1}^L \mu(y^j) G_j, \end{aligned} \quad (2.10)$$

where  $\mathbf{c} = (c_1, \dots, c_L)$  and  $\mathbf{g} = (g_1, \dots, g_L)$ . Note that the parameters  $c_j$ ,  $g_j$ , and  $G_j$  in each Lagrange function  $l_j$  are uniquely determined when the points  $y^j$  are given, regardless of the function  $f$ .

Since we do not have any prior assumption for the distributions of  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$ , we assign non-informative prior distributions for them. In doing this, the joint posterior distributions of  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  are derived as

$$\begin{aligned}\boldsymbol{\Sigma}|X^N &\sim \text{Wishart}_L(\hat{\boldsymbol{\Sigma}}, N + L - 2), \\ \boldsymbol{\mu}|\boldsymbol{\Sigma}, X^N &\sim N(\bar{\boldsymbol{\mu}}, \boldsymbol{\Sigma}/N).\end{aligned}\tag{2.11}$$

Here the Wishart distribution  $\text{Wishart}_p(\boldsymbol{\nu}, m)$  has covariance matrix  $\boldsymbol{\nu}$  and  $m$  degrees of freedom. The Wishart distribution is a multivariate generalization of the  $\chi^2$  distribution.

The distribution of the mean value  $\boldsymbol{\mu}$  is of most interest to us. When the sample size is large, we can replace the covariance matrix  $\boldsymbol{\Sigma}$  in (2.11) with the sample covariance matrix  $\hat{\boldsymbol{\Sigma}}$ , and asymptotically derive the posterior distribution of  $\boldsymbol{\mu}|X^N$  as

$$\boldsymbol{\mu}|X^N \sim N(\bar{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}/N).\tag{2.12}$$

It should be noted that, with an exact computation, the marginal distribution of  $\boldsymbol{\mu}|X^N$  inferred by (2.11) (eliminating  $\boldsymbol{\Sigma}$ ) is,

$$\boldsymbol{\mu}|X^N \sim St_L(\bar{\boldsymbol{\mu}}, N\hat{\boldsymbol{\Sigma}}^{-1}, N - 1),\tag{2.13}$$

where a random variable with Student's t-distribution  $St_L(\boldsymbol{\mu}, \boldsymbol{\kappa}, m)$  has mean  $\boldsymbol{\mu}$ , precision  $\boldsymbol{\kappa}$ , and  $m$  degrees of freedom. The normal formulation (2.12) is more convenient to manipulate than the t-version (2.13), and the results of both versions turn out to be close [9]. Therefore, in our work, we will use the normal distribution (2.12).

Combining (2.10) and (2.12), the posterior distributions of  $c_k^\infty, g_k^\infty$  and  $G_k^\infty$  are normal-like distributions:

$$c_k^\infty|X^N \sim N(\mathbf{c}\bar{\boldsymbol{\mu}}^T, \mathbf{c}\hat{\boldsymbol{\Sigma}}\mathbf{c}^T/N),\tag{2.14}$$

$$g_k^\infty|X^N \sim N(\mathbf{g}\bar{\boldsymbol{\mu}}^T, \mathbf{g}\hat{\boldsymbol{\Sigma}}\mathbf{g}^T/N),\tag{2.15}$$

$$G_k^\infty|X^N \sim MN\left(\sum_{j=1}^L \bar{\mu}(y^j)G_j, \mathbf{P}^T \hat{\boldsymbol{\Sigma}} \mathbf{P}/N, \mathbf{P}^T \hat{\boldsymbol{\Sigma}} \mathbf{P}/N\right),\tag{2.16}$$

where the  $L \times N$  matrix  $\mathbf{P} = (G_1\mathbf{1}, \dots, G_L\mathbf{1})^T$ . The matrix normal distribution  $MN(\boldsymbol{\mu}, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2)$  has parameters mean  $\boldsymbol{\mu}$ , left variance  $\boldsymbol{\nu}_1$ , and right variance  $\boldsymbol{\nu}_2$  [7]. In (2.16), because  $G_j$  are symmetric, the left variance and right variance coincide.

While the multivariate normal assumption (2.6) is not always valid, several relevant points indicate that it is likely to be satisfied in practice [2].

- The form (2.6) is only used to derive the (normal) posterior distribution  $\boldsymbol{\mu}|X$ .
- Other types of distribution assumptions may be appropriate in different circumstances. For example, when a simulation output follows a Bernoulli 0-1 distribution, then it would be easier to perform parameter analysis using beta prior and posterior distributions. The normal assumption (2.6) is the more relevant to continuous simulation output with unknown mean and variance.
- The normal assumption is asymptotically valid for many applications. Many regular distributions, such as distributions from the exponential family, are normal-like distributions. The analysis using normal distributions is asymptotically correct.

## 2.2 The core algorithm

In this section, we present an algorithm outline based on the general model-based approach, omitting specific details of UOBYQA. Interested readers may refer to Powell's paper [27] for further details.

Starting the algorithm requires an initial trial point  $x_0$  and an initial trust region radius  $\Delta_0$ . As in a classical trust region method, a new promising point is determined from a subproblem:

$$\min_{s \in \mathbb{R}^n} Q_k^N(x_k + s), \quad \text{subject to } \|s\| \leq \Delta_k.\tag{2.17}$$

The new solution  $s^{*,N}$  is accepted (or not) by evaluating the ‘degree of agreement’ between  $\hat{f}^N$  and  $Q_k^N$ :

$$\rho_k^N = \frac{\hat{f}^N(x_k) - \hat{f}^N(x_k + s^{*,N})}{Q_k^N(x_k) - Q_k^N(x_k + s^{*,N})}. \quad (2.18)$$

If the ratio  $\rho_k^N$  is large enough, which indicates a good agreement between the quadratic model  $Q_k^N$  and the function  $\hat{f}^N$ , the point  $x_k + s^{*,N}$  is accepted into the set  $\mathcal{I}_k$ . We introduce the following lemma concerning the ‘sufficient reduction’ within a trust region step. This is an important but standard result in the trust region literature.

**Lemma 3** *The solution  $s_k^{*,N}$  of the subproblem (2.17) satisfies*

$$Q_k^N(x_k) - Q_k^N(x_k + s^{*,N}) \geq \kappa_{mdc} \|g_k^N\| \min \left[ \frac{\|g_k^N\|}{\kappa_{Qh}}, \Delta_k \right] \quad (2.19)$$

for some constant  $\kappa_{mdc} \in (0, 1)$  independent of  $k$ .

*Proof* For the *Cauchy point*  $x_k + s_c^N$  defined as the minimizer of the model in the trust region along the steepest decent direction, we have a corresponding reduction [22]

$$Q_k^N(x_k) - Q_k^N(x_k + s_c^{*,N}) \geq \frac{1}{2} \|g_k^N\| \min \left[ \frac{\|g_k^N\|}{\kappa_{Qh}}, \Delta_k \right]. \quad (2.20)$$

Since the solution  $s^{*,N}$  of the subproblem yields an even lower objective value of  $Q_k^N$ , we have the inequality (2.19). The complete proof can be found in [24].  $\square$

Comment 1: Lemma 3 is generally true for models  $Q_k^N$  and  $Q_k^\infty$ .

Comment 2: There are issues concerning setting the values of  $\kappa_{mdc}$  and  $\kappa_{Qh}$  in an implementation. For  $\kappa_{mdc}$ , we use a safeguard value of 0.49, which is slightly smaller than  $\frac{1}{2}$ . This value is true for Cauchy points, so is valid for the solutions of the subproblem. For  $\kappa_{Qh}$ , we update it as the algorithm proceeds

$$\kappa_{Qh} := \max(\kappa_{Qh}, \|G_k^N\|), \quad (2.21)$$

that is,  $\kappa_{Qh}$  is updated whenever a new  $G_k^N$  is generated. Assumption 3 ensures the boundedness of the sampled Hessian and prevents the occurrence of ill-conditioned problems. It is hard to find a good value of  $\kappa_{Qh}$  satisfying Assumption 3, but in practice the above scheme updates the value very infrequently.

It may happen that the quadratic model becomes inadequate after a potential step. Accordingly, UOBYQA first checks and improves the adequacy of  $\mathcal{I}_k$  before the trust region radius is updated following standard trust region rules. Whenever a new point  $x^+$  enters (the point  $x^+$  may be the solution point  $x_k + s^{*,N}$  or a replacement point to improve the geometry), the agreement is rechecked to determine the next iterate.

We now present the extended UOBYQA algorithm that uses the VNSP scheme that we describe in the next section. The constants associated with the trust region update are:

$$0 < \eta_0 \leq \eta_1 < 1, 0 < \gamma_0 \leq \gamma_1 < 1 \leq \gamma_2, \epsilon_1 > 0 \text{ and } \epsilon_2 \geq 1.$$

**Algorithm 1** *Choose a starting point  $x_0$ , an initial trust region radius  $\Delta_0$  and a termination trust region radius  $\Delta_{end}$ .*

1. *Generate initial trial points in the interpolation set  $\mathcal{I}_k$ . Determine the first iterate  $x_1 \in \mathcal{I}_k$  as the best point in  $\mathcal{I}_k$ .*
2. *For iterations  $k = 1, 2, \dots$* 
  - (a) *Determine  $N_k$  via the VNSP scheme in Section 2.3.*
  - (b) *Construct a quadratic model  $Q_k^{N_k}$  of the form (2.1) which interpolates points in  $\mathcal{I}_k$ . If  $\|g_k^{N_k}\| \leq \epsilon_1$  and  $\mathcal{I}_k$  is inadequate in  $\mathcal{B}_k(\epsilon_2 \|g_k^{N_k}\|)$ , then improve the quality of  $\mathcal{I}_k$ .*
  - (c) *Solve the trust region subproblem (2.17). Evaluate  $\hat{f}^{N_k}$  at the new point  $x_k + s^{*,N_k}$  and compute the agreement ratio  $\rho_k^{N_k}$  in (2.18).*



- (d) If  $\rho_k^{N_k} \geq \eta_1$ , then insert  $x_k + s^{*,N_k}$  into  $\mathcal{I}_k$ . If a point is added to the set  $\mathcal{I}_k$ , another element in  $\mathcal{I}_k$  should be removed to maintain the cardinality  $|\mathcal{I}_k| = L$ . If  $\rho_k^{N_k} < \eta_1$  and  $\mathcal{I}_k$  is inadequate in  $\mathcal{B}_k$ , improve the quality of  $\mathcal{I}_k$ .
- (e) Update the trust region radius  $\Delta_k$ :

$$\Delta_{k+1} \begin{cases} \in [\Delta_k, \gamma_2 \Delta_k], & \text{if } \rho_k^{N_k} \geq \eta_1; \\ \in [\gamma_0 \Delta_k, \gamma_1 \Delta_k], & \text{if } \rho_k^{N_k} < \eta_1 \text{ and } \mathcal{I}_k \text{ is adequate in } \mathcal{B}_k(\Delta_k); \\ = \Delta_k, & \text{otherwise.} \end{cases} \quad (2.22)$$

- (f) When a new point  $x^+$  is added into  $\mathcal{I}_k$ , if

$$\hat{\rho}_k^{N_k} = \frac{\hat{f}^{N_k}(x_k) - \hat{f}^{N_k}(x^+)}{Q_k^{N_k}(x_k) - Q_k^{N_k}(x_k + s^{*,N_k})} \geq \eta_0, \quad (2.23)$$

then  $x_{k+1} = x^+$ , otherwise,  $x_{k+1} = x_k$ .

- (g) Check whether any of the termination criteria is satisfied, otherwise repeat the loop. The termination criteria include  $\Delta_k \leq \Delta_{\text{end}}$  and hitting the maximum limit of function evaluations.

3. Evaluate and return the final solution point.

Note that in the algorithm a *successful* iteration is claimed only if the new iterate  $x_{k+1}$  satisfies the condition

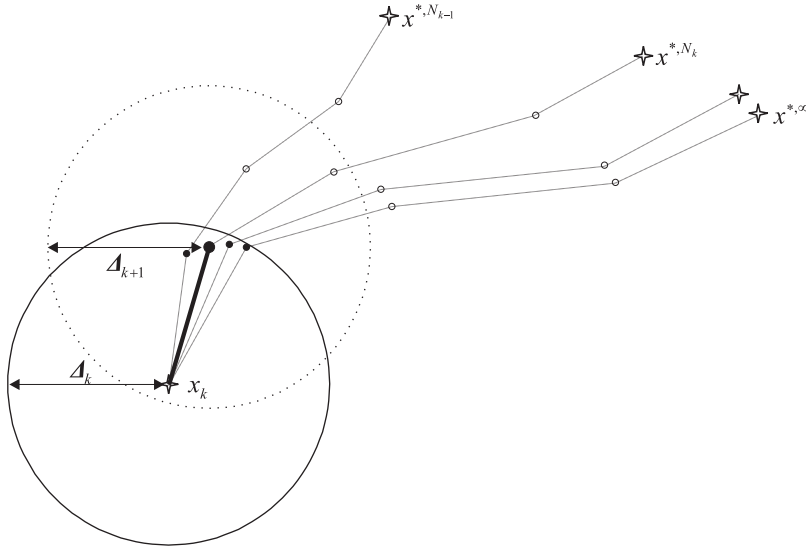
$$\hat{\rho}_k^{N_k} \geq \eta_0,$$

otherwise, the iteration is called *unsuccessful*.

### 2.3 Bayesian VNSP scheme

We have implemented the VNSP scheme within UOBYQA because UOBYQA is a self-contained algorithm that includes many nice features such as initial interpolation point design, adjustment of the trust region radii and geometry improvement of the interpolation set.

The goal of a VNSP scheme is to determine the suitable sample number  $N_k$  to be applied at iteration  $k$ . As a consequence, the algorithm, performing on averaged sample function  $\hat{f}^{N_k}$ , produces solutions  $x_k$  that converge to  $x^{*,N_\infty} = x^{*,\infty}$  (see Figure 3).



**Fig. 3** Choose the correct  $N_k$  and move the next iterate along the averaged sample function  $\hat{f}^{N_k}$ .

In our algorithm,  $Q_k^N(x_k) - Q_k^N(x_k + s^{*,N})$  is the observed model reduction, which serves to promote the next iterate (i.e., used to compute the agreement  $\rho_k^N$  in (2.18)). The key idea for the global convergence of algorithm is that, by replacing  $g_k^N$  with  $g_k^\infty$  in (2.19), we force the model reduction  $Q_k^N(x_k) - Q_k^N(x_k + s^{*,N})$  to regulate the size of  $\|g_k^\infty\|$ , and so drive  $\|g_k^\infty\|$  to zero. We present the modified ‘sufficient reduction’ criterion:

$$Q_k^N(x_k) - Q_k^N(x_k + s^{*,N}) \geq \kappa_{mdc} \|g_k^\infty\| \min \left[ \frac{\|g_k^\infty\|}{\kappa_{Qh}}, \Delta_k \right]. \quad (2.24)$$

Lemma 2 and 3 imply that increasing the replication number  $N$  lessens the bias between the quadratic models  $Q_k^N$  and  $Q_k^\infty$ , and is likely to produce a more precise step length  $s^{*,N}$ , close to  $s^{*,\infty}$ . The criterion will be eventually satisfied when  $N \rightarrow \infty$ .

To ensure the ‘sufficient reduction’ criterion (2.24) is satisfied accurately, we require

$$\begin{aligned} Pr(E_k^N) &= Pr \left( Q_k^N(x_k) - Q_k^N(x_k + s^{*,N}) < \kappa_{mdc} \|g_k^\infty\| \min \left[ \frac{\|g_k^\infty\|}{\kappa_{Qh}}, \Delta_k \right] \right) \\ &\leq \alpha_k, \end{aligned} \quad (2.25)$$

where the event  $E_k^N$  is defined as the failure of (2.24) for the current  $N$  and  $\alpha_k$  is the significance level. The probability is taken over the sample path space  $\tilde{\Omega}$ . In practice, the risk  $Pr(E_k^N)$  is difficult to evaluate because 1) it requires multiple sample paths, while the available data is limited to one sample path, and 2) we do not know the explicit form of  $Q_k^\infty$  (and hence  $g_k^\infty$ ).

By adapting knowledge from Bayesian inference, we approximate the risk value by a Bayesian posterior estimation based on the current observations  $X^N$

$$Pr(E_k^N) \approx Pr(E_k^N | X^N). \quad (2.26)$$

The value  $Pr(E_k^N | X^N)$  is thus called *Bayes risk*, which depends on a particular sample path. In the Bayesian perspective, the unknown quantities, such as  $f(x)$  and  $g_k^\infty$ , are considered as random variables, whose posterior distributions are inferred by Bayes’ rule. Given the observations  $X^N$ , we have

$$\begin{aligned} Pr(E_k^N | X^N) &= Pr \left( Q_k^N(x_k) - Q_k^N(x_k + s^{*,N}) < \kappa_{mdc} \|g_k^\infty\| \min \left[ \frac{\|g_k^\infty\|}{\kappa_{Qh}}, \Delta_k \right] \middle| X^N \right) \\ &= Pr \left( Q_k^N(x_k) - Q_k^N(x_k + s^{*,N}) < \kappa_{mdc} \|g_k^\infty | X^N\| \min \left[ \frac{\|g_k^\infty | X^N\|}{\kappa_{Qh}}, \Delta_k \right] \right). \end{aligned} \quad (2.27)$$

The left hand side  $Q_k^N(x_k) - Q_k^N(x_k + s^{*,N})$  of the inequality becomes a fixed quantity given  $X^N$ . The probability evaluation is computed with respect to the posterior distribution  $g_k^\infty | X^N$ . Here we show the fact:

**Lemma 4** *The Bayes risk  $Pr(E_k^N | X^N)$  converges to zero as  $N \rightarrow \infty$ .*

*Proof* For simplicity in notation, let  $A^N = \|g_k^\infty | X^N\| \min \left[ \frac{\|g_k^\infty | X^N\|}{\kappa_{Qh}}, \Delta_k \right]$  be a sequence of random variables, and  $b^N = Q_k^N(x_k) - Q_k^N(x_k + s^{*,N})$  be a sequence of scalars. As shown in (2.15), as  $N \rightarrow \infty$  the distribution  $g^\infty | X^N$  converges to a delta distribution.  $A^N$  also converges to a delta distribution  $A^\infty$  centered at  $\|g_k^\infty\| \min \left[ \frac{\|g_k^\infty\|}{\kappa_{Qh}}, \Delta_k \right]$ . Therefore,  $A^\infty$  is essentially a constant with zero variance. We can rewrite the Bayes risk in (2.27) as follows:

$$\begin{aligned} Pr(E_k^N | X^N) &= Pr(b^N < \kappa_{mdc} A^N) \\ &= Pr \left( (b^N - b^\infty) + \left( b^\infty - \frac{1}{2} A^\infty \right) + \left( \frac{1}{2} A^\infty - \kappa_{mdc} A^\infty \right) < \kappa_{mdc} (A^N - A^\infty) \right) \\ &= Pr \left( A^N - A^\infty > \frac{(b^N - b^\infty) + \left( b^\infty - \frac{1}{2} A^\infty \right) + \left( \frac{1}{2} A^\infty - \kappa_{mdc} A^\infty \right)}{\kappa_{mdc}} \right). \end{aligned}$$

As  $N \rightarrow \infty$ ,  $b^N - b^\infty$  converges to zero,  $b^\infty - \frac{1}{2}A^\infty \geq 0$  by Lemma 3, and  $\frac{1}{2}A^\infty - \kappa_{mdc}A^\infty$  converges to a strictly positive value because  $\kappa_{mdc} < \frac{1}{2}$ . Thus the right hand side of the inequality converges to a strictly positive value. Showing the Bayes risk converges to zero is equivalent to showing the random variable  $A^N$  converges to  $A^\infty$  in probability.

If we denote  $a^N = \mathbb{E}[A^N]$ , then  $a^N \rightarrow \mathbb{E}[A^\infty] = A^\infty$  (Theorem (3.8) p17 [10]). For a given positive value  $\varepsilon > 0$ , there exists a large enough  $N'$  such that when  $N > N'$  we have  $|a^N - A^\infty| \leq \varepsilon/2$ . If  $N > N'$ ,

$$\begin{aligned} Pr(A^N - A^\infty > \varepsilon) &\leq Pr(|A^N - A^\infty| > \varepsilon) \\ &= Pr(|A^N - a^N + a^N - A^\infty| > \varepsilon) \\ &\leq Pr(|A^N - a^N| + |a^N - A^\infty| > \varepsilon) \\ &\leq Pr(|A^N - a^N| > \varepsilon/2) \\ &\leq (2/\varepsilon)^2 var(A^N). \end{aligned}$$

The last inequality is by the Chebyshev's inequality [10]. Because  $var(A^N)$  decreases to zero, we have  $Pr(A^N - A^\infty > \varepsilon)$  decreases to zero and  $A^N$  converges to  $A^\infty$  in probability. The proof of the lemma follows.  $\square$

Lemma 4 guarantees that  $Pr(E_k^N | X^N) \leq \alpha_k$  will eventually be satisfied when  $N$  is large enough.

In Section 2.1, we derived the posterior distributions for the parameters of  $Q_k^\infty$ . These distributions can be plugged in (2.27) to evaluate the Bayes risk. However, the exact evaluation of the probability is hard to compute, especially involving the component  $\kappa_{mdc} \|g_k^\infty | X^N\| \min \left[ \frac{\|g_k^\infty | X^N\|}{\kappa_{Qh}}, \Delta_k \right]$ . Instead we use the Monte Carlo method to approximate the probability value: we generate  $M$  random samples from the posterior distribution of  $g_k^\infty | X^N$ . Based on the samples, we check the event of 'sufficient reduction' and make a count on the failed cases:  $M_{fail}$ . The probability value in (2.27) is then approximated by

$$Pr(E_k^N | X^N) \approx \frac{M_{fail}}{M}. \quad (2.28)$$

The approximation becomes accurate as  $M$  increases. Normally, we use a large value  $M = 500$ . Note that this does not require any new evaluations of the sample response function, but instead samples from the inferred Bayesian distribution  $g_k^\infty | X^N$ . We actually enforce a stricter accuracy on the fraction value for reasons that will be described below:

$$\frac{M_{fail}}{M} \leq \frac{\alpha_k}{2}. \quad (2.29)$$

A complete description of our Bayesian VNSP scheme follows:

*The VNSP scheme* At the  $k$ th iteration of the algorithm, start with  $N = N_{k-1}$ .

Loop

1. Evaluate  $N$  replications at each point  $y^j$  in the interpolation set  $\mathcal{I}_k$ , to construct the data matrix  $X^N$ . Note: data from previous iterations can be included.
2. Construct the quadratic model  $Q_k^N$  and solve the subproblem for  $x_k + s^{*,N}$ .
3. Update the value of  $\kappa_{Qh}$  by (2.21).
4. Compute the Bayesian posterior distributions for the parameters of  $Q_k^\infty$  as described above.
5. Validate the Monte Carlo estimate (2.29). If the criterion is satisfied, then stop with  $N_k = N$ ; otherwise increase  $N$ , and repeat the loop.

Since a smaller  $N_k$  is preferable, a practical approach is to sequentially allocate computing resources: starting with  $N = N_{k-1}$ , we decide to increase  $N$  or keep  $N$  by checking (2.29). If rejected,  $N$  is updated as

$$N := N \cdot \beta,$$

where  $\beta$  is an incremental factor. Otherwise, the current  $N$  is used as the sample number  $N_k$  at iteration  $k$ .

Two approximation steps (2.26) and (2.28) are employed in the computation. The following assumptions formally guarantee that risk  $Pr(E_k^N)$  is eventually approximated by the Monte Carlo fraction value  $M_{fail}/M$ .

**Assumption 4** *The difference between the risk  $Pr(E_k^N)$  and the Monte Carlo estimation value is bounded by  $\frac{\alpha_k}{2}$*

$$\left| Pr(E_k^N) - \frac{M_{fail}}{M} \right| \leq \frac{\alpha_k}{2}.$$

When  $M \rightarrow \infty$ ,  $\frac{M_{fail}}{M}$  approaches the Bayes risk  $Pr(E_k^N|X^N)$ . The assumption essentially guarantees the Bayes risk  $Pr(E_k^N|X^N)$  is a good approximation of the real risk  $Pr(E_k^N)$ . Under this assumption and the criterion (2.29), it implies

$$|Pr(E_k^N)| \leq \left| Pr(E_k^N) - \frac{M_{fail}}{M} \right| + \left| \frac{M_{fail}}{M} \right| \leq \frac{\alpha_k}{2} + \frac{\alpha_k}{2} = \alpha_k,$$

which guarantees the accuracy of the ‘sufficient reduction’ criterion (2.25). The algorithm enforces (2.29) and the convergence proof can thus use the criterion (2.25).

**Assumption 5** *The sequence of significance level values  $\{\alpha_k\}$  satisfy the property:*

$$\sum_{k=1}^{\infty} \alpha_k < \infty. \quad (2.30)$$

The assumption necessitates a stricter accuracy to be satisfied as the algorithm proceeds, which allows the use of the *Borel-Cantelli Lemma* in probability theory.

**Lemma 5 ((1st) Borel-Cantelli Lemma)** *Let  $\{E_k^N\}$  be a sequence of events, and the sum of the probabilities of  $E_k^N$  is finite, then the probability of infinitely many  $E_k^N$  occur is 0.*

*Proof* See the book by Durrett [10].  $\square$

Consider the event  $E_k^N$  to be the failure to satisfy the ‘sufficient reduction’ criterion (2.24). Given the error rate (2.25) and Assumption 5, the Borel-Cantelli Lemma provides that the events  $E_k^N$  only happen finitely many times w.p.1. Therefore, if we define  $K$  as the first successful index after all failed instances, then (2.24) is satisfied w.p.1 for all iterations  $k \geq K$ . We will use this without reference in the sequel.

Finally, we will require the following uniformity assumptions to be valid in the convergence proof.

**Assumption 6** *Given two points  $x_1, x_2 \in \mathbb{R}^n$ , the sample response difference of the two points is  $F(x_1, \xi(\omega)) - F(x_2, \xi(\omega))$ . We assume that the 2nd and 4th central moments of the sample response difference are uniformly bounded. For simplicity, we denote the  $i$ th central moment of a random variable  $Z$  as  $\varphi_i(Z)$ , that is*

$$\varphi_i(Z) = \mathbb{E}[(Z - \mathbb{E}Z)^i].$$

*Then the assumptions are, for any  $x_1, x_2 \in \mathbb{R}^n$ ,*

$$\varphi_2(F(x_1, \xi(\omega)) - F(x_2, \xi(\omega))) \leq \kappa_{\sigma^2} \quad (2.31)$$

$$\varphi_4(F(x_1, \xi(\omega)) - F(x_2, \xi(\omega))) \leq \kappa_{\sigma^4} \quad (2.32)$$

*for some constants  $\kappa_{\sigma^2}$  and  $\kappa_{\sigma^4}$ .*

Note that difference of the underlying function is the mean of the sample response difference

$$f(x_1) - f(x_2) = \mathbb{E}[F(x_1, \xi(\omega)) - F(x_2, \xi(\omega))].$$

The assumptions in fact constrain the gap between the change of the sample response function and the change of the underlying function. The 4th central moment exists for almost all statistical distributions. In Assumption 6, we consider two points  $x_1$  and  $x_2$ , because we would like to constrain their correlations (covariance, high order covariance) as well.

Moreover, for the averaged sample function  $\hat{f}^N(x)$ ,

$$\begin{aligned}
& \varphi_4 \left( \hat{f}^N(x_1, \xi(\omega)) - \hat{f}^N(x_2, \xi(\omega)) \right) \\
&= \frac{1}{N^3} \varphi_4(F(x_1, \xi(\omega)) - F(x_2, \xi(\omega))) + \frac{3(N-1)}{N^3} \varphi_2^2(F(x_1, \xi(\omega)) - F(x_2, \xi(\omega))) \\
&= \frac{1}{N^2} \left( \frac{1}{N} \varphi_4(F(x_1, \xi(\omega)) - F(x_2, \xi(\omega))) + \frac{3(N-1)}{N} \varphi_2^2(F(x_1, \xi(\omega)) - F(x_2, \xi(\omega))) \right) \\
&\leq \frac{1}{N^2} (\kappa_{\sigma^4} + 3\kappa_{\sigma^2}^2). \tag{2.33}
\end{aligned}$$

Therefore, Assumption 6 implies that the 4th central moment of the change of averaged sample function decreases quadratically fast with the sample number  $N$ .

### 3 Convergence Analysis of the Algorithm

Convergence analysis of the general model-based approach is given by Conn, Scheinberg, and Toint in [5]. Since the model-based approach is in the trust region framework, their proof of global convergence follows general ideas for the proof of the standard trust region method [22,24].

We start by showing that there is at least one stationary accumulation point. The stationary point of a function is a point at which the gradient of the function is zero. The idea is to first show that the gradient  $g_k^\infty$ , driven by the ‘sufficient reduction’ criterion (2.24), converges to zero, and then prove that  $\|\nabla f(x_k)\|$  converges to zero as well.

**Lemma 6** *Assume Assumptions 1–6 hold. If  $\|g_k^\infty\| \geq \epsilon_g$  for all  $k$  and for some constant  $\epsilon_g > 0$ , then there exists a constant  $\epsilon_\Delta > 0$  such that w.p.1,*

$$\Delta_k > \epsilon_\Delta, \text{ for all } k \geq K. \tag{3.1}$$

*Proof* Given the condition  $\|g_k^\infty\| \geq \epsilon_g$ , we will show that the corresponding  $\Delta_k$  cannot become too small, therefore, we can derive the constant  $\epsilon_\Delta$ .

Let us evaluate the following term associated with the agreement level

$$|\rho_k^{N_k} - 1| = \left| \frac{\hat{f}^{N_k}(x_k + s^{*,N_k}) - Q_k^{N_k}(x_k + s^{*,N_k})}{Q_k^{N_k}(x_k) - Q_k^{N_k}(x_k + s^{*,N_k})} \right|. \tag{3.2}$$

By Lemma 1, we compute the error bound for the numerator

$$\left| \hat{f}^{N_k}(x_k + s^{*,N_k}) - Q_k^{N_k}(x_k + s^{*,N_k}) \right| \leq \kappa_{em} \max[\Delta_k^2, \Delta_k^3]. \tag{3.3}$$

Note that when  $\Delta_k$  is small enough, satisfying the condition

$$\Delta_k \leq \min \left[ 1, \frac{\kappa_{mdc} \epsilon_g (1 - \eta_1)}{\max[\kappa_{Qh}, \kappa_{em}]} \right], \tag{3.4}$$

according to the facts  $\eta_1, \kappa_{mdc} \in (0, 1)$  and  $\|g_k^\infty\| \geq \epsilon_g$ , we deduce

$$\Delta_k \leq \frac{\|g_k^\infty\|}{\kappa_{Qh}}. \tag{3.5}$$

For the denominator in (3.2), our ‘sufficient reduction’ criterion (2.24) provides a lower bound for  $Q_k^{N_k}(x_k) - Q_k^{N_k}(x_k + s^{*,N_k})$ . When  $k \geq K$  the inequality holds w.p.1

$$Q_k^{N_k}(x_k) - Q_k^{N_k}(x_k + s^{*,N_k}) \geq \kappa_{mdc} \|g_k^\infty\| \min \left[ \frac{\|g_k^\infty\|}{\kappa_{Qh}}, \Delta_k \right] = \kappa_{mdc} \|g_k^\infty\| \Delta_k. \tag{3.6}$$

Combining (3.2), (3.3), (3.4) and (3.6), the following inequality holds w.p.1 for iteration  $k \geq K$

$$\begin{aligned}
|\rho_k^{N_k} - 1| &= \left| \frac{\hat{f}^{N_k}(x_k + s^{*,N_k}) - Q_k^{N_k}(x_k + s^{*,N_k})}{Q_k^{N_k}(x_k) - Q_k^{N_k}(x_k + s^{*,N_k})} \right| \\
&\leq \frac{\kappa_{em} \max[\Delta_k^2, \Delta_k^3]}{\kappa_{mdc} \|g_k^\infty\| \Delta_k} \\
&\leq \frac{\kappa_{em} \Delta_k}{\kappa_{mdc} \|g_k^\infty\|} \\
&\leq 1 - \eta_1.
\end{aligned} \tag{3.7}$$

The criterion  $\rho_k^{N_k} \geq \eta_1$  implies the identification of a good agreement between the model  $Q_k^{N_k}$  and the function  $\hat{f}^{N_k}$ , which will induce an increase of the trust region radius  $\Delta_{k+1} \geq \Delta_k$  (2.22). We thus have

$$\rho_k^{N_k} \geq \eta_1 \text{ valid w.p.1 for all } k \geq K.$$

According to (3.4), it is equivalent to say that  $\Delta_k$  can shrink only when

$$\Delta_k \geq \min \left[ 1, \frac{\kappa_{mdc} \epsilon_g (1 - \eta_1)}{\max[\kappa_{Qh}, \kappa_{em}]} \right].$$

We therefore derive a lower bound for  $\Delta_k$ :

$$\Delta_k > \epsilon_\Delta = \gamma_0 \min \left[ 1, \frac{\kappa_{mdc} \epsilon_g (1 - \eta_1)}{\max[\kappa_{Qh}, \kappa_{em}]} \right], \text{ for } k \geq K. \tag{3.8}$$

□

**Theorem 1** *Assume Assumptions 1–6 hold. Then, w.p.1*

$$\liminf_{k \rightarrow \infty} \|g_k^\infty\| = 0. \tag{3.9}$$

*Proof* We prove the statement (3.9) by contradiction. Suppose there is  $\epsilon_g > 0$  such that

$$\|g_k^\infty\| \geq \epsilon_g. \tag{3.10}$$

By Lemma 6, we have w.p.1,  $\Delta_k > \epsilon_\Delta$  for  $k \geq K$ .

We first show there exists only finitely many successful iterations. If not, suppose we have infinitely many successful iterations. At each successful iteration  $k \geq K$ , by (2.18), (2.24), (3.10) and  $\Delta_k > \epsilon_\Delta$ , the inequality

$$\begin{aligned}
\hat{f}^{N_k}(x_k) - \hat{f}^{N_k}(x_{k+1}) &\geq \eta_0 \left[ Q_k^{N_k}(x_k) - Q_k^{N_k}(x_k + s^{*,N_k}) \right] \\
&\geq \eta_0 \kappa_{mdc} \epsilon_g \min \left[ \frac{\epsilon_g}{\kappa_{Qh}}, \epsilon_\Delta \right]
\end{aligned} \tag{3.11}$$

holds w.p.1.

We will discuss two situations here: (a) when the limit of the sequence  $\lim_{k \rightarrow \infty} N_k = N_\infty$  is a finite number, and (b) when  $N_\infty$  is infinite. Both situations are possible in our algorithm. For simplicity, we denote  $\mathcal{S}$  as the index set of successful iterations and define

$$\epsilon_d := \eta_0 \kappa_{mdc} \epsilon_g \min \left[ \frac{\epsilon_g}{\kappa_{Qh}}, \epsilon_\Delta \right],$$

the positive reduction in right hand side of (3.11).

*Situation (a):* If  $N_\infty < \infty$ , then there exists an index  $\tilde{K} \geq K$  such that  $N_k = N_\infty$  for  $k \geq \tilde{K}$ . Since  $\{\hat{f}^{N_\infty}(x_k) | k \geq \tilde{K}\}$  is monotonically decreasing

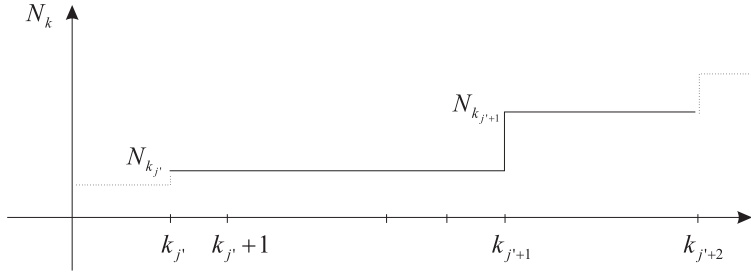
$$\begin{aligned} \hat{f}^{N_\infty}(x_{\tilde{K}}) - \hat{f}^{N_\infty}(x_{\tilde{K}+1}) &\geq \sum_{\substack{k \geq \tilde{K}, k \leq \hat{K}, \\ k \in \mathcal{S}}} \hat{f}^{N_\infty}(x_k) - \hat{f}^{N_\infty}(x_{k+1}) \\ &\geq t(\hat{K})\epsilon_d, \end{aligned} \quad (3.12)$$

where  $\hat{K}$  is a large index in  $\mathcal{S}$  and  $t(\hat{K})$  is a count number of indexes in the summation term. Since  $\hat{f}^{N_\infty}$  is bounded below (Assumption 2), we know that  $\hat{f}^{N_\infty}(x_{\tilde{K}}) - \hat{f}^{N_\infty}(x_{\tilde{K}+1})$  is a finite value. However, the right hand side goes to infinity because there are infinitely many indexes in  $\mathcal{S}$  w.p.1 ( $t(\hat{K}) \rightarrow \infty$ , as  $\hat{K} \rightarrow \infty$ ). This induces a contradiction, therefore, there are only a finite number of successful iterations.

*Situation (b):* For this situation,  $N_\infty = \infty$ . Let us define a specific subsequence of indexes  $\{k_{j'} | k_{j'} \geq K\}$  (see Figure 4), indicating where there is a jump in  $N_k$ , i.e., a truncated part of subsequence is

$$\dots < N_{k_{j'}} = N_{k_{j'}+1} = \dots = N_{k_{j'}+1-1} < N_{k_{j'}+1} = \dots.$$

Let  $\mathcal{S}'$  be a subset of  $\{k_{j'}\}$ , including  $k_{j'}$  if there is at least one successful iteration in  $\{k_{j'}, \dots, k_{j'}+1-1\}$ .



**Fig. 4** Illustration of the subsequence  $\{k_{j'}\}$

This implies

$$x_{k_{j'}+1} \begin{cases} \neq x_{k_{j'}}, & \text{for } k_{j'} \in \mathcal{S}'; \\ = x_{k_{j'}} \text{ (unchanged)}, & \text{for } k_{j'} \notin \mathcal{S}'. \end{cases}$$

For  $k_{j'} \in \mathcal{S}'$ , sum the inequality (3.11) for  $k \in \{N_{k_{j'}}, \dots, N_{k_{j'}+1}-1\}$  to derive

$$\begin{aligned} \hat{f}^{N_{k_{j'}}}(x_{k_{j'}}) - \hat{f}^{N_{k_{j'}}}(x_{k_{j'}+1}) &\geq \sum_{\substack{k \geq k_{j'}, k \leq k_{j'}+1-1 \\ k \in \mathcal{S}'}} \hat{f}^{N_{k_{j'}}}(x_k) - \hat{f}^{N_{k_{j'}}}(x_{k+1}) \\ &\geq \epsilon_d. \end{aligned} \quad (3.13)$$

We want to quantify the difference between  $\hat{f}^{N_{k_{j'}}}(x_{k_{j'}}) - \hat{f}^{N_{k_{j'}}}(x_{k_{j'}+1})$  and  $f(x_{k_{j'}}) - f(x_{k_{j'}+1})$ . The idea behind this is that moving from  $x_{k_{j'}}$  to  $x_{k_{j'}+1}$ , the function  $\hat{f}^{N_{k_{j}'}}$  decreases, and so does the underlying function  $f$ . Since infinitely many decrement steps for  $f$  are impossible, we derive a contradiction.

Define the event  $\hat{E}_{k_{j'}}$  as the occurrence of  $\hat{f}^{N_{k_{j'}}}(x_{k_{j'}}) - \hat{f}^{N_{k_{j'+1}}}(x_{k_{j'+1}}) \geq \epsilon_d$  while  $f(x_{k_{j'}}) - f(x_{k_{j'+1}}) \leq \frac{\epsilon_d}{2}$ . The probability of event

$$\begin{aligned}
& Pr\left(\hat{E}_{k_{j'}}\right) \\
& \leq Pr\left(\left(\hat{f}^{N_{k_{j'}}}(x_{k_{j'}}) - \hat{f}^{N_{k_{j'+1}}}(x_{k_{j'+1}})\right) - \left(f(x_{k_{j'}}) - f(x_{k_{j'+1}})\right) \geq \frac{\epsilon_d}{2}\right) \\
& \leq Pr\left(\left|\left(\hat{f}^{N_{k_{j'}}}(x_{k_{j'}}) - \hat{f}^{N_{k_{j'+1}}}(x_{k_{j'+1}})\right) - \left(f(x_{k_{j'}}) - f(x_{k_{j'+1}})\right)\right| \geq \frac{\epsilon_d}{2}\right) \\
& = Pr\left(\left(\left(\hat{f}^{N_{k_{j'}}}(x_{k_{j'}}) - \hat{f}^{N_{k_{j'+1}}}(x_{k_{j'+1}})\right) - \left(f(x_{k_{j'}}) - f(x_{k_{j'+1}})\right)\right)^4 \geq \left(\frac{\epsilon_d}{2}\right)^4\right) \\
& \leq \frac{16}{\epsilon_d^4} \cdot \mathbb{E}\left[\left(\hat{f}^{N_{k_{j'}}}(x_{k_{j'}}) - \hat{f}^{N_{k_{j'+1}}}(x_{k_{j'+1}})\right) - \left(f(x_{k_{j'}}) - f(x_{k_{j'+1}})\right)\right]^4 \\
& = \frac{16}{\epsilon_d^4} \cdot \varphi_4\left(\hat{f}^{N_{k_{j'}}}(x_{k_{j'}}) - \hat{f}^{N_{k_{j'+1}}}(x_{k_{j'+1}})\right) \\
& \leq \frac{16\left(\kappa_{\sigma^4} + 3\kappa_{\sigma^2}^2\right)}{\epsilon_d^4(N_{k_{j'}})^2}.
\end{aligned}$$

The third inequality is due to Markov's inequality [10]. The random quantity  $\hat{f}^{N_{k_{j'}}}(x_{k_{j'}}) - \hat{f}^{N_{k_{j'+1}}}(x_{k_{j'+1}})$  has mean value  $f(x_{k_{j'}}) - f(x_{k_{j'+1}})$ . The last inequality is due to the implication of Assumption 6, see (2.33).

The result implies that probability of the event  $\hat{E}_k$  decreases quadratically fast with  $k$ . Since the sum of the probability values is finite

$$\sum_{\substack{j'=1 \\ k_{j'} \in \mathcal{S}'}}^{\infty} Pr\left(\hat{E}_{k_{j'}}\right) \leq \sum_{\substack{j'=1 \\ k_{j'} \in \mathcal{S}'}}^{\infty} \frac{16\left(\kappa_{\sigma^4} + 3\kappa_{\sigma^2}^2\right)}{\epsilon_d^4(N_{k_{j'}})^2} < \infty,$$

applying the Borel-Cantelli Lemma again, the event  $\hat{E}_{k_{j'}}$  occurs only finitely many times w.p.1. Thus, there exists an index  $\bar{K}$ , such that

$$f(x_{k_{j'}}) - f(x_{k_{j'+1}}) \geq \frac{\epsilon_d}{2}, \text{ for all } \{k_{j'} | k_{j'} \geq \bar{K}, k_{j'} \in \mathcal{S}'\} \text{ w.p.1.}$$

Playing the same trick as before, by summing over all  $k_{j'} \geq \bar{K}$ , we derive that w.p.1

$$\begin{aligned}
f(x_{\bar{K}}) - f(x_{\bar{K}+1}) & \geq \sum_{\substack{k_{j'} \geq \bar{K}, k_{j'} \leq \hat{K} \\ k_{j'} \in \mathcal{S}'}} f(x_{k_{j'}}) - f(x_{k_{j'+1}}) \\
& \geq t(\hat{K}) \frac{\epsilon_d}{2}.
\end{aligned} \tag{3.14}$$

The left hand side is a finite value, but the right hand side goes to infinity. This contradiction also shows that the number of successful iterations is finite.

Combining the two situations above, we must have infinitely many unsuccessful iterations when  $k$  is sufficiently large. As a consequence, the trust region radius  $\Delta_k$  decreases to zero

$$\lim_{k \rightarrow \infty} \Delta_k = 0,$$

which contradicts the statement that  $\Delta_k$  is bounded below (3.8). Thus (3.10) is false, and the theorem is proved.  $\square$



**Theorem 2** Assume Assumptions 1–6 hold. If

$$\liminf_{j \rightarrow \infty} \|g_{k_j}^\infty\| = 0 \text{ w.p.1} \quad (3.15)$$

holds for a subsequence  $\{k_j\}$ , then we also have

$$\liminf_{j \rightarrow \infty} \|\nabla f(x_{k_j})\| = 0 \text{ w.p.1.} \quad (3.16)$$

*Proof* Due to the fact  $\lim_{j \rightarrow \infty} \Delta_{k_j} = 0$ , Lemma 1 guarantees that the difference between  $\|g_{k_j}^\infty\|$  and  $\|\nabla f(x^{k_j})\|$  is small. Thus the assertion (3.16) follows. The details of the proof refer to Theorem 11 in [5].  $\square$

**Theorem 3** Assume Assumptions 1–6 hold. Every limit point  $x^*$  of the sequence  $\{x_k\}$  is stationary.

*Proof* The procedure of proof is essentially the same as given for Theorem 12 in [5]. However, we use the ‘sufficient reduction’ inequalities (3.12) when  $N_\infty$  is finite and (3.14) when  $N_\infty$  is infinite.  $\square$

## 4 Numerical Results

We apply the new UOBYQA algorithm implementing the VNSP scheme to several numerical examples. The noisy test functions are altered from deterministic functions with artificial randomness.

The first numerical function we employed was the well-known extended Rosenbrock function. The random term was added only to the first component of the input variable. Define

$$\hat{x}(x, \xi(\omega)) := (x_{(1)}\xi(\omega), x_{(2)}, \dots, x_{(n)})$$

and the corresponding function becomes

$$F(x, \xi(\omega)) = \sum_{i=1}^{n-1} 100(\hat{x}_{(i+1)} - \hat{x}_{(i)}^2)^2 + (\hat{x}_{(i)} - 1)^2. \quad (4.1)$$

We assume  $\xi(\omega)$  is a normal variable centered at 1:

$$\xi(\omega) \sim N(1, \sigma^2).$$

As a general setting, the initial and end trust region radius  $\Delta_0, \Delta_{end}$  were set to 2 and  $1.0e - 5$ , respectively. Implementing the algorithm required a starting value  $N_0 = 3$ , which was used to estimate the initial sample mean and sample covariance matrix. We believe such a value is the minimum required for reasonable estimates. Larger values of  $N_0$  would in most cases lead to wasted evaluations.  $M = 500$  (see (2.28)) trials samples were generated to evaluate the Bayes probability (2.27) in the VNSP procedure. To satisfy Assumption 5, the sequence  $\{\alpha_k\}$  was pre-defined as

$$\alpha_k = 0.5 \times (0.98)^k.$$

Table 1 presents the details about a single-run of the new algorithm on the two-dimensional Rosenbrock function with  $\sigma^2 = 0.01$ . The starting point was chosen to be (-1,1.2), and the maximum number of function evaluations was 10000. We recorded the iteration number  $k$  when there was a change in  $N_k$ . For example,  $N_k$  remained at 3 in iterations 1–19, and  $N_k$  changed to 4 at iteration 20. Since in the first 19 iterations, the averaged sample function was  $\hat{f}^3$ , all the steps were taken regarding  $\hat{f}^3$  as the objective function. Therefore, it was observed that the iterates  $x_k$  moved toward the solution  $x^{*,3}$  of the averaged sample problem (1.3) with  $N = 3$ . In Table 2 we present the corresponding sample-path solution of the optimization problem (1.3). For example,  $x^{*,3} = (0.5415, 0.2778)$ . Note

**Table 1** The performance of the new algorithm for the noisy Rosenbrock function, with  $n = 2$  and  $\sigma^2 = 0.01$ .

Iteration $k$	$N_k$	FN	$x_k$	$\hat{f}^{N_k}(x_k)$	$\Delta_k$
0	3	3	(-1.0000,1.2000)	11.7019	2.0
19	3	81	(0.5002,0.2449)	0.3616	0.1
20	4	91	(0.5002,0.2449)	0.4904	0.05
21	5	102	(0.5208,0.2904)	0.4944	0.02
22	22	226	(0.5082,0.2864)	0.4018	0.02
23	22	248	(0.5082,0.2864)	0.4018	0.02
24	30	326	(0.5082,0.2864)	0.5018	0.02
29	30	476	(0.4183,0.1862)	0.4447	0.04
30	113	1087	(0.4328,0.1939)	0.4290	0.02
31	113	1200	(0.4328,0.1939)	0.4290	0.02
32	221	1848	(0.4328,0.1939)	0.4437	0.02
33	604	4750	(0.4328,0.1939)	0.4601	0.01
35	604	5958	(0.4276,0.1837)	0.4569	0.0125
36	845	8249	(0.4197,0.1774)	0.4556	0.0101
37	1183	10277	(0.4172,0.1760)	0.4616	0.0101

that, in order to derive the solution to  $f$  in the two dimensional problem, the noisy Rosenbrock function was rearranged as

$$\begin{aligned} f(x) &= \mathbb{E} \left[ 100(\hat{x}_{(2)} - \hat{x}_{(1)}^2)^2 + (\hat{x}_{(1)} - 1)^2 \right] \\ &= 100x_{(2)}^2 + 1 - 2x_{(1)}\mathbb{E}[\xi] + (-200x_{(2)}x_{(1)}^2 + x_{(1)}^2)\mathbb{E}[\xi^2] + 100x_{(1)}^4\mathbb{E}[\xi^4]. \end{aligned}$$

By plugging the values  $\mathbb{E}[\xi] = 1$ ,  $\mathbb{E}[\xi^2] = 1.01$ , and  $\mathbb{E}[\xi^4] = 1.0603$ , we obtained the solution  $x^{*,\infty} = (0.4162, 0.1750)$ , which was different from the deterministic Rosenbrock solution  $(1, 1)$ . For different  $N_k$ , the averaged function  $\hat{f}^{N_k}$  might vary greatly. In Table 1, we observe that  $x_{19} = x_{20} = (0.5002, 0.2449)$ . The value of  $\hat{f}^{N_{19}}(x_{19})$  is 0.3616, while the value of  $\hat{f}^{N_{20}}(x_{20})$  is 0.4904. It shows that the algorithm actually worked on objective functions with increasing accuracy.

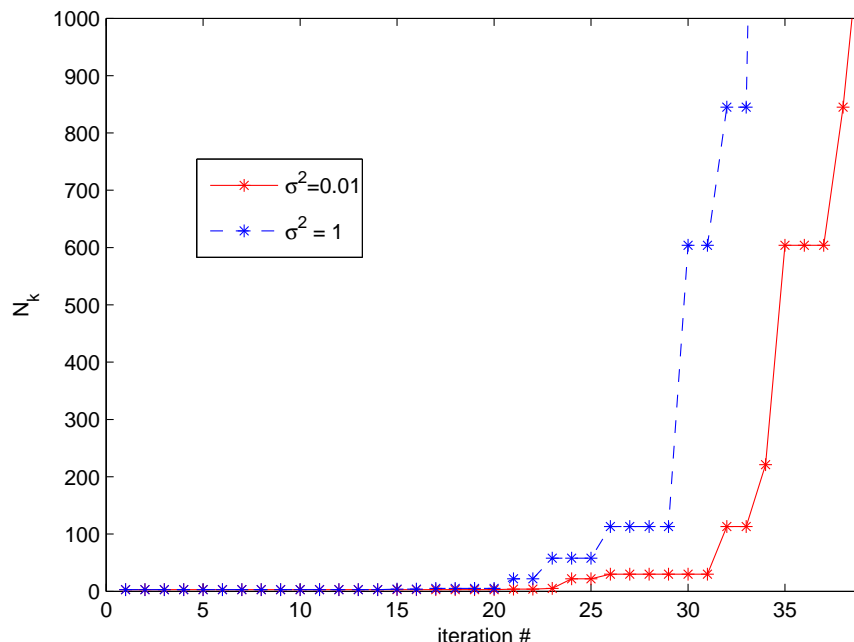
**Table 2** Averaged sample-path solution with different sample number  $N$

$N$	$x^{*,N}$	$\hat{f}^{N_k}(x^{*,N})$
3	(0.5415,0.2778)	0.3499
4	(0.4302,0.1922)	0.4412
5	(0.4218,0.1936)	0.4395
22	(0.4695,0.2380)	0.3892
30	(0.4222,0.1896)	0.4446
113	(0.4423,0.2027)	0.4286
221	(0.4331,0.1910)	0.4427
604	(0.4226,0.1798)	0.4567
845	(0.4236,0.1807)	0.4556
1183	(0.4174,0.1761)	0.4615
$\infty$	(0.4162,0.1750)	0.4632

As shown in Table 1, the algorithm used a small  $N_k$  to generate new iterates in the earlier iterations. Only 476 function evaluations were applied for the first 29 iterations. This implies that when noisy effects were small compared to the large change of function values, the basic operation of the method was unchanged and  $N_k = N_0$  samples were used. As the algorithm proceeded, the demand for accuracy increased, therefore,  $N_k$  increased as well as the total number of function evaluations. We obtained very good solutions. At the end of the algorithm, we generated a solution  $x_{37} = (0.4172, 0.1760)$ , which is close to the averaged sample-path solution  $x^{*,N=1183} = (0.4174, 0.1761)$  and is better than the solution  $x^{*,N=845} = (0.4236, 0.1807)$ . In a standard sample-path optimization method, assuming that there are around 40 iterations in the algorithm, we need  $845 \times 40 = 33800$

function evaluations for the solution  $x^{*,N=845}$  and  $1183 \times 40 = 43720$  for the solution  $x^{*,N=1183}$ . Our algorithm indeed saved a significant amount of function operations.

To study the changes of  $N_k$ , in Figure 5, we plot  $N_k$  against the iteration number for two problems. One is a high volatility case with  $\sigma^2 = 1$  and the other is a low volatility case with  $\sigma^2 = 0.01$ . In both problems,  $N_k$  was 3 for the first 20 iterations, when the noise is not the dominating factor. In the later iterations, the noise became significant and we observe that the demand for  $N_k$  increased faster for the high volatility case. If we restricted the total function evaluations to be 10000, the high volatility case resulted in an early termination at the 34th iteration.



**Fig. 5** Compare changes of  $N_k$  with different levels of noise

We applied the algorithm to both 2 and 10 dimensional problems. Increasing the dimension significantly increased computational burden. The problem with dimension  $n = 10$  is already very hard to tackle. Even in the deterministic case, the standard UOBYQA requires around 1400 iterations to terminate at  $\Delta_{end} = 0.0001$ . In Table 3, we record a summary of the algorithm applied to the Rosenbrock function with different dimensions and noise levels. For comparisons, we include the result of the standard sample-path methods with fixed numbers of samples: 10, 100, and 1000. The statistical results are based on 10 replications of the algorithm. The variance of the error is small, showing that the algorithm was generally stable. For  $n = 10$  and  $\sigma^2 = 1$ , we notice a big mean error 2.6 and a relatively small variance of error 0.10. This is due to the earlier termination of the algorithm when  $\sigma^2$  is large (we used a limit of 20000 function evaluations in this case). There are two reasons why the standard sample-path methods yield relatively larger errors. 1) Methods SP(10) and SP(100) do not provide accurate averaged sample functions  $\hat{f}^N$ . 2) For a large sample number  $N$ , the iteration number of the algorithm is limited. For example, we can expect SP(100) is limited to 200 iterations and SP(1000) is limited to 20 iterations. Increasing the total number of function evaluations can significantly improve the performance of the sample path optimization methods. For example, if we allow 2,000,000 total function evaluations for the 10 dimensional case and the noise level  $\sigma^2 = 1$ , the mean error of SP(100) and SP(1000) are 1.6, 7.5, respectively. The VSNP method performs better than this.

**Table 3** Statistical summary

n	Noise level $\sigma^2$	VNSP		SP(10)	SP(100)	SP(1000)
		Mean error	Variance of error	Mean error	Mean error	Mean error
2	0.01	1.1e-5	1.2e-5	0.035	0.0045	7.9e-5
2	0.1	8.9e-5	3.3e-5	0.079	0.0067	4.2e-4
2	1	1.1e-4	8.2e-5	0.098	0.0088	8.9e-4
10	0.01	0.054	0.067	0.44	28	120
10	0.1	0.087	0.060	2.1	44	129
10	1	2.6	0.10	14	32	145

For another test example, we refer back to the toy example in Section 1. The objective function is only affected by ‘white noise’

$$F(x, \xi(\omega)) = \phi(x) + \xi(\omega).$$

We will show  $N_k$  is unchanged for every iteration, that is,  $N_1 = N_2 = \dots = N_\infty$ . At iteration  $k$ , the function outputs at points  $y^j$  in  $\mathcal{I}_k$  are entirely correlated. As a result, the sample covariance matrix  $\hat{\Sigma}$  (2.8) is a rank-one matrix, whose elements are all identical  $\hat{\Sigma}(i, j) = a$ ,  $i, j = 1, 2, \dots, L$ , where  $a = \text{var}[(\xi_1, \dots, \xi_{N_k})]$ . Thus, the matrix can be decomposed as

$$\hat{\Sigma} = \mathbf{1} \cdot a \cdot \mathbf{1}^T. \quad (4.2)$$

Plug (4.2) into (2.15), we obtain the posterior covariance of  $g_k^\infty$

$$\text{cov}(g_k^\infty | X^N) = (\mathbf{g} \cdot \mathbf{1})^T \cdot a \cdot (\mathbf{g} \cdot \mathbf{1}) = (\mathbf{0})^T \cdot a \cdot \mathbf{0} = \mathbf{0}_{L \times L},$$

which implies  $g_k^\infty$  is not random and  $g_k^\infty = g_k^{N_k}$ . As a consequence, in the VNSP scheme, the mechanism will not increase  $N_k$  because the criterion (2.24) is always satisfied.

The fact  $\mathbf{g} \cdot \mathbf{1} = \sum_{j=1}^L g_j = \mathbf{0}$  is a property of Lagrange functions. The proof is simple - the sum of Lagrange functions  $\sum_{j=1}^L l_j(x)$  is the unique quadratic interpolant of a constant function  $\hat{g}(x) = 1$  at the points  $y^j$ , because  $\sum_{j=1}^L l_j(y^j) = 1 = \hat{g}(y^j)$ ,  $j = 1, \dots, L$ . Therefore, the gradient of the interpolant  $\sum_{j=1}^L g_j = \mathbf{0}$ .

In practice, the behavior of the toy example occurs rarely. We present it here to show that our algorithm indeed checks the uncertainty of each iterate  $x_k$ , but not that of objective value  $\hat{f}^{N_k}(x_k)$ .

## 5 Conclusions

This paper proposes and analyzes a variable number sample-path scheme for optimization of noisy functions. The VNSP scheme applies analytical Bayesian inference to determine an appropriate number of samples  $N_k$  to use in each iteration. For the purpose of convergence, we only allow  $N_k$  to be non-decreasing. As the iterations progress, the algorithm automatically increases  $N_k$  and thus adaptively produces more accurate objective function evaluations. The key idea of choosing an appropriate  $N_k$  in the VNSP scheme is to test the Bayes risk of satisfying a ‘sufficient reduction’ criterion. Under appropriate assumptions, the global convergence of the algorithm is guaranteed:

$$\lim_{k \rightarrow \infty} x_k = x^{*, N_\infty} = x^{*, \infty}.$$

UOBYQA implements the Moré and Sorensen method [23] to handle the trust region subproblem. Extending our algorithm to constrained optimization problems requires corresponding tools to solve a constrained subproblem

$$\min_{x \in \mathcal{S}} Q_k(x), \text{ s.t. } \|x - x_k\| \leq \Delta_k, x \in \mathcal{S},$$

where  $\mathcal{S}$  is a feasible set for  $x$ . An efficient derivative free algorithm for obtaining a global solution to the problem is not yet available. On the other hand, the techniques outlined here have potential

even for constrained optimization problems, because they are couched in standard trust region theory which has become prevalent in algorithm design for such problems.

The VNSP scheme can be generalized to other model-based algorithms, such as the WEDGE algorithm. Our modifications are not intended to be applied to linear model based algorithms, since linear models are more sensitive to noise. In a stochastic situation, quadratic models are robust against noise and preferable to use. Some algorithms may use less than  $L = \frac{1}{2}(n+1)(n+2)$  initial points to construct quadratic models. For example, NEWUOA uses  $2n+1$  points for the initial model and updates the models while minimizing the change in Frobenius norm of the curvature. The VNSP scheme should be altered to accommodate this different approach, but this will require further analysis.

The new algorithm has broad practical applications. For example, we have successfully applied it to seek the optimal design of an interstitial coaxial antenna, which is used in microwave ablation treatment for hepatic cancer [29]. Since the permittivity and electric conductivity vary among patients, the optimal design is required to perform well in the averaged sense. Further applications will be addressed in future work.

**Acknowledgements** The authors would like to thank three anonymous referees for detailed comments that helped improve the accuracy and correctness of the paper.

## References

1. H.-C. Chen, C.-H. Chen, and E. Yucesan. An asymptotic allocation for simultaneous simulation experiments. In *Proceedings of the 1999 Winter Simulation Conference*, pages 359–366, 1999.
2. S. Chick. Personal communication.
3. S. E. Chick and K. Inoue. New procedures to select the best simulation system using common random numbers. *Management Science*, 47(8):1133–1149, 2001.
4. S. E. Chick and K. Inoue. New two-stage and sequential procedures for selecting the best simulated system. *Operations Research*, 49:1609–1624, 2003.
5. A. R. Conn, K. Scheinberg, and P. L. Toint. On the convergence of derivative-free methods for unconstrained optimization. *Approximation Theory and Optimization, Tributes to M. J. D. Powell*, edited by M. D. Buhmann and A. Iserles, pages 83–108, 1996.
6. A. R. Conn and Ph. L. Toint. An algorithm using quadratic interpolation for unconstrained derivative free optimization. In G. Di Pillo and F. Giannessi, editors, *Nonlinear Optimization and Applications*, pages 27–47. Plenum Press, New York, 1996.
7. A. P. Dawid. Some matrix-variate distribution theory: Notational considerations and a bayesian application. *Biometrika*, 68(1):265–74, 1981.
8. G. Deng and M. C. Ferris. Adaptation of the UOBYQA algorithm for noisy functions. In L. F. Perrone, F. P. Wieland, B. G. Lawson J. Liu, D. M. Nicol, and R. M. Fujimoto, editors, *Proceedings of the 2006 Winter Simulation Conference*, pages 312–319.
9. G. Deng and M. C. Ferris. Extension of the DIRECT optimization algorithm for noisy functions. 2007. Submitted to 2007 Winter Simulation Conference.
10. R. Durrett. *Probability: Theory and Examples*. Duxbury Press, Pacific Grove, California, 3rd edition, 2004.
11. M. C. Ferris, T. S. Munson, and K. Sinapiromsaran. A practical approach to sample-path simulation optimization. In J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, editors, *Proceedings of the 2000 Winter Simulation Conference*, pages 795–804, Orlando, Florida, 2000. Omnipress.
12. M. Fu. Optimization via simulation: A review. *Annals of Operations Research*, 53:199–248, 1994.
13. A. Gosavi. *Simulation-Based Optimization: Parametric Optimization Techniques and Reinforcement Learning*. Kluwer Academic Publishers, Norwell, MA, USA, 2003.
14. G. Gürkan, A. Yonca Özge, and S. M. Robinson. Sample-path solution of stochastic variational inequalities, with applications to option pricing. In D. T. Brunner J. M. Charnes, D. J. Morrice and J. J. Swain, editors, *Proceedings of the 1996 Winter Simulation Conference*, pages 337–344, 1996.
15. G. Gürkan, A. Yonca Özge, and S. M. Robinson. Solving stochastic optimization problems with stochastic constraints: An application in network design. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, editors, *Proceedings of the 1999 Winter Simulation Conference*, pages 471–478, 1999.
16. T. Homem-de-Mello. Variable-sample methods for stochastic optimization. *ACM Transactions on Modeling and Computer Simulation*, 13(2), 2003.
17. T. Homem-de-Mello. On rates of convergence for stochastic optimization problems under non-I.I.D. sampling. *submitted for publication*, 2006.
18. S.-H. Kim and B. L. Nelson. Selecting the best system: Theory and methods. In *Proceedings of the 2003 Winter Simulation Conference*, pages 101–112, 2003.

19. A. J. Kleywegt, A. Shapiro, and T. Homem-de-Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12:479–502, 2001.
20. A. Law and W. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, New York, third edition, 2000.
21. M. Marazzi and J. Nocedal. Wedge trust region methods for derivative free optimization. *Mathematical Programming*, 91:289–305, 2002.
22. J. J. Moré. Recent developments in algorithms and software for trust region methods. In *Mathematical Programming: The State of the Art*, pages 258–287. Springer Verlag, 1983.
23. J. J. Moré and D. C. Sorensen. Computing a trust region step. *SIAM Journal on Scientific and Statistical Computing*, 4:553–572, 1983.
24. J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, second edition, 2006.
25. E. L. Plambeck, B. R. Fu, S. M. Robinson, and R. Suri. Throughput optimization in tandem production lines via nonsmooth programming. In J. Schoen, editor, *Proceedings of the 1993 Summer Computer Simulation Conference*, pages 70–75, San Diego, California, 1993. Society for Computer Simulation.
26. E. L. Plambeck, B. R. Fu, S. M. Robinson, and R. Suri. Sample-path optimization of convex stochastic performance functions. *Mathematical Programming*, 75:137–176, 1996.
27. M. J. D. Powell. UOBYQA: Unconstrained optimization by quadratic approximation. *Mathematical Programming*, 92:555–582, 2002.
28. M. J. D. Powell. The NEWUOA software for unconstrained optimization with derivatives. *DAMTP Report 2004/NA05, University of Cambridge*, 2004.
29. P. Prakash, G. Deng, M. C. Converse, J. G. Webster, David M. Mahvi, and M. C. Ferris. Design optimization of a robust sleeve antenna for hepatic microwave ablation. Working paper.
30. S. M. Robinson. Analysis of sample-path optimization. *Mathematics of Operations Research*, 21:513–528, 1996.
31. A. Ruszczyński and A. Shapiro. Stochastic programming models. In *Stochastic Programming*, pages 1–63. Elsevier, Amsterdam, Netherlands, 2003.
32. A. Shapiro. Asymptotic behavior of optimal solutions in stochastic programming. *Mathematics of Operations Research*, 18:829–845, 1993.
33. A. Shapiro. Statistical inference of stochastic optimization problems. In S. P. Uryasev, editor, *Probabilistic Constrained Optimization: Methodology and Applications*, pages 91–116. Kluwer Academic Publishers, 2000.
34. A. Shapiro. Monte Carlo sampling approach to stochastic programming. In J. P. Penot, editor, *ESAIM: Proceedings*, volume 13, pages 65–73, 2003.
35. A. Shapiro. Monte Carlo sampling methods. In *Stochastic Programming*, pages 1–63. Elsevier, Amsterdam, Netherlands, 2003.
36. A. Shapiro and T. Homem-de-Mello. A simulation-based approach to two-stage stochastic programming with recourse. *Mathematical Programming: Series A*, 81(3):301–325, 1998.