Variable-Number Sample-Path Optimization*

Geng Deng† Michael C. Ferris‡

June 28, 2006

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 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.

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* [9, 10, 17] is an emerging field which integrates optimization techniques into the 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

†Department of Mathematics, University of Wisconsin, 480 Lincoln Drive, Madison, WI 53706

‡Computer Sciences Department, University of Wisconsin, 1210 West Dayton Street, Madison, WI 53706

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

$$\min_{x\in\mathbb{R}^n} F(x) = \mathbb{E}[f(x,\omega)] = \int_\Omega f(x,\omega)P(d\omega). \tag{1.1}$$

Here ω is a random variable defined in the probability space $(\Omega, \mathcal{F}, P)$. The sample response function $f(x,\omega)$ takes two inputs, the simulation parameters $x \in \mathbb{R}^n$ and a random sample from the distribution of ω. A basic assumption requires that the sample function $f(x,\omega)$ is measurable, so that the expectation $\mathbb{E}[f(x,\omega)]$ exists for every x. Given a random sample realization ω_i, $f(x,\omega_i)$ 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.

The *sample-path method* is a well-recognized method in simulation-based optimization [8, 11, 12, 21, 22, 25]. It is sometimes called the *Monte Carlo Sampling Approach* [28] or the *Sample Average Approximation Method* [13, 14, 16, 27]. 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

$$F(x) \approx \bar{f}^N(x) := \frac{1}{N}\sum_{i=1}^{N} f(x,\omega_i), \tag{1.2}$$

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

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

which serves as a substitute for (1.1). The 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 tools) to be used. Convergence proofs of the sample-path method are given in [25, 26]. Suppose there is a unique solution x^* to the problem (1.1), then under assumptions such as the sequence of functions $\{\bar{f}^N\}$ epiconverges to the limit function $\bar{f}^\infty = F(x)$ the optimal solution sequence $\{x^{*,N}\}$ converges to x^* almost surely. See Figure 1 for the illustration of the sample-path optimization method.

Our motivation for the 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 $\bar{f}^N(x_k)$ is required (with N large). The new VNSP

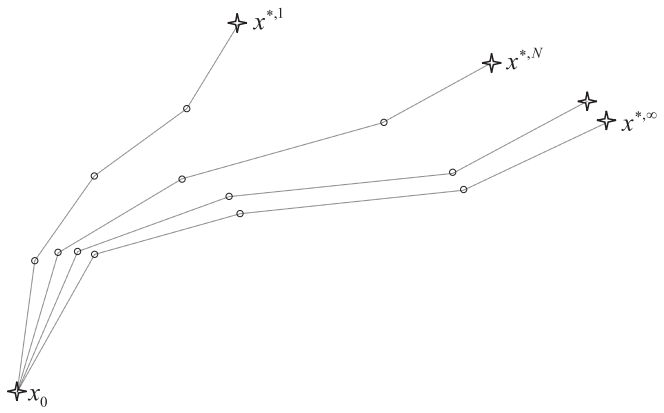

Figure 1: Mechanism of the sample-path optimization method. Starting from x_0, for a given N, a deterministic algorithm is applied to solve the sample-path problem. The sequence of solutions $\{x^{*,N}\}$ converges to the true solution $x^{*,\infty} = x^*$.

scheme is designed to generate different numbers of sample paths (N) at each iteration. Denoting N_k as the number of sample paths at iteration k, the VNSP scheme integrates Bayesian techniques to establish satisfactory critera for N_k, which accordingly ensure the accuracy of the approximation of $\bar{f}^N(x)$ 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.

Another 'variable-sample' scheme for sample-path optimization was proposed by Homem-de-Mello in [13]. In his work, the sequence of $\{N_k\}$ is pre-scheduled to increase at a certain rate, with $\lim_{k\to\infty} N_k = \infty$. The rate assumption for the sequence is crucial to show the convergence property of the algorithm. Our VNSP scheme is significantly different: N_k in our scheme is validated based on an instant inspection of the current step. As a consequence, $\{N_k\}$ is a non-decreasing sequence with the limit value N_∞, being either finite or infinite. Here is a toy example showing that the limit sample path number N_∞ in our algorithm can be finite. Consider a simulation system only with 'white noise':

$$f(x, \omega) = g(x) + \omega,$$

where $g(x)$ is a deterministic function and $\omega \sim N(0, \sigma^2)$. As a result, the minimizer of each piece $f(x, \omega_i) = g(x) + \omega_i$ coincides with the minimizer of $F(x) = g(x)$. The sample-path method thus yields solutions to (1.2): $x^{*,1} = x^{*,2} = \cdots = x^{*,\infty}$. In this case, our VNSP scheme turns out to generate a constant sequence of sample path numbers $N_k : N_1 = N_2 = \cdots = N_\infty$, while for the toy example, the 'variable-sample' scheme in [13] still necessitates $\lim_{k\to\infty} N_k = \infty$.

We apply Powell's UOBYQA algorithm (*Unconstrained Optimization BY Quadratic Approximation*) [23] as our base sample-path optimization solver. The algorithm is a derivative-free approach, thus is a good fit for the optimization problem (1.3).

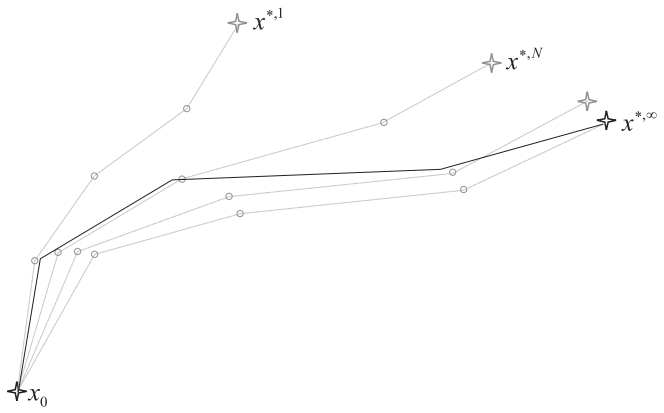

Figure 2: Mechanism of the new sample-path method with the VNSP scheme. Starting from x_0, the algorithm generates its iterates across different averaged sample paths. In an intermediate iteration k, it first computes a satisfactory N_k which guarantees certain level of accuracy, then an optimization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

It is designed to solve nonlinear problems with a moderate number of dimensions. The general structure of UOBYQA follows a *model-based approach* [4, 5], which constructs a chain of local quadratic models that approximate the objective function. The method is an iterative algorithm in a trust region framework [20], 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 includes WEDGE [18] and NEWUOA [24].

Sections of the paper are arranged as follows. In Section 2.1 we will provide the outline of the new algorithm, with a realization of the VNSP scheme. In Section 2.2, we describe the Bayesian VNSP scheme to determine the suitable value of N_k at iteration k. In Section 3, we analyze the global convergence properties of the algorithm. Finally, in Section 4, we discuss several numerical results on test functions.

2 The Extended UOBYQA Algorithm

As we have mentioned before, UOBYQA is essentially in the framework of a model-based approach. The idea of the model-based approach centers around 'interpolation model construction and updating'. A general framework of this approach is given by Conn and Toint in [5], and convergence analysis is presented in [4]. In our extension of UOBYQA, we inherent several basic assumptions of the objective function from [4].

Assumption 1. *The underlying function $F(x)$ is twice continuously differentiable*

and its gradient and Hessian are uniformly bounded in $\mathbb{R}^n$. *There exist constants* $\kappa_{Fg} > 0$ *and* $\kappa_{Fh} > 0$, *such that for each* $x \in \mathbb{R}^n$, *the inequalities hold*

$$\|\nabla F(x)\| \leq \kappa_{Fg}$$

and

$$\|\nabla^2 F(x)\| \leq \kappa_{Fh}.$$

Assumption 2. *The underlying function* $F(x)$ *and each sample response function* $f(x, \omega_i)$ *are bounded below on* $\mathbb{R}^n$.

We implement the VNSP scheme based on UOBYQA because it is a self-contained algorithm that also includes many nice features such as initial interpolation point design, adjustment of the trust region radii and geometry improvement of the interpolation set, etc. In Section 2.1, we present the algorithm outline based on the general model-based approach, where many specific details of UOBYQA are ignored. Interested readers may refer to Powell's paper [23].

We mention the notion of adequacy of the interpolation points in a ball

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

The definition of this notion is given in [4] and the key property that we use is found in Lemma 2 of Section 3. The paper [4] shows a mechanism that will generate adequate interpolation points after a finite number of applications. UOBYQA applies a heuristic procedure, which may not guarantee these properties, but is very effective in practice. Since this point is orthogonal 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.

2.1 Outline of the new algorithm

Starting the algorithm requires an initial trial point x_0 and an initial trust region radius Δ_0. At each iteration k, the correct sample path number N_k is computed by the Bayesian VNSP scheme in Section 2.2, resulting in an averaged sample-path function $\bar{f}^{N_k}$ (an approximation to $\bar{f}^{\infty} = F(x)$ in (1.2)). The derivative estimate of $\bar{f}^{N_k}$ is contained in a quadratic model

$$Q_k^{N_k}(x_k + s) = c_k^{N_k} + \left(g_k^{N_k}\right)^T s + \frac{1}{2} s^T G_k^{N_k} s, \tag{2.1}$$

which is constructed by interpolating a set of well-positioned points $\mathcal{I} = \{y^1, y^2, \cdots, y^L\}$,

$$Q_k^{N_k}(y^i) = \bar{f}^{N_k}(y^i), \quad i = 1, 2, \cdots, L.$$

The point x_k acts as the center of the trust region. The coefficient $c_k^{N_k}$ is a scalar, $g_k^{N_k}$ is a vector in $\mathbb{R}^n$, and $G_k^{N_k}$ is an $n \times n$ real matrix. The interpolation model is expected to approximate $\bar{f}^{N_k}$ well around the base point x_k, such that the parameters $c_k^{N_k}, g_k^{N_k}$

and $G_k^{N_k}$ approximate the Taylor expansion coefficients of $\bar{f}^{N_k}$ around x_k. To ensure a unique quadratic interpolator, the number of interpolating points should satisfy

$$L = \frac{1}{2}(n+1)(n+2). \tag{2.2}$$

Note that the model construction step (2.1) does not require evaluations of the gradient or the Hessian.

For each quadratic interpolation model, we require that the Hessian matrix is uniformly bounded.

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

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

As in a classical trust region method, a new promising point is determined from a subproblem:

$$\min_{s \in \mathbb{R}^n} Q_k^{N_k}(x_k + s), \quad \text{subject to } \|s\| \leq \Delta_k. \tag{2.3}$$

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

$$\rho_k^{N_k} = \frac{\bar{f}^{N_k}(x_k) - \bar{f}^{N_k}(x_k + s^{*,N_k})}{Q_k^{N_k}(x_k) - Q_k^{N_k}(x_k + s^{*,N_k})}. \tag{2.4}$$

If the ratio $\rho_k^{N_k}$ is large enough, the point $x_k + s^{*,N_k}$ is accepted into the set $\mathcal{I}$, otherwise, the geometry of $\mathcal{I}$ should be improved when necessary. The trust region radius is then updated following standard trust region rules. Whenever a new point x^+ enters (the point x^+ may be the solution point $x_k + s^{*,N_k}$ or a replacement point to improve the geometry), the agreement should be rechecked to determine the next iterate.

We now present the extended UOBYQA algorithm, implementing the VNSP scheme. 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 Δ_0 and a termination trust region radius Δ_{end}.*

1. *Generate initial trial points in the interpolation set $\mathcal{I}$. Determine the first iterate $x_1 \in \mathcal{I}$ as the best point in $\mathcal{I}$.*

2. *For iterations $k = 1, 2, \ldots$*

   (a) *Determine N_k via the VNSP scheme in Section 2.2.*

(b) *Construct a quadratic model* $Q_k^{N_k}$ *of the form (2.1) which interpolates points in* $\mathcal{I}$. *If* $\|g_k^{N_k}\| \leq \epsilon_1$ *and* $\mathcal{I}$ *is inadequate in* $\mathcal{B}_k(\epsilon_2\|g_k^{N_k}\|)$, *then improve the quality of* $\mathcal{I}$.

(c) *Solve the trust region subproblem (2.3). Evaluate* $\bar{f}^{N_k}$ *at the new point* $x_k + s^{*,N_k}$ *and compute the agreement ratio* $\rho_k^{N_k}$ *in (2.4).*

(d) *If* $\rho_k^{N_k} \geq \eta_1$, *then insert* $x_k + s^{*,N_k}$ *into* $\mathcal{I}$. *If a point is added to the set* $\mathcal{I}$, *another element in* $\mathcal{I}$ *should be removed to maintain the cardinality* $|\mathcal{I}| = L$. *Improve the quality of* $\mathcal{I}$ *if necessary.*

(e) *Update the trust region radius* Δ_k*:*

$$\Delta_{k+1} \begin{cases} \in [\Delta_k, \gamma_2\Delta_k], & \text{if } \rho_k^{N_k} \geq \eta_1; \\ \in [\gamma_0\Delta_k, \gamma_1\Delta_k], & \text{if } \rho_k^{N_k} < \eta_1 \text{ and } \mathcal{I} \text{ is adequate in } \mathcal{B}_k(\Delta_k); \\ = \Delta_k, & \text{otherwise.} \end{cases} \tag{2.5}$$

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

$$\hat{\rho}_k^{N_k} = \frac{\bar{f}^{N_k}(x_k) - \bar{f}^{N_k}(x^+)}{Q_k^{N_k}(x_k) - Q_k^{N_k}(x_k + s^{*,N_k})} \geq \eta_0, \tag{2.6}$$

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* $\Delta_k \leq \Delta_{end}$ *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*.

2.2 Bayesian VNSP scheme

The goal of the VNSP scheme is to determine the suitable sample path number N_k to be applied at iteration k. As a consequence, the algorithm, performing on the average sample path $\bar{f}^{N_k}$, produces solutions x_k that converge to $x^{*,N_\infty} = x^{*,\infty}$ (see Figure 3). As we have known, increasing the number N_k lessens the bias between the quadratic model $Q_k^{N_k}$ and the 'expected' quadratic model Q_k^∞, and is likely to produce a more precise step length s^{*,N_k}, close to $s^{*,\infty}$. The difficulty lies in that the expected function $\bar{f}^\infty$ is not numerically available, neither is the interpolative model Q_k^∞.

At iteration k, the aim is to limit N_k while simultaneously maintaining high accuracy of the algorithm. A practical approach is to sequentially allocate resources: we

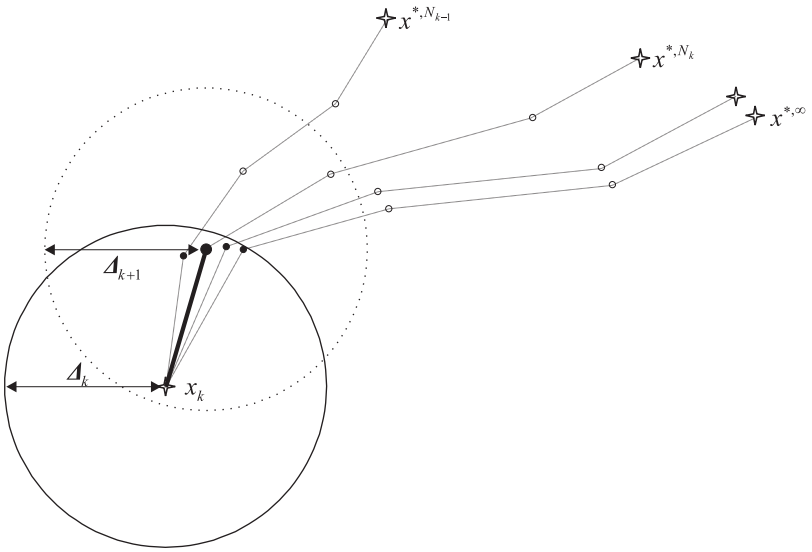

Figure 3: Choose the correct N_k and move the next iterate along the averaged sample-path function $\bar{f}^{N_k}$.

evaluate and check a satisfactory criterion for the current N_k. If rejected, we increase N_k to improve the accuracy. Typically, N_k is updated as

$$N_k := N_k \cdot \beta,$$

where β is an incremental factor. Before we describe the statistically valid satisfactory criterion, we will 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 1. *At iteration k, if $Q_k^{N_k}$ has the form of (2.1), then the solution s_k^{*,N_k} of the subproblem (2.3) satisfies*

$$Q_k^{N_k}(x_k) - Q_k^{N_k}(x_k + s^{*,N_k}) \geq \kappa_{mdc} \|g_k^{N_k}\| \min\left[\frac{\|g_k^{N_k}\|}{\kappa_{Qh}}, \Delta_k\right] \tag{2.7}$$

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

Proof. For the *Cauchy Point* $x_k + s_c^{N_k}$ defined as the minimizer of the model in the trust region along the steepest decent direction, we have a corresponding reduction [19]

$$Q_k^{N_k}(x_k) - Q_k^{N_k}(x_k + s_c^{*,N_k}) \geq \frac{1}{2} \|g_k^{N_k}\| \min\left[\frac{\|g_k^{N_k}\|}{\kappa_{Qh}}, \Delta_k\right]. \tag{2.8}$$

Since the solution s^{*,N_k} of the subproblem yields an even lower objective value of $Q_k^{N_k}$, we have the inequality (2.7). The complete proof can be found in [20]. □

There are issues concerning setting the values of κ_{mdc} and κ_{Qh} in an implementation. For κ_{mdc}, we use the value of $\frac{1}{2}$. This value is true for Cauchy points, so is valid for the solutions of the subproblem. For κ_{Qh}, we update it as the algorithm proceeds

$$\kappa_{Qh} := \max\left(\kappa_{Qh}, \|G_k^{N_k}\|\right), \tag{2.9}$$

that is, κ_{Qh} is updated whenever a new $G_k^{N_k}$ is generated.

In the algorithm, $Q_k^{N_k}(x_k) - Q_k^{N_k}(x_k + s^{*,N_k})$ is the observed model reduction, which serves to promote the next iterate (i.e. used in computing the agreement $\rho_k^{N_k}$ in (2.4)). Lemma 1 implies that the reduction also regulates the size of $\|g_k^{N_k}\|$, and further drives $\|g_k^{N_k}\|$ to zero. Therefore, by replacing $g_k^{N_k}$ with g_k^{∞}, we come up with a gauge for the goodness of N_k by checking whether the model reduction $Q_k^{N_k}(x_k) - Q_k^{N_k}(x_k + s^{*,N_k})$ regulates the expected norm $\|g_k^{\infty}\|$ or not. We introduce the following satisfactory criterion for N_k:

$$Pr\left(Q_k^{N_k}(x_k) - Q_k^{N_k}(x_k + s^{*,N_k}) \geq \kappa_{mdc}\|g_k^{\infty}\| \min\left[\frac{\|g_k^{\infty}\|}{\kappa_{Qh}}, \Delta_k\right]\right) \geq 1 - \alpha_k, \tag{2.10}$$

where α_k indicates the significance level. This criterion is a probabilistic statement because we do not know the explicit form of Q_k^{∞} (and hence g_k^{∞}). The structure of Q_k^{∞} must be estimated based on existing data, thus, it is subject to randomness and imprecise. The existing data we use is accumulated into an $N_k \times L$ matrix X, where

$$X_{ij} = f(y^j, \omega_i), \;\; i = 1, \cdots, \; N_k, j = 1, \cdots, L,$$

and L is the cardinality of the set $\mathcal{I}$ defined in (2.2)). This data is available before the construction of the model $Q_k^{N_k}$.

For the purpose of global convergence (Section 3), we require the following assumption.

Assumption 4. *The sequence of significance level values α_k satisfies the property:*

$$\sum_{k=1}^{\infty} \alpha_k < \infty. \tag{2.11}$$

The assumption necessitates a stricter satisfactory criterion for N_k as k increases.

In the remainder of this section, we focus on constructing a Bayesian estimator for the formulation of Q_k^{∞} (which implicitly defines g_k^{∞}), with the goal of validating the criterion (2.10). Bayesian techniques are used extensively in simulation output analysis. Chick [2, 3] has implemented Bayesian estimation in ordering discrete simulation systems (ranking and selection [1, 15]). Deng and Ferris [7] propose a similar Bayesian analysis to evaluate the stability of surrogate models.

Let us assume the simulation output at points of $\mathcal{I}$

$$\boldsymbol{f} = (f(y^1, \omega), f(y^2, \omega), \cdots, f(y^L, \omega))$$

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

$$\boldsymbol{f} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{2.12}$$

Since the simulation outcomes are correlated, the covariance matrix is typically not a diagonal matrix. (Though the multivariate normal assumption (2.12) is not precise, in our later analysis, it is only used to derive the parameters of the multivariate normal vector $\boldsymbol{\mu}$, whose normality is the result of the Central Limit Theorem.)

We delve into the detailed steps of quadratic model construction in the UOBYQA algorithm. Using the mean values of $\boldsymbol{f}$, the quadratic model Q_k^∞ is expressed as a linear combination of Lagrange functions $l_j(x)$,

$$Q_k^\infty(x) = \sum_{j=1}^{L} \bar{f}^\infty(y^j) l_j(x) = \sum_{j=1}^{L} \mu(y^j) l_j(x), \quad x \in \mathbb{R}^n. \tag{2.13}$$

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

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

that has the property

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

where δ_{ij} is 1 if $i = j$ and 0 otherwise. It follows from (2.1) and (2.13) that the parameters of Q_k^∞ are derived as

$$\begin{aligned} c_k^\infty = \boldsymbol{c}\boldsymbol{\mu}^T,\ g_k^\infty = \boldsymbol{g}\boldsymbol{\mu}^T, \\ \text{and } G_k^\infty = \sum_{j=1}^{L} \mu(y^j) G_j, \end{aligned} \tag{2.14}$$

where $\boldsymbol{c} = (c_1, \cdots, c_L)$ and $\boldsymbol{g} = (g_1, \cdots, g_L)$. 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 $\bar{f}^\infty$.

In the Bayesian framework, the unknown mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{f}$ are considered as random variables, whose distributions are inferred by Bayes' rule. 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 $\boldsymbol{s}_i = (f(y^1, \omega_i), \cdots, f(y^L, \omega_i)), i = 1, \cdots, N_k$, so that $X = (\boldsymbol{s}_1, \cdots, \boldsymbol{s}_{N_k})^T$. The sample mean and sample covariance matrix are calculated as

$$\begin{aligned} \bar{\boldsymbol{\mu}} &= \sum_{i=1}^{N_k} \boldsymbol{s}_i / N_k \\ &= (\bar{f}^{N_k}(y^1), \cdots, \bar{f}^{N_k}(y^L)), \end{aligned} \tag{2.15}$$

and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N_k - 1} \sum_{i=1}^{N_k} (\boldsymbol{s}_i - \bar{\boldsymbol{\mu}})^T (\boldsymbol{s}_i - \bar{\boldsymbol{\mu}}). \tag{2.16}$$

For the validity of $\hat{\boldsymbol{\Sigma}}$, we introduce the following assumption.

Assumption 5. *The variance of the sample response function is uniformly bounded, that is, for each $x \in \mathbb{R}^n$,*

$$var(f(x, \omega)) \leq \kappa_{\sigma^2} \tag{2.17}$$

for some constant κ_{σ^2}.

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

$$\begin{aligned} \boldsymbol{\Sigma}|X &\sim Wishart_L(\hat{\boldsymbol{\Sigma}}, N_k + L - 2), \\ \boldsymbol{\mu}|\boldsymbol{\Sigma}, X &\sim N(\bar{\boldsymbol{\mu}}, \boldsymbol{\Sigma}/N_k). \end{aligned} \tag{2.18}$$

Here the Wishart distribution $Wishart_p(\boldsymbol{\nu}, m)$ has covariance matrix $\boldsymbol{\nu}$ and m degrees of freedom. The Wishart distribution is a multivariate generalization of the χ^2 distribution. The notation ' $|X$' stands for 'posterior distribution with the knowledge of data'.

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 (2.18) with the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$, and asymptotically derive the posterior distribution of $\boldsymbol{\mu}|X$ as

$$\boldsymbol{\mu}|X \sim N(\bar{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}/N_k). \tag{2.19}$$

Moreover, using an exact computation, the marginal distribution of $\boldsymbol{\mu}|X$ inferred by (2.18) (eliminating $\boldsymbol{\Sigma}$) is,

$$\boldsymbol{\mu}|X \sim St_L(\bar{\boldsymbol{\mu}}, N_k \hat{\boldsymbol{\Sigma}}^{-1}, N_k - 1), \tag{2.20}$$

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 exact formulation (2.20) matches the result of the frequentists' estimation in constructing confidence region for $\boldsymbol{\mu}$.

Although the normal formulation (2.19) is not as precise, it is more convenient to manipulate than the t-version (2.20), and the results of both versions turn out to be very close. Therefore, in our work, we will use the normal distribution (2.19).

If we treat the c_k^∞, g_k^∞ and G_k^∞ as random variables from a Bayesian perspective, according to (2.14) and (2.19), they are normal-like variables:

$$c_k^\infty|X \sim N(\boldsymbol{c}\bar{\boldsymbol{\mu}}^T, \boldsymbol{c}\hat{\boldsymbol{\Sigma}}\boldsymbol{c}^T/N_k), \tag{2.21}$$

$$g_k^\infty|X \sim N(\boldsymbol{g}\bar{\boldsymbol{\mu}}^T, \boldsymbol{g}\hat{\boldsymbol{\Sigma}}\boldsymbol{g}^T/N_k), \tag{2.22}$$

$$G_k^\infty|X \sim MN(\sum_{j=1}^{L} \bar{\mu}(y^j) G_j, \boldsymbol{P}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{P}/N_k, \boldsymbol{P}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{P}/N_k), \tag{2.23}$$

where the $L \times N_k$ matrix $\boldsymbol{P} = (G_1\mathbf{1}, \cdots, G_L\mathbf{1})^T$. 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 $\boldsymbol{\nu}_2$ [6]. In (2.23), because G_j are symmetric, the left variance and right variance coincide.

Up to now, we have successfully derived the posterior distributions for the parameters of Q_k^∞. These distributions can be applied in (2.10) to evaluate the probability of the 'sufficient reduction' event. Furthermore, because g_k^∞ is a multivariate normal variable, we know that $\kappa_{mdc}\|g_k^\infty\| \min\left[\frac{\|g_k^\infty\|}{\kappa_{Qh}}, \Delta_k\right]$ is a χ^2-like variable. However, the exact evaluation of the probability in (2.10) is very complex, especially involving the component $\min\left[\frac{\|g_k^\infty\|}{\kappa_{Qh}}, \Delta_k\right]$. Instead we use the Monte Carlo method to approximate the result: we generate N_t random samples from the posterior distribution of $\boldsymbol{\mu}|X$ (2.18). Based on the samples, we evaluate the event of 'sufficient reduction' in (2.10) and have a count on the successful cases: N_{succ}. The probability value in (2.10) is then approximated by

$$Pr(\text{sufficient reduction}) \approx \frac{N_{succ}}{N_t}. \tag{2.24}$$

N_t should be sufficiently large, i.e. we typically use 500.

We finalize our Bayesian VNSP scheme discussion with a complete description.

The VNSP scheme At the kth iteration of the algorithm, start with $N_k = N_{k-1}$. Loop

1. Evaluate N_k replications at each point y^j in the interpolation set $\mathcal{I}$, to construct the data matrix X. Note: data from previous iterations can be included.
2. Construct the quadratic model $Q_k^{N_k}$ and solve the subproblem for $x_k + s^{*,N_k}$.
3. Update the value of κ_{Qh} by (2.9).
4. Compute the Bayesian posterior distributions for the parameters of Q_k^∞ as described above.
5. Use the Monte Carlo method to compute the probability of 'sufficient reduction' in criterion (2.10).
6. If the probability value is greater than $1-\alpha_k$, then stop; otherwise increase N_k, and repeat the loop.

3 Convergence Analysis of the Algorithm

Convergence analysis of the general model-based approach is given by Conn, Scheinberg, and Toint in [4]. 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 [19, 20].

As a key component of the analysis, they address the difference of using the classical Taylor expansion model

$$\hat{Q}_k^{N_k}(x_k+s) = \bar{f}^{N_k}(x_k) + \nabla \bar{f}^{N_k}(x_k)^T s + \frac{1}{2} s^T \nabla^2 \bar{f}^{N_k}(x_k) s$$

and the interpolative quadratic model $Q_k^{N_k}$. The model $\hat{Q}_k^{N_k}$ shares the same gradient $\nabla \bar{f}^{N_k}(x_k)$ at x_k with the underlying function, while for the interpolative model $Q_k^{N_k}$, its gradient $g_k^{N_k}$ 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 $Q_k^{N_k}$ is also a decent approximation model.

Lemma 2. *(Theorem 4 in [4]) Assume Assumptions 1-3 hold and $\mathcal{I}$ is adequate in the trust region $\mathcal{B}_k(\Delta_k)$. Suppose at iteration k, $Q_k^{N_k}$ is the interpolative approximation model for the function $\bar{f}^{N_k}$, then the bias of the function value and the gradient are bounded within the trust region. There exist constants κ_{em} and κ_{eg}, for each $x \in \mathcal{B}_k(\Delta_k)$, the following inequalities hold*

$$|\bar{f}^{N_k}(x) - Q_k^{N_k}(x)| \leq \kappa_{em} \max[\Delta_k^2, \Delta_k^3] \tag{3.1}$$

and

$$||\nabla \bar{f}^{N_k}(x) - g_k^{N_k}|| \leq \kappa_{eg} \max[\Delta_k, \Delta_k^2]. \tag{3.2}$$

In fact, the proof of Lemma 2 is associated with manipulating Newton polynomials instead of Lagrange functions. Because the quadratic model is unique via interpolation (by choice of L), the results are valid regardless of how the model is constructed.

In the VNSP scheme, since the criterion of sufficient reduction (2.10) is satisfied with a certain probability, the *Borel-Cantelli Lemma* from *Probability Theory* is useful in our proof.

Lemma 3 ((1st) Borel-Cantelli Lemma)**.** *If the events E_k are independent, and the sum of the probabilities of E_k is finite, then with probability 1 (w.p.1), only finitely many E_k occur.*

A corresponding 2nd Borel-Cantelli Lemma states that if the sum of probabilities of E_k is infinite, then w.p.1, infinitely many E_k occur.

We start by showing that there is at least one critical accumulation point. The idea is to first show that the gradient g_k^∞ driven by the sufficient reduction criterion (2.10) converges to zero, and then prove that $\|\nabla \bar{f}^\infty(x_k)\|$ converges to zero as well.

Lemma 4. *Assume Assumptions 1–5 hold. If $\|g_k^\infty\| \geq \epsilon_g$ for all k and for some constant $\epsilon_g > 0$, then there exists a constant $\epsilon_\Delta > 0$ and a sufficiently large index K such that w.p.1,*

$$\Delta_k > \epsilon_\Delta, \text{ for all } k \geq K. \tag{3.3}$$

Proof. Under the assumption $\|g_k^\infty\| \geq \epsilon_g$, we will show that the corresponding Δ_k cannot become too small, therefore, we can derive the constant ϵ_Δ. Let us first evaluate the following term associated with the agreement level

$$|\rho_k^{N_k} - 1| = \left| \frac{\bar{f}^{N_k}(x_k + s^{*,N_k}) - Q_k^{N_k}(x_k + s^{*,N_k})}{Q_k^{N_k}(x_k) - Q_k^{N_k}(x_k + s^{*,N_k})} \right|. \tag{3.4}$$

By Lemma 2, we compute the error bound for the numerator

$$\left| \bar{f}^{N_k}(x_k + s^{*,N_k}) - Q_k^{N_k}(x_k + s^{*,N_k}) \right| \leq \kappa_{em} \max[\Delta_k^2, \Delta_k^3]. \tag{3.5}$$

Note that when Δ_k is small enough, satisfying the condition

$$\Delta_k \leq \min \left[1, \frac{\kappa_{mdc}\epsilon_g(1-\eta_1)}{\max[\kappa_{Qh}, \kappa_{em}]} \right], \tag{3.6}$$

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

$$\Delta_k \leq \frac{\|g_k^\infty\|}{\kappa_{Qh}}. \tag{3.7}$$

For the denominator in (3.4), our 'sufficient reduction' criterion (2.10) provides a probabilistic bound. With probability at least $1 - \alpha_k$, the inequality

$$Q_k^{N_k}(x_k) - Q_k^{N_k}(x_k + s^{*,N_k}) \geq \kappa_{mdc}\|g_k^\infty\| \min \left[\frac{\|g_k^\infty\|}{\kappa_{Qh}}, \Delta_k \right] = \kappa_{mdc}\|g_k^\infty\|\Delta_k \tag{3.8}$$

holds.

Combining (3.4), (3.5), (3.6) and (3.8), the following inequality holds with probability greater than $1 - \alpha_k$

$$\begin{aligned} |\rho_k^{N_k} - 1| &= \left| \frac{\bar{f}^{N_k}(x_k + s^{*,N_k}) - Q_k^{N_k}(x_k + s^{*,N_k})}{Q_k^{N_k}(x_k) - Q_k^{N_k}(x_k + s^{*,N_k})} \right| \\ &\leq \frac{\kappa_{em} \max[\Delta_k^2, \Delta_k^3]}{\kappa_{mdc}\|g_k^\infty\|\Delta_k} \\ &\leq \frac{\kappa_{em}\Delta_k}{\kappa_{mdc}\|g_k^\infty\|} \\ &\leq 1 - \eta_1. \end{aligned} \tag{3.9}$$

This further gives

$$Pr\left(\rho_k^{N_k} \geq \eta_1\right) \geq Pr\left(|\rho_k^{N_k} - 1| \leq 1 - \eta_1\right) \geq 1 - \alpha_k. \tag{3.10}$$

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 $\bar{f}^{N_k}$, which will induce an increase of the trust region

radius $\Delta_{k+1} \geq \Delta_k$ (2.5). However, in (3.10) the criterion $\rho_k^{N_k} \geq \eta_1$ is only valid with the probability $\geq 1 - \alpha_k$. Defining the event $E_k := \{\rho_k^{N_k} \geq \eta_1\}$, under Assumption 4, the Borel-Cantelli Lemma gives that the failed events E_k only occur finitely many times w.p.1. Therefore, if we define K as the first successful index after all the failed events, we thus have w.p.1,

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

According to (3.6), it is equivalent to say that Δ_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 Δ_k:

$$\Delta_k > \epsilon_\Delta = \gamma_0 \min\left[1, \frac{\kappa_{mdc}\epsilon_g(1-\eta_1)}{\max[\kappa_{Qh}, \kappa_{em}]}\right], \text{ for } k \geq K. \tag{3.11}$$

□

Theorem 1. *Assume Assumptions 1–5 hold. Then there is a subsequence of* $\{g_k^\infty\}$ *converges to zero*

$$\liminf_{k\to\infty} \|g_k^\infty\| = 0. \tag{3.12}$$

Proof. We prove the statement (3.12) by contradiction. Suppose there is $\epsilon_g > 0$ such that

$$\|g_k^\infty\| > \epsilon_g. \tag{3.13}$$

By Lemma 4, 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 (2.4), (2.10), (3.13) and $\Delta_k > \epsilon_\Delta$, the inequality

$$\begin{aligned} \bar{f}^{N_k}(x_k) - \bar{f}^{N_k}(x_{k+1}) &\geq \eta_0\left[Q_k^{N_k}(x_k) - Q_k^{N_k}(x_k + s^{*,N_k})\right] \\ &\geq \eta_0\kappa_{mdc}\epsilon_g \min\left[\frac{\epsilon_g}{\kappa_{Qh}}, \epsilon_\Delta\right] \end{aligned} \tag{3.14}$$

holds with probability $\geq 1 - \alpha_k$.

We will discuss two situations here: (a) when the limit of the sequence $\lim_{k\to\infty} N_k = N_\infty$ is a finite number, and (b) when N_∞ 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 (3.14).

Situation (a): If $N_\infty < \infty$, then there exists an index $\tilde{K}$ such that $N_k = N_\infty$ for $k \geq \tilde{K}$. By Assumption 4, the Borel-Cantelli Lemma shows that (3.14) is true for infinitely many successful iterations w.p.1. Let $\mathcal{K} \subset \mathcal{S}$ denote the index of these iterations, and sum over indexes in $\mathcal{K}$. Since $\{\bar{f}^{N_\infty}(x_k) | \, k \geq \tilde{K}\}$ is monotonically decreasing

$$\begin{aligned} \bar{f}^{N_\infty}(x_{\tilde{K}}) - \bar{f}^{N_\infty}(x_{\hat{K}+1}) &\geq \sum_{\substack{k \geq \tilde{K}, k \leq \hat{K}, \\ k \in \mathcal{K}}} \bar{f}^{N_\infty}(x_k) - \bar{f}^{N_\infty}(x_{k+1}) \\ &\geq t(\hat{K})\epsilon_d, \end{aligned} \tag{3.15}$$

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

Situation (b): For this situation, $N_\infty = \infty$. By the Borel-Cantelli Lemma and Assumption 4, (3.14) fails only for a finite number of iterations w.p.1. Therefore, there exists a sufficiently large index $\bar{K}$ (which is greater than all the failed indices), such that (3.14) holds for all successful iteration $k \geq \bar{K}$ w.p.1.

Let us define a specific subsequence of indexes $\{k_{j'} | \, k_{j'} \geq \bar{K}\}$ (see Figure 4), indicating where there is a jump in N_k, i.e. a truncated part of subsequence is

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

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

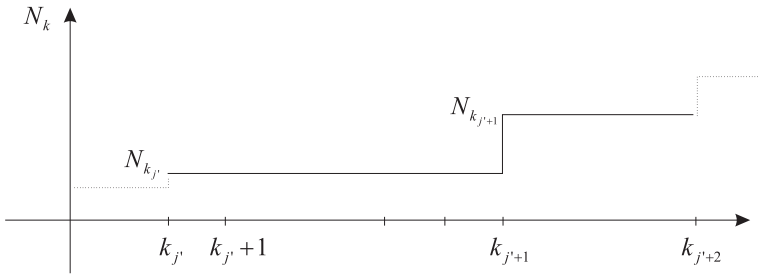

Figure 4: Illustration of the subsequence $\{k_{j'}\}$

in $\{k_{j'}, \ldots, k_{j'+1} - 1\}$. 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 (3.14) for $k \in \{N_{k_{j'}}, \ldots, N_{k_{j'+1}-1}\}$ to derive

$$\begin{aligned} \bar{f}^{N_{k_{j'}}}(x_{k_{j'}}) - \bar{f}^{N_{k_{j'}}}(x_{k_{j'+1}}) &\geq \sum_{\substack{k \geq k_{j'}, k \leq k_{j'+1}-1 \\ k \in \mathcal{S}}} \bar{f}^{N_{k_{j'}}}(x_k) - \bar{f}^{N_{k_{j'}}}(x_{k+1}) \\ &\geq \epsilon_d. \end{aligned} \tag{3.16}$$

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

According to the Central Limit Theorem, when $N_{k_{j'}}$ is large, the difference is approximately a normal variable

$$\left(\bar{f}^{N_{k_{j'}}}(x_{k_{j'}}) - \bar{f}^{N_{k_{j'}}}(x_{k_{j'+1}})\right) - \left(\bar{f}^{\infty}(x_{k_{j'}}) - \bar{f}^{\infty}(x_{k_{j'+1}})\right) \sim N\left(0, \frac{\sigma^2(x_{k_{j'}}) + \sigma^2(x_{k_{j'+1}})}{N_{k_{j'}}}\right),$$

where $\sigma^2(\cdot)$ represents the variance of $f(\cdot, \omega)$ at x, and is bounded by κ_{σ^2} according to Assumption 5. Consider the probability of 'insufficient reduction' of $\bar{f}^{\infty}(x_{k_{j'}}) - \bar{f}^{\infty}(x_{k_{j'+1}})$ for $k_{j'} \in \mathcal{S}'$:

$$\begin{aligned} & Pr\left(\bar{f}^{\infty}(x_{k_{j'}}) - \bar{f}^{\infty}(x_{k_{j'+1}}) \leq \frac{\epsilon_d}{2}\right) \\ = \; & Pr\left(\bar{f}^{N_{k_{j'}}}(x_{k_{j'}}) - \bar{f}^{N_{k_{j'}}}(x_{k_{j'+1}}) + N\left(0, \frac{\sigma^2(x_{k_{j'}}) + \sigma^2(x_{k_{j'+1}})}{N_{k_{j'}}}\right) \leq \frac{\epsilon_d}{2}\right) \\ \leq \; & Pr\left(N\left(0, \frac{\sigma^2(x_{k_{j'}}) + \sigma^2(x_{k_{j'}+1})}{N_{k_{j'}}}\right) \leq -\frac{\epsilon_d}{2}\right) \\ \leq \; & Pr\left(N\left(0, \frac{2\kappa_{\sigma^2}}{N_{k_{j'}}}\right) \leq -\frac{\epsilon_d}{2}\right) = Pr\left(N\left(0, \frac{1}{N_{k_{j'}}}\right) \leq -\frac{\sqrt{2\kappa_{\sigma^2}}\epsilon_d}{4\kappa_{\sigma^2}}\right). \end{aligned}$$

The Chernoff's bound in *Large Deviation Theory* implies that the above probability decreases exponentially fast with respect to $N_{k_{j'}}$ [13]

$$Pr\left(N\left(0, \frac{1}{N_{k_{j'}}}\right) \leq -\frac{\sqrt{2\kappa_{\sigma^2}}\epsilon_d}{4\kappa_{\sigma^2}}\right) \leq \exp\left(-N_{k_{j'}} I\left(\frac{\sqrt{2\kappa_{\sigma^2}}\epsilon_d}{4\kappa_{\sigma^2}}\right)\right).$$

Here

$$I(z) = \sup_{t \in \mathbb{R}}(tz - logM(t))$$

is called *rate function* and

$$M(t) = \mathbb{E}[e^{tN(0,1)}] = e^{\frac{1}{2}t^2}$$

is the *moment generating function* of $N(0,1)$. The value of $I\left(\frac{\sqrt{2\kappa_{\sigma^2}}\epsilon_d}{4\kappa_{\sigma^2}}\right)$ is therefore a positive scalar.

Since the sum of probabilities is bounded

$$\sum_{\substack{j'=1\\ k_{j'}\in\mathcal{S}'}}^{\infty} Pr\left(\bar{f}^{\infty}(x_{k_{j'}}) - \bar{f}^{\infty}(x_{k_{j'+1}}) \leq \frac{\epsilon_d}{2}\right) \leq \sum_{\substack{j'=1\\ k_{j'}\in\mathcal{S}'}}^{\infty} \exp\left(-N_{k_{j'}} I\left(\frac{\sqrt{2\kappa_{\sigma^2}}\epsilon_d}{4\kappa_{\sigma^2}}\right)\right) < \infty,$$

applying the Borel-Cantelli Lemma again, the event $E_{k_{j'}} = \{\bar{f}^{\infty}(x_{k_{j'}}) - \bar{f}^{\infty}(x_{k_{j'}+1}) \leq \epsilon_d/2\}$ occurs only finitely many times w.p.1. Thus, there exists an index $\bar{K}_1 \geq \bar{K}$, such that

$$\bar{f}^{\infty}(x_{k_{j'}}) - \bar{f}^{\infty}(x_{k_{j'+1}}) \geq \frac{\epsilon_d}{2}, \text{ for all } \{k_{j'} | k_{j'} \geq \bar{K}_1, k_{j'} \in \mathcal{S}'\} \quad w.p.1.$$

Playing the same trick as before, by summing over all $k_{j'} \geq \bar{K}_1$, we derive that

$$\begin{aligned} \bar{f}^{\infty}(x_{\bar{K}_1}) - \bar{f}^{\infty}(x_{\hat{K}+1}) &\geq \sum_{\substack{k_{j'}\geq\bar{K}_1, k_{j'}\leq\hat{K}\\ k_{j'}\in\mathcal{S}'}} \bar{f}^{\infty}(x_{k_{j'}}) - \bar{f}^{\infty}(x_{k_{j'}+1}) \\ &\geq t(\hat{K})\eta_0\frac{\epsilon_d}{2}, \end{aligned} \tag{3.17}$$

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 Δ_k decreases to zero

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

which contradicts the statement that Δ_k is bounded below (3.11). Thus (3.13) is false, and the theorem is proved. □

Theorem 2. *Assume Assumptions 1–5 hold. If*

$$\liminf_{k\to\infty} \|g_{k_j}^{\infty}\| = 0 \tag{3.18}$$

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

$$\liminf_{k\to\infty} \|\nabla\bar{f}^{\infty}(x_{k_j})\| = 0. \tag{3.19}$$

Proof. Because of the fact $\lim_{j\to} \Delta_{k_j} = 0$, Lemma 2 guarantees that the difference between $\|g_{k_j}^{\infty}\|$ and $\|\nabla\bar{f}^{\infty}(x^{k_j})\|$ is small. Thus the assertion (3.19) follows. The details of the proof refer to Theorem 11 in [4]. □

Theorem 3. *Assume Assumptions 1–5 hold. Every limit point x^* of the sequence x_k is critical.*

Proof. The procedure of proof is essentially the same as given for Theorem 12 in [4]. However, we use the 'sufficient reduction' inequalities (3.15) when N_{∞} is finite and (3.17) when N_{∞} is infinite. □

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

$$\hat{x}(x,\omega) := (x_1\omega, x_2, \cdots, x_n)$$

and the corresponding function becomes

$$f(x,\omega) = \sum_{i=1}^{n-1} 100(\hat{x}_{i+1} - \hat{x}_i^2)^2 + (\hat{x}_i - 1)^2. \tag{4.1}$$

We assume ω is a normal variable centered at 1:

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

As a general setting, the initial and end trust region radius Δ_0, Δ_{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. The number of trial samples $N_t = 500$ (see (2.24)) were generated to evaluate the probability of the 'sufficient reduction' event (2.10) in the VNSP procedure. To satisfy Assumption 4, 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 $\bar{f}^3$, all the steps were taken regarding $\bar{f}^3$ as the objective function. Therefore, it was observed that the iterates x_k moved toward the solution $x^{*,3}$ of the averaged sample path 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 that, in order to derive the solution to $\bar{f}^\infty$ in the two dimensional problem, the noisy Rosenbrock function was rearranged as

$$\begin{aligned} \bar{f}^\infty(x) &= \mathbb{E}\left[100(\hat{x}_2 - \hat{x}_1^2)^2 + (\hat{x}_1 - 1)^2\right] \\ &= 100x_2^2 + 1 - 2x_1\mathbb{E}[\omega] + (-200x_2x_1^2 + x_1^2)\mathbb{E}[\omega^2] + 100x_1^4\mathbb{E}[\omega^4]. \end{aligned}$$

By plugging the values $\mathbb{E}[\omega] = 1, \mathbb{E}[\omega^2] = 1.01$, and $\mathbb{E}[\omega^4] = 1.0603$, 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 $\bar{f}^{N_k}$ might vary greatly.

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	$f^{N_k}(x_k)$	Δ_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

In Table 1, we observe that $x_{19} = x_{20} = (0.5002, 0.2449)$. The value of $\bar{f}^{N_{19}}(x_{19})$ is 0.3616, while the value of $\bar{f}^{N_{20}}(x_{20})$ is 0.4904. It shows that the algorithm actually worked on objective functions with increases due to randomness.

Table 2: Averaged sample-path solution with different sample number N

N	$x^{*,N}$	$f^{N_k}(x^{*,N})$
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
∞	(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 $N_k = N_0$ sample paths 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 a early termination at the 34th iteration.

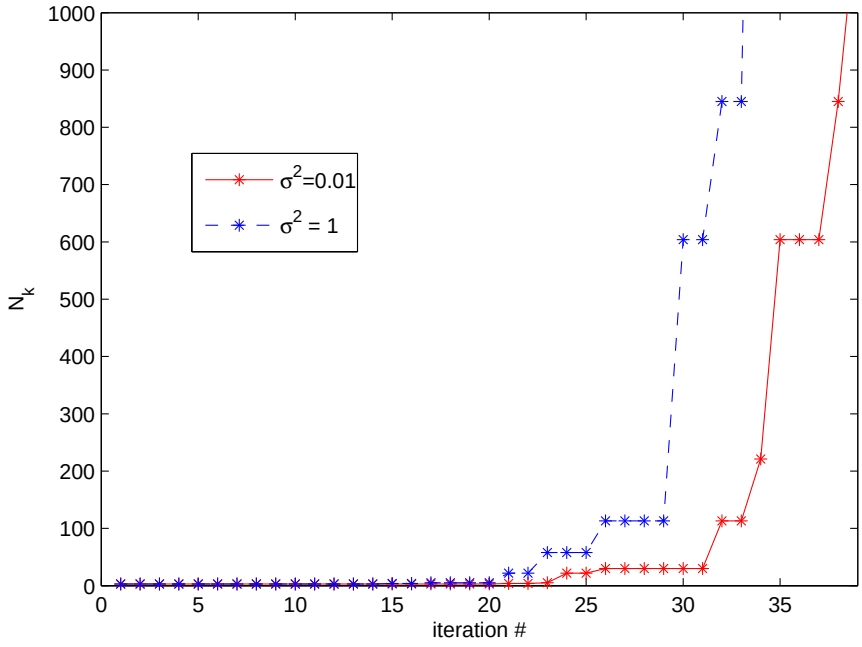

Figure 5: Compare changes of N_k with different levels of noise

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, it 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. The statistical results were based on 10 replications of the algorithm, 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 σ^2 is large (we used a limit of 20000 function evaluations in this case).

Table 3: Statistical summary

N	Noise level σ^2	Mean error	Variance of error
2	0.01	1.1e-5	1.2e-5
2	0.1	8.9e-5	3.3e-5
2	1	1.1e-4	8.2e-5
10	0.01	0.054	0.067
10	0.1	0.087	0.060
10	1	2.6	0.10

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

$$f(x, \omega) = g(x) + \omega.$$

We will show N_k is unchanged for every iteration, that is, $N_1 = N_2 = \cdots = N_\infty$. At iteration k, the function outputs at points y^j in $\mathcal{I}$ are entirely correlated. As a result, the sample covariance matrix $\hat{\mathbf{\Sigma}}$ (2.16) is a rank-one matrix, whose elements are all identical $\hat{\mathbf{\Sigma}}(i, j) = a,\ i, j = 1, 2, \cdots, L$, where $a = var[(\omega_1, \cdots, \omega_{N_k})]$. Thus, the matrix can be decomposed as

$$\hat{\mathbf{\Sigma}} = \mathbf{1} \cdot a \cdot \mathbf{1}^T. \tag{4.2}$$

Plug (4.2) into (2.22), we obtain the posterior covariance of g_k^∞

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

which implies g_k^∞ 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 (2.10) is always satisfied.

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

5 Conclusions

This paper proposes and analyzes a variable number sample-path schem 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 switches to more accurate objective functions, whose accuracy increased with N_k. The key idea of choosing an appropriate N_k in the VNSP scheme is to test whether the 'sufficient reduction' criterion is satisfied for the 'expected' model Q_k^∞. Although the criterion is a probabilistic statement, under Assumption 4, the global convergence of algorithm is guaranteed:

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

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.

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. Because the permittivity and electric conductivity vary among patients, we want the optimal design to perform well in the averaged sense. Further applications will be addressed in future work.

References

[1] H.-C. Chen, C.-H. Chen, and E. Yucesan. An asymptotic allocation for simultaneous simulation experiments. In D. T. Sturrock P. A. Farrington, H. B. Nembhard and G. W. Evans, editors, *Proceedings of the 1999 Winter Simulation Conference*, 1999.

[2] 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.

[3] S. E. Chick and K. Inoue. New two-stage and sequential procedures for selecting the best simulated system. *Operations Research*, 49:1609–1624, 2003.

[4] 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.

[5] 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.

[6] A. P. Dawid. Some matrix-variate distribution theory: Notational considerations and a bayesian application. *Biometrika*, 68(1):265–74, 1981.

[7] G. Deng and M. C. Ferris. Adaptation of the UOBQYA algorithm for noisy functions. In B. Lawson, J. Liu, F. Perrone, and F. Wieland, editors, *Proceedings of the 2006 Winter Simulation Conference*, Orlando, Florida, 2006. Omnipress.

[8] 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.

[9] M. Fu. Optimization via simulation: A review. *Annals of Operations Research*, 53:199–248, 1994.

[10] A. Gosavi. *Simulation-Based Optimization: Parametric Optimization Techniques and Reinforcement Learning*. Kluwer Academic Publishers, 2003.

[11] G. Gürkan, A. Yonca Özge, and S. M. Robinson. Sample-path optimization in simulation. In J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila, editors, *Proceedings of the 1994 Winter Simulation Conference*, 1994.

[12] 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*, 1999.

[13] T. Homem–de–Mello. Variable-sample methods for stochastic optimization. *ACM Transactions on Modeling and Computer Simulation*, 13(2):108–133, 2003.

[14] T. Homem–de–Mello. On rates of convergence for stochastic optimization problems under non-I.I.D. sampling. *submitted for publication*, 2006.

[15] S.-H. Kim and B. L. Nelson. Selecting the best system: Theory and methods. In S. Chick, P. J. Sanchez, D. Ferrin, and D. J. Morrice, editors, *Proceedings of the 2003 Winter Simulation Conference*, 2003.

[16] 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.

[17] A. Law and W. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, New York, third edition, 2000.

[18] M. Marazzi and J. Nocedal. Wedge trust region methods for derivative free optimization. *Mathematical Programming*, 91:289–305, 2002.

[19] 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.

[20] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 1999.

[21] 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.

[22] 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.

[23] M. J. D. Powell. UOBYQA: unconstrained optimization by quadratic approximation. *Mathematical Programming*, 92:555–582, 2002.

[24] M. J. D. Powell. The NEWUOA software for unconstrained optimization with derivatives. *DAMTP Report 2004/NA05, University of Cambridge*, 2004.

[25] S. M. Robinson. Analysis of sample-path optimization. *Mathematics of Operations Research*, 21:513–528, 1996.

[26] A. Shapiro. Asymptotic behavior of optimal solutions in stochastic programming. *Mathematics of Operations Research*, 18:829–845, 1993.

[27] 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.

[28] A. Shapiro. Monte carlo sampling approach to stochastic programming. In J. P. Penot, editor, *ESAIM:Proceedings*, volume 13, pages 65–73, 2003.