Algorithm XXX: QNSTOP—Quasi-Newton Algorithm for Stochastic Optimization

BRANDON D. AMOS, DAVID R. EASTERLING, LAYNE T. WATSON
Virginia Polytechnic Institute and State University
WILLIAM I. THACKER
Winthrop University
and
BRENT S. CASTLE, MICHAEL W. TROSSET
Indiana University

QNSTOP consists of serial and parallel (OpenMP) Fortran 2003 codes for the quasi-Newton stochastic optimization method of Castle and Trosset. For stochastic problems, convergence theory exists for the particular algorithmic choices and parameter values used in QNSTOP. Both the parallel driver subroutine, which offers several parallel decomposition strategies, and the serial driver subroutine can be used for stochastic optimization or deterministic global optimization, based on an input switch. QNSTOP is particularly effective for "noisy" deterministic problems, using only objective function values. Some performance data for computational systems biology problems is given.

Categories and Subject Descriptors: J.2 [**Computer Applications**]: Physical Science and Engineering — *Mathematics*; G.3 [**Mathematics of Computing**]: Probability and Statistics; G.4 [**Mathematics of Computing**]: Mathematical Software

General Terms: Algorithms, Design, Documentation

Additional Key Words and Phrases: derivative-free, deterministic global optimization, quasi-Newton, response surface, stochastic optimization

This work was supported in part by Air Force Research Laboratory Grant FA8650-09-2-3938 and Air Force Office of Scientific Research Grant FA9550-09-1-0153.
Authors' addresses: B. D. Amos, D. R. Easterling, Department of Computer Science, L. T. Watson, Departments of Computer Science, Mathematics, and Aerospace and Ocean Engineering, Virginia Polytechnic Institute & State University, Blacksburg, VA 24061; e-mail: `bdamos@vt.edu`, {`dreast, ltw`}`@cs.vt.edu`; W. I. Thacker, Computer Science Department, Winthrop University, Rock Hill, SC 29733; e-mail: `thackerw@winthrop.edu`; B. S. Castle, School of Informatics and Computing, M. W. Trosset, Department of Statistics, Indiana University, Bloomington, IN 47405; e-mail: `mtrosset@indiana.edu`.
Permission to make digital/hard copy of part or all of this work for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication, and its date appear, and notice is given that copying is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires specific permission and/or fee.
© 2014 by the Association for Computing Machinery, Inc.

1. INTRODUCTION

Given the function $f : \mathbb{R}^p \to \mathbb{R}$ and $\ell,\ u \in \mathbb{R}^p$, $\ell < u$, the problem under consideration is

$$\min_{x \in B} f(x)$$

over the box $B = \{x \in \mathbb{R}^p \mid \ell \leqq x \leqq u\}$. The types of functions of interest are where f is either stochastic (f itself, or observations of $f(x)$) or deterministic but noisy (having large local total variation). Although the algorithm here (QNSTOP) was conceived for stochastic optimization, it has two features that make it attractive for globally optimizing noisy deterministic objectives. First, because QNSTOP smooths (by regression) observed responses to construct semilocal approximations, it automatically filters high-frequency oscillations in the objective. Second, if the designs used by QNSTOP to obtain information in the current ellipsoidal design region $E_k(\tau_k)$ are space-filling (QNSTOP elects to draw random samples from a uniform distribution on $E_k(\tau_k)$, then edits them to increase the minimum interpoint distance between the proposed design sites), then QNSTOP may serendipitously discover unexpectedly small objective values within the semilocal region $E_k(\tau_k)$ used for smoothing. Plausible competitors to QNSTOP are Spall's Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm [Spall 1987, 1992, 1998], which constructs gradient estimates from just two function evaluations, and Kelley's implicit filtering algorithm [Gilmore and Kelley 1995], which relies on coarse stencil-based finite differencing to construct a descent direction. Both of these are philosophically quite different from QNSTOP, as explained in detail later.

The following sections provide background, discuss varying philosophies of stochastic optimization, describe QNSTOP in detail, and provide some performance data on difficult systems biology problems.

2. STOCHASTIC OPTIMIZATION BACKGROUND

Stochastic optimization is optimization when function evaluation is uncertain, i.e., corrupted by the presence of random noise. For example, suppose that one seeks to minimize $\mu : \mathbb{R}^p \to \mathbb{R}$. Given $x \in \mathbb{R}^p$, one would like to observe $\mu(x)$; instead, one observes $\mu(x) + \epsilon_x$, where ϵ_x is a random variable. This is the case of additive random noise. In this case, the underlying objective function μ is often called the regression function (in the stochastic approximation literature) or the response surface (in the response surface methodology literature).

One typically imposes various assumptions on the ϵ_x. The assumption $E(\epsilon_x) = 0$, from which it follows that the expected value of an observation is the true value of the objective function, is inevitable. One might also assume that $\epsilon_x \sim \text{Normal}(0, \sigma_x^2)$ (normality), that $\text{Var}(\epsilon_x) = \sigma_x^2$ does not depend on x (homoscedasticity), and that the ϵ_x are independent (white noise). The preceding set of assumptions is referred to as the *standard example*.

Random noise may not be additive. Because there is no elegant way to catalog the many random mechanisms by which a deterministic objective function might be corrupted, the concept of optimizing in the presence of random noise is somewhat

elusive. The usual approach is to begin with what one observes, not with what one seeks to optimize. Given $x \in \mathbb{R}^p$, suppose that one observes a random variable Y_x. One then *defines* the objective function to be $\mu(x) = EY_x$. However, there are a number of meaningful optimization problems for which function evaluation is uncertain that are more naturally expressed in a slightly different setting.

Let

$$\mathcal{P} = \{P(\cdot; x) \mid x \in \mathcal{C} \subseteq \mathbb{R}^p\}$$

denote a family of probability distributions indexed by x. Assume that the $P(\cdot; x)$ are completely unknown or analytically intractable, but that one can sample from any specified $P(\cdot; x)$. The first case might arise as one varies the prescribed operating characteristics of a manufacturing facility in search of an optimum. This is a typical concern of response surface methodology. In this case, observations are generated by a physical process for which a formal mathematical description is not available. The second case might arise when one is tuning the parameters of a simulated stochastic process, searching for settings that produce simulated data sets that resemble an actual data set. This is a useful approach to parameter estimation when the statistical model is defined implicitly, i.e., in terms of a generating stochastic mechanism rather than by specifying a parametric family of probability distributions. See, for example, Diggle and Gratton [1984] and Thompson [2000]. In neither case can one manipulate the $P(\cdot; x)$ as mathematical objects; instead one must rely on random sampling to obtain information about them.

Now let $T : \mathcal{P} \to \mathbb{R}$ and let $f(x) = T(P(\cdot; x))$. One seeks local solutions of

$$\min_{P \in \mathcal{P}} T(P), \tag{1}$$

or, equivalently, of

$$\min_{x \in \mathcal{C}} f(x). \tag{2}$$

Additional smoothness assumptions are imposed on T or f, as needed.

As stated, Problems (1) and (2) are unambiguous, deterministic optimization problems. They become stochastic when one cannot manipulate the $P(\cdot; x)$ as mathematical objects. When one must estimate $f(x) = T(P(\cdot; x))$ from a random sample

$$\omega_1(x), \ldots, \omega_n(x) \sim P(\cdot; x), \tag{3}$$

then function evaluation is uncertain and Problems (1) and (2) are stochastic optimization problems.

Given an independent and identically distributed random sample (3), let

$$\hat{P}_n(\cdot; x) = \sum_{i=1}^{n} \frac{1}{n} \delta_{\omega_i(x)}$$

denote the empirical distribution of the sample, i.e., the discrete probability distribution that assigns probability $1/n$ to each $\omega_i(x)$. In the case of univariate probability distributions, the empirical distribution is usually identified as the empirical cumulative distribution function (cdf), i.e., the function (of y)

$$\hat{P}_n\left(\omega(x) \leq y; x\right) = \frac{\#\{\omega_i(x) \leq y\}}{n}.$$

Let $T_n(\omega_1(x), \ldots, \omega_n(x))$ denote a statistic, i.e., a real-valued quantity calculated from the sample. Then von Mises [1947] observed that many useful statistics are of the form

$$T_n\left(\omega_1(x), \ldots, \omega_n(x)\right) = T\left(\hat{P}_n(\cdot; x)\right),$$

in which context T is often called a statistical functional. The theory of statistical functionals provides an elegant and useful framework in which to consider stochastic optimization.

Example 1. To recover the standard example from this new perspective, let $\mu : \mathbb{R}^p \to \mathbb{R}$ and $\sigma > 0$ be fixed but unknown. Let

$$\mathcal{P} = \left\{P(\cdot; x) = \text{Normal}\left(\mu(x), \sigma^2\right) \mid x \in \mathbb{R}^p\right\}$$

and let

$$T(P) = \int_{-\infty}^{\infty} \omega P(d\omega).$$

Then

$$f(x) = T\left(P(\cdot; x)\right) = \int_{-\infty}^{\infty} \omega P(d\omega; x) = \mu(x),$$

as desired. One cannot evaluate $\mu(x)$, but one can draw a random sample (3) and use it to estimate $\mu(x)$, e.g., by computing the sample mean,

$$\bar{\omega}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \omega_i(x).$$

In fact, because

$$\sqrt{n}\left[\bar{\omega}_n(x) - \mu(x)\right] \sim \text{Normal}(0, \sigma^2),$$

one can estimate $\mu(x)$ as accurately as one pleases by choosing n sufficiently large. Notice that T is a classic example of a statistical function:

$$T\left(\hat{P}_n(\cdot; x)\right) = \int_{-\infty}^{\infty} \omega \hat{P}_n(d\omega; x) = \frac{1}{n}\sum_{i=1}^{n} \omega_i(x) = \bar{\omega}_n(x).$$

Example 2. There is special interest in stochastic optimization problems that arise when estimating the parameters of a stochastic process that is easily simulated but

analytically intractable. For example, Atkinson, Bartoszynski, Brown, and Thompson [1983] modeled two possible mechanisms for tumor recurrence, metastasis (tumors that grow from cells that break off from a primary tumor and lodge elsewhere in the body) and a systemic mechanism that generates multiple primary tumors. Assume the following:

1. Each tumor originates from a single cell and grows exponentially at rate θ_1.
2. Occurrence of systemic tumors is a Poisson process with rate θ_2.
3. Detection of tumor j is a nonhomogeneous Poisson process with rate $\theta_3 Y_j(t)$, where $Y_j(t)$ is the size of tumor j at time t.
4. Until the removal of the primary tumor, metastasis is a nonhomogeneous Poisson process with rate $\theta_4 Y_0(t)$.

Let $\texttt{Time} \sim P(\cdot;\theta)$ denote the time from detection of the first tumor to detection of the second tumor, where $\theta = \left(\theta_1, \theta_2, \theta_3, \theta_4\right)$. $P(\cdot;\theta)$ is (nearly) intractable, but easily sampled by stochastic simulation. The random variable `Time` was observed for 116 breast cancer patients. Let $\hat{Q}$ denote the empirical distribution of these times and let Δ denote a measure of discrepancy between two probability measures, e.g., the Kolmogorov-Smirnov criterion or the Cramér-von-Mises criterion. One would like to estimate θ by *minimum distance estimation*, i.e., by minimizing

$$f(\theta) = T\left(P(\cdot;\theta)\right) = \Delta\left(P(\cdot;\theta), \hat{Q}\right),$$

but evaluation of f is intractable. Instead, estimate $f(\theta)$ with

$$\hat{f}_n(\theta) = T\left(\hat{P}_n(\cdot;\theta)\right) = \Delta\left(\hat{P}_n(\cdot;\theta), \hat{Q}\right),$$

where $\hat{P}_n$ is the empirical distribution of a simulated sample. With this substitution, the problem of minimum distance estimation becomes a problem of stochastic optimization. Furthermore—and this is the very point that motivated Atkinson et al.—the objective function is sufficiently complicated that it is best treated as a black box.

Example 3. Engineers increasingly rely on computer simulation to develop new products and to understand emerging technologies. In practice, this process is permeated with uncertainty: manufactured products deviate from designed products; actual products must perform under a variety of operating conditions. Most of the computational tools developed for design optimization ignore or abuse the issue of uncertainty, whereas traditional methods for managing uncertainty are often prohibitively expensive.

Robust design optimization (RDO) requires the simultaneous manipulation of design variables and noise variables. Using ideas from statistical decision theory, the problem of robust design can be formulated as an optimization problem. Consider loss functions of the form $L : A \times B \to \Re$, where $a \in A$ represents decision variables, inputs (designs) controlled by the engineer; $b \in B$ represents uncertainty, inputs not controlled by the engineer; and $L(a;b)$ quantifies the loss that accrues from

design a when conditions b obtain. The (unattainable) goal is to find $a^* \in A$ such that, for every $b \in B$,

$$L(a^*; b) \leq L(a; b) \qquad \forall\, a \in A.$$

The unsolvable problem of finding $a^* \in A$ that simultaneously minimizes $L(a; b)$ for each $b \in B$ is the central problem of statistical decision theory: find a decision rule that simultaneously minimizes risk for every possible state of nature. A standard way of negotiating this problem is to replace each $L(a; \cdot)$ with a real valued attribute of it. Thus, Bayes principle results in the optimization problem

$$\min_{a \in A} f(a) = \int_B L(a; b) p(b)\, db, \tag{4}$$

where p denotes a probability density function on B. If f is evaluated by Monte Carlo integration, then (4) becomes a stochastic optimization problem. In previous work, Kugele, Trosset, and Watson [2008] attempted to solve (4) using traditional algorithms for numerical optimization and concluded that they were ineffective. This RDO example has directly available gradient information, which would be used in lieu of the gradient estimation algorithm built into QNSTOP. Thus QNSTOP would have to be modified slightly for problems where gradient information is directly available.

The theory of statistical functionals provides a way to extend many features of the standard example. In what follows, the univariate probability distributions $P(\cdot; x)$ and $\hat{P}_n(\cdot; x)$ are identified with their corresponding cumulative distribution functions. Let

$$D_n = \sup_y \left| \hat{P}_n\left(\omega(x) \leq y; x\right) - P\left(\omega(x) \leq y; x\right) \right|.$$

The Glivenko-Cantelli Theorem states that $P\left(D_n \to 0\right) = 1$; hence, if T is continuous in a suitable sense, then one should find that

$$T\left(\hat{P}_n\left(\cdot; x\right)\right) \xrightarrow{P} T\left(P\left(\cdot; x\right)\right).$$

This says that one can consistently estimate $f(x) = T(P(\cdot; x))$ by sampling from $P(\cdot; x)$. In fact, one can usually say considerably more. The theory of statistical functionals is primarily concerned with connecting the differentiability of T to the asymptotic normality of $T(\hat{P}_n(\cdot; x))$. See, for example, Fernholz [1983].

2.1 Stochastic Approximation Versus Response Surface Methodology

There are two fundamental approaches to solving stochastic optimization problems, stochastic approximation (SA) and response surface methodology (RSM). Both SA and RSM originated in the early 1950s. For SA, the seminal papers are Robbins and Monro [1951], Kiefer and Wolfowitz [1952], Blum [1954], and Dvoretsky [1956]. See Kushner and Yin [1997], Spall [2003], and Marti [2005] for modern surveys. For

RSM, the seminal paper is Box and Wilson [1951]. See Myers and Montgomery [1995] for a modern survey.

Both SA and RSM evolved from attempts to adapt the method of steepest descent for numerical optimization. Both approaches construct local models (typically linear, but occasionally quadratic) of the objective function. Because the objective function cannot be manipulated directly, derivatives are not available and cannot be used to construct the local models. SA constructs local models from estimated derivatives, obtained by finite differencing. RSM constructs local models directly, from designed regression experiments.

In numerical optimization, the magnitude of the differences used in finite differencing schemes is extremely small. When function evaluation is corrupted by random noise, trends in the objective function cannot be detected with such small differences. Furthermore, once a descent direction has been estimated, line searches cannot reliably determine an optimal step length. As a result, SA relies on predetermined decreasing sequences of differences and step length multipliers. Convergence to a local solution is guaranteed by controlling the behavior of these sequences. Traditionally, the differences are $\mathcal{O}(1/k^3)$ and the step length multipliers are $\mathcal{O}(1/k)$, where k is the iteration counter.

SA relies on averaging. The models constructed for individual iterations may be quite crude (Spall's simultaneous perturbation stochastic approximation (SPSA) algorithm estimates a gradient from just two function evaluations); SA succeeds by taking a large number of steps. For fixed budgets, it may be better to choose $n = 1$ in (3) and take a great many steps than to choose $n \gg 1$ and settle for fewer steps of higher quality. One of the most significant advances in SA is due to Polyak and Juditsky [1992], who demonstrated that convergence could by accelerated by using larger step length multipliers and averaging the sequence of iterates.

In contrast, RSM typically takes a small number of carefully chosen steps. Whereas SA has produced a huge literature on asymptotic convergence theory, RSM has produced a huge literature on experimental design. There is virtually no overlap between the SA and RSM literatures.

3. QUASI-NEWTON METHODS FOR STOCHASTIC OPTIMIZATION

Both RSM and SA mimic the method of steepest descent, but numerical optimization has advanced dramatically since the 1950s and the method of steepest descent is no longer the state of the art. Quasi-Newton methods for stochastic optimization (QNSTOP) synthesize ideas from RSM (semilocal approximations constructed from designed experiments by regression, confidence sets for constrained minimizers, ridge analysis) and SA (convergence analysis), combining them with ideas from modern numerical optimization (trust regions, secant updates).

QNSTOP is distinct from, but closely related to, three previous trust region methods for stochastic optimization. First, Lawera and Thompson [1993] described a response surface method based on ideas in [Box and Hunter 1957]. Significant innovations include adaptive experimental designs and quasi-trust region step length control.

Second, Deng and Ferris [2006] proposed three novel modifications to Powell's [2002] UOBYQA (unconstrained optimization by quadratic approximation) algorithm for numerical optimization, endeavoring to adapt it for stochastic optimization. Their algorithm observes the response at each design site multiple times and interpolates the mean responses. A heuristic is used to determine how many observations should be taken at each design site so that the quadratic model and the constrained minimizer are stable. The constrained minimizer of the quadratic model is computed in the same way as in UOBYQA; however, a novel heuristic is used to decide whether to update the current iterate with the minimizer or leave it unchanged. They also describe termination criteria specific to the stochastic setting based upon having similar mean responses amongst a large portion of sites on the boundary of the trust region.

Finally, Chang, Hong, and Wan [2007] and Chang and Wan [2009] proposed the STRONG and STRONG-X algorithms. STRONG assumes normally distributed function evaluation errors, while STRONG-X relaxes this assumption to additive errors with bounded variance. Both algorithms adapt the standard two-phase RSM approach and utilize trust regions to control progress. The first phase constructs a linear model fit partially by least squares to multiple observations at design sites in an appropriate design (the authors recommend a fractional factorial or factorial design plus the current iterate). A line search is used in the direction of negative gradient within the trust region to choose the subsequent iterate. The second phase constructs a quadratic model by least squares. If sufficient progress is made, the algorithm steps to the Cauchy point, i.e., the minimizer of the quadratic in the direction of steepest descent subject to the trust region constraint. Heuristics are used to determine whether sufficient progress was obtained in each phase.

QNSTOP is a class of quasi-Newton methods originally developed for stochastic optimization, but which can also be used for deterministic global optimization with minor variations at certain steps. Both uses supported by the code are described simultaneously in what follows. In iteration k, QNSTOP methods compute the gradient vector $\hat{g}_k$ and Hessian matrix $\hat{H}_k$ of a quadratic model

$$\widehat{m}_k(X - X_k) = \hat{f}_k + \hat{g}_k^T (X - X_k) + \frac{1}{2} (X - X_k)^T \hat{H}_k (X - X_k), \tag{5}$$

of the objective function f centered at X_k, where $\hat{f}_k$ is generally not $f(X_k)$.

In the unconstrained context, QNSTOP methods progress by

$$X_{k+1} = X_k - \left[\hat{H}_k + \mu_k W_k\right]^{-1} \hat{g}_k, \tag{6}$$

where μ_k is the Lagrange multiplier of a trust region subproblem and W_k is a scaling matrix. In the case where the feasible set Θ is a convex subset of $\mathbb{R}^p$, consider an algorithm of the form

$$X_{k+1} = \left(X_k - \left[\hat{H}_k + \mu_k W_k\right]^{-1} \hat{g}_k\right)_\Theta,$$

where $(\cdot)_\Theta$ denotes projection onto Θ.

3.1 Estimating the Gradient

Following a response surface methodology approach, QNSTOP designs regression experiments in a region of interest containing the current iterate. QNSTOP uses an ellipsoidal design region centered at the current iterate $X_k \in \mathbb{R}^p$. Let

$$W_\gamma = \left\{W \in \mathbf{R}^{p\times p} : W = W^T,\ \det(W) = 1,\ \gamma^{-1} I_p \preceq W \preceq \gamma I_p\right\}$$

for some $\gamma \geq 1$ where I_p is the $p \times p$ identity matrix. A typical value for γ is 20. The elements of the set W_γ are valid scaling matrices that control the shape of the ellipsoidal design regions with eccentricity constrained by γ. Let the ellipsoidal design regions

$$E_k(\tau_k) = \left\{X \in \mathbb{R}^p : (X - X_k)^T W_k (X - X_k) \leq \tau_k^2\right\}$$

where $W_k \in W_\gamma$. In the deterministic case $\tau_k = \tau_0 > 0$ is fixed if there is no gain, otherwise for gain $\zeta > 0$ (an input parameter)

$$\tau_k = \frac{\zeta}{\zeta + k}\tau_0.$$

In the stochastic case, the convergence theory requires that τ_k be decayed according to the formula $\tau_k = a(k+1)^{-b}$, where $a > 0$ and $b \in (0, 0.5)$.

In each iteration, QNSTOP methods choose a set of N_k design sites $\{X_{k1}, \ldots, X_{kN_k}\} \subset E_k(\tau_k) \cap \Theta$. In this implementation $N = N_k$ is fixed for each $k = 1, 2, \ldots$ and $X_{k1}, \ldots, X_{kN} \in E_k(\tau_k) \cap \Theta$ are uniformly sampled in each iteration.

Let $Y_k = (y_{k1},\ldots,y_{kN})^T$ denote the N-vector of responses where $y_{ki} = F(X_{ki}) +$ noise. The response surface is modeled by the linear model $y_{ki} = \hat{f}_k + X_{ki}^T \hat{g}_k + \epsilon_{ki}$ where ϵ_{ki} accounts for lack of fit. Let $\bar{X}_k = N^{-1}\sum_{i=1}^N X_{ki}$. The least squares estimate of the gradient $\hat{g}_k$, ignoring the estimate for $\hat{f}_k$, is obtained by observing the responses and solving

$$\left(D_k^T D_k\right)\hat{g}_k = D_k^T Y_k \tag{7}$$

where

$$D_k = \begin{bmatrix} \left(X_{k1} - \bar{X}_k\right)^T \\ \vdots \\ \left(X_{kN} - \bar{X}_k\right)^T \end{bmatrix}.$$

3.2 Updating the Model Hessian Matrix

In the stochastic context, QNSTOP methods constrain the Hessian matrix update to satisfy

$$-\eta I_p \preceq \hat{H}_k - \hat{H}_{k-1} \preceq \eta I_p \tag{8}$$

for some $\eta \geq 0$. Conceptually, this prevents the quadratic model from changing drastically from one iteration to the next. A variation of the SR1 (symmetric, rank one) update $\hat{H}_k$ that satisfies this constraint is computed as the solution to the problem

$$\begin{aligned} &\min_{H \in \mathbf{R}^{p \times p}} \ \left\| H\left(X_k - X_{k-1}\right) - \left(\hat{g}_k - \hat{g}_{k-1}\right) \right\|^2 \\ &\text{subject to } H = H^T, \ \text{rank}\left(H - \hat{H}_{k-1}\right) = 1, \ -\eta I_p \preceq H - \hat{H}_{k-1} \preceq \eta I_p. \end{aligned}$$

This problem has an easily computed explicit solution. However, the constraint (8) is simply relaxed in the deterministic case and the BFGS update is used, i.e., with the Hessian matrix updated according to

$$\hat{H}_k = \hat{H}_{k-1} - \frac{\hat{H}_{k-1} s_k s_k^T \hat{H}_{k-1}}{s_k^T \hat{H}_{k-1} s_k} + \frac{\nu_k \, \nu_k^T}{\nu_k^T \, s_k},$$

where $s_k = X_k - X_{k-1}$, $\nu_k = \hat{g}_k - \hat{g}_{k-1}$.

3.3 Step Length Control

QNSTOP methods utilize an ellipsoidal trust region concentric with the design region for controlling step length. In the deterministic case, the trust region ellipsoid radius ρ_k is taken equal to the design ellipsoid radius τ_k, and the following optimization problem is solved:

$$\min_{X \in E_k(\rho_k)} \quad \hat{g}_k^T \left(X - X_k\right) + \frac{1}{2} \left(X - X_k\right)^T \hat{H}_k \left(X - X_k\right). \tag{9}$$

The solution to (9) is on the arc

$$X(\mu) = X_k - \left[\hat{H}_k + \mu W_k\right]^{-1} \hat{g}_k. \tag{10}$$

It remains to estimate μ_k such that $X(\mu_k)$ solves (9). Using Lemma 6.4.1 from [Dennis and Schnabel 1983] and a little manipulation, it can be established that there is a unique $\mu_k \geq 0$ such that $\|X(\mu_k) - X_k\|_{W_k} = \rho_k$, unless $\|X(0) - X_k\|_{W_k} \leq \rho_k$ in which case $\mu_k = 0$. Estimating μ_k is difficult, but well understood. Chapter 7 in [Conn, Gould, and Toint 2000] is a comprehensive treatment. In particular, Algorithm 7.3.6 in [Conn, Gould, and Toint 2000] is robust and easily implemented.

In the stochastic case, the trust region ellipsoid radius ρ_k is different from the design ellipsoid radius τ_k, but rather than updating the trust region radius ρ_k and then solving for the Lagrange multiplier μ_k from (10), μ_k is directly updated, thereby defining the trust region radius implicitly rather than explicitly. Specifically, fix $c \geq 0$ and $d > \eta\gamma$, set $\mu_k = d(c + k + 1)$, and solve (6) to obtain X_{k+1}, the next iterate. Then $\rho_k = \|X_{k+1} - X_k\|_{W_k}$ is indirectly defined by μ_k. This strategy

is dictated by the convergence theory of Castle [2012] that requires control of the Lagrange multipliers.

3.4 Updating the Experimental Design Region

The QNSTOP approach to constructing the ellipsoidal design regions is now described. To motivate the approach, consider Example 1 with μ quadratic and the problem of minimizing μ subject to an ellipsoidal constraint. If a quadratic model is estimated by least squares regression, then the method of Stablein et al. [1983] can be used to derive a nonlinear inequality that characterizes a confidence set for the constrained minimizer of μ. The confidence set itself is intractable, but a convenient ellipsoidal approximation of it is available.

QNSTOP mimics the construction described above to construct a new ellipsoid from an ellipsoidal trust region subproblem.. Because QNSTOP constructs a linear model by least squares regression, then updates the model Hessian matrix by a secant update, the interpretation of the ellipsoid as a confidence set is somewhat more tenuous. Regardless, the approximation for the covariance matrix of $\nabla \widehat{m}_k(X_{k+1} - X_k)$,

$$V_k = 4\sigma^2 (D_k^T D_k)^{-1}, \tag{11}$$

is computed, where σ^2 is the ordinary least squares estimate of the variance. Then

$$E_{k+1}(\chi_{p,1-\alpha}) = \left\{ X \in \mathbb{R}^p : (X - X_{k+1})^T W_{k+1} (X - X_{k+1}) \leq \chi^2_{p,1-\alpha} \right\},$$

where

$$W_{k+1} = \left(\hat{H}_k + \mu_k W_k \right)^T V_k^{-1} \left(\hat{H}_k + \mu_k W_k \right)$$

and $\chi^2_{p,1-\alpha}$ is the $1-\alpha$ quantile of a chi-squared distribution with p degrees of freedom.

Castle [2012] discovered that strict use of the above updates for W_{k+1} can lead to degenerate ellipsoids. To ensure useful design ellipsoids and guarantee convergence, the constraints $\gamma^{-1} I_p \preceq W_{k+1} \preceq \gamma I_p$ and $\det(W_{k+1}) = 1$ are enforced by modifying the eigenvalues—hence, the definition of $\mathrm{W}_\gamma \ni W_{k+1}$.

3.5 Algorithm Summary

The Fortran code takes as optional arguments all the parameters mentioned above, as well as a few more not mentioned (e.g., one can bound the eccentricity of V_k in (11)). The only required arguments are those defining the problem and a mode—global deterministic or stochastic. Optional arguments not defined default to reasonable values. In both modes it is generally desirable to run QNSTOP from multiple start points, and the code provides several different ways to acquire these start points. The algorithm described below is repeated for each start point.

Step 0 (initialization): Given a function evaluation budget $\tilde{B}$ per start point and operating mode (deterministic or stochastic), set values for $\tau_0 > 0$, $\mu_0 > 0$, $\gamma \geq 1$, $\eta \geq 0$, $\zeta \geq 0$, N, X_0, $k := 0$, $W_0 := \hat{H}_0 := I_p$.

Step 1 (regression experiment): Depending on the mode, compute τ_k. Uniformly sample $\{X_{k1}, \ldots, X_{kN}\} \subset E_k(\tau_k) \cap \Theta$. Observe the response vector $Y_k = (y_{k1}, \ldots, y_{kN})^T$. Compute $\hat{g}_k$ by solving (7).

Step 2 (secant update): If $k > 0$, compute the model Hessian matrix $\hat{H}_k$ using BFGS (deterministic) or SR1 variant (stochastic) update.

Step 3 (update iterate): Compute μ_k depending on the mode as described in Section 3.3, solve $[\hat{H}_k + \mu_k W_k]s_k = -\hat{g}_k$, and compute $X_{k+1} = \big(X_k + s_k\big)_\Theta$.

Step 4 (update subsequent design ellipsoid): Compute $W_{k+1} \in W_\gamma$ using the approach described in Section 3.4.

Step 5: If $(k+2)(N+1)+1 < \tilde{B}$ then increment k by 1 and go to **Step 1**. Otherwise, the algorithm terminates. (f is also observed at each ellipsoid center X_k.)

As a practical matter to deal with variable scaling, the feasible set (box) $\Theta = B = \{x \in \mathbb{R}^p \mid \ell \leqq x \leqq u\}$ is mapped to the unit hypercube $[0,1]^p$ internally by the code, and the algorithm effectively operates on $[0,1]^p$. All input and output is in the original problem coordinate system.

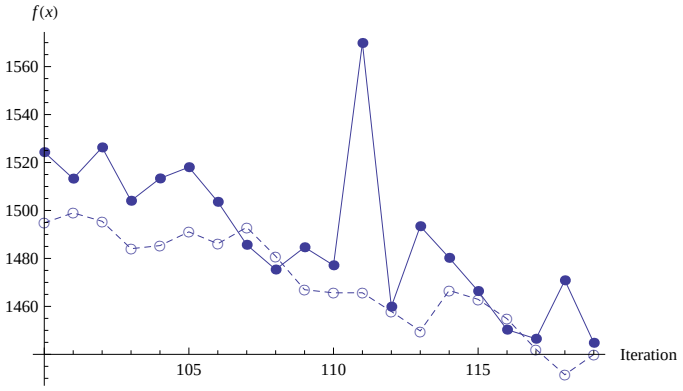

Fig. 1. A typical QNSTOP progression.

Figure 1 shows a typical progression of QNSTOP over 20 iterations, from a difficult biomechanics problem described in [Radcliffe et al. 2010, Easterling et al. 2014]. The solid line represents the lowest value found among 200 design sites for that iteration, while the dotted line represents the corresponding minimum found by the minimizer of the quadratic model. Note that while at times the model will give an imperfect minimum, the overall downward trend is significant.

3.6 Convergence Theory

The convergence theory for QNSTOP [Castle 2012] requires certain assumptions (stated precisely below) and certain conditions on the various parameters (stated earlier in this section in reference to the "stochastic case"). These assumptions

are typical in the analysis of stochastic approximation algorithms [Castle 2012]. Precisely, using the notation in Sections 2 and 3, assume

(1) the decaying τ_k and increasing μ_k have the earlier stated forms for the stochastic case;

(2) the gradient estimate $\hat{g}_k$ used in the quadratic model $\hat{m}_k$ is independent of the gradient estimate $\check{g}_k$ used to construct $\hat{H}_k$ (achieved by having two observed responses at each design site X_{ki} — $\hat{g}_k = \check{g}_k$ has been used in practice with no apparent ill effects);

(3) for each k and design points $\{X_{k1}, \ldots, X_{kN}\} \subset E_k(\tau_k) \cap \Theta$, the scaled design matrix

$$\frac{1}{2\tau_k \gamma^{1/2}} \begin{bmatrix} \big(X_{k1} - \bar{X}_k\big)^T \\ \vdots \\ \big(X_{kN} - \bar{X}_k\big)^T \end{bmatrix}$$

has singular values bounded below by $\Pi > 0$;

(4) $f(x) = T\big(P(\cdot; x)\big)$ with observations $\hat{f}_n(x) = T\big(\hat{P}_n(\cdot; x)\big) = T\big(P(\cdot; x)\big) + \epsilon_x$;

(5) the objective function f is twice continuously differentiable, bounded from below, and $\|\nabla^2 f(x)\| \leq L < \infty$ for some L and all $x \in \mathbb{R}^p$;

(6) the observed errors have zero mean and finite variance, i.e., $E[\epsilon_x] = 0$ and $E[\epsilon_x^2] \leq c_\epsilon$;

(7) the objective function has a unique minimizer θ^*,

$$\inf_{\|x-\theta^*\|>\phi} \|\nabla f(x)\| > 0,$$

and

$$\inf_{\|x-\theta^*\|>\phi} \|f(x) - f(\theta^*)\| > 0$$

for some $\phi > 0$.

Then the iterates X_k generated by QNSTOP converge almost surely to the unique minimizer θ^* of f.

The multivariate Kiefer-Wolfowitz algorithm for stochastic approximation is

$$\theta_{k+1} = \theta_k - \frac{b_k}{2c_k} \begin{pmatrix} \hat{f}_1(\theta_k + c_k e_1) - \hat{f}_1(\theta_k - c_k e_1) \\ \vdots \\ \hat{f}_1(\theta_k + c_k e_p) - \hat{f}_1(\theta_k - c_k e_p) \end{pmatrix},$$

where $e_1, \ldots, e_p$ are unit vectors in the coordinate directions, $c_k > 0$ controls the width of the finite differencing interval, and $b_k > 0$ controls step length. Choosing $\mu_k = 1/b_k$, $\eta = 0$ (entailing $\hat{H}_k = \hat{H}_0$), $\gamma = 1$ (entailing $W_k = I_p$, which results in spherical experimental regions), and $N = 2p$ design sites at $\theta_k \pm c_k e_i$ yields Kiefer-Wolfowitz as a special case of QNSTOP. Allowing $\gamma > 1$ and placing the $2p$ design sites at the endpoints of the resulting ellipsoid's axes permits simulation

experiments that investigate the value of replacing spherical design regions with elliptical regions that adapt to the contours of the objective function. Allowing $\eta > 0$ permits simulation experiments that investigate the value of using (some) second-order information. Castle's [2012] simulation experiments suggest that both innovations have virtue.

4. PARALLEL IMPLEMENTATION

QNSTOP, unlike, say, the massively parallel direct search code VTDIRECT95 [Jones et al. 1993, Jones 2001, Deng and Ferris 2007, He et al. 2009], requires no exotic data structures or sophisticated communication management. There are just three potentially significant sources of parallelism: the individual function evaluations $f(X_{ki})$, the loop ($i = 1, \ldots, N$) over the samples in an experimental design, and the loop over the start points (of size `NSTART`). These three levels are nested. If each evaluation $f(X_{ki})$ were a large scale parallel simulation using MPI, then a master-slave paradigm with the master farming out points X_{ki} to the slaves for evaluation is a reasonable approach entirely based on MPI. If the distributed memory nodes are multicore, then a mixed programming model makes sense, but the shared memory (OpenMP) component would be within the function evaluations, not at the level of the two outer loops. On a large shared memory machine, there will be ample parallelism at the two outer loops, motivating an OpenMP approach.

Due to the exception handling limitations of OpenMP threads, the logical flow of the parallel driver subroutine `QNSTOPP` is significantly different from that of the serial (without OpenMP directives) driver subroutine `QNSTOPS`. Consequently the serial version `QNSTOPS` execution is more efficient than the parallel version `QNSTOPP` execution with a single thread. This is the justification for providing both serial and parallel subroutines, even though in principle the OpenMP code `QNSTOPP` can be run serially.

The parallel (OpenMP) implementation of QNSTOP has four choices for parallelization, controlled by an optional argument to the Fortran subroutine `QNSTOPP`: (1) serial (no parallelization at all, the default), (2) parallelize only the outer loop over the start points, (3) parallelize only the second outermost loop over the experimental design samples, or (4) do both (2) and (3). The choice (4), because of nesting, could generate a very large number of threads, so should be used with care. Figures 2–4 show speedup results for a eukaryotic cell cycle model problem [Oguz et al. 2013] from the systems biology literature. The model is a system of 26 stiff ODEs with 149 parameters. There is experimental data on 119 mutants, each of which corresponds to a modification of the base (or "wild type") system of ODEs, and the optimization problem is to estimate the 149 parameters so as to best fit the experimental data for all the mutants. Each mutant is classified as "viable", "inviable", or "neither", and the objective function value at a particular 149-dimensional parameter vector is simply the (negative) count of how many mutants' behavior is matched by the model. One objective function evaluation $f(X)$ on a single PowerPC G4 processor typically takes about 15 s, but can take an order

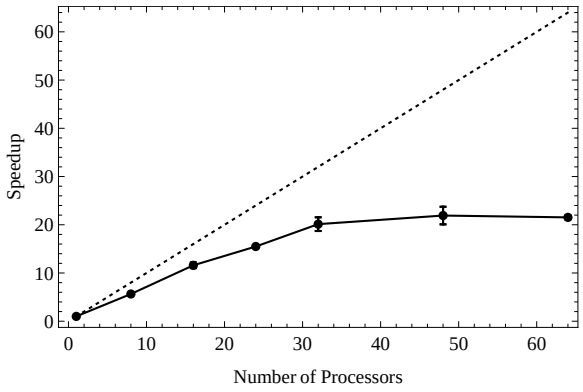

Fig. 2. Speedup of the parallel QNSTOPP over the serial QNSTOPS for the cell cyle problem with OMP = 1 (parallel loop over start points). The mean speedup is plotted with error bars at one standard deviation.

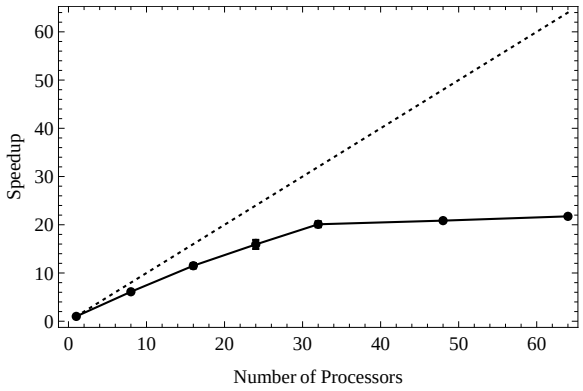

Fig. 3. Speedup of the parallel QNSTOPP over the serial QNSTOPS for the cell cyle problem with OMP = 2 (parallel loop over design ellipsoid sample points). The mean speedup is plotted with error bars at one standard deviation.

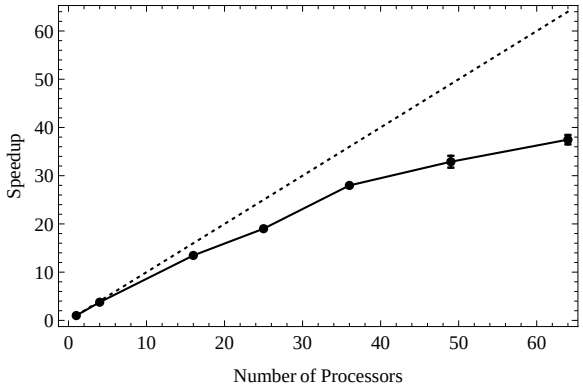

Fig. 4. Speedup of the parallel QNSTOPP over the serial QNSTOPS for the cell cyle problem with OMP = 3 (both OMP = 1 and OMP = 2, nesting). The mean speedup is plotted with error bars at one standard deviation.

of magnitude more depending on the parameter vector, due to the different ODE trajectories (being tracked with LSODAR).

The optional argument `OMP`, referenced in Figs. 2–4, defining the parallel decomposition has the values 1, 2, 3 corresponding to dynamically scheduled loop parallelization over the start points, design ellipsoid sample points, or both, respectively. For these experiments, the number of start points is `NSTART` = 64 and the number of design ellipsoid sample points (at which the objective function is observed) is `N` = 256. Each data point represents the mean of three runs (for which the variance is so small that the point is shown without error bars) or five runs (point shown with error bars). It is not surprising that OpenMP nesting (`OMP` = 3) performs significantly better than no nesting, since there are fewer threads (square root of the total number of threads) at each level of parallelism. The speedup plots (Figs. 2–4) are consistent with Amdahl's Law, and show the limitations of coarse grained parallelization (even with dynamic loop scheduling) when there is limited problem parallelism and the function evaluation times are highly variable (typical of optimization problems with black box simulation function values).

5. PERFORMANCE

The systems biology literature on cell cycle models contains a parameter vector X^0 (called the TL set) obtained by biochemistry knowledge and manual twiddling, considered in the ballpark of the correct values. Searches for the optimal parameter vector generally are conducted in a box defined by X^0 plus or minus some percent of X^0, say 20%, 40%, 90% defining the boxes $[0.8\,X^0, 1.2\,X^0]$, $[0.6\,X^0, 1.4\,X^0]$, $[0.1\,X^0, 1.9\,X^0]$, respectively. For the particular model known as "Oak's deterministic model" [Oguz et al. 2013], the best known value of $f(X)$ is -110 (obtained using LSODAR, or -111 obtained using a less accurate fixed step Euler method as done by Oguz et al. [2013]), where $f(X^0) = -73$. Using `NSTART` = 84 and `N` = 225 (from the statistical rule of thumb that at least $1.5p$ data points are needed to estimate p parameters, and the model gradient $\hat{g}_k$ here has dimension $p = 149$), Figs. 5–7 show the iteration histories for three start points (out of 84) for each of the three ±20%, ±40%, ±90% boxes, running QNSTOP in deterministic global optimization mode with the other relevant algorithm parameters shown in the figure legends. These legends list the subroutine `QNSTOP[P|S]` arguments: `N` is the number of design ellipsoid sample points; `TAU` is the initial ellipsoid radius τ_0; `GAIN` is the gain ζ (cf. §3.1); [`LB`, `UB`] is the feasible box; `SWITCH` controls how start points are provided, with values 1, 2, 3, 4 corresponding to a single start point `XI`, a given list of start points, an automatically generated Latin hypercube design (containing `XI`) of start points, adaptive generation of a sequence of start points (beginning

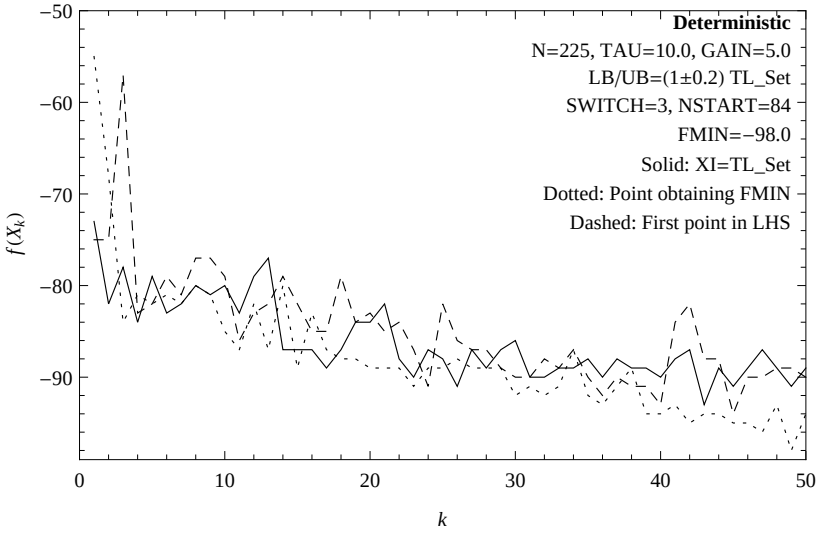

Fig. 5. Execution traces of QNSTOP in deterministic mode for three start points in the ±20% box. One trace starts at the center of the box (where $f(X) = -73$) and another trace contains the best point of the entire run (where $f(X) = -98$). Another run with TAU = 2.2 (scaled from TAU = 10.0 for the ±90% box) yielded a best value of −97.

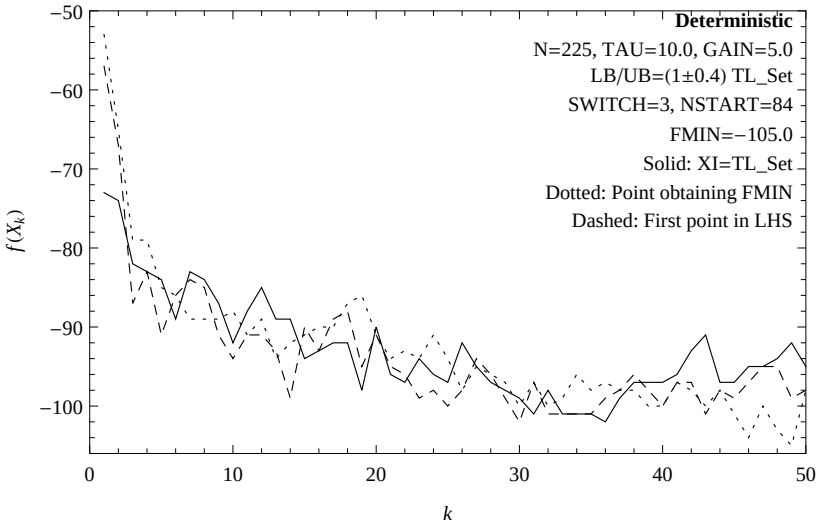

Fig. 6. Execution traces of QNSTOP in deterministic mode for three start points in the ±40% box. One trace starts at the center of the box (where $f(X) = -73$) and another trace contains the best point of the entire run (where $f(X) = -105$). Another run with TAU = 4.4 (scaled from TAU = 10.0 for the ±90% box) yielded a best value of −104.

with `XI`) by a user provided procedure, respectively; `NSTART` is the number of start points (for `SWITCH` = 3 or 4); `XI` is the initial specified start point.

The trajectories for *all* start points are similar to the general downward trend of the three start point trajectories shown. The best values found for $f(X)$ during the

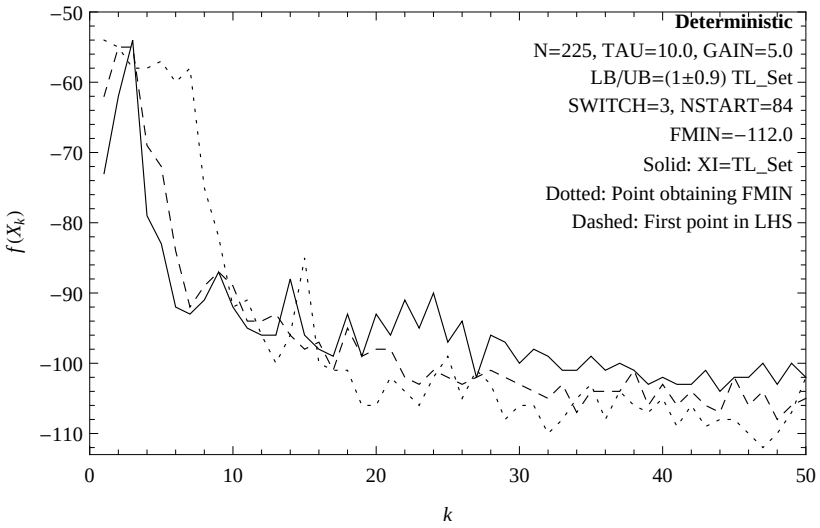

Fig. 7. Execution traces of QNSTOP in deterministic mode for three start points in the ±90% box. One trace starts at the center of the box (where $f(X) = -73$) and another trace contains the best point of the entire run (where $f(X) = -112$).

Table I. Statistics for Best $f(X)$ Value Found with Each of the 84 Starting Points, for Each of the ±20%, ±40%, ±90% Boxes.

box	min	median	max	$\bar{\sigma}$	mode
±20%	−98	−92	−88	1.97	G
±40%	−105	−100	−95	2.19	G
±90%	−112	−105	−55	7.42	G
±90%	−109	−101	−55	18.88	S

three runs for the ±20%, ±40%, ±90% boxes were −98, −105, −112, respectively, improving on the best known value in the literature. For the runs depicted in Figs. 5–7, Table I gives the statistics for the best $f(X)$ value found with each of the 84 starting points. The global deterministic (stochastic) mode is denoted by 'G' ('S').

Figure 8 shows a trace plot for the stochastic mode (S) for the ±90% box similar to Fig. 7 for the global deterministic mode (G), and the statistics for that stochastic mode run are included in Table I. Execution traces and statistics for the stochastic mode for the ±20% and ±40% boxes are what would be expected for these smaller boxes, and thus are omitted. Since the stochastic mode has to protect against unknown random fluctuations, the convergence is much slower than for the global deterministic mode (for this deterministic cell cycle problem). Castle [2012] reports results for QNSTOP in stochastic mode applied to a truly stochastic tumor growth model.

Comparison of QNSTOP to other algorithms, both deterministic and nondeterministic, is not done here since that has already been done in the literature [Easterling et al. 2014] for some very hard "noisy" scientific optimization problems.

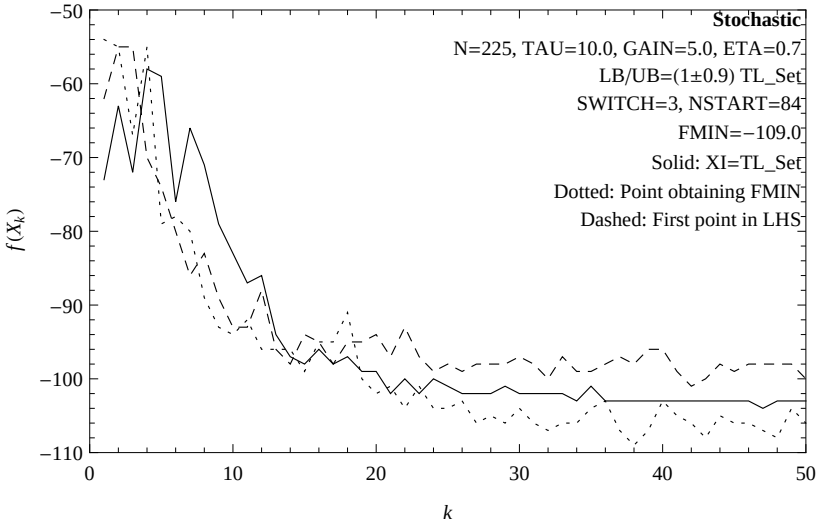

Fig. 8. Execution traces of QNSTOP in stochastic mode for three start points in the ±90% box. One trace starts at the center of the box (where $f(X) = -73$) and another trace contains the best point of the entire run (where $f(X) = -109$).

BIBLIOGRAPHY

Atkinson, E. N., Bartoszyński, B., Brown, B. W., and Thompson, J. R. 1983. Simulation techniques for parameter estimation in tumor related stochastic processes. In *Proc. 1983 Computer Simulation Conference*, North-Holland, New York, 754–757.

Blum, J. R. 1954. Multidimensional stochastic approximation methods. *Annals of Mathematical Statistics 25*, 737–744.

Box, G. E. P. and Hunter, J. S. 1957. Multi-factor experimental designs for exploring response surfaces. *Annals of Mathematical Sciences 28*, 195–241.

Box, G. E. P. and Wilson, K. B. 1951. On the experimental attainment of optimum conditions. *J. Royal Statistical Society, Series B, 13*, 1–45.

Castle, B. S. 2012. Quasi-Newton methods for stochastic optimization and proximity-based methods for disparate information fusion. Ph.D. thesis, Indiana University, Bloomington, IN.

Chang, K. H., Hong, L. J. and Wan, H. 2007. Stochastic trust region gradient-free method (STRONG) —a new response-surface-based algorithm in simulation optimization. In *Proceedings of the 2007 Winter Simulation Conference*, S. G. Henderson, B. Biller, M. H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, eds., 346–354.

Chang, K. H. and Wan, H. 2009. Stochastic trust region response surface convergent method for generally-distributed response surface. In *Proceedings of the 2009 Winter Simulation Conference*, M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds., 563–573.

Conn, A. R., Gould, N. I. M., and Toint, P. L. 2000. *Trust-Region Methods*. MPS-SIAM Series on Optimization, SIAM, Philadelphia.

Deng, G. and Ferris, M. C. 2006. Adaptation of the UOBYQA algorithm for noisy functions. In *Proceedings of the 2006 Winter Simulation Conference*, L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds., 312–319.

Deng, G. and Ferris, M. C. 2007. Extension of the DIRECT optimization algorithm for noisy functions. In *Proceedings of the 2007 Winter Simulation Conference*, S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, eds., 497–504.

Dennis, J. E. and Schnabel, R. B. 1983. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, New Jersey.

Diggle, P. J. and Gratton, R. J. 1984. Monte Carlo methods of inference for implicit statistical models. *J. Royal Statistical Society, Series B, 46*, 193–227.

Dvoretsky, A. 1956. On stochastic approximation. In *Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 39–55.

Easterling, D. R., Watson, L. T., Madigan, M. L., Castle, B. S., and Trosset, M. W. 2014. Parallel deterministic and stochastic global minimization of functions with very many minima. *Comput. Optim. Appl. 57*, 2, 469–492.

Fernholz, L. T. 1983. *von Mises Calculus for Statistical Functionals*. Springer-Verlag, New York.

Gilmore, P. and Kelley, C. T. 1995. An implicit filtering algorithm for optimization of functions with many local minima. *SIAM J. Optim. 5*, 2, 269–285.

He, J., Watson, L. T., and Sosonkina, M. 2009. Algorithm 897: VTDIRECT95: serial and parallel codes for the global optimization algorithm DIRECT. *ACM Trans. Math. Software 36*, Article 17, 1–24.

Jones, D. R. 2001. The DIRECT global optimization algorithm. In *Encyclopedia of Optimization*, Vol. 1, Kluwer Academic Publishers, Dordrecht, 431–440.

Jones, D. R., Pertunen, C. D., and Stuckman, B. E. 1993. Lipschitzian optimization without the Lipschitz constant. *J. Optimization Theory and Applications 79*, 1, 157–181.

Kiefer, J. and Wolfowitz, J. 1952. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics 23*, 462–466.

Kugele, S. C., Trosset, M. W., and Watson, L. T. 2008. Numerical integration in statistical decision-theoretic methods for robust design optimization. *Structural Multidisciplinary Optim. 36*, 457–475.

Kushner, H. J. and Yin, G. G. 1997. *Stochastic Approximation Algorithms and Application*. Springer, New York.

Lawera, M. and Thompson, J. R. 1993. A parallelized, simulation based algorithm for parameter estimation. In *Proceedings of the Thirty-Eighth Conference on the Design of Experiments in Army Research Development and Testing*, B. Bodt, ed., 321–341.

Marti, K. 2005. *Stochastic Optimization Methods*. Springer, Berlin.

Myers, R. H. and Montgomery, D. C. 1995. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. John Wiley & Sons, New York.

Oguz, C., Laomettachit, T., Chen, K. C., Watson, L. T., Baumann, W. T., and Tyson, J. J. 2013. Optimization and model reduction in the high dimensional parameter space of a budding yeast cell cycle model. *BMC Systems Biol. 7:53*, 1–17.

Polyak, B. T. and Juditsky, A. B. 1992. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optimization 30*, 838–855.

Powell, M. J. D. 2002. UOBYQA: Unconstrained optimization by quadratic approximation. *Math. Prog. 92*, 555–582.

Radcliffe, N. R., Easterling, D. R., Watson, L. T., Madigan, M. L., and Bieryla, K. A. 2010. Results of two global optimization algorithms applied to a problem in biomechanics. in *Proc. 2010 Spring Simulation Multiconference, High Performance Computing Symp*, A. Sandu, L. Watson, and W. Thacker (eds), Soc. for Modelling and Simulation Internat., Vista, CA, 117–123.

Robbins, H. and Monro, S. 1951. A stochastic approximation method. *Annals of Mathematical Statistics 22*, 400–407.

Spall, J. C. 1987. A stochastic approximation technique for generating maximum likelihood parameter estimates. In *Proc. American Control Conference*, Minneapolis, MN, June 10-12, 1161–1167.

Spall, J. C. 1992. Multivariate stochastic approximation using simultaneous perturbation gradient approximation. *IEEE Trans. Autom. Control 37*, 3, 332–341.

Spall, J. C. 1998. Implementation of the simultaneous perturbation algorithm for stochastic optimization. *IEEE Trans. Aerospace Electronic Systems 34*, 3, 817–823.

Spall, J. C. 2003. *Introduction to Stochastic Search and Optimization*. John Wiley & Sons, New York.

Stablein, D. M., Carter, Jr., W. H., and Wampler, G. L. 1983. Confidence regions for constrained optima in response-surface experiments. *Biometrics 39*, 759–763.

Thompson, J. R. 2000. *Simulation: A Modeler's Approach*. John Wiley & Sons, New York, NY.

von Mises, R. 1947. On the asymptotic distribution of differentiable statistical functions. *Annals of Mathematical Statistics 18*, 309–348.