

ARGONNE NATIONAL LABORATORY
9700 South Cass Avenue
Argonne, Illinois 60439

Non-intrusive Termination of Noisy Optimization *

Jeffrey Larson[†] and Stefan M. Wild[‡]

Mathematics and Computer Science Division

Preprint ANL/MCS-P1887-0511.

May 2011

*This work was supported by the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract DE-AC02-06CH11357.

[‡]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439.

[†]Department of Mathematical & Statistical Sciences, University of Colorado Denver, Denver, CO 80217.

Contents

1	Introduction and Motivation	1
2	Background	3
3	Stopping Tests	5
3.1	$f_i^{*/f}$ test	6
3.2	Max-Difference- f test	6
3.3	Max-Distance- x test	7
3.4	Max-Distance- x_i^* test	7
3.5	Max-Budget test	7
3.6	Tests based on estimates of the noise	7
3.7	Relationship to loss functions	8
4	Numerical Experiments	9
4.1	Accuracy profiles for the ϕ_1 family	10
4.2	Performance profiles for the ϕ_1 family	12
4.3	Accuracy and performance plots for the ϕ_2 family	14
4.4	Across-family comparisons	15
4.5	Deterministic noise	16
5	Conclusions	16

Non-intrusive Termination of Noisy Optimization*

May 31, 2011

Jeffrey Larson and Stefan M. Wild

Abstract

Significant savings can be gained from terminating the optimization of a computationally expensive function well before traditional criteria, such as a maximum budget of evaluations, are satisfied. Early termination is especially desirable for noisy functions, where a solver could potentially proceed indefinitely while seeing changes insignificant relative to the noise. In this paper we consider general termination tests that can be used in conjunction with any solver’s built-in termination criteria. We propose parameterized families of termination tests, analyze their properties, and illustrate how they can employ an estimate of the function’s noise level. Using a set of benchmark problems with both stochastic and deterministic noise, we compare the tests and their sensitivities to parameters in terms of both accuracy and efficiency. Recommendations are made for using the proposed tests in practice.

1 Introduction and Motivation

The optimization of real-world, computationally expensive functions invariably leads to the difficult question of when an optimization procedure should be terminated. Algorithm developers and the mathematical optimization community at large typically assume that the optimization is terminated when either a measure of criticality (gradient norm, mesh size, etc.) is satisfied or a user’s computational budget (number of evaluations, wall clock time, etc.) is exhausted. This assumption is made largely for convenience and for generalizability across many problem domains, since the latter condition allows a user to assert full control over the optimization.

For a large class of problems, however, the user may not have a well-defined computational budget and instead demand a termination test t solving

$$\begin{aligned} \min_t \quad & \text{Computational expense}(t) \\ \text{s.t.} \quad & \text{Acceptable accuracy of the solution}(t), \end{aligned} \tag{1}$$

with the criticality measure of the solver employed typically chosen with the accuracy constraint in mind. Examples of such accuracy-based criticality tests are discussed in detail by Gill, Murray, and Wright [9, Section 8.2.3].

*Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439. This work was supported by the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract DE-AC02-06CH11357. Performed while the first author was visiting the Mathematics and Computer Science Division at Argonne National Laboratory.

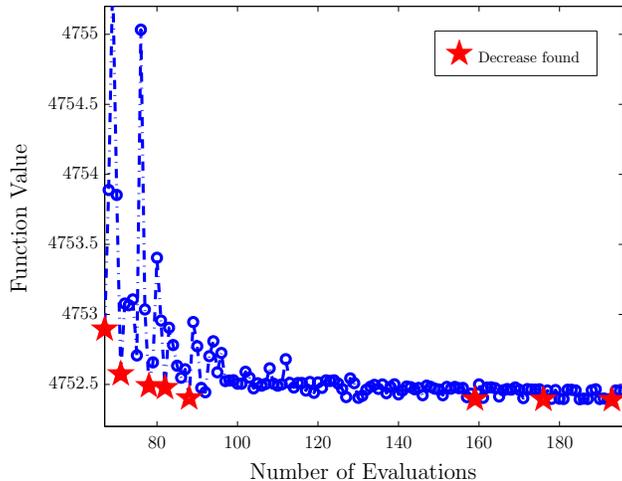


Fig. 1: Noisy trajectory of function values for an expensive nuclear physics problem.

The main difficulties arising from this approach are a result of (1) possibly being poorly formulated. The computational expense could be unbounded because an a priori user-defined accuracy is unrealistic for the problem/solver pair or, worse still, unknown. Furthermore, a user may have difficulty translating the criticality measures provided by a solver, which are generally based on assumptions of smoothness and infinite-precision calculations, into practical metrics on the solution accuracy.

In Fig. 1 we illustrate the challenges in this area with an example from nuclear physics, similar to the minimization problems considered in [16]. Each of the function values shown is obtained from running a deterministic simulation for one minute on a 640-core cluster. Stopping the optimization sooner than 200 function evaluations not only would return a solution faster but also would free the cluster for other applications and/or result in a savings in energy, an increasingly crucial factor in high-performance computing.

If we assume that the optimization shown in Fig. 1 has not been terminated by a solver’s criticality measures or a user’s computational budget, the question is then whether termination should occur for other reasons. For example, if only the first three digits of the simulation output were computed stably, one may want to terminate the optimization sooner (perhaps even before the data from this figure were generated) than if computational noise corrupted only the eighth digit of the output. Alternatively, the behavior shown could mean the solver in question has stagnated (because of noise, errors in the simulation, a limitation of the solver, etc.), and hence examining the solution and/or restarting the optimization could be a more effective use of the remaining computational budget. Wright [23] refers to this stalled progress as *perseveration* and notes that there is “no fully general way to define ‘insufficient progress.’ ” Even so, it may be advantageous to use knowledge of the uncertainty or accuracy of a given function evaluation when making such a decision.

In the remainder of this paper we explore these issues and propose termination criteria

that can be easily incorporated on top of a user’s solver of choice. In [8], Fletcher summarizes the challenges at hand (in the case of round-off errors alone):

Some consideration has to be given to the effects of round-off near the solution, and to terminate when it is judged that these effects are preventing further progress. It is difficult to be certain what strategy is best in this respect.

Moreover, Gill, Murray, and Wright [9] stress that

no set of termination criteria is suitable for all optimization problems and all methods.

This sentiment is shared by Powell [19] who says

it is believed that it is impossible to choose such a convergence criterion which is effective for the most general function ... so a compromise has to be made between stopping the iterative procedure too soon and calculating f an unnecessarily large number of times.

Consequently, we will consider tests that allow for the use of estimates of the noise particular to a problem. Furthermore, our criteria are not intended as substitutes for a computational budget or a solver’s built-in criticality tests, which we consider to be important safeguards. Likewise, the termination problem can be viewed as a real-time control problem depending on complete knowledge of the solver’s decisions, but we resist this urge for purposes of portability and applicability.

We provide background on previous work and introduce notation in Section 2. The families of stopping tests we propose in Section 3 do not provide guarantees on the quality of the solution, although doing so may be the role of a solver’s built-in criteria. Instead, the proposed tests are parameterized in order to quantify a user’s trade-off between the benefit of achieving additional decrease and the cost of additional evaluations, while requiring a minimal amount of information from the solver. Equally important, our results in Section 4 comparing the quality of these families of stopping tests focus on local optimization. While our results can be incorporated in a local subroutine of any global search algorithm, the tests proposed in Section 3 are unable to distinguish between exploration and refinement phases in their current form. We summarize our results in Section 5 and provide recommendations when implementing these tests.

2 Background

Our preliminary discussion is limited to optimization methods that do not explicitly require derivative information. While our work can be extended to incorporate noisy gradient information, the derivatives of noisy functions are typically even noisier than the function.

Derivative-free optimization methods are often favored for their perceived ability to handle noisy functions. Although asymptotic convergence of these methods is generally proved assuming a smooth function, adjustments are frequently made to accommodate

noise. For example, in the case of stochastic noise, replications of function evaluations can be used to modify existing methods (e.g., [4] modifying UOBYQA [20], [5, 1] modifying DIRECT [13], and [22] modifying Nelder-Mead (see, e.g., [3])). However, stopping criteria for these methods involve limited knowledge of the noise and indicate the wide variety of stopping tests used in practice. In [1], optimization is stopped when adjacent points are within 10^{-4} of each other, whereas [5] allows stopping when the best function value has not been improved after some number of consecutive iterations. To limit the number of stochastic replications, the authors of [4] and [22] adjust the maximum number of allowed replications at a particular point based on the variance of the noise.

Deterministic noise is far less understood than its stochastic counterpart [18]. Not surprising, even less knowledge of the magnitude of noise is used for problems with deterministic objectives. When low-amplitude noise is present, Kelley [15] proposes a restart technique for Nelder-Mead but terminates when sufficiently small differences exist in the simplicial function values, independent of the magnitude of the noise. Implicit filtering [14] has numerous termination possibilities (small function value differences on a stencil, a small change in the best function value from one iteration to the next, etc.) but none that are explicitly related by the author to the magnitude of the noise. A similar implicit relationship to noise can be seen in [10], where treed Gaussian process models for optimization are terminated when a maximum improvement statistic is sufficiently small. The authors of SNOBFIT [12] suggest stopping when the best point has not changed for a number of consecutive SNOBFIT calls.

Our work more closely follows that of Gill et. al [9], where an entire section (8.2) is devoted to properties of the computed solution. The authors there recommend terminating Nelder-Mead-like algorithms when the maximum difference between function values on the simplex is less than a demanded accuracy weighted by the best function value on the simplex.

The only other direct relationship between stopping criteria and a magnitude of noise that we are aware of are in [18, Section 9] and [11]. In [18], a stochastic model of the noise is used to estimate the *noise level* of a function value $f(x)$ by difference table-based approximations of the standard deviation $(\text{Var}\{f(x)\})^{1/2}$. Results are validated for deterministic f . As an example application, the authors terminate a Nelder-Mead method on an ODE-based problem when consecutive decreases are less than a factor of the noise level. The authors of [11] perturb bound-constrained problems so the incumbent iterate is the exact solution to this new problem. An algorithm can then be terminated when the size of this perturbation first decreases below the error in the problem.

Before proceeding, we define the notation employed throughout. We let \mathbb{R}_+ denote the nonnegative reals and \mathbb{N} denote the natural numbers. We let $\{x_1, \dots, x_m\} \subset \mathbb{R}^n$ and $\{f_1, \dots, f_m\} \in \mathbb{R}$ be a sequence of points and corresponding function values produced by a local minimization solver, and we collect the data from the first i evaluations in $\mathcal{F}_i = \{(x_1, f_1), \dots, (x_i, f_i)\}$. The best function value in the first i evaluations is given by $f_i^* = \min_{1 \leq j \leq i} \{f_j\}$, with x_i^* denoting the point corresponding to f_i^* . Accordingly, the sequence $\{f_i^*\}$ is nonincreasing. Unless otherwise stated, $\|\cdot\|$ denotes the standard Euclidean distance.

We let $\hat{\varepsilon}_{i_r}$ be an estimate of the relative noise at f_i . This estimate may come from experience, numerical analysis of the underlying processes in computing f_i , or appropriate scaling (by $|f_i|$) of the noise-level estimates from the method proposed in [18]. In the case of stochastic noise, $\hat{\varepsilon}_{i_r}$ can be viewed as the standard deviation of f_i relative to the magnitude of f_i .

Favorable properties of a termination test include scale and shift invariance, so that the test would terminate after the same number of evaluations for any affine transformation of the objective function. We use the following proposition to aid in the subsequent analysis of scale and shift invariance.

Proposition 2.1 *The quantity $\hat{\varepsilon}_{i_r}$ is scale invariant in f , and the product $\hat{\varepsilon}_{i_r}|f_i|$ is shift invariant in f .*

3 Stopping Tests

In this section we define families of termination tests and provide motivation for their use. Each family can be defined through an extended-value function ϕ mapping to $\mathbb{R} \cup \{+\infty\}$. The associated termination test stops after i^* evaluations, where i^* is the solution to the hitting problem

$$\min_i \{i : \phi(\mathcal{F}_i; \nu_{\mathcal{F}_i}, \eta) \leq 0\}. \quad (2)$$

Members of a family of tests are determined by different values of the parameter vector $(\nu_{\mathcal{F}_i}, \eta)$, with $\nu_{\mathcal{F}_i}$ and η denoting parameters that are (possibly) dependent on \mathcal{F}_i and independent of \mathcal{F}_i , respectively. Since ϕ quantifies the progress of an algorithm (through the history of function values and points), each family of tests is designed to determine when continuing with the present course is likely wasteful as measured by the parameters in $(\nu_{\mathcal{F}_i}, \eta)$.

It is often useful to consider how a test will change if the underlying function undergoes an affine change. We will say that a test is *scale invariant* if

$$\phi(\mathcal{F}_i; \nu_{\mathcal{F}_i}, \eta) = \phi(\alpha\mathcal{F}_i; \nu_{\alpha\mathcal{F}_i}, \eta) \quad \forall \alpha > 0,$$

where $\alpha\mathcal{F}_i \equiv \{(x_1, \alpha f_1), \dots, (x_i, \alpha f_i)\}$. Similarly, we will call a test *shift invariant* if

$$\phi(\mathcal{F}_i; \nu_{\mathcal{F}_i}, \eta) = \phi(\mathcal{F}_i + \beta; \nu_{\mathcal{F}_i + \beta}, \eta) \quad \forall \beta,$$

where $\mathcal{F}_i + \beta \equiv \{(x_1, f_1 + \beta), \dots, (x_i, f_i + \beta)\}$. Similar affine changes to $\{x_1, \dots, x_i\}$ could be considered but are not central to the present discussion, and hence all notions of invariance here are relative to the function f .

Similarly, it is useful to consider whether ϕ is monotone in some of its parameters. Monotonicity of ϕ is desirable because it results in the same form of monotonicity for the corresponding number of evaluations i^* . For example, if ϕ is monotonically increasing in a scalar parameter η , then increasing η results in a more conservative test because the solution to (2) is at least as large. As a consequence, if ϕ is monotonically increasing in η and the

test $\phi(\cdot; \cdot, \eta_1)$ is always satisfied on a set of problems, it is not necessary to consider $\eta > \eta_1$ values on that set of problems.

We now define several families of termination tests and discuss their properties and underlying motivation. All of these tests assume no knowledge of the inner workings of the algorithm they are terminating, but such knowledge might lead to appropriate modifications. For example, if the method uses a simplex, rather than stopping when the last κ function evaluations are within a factor of the noise, one could stop when the last κ simplex vertices are within a factor of the noise (essentially a modification of the proposed rule in [9]).

3.1 $f_i^{*'} test$

$$\phi_1(\mathcal{F}_i; \nu_{\mathcal{F}_i}, \kappa, \mu) \equiv \begin{cases} \frac{f_{i-\kappa+1}^* - f_i^*}{\kappa} - \mu |f_i^*| \nu_{\mathcal{F}_i} & \text{if } i \geq \kappa, \\ \infty & \text{else,} \end{cases} \quad \text{with } \nu_{\mathcal{F}_i}, \mu \in \mathbb{R}_+, \kappa \in \mathbb{N}. \quad (3)$$

This family of tests is designed to stop when the average relative change in f^* over the last κ evaluations is less than $\mu \nu_{\mathcal{F}_i}$. The integer κ can be thought of as a backward difference parameter for estimating the change in the best function value with respect to the number of evaluations.

We note that ϕ_1 is monotonically decreasing in μ since, for fixed κ , \mathcal{F}_i , and $\nu_{\mathcal{F}_i}$,

$$\mu_1 \leq \mu_2 \implies \phi_1(\mathcal{F}_i; \nu_{\mathcal{F}_i}, \kappa, \mu_1) \geq \phi_1(\mathcal{F}_i; \nu_{\mathcal{F}_i}, \kappa, \mu_2).$$

ϕ_1 is also monotonically decreasing in $\nu_{\mathcal{F}_i}$ but is not monotone in κ . Members of this family are scale invariant provided that $\nu_{\mathcal{F}_i}$ is, and shift invariant provided that $|f_i^*| \nu_{\mathcal{F}_i}$ is.

We consider two special cases. When $\nu_{\mathcal{F}_i} = 1$ (or any constant), we obtain tests that are scale invariant but not shift invariant and stop if the average relative change in the best function value drops below μ . If $\nu_{\mathcal{F}_i} = \hat{\varepsilon}_{i_r}$, the tests are scale and shift invariant by Proposition 2.1 and stop an algorithm if the average relative change becomes less than a factor μ times the relative noise.

3.2 Max-Difference- f test

$$\phi_2(\mathcal{F}_i; \nu_{\mathcal{F}_i}, \kappa, \mu) \equiv \begin{cases} \max_{i-\kappa+1 \leq j \leq i} |f_j - f_i^*| - \mu |f_i^*| \nu_{\mathcal{F}_i} & \text{if } i \geq \kappa, \\ \infty & \text{else,} \end{cases} \quad \text{with } \nu_{\mathcal{F}_i}, \mu \in \mathbb{R}_+, \kappa \in \mathbb{N}. \quad (4)$$

This family of tests stops when κ consecutive function values are within $\mu |f_i^*| \nu_{\mathcal{F}_i}$ of f_i^* .

One can show that ϕ_2 is monotonically decreasing in both μ and $\nu_{\mathcal{F}_i}$ and monotonically increasing in κ since

$$\kappa_1 \leq \kappa_2 \implies \max_{i-\kappa_1 \leq j \leq i} |f_j - f_i^*| \leq \max_{i-\kappa_2 \leq j \leq i} |f_j - f_i^*|.$$

We also note that if ϕ_2 is modified so that f_j is replaced by f_j^* , we obtain a test equivalent to $\phi_1(\mathcal{F}_i; \nu_{\mathcal{F}_i}, \kappa, \mu)$. Members of this family are scale invariant provided that $\nu_{\mathcal{F}_i}$ is, and shift invariant provided that $|f_i^*| \nu_{\mathcal{F}_i}$ is.

We examine two special cases. If $\nu_{\mathcal{F}_i} = 1$ (or any constant), ϕ_2 is scale invariant but not shift invariant; this family $\phi_2(\mathcal{F}_i; 1, \kappa, \mu)$ terminates when the last κ function values differ by less than a factor μ relative to the best function value so far. If $\nu_{\mathcal{F}_i} = \hat{\varepsilon}_{i_r}$, the resulting tests are scale and shift invariant (by Proposition 2.1) and terminate when the relative change in the last κ function values is less than a factor μ of the noise.

3.3 Max-Distance- x test

$$\phi_3(\mathcal{F}_i; \kappa, \mu) \equiv \begin{cases} \max_{i-\kappa+1 \leq j, k \leq i} \|x_j - x_k\| - \mu & \text{if } i \geq \kappa \\ \infty & \text{else,} \end{cases} \quad \text{with } \mu \in \mathbb{R}_+, \kappa \in \mathbb{N}. \quad (5)$$

This family stops when κ consecutive x -values are within a distance μ of each other and is analyzed with ϕ_4 below.

3.4 Max-Distance- x_i^* test

$$\phi_4(\mathcal{F}_i; \kappa, \mu) \equiv \begin{cases} \max_{i-\kappa+1 \leq j \leq i} \|x_j^* - x_i^*\| - \mu & \text{if } i \geq \kappa, \\ \infty & \text{else,} \end{cases} \quad \text{with } \mu \in \mathbb{R}_+, \kappa \in \mathbb{N}. \quad (6)$$

This family stops when κ consecutive x_i^* -values are within a distance μ of each other. In general, members of both of the families defined by ϕ_3 and ϕ_4 are not scale (shift) invariant unless the procedure generating $\{x_i\}_i$ is scale (shift) invariant in f . Both ϕ_3 and ϕ_4 are monotonically decreasing in μ and monotonically increasing in κ . We examined a test using $\max_{i-\kappa+1 \leq j \leq i} \|x_j - x_i^*\|$ but found the performance to be similar to that of ϕ_3 .

3.5 Max-Budget test

$$\phi_5(\mathcal{F}_i; \kappa) \equiv \begin{cases} 0 & \text{if } i \geq \kappa, \\ \infty & \text{else,} \end{cases} \quad \text{with } \kappa \in \mathbb{N}. \quad (7)$$

As a point of reference, we include the family corresponding to stopping after a budget of κ evaluations. This commonly used test is trivially scale and shift invariant.

3.6 Tests based on estimates of the noise

The families of tests introduced above have been broadly parameterized to capture a wide range of behaviors. We now provide motivation for using an estimate of the noise in some of tests.

Using $\nu_{\mathcal{F}_i} = \hat{\varepsilon}_{i_r}$ in the special cases of ϕ_1 and ϕ_2 has the benefit of the resulting tests being both scale and shift invariant. Furthermore, the first term in the definition of ϕ_1 and ϕ_2 is strongly correlated with the magnitude of the noise. This feature is illustrated in Fig. 2, which demonstrates running a Nelder-Mead method on a 10-dimensional convex quadratic for levels of stochastic relative noise differing by an order of magnitude. Fig. 2 (left) shows the first term of ϕ_1 , $\frac{f_{i-\kappa+1}^* - f_i^*}{\kappa}$, plotted as a function of the number of evaluations

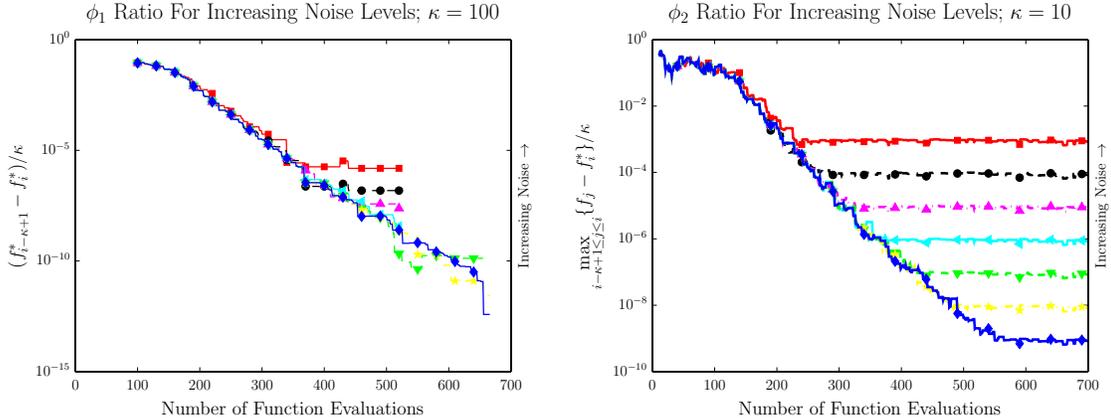


Fig. 2: First terms in ϕ_1 (left, with $\kappa = 100$) and ϕ_2 (right, with $\kappa = 10$) (on a \log_{10} scale) when minimizing a 10-dimensional convex quadratic with stochastic relative noise of different magnitudes. The asymptotes of the quantities shown tend to be separated by the differences in magnitudes of the noise.

i. Here we see that the quantity generally flattens out at increments separated by the same order of magnitude as the seven noise levels. This correlation is even more evident in Fig. 2 (right) when the first term of ϕ_2 , $\max_{i-\kappa+1 \leq j \leq i} |f_j - f_i^*|$, is considered.

Consequently, in the numerical tests in Section 4, we restrict our attention to tests based on ϕ_1 and ϕ_2 for which $\nu_{\mathcal{F}_i} = \hat{\varepsilon}_{i_r}$. We note that a larger κ is required in Fig. 2 (left) to prevent the first term in ϕ_1 from prematurely taking a zero value; dependence on parameters like κ is discussed further in Section 4. We examined plots similar to those in Fig. 2 for the first terms of ϕ_3 and ϕ_4 but found no such relationship with the noise level. As a result, we have chosen to not include constants of the form $\nu_{\mathcal{F}_i}$ in the definitions of ϕ_3 and ϕ_4 .

3.7 Relationship to loss functions

Ideally, an algorithm should stop when the cost of performing additional function evaluations outweighs additional improvements in the function value. When such a trade-off can be quantified, this problem becomes one of *optimal stopping* [21]. Results in the literature typically focus on cases when the distribution of the stochastic improvement is known. We briefly illustrate a connection to a simple loss function employed in optimal stopping with our tests.

We focus on the case when the cost of an additional evaluation is constant. This can be viewed as treating the computational expense per function evaluation as constant, but the cost and the tests proposed here could be suitably modified as an algorithm enters a subdomain where the cost of an evaluation changes. Given a sequence $\{f_j\}$, the loss function

$$L(i, c) = \min_{1 \leq j \leq i} \{f_j\} + c \cdot i \quad (8)$$

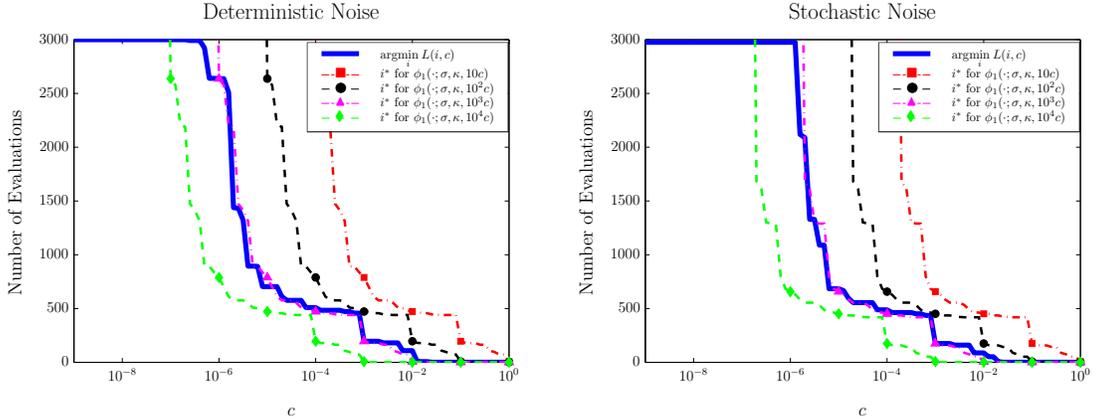


Fig. 3: Number of evaluations i^* for a termination test based on (3) that is parameterized by c . The plots show remarkably similar behavior to the number of evaluations that minimize $L(\cdot, c)$ in (8).

provides a measure of the success of stopping after i evaluations when the cost per evaluation (relative to f^*) is c . This loss function appears in the optimal stopping literature as the house-selling problem [2], where $\{f_j\}$ are assumed to be independent and identically distributed random variables.

Fig. 3 shows the minimizer of $L(\cdot, c)$ for a variety of c values on a sequence $\{f_j\}_{j=1}^{3000}$ output by a direct search solver on a nonlinear function with deterministic (left) and stochastic (right) noise. We compare this minimizer with the number of evaluations i^* defined by (2) for the family ϕ_1 when c is used as a linear multiplier for the parameter μ . Fig. 3 shows a strong correlation between the behavior of $\text{argmin}_i L(i, c)$ and the termination test defined by ϕ_1 using an estimate of the noise and an appropriate choice of the parameters (μ, κ) . This illustrates how varying the parameters in the proposed families can be closely related to the cost of performing an evaluation.

4 Numerical Experiments

We now demonstrate the merits of the proposed tests and explore the effect of changing the associated parameter values by considering outputs generated by a set of derivative-free optimization solvers on a collection of noisy test problems.

We consider the collection of unconstrained least-squares problems used in [17], with each function taking the form

$$f(x) = 1 + (1 + \sigma g(x)) \sum_{i=1}^m F_i^s(x)^2, \quad (9)$$

where each F^s is a smooth, deterministic function and $\sigma \ll 1$ is a positive scalar used to control the amplitude of the noise being added to $f^s(x) = \sum_{i=1}^m F_i^s(x)^2$. We begin our

study by considering stochastic noise, so that $g(x)$ represents independent and identically distributed (iid) random variables with variance $\text{Var}\{g(x)\} = 1$. As a result, the relative noise of these test functions is simply σ and, hence, independent of x . A constant was added in (9) so that the relative noise is consistently defined; such shifts are commonly performed in accuracy measures (see, e.g., [6]).

To examine the tests on a diverse set of local methods, we consider sequences $\{f_j\}$ produced by different derivative-free optimization solvers. Since the relative merits of these solvers is not the focus of this study, we do not explicitly list these solvers, but we note that they come from a variety of classes, including model-based methods, implementations of Nelder-Mead, pattern search methods, and methods that cross these classes.

To more accurately study the effect of our tests, we have made the built-in termination criteria of these solvers as ambitious as possible in an attempt to remove their influence. Hence, we ran each solver until either it crashed (e.g., for a numerical reason, such as the simplex sides being dropped sufficiently below machine precision) or a maximum budget of 5,000 function evaluations was achieved. This budget of evaluations is significantly larger than the one considered in [17], and we consider it to be more than sufficient for the problems in this set, which range in dimension from $n = 2$ to $n = 12$. We denote the maximum number of function evaluations (either 5,000, or fewer if the solver crashed) by i_{\max} .

We then have a set of 318 nonnegative sequences $\{f_j\}_{j=1}^{i_{\max}}$, which constitute our set of problems \mathcal{P} . We use these problems to examine the performance of a set of tests \mathcal{T} , defined as members of the families proposed in Section 3. For a test $t \in \mathcal{T}$ and problem $p \in \mathcal{P}$, we denote $i_{p,t}^*$ to be the number of function values after which test t would stop on problem p . If the test is not satisfied before the maximum number of evaluations i_{\max} of problem p , we let $i_{p,t}^* = i_{\max}$ to mirror what would be done in practice.

4.1 Accuracy profiles for the ϕ_1 family

Termination criteria that are too easily satisfied have limited practicality since they could stop with a function value far from the minimum. We will measure this ability by considering the relative difference between $f_{i_{p,t}^*}^*$ and $f_{i_{\max}}^*$,

$$e_{p,t} = \begin{cases} \infty & \text{if } i_{p,t}^* = i_{\max}, \\ \frac{f_{i_{p,t}^*}^* - f_{i_{\max}}^*}{f_{i_{p,t}^*}^*} & \text{if } i_{p,t}^* < i_{\max}. \end{cases} \quad (10)$$

We note that, with the exception of the case $i_{p,t}^* = i_{\max}$, (10) is the relative error $\text{re}(\alpha, \beta) = \frac{|\alpha - \beta|}{\max(|\alpha|, |\beta|)}$ but exploiting the fact that the sequence $\{f_j\}$ is monotone and strictly positive. It follows that $e_{p,t} \in [0, 1] \cup \{\infty\}$. The exception $i_{p,t}^* = i_{\max}$ is made in order to focus on problems where the test terminated short of the maximum budget i_{\max} .

For this study, we consider the termination by test t to have occurred with acceptable accuracy on problem p if $e_{p,t}$ is within a small multiple of the relative noise for problem p .

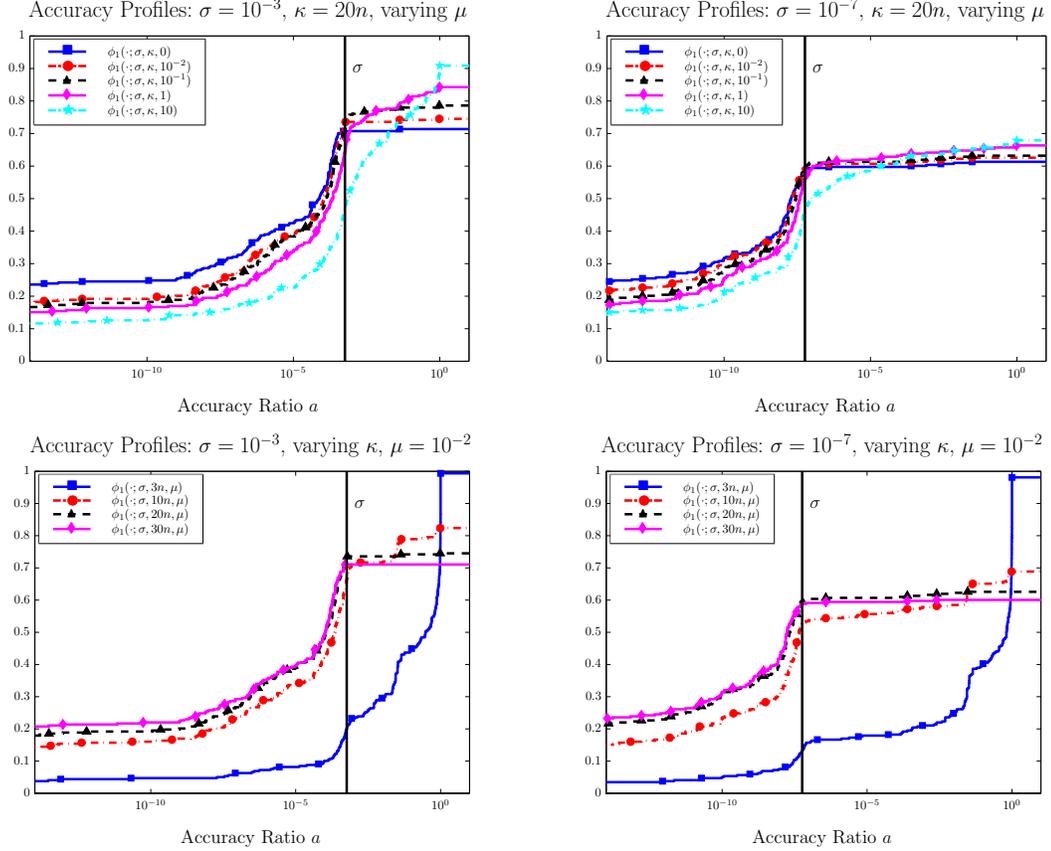


Fig. 4: Accuracy profiles for members of the ϕ_1 family on problems (9) with two different magnitudes of (known) stochastic relative noise σ . In the top plots, κ is held fixed and the shown members have different μ values. In the bottom plots, μ is held fixed and the shown members have different κ values.

For each test t considered, we plot the cumulative distribution function

$$\omega_t(a) = \frac{1}{|\mathcal{P}|} |\{p \in \mathcal{P} : e_{p,t} \leq a\}|, \quad a \in [0, 1],$$

where $|\cdot|$ denotes the cardinality of a set. A test successfully terminates within the level of the noise on the collection of problems \mathcal{P} if its *accuracy profile* has large values of $\omega_t(c\sigma)$ for constants $c \geq 1$ of modest size. On the other hand, as we discuss next, smaller values of $\omega_t(c\sigma)$ are desired for $c \ll 1$ since otherwise it is possible that a test should have stopped sooner.

In Fig. 4 we consider stochastic problems of the form (9) with two different levels of relative noise, $\sigma = 10^{-3}$ (left) and $\sigma = 10^{-7}$ (right). We focus on the family of tests given by ϕ_1 in (3), with the relative noise σ being known exactly. The top plots of Fig. 4 study the effects of varying the parameter μ , while the bottom plots study the effects of varying the parameter κ . The vertical lines in Fig. 4 denote $a = \sigma$ as a point of reference.

The left asymptote of ω_t shows the fraction of problems on which the test stopped after it reached the minimum value $f_{i_{\max}}^*$, while the right asymptote is the fraction of problems for which the test was satisfied with $i_{p,t}^* < i_{\max}$. Values $e_{p,t} \approx 1$ correspond to cases when test t stopped with a function value well above $f_{i_{\max}}^*$; these values indicate that the test is too easily satisfied on problem p .

For the top two plots in Fig. 4, we see the ϕ_1 family for various values of μ with κ fixed to 20 times the dimension n of each problem. We note that even though ϕ_1 is monotone in μ , the accuracy profiles ω_t can cross because the relative error measure $e_{p,t}$ does not preserve the monotonicity. Fig. 4 shows that as the tests get less conservative (as μ grows), a number of problems are terminated well before the relative error is on the order of the noise. On the other hand, not much is gained by setting μ less than 10^{-1} or 10^{-2} .

The bottom two plots in Fig. 4 show ϕ_1 family members for fixed $\mu = 10^{-2}$ and various values of κ , which can be thought of as a backward difference parameter. Little improvement is seen for $\kappa > 20n$, but a marked decrease in accuracy occurs when $\kappa < 10n$. Many problems are stopping with a large relative error, in part because f_j^* can remain unchanged for many consecutive j . For example, the lower left plot shows that for noise affecting the third digit, on all 318 problems f_j^* remained unchanged for $3n$ consecutive evaluations before $j = i_{\max}$ was reached.

4.2 Performance profiles for the ϕ_1 family

While accuracy profiles can quantify when a test stops too soon, they may not reveal which tests require excessive function evaluations to achieve high accuracies. For example, the maximum budget test ϕ_5 trivially achieves ideal accuracy but can make many more evaluations than are required to get sufficient accuracy.

We use performance profiles [7] to compare different stopping rules in terms of both accuracy and the number of function evaluations required. A performance profile requires a convergence test as well as a performance measure $r_{p,t}$ for each problem $p \in \mathcal{P}$ and test $t \in \mathcal{T}$. We use the number of evaluations $i_{p,t}^*$ as our performance measure and a convergence test requiring that the solution obtained is within a factor τ of the final one,

$$f_{i_{p,t}^*}^* - f_{i_{\max}}^* \leq \tau |f_{i_{p,t}^*}^*| \hat{\epsilon}_{i_r}. \quad (11)$$

The convergence test has the effect of setting the performance measure $i_{p,t}^* = \infty$ whenever the original $i_{p,t}^*$ does not satisfy (11). The performance ratio

$$r_{p,t} = \begin{cases} \frac{i_{p,t}^*}{\min\{i_{p,\hat{t}}^* : \hat{t} \in \mathcal{T}, (11) \text{ satisfied for } (p, \hat{t})\}} & \text{if (11) is satisfied for } (p, t) \\ \infty & \text{else} \end{cases}$$

measures the relative performance on problem p of test t when compared with the other tests in \mathcal{T} .

The performance profile

$$\rho_t(\alpha) = \frac{|\{p \in \mathcal{P} : r_{p,t} \leq \alpha\}|}{|\mathcal{P}|}$$

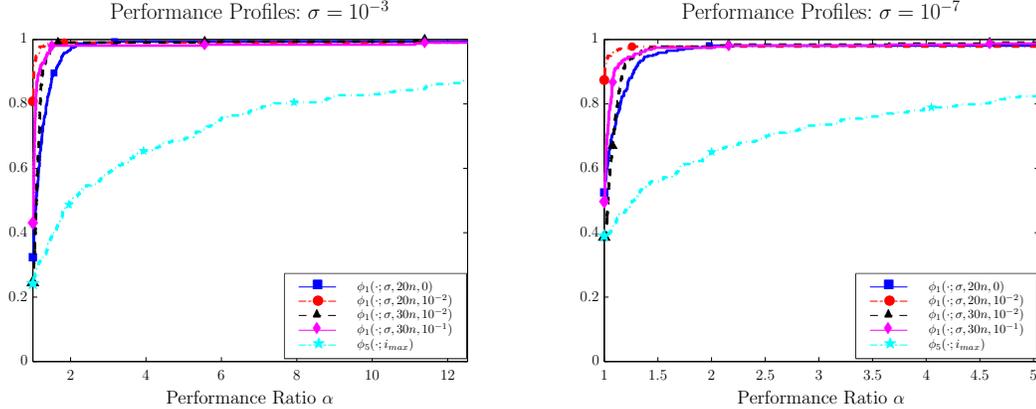


Fig. 5: Performance profiles for the most accurate ϕ_1 tests on problems (9) with two different magnitudes of (known) stochastic relative noise σ . Note that the α -axis has been truncated for each plot; ϕ_5 eventually terminates all of the problems and thus has a profile that will reach the value 1; all other tests change by less than .01.

then represents the fraction of problems where test t satisfied the accuracy requirement (11) with a number of evaluations within a factor α of the best-performing, sufficiently accurate test. Larger values of $\rho_t(\alpha)$ are hence better, with $\rho_t(1)$ being the fraction of problems where t has successfully terminated first among all tests in \mathcal{T} and $\lim_{\alpha \rightarrow \infty} \rho_t(\alpha)$ being the fraction of problems for which t satisfied (11).

Fig. 5 shows the performance profiles for the most accurate ϕ_1 family members for a convergence level $\tau = 1$ in (11) and two levels of noise σ . We include $\phi_5(\cdot; i_{\max})$ in \mathcal{T} as a point of reference to indicate an upper bound on the fraction of problems that all other tests may not have terminated with $i_{p,t}^* < i_{\max}$; this strategy also ensures that at least one test in \mathcal{T} will satisfy (11) for any $\tau \geq 0$.

These performance profiles illustrate that some members of the ϕ_1 family of tests require a fraction of the full i_{\max} evaluations. This is the case especially for larger magnitudes of noise, where less accurate solutions are demanded, and this advantage can be extended if τ is increased in (11). As τ decreases, more conservative tests become more appealing because the convergence test (11) is more difficult to satisfy. Likewise, as the noise decreases, (11) demands more accurate solutions, and it becomes necessary to perform i_{\max} evaluations on a larger share of the problems. Although we have examined performance profiles for various τ , we fix $\tau = 1$ for the remainder of the paper.

These performance profiles also demonstrate how more liberal stopping rules can be more successful than the accuracy profiles reveal. For example, in Fig. 4, $\phi_1(\cdot; \cdot, 20n, 0)$ and $\phi_1(\cdot; \cdot, 20n, 10^{-2})$ looked nearly identical in terms of their accuracy, but in Fig. 5 we see a marked difference in the performance measures. The right asymptotes of their performance profiles are nearly identical, a reflection of their accuracy profiles at $a = \sigma$, but the rest of the profiles show that $\phi_1(\cdot; \cdot, 20n, 10^{-2})$ uses considerably fewer function evaluations to satisfy

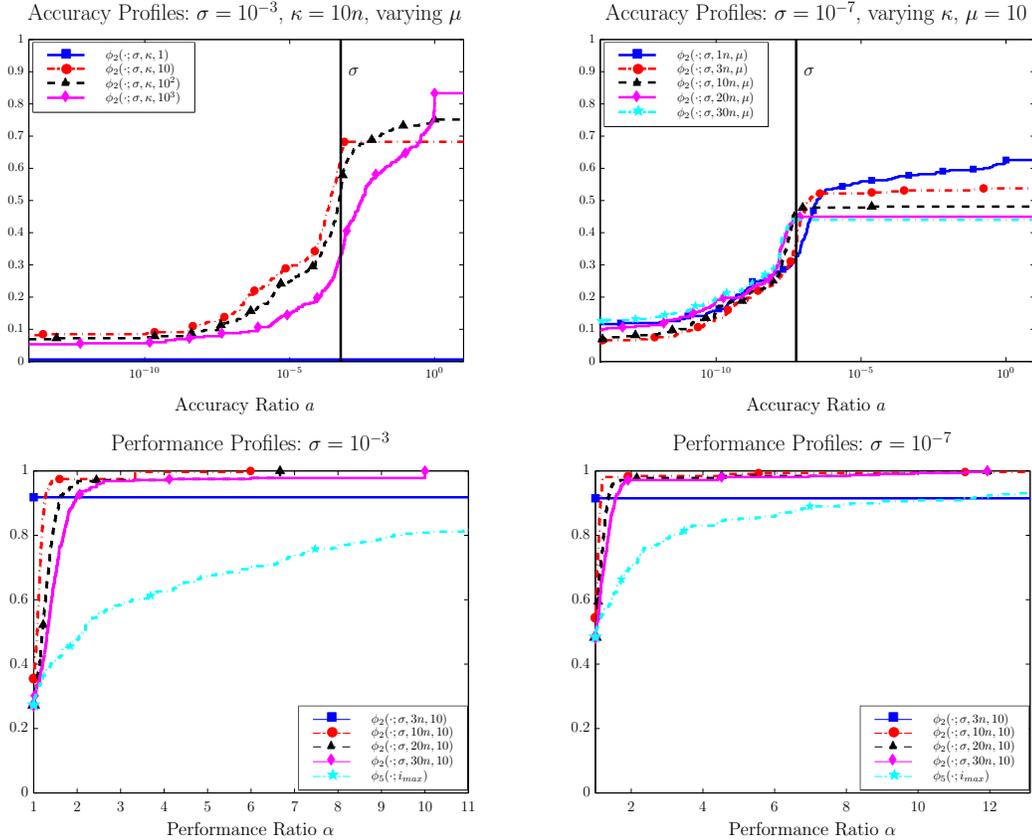


Fig. 6: Accuracy (top) and performance (bottom) profiles for the ϕ_2 family on problems (9) with two different magnitudes of stochastic relative noise σ as κ and μ are varied.

this accuracy requirement. Because of this high accuracy and performance, we consider $\phi_1(\cdot; \cdot, 20n, 10^{-2})$ to be the best stopping rule in its family.

4.3 Accuracy and performance plots for the ϕ_2 family

Having outlined our procedure for determining what constitutes good members of the ϕ_1 family, we can now quickly do so for the family based on ϕ_2 in (4).

The accuracy profiles in the upper left plot of Fig. 6 show that the $\phi_2(\cdot; \cdot, 10n, 1)$ test was satisfied on less than 5% of the problems, and so $\mu \leq 1$ has little relevance for this family. In our experience, decreasing κ did not alleviate this problem for small μ . In general, ϕ_2 tends to be much more sensitive to the value κ than are tests based on ϕ_1 . We also see that ϕ_2 is more accurate at smaller values of κ than ϕ_1 was; $\kappa = 3n$ is now a more competitive parameter choice. This trade-off in accuracy comes at the cost of the ϕ_2 tests being more conservative and, hence, satisfied on fewer of the problems.

We again use performance profiles to measure whether tests are overly conservative. As indicated by the larger range for α in the bottom two plots of Fig. 6, the ϕ_2 family of tests are

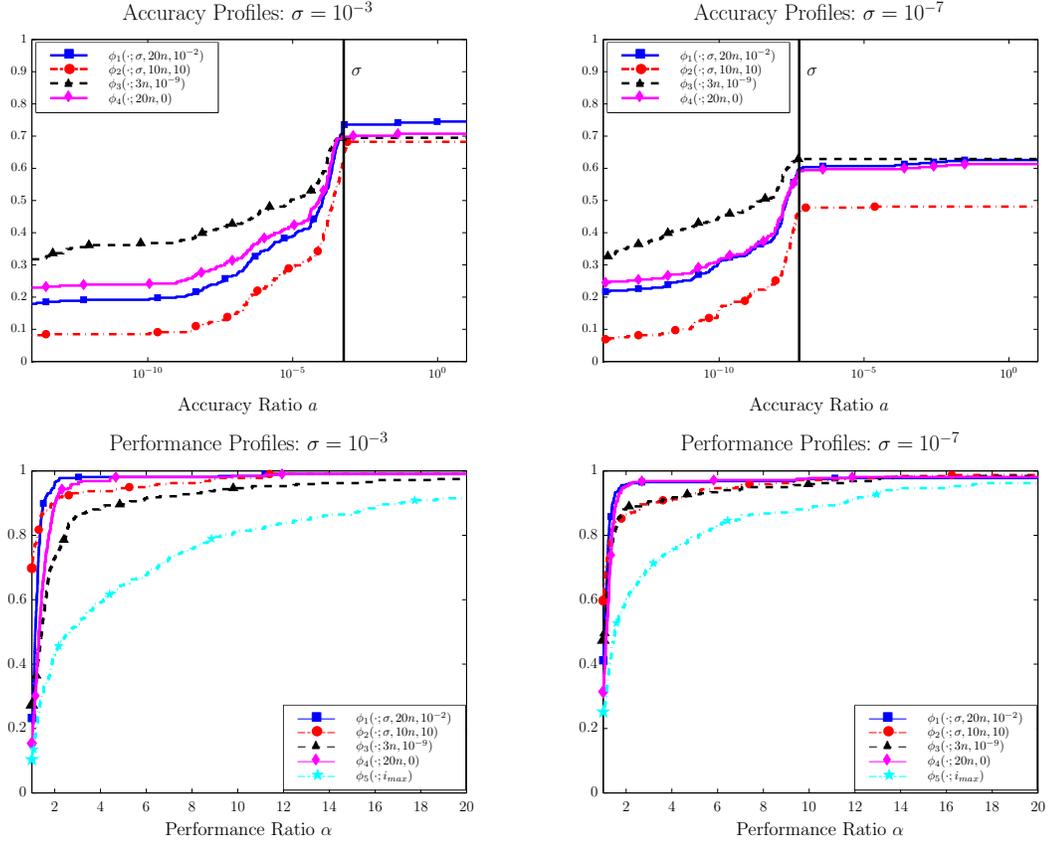


Fig. 7: Accuracy (top) and performance (bottom) profiles for the best tests on problems (9) with two different magnitudes of stochastic relative noise σ . The horizontal axes on the performance profiles are truncated for clarity; ϕ_5 eventually achieves a value of 1; all other tests change by less than .03.

more difficult to satisfy overall, and the number of function evaluations required compares slightly less favorably with i_{\max} than for the ϕ_1 family. We also see that $\phi_2(\cdot, \cdot, 3n, 10)$ tends to be the most liberal test, in part because it requires fewer consecutive evaluations than the other members shown as indicated by the value of κ , but that $\phi_2(\cdot, \cdot, 10n, 10)$ requires just a small increase in the function values while solving a greater fraction of problems overall. Based on our computational experience, we consider $\phi_2(\cdot, \cdot, 10n, 10)$ to be the best test in this family for these problems.

4.4 Across-family comparisons

We performed similar comparisons for the members of the ϕ_3 and ϕ_4 families, but the analysis is identical to what has been presented above. For the benchmark problems \mathcal{P} , we found $\phi_3(\cdot, \cdot, 3n, 10^{-9})$ and $\phi_4(\cdot, \cdot, 20n, 0)$ to be the best among their respective families.

Having identified the best members of each family of tests, we compare them head-to-

head in Fig. 7. The top two plots of Fig. 7 demonstrate that when the four tests considered stop with fewer than i_{\max} evaluations, they all tend to have obtained a solution within the level of the noise, ϕ_3 being the slight loser of the group because of the .05 jump past $a = \sigma$. The $\phi_1(\cdot; \cdot, 20n, \cdot)$ test tends to be the most successful in this metric, because it terminates on a larger fraction of problems while still being accurate.

On the other hand, the lower two plots of Fig. 7 show that the test based on ϕ_2 generally requires fewer evaluations to be satisfied. The tests based on ϕ_3 and ϕ_4 both tend to require more evaluations and satisfy the convergence test on a smaller fraction of the problems.

4.5 Deterministic noise

We now consider how these tests perform in the presence of deterministic noise by using functions of the form (9), with a deterministic g . To model deterministic noise, we use the same g combining high-frequency and lower-frequency nonsmooth oscillations as used in [17], with $g : \mathbb{R}^n \rightarrow [-1, 1]$ defined by the cubic Chebyshev polynomial $g(x) = \xi(x)(4\xi(x)^2 - 3)$, where

$$\xi(x) = 0.9 \sin(100\|x\|_1) \cos(100\|x\|_\infty) + 0.1 \cos(\|x\|_2).$$

Using the technique in [18], we consistently estimated the relative noise in the 318 resulting problems to be of the order 0.6σ , provided that the sampling distance is appropriately chosen.

The accuracy profiles in Fig. 8 show a mild decrease in accuracy for the best tests compared with stochastic noise in Fig. 7. As a result, we see that on just over 10% (20%) of the problems, the test based on ϕ_1 (ϕ_2) now terminates while not satisfying the convergence test (11) with $\tau = 1$ when $\sigma = 10^{-3}$. An improvement upon these tests is discussed in the next section.

5 Conclusions

In this paper we have considered parameterized families of termination tests that require solely a history of evaluations (points and function values) from an optimization solver. Our analysis and experiments show how values for these parameters can be changed to reflect a user’s view of the expense of an additional function evaluation and the accuracy demanded, the two characteristics that form the basis for (1).

Our study of stochastic noise confirmed that tests based on x values do not perform as well as tests based on function values when the accuracy of the final function value is the primary metric.

We found that our tests based on function values (ϕ_1 and ϕ_2) can be sensitive to how far back in the history these values are examined, as described by the parameter κ . This result is important since previous work generally focused on successive decreases [17] or values on a simplex or stencil [9]. The choice of κ must balance the competing demands of terminating prematurely and identifying potentially irreparable stagnation.

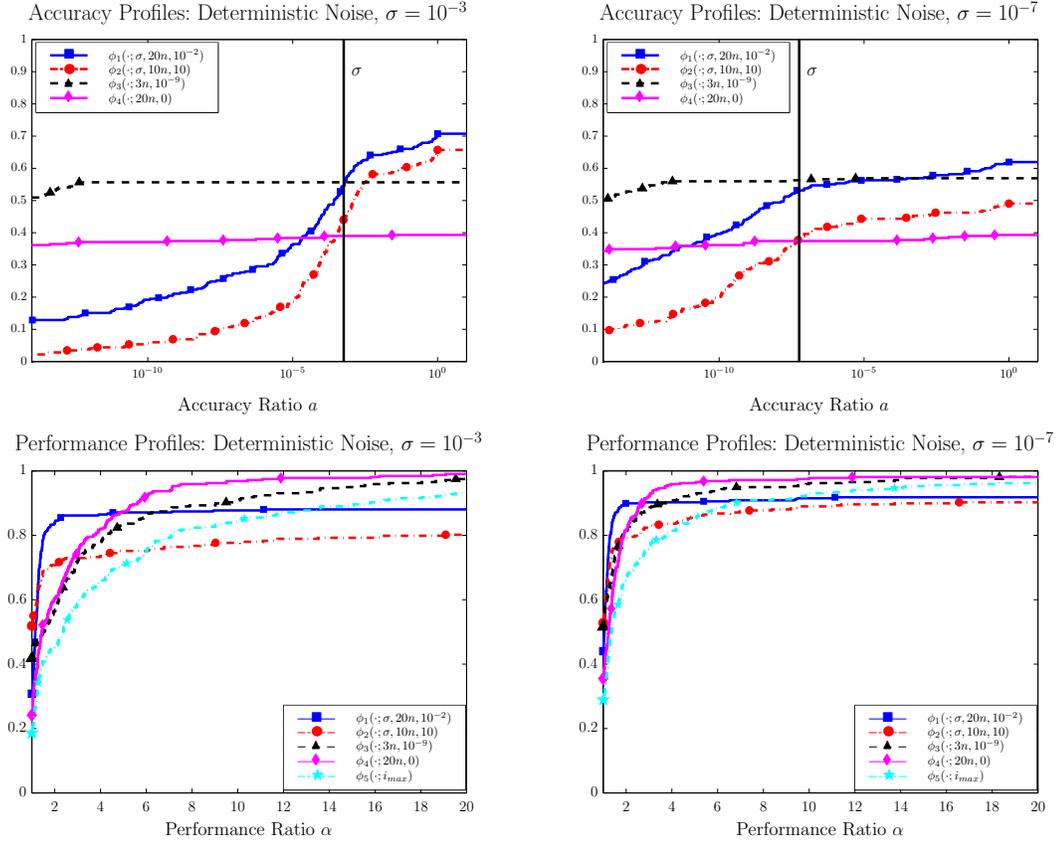


Fig. 8: Accuracy (top) and performance (bottom) profiles for the best tests on problems (9) with two different magnitudes of deterministic noise. The horizontal axes on the performance profiles are truncated for clarity; ϕ_5 eventually achieves a value of 1; all other tests change by less than .03.

In our study of stochastic noise we encountered a nontrivial number of problems where a solver found no change in f^* for 300 or 400 evaluations but then found change in the first digit of the solution. Hence we recommend a baseline value of $\kappa = 20n$, with a corresponding μ less than or equal to 0.1 for ϕ_1 . Good performance for ϕ_2 can still be seen for κ less than $10n$, but this test is more sensitive to μ values. For $\mu \ll 10$, ϕ_2 tests are rarely satisfied, whereas for $\mu \gg 10$, termination can occur with an inaccurate solution.

We have also seen that fewer problems are terminated before the budget constraint as noise (measured by σ) becomes small. In these cases, however, a solver's built-in termination criteria should be satisfied more easily.

In general, we found that the tests based on ϕ_1 and ϕ_2 were also able to stop considerably short of the maximum budget on the deterministic problems examined but that this result was obtained at the cost of lower accuracy. We therefore recommend that these tests be made more conservative for deterministic noise and again allow the solver's built-in tests to stop a run when necessary. The effect is shown in Fig. 9, where we see that tests based on

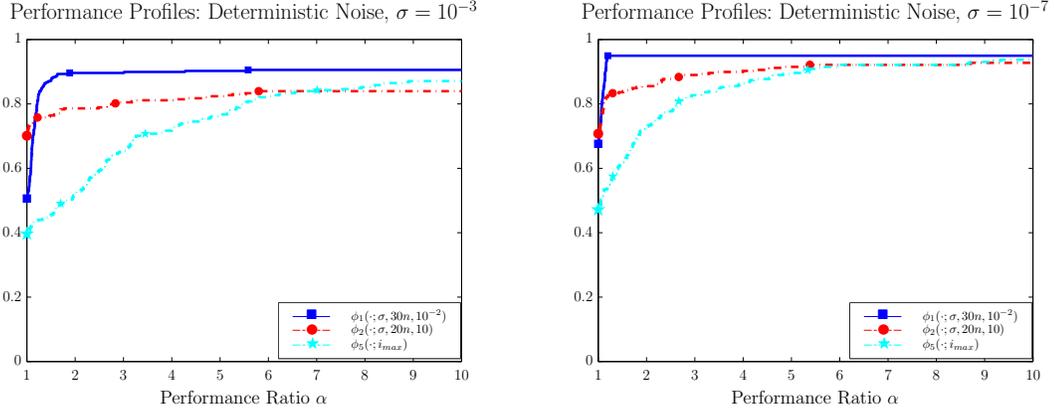


Fig. 9: Performance profiles for more conservative tests on problems (9) with two different magnitudes of deterministic noise. The horizontal axes on the performance profiles are truncated for clarity; ϕ_5 eventually achieves a value of 1; all other tests change by less than .03.

ϕ_1 and ϕ_2 perform better when κ is increased by $10n$. In practice, one would also need an estimate of the relative noise $\hat{\varepsilon}_{i_r}$, but our results of ϕ_1 and ϕ_2 varying the linear multiplier of the noise, μ , show that the test remain relatively stable if $\hat{\varepsilon}_{i_r}$ is estimated within an order of magnitude.

Our last comments underscore that we do not consider the termination of noisy optimization problems to be a solved problem. To the contrary, there is no free lunch: any test that seeks to save expensive evaluations will always sacrifice something in terms of accuracy, and any test pursuing accuracy in general will pay through function values. Accepting this trade-off, and choosing a test with the appropriate balance specific to a problem in hand, is often the best one can hope for.

References

- [1] Anderson, E.J., Ferris, M.C.: A direct search algorithm for optimization with noisy function evaluations. *SIAM J. Optimization* **11**, 837–857 (2000). DOI 10.1137/S1052623496312848
- [2] Chow, Y.S., Robbins, H.: On optimal stopping rules. *Probability Theory and Related Fields* **2**, 33–49 (1963). DOI 10.1007/BF00535296
- [3] Conn, A.R., Scheinberg, K., Vicente, L.N.: *Introduction to Derivative-Free Optimization*. MPS/SIAM Series on Optimization. SIAM, Philadelphia (2009)
- [4] Deng, G., Ferris, M.C.: Adaptation of the UOBYQA algorithm for noisy functions. In: *Proceedings of the Winter Simulation Conference*, pp. 312–319 (2006). DOI 10.1109/WSC.2006.323088
- [5] Deng, G., Ferris, M.C.: Extension of the direct optimization algorithm for noisy functions. In: *Proceedings of the Winter Simulation Conference*, pp. 497–504 (2007). DOI 10.1109/WSC.2007.4419640
- [6] Dennis, J.E., Schnabel, R.B.: *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. SIAM, Philadelphia (1996)
- [7] Dolan, E.D., Moré, J.J.: Benchmarking optimization software with performance profiles. *Math. Programming* **91**, 201–213 (2002). DOI 10.1007/s101070100263
- [8] Fletcher, R.: *Practical Methods of Optimization*, 2nd edn. John Wiley & Sons, New York (1987)
- [9] Gill, P.E., Murray, W., Wright, M.H.: *Practical Optimization*. Academic Press, London (1981)
- [10] Gramacy, R.B., Taddy, M.A.: Categorical inputs, sensitivity analysis, optimization and importance tempering with `tgp` version 2, an R package for treed Gaussian process models. *J. Statistical Software* **33**(6), 1–48 (2010). DOI 10.1198/jcgs.2010.192jss
- [11] Gratton, S., Mouffe, M., Toint, P.: Stopping rules and backward error analysis for bound-constrained optimization. *Numerische Mathematik* pp. 1–25 (2011). DOI 10.1007/s00211-011-0376-1
- [12] Huyer, W., Neumaier, A.: SNOBFIT – Stable noisy optimization by branch and fit. *ACM Trans. Math. Softw.* **35**, 9:1–9:25 (2008). DOI 10.1145/1377612.1377613
- [13] Jones, D.R., Perttunen, C.D., Stuckman, B.E.: Lipschitzian optimization without the Lipschitz constant. *J. Optimization Theory and Applications* **79**, 157–181 (1993). DOI 10.1007/BF00941892

- [14] Kelley, C.T.: Users Guide for imfil version 1. Available at www4.ncsu.edu/~ctk/imfil.html
- [15] Kelley, C.T.: Detection and remediation of stagnation in the Nelder-Mead algorithm using a sufficient decrease condition. *SIAM J. Optimization* **10**(1), 43–55 (1999). DOI 10.1137/S1052623497315203
- [16] Kortelainen, M., Lesinski, T., Moré, J., Nazarewicz, W., Sarich, J., Schunck, N., Stoitsov, M.V., Wild, S.: Nuclear energy density optimization. *Phys. Rev. C* **82**(2), 024,313 (2010). DOI 10.1103/PhysRevC.82.024313
- [17] Moré, J.J., Wild, S.M.: Benchmarking derivative-free optimization algorithms. *SIAM J. Optimization* **20**(1), 172–191 (2009). DOI 10.1137/080724083
- [18] Moré, J.J., Wild, S.M.: Estimating computational noise. *SIAM J. Scientific Computing* **33**(3), 1292–1314 (2011). DOI 10.1137/100786125
- [19] Powell, M.J.D.: An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal* **7**(2), 155–162 (1964). DOI 10.1093/comjnl/7.2.155. URL <http://comjnl.oxfordjournals.org/content/7/2/155.abstract>
- [20] Powell, M.J.D.: UOBYQA: Unconstrained optimization by quadratic approximation. *Math. Programming* **92**, 555–582 (2002). DOI 10.1007/s101070100290
- [21] Shiriyayev, A.: *Optimal Stopping Rules*. Springer-Verlag, New York (1978)
- [22] Tomick, J.J., Arnold, S.F., Barton, R.R.: Sample size selection for improved Nelder-Mead performance. In: *Proceedings of the Winter Simulation Conference*, pp. 341–345. IEEE Computer Society (1995). DOI 10.1145/224401.224630
- [23] Wright, M.H.: Using randomness to avoid perseveration in direct search methods. Presentation at “The International Symposium on Mathematical Programming” (2009)

<p>The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory (“Argonne”) under Contract DE-AC02-06CH11357 with the U.S. Department of Energy. The U.S. Government retains for itself, and others acting on its behalf, a paid-up, nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.</p>
--