Multi-objective optimisation in the presence of uncertainty

Jonathan E. Fieldsend
Department of Computer Science
University of Exeter, Exeter, UK EX4 4QF
J.E.Fieldsend@exeter.ac.uk

Richard M. Everson
Department of Computer Science
University of Exeter, Exeter, UK EX4 4QF
R.M.Everson@exeter.ac.uk

Abstract- There has been only limited discussion on the effect of uncertainty and noise in multi-objective optimisation problems and how to deal with it. Here we address this problem by assessing the probability of dominance and maintaining an archive of solutions which are mutually non-dominating with some known probability.

We examine methods for estimating the probability of dominance. These depend crucially on estimating the effective noise variance and we introduce a novel method of learning the variance during optimisation.

Probabilistic domination contours are presented as a method for conveying the confidence that may be placed in objectives that are optimised in the presence of uncertainty.

1 Introduction

Evolutionary computation (EC) techniques are now extensively used when attempting to discover the optimal or near optimal parameterisation for problems with unknown or complex function transformation from parameters to objective(s) (see for instance the Adaptive Computing in Design and Manufacture series (Parmee, 2004)). Almost all optimisation procedures search the parameter space by evaluating the objectives for a given parameterisation before proposing a new, hopefully better, parameterisation. It is generally assumed that repeated evaluation of the objectives for a single parameterisation yields the same objective values. However, a special, but not insubstantial, class of these problems exists in which there is additional uncertainty in the veracity of the results obtained from the system model. One clear example arises when measurement error or stochastic elements in physical system leads to different results for repeated evaluations at the same parameter values (Büche *et al.* , 2002; Stagge, 1998). Our own interest in this topic arises from the optimisation of classification error rates in pattern recognition tasks: precise error rates depend upon the particular data set used; different, but statistically equivalent data sets, arising from bootstrap samples yield different error rates (Fieldsend & Everson, 2004; Everson & Fieldsend, 2005) and it is important to evaluate the uncertainty associated with optimal classifiers.

Work is ongoing in the scalar optimisation EC community to tackle these types of problem, and indicates some elitist techniques can prove fragile (Rana *et al.* , 1996; Di Pietro *et al.* , 2004), as there is no longer the guarantee that fitness of the elite solution improves with generation. As most modern multi-objective evolutionary algorithms (MOEAs) rely heavily on elitism, this should be of concern to practitioners dealing with uncertain multi-objective problems. In this paper we model the uncertainty in the objectives as observational noise. We describe, in Section 2, a sequence of methods for use in multi-objective optimisation depending on the amount of information that is known about the noise characteristics. These methods are illustrated on standard test problems modified by the addition of noise.

Teich (2001) and Hughes (2001) have each addressed multiobjective optimisation with uncertainty and here we extend their work by relaxing the assumptions made about the noise characteristics. Hughes (2001) makes the important distinction between two sorts of uncertainty. First, the function being optimised may be in error, that it is it may not be a faithful model of the system which should be optimised, but re-evaluations of the function with the same parameters yield identical results. Secondly, the function being evaluated may be noisy; repeated evaluations giving different results. It is this latter case that we address here, although the methods may be applied to the former situation if information about inaccuracies in the model is available.

Before discussing the effect of uncertainty in the evaluation of objectives, we briefly review the ideas of dominance and Pareto optimality which are central to multi-objective optimisation.

A general multi-objective optimisation problem seeks to simultaneously extremise D objectives: $f_d(\mathbf{x})$, $d = 1, \dots, D$ where each objective depends upon a vector $\mathbf{x} = (x_1, \dots, x_P)$ of P parameters or decision variables. In this paper we are concerned with the problem in which the objectives themselves are unobservable, but instead we have access to y_d the objectives contaminated by *observational* noise ϵ:

$$y_d = f_d(\mathbf{x}) + \epsilon_d \tag{1}$$

The parameters may also be subject to the J constraints:

$$e_j(\mathbf{x}) \geq 0, \qquad j = 1, \dots J \tag{2}$$

which, for simplicity, we assume can be evaluated precisely. The multi-objective optimisation problem may thus be expressed as:

$$\text{minimise} \quad \mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_D(\mathbf{x})) \tag{3}$$

$$\text{subject to} \quad \mathbf{e}(\mathbf{x}) = (e_1(\mathbf{x}), \dots, e_J(\mathbf{x})) \geq 0. \tag{4}$$

When faced with only a single objective an optimal solution is one which minimises the objective given the model constraints. However, when there is more than one objective to be minimised solutions may exist for which performance on one objective cannot be improved without reducing performance on at least one other. Such solutions are said to be *Pareto optimal*. The set of all Pareto optimal solutions is said to form the Pareto front.

The notion of *dominance* may be used to make Pareto optimality clearer. Assuming that the goal is to minimise the objectives and there is no noise, a decision vector $\mathbf{x}$ is said to *strictly dominate* another $\mathbf{z}$ (denoted $\mathbf{x} \prec \mathbf{z}$) iff

$$\begin{array}{c} f_d(\mathbf{x}) \le f_d(\mathbf{z}) \quad \forall d = 1, \ldots, D \quad \text{and} \\ f_d(\mathbf{x}) < f_d(\mathbf{z}) \quad \text{for some } d. \end{array} \tag{5}$$

Less stringently, $\mathbf{x}$ *weakly dominates* $\mathbf{z}$ (denoted $\mathbf{x} \preceq \mathbf{z}$) iff

$$f_d(\mathbf{x}) \le f_d(\mathbf{z}) \quad \forall d = 1, \ldots, D. \tag{6}$$

A set of decision vectors F is said to be a *non-dominated set* if no member of the set is dominated by any other member:

$$\mathbf{x} \not\prec \mathbf{z} \quad \forall \mathbf{x}, \mathbf{z} \in F. \tag{7}$$

A solution to the minimisation problem (3) is thus *Pareto optimal* if it is not dominated by any other feasible solution, and the non-dominated set of all Pareto optimal solutions is the Pareto front $\mathcal{P}$.

Elitist MOEAs generally maintain a non-dominated set or archive F of solutions which form the estimated Pareto front. As the optimisation proceeds new solutions are generated (either by copying and perturbing solutions in F (e.g., (Knowles & Corne, 2000)) or by mutating and recombining solutions in a search population (e.g., (Zitzler & Thiele, 1999; Deb *et al.* , 2000)). If the new solution, say $\mathbf{x}'$ is not dominated by a member of F then $\mathbf{x}'$ is added to F, and any solutions in F that are dominated by $\mathbf{x}'$ are deleted from F. In this way F is always a non-dominated set that cannot move away from the true Pareto front $\mathcal{P}$.

Taking as a starting point an elitist MOEA, the effect of noise within the evaluation process is twofold. Firstly, the optimising procedure is affected as solutions which should be added to F may be rejected as the noise may represent them as dominated by an element of F, and conversely, solutions that should not have been entered into the archive may be entered, as noise makes them seem better than they are. This therefore has the effect of reducing the algorithm efficiency. Secondly, the final archive may over estimate the true Pareto front due to "lucky" noise realisations making an objective better than is really feasible. In addition, the archive may also contain a large number of solutions which would actually be dominated by other archive members (if evaluated without noise).

2 Probabilistic dominance

The crucial operations in elitist multi-objective optimisation are whether to add a proposed solution $\mathbf{x}'$ to the archive and which solutions to delete from the archive because they are dominated by $\mathbf{x}'$. If evaluation of the objectives is uncertain we can only speak about the probability of dominance rather than outright dominance. We use the notation $\mathbf{x} \prec^{\alpha} \mathbf{x}'$ to mean that $p(\mathbf{x} \prec \mathbf{x}') \ge \alpha$; clearly when $\alpha = 1$ probabilistic dominance is equivalent to the usual deterministic dominance.

Probabilistic dominance allows us to use the usual deterministic elitist algorithms, but we maintain an archive in which we have a degree of *confidence*. Thus we only include a proposal $\mathbf{x}'$ in the archive if the *total* probability that it is dominated by another point in the archive is less than $1 - \alpha$.

Therefore a proposal $\mathbf{x}'$ is added to the archive F if

$$\sum_{\mathbf{x} \in F} p(\mathbf{x} \prec \mathbf{x}') < 1 - \alpha. \tag{8}$$

Likewise, once a new point has been accepted into the archive we should delete from F any solutions $\mathbf{x}$ for which:

$$\sum_{\mathbf{z} \in F} p(\mathbf{z} \prec \mathbf{x}) \ge \alpha. \tag{9}$$

These criteria could be computationally expensive because of the need to compute the probability of dominance between each potential new entrant to the archive and all the members of the archive. However, the time-cost can be reduced at the expense of space by keeping a table of $p(\mathbf{x} \prec \mathbf{z})$ for all $\mathbf{x}$ and $\mathbf{z}$ in F. Every time there is a new entrant to F a new row and column of the table need to be updated (but the entries have to be computed anyway to determine whether to accept the entrant and possibly delete now-dominated members of F); and a row must be deleted when a member is deleted.

It is possible that we might want to remove the summation from (9), so that we reject a point if the probability that it is dominated by any other single archive member is greater than α. This would remove problems with (9) if the size of the archive becomes large, but would leave open the possibility of a point being *almost* dominated by lots of archive members at the α level.

The imperative issue in using the probabilistic dominance framework is obviously how to calculate $p(\mathbf{x} \prec \mathbf{x}')$. This is dependant on what *a priori* knowledge is available about the system noise (or what assumptions it is reasonable to make). Various methods for its calculation in different situations are now discussed.

2.1 Unknown noise

The most severe case is when the noise properties are completely unknown. It may be asymmetric and there may be dependencies between noise in different objectives. In this case the degree of dominance between two solutions can be estimated by repeated evaluation at the fixed parameter values $\mathbf{x}$ and $\mathbf{x}'$. Suppose that $\{\mathbf{y}_i\}_{i=1}^{n}$, are the objectives evaluated n times at $\mathbf{x}$, and $\{\mathbf{y}'_i\}_{i=1}^{n'}$ are n' evaluations at $\mathbf{x}'$, then the probability that $\mathbf{x}$ dominates $\mathbf{x}'$ is estimated by the fraction of times that $\mathbf{y}_i$ dominates $\mathbf{y}'_j$:

$$p(\mathbf{x} \prec \mathbf{x}') = \frac{1}{nn'} \sum_{i=1}^{n} \sum_{j=1}^{n'} I(\mathbf{y}_i \prec \mathbf{y}'_j) \tag{10}$$

where $I(\cdot)$ is the indicator function.

Estimating probabilistic dominance by this sampling method clearly requires several evaluations of the objectives at both $\mathbf{x}$ and $\mathbf{x}'$, which may be prohibitively expensive. This cost can be substantially reduced if it is known or it can be assumed that the noise corrupting each objective is independent. In this case the evaluations for each objective dimension can be permuted to form additional samples.

2.2 Independent noise for each objective

If the noise on contaminating the objectives can be assumed to be independent, the probability of dominance decomposes into a product of probabilities for each dimension:

$$p(\mathbf{x} \prec \mathbf{x}') = \prod_{i=1}^{D} p(f_d(\mathbf{x}) < f_d(\mathbf{x}')) \tag{11}$$

Each of the constituent probabilities $p(f_d(\mathbf{x}) < f_d(\mathbf{x}'))$ is:

$$\int_{-\infty}^{\infty} p(f_d(\mathbf{x}) \,|\, Y_d) \int_{f_d(\mathbf{x})}^{\infty} p(f(\mathbf{x}') \,|\, Y'_d)\, df_d(\mathbf{x}')\, df_d(\mathbf{x}) \tag{12}$$

where Y_d and Y'_d represent the evaluations of $\mathbf{f}(\mathbf{x})$ and $\mathbf{f}(\mathbf{x}')$ respectively. These integrals can be computed if additional information about the noise distributions is known. Hughes (2001) has addressed the case of Gaussian noise with known noise variance, which we review and extend below. An important simplification occurs when the noise is known to be bounded: Teich (2001) has modelled the noise as uniform and has proposed modified archive acceptance schemes for various MOEAs, similar to those discussed above (8) and (9). However the boundedness of the noise means that if the solutions are sufficiently well separated in objective space dominance or lack of dominance may be decided with no uncertainty.

2.3 Gaussian noise with known variance

An attractive and often reasonable assumption is that the noise is Normally distributed about the true objective value; thus, dropping the indices indicating the particular objective,

$$p(y \,|\, f(\mathbf{x})) = \mathcal{N}(y \,|\, f(\mathbf{x}), \sigma^2_{\mathbf{x}}) \tag{13}$$

where

$$\mathcal{N}(y \,|\, \mu, \sigma) = (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \tag{14}$$

is the standard normal density with mean μ and standard deviation σ^2. Note that (13) allows for the possibility that the noise variance $\sigma_{\mathbf{x}}$ may vary with location. Using (13) in (12) gives the probability of dominance in terms of the error function (Hughes, 2001):

$$p(f(\mathbf{x}) < f(\mathbf{x}')) = \frac{1}{2}\left[1 + \mathrm{erf}\left(m/\sqrt{2}\right)\right] \tag{15}$$

where

$$m = \frac{y(\mathbf{x}') - y(\mathbf{x})}{\sqrt{\sigma^2_{\mathbf{x}} + \sigma^2_{\mathbf{x}'}}}. \tag{16}$$

Clearly if $\mathbf{x}' = \mathbf{x}$ then $m = 0$ and $p(f(\mathbf{x}) < f(\mathbf{x}') = 1/2$ as expected by symmetry. Consequently $p(\mathbf{f}(\mathbf{x}) < \mathbf{f}(\mathbf{x}')) = 2^D$ and the probability that $\mathbf{x}'$ and $\mathbf{x}$ are mutually non-dominating is $1 - \frac{1}{2^{D-1}}$.

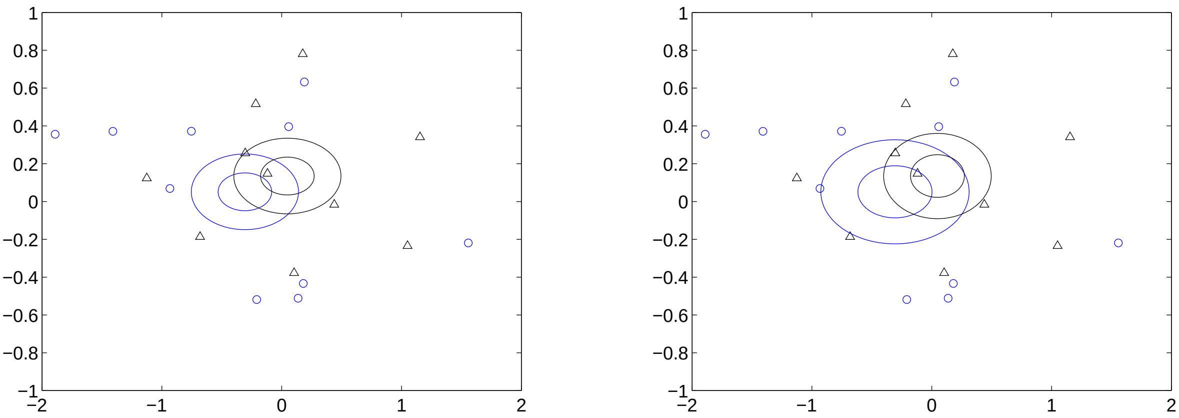
The uncertainty in whether one solution dominates another may be reduced by re-evaluating the solutions at $\mathbf{x}$ and $\mathbf{x}'$. As above, $\{\mathbf{y}_i\}_{i=1}^{n}$ are the objectives evaluated n times at $\mathbf{x}$, and $\{\mathbf{y}'_j\}_{j=1}^{n'}$ are n' evaluations at $\mathbf{x}'$. Then the uncertainty in the location of $f(\mathbf{x})$ is reduced proportional to the square root of the number of repeated measurements:

$$p(f(\mathbf{x}) \,|\, y_1, \ldots, y_n, \sigma^2_{\mathbf{x}}) = \mathcal{N}(f(\mathbf{x}) \,|\, \bar{y}, \sigma^2_{\mathbf{x}}/n) \tag{17}$$

where $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$ is the mean estimate for $f(\mathbf{x})$, with an analogous expression for $\mathbf{x}'$. The probability of dominance is therefore calculated by (16) but with

$$m = \frac{\bar{y}' - \bar{y}}{\sqrt{\frac{\sigma^2_{\mathbf{x}}}{n} + \frac{\sigma^2_{\mathbf{x}'}}{n'}}}. \tag{18}$$

These expressions show that the probability of one solution dominating another can be calculated with increasing accuracy as the number of times each solution is re-evaluated is increased. Also, as might be expected the best estimate for the $f(\mathbf{x})$ is just the sample mean $\bar{y}$. However, this analysis is based upon the variance of the contaminating noise being known. Figure 1 illustrates the importance of having an accurate knowledge of the variance. The figure shows $n = n' = 10$ noisy evaluations of two objectives at $\mathbf{f}(\mathbf{x}) = (0.0, 0.0)$ and $\mathbf{f}(\mathbf{x}') = (0.1, 0.1)$. The noise is Gaussian with covariance matrices $\mathbf{\Sigma}_{\mathbf{x}} = \mathbf{\Sigma}_{\mathbf{x}'} = \left[\begin{smallmatrix} 0.5 & 0 \\ 0 & 0.1 \end{smallmatrix}\right]$. Contours in Figure 1a show one and two times the standard error, using the known variances, centred on the sample mean estimates of $\mathbf{f}(\mathbf{x})$ and $\mathbf{f}(\mathbf{x}')$. The contours depicted in the right panel show the estimates using the variances estimated from the samples themselves. With these true standard errors the $p(\mathbf{f}(\mathbf{x}) \prec \mathbf{f}(\mathbf{x}')) = 0.63$ and $p(\mathbf{f}(\mathbf{x}') \prec \mathbf{f}(\mathbf{x}')) = 0.04$ The variances estimated from the data samples themselves are used in Figure 1b. Clearly the estimates for $\mathbf{f}(\mathbf{x})$ are poor and give a misleading estimates of the probability of dominance; in this case $p(\mathbf{f}(\mathbf{x}) \prec \mathbf{f}(\mathbf{x}')) = 0.56$ and $p(\mathbf{f}(\mathbf{x}') \prec \mathbf{f}(\mathbf{x}')) = 0.06$. This example highlights the importance of accurately estimating the variance and we now describe a straightforward Bayesian scheme which can be used for estimating it as the optimisation proceeds.



(a)

(a)

Figure 1: $n = n' = 10$ re-evaluations of operating points $(0, 0)$ and $(0.1, 0.1)$, with added Gaussian noise with variances 0.5 and 0.1. Contours show the standard error and two times standard error plotted around each mean. *a)* Using true underlying variance. *b)* Variances estimated from each sample group.

2.4 Gaussian noise with unknown variance

During the course of an optimisation we envisage evaluating the objectives for each parameter set n times. Although our approximation of $\mathbf{f}(\mathbf{x})$ cannot incorporate information from re-evaluations at different parameterisations, if it is assumed that the variance of the corrupting noise is constant across parameter space then the repeated evaluations can be combined to learn the variance, σ^2.

Consider the re-evaluations $Y = \{y_i\}_{i=1}^n$ for a fixed parameter set $\mathbf{x}$. The likelihood of observing these evaluations is:

$$p(Y \mid \sigma^2, f(\mathbf{x})) = \prod_{i=1}^{n} p(y_i \mid \sigma^2, f(\mathbf{x})) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f(\mathbf{x}))^2 \right\}. \tag{19}$$

If we have some prior idea of the probability density, $p(\sigma^2, f(\mathbf{x}))$, of σ^2 and $f(\mathbf{x})$ it may be combined in Bayes rule with the likelihood to yield a posterior density, which incorporates the prior information and the likelihood:

$$p(\sigma^2, f(\mathbf{x}) \mid Y) = \frac{p(Y \mid \sigma^2, f(\mathbf{x}))p(\sigma^2, f(\mathbf{x}))}{p(Y)}. \tag{20}$$

In general we have no *a priori* belief about $f(\mathbf{x})$ and so choose the prior to be uniform on some suitably large interval of the real line. In some circumstances more may be known about the location of $f(\mathbf{x})$ (for example, that $f(\mathbf{x})$ lies in some interval or is positive) in which case more informative priors can be used. We choose a conjugate prior (Bernardo & Smith, 1994) for σ^2, namely the inverse gamma prior:

$$p(\sigma^2) = IG(\sigma^2 \mid a, b) = \frac{b^a}{\Gamma(a)} (\sigma^2)^{-(a+1)} \exp\left\{ -b/\sigma^2 \right\}. \tag{21}$$

The hyperparameters a and b control the shape and width of the prior; as $a, b \to 0$ the prior becomes non-informative, favouring no particular scale.

Combining the prior and likelihood to evaluate the posterior shows that the joint posterior is a Normal Inverse Gamma (NIG) density:

$$p(f(\mathbf{x}), \sigma^2 \mid Y) = \mathcal{N}(f(\mathbf{x}) \mid \bar{y}, \sigma^2/n) IG(\sigma \mid a', b'). \tag{22}$$

The posterior parameters are

$$a' = a + \frac{n-1}{2} \tag{23}$$

$$b' = b + \frac{1}{2}(n-1)S^2 \tag{24}$$

where

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{25}$$

is the unbiased estimator of the sample variance. The posterior marginal density $p(\sigma^2 \mid Y)$ is again an inverse gamma density with parameters a' and b'. The expected value of the variance is thus (Bernardo & Smith, 1994, page 119):

$$E[\sigma^2] = \frac{b'}{a'-1}. \tag{26}$$

This estimate of the expected variance can then be used in (15) and (18) to estimate domination probabilities. Note that (23) and (24) show that at least two evaluations must be made to contribute to the variance estimate; two evaluations are clearly necessary to obtain information about the spread of the noise.

The power of this method derives from the fact that the posterior density for σ^2 has the same form as the prior density, namely an inverse gamma density. This means that the

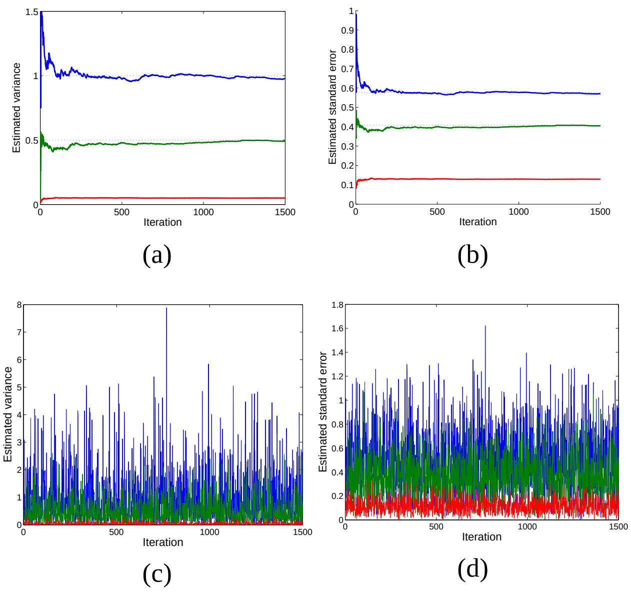


Figure 2: Noise variance (*a*) and standard error of objectives (*b*) using the Bayesian learning during optimisation. Sample estimates of noise variance (*c*) and standard error objective (*d*) $n = n' = 3$ evaluations made at each parameterisation.

posterior from set of evaluations can be used as the prior for the next set at a new parameter value, so that information on the variance is accumulated throughout the optimisation. Note that information about the successive objectives is not propagated because successive evaluations may be in quite different objective-space locations and propagating this information would unwarrantedly bias the search. The estimation of the variances is illustrated in Figures 2 and 3 which, for a three objective test problem described fully below, show the estimated noise variance and the estimated standard errors $\sigma_{\mathbf{x}}/\sqrt{n}$ and $\sigma_{\mathbf{x}'}/\sqrt{n'}$ as the optimisation proceeds. Also shown in the figures are the estimates of the noise variances and standard errors using just the sample estimates of the variances from each group of n or n' evaluations. As the figures show, the noise estimates converge rapidly to the true values, permitting accurate estimates of domination to be made. In contrast, the per group of evaluation sample estimates, which do not propagate information, do not converge and thus lead to poor estimates of domination throughout the optimisation.

In fact the marginal density of $f(\mathbf{x})$ is a student-t density (see, for example, (Bernardo & Smith, 1994)):

$$p(f(\mathbf{x}) \,|\, Y) = \mathrm{St}(f(\mathbf{x}) \,|\, \bar{y}, \frac{na'\sigma_{\mathbf{x}}^2}{b'}, 2a'). \tag{27}$$

The mean of this density is just $\bar{y}$ and its variance is

$$\mathrm{var}[f(\mathbf{x})] = \frac{1}{n}\frac{2a'}{(2a'-2)}\sigma_{\mathbf{x}}^2 \tag{28}$$

which can be used in place of the standard error. The tails of the student-t density decay more slowly than a Gaussian density, reflecting the additional uncertainty in $f(\mathbf{x})$ than

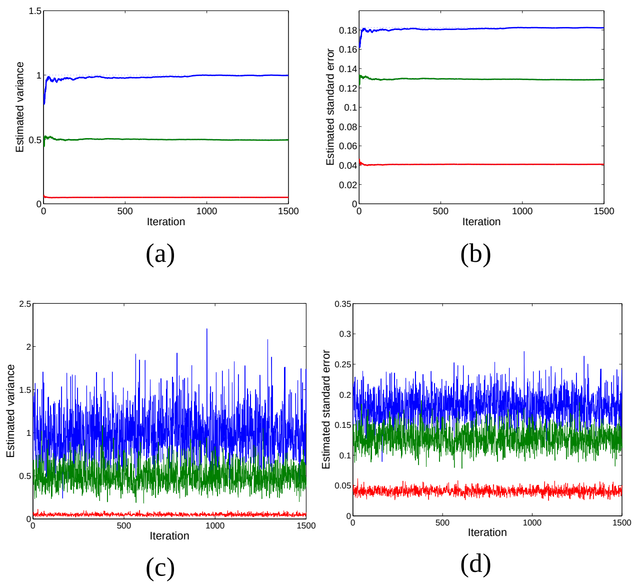


Figure 3: Noise variance (*a*) and standard error of objectives (*b*) using the Bayesian learning during optimisation. Sample estimates of noise variance (*c*) and standard error objective (*d*) $n = n' = 30$ evaluations made at each parameterisation.

if the noise variance were known. Properly the student-t density should therefore be used in (12). However, as the shape parameter $2a'$ becomes large the student-t approaches the Gaussian density. Reference to (23) shows that $2a'$ increases like the total number of evaluations for each objective, so approximating $p(f_d(\mathbf{x}) \,|\, Y)$ by Gaussian densities rapidly becomes a good approximation.

2.5 Variable noise

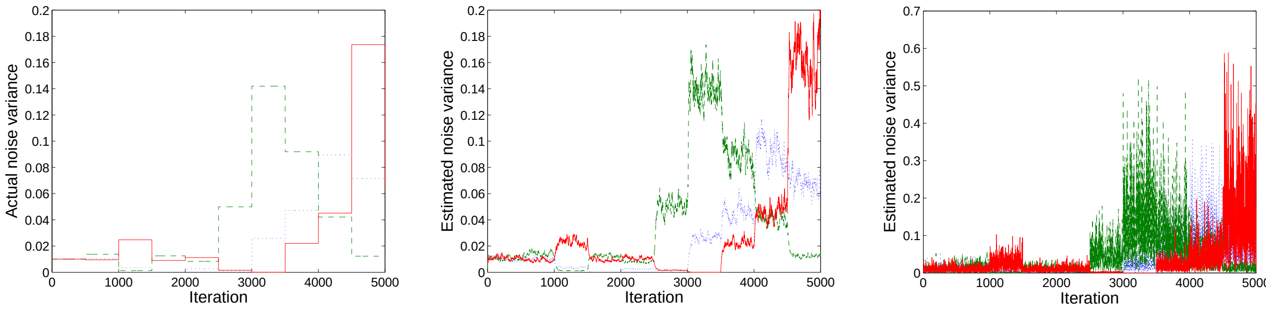
So far we have assumed that the noise for each objective is constant for all parameter values, although in practise this may not be the case, particularly if the noise enters the problem through the parameterisation. The assumption of constant noise in this case can be dangerous as it may lead to over-optimistic insertion of solutions into the archive. Although more sophisticated modelling schemes can be devised, the Bayesian updates for the inverse gamma parameters (equations (23) and (24)) may be modified to discount the contribution of historical data as follows:

$$a' = \eta a + \frac{n-1}{2} \tag{29}$$

$$b' = \eta b + \frac{1}{2}(n-1)S^2. \tag{30}$$

When $\eta = 0$ all prior information is forgotten and the measured noise variance is the sample variance, whereas $\eta = 1$ recovers the constant noise situation, in which all samples count equally regardless of how recently they were measured. Intermediate values of η exponentially discount historical samples.

Figure 4 illustrates the estimation of time varying noise on the modified DLTZ2 problem described below. As Figure



(a) (b) (c)

Figure 4: Time varying noise. *a:* True noise variance at each iteration. *b:* Estimates of noise variance for three objectives using Bayesian updates. *c:* Sample estimates of noise variance $n = 5$ evaluations taken for each parameterisation.

4a shows, the noise for each of the three objectives was varied in a step-wise manner every 500 iterations. Figure 4b shows that the modified scheme with $\eta = 0.95$ is well able to track the noise variance, while the variance estimates using no historical information (Figure 4c) are prone to extreme fluctuations.

3 Examples of optimisation with noise

Here we illustrate optimising with uncertainty in the multi-objective domain, with Guassian noise of unknown variance.

We create a modified version of the DTLZ2 test function (Deb *et al.* , 2002), which is formulated in its standard 3-objective parameterisation. Gaussian noise is then added at each evaluation, with variances 1, 0.5 and 0.05 added to the three different dimensions. The formulation of this augmented test function is given in Table 1, and is implemented here with $K = 8$.

Table 1: Test problem DTLZ2 of Deb *et al.* (2002) for 3 objectives, augmented with Guassian noise with variances v_1, v_2 and v_3.

$$f_1(\mathbf{x}) = \cos(x_1\pi/2)\cos(x_2\pi/2)(1+g(\mathbf{x})) + \mathcal{N}(0, v_1)$$
$$f_2(\mathbf{x}) = \cos(x_1\pi/2)\sin(x_2\pi/2)(1+g(\mathbf{x})) + \mathcal{N}(0, v_2)$$
$$f_3(\mathbf{x}) = \sin(x_1\pi/2)(1+g(\mathbf{x})) + \mathcal{N}(0, v_3)$$
$$g(\mathbf{x}) = \sum_{k=3}^{K}(x_k - 0.5)^2$$
$$0 \le x_k \le 1, \forall k$$

We then run a simple (1+1)–ES MOEA (as described in (Fieldsend & Everson, 2004; Everson & Fieldsend, 2005)) for 1500 iterations and view the estimated noise variance and standard error during this time - in relation to the true (known) values. Figure 2 shows plots of these values using $n = 3$ evaluations for each parameterisation, both using the Bayesian update method (Figures 2a and 2b), and by re-estimating at each iteration empirically from the samples at that iteration (Figures 2c and 2d). Figure 3 shows similar plots, but using $n = 30$ evaluations for each parameterisation instead of $n = 3$. Both figures show that after only a relatively few iterations a good approximation to the underlying noise properties is being made by the Bayesian update method. The variability of simply re-estimating the variance at each iteration from the samples at that iteration on the other hand generates a very volatile estimate of the variance, even when 30 samples are taken at each generation.

Figure 5 shows the fronts obtained for this problem when using Gaussian noise with variances 0.1 added to the three different objectives. The noiseless front for this problem is a shell of radius 1 in the positive octant, centred on the origin. The top row of figures correspond to using the Bayesian variance update methodology and 5 re-evaluations per iteration after 5000 iterations. The bottom row of figures correspond to using a traditional dominance form and archive update (a single evaluation per parameterisation), which is run for 25000 iterations (meaning both were run for the same number of function evaluations). As can be seen, the re-evalution method with variance update produces an archive front which is representative of the *true* operating front and the points are distributed across the front. The standard dominance implementation on the other hand creates a front which mis-directs the user – Figure 5c has many point in the unobtainable negative region of objective space, and the true operating points in Figure 5d lie much further away from their corresponding points in Figure 5c than those in Figures 5a and 5b. Points are also on average further away from the optimal front in 5d than 5b.

4 Representing uncertain fronts

Although evaluating the uncertainty in the *true* objectives is useful when optimising with uncertainty, it is important to evaluate and convey the uncertainty in the final front obtained. In order to do this we can evaluate the probability that a particular set objective values is dominated by a parameterisation on the front. This permits assessment of the degree of confidence that the performance will be better than the set of objective values.

In order to make this idea more concrete, we illustrate it

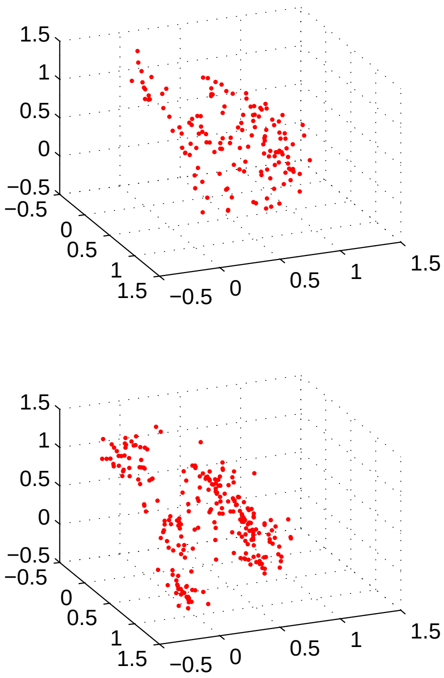


Figure 5: Fronts found during two different search methods – re-evaluations with Bayesian update of variance estimate, and standard (single evaluation) method. *Top left:* Mean of evaluations (5 evaluations per parameterisation). *Top right:* Actual (noiseless) evaluations. *Bottom left:* Assigned objective evaluations (1 evaluation per parameterisation). *Bottom right:* Actual (noiseless) evaluations.

on the optimisation of the Receiver Operating Curve for a classifier used for Short Term Conflict Alert (STCA) systems alerting whether pairs of aircraft are likely to become dangerously close (Fieldsend & Everson, 2004; Everson & Fieldsend, 2005). In this problem we aim to maximise the number of true positives while minimising the number of false positives. Uncertainty is present because the classifier is trained and evaluated on a finite-sized dataset. The uncertainty may be assessed by evaluating the classification rates on bootstrap samples of the data or by using a normal approximating to the variance of the bootstrap sample (Everson & Fieldsend, 2005). The estimated Pareto front was located using a standard MOEA; then by drawing samples for every solution in the estimated Pareto front F, we estimate the probability that any operating point is dominated by a solution on the front. Figure 6 shows the *probabilistic domination contours* for the STCA system obtained using this method. Clearly, quoting a true/false positive rates based on the 50% contour may be over-optimistic in the light of the uncertainty in the objectives and a more conservative assessment would report the rates for the 90% contour along with the probability of dominance. We point out the similarity of these probabilistic domination contours to the attainment surface technique for repeated runs of MOEAs (Grunert da Fonseca *et al.* , 2001; Fonseca *et al.* , 2005).

5 Conclusion

In this paper we have presented methods for multi-objective optimisation with uncertain objectives. When the noise characteristics are unknown sampling methods may be used to assess the probabilities of dominance, but these may be time-consuming when objective evaluations are expensive. If the noise can be assumed to be Normally distributed, we have presented a Bayesian algorithm for learning the noise variance. Empirical results indicate that this effectively incorporates historical information and rapidly converges to the correct value. A simple modification of this method permits the tracking of variable noise, which was identified as a problem by Di Pietro *et al* (2004).

We have assumed throughout that the noise contaminating each objective is independent of the noise on other objectives. This assumption is may be too strong in some cases, particularly for parametrically introduced noise. The Bayesian scheme can straightforwardly be modified to handle this case by using inverted Wishart densities in place of the inverse Gamma densities. However, handling non-Gaussian noise in an efficient manner remains an active area of research.

Finally, we advocate the use of probabilistic domination contours for the presentation of estimated Pareto fronts as these permit assessment of the degree of confidence that

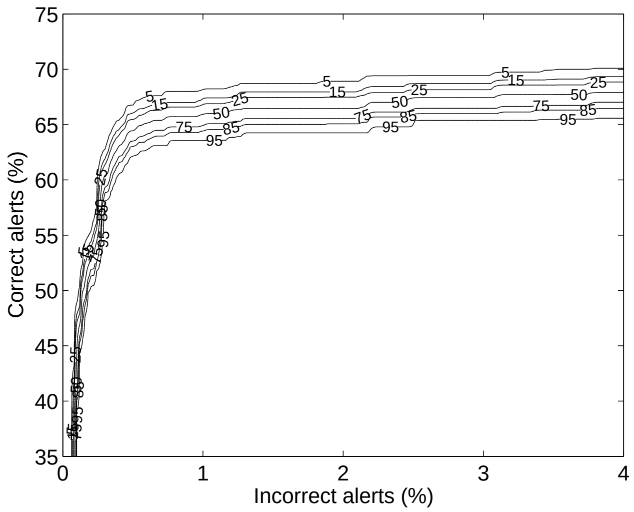


Figure 6: Probabilistic domination contours for correct alerts (true positive) and incorrect alerts (false positives) for the STCA system described in (Everson & Fieldsend, 2005).

may be placed in an operating point.

Acknowledgements

This work was supported in part by the EPSRC, grant GR/R24357/01. We thank Trevor Bailey, Adolfo Hernandez, Wojtek Krzanowski, Derek Partridge, Vitaly Schetinin and Jufen Zhang for their helpful comments.

References

Bernardo, J.M., & Smith, A.F.M. 1994. *Bayesian Theory*. Wiley.

Büche, D., Stoll, P., & Koumoutsakos, P. 2002. Multi-objective evolutionary algorithm for the optimization of noisy combustion. *IEEE Transactions on Systems, Man and Cybernetics*, **32**(4).

Deb, K., Agrawal, S., Pratap, A., & Meyarivan, T. 2000. A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II. *Pages 849–858 of: Proceedings of Parallel Problem Solving from Nature - PPSN VI.* Springer.

Deb, K., Thiele, L., Laumanns, M., & Zitzler, E. 2002. Scalable Multi–Objective Optimization Test Problems. *Pages 825–830 of: Congress on Evolutionary Computation (CEC'2002)*, vol. 1.

Di Pietro, A., While, L., & Barone, L. 2004. Applying Evolutionary Algorithms to Problems with Noisy, Time-consuming Fitness Functions. *Pages 1254–1261 of: Proceedings of the 2004 Congress on Evolutionary Computation. CEC2004,*, vol. 2. IEEE.

Everson, R.M., & Fieldsend, J.E. 2005. Multi-Objective Optimisation of Safety Related Systems: An Application to Short Term Conflict Alert. *IEEE Transactions on Evolutionary Computation*. Under review. Draft available from `http://www.dcs.ex.ac.uk/academics/reverson`.

Fieldsend, J.E., & Everson, R.M. 2004. ROC Optimisation of Safety Related Systems. *Pages 37–44 of:* Hernández-Orallo, J., Ferri, C., Lachiche, N., & Flach, P. (eds), *Proceedings of ROCAI 2004, part of the 16th European Conference on Artificial Intelligence (ECAI)*.

Fonseca, C.M., Grunert da Fonseca, V., & Paquete, L. 2005. Exploring the Performance of Stochastic Multiobjective Optimisers with Second-Order Attainment Function. *Pages 250–264 of:* Coello Coello, C.A., Hernández Aguirre, A., & Zitzler, E. (eds), *Evolutionary Multi-Criterion Optimization, EMO 2005.* LNCS, no. 3410. Springer.

Grunert da Fonseca, V., Fonseca, C.M., & Hall, A.O. 2001. Inferential performance assessment of stochastic optimisers and the attainment function. *Pages 213–225 of:* Zitzler, E., Deb, K., Thiele, L., Coello Coello, C.A., & Corne, D. (eds), *Evolutionaty Multi-Criterion Optimization, EMO 2001.* LNCS, no. 1993. Springer.

Hughes, E.J. 2001. Evolutionary multi-objective ranking with uncertainty and noise. *Pages 329–342 of:* Zitzler, E., Deb, K., Thiele, L., Coello Coello, C.A., & Corne, D. (eds), *Evolutionary multi-criterion optimization, emo 2001.* LNCS, vol. 1993. Springer.

Knowles, J.D., & Corne, D. 2000. Approximating the Non-dominated Front Using the Pareto Archived Evolution Strategy. *Evolutionary Computation*, **8**(2), 149–172.

Parmee, I.C. (ed). 2004. *Adaptive Computing in Design and Manufacture VI.* Springer.

Rana, S., Whitley, L.D., & Cogswell, R. 1996. Searching in the Presence of Noise. *Pages 198–207 of: Parallel Problem Solving from Nature, 4.* Springer Verlag.

Stagge, P. 1998. Averaging Efficiently in the Presence of Noise. *Pages 188–200 of:* Eiben, A.E., Bäck, T., Schoenauer, M., & Schwefel, H.-P. (eds), *PPSN V: Proceedings of the 5th International Conference on Parallel Problem Solving from Nature.* Springer.

Teich, J. 2001. Pareto-front exploration with uncertain objectives. *Pages 314–328 of:* Zitzler, E., Deb, K., Thiele, L., Coello Coello, C.A., & Corne, D. (eds), *Evolutionary Multi-Criterion Optimization, EMO 2001.* LNCS, vol. 1993. Springer.

Zitzler, E., & Thiele, L. 1999. Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Approach. *IEEE Transactions on Evolutionary Computation*, **3**(4), 257–271.