Editorial Manager(tm) for Optimization and Engineering
Manuscript Draft

Manuscript Number: OPTE538

Title: Engineering Design Applications of Surrogate-Assisted Optimization Techniques

Article Type: Special 10th Anniversay Issue

Keywords: surrogate modeling; optimization; sampling plans; expected improvement; multi-objective optimization

Corresponding Author: Dr. Andras Sobester,

Corresponding Author's Institution: University of Southampton

First Author: Andras Sobester

Order of Authors: Andras Sobester; Alexander I Forrester; David J Toal; Es Tresidder; Simon Tucker

Noname manuscript No.
(will be inserted by the editor)

Engineering Design Applications of Surrogate-Assisted Optimization Techniques

András Sóbester ·
Alexander I. J. Forrester ·
David J. J. Toal ·
Es Tresidder ·
Simon Tucker

Received: date / Accepted: date

Abstract The construction of models aimed at learning the behaviour of a system whose responses to inputs are expensive to measure is a branch of statistical science that has been around for a very long time. Geostatistics has pioneered a drive over the last half century towards a better understanding of the accuracy of such 'surrogate' models of the expensive function. Of particular interest to us here are even more recent advances related to exploiting such formulations in an optimization context. While the classic goal of the modelling process has been to achieve a uniform prediction accuracy across the domain, an economical optimization process may aim to bias the distribution of the learning budget towards promising basins of attraction. This can only happen, of course, at the expense of the global exploration of the space and thus finding the best balance may be viewed as an optimization problem in itself. We examine here some of the state-of-the-art solutions to this type of balancing exercise through the prism of several simple, illustrative problems, followed by two 'real world' applications: the design of a regional airliner wing and the multi-objective search for a low environmental impact house.

Keywords surrogate modeling · optimization · sampling plans · expected improvement · multi-objective optimization

A. Sóbester, A. I. J. Forrester, D. J. J. Toal
University of Southampton
School of Engineering Sciences
Tel.: +44 23 8059 2350
Fax: +44 23 8059 4813
E-mail: a.sobester, alexander.forrester, d.j.j.toal@soton.ac.uk

E. Tresidder, S. Tucker
Graduate School of the Environment
Centre for Alternative Technology
Machynlleth, UK
Tel.: +44 1654 703065
E-mail: simon.tucker@cat.org.uk

1 Background and a Key Assumption

The origins of many of the surrogate-based optimization techniques in use today can be traced back to geology – more specifically to the science of *geostatistics*, which has played an important role in mining engineering. Although the applications of geostatistics vary, the fundamental problem is usually formulated as follows. The optimum location is sought for a mineral extraction operation – this is usually the maximum ore grade area. The ore grade in a given location can be obtained through drilling a borehole, but this is an expensive operation so it must be comissioned sparingly. The geostatistical solution is to build up a spatial model of ore grade distribution based on the few known borehole values and use the predictions of this model as a guide to identifying the best mining location, or, if the budget permits it, the most informative locations for further boreholes.

The central assumption behind geostatistical models is that the ore grade in a given location is correlated with that measured at a nearby borehole, as well as with its distance from that borehole. Some argue that this is a false premise, particularly for more relaxed definitions of 'nearby'. In other sciences, however, such specifically geological objections are of little consequence – if the response functions being modeled can be assumed to behave in a smooth, continuous manner, spatial statistics can be a very powerful tool in the optimization of expensive black-box functions and the relevant techniques, developed in geostatistics or elsewhere, can be deployed with few reservations. This is the stance we adopt here.

We present two case studies highlighting the use of surrogate model-based optimization algorithms in 'real-world' design problems (Sections 5 and 6), preceded by a roundup of some of the key ingredients of these procedures (Sections 2, 3 and 4). We begin by considering the issue of where to make the initial measurements of the expensive objective function.

2 Sampling the Design Space

It almost goes without saying that the pre-requisite of a successful surrogate-based optimization process is a surrogate $\widehat{f}$ that generalizes reasonably well, that is, it is capable of predicting fairly accurately at sites other than those included in the *sampling plan* where the objective function value has been evaluated directly. At the very least, the surrogate has to be capable of predicting the *trends* of the landscape accurately, while the precise scaling of the surface is less important from an optimization perspective.

In turn, the generalization properties of the surrogate are intimately linked with the space-filling properties of the sampling plan. It makes intuitive sense to spread out the designs included in the initial sampling plan in a manner that will give as uniform a coverage of the design domain D as possible. However, the exact definition of 'uniform' is neither immediately obvious, nor is it always easy to translate into obtaining an initial sample $\mathbf{X} = \left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots \mathbf{x}^{(n)}\right\}$.

One of the reasons for this is the relative sparsity of these observations – after all, at this stage we are usually aiming to build an approximation that is just about accurate enough to give the optimization process an initial handhold and to give us a general idea as to what the landscape is likely to look like (in terms of general trends, multi-modality, range of values, etc.).

The other reason is the often very high number of design variables and this, due to the 'curse of dimensionality' it brings with it, is a very significant aspect that will make or break the subsequent optimization process too. It therefore makes sense to minimize k, the number of design variables, at the outset. It is worth bearing in mind that if a certain level of prediction accuracy is achieved by sampling a one-variable space in n locations, to achieve the same sample density in a k-dimensional space, n^k observations are required. There are two fundamental ways of reducing k: first, by a judicious parameterization of the design and second, by *screening* [Welch et al (1992)] variables for their impact on the objective function [see, for example, the work of Morris and Mitchell (1995) on how to achieve the latter through a small number of observations and with minimal assumptions regarding the objective function landscape].

With the set of k important variables established, a small sampling plan $\mathbf{X} = \left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots \mathbf{x}^{(n)}\right\}$ can be created, which, along with the model fitted to the corresponding responses $y^{(1)} = f(\mathbf{x}^{(1)}), y^{(2)} = f(\mathbf{x}^{(2)}), \dots y^{(n)} = f(\mathbf{x}^{(n)})$, will form the starting point of the optimization process. Just how small should this initial sampling plan be is an open question, though there is some empirical evidence [Sóbester et al (2003)] that a good, relatively problem-independent choice might be around 30% of the overall computational budget.

We mentioned earlier the difficulties of defining the 'space-fillingness' of sampling plans. A naive approach could be to use what statisticians call a *full factorial* plan, which samples all variables at all levels, giving a uniform grid. While uniform, such plans tend to require many points, especially if the number of variables is high (the curse of dimensionality again!). They also suffer from having bad projective properties, that is, when projected onto the axes, many points will overlap. *Latin hypercubes* [McKay et al (1979)] offer a cure for both of these problems, as they can be made up of any number of points and the points are uniformly distributed along all of the axes.

An important point to be made here is that uniform projection properties are desirable, but they do not equate to 'space-fillingness'. Random Latin hypercubes are easy to construct by collating random permutations, but some of these will fill the space more uniformly than others and much effort has been devoted over the last decades to finding a recipe for choosing the best sampling plan across the space of all Latin hypercubes of a given size and dimensionality.

We shall not delve into the details and comparative merits of Latin hypercube sampling plan optimization techniques [see, e.g., Forrester et al (2008) for more details on the former], we merely note that for the experiments described here we adopt the widely used *maximin* criterion introduced by Johnson et al (1990), defined as follows. Let $d_1, d_2, \dots d_m$ be the list of the unique values of distances between all possible pairs of points in a sampling plan $\mathbf{X}$, sorted

in ascending order. Further, let $J_1, J_2, \ldots J_m$ be defined such that J_j is the number of pairs of points in $\mathbf{X}$ separated by the distance d_j. We will call $\mathbf{X}$ a maximin plan amongst all available plans if it maximizes d_1 and, among plans for which this is true, minimizes J_1. It can be shown that this is equivalent to the so-called *D-optimality* criterion used in linear regression.

3 Modeling Approaches

Once the sampling plan is constructed, we can build our initial approximation $\widehat{f}$ of the expensive objective function. There is an infinity of functions we could conceive with the property that they reproduce (interpolate) the set of observations $\left\{\mathbf{x}^{(i)} \rightarrow y^{(i)} = f(\mathbf{x}^{(i)}) \mid i = 1 \ldots n\right\}$ based on the sampling plan. However, the vast majority of these functions would be nonsensical and they would generalise very poorly – that is, they would be practically useless at predicting the function value at other sites, which is, of course, their *raison d'etre*.

One way of generating approximations with good generalisation properties is to choose a class of parametric functions that tend to emulate well the types of objective functions we are likely to wish to optimize. Fitting the function to the data generated according to the sampling plan involves estimating the parameters of such functions to maximize the generalisation properties of the approximation [via, for example, *cross-validation* or *likelihood maximization*, see Hastie et al (2001), Cherkassky and Mulier (1998)].

In terms of the generic shape of the surrogate, a popular choice for objective functions that occur in engineering sciences is the linear combination of *basis functions* of various shapes and supports. Perhaps the simplest such formulation is that of the *radial basis function* approximator

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n_b} w_i \phi(\|\mathbf{x} - \mathbf{x}_c^{(i)}\|), \tag{1}$$

where the bases ϕ, which can take a variety of forms, are defined in terms of the distance of the predicted point from a set of n_b basis function centres $\mathbf{x}_c$ (often chosen to coincide with the sampling plan points, thus allowing the construction of an interpolating model).

From an optimization perspective, beyond good generalisation, the approximator should permit relatively straightforward computation of an error measure, that is, another function $\widehat{s}(\mathbf{x})$ that quantifies our confidence in the values predicted by $\widehat{f}$. If the responses we are fitting the function to are assumed to be realisations of stochastic processes (an artifice often used even when these points come from deterministic computer simulations), a Gaussian radial basis function model can be used effectively to estimate prediction errors – these predictors have the form of (1), with $\phi(r) = \exp\left(-(r^2)/(2\sigma^2)\right)$. There is a single parameter that has to be estimated here, σ, to maximize the generalisation properties of the predictor. This can be a good thing, as it

makes the training process quite easy – it is, essentially, a one dimensional optimization problem. On the other hand, it suggests, that the flexibility of the model is somewhat limited. In other words, whether we choose the likelihood maximization approach or some other parameter estimation technique, there is only so much generalisation ability that we can achieve.

Kriging, one of the techniques originated in geostatistics, as discussed in the introduction [named after mining engineer D. G. Krige, whose work in this field stretches back to the 50s – see, for example, Krige (1951)] is a similarly structured but more flexible model, which permits differential shape control in all of the dimensions of the search space. A kriging predictor is thus far more difficult to fit (at least k model parameters to estimate), but promises better generalisation, a property that has undoubtedly contributed to the popularity of the technique in the design optimization community, gaining momentum from the late 90s onwards [see Jones et al (1998); Simpson and Mistree (2001); Forrester and Keane (2008)].

Kriging also permits fitting to multiple sets of data representing the same function, but at *different levels of fidelity* – this is possible via *co-kriging* [Kennedy and O'Hagan (2000); Forrester et al (2007)], a formulation, which can also be used to build a model based on a function and its gradients [Chung and Alonso (2002)], on values of the same quantity obtained from different sources [Krajewski (1987)], etc.

We are unlikely to find the global optimum simply by searching our surrogate of the objective function: the surrogate is likely to have some inaccuracies and, depending on the level of sampling, may not contain the same basins of attraction as the true function [Jones (2001)]. We therefore enhance the surrogate by additional sampling of the objective function, and it is the selection of the position of these *infill points* which we turn our attention to next.

4 Towards a Global Optimum

4.1 Balancing local exploitation with global exploration

The most obvious way to position an infill point is to exploit the information at hand and place the point at the minimum of the surrogate. Such a strategy, when coupled with an interpolating surrogate (e.g. a radial basis function), will quickly descend into the basin of attraction (e.g., see Figure 1). However, in multi-modal landscapes, this may well not be the global optimum. We need a method that can branch out from the predicted minimum and explore other areas of the landscape.

As we indicated earlier, Gaussian process based models allow us to calculate error estimates for our surrogate, and we can use these to highlight areas where we are the least certain about the shape of the landscape. Positioning infill points in these areas will increase the global accuracy of the model, improving its ability to identify the region of the global optimum. However, a pure maximum estimated error based infill strategy will only *explore* the

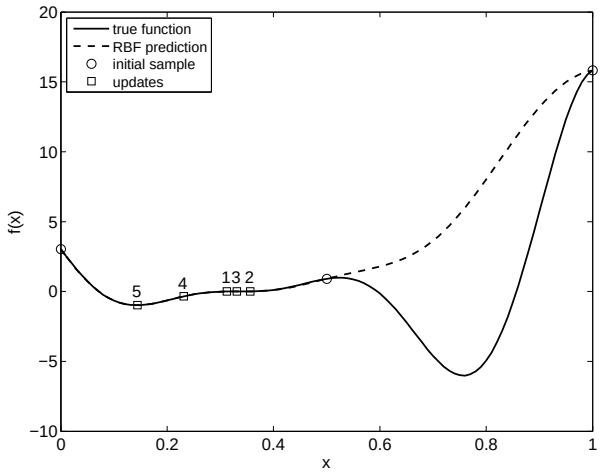

Fig. 1 A minimum Gaussian RBF prediction based infill strategy starting from a three point sample followed by five infill points fails to find the global optimum of the function.

model, while we need some element of *exploitation* to find the bottom of the lowest basin of attraction once the exploration has located it. The key to a successful infill strategy is balancing exploration and exploitation. We must never completely trust the model at a value of $\mathbf{x}$ where we have not sampled, but we must trust it sufficiently to fully exploit promising basins of attraction.

A popular infill strategy in both academia and industry is to maximize the expectation of the improvement [Matheron (1963)] over the best point found so far:

$$E[I(\mathbf{x})] = \begin{cases} (y_{\min} - \widehat{y}(\mathbf{x}))\Phi\left(\frac{y_{\min}-\widehat{y}(\mathbf{x})}{\widehat{s}(\mathbf{x})}\right) + \widehat{s}(\mathbf{x})\phi\left(\frac{y_{\min}-\widehat{y}(\mathbf{x})}{\widehat{s}(\mathbf{x})}\right) & \text{if} \quad \widehat{s}(\mathbf{x}) > 0 \\ 0 & \text{if} \quad \widehat{s}(\mathbf{x}) = 0 \end{cases} \tag{2}$$

where $\Phi(.)$ and $\phi(.)$ are the normal cumulative distribution function and probability density function (pdf) respectively, $\widehat{y}(\mathbf{x})$ is the Gaussian process based prediction (its mean) and $\widehat{s}(\mathbf{x})$ is the estimated error. Figure 2 shows the progress of a $\max\{E[I(\mathbf{x})]\}$ infill strategy which finds the region of the global optimum of the function where pure exploitation failed in Figure 1.

This is an attractive infill criterion since, not only does it offer a good balance between exploitation and exploration, it is also, in itself, an informative quantity of how the optimization is progressing. Caution should be taken though, because the expected improvement is often lower than the actual improvement that might be obtained if the infill strategy were continued further. This is because $\widehat{s}(\mathbf{x})$ is often an under estimator. As such, although the expected improvement is a very good infill criterion, it is often a poor convergence criterion. Because of its reliance on the Gaussian process error estimates, the method can be slow to stop exploiting and resume exploitation of

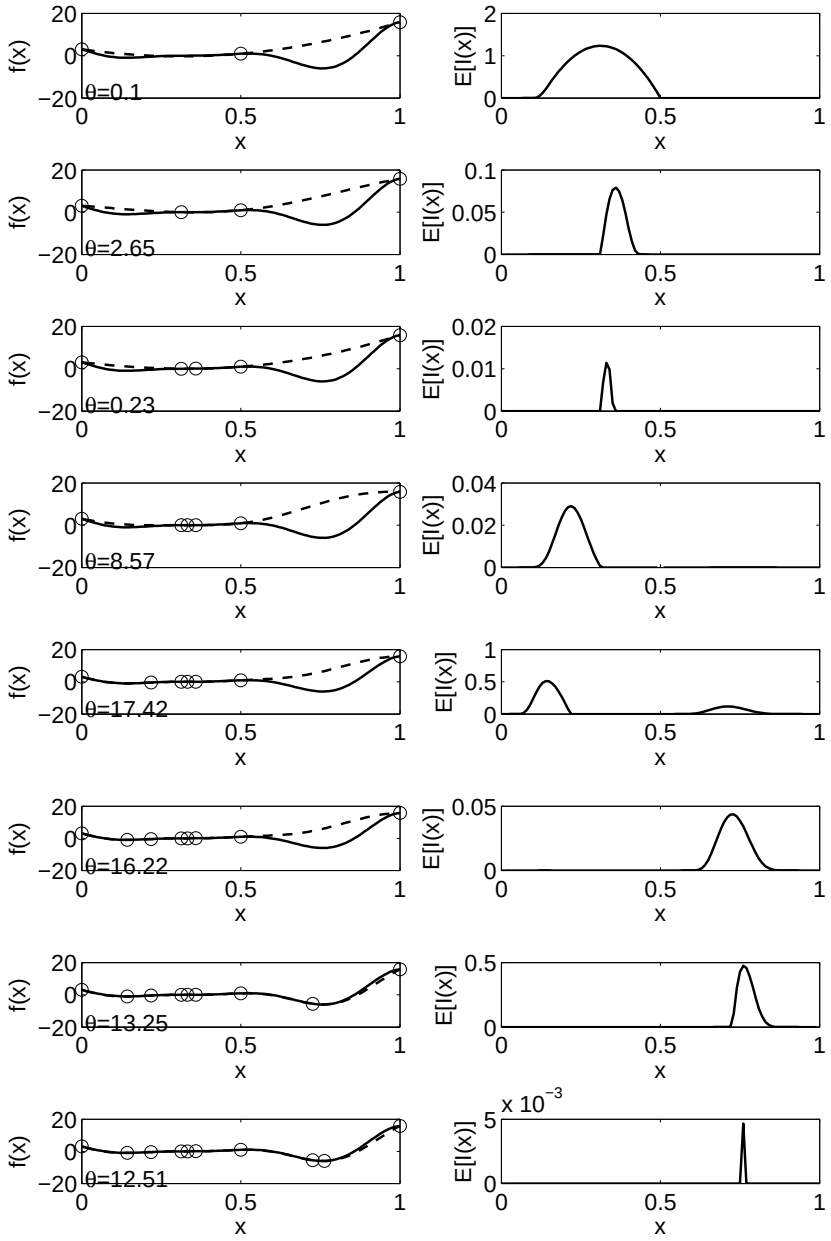

Fig. 2 A $\max\{E[I(\mathbf{x})]\}$ infill strategy starting from a three point sample converges towards the global optimum of the function.

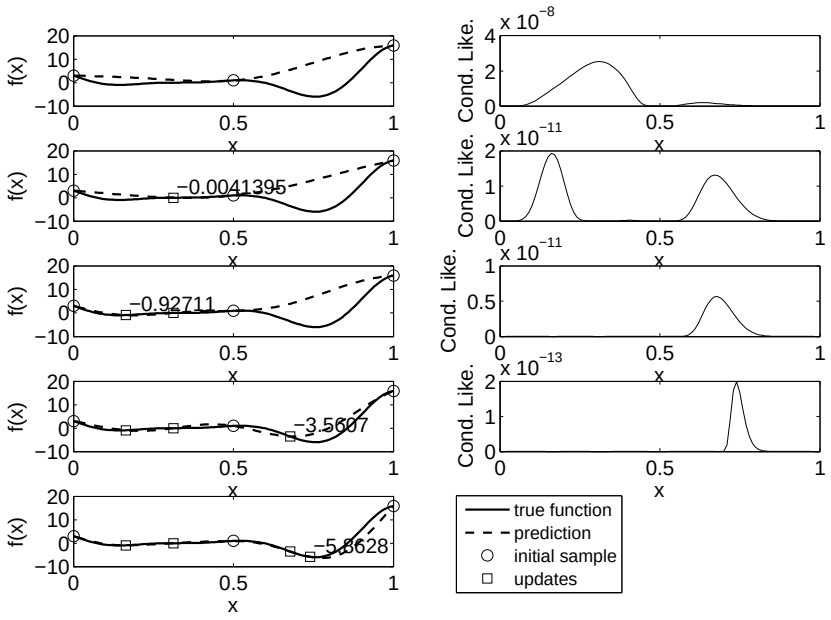

Fig. 3 A Gaussian RBF goal seeking infill strategy starting from a three point sample quickly finds the only region containing a value equal to or less than the goal.

deceptive functions if a local basin is not, in fact, the global optimum. In some unlikely situations, where a deceptive function is compounded by an unlucky sample, $\max\{E[I(\mathbf{x})]\}$ may in fact fail to find even a local optimum.

The frailties of $\max\{E[I(\mathbf{x})]\}$ lead us to another infill strategy - *goal seeking* [Jones (2001)]. If we are able to hazard a guess as to what the minimum value of the objective function might be, or, at least, a value we would be happy with – perhaps a percentage improvement in performance over a current product – we can use this as a goal for the infill criterion to search for. We do this using the notion of a *conditional likelihood*. We posit the hypothesis that our prediction passes through our goal at a given $\mathbf{x}$ and maximize the likelihood by varying the model parameters (e.g. the Gaussian kernel variance σ^2). We can compare this to likelihoods of other $\mathbf{x}$'s to find the $\mathbf{x}$ which maximizes the likelihood conditional upon the prediction passing through the goal. We position the next infill point at this $\mathbf{x}$. This infill criterion is not reliant upon accurate error estimates and is a stopping criterion in itself. Figure 3 shows how a goal seeking approach, with a goal of $f(\mathbf{x}) = 5$ finds the region of the global optimum; the only region containing a value equal to or less than the goal.

4.2 Dealing with constraints

Let us now consider how constraints can be managed in a surrogate-based global search. The most commonly used method of avoiding regions which violate constraints is through the application of penalty functions. In most cases penalty functions can be applied in the usual manner, the exception being when using a $\max\{E[I(\mathbf{x})]\}$ or $\max\{P[I(\mathbf{x})]\}$ (probability of improvement) based search. Here, y_{min} should be replaced with the minimum observed function value which satisfies the constraint.

A more elegant method of applying a constraint is to multiply $E[I(\mathbf{x})]$ by the probability of the constraint being met:

$$P\left[g(\mathbf{x}) < c\right] = \Phi\left(\frac{c - \widehat{g}(\mathbf{x})}{\widehat{s}(\mathbf{x})}\right). \tag{3}$$

where $g(\mathbf{x})$ is the constraint function and c is the constraint limit, i.e. the constraint function must be below this value [Schonlau (1997)]. The first plot of Figure 4 shows the one variable function from the previous examples along with a constraint function (simply the negative of the objective minus a constant) and Kriging predictions of the two based on the four sample points shown. The dashed line represents the constraint limit. The second plot shows $E[I(\mathbf{x})]$, the third plot shows the probability of meeting the constraint, and the fourth plot shows the product of these – our constrained expected improvement. Note how multiplying by the probability forces the expectation away from $x = 0.66$ where it is known that the constraint is violated. In figure 5 we see that the first and second infill point satisfy the constraint, but fail to find the global optimum. The third infill point actually violates the constraint, but improves both predictions in this area such that the fourth infill point, based on the $E[I(\mathbf{x})]P\left[g(\mathbf{x}) < c\right]$ shown, is positioned on the right hand side of the constraint boundary (the ninth infill point finds the, rather deceptive, global optimum in this case). Note that in some situations this method will suffer from the same problems as unconstrained $\max\{E[I(\mathbf{x})]\}$ searches.

4.3 Multiple objective functions

Recently the literature has been very active on the subject of multi-objective optimization, with multi-objective genetic algorithms such as NSGA-II [Deb et al (2002)] proving popular. Such methods can be directly applied to the search of surrogate model predictions. However, they can be slow or fail to find all areas of non-dominated solutions – so-called Pareto optimal solutions. Keane (2006) constructed a dual-objective probability of improvement formulation (open to extension to further objectives) using a two dimensional Gaussian probability density function of independent objectives:

$$P[Y_1(x) < \mathbf{y}_1^* \cap Y_2(x) < \mathbf{y}_2^*] = \Phi\left(\frac{y_1^{*(i)} - \widehat{y}_1(x)}{\widehat{s}_1(x)}\right) \tag{4}$$

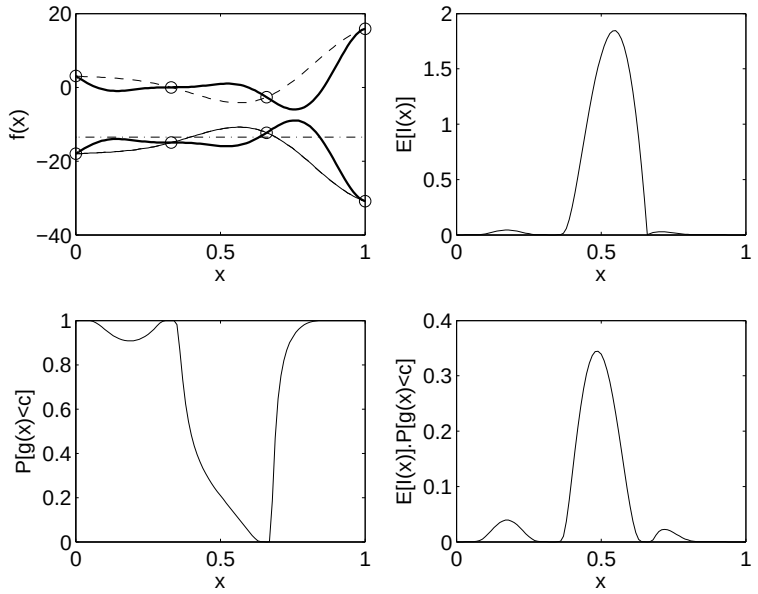

Fig. 4 Predictions of the objective and constraint functions based on four sample points, with the constraint limit shown as a dashed line (first plot), the unconstrained $E[I(\mathbf{x})]$ (second plot), the probability of meeting the constraint (third plot) and the constrained expected improvement (final plot).

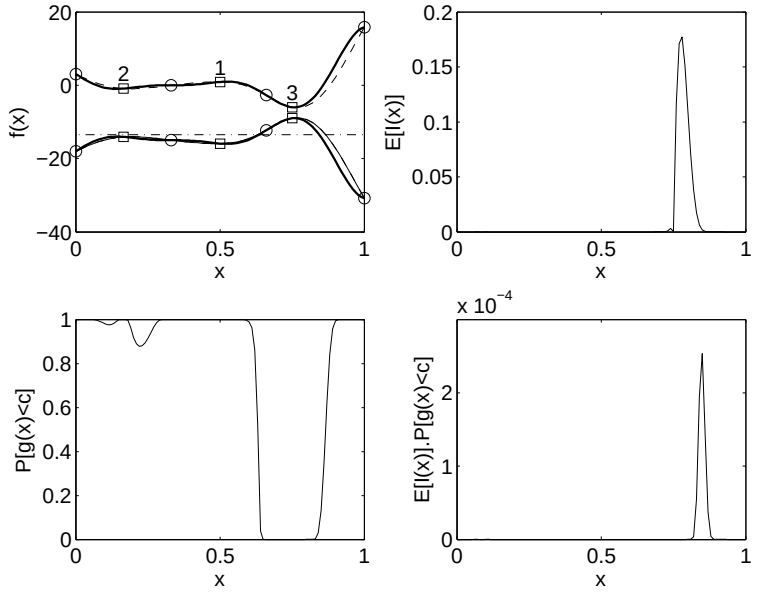

Fig. 5 The build up of the constrained expected improvement after three infill points have been applied.

$$+\sum_{i=1}^{m-1}\left\{\Phi\left(\frac{y_1^{*(i+1)}-\widehat{y}_1(x)}{\widehat{s}_1(x)}\right)-\Phi\left(\frac{y_1^{*(i)}-\widehat{y}_1(x)}{\widehat{s}_1(x)}\right)\right\}$$
$$\times\Phi\left(\frac{y_2^{*(i+1)}-\widehat{y}_2(x)}{\widehat{s}_2(x)}\right)$$
$$+\left\{1-\Phi\left(\frac{y_1^{*(m)}-\widehat{y}_1(x)}{\widehat{s}_1(x)}\right)\right\}\Phi\left(\frac{y_2^{*(m)}-\widehat{y}_2(x)}{\widehat{s}_2(x)}\right).$$

Where $\mathbf{y}_1^* = \left\{y_1^{*(1)}, y_1^{*(2)}, \ldots, y_1^{*(m)}\right\}$ is the current set of non-dominated objective function values for objective one and $\mathbf{y}_1^*$ is the corresponding set of values for objective two.

The first plot in Figure 6 shows two Kriging predictions of two objective functions (the first function is that used in the previous examples) based on a four point initial sample. The first infill point (found in this case by maximizing the related dual-objective expected improvement Keane (2006), which is more appropriate for this simple 1D example, as probability of improvement tends to exploitation in such simple problems) is to be placed at $x = 0.52$. The bottom three plots in Figure 6 indicate why. They show the two dimensional pdfs under the current Pareto front at $x = 0.25, 0.5, 0.75$. These pdfs are centred around the Kriging predictions at these points $\widehat{y}_1(x), \widehat{y}_2(x)$ and their variance is the Kriging errors $\widehat{s}_1^2(x), \widehat{s}_2^2(x)$. To find the probability of improvement over the three non-dominated points, this pdf is integrated under the Pareto front (to find the expected improvement we then find the moment of this area about the closest non-dominated point to the centroid of the integral). It can be seen in Figure 6 that the pdf at $x = 0.5$ has the greatest area under the Pareto front and an infill point is duly selected. Clearly this point is not Pareto optimal, but following this the search finds the area of non-dominated points in the $x = 0.25$ region. Figure 7 shows the selection of the fourth infill point. Here there is a high probability of improvement for $x = 0.25$ and $x = 0.75$ (there is virtually no probability of improvement at $x = 0.5$ as we have already sampled near to here). In fact the probability is marginally higher at $x = 0.25$, but the longer moment arm from the centroid of the pdf at $x = 0.75$ to the first dominated point means that the infill point is placed at $x = 0.74$.

Clearly the ability to balance exploitation and exploitation in a multi-objective sense when using surrogate models is an attractive prospect. The method can also be extended to constrained optimization by a simple probability of feasibility multiple.

Having covered some of the key challenges and techniques of surrogate model-based engineering design optimization, let us now revisit them once more, this time in the light of some more complex case studies.

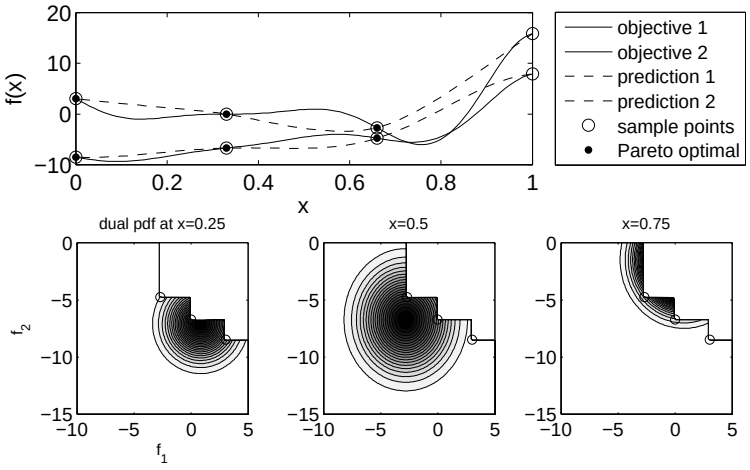

Fig. 6 Initial Kringing predictions based on four sample points and the pdfs under the Pareto front, indicating that the highest probability of improvement is in the $x = 0.5$ region.

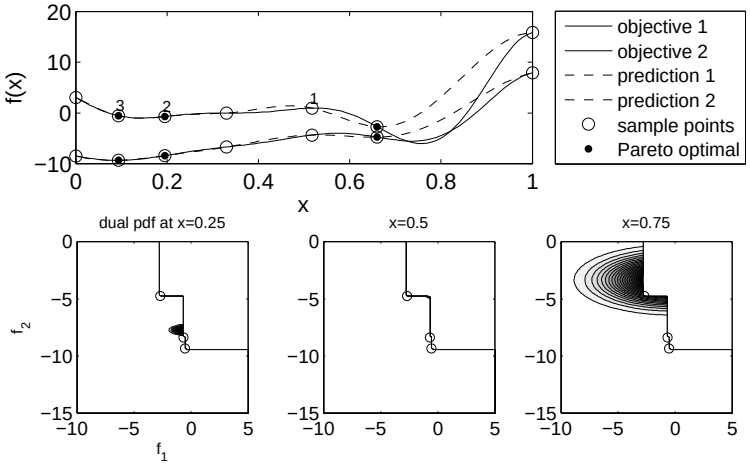

Fig. 7 After three infill points the search finds the area of Pareto optimal solutions in the $x = 0.75$ region based on the dual pdfs shown.

5 Aerodynamic Optimization of a Regional Airliner Wing

Let us consider the following wing design problem. The aircraft is a small regional turboprop airliner, cruising at an altitude of 26,000 feet at a speed of 260 knots, with a typical cruise weight of 10t. The aspect ratio and the dihedral of the wing are fixed (at 10.3 and 7° respectively) and, as a result of fuel capacity considerations, the volume of the wing is fixed too at 3.25 m^3. The root and tip cross sections (which will determine all other cross sections

through spanwise linear interpolation) must be designed for minimum cruise drag (at the fixed lift value of 10t).

There are two essential pre-requisites for the solution of this problem: a parametric wing geometry description and a drag prediction capability. We discuss these next.

5.1 A Parametric Wing

Parametric lifting surface geometries are amongst the most widely studied shapes in design optimization, efforts in this direction going back far beyond the advent of computational design search (consider, for instance, the ubiquitous NACA profiles of the 1930s). While a comprehensive review of the correspondingly weighty literature on the subject would be misplaced here, it is worth outlining at least the highest level of one of the many possible taxonomies of existing approaches.

We could split parametric geometries into two broad categories. First, a number of generic formulations can be applied to the description of lifting surfaces: Non-Uniform Rational B-Spline (NURBS) surfaces, a *de facto* standard in CAD systems, are a frequently encountered example [see Samareh (2001) for more]. The second category comprises models specific to aerodynamic design, that is, expressions whose structure was specifically designed to allow the reproduction of (classes of) aerodynamic shapes. A classic example is the 'bump' function parameterization of Hicks and Henne (1978). The advantage of the former category is seamless integration with CAD engines. The latter group of techniques yields design spaces tailored to aerodynamic design applications, that is, design spaces that only include a relatively small percentage of 'aerodynamically nonsensical' shapes – a key advantage when considered in the light of the computational cost implications of searching unnecessarily expansive design spaces[1].

It is for this reason that here we opt for the second category. More specifically, we define the airfoil sections by means of a Kulfan transformation [also known as the Class-Shape Transformation or CST for short, see Kulfan (2008)], using the 'airfoil' class function. CST is, essentially, a two level approximation model, where a so-called *class function* captures the essential, shared features of the family of shapes being considered (in this case, the family of airfoils), with a set of *shape functions* approximating the more specific detail. In common with Kulfan's original paper, we use a set of Bernstein polynomials as shape functions, their chief attraction being that for any chosen order they always add up to one, thus providing an obvious set of baseline values (with all Bernstein coefficients set to one, the approximation will simply yield the class function, a basic, symmetrical airfoil). Here is the Kulfan Transformation of the upper surface of a generic airfoil (the lower surface being defined in the same way):

[1] A hybrid approach is conceivable too, aimed at capturing the advantages of both classes of models – see Sóbester (2009).

$$z^{\mathrm{u}}(x,\ v_0^{\mathrm{u}}, v_1^{\mathrm{u}}, \ldots v_{n_{\mathrm{BP}}^{\mathrm{u}}}^{\mathrm{u}}, v_{\mathrm{LE}}^{\mathrm{u}}) = \underbrace{\sqrt{x}(1-x)}_{\text{class function}} \underbrace{\sum_{r=0}^{n_{\mathrm{BP}}^{\mathrm{u}}} v_r^{\mathrm{u}}\ C_{n_{\mathrm{BP}}^{\mathrm{u}}}^{r} x^r (1-x)^{n_{\mathrm{BP}}^{\mathrm{u}}-r}}_{\text{scaled Bernstein partition of unity}} + \underbrace{x\sqrt{1-x}\ v_{\mathrm{LE}}^{\mathrm{u}}(1-x)^{n_{\mathrm{BP}}^{\mathrm{u}}}}_{\text{leading edge camber line shaping term}}, \quad (5)$$

where $C_{n_{\mathrm{BP}}^{\mathrm{u}}}^{r} = \frac{n_{\mathrm{BP}}^{\mathrm{u}}!}{r!(n_{\mathrm{BP}}^{\mathrm{u}}-r)!}$. In addition to the class- and shape functions we also adopt a supplementary term for the more effective shaping of the camber line near the leading edge of the airfoil[2]. One could potentially link a design variable to each of the $n_{\mathrm{BP}}^{\mathrm{u}} + 2$ available degrees of freedom ($n_{\mathrm{BP}}^{\mathrm{u}} + n_{\mathrm{BP}}^{\mathrm{l}} + 4$, considering both surfaces), but here we opt (arbitrarily) to work with a subset. More specifically, choosing $n_{\mathrm{BP}}^{\mathrm{u}} = 3$ and $n_{\mathrm{BP}}^{\mathrm{u}} = 4$ (experience suggests that lower parametric airfoil surfaces in this class require slightly more flexibility than the upper surfaces), we fix the value of the first of the lower surface coefficients (which, effectively, determines the lower leading edge radius) on both the root and the tip airfoil. The remaining ten coefficients (per airfoil) are allowed to deviate (up or down) from a set of central (baseline) values by 0.1 at the root and by 0.07 at the tip (the maximum deviations being limited here for structural reasons). We take these central values from the the Kulfan transformations of two classic airfoils: NACA63A418 (root) and NACA63A412 (tip).

We impose a proportionality link between the root and tip deviations. Thus, for example, variable one (x_1) will determine the deviations from the baseline values of the first Bernstein coefficients on the upper surfaces of both the root and tip airfoils, so

$$v_1^{\mathrm{uROOT}} = -0.1 + x_1 * 0.2 + v_1^{\mathrm{uNACA63A418}} \quad (6)$$

and

$$v_1^{\mathrm{uTIP}} = -0.07 + x_1 * 0.14 + v_1^{\mathrm{uNACA63A412}}, \quad \text{where } x_1 \in [0,1], \quad (7)$$

with variables two through ten defined in the same way. With the airfoil sections thus described (recall that the spanwise variation of the polynomial coefficients is linear) and the volume and the aspect ratio fixed, the ten design variables unequivocally define the wing. Let us now consider the physics-based analysis of candidate wing designs.

[2] Note also that this formulation assumes a sharp trailing edge – a third term could be added if a finite thickness trailing edge was needed.

5.2 Notes on the Flow Analysis

In order to compute the cruise drag associated with instances of the parametric geometry described above, we employ an inviscid, full-potential, three dimensional flow solver developed by the Engineering Sciences Data Unit ESDU (2002).

The code generates meshes via a conformal mapping scheme. The flow equations are solved over this mesh through a finite difference algorithm including a three-level multi-grid scheme. The finest grid, corresponding to the original mesh of up to 115,200 cells is employed at the final stage of the computation with medium and coarse grids, of 14,400 and 7,200 cells respectively, employed in the preceding stages. Full convergence can typically be achieved in around 800 iterations, with 200 iterations using the coarse grid and 100 using the medium grid. A post-processor evaluates both the trailing vortex and wave drag components of a wing's inviscid drag coefficient. Trailing vortex drag is calculated using a method based on linearised theory and therefore ignores the effects of rolling-up and downward deflection of the trailing vortex sheet – this is "Model A" of Ashill and Fulker (1987). Wave drag is calculated via both the "first-order" and the "improved" methods of Lock (1985).

As it stands the FP wing analysis package will only predict vortex and wave drag. However, ESDU recently published a method for the prediction of the viscous drag coefficient for a wing in shock-free and attached flow. This can then be combined with the inviscid drag components from FP to obtain a prediction of the total drag for a wing.

The viscous drag coefficient is estimated assuming fully attached, shock-free flow, through another ESDU scheme [ESDU (2008)], whereby the wing's minimum profile drag coefficient, C_{DPmin0}, increment in profile drag due to twist, $(\Delta C_{DPmin})_{\varepsilon}$, lift-dependent viscous drag factor, K_{visc} and lift coefficient for minimum viscous drag, C_{Lmin} are estimated. These coefficients are then used to calculate the viscous drag coefficient of a wing via

$$C_{Dvisc} = C_{DPmin0} + (\Delta C_{DPmin})_{\varepsilon} + K_{visc}(C_L - C_{Lmin})^2, \qquad (8)$$

where C_L is the lift coefficient of the wing.

The method has been validated against experimental drag polars for a range of wings [ESDU (2008); Toal (2009)] and has been demonstrated to be of suitable accuracy (for instance, it predicts the minimum drag coefficient of the Brebner wing [Brebner and Wyatt (1961)] to within 9 counts).

5.3 Wing Design Search

We have indicated earlier that a one to two split in the sampling plan/update sequence division of computational effort appears to work well across a spectrum of objective landscape modalities. In that spirit we allocated 30 of a total budget of 90 runs of the full potential analysis code to a Morris-Mitchell-optimal Latin hypercube sampling plan, with the remaining 60 allocated to a

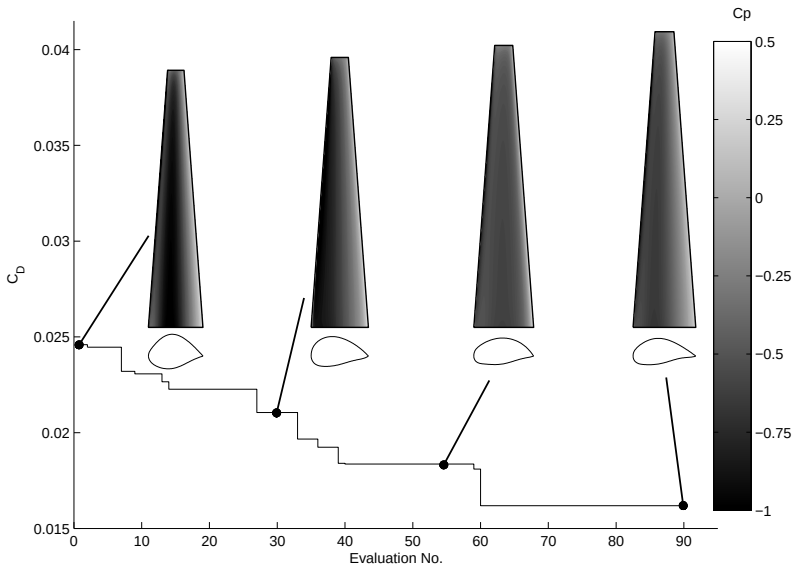

Fig. 8 Expected improvement-based optimization history. The airfoils are not shown to scale – their vertical coordinates are exaggerated to highlight shape variations.

an expected improvement-based design exploration sequence, as per the formulation described in Section 4.1, using a genetic algorithm / simplex local search combination to find the global optimum of the expected improvement landscape for the siting of the next infill point. The resulting optimization history, in terms of the current lowest drag (for 10t of lift), is depicted by Figure 8, which also highlights some of the representative designs obtained along the way.

As shown by the pressure coefficient (C_P) contours of these representative wings, the trend is towards larger (both in terms of span and chord), more lightly and more uniformly loaded wings, with the airfoils getting progressively thinner in the course of the search. Through the optimization procedure the drag coefficient C_D drops considerably, from an initial value of 0.026 to an optimum of 0.016.

We have indicated earlier that when a realistic goal is known in advance, *goal seeking* can be considered as an alternative to expected improvement updating of the kriging model. Figure 9 depicts the results of a goal seeking procedure performed on the same wing design problem (and same initial sampling plan as before), targeting a 40% cut on the total drag. As the comparative histories indicate, goal seeking does, in fact, produce a better result; moreover, it finds this more quickly.

Such evidence, of course, is insufficient to draw far-reaching conclusions on the comparative performances of expected improvement-based update cycles and goal seeking. It is, nevertheless, an indication that the latter is a route

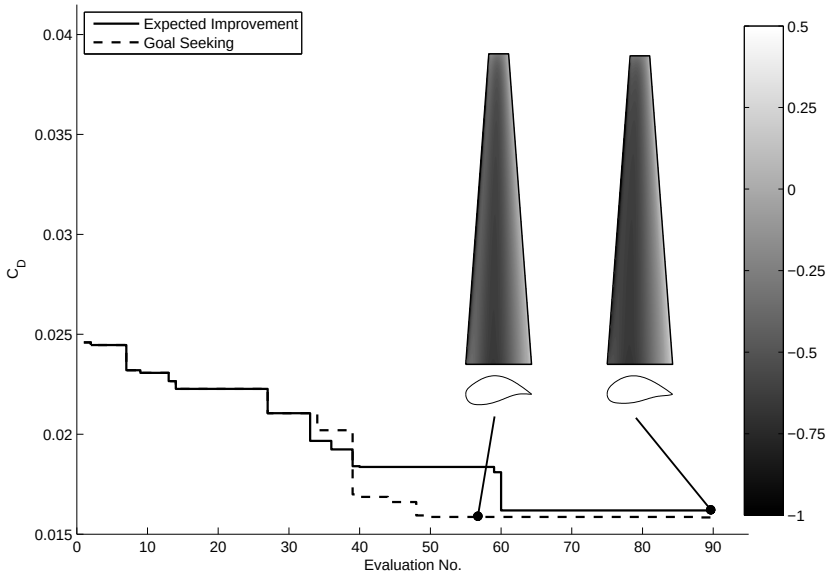

Fig. 9 Expected improvement and goal seeking optimization histories compared. With a target improvement of 40%, goal seeking finds a better optimum in fewer updates.

worth considering when a goal is available. We next move on to a multi-objective problem, the optimal design of a low environmental impact house.

6 Minimizing Cost and Environmental Impact – a Case of Multi-objective Trade-offs

Amidst increasing concerns about climate change, and a push to reduce global CO_2 emissions, the design of more efficient buildings has become increasingly important. In the UK, the energy used in buildings accounts for nearly half of total energy use [Communities and Government (2009)] and consequently a large proportion of total CO_2 emissions.

Substantial improvements in the energy efficiency of buildings are available, but the design of low-energy buildings is a complex optimization problem, with many variables interacting with each other in difficult-to-predict and non-linear ways [Wright et al (2002); Coley and Schukat (2002)]. In addition, energy-efficiency objectives often conflict with cost-minimization objectives [Wang et al (2005)], and as such hinder the widespread adoption of low-energy building designs.

This case-study focused on minimizing the CO_2 emissions resulting from the energy used to heat, light and cool a completed building. Other energy uses in buildings (hot water, appliance use, cooking) are less affected by the design of the building and so were omitted. The CO_2 emissions resulting from the construction of the building were also omitted.

6.1 Building energy demands

The energy used to heat, light and cool a building of a given shape and size is affected by many variables related to the design and specification of the building. The effect of each of these design variables will depend on both the use for which the building is intended, the geographical situation of the building, and also the values of other variables. A summary of how these design variables might be expected to affect energy demand in a building is presented below.

- Glazing - Additional glazing will reduce the need for artificial lighting up to a certain percentage of glazing, above which additional glazing will make little difference to artificial lighting needs. Higher levels of glazing will increase heat losses from the building if they are replacing well insulated walls, but will also increase heat gains if placed facing the sun, leading to opportunities for passive solar heating in winter, but also to overheating risks in summer.
- Thermal mass - In some climates internal thermal mass can reduce heating energy demands by storing heat won through solar gains if used in conjunction with large southerly windows (in the northern hemisphere), but otherwise may have little effect on heating energy use. Internal thermal mass can also be used to reduce overheating problems in some climates.
- Window thermal resistance - Improving the thermal resistance of the glazing will decrease heat losses but in some instances will also decrease solar gains and light transmission.
- Building air-tightness and ventilation - Decreasing air infiltration and ventilation will tend to reduce the energy required to heat and cool a building.
- Insulation - Increasing insulation in the building envelope will tend to decrease the need for heating and cooling the building.

The relative importance of each of the above design variables will vary according to the latitude of the building, the local weather, the position of the building relative to other buildings and obstacles, the efficiency of heating, lighting and cooling appliances and the way in which the building is used. Because of this, and because adjusting several of the above design variables involves accepting a trade-off between a desired effect and an unintended and contradictory effect (e.g. increasing window thermal resistance reduces heat losses but also reduces solar heat gains), specifying the above design variables at or near to optimum combinations is a difficult task.

6.2 Parametric house

The building design used for this case study is relatively simple, with a single 8m by 8m by 3.5m room oriented so that each wall faced due north, south, east or west, and with a flat roof. The heating, lighting ventilation and cooling use parameters and schedules for the building were set to mimic a house being

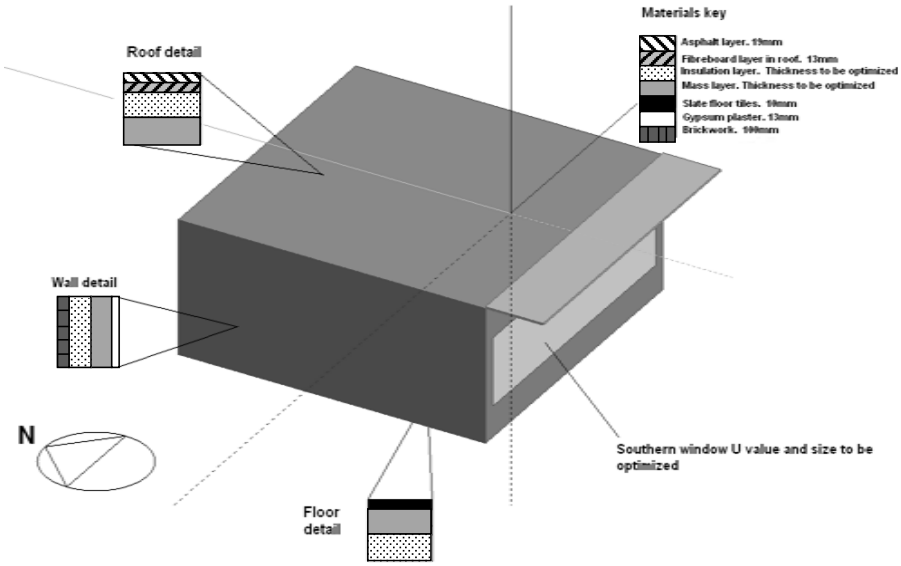

Fig. 10 The building model used in Energy+, showing the basic building shape, a 1.5 m long shade over the southern window, and with construction elements shown schematically according to the key.

lived in by a professional couple, with efficient lighting, heating and cooling. The location of the building was set as London, UK. Assumptions were made about the efficiency of plant and the CO_2 emissions associated with gas and electricity. The air infiltration rate was fixed.

The basic building model is shown in Figure 10 and the building variables to be optimized were as follows:

- Southern wall glazing extent (to vary between 0% and 100% glazed),
- Southern glazing thermal resistance (U value to vary between 0.8 and 1.8 $W \cdot K^{-1} \cdot m^{-2}$),
- insulation thickness on each aspect and on the roof and floor (6 variables, to vary between 0 and 1 m thickness), and
- internal thermal mass thickness on the floor and north wall (2 variables, to vary between 0 and 1 m thickness).

6.3 Notes on the CO_2 and Cost analysis

A simple mathematical cost model was constructed based on industry-supplied information on the costs of different standards of glazing, insulation (XPS polystyrene), thermal mass (concrete blocks) and other material elements of the wall construction. The simulation of the building's energy demands was made using Energy+, a well established and tested dynamic thermal simulation engine [USDE (2010); Crawley et al (2001); Neymark et al (2002)].

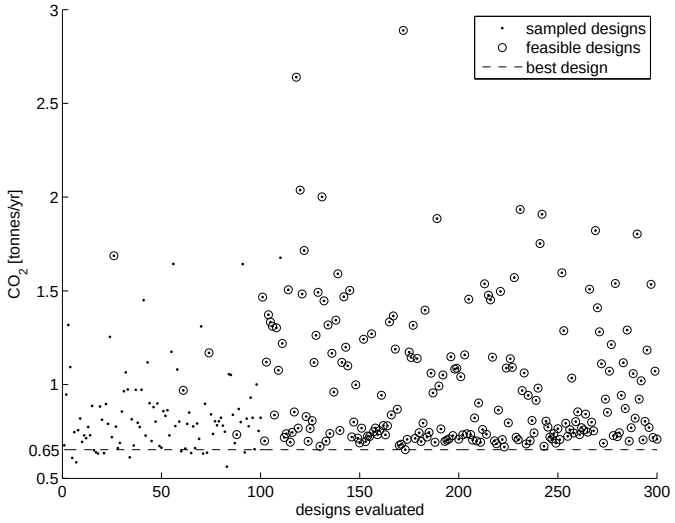

Fig. 11 CO_2 values for designs simulated in the $\max_{\mathbf{x}} E[I(\mathbf{x})]P[g(\mathbf{x}) < c]$ infill criterion based optimization.

This tool couples local weather files with a user-defined model of the building construction and use, in order to calculate the energy being used to keep the building within user defined ranges of temperature and light levels. These calculations were made at 15minute intervals over the course of a whole year to give annual demand for heating, cooling and lighting. These demands were then used to generate figures for CO_2 emissions from each source using a simple mathematical model taking into account the plant efficiency and CO_2 content of different fuels.

6.4 Building design: constrained optimization

Our first optimization case assumes a maximum budget for insulation, thermal mass, other wall materials and windows of £25,000. For this cost limit we wish to minimize CO_2 emissions, as calculated by Energy+. With a budget of 300 simulations we choose 100 initial sample points using, as for the wing optimization, a Morris-Mitchell-optimal Latin hypercube. We follow this with 200 infill points found by maxmizing $E[I(\mathbf{x})]P[g(\mathbf{x}) < c]$ (see equation 3), where $c = 25000$. CO_2 values for designs simulated are shown in Figure 11, where circled points represent designs which meet the £25,000 constraint.

Following the initial sample, only one out of 200 designs failed to meet the constraint. The optimum design was found after only the 73$^{\text{rd}}$ infill point, indicating that the number of simulations used here is well in excess of that required to solve this problem. However, the large number of designs around the same performance and price point do, in fact, have a good deal of variation in

the design variables themselves. This is a particularly interesting result from a design viewpoint, in that it indicates a plurality of solutions will be possible as opposed to fixing rigid limits on permissible values of building variables. Given that regulations for limiting carbon emissions from buildings are becoming increasingly onerous, and that the search space for optimal solutions is so large, it appears that this technique has the potential to offer the building designer useful information. It is though likely to be preferable to display a CO_2 and cost tradeoff to the designer by considering the two as distinct objectives.

6.5 Building design: multi-objective optimization

Using on the same initial sampling plan, we now choose our 200 infill points based on maximizing the dual objective probability of improvement (see equation 5). The results of this multi-objective optimization (Figure 12) show that relatively large, and low cost, CO_2 improvements could be made to the building design between the costs of £5,000 and £15,000, and that from this point improvements in CO_2 performance were much more costly. Note that, even though we are optimizing across two objectives and using the same simulation budget, the Pareto front passes through the best design found in the constrained optimization ($CO_2 = 0.65t/yr$). The general trends in building designs as we move up the ranked list of non-dominated designs are shown in Figure 13.

The highest CO_2 emission non-dominated designs had very low levels of insulation on all aspects, low levels of thermal mass assigned to the floor and north wall and large southern windows with low thermal resistance (high U value). To reduce CO_2 emissions from this point, non-dominated designs involved decreasing southerly glazing extent, improving the thermal resistance of the windows, increasing insulation thicknesses on all aspects and increasing the total amount of thermal mass. Interestingly, the optimization does not tend towards what would traditionally be viewed as high-performance windows. Windows with a U value as low as 0.8 $W \cdot K^{-1} \cdot m^{-2}$ were available in the optimization, but instead the lowest CO_2 emission buildings show U values of around 1.62 $W \cdot K^{-1} \cdot m^{-2}$. It appears the methodology is optimizing the window U value to balance the conflicting objectives of increased solar gain offered by poorly insulated windows and the decreased heat losses offered by well insulated windows, and also to balance these performance variables with cost.

An extremely important first step in the design of 'green' housing is the ability to present the trade-off between cost and environmental impact in a form understandable to the expert and the layperson alike. We believe that a Pareto-type framework is ideal for this purpose and, while building this type of trade-off model can often be prohibitively expensive, the above analysis indicates, that a potentially affordable route is through a multi-objective expected improvement type scheme. In the interest of clarity we have presented

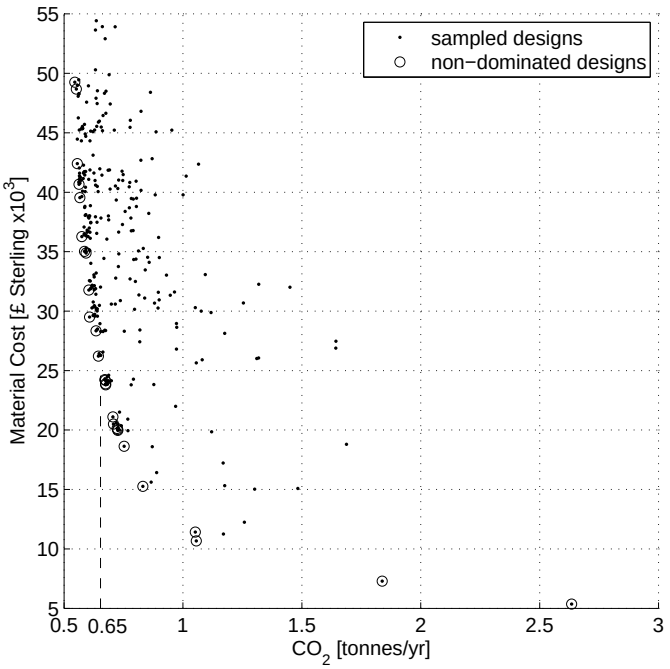

Fig. 12 CO_2/cost tradeoff produced by the dual objective probability of improvement infill criterion based search. The dashed line at $CO_2 = 0.65t/yr$ indicates the best design found for the £25,000 constrained optimization.

a relatively simple case study here, but the approach described here is scalable to more expensive objectives and larger numbers of design variables.

7 Conclusions

The high computational expense of measures of merit, constraints drawing awkward boundaries around design spaces and multiple, competing objectives are some of the key challenges of design optimization. Over these pages we have outlined a toolkit designed to tackle these problems through surrogate modeling and we deployed these tools on two problems that, we believe, are representative of modern, 'real-world' design problems. The techniques used here do not represent the only viable strategy for tackling expensive, black-box type search problems. They do, however, represent a distillation of the authors' combined experience in tackling this class of problems and we hope that the two case studies provide compelling evidence for the feasibility of surrogate model-based global searches as an alternative to more conventional, direct optimization methods. We also hope to have made a case for the generic nature of the approaches described here – the success of this effort will be measured by the reader considering to apply these methods to their own problems.

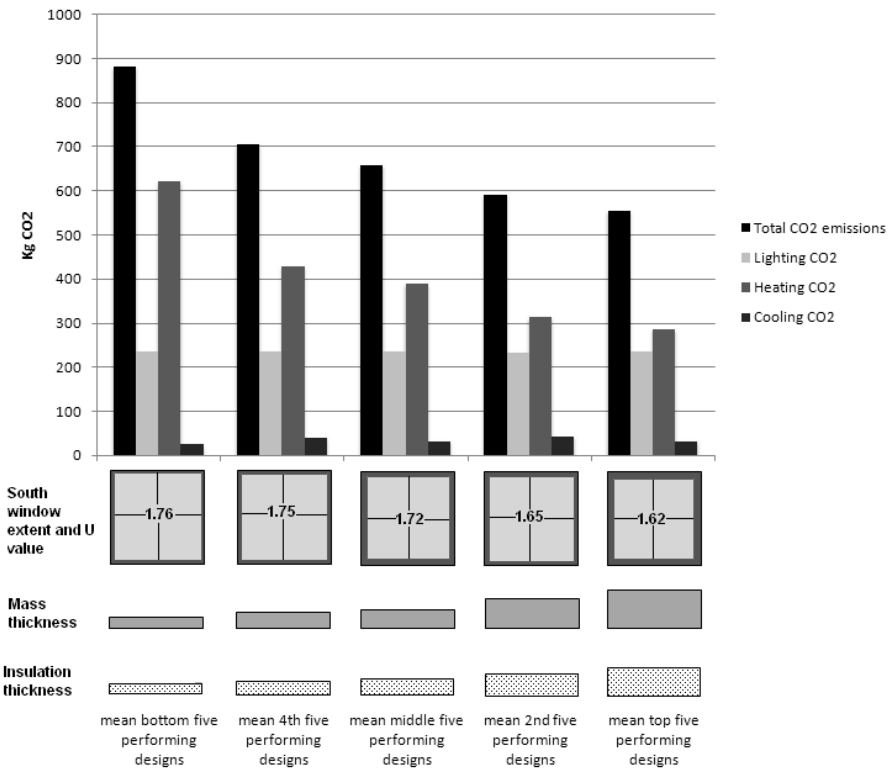

Fig. 13 Average design and CO_2 emissions for groups of five building designs, from the five worst CO_2 emission designs on the left, through to the lowest CO_2 emission designs on the right. The two highest CO_2 emission designs were omitted from the figure to allow a linear scale for CO_2 emissions to be used. The south window extent shows the size of the southern glazing area relative to the southern wall, and the U value is shown in $W \cdot K^{-1} \cdot m^{-2}$ as a figure in the middle of the window icon. Average thermal mass thickness for the north wall and floor is shown by the height of the grey boxes (relative thicknesses to scale, not absolute thicknesses). Average insulation thickness for all walls, the floor and the roof is shown in the same way by the yellow boxes.

Acknowledgements The first author's work has been supported by the Royal Academy of Engineering and the Engineering and Physical Sciences Research Council.

References

Ashill P, Fulker J (1987) Calculation of the Viscous and Vortex Drag Components of Wing/Body Configurations. RAE Technical Report TR 87028

Brebner G, Wyatt L (1961) Boundary layer measurements at low speed on two wings of 45 and 55 degrees sweep. Tech. Rep. ARC-CP-554, Aeronautical Research Council

Cherkassky V, Mulier F (1998) Learning from Data Concepts, Theory, and Methods. John Wiley and Sons

Chung H, Alonso J (2002) Using gradients to construct cokriging approximation models for high-dimensional design optimization problems. AIAA 40th Aerospace Sciences Meeting and Exhibit, AIAA-2002-0317, Reno, NV

Coley D, Schukat S (2002) Low-energy design: combining computer-based optimisation and human judgement. Building and Environment 37(12):1241–1247

Communities, Government L (2009) Improving the energy efficiency of our homes and buildings: Energy certificates and air-conditioning inspections for buildings: http://www.communities.gov.uk/documents/planningandbuilding/pdf/714826.pdf, last accessed December 2009

Crawley D, Lawrie L, Winkelmann F, Buhl W, Huang Y, Pedersen C, Strand R, Liesen R, Fisher D, Witte M, Glazer J (2001) Energyplus: creating a new-generation building energy simulation program. Energy and Buildings 33:319–331

Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: NSGA II. IEEE Transactions on Evolutionary Computation 6(2):182–197

ESDU (2002) Full-Potential (FP) Method for Three-Dimensional Wings and Wing-Body Combinations - Inviscid Flow Part 1: Principles and Results. ESDU-02013, London

ESDU (2008) Wing Viscous Drag Coefficient in Shock-Free Attached Flow. ESDU-07002, London

Forrester AIJ, Keane AJ (2008) Recent advances in surrogate-based optimization. Progress in Aerospace Sciences 45(1-3):50–79

Forrester AIJ, Sóbester A, Keane AJ (2007) Multi-fidelity optimization via surrogate modelling. Proc R Soc A 463(2088)):3251–3269

Forrester AIJ, Sóbester A, Keane AJ (2008) Engineering Design via Surrogate Modeling – a Practical Guide. Wiley-Blackwell

Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning. Springer-Verlag

Hicks RM, Henne PA (1978) Wing design by numerical optimization. Journal of Aircraft 15:407–412

Johnson ME, Moore LM, Ylvisaker D (1990) Minimax and maximin distance designs. Journal of Statistical Planning and Inference 26:131–148

Jones DR (2001) A taxonomy of global optimization methods based on response surfaces. Journal of Global Optimisation 21:345–383

Jones DR, Schonlau M, Welch WJ (1998) Efficient global optimization of expensive black-box functions. Journal of Global Optimization 13(4):455–492

Keane AJ (2006) Statistical improvement criteria for use in multiobjective design optimization. AIAA Journal 44(4):879–891

Kennedy MC, O'Hagan A (2000) Predicting the output from complex computer code when fast approximations are available. Biometrika 87(1):1–13

Krajewski WF (1987) Cokriging radar-rainfall and rain-gauge data. Journal of Geophycisal Research – Atmospheres 92(D8):9571–9580

Krige DG (1951) A statistical approach to some basic mine valuation problems on the witwatersrand. Journal of the Chemical, Metallurgical and Mining Engineering Society of South Africa 52(6):119–139

Kulfan BM (2008) Universal parametric geometry representation method. Journal of Aircraft 45(1):142–158, dOI: 10.2514/1.29958

Lock R (1985) Prediction of the Drag of Wings at Subsonic Speeds by Viscous/Inviscid Interaction Techniques. Paper 10, AGARD Report 723

Matheron G (1963) Principles of geostatistics. Economic Geology 58:1246–1266

McKay MD, Conover WJ, Beckman RJ (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21:239–245

Morris MD, Mitchell TJ (1995) Exploratory designs for computational experiments. Journal of Statistical Planning and Inference 43:381–402

Neymark J, Judkoff R, Knabe G, Le H, Rig M, Glass A, Zweifel G (2002) Applying the building energy simulation test (BESTEST) diagnostic method to verification of space conditioning equipment models used in whole-building energy simulation programs. Energy and Buildings 34:917–931

Samareh J (2001) Survey of shape parameterization techniques for high-fidelity multidisciplinary shape optimization. AIAA Journal 29(5):877–884

Schonlau M (1997) Computer experiments and global optimization. PhD thesis, University of Waterloo, Waterloo, Ontario, Canada

Simpson T, Mistree F (2001) Kriging models for global approximation in simulation-based multidisciplinary design optimization. AIAA Journal 39(12):2233–2241

Sóbester A (2009) Concise airfoil representation via case-based knowledge capture. AIAA Journal 47(5):1209–1218

Sóbester A, Leary SJ, Keane AJ (2003) On the design of optimization strategies based on global response surface approximation models. Journal of Global Optimization 33(1):31–59

Toal D (2009) Proper Orthogonal Decomposition and Kriging Strategies for Design. Ph.D Thesis

USDE (2010) Energy plus simulation tool, available to download at http://apps1.eere.energy.gov/buildings/energyplus/, last accessed March 2010

Wang W, Zmeureanu R, Rivard H (2005) Applying multi-objective genetic algorithms in green building design optimization. Building and Environment 40:1512–1525

Welch WJ, Buck RJ, Sacks J, Wynn HP, Mitchell TJ, Morris MD (1992) Screening, predicting, and computer experiments. Technometrics 34:15–25

Wright A, Loosemore H, Farmani R (2002) Optimization of building thermal design and control by multi-criterion genetic algorithm. Energy and Buildings 34:959–972