# OPTIMAM PARTEM ELEGIT[1]

**D. SPREVAK**.
Faculty of Informatics. University of Ulster.
N. Ireland. UK.
D.Sprevak@ulst.ac.uk

**R.S. FERGUSON**.
School of Electrical Engineering.
Queen's University Belfast-N. UK
R.Ferguson@qub.ac.uk

## ABSTRACT

This paper discusses some issues in numerical optimisation. It illustrates graphically the rationale behind some optimisation techniques. It shows the perils that await the unwary when extrapolating using functions whose parameters have been specified by choosing the values, which minimize a sum of squares of errors.

[1] Choose the better part. (Luke 10:42)

# Introduction

The wisdom of the command: ‘ choose the best part’, should be obvious to all. Optimisation is the branch of mathematics which deals with the techniques for locating the maximum or the minimum of a function, i.e. ‘ the best part’’.

There is the common misconception that to determine the location of the minimum of a function of several variables, $f(x_1, x_2, ..., x_n)$, one simply needs to solve the system of non-linear equations formed by setting to zero the partial derivatives of $f$:

$$F_i(x_1, x_2, ..., x_n) = \frac{\partial f(x_1, x_2, ..., x_n)}{\partial x_i} = 0, i = 1,2,...n$$

However, to solve such a system, usually, one needs to use a numerical procedure. Efficient numerical methods to do this are based on finding the minimum of

$$S = \sum_{i=1}^{n} F_i(x_1, x_2, ..., x_n)^2$$

Thus, numerical optimisation is required for solving systems of non-linear equations and not the other way around.

The computational methods for solving optimization problems are generally known as hill-climbing techniques that is because they mimic the strategy that a climber may use in trying to reach the summit of a mountain. Different strategies are open to the climber to reach the summit and we shall illustrate the rationale behind some of them.

Optimisation is frequently used to fit models to data with the intention of summarizing, interpolating or extrapolating from the observations. Extrapolation carries the implication that the estimated parameters are physically meaningful. However, it is very possible that parameters which produce a very good fit to the data lead to disastrously unsuitable extrapolations. Then, when is it safe to extrapolate? The paper discusses, through examples, the issues involved.

# Finding the best part

Let us consider the simplest strategy for locating the optimum of a non-linear function using a hill climbing technique. Consider that a climber is trying to reach the summit (maximisation) of a hill, or the bottom of the hill (minimisation), without a map and in dense fog. The climber can rely on an altimeter to measure altitude and a compass, which allows him to maintain a fixed direction. Measuring is time consuming, but movement itself is easy. The climber wishes to move as fast as possible. What is the best strategy?

It seems that the simplest approach would be to move along an arbitrary direction, such as the north-south line making regular measurements of the altitude until the highest point on the line is reached. Starting from this new point the same operation can be carried out along the east-west direction. This process of alternating searching along fixed directions ultimately will take the climber to the summit.

The algorithmic implementation of such simple procedure is known as the univariate search. To illustrate it we consider a problem presented by Box et al [1]. We wish to specify a function that relates the concentration $h$ of a chemical substance with time. The function is of the form:

$$h = \frac{b_1}{(b_1 - b_2)}(e^{-b_2 t} - e^{-b_1 t})$$

where, $b_1$ and $b_2$ are parameters which need to be estimated. Given a set of observed values for $h$ and $t$, a common procedure is to estimate the $b$ s by the method of least squares. That is: minimise the sum of the squared differences between the observed values and the predicted ones. That is, we want the location of the minimum of

$$f(x_1, x_2) = \sum_{i=1} (y_i - h(x_1, x_2))^2$$

where $x_1$ and $x_2$ stand for the possible values that we can, respectively, assign to $b_1$ and $b_2$; $y_i$ correspond to the observed concentration at time $t_i$. A set of observations is listed in Table 1. Let us consider finding values for the betas by minimizing $f$ using only the first six pairs of values of the data set.

**Table 1.** Observed concentration values $y_i$ at times $t_i$.

| $t_i$ | 0.0625 | 0.125 | 0.25 | 0.50 | 1.00 | 2.00 | 4.00 | 5.00 | 6.00 | 7.00 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | 0.01 | 0.02 | 0.08 | 0.15 | 0.22 | 0.51 | 0.48 | 0.29 | 0.20 | 0.12 |

The shape of the function $f$ is illustrated by its contours, shown in Figure 1(a). The picture also gives the path to the minimum using the univariate strategy. It is obvious from the graph that the path to the optimum requires a large number of short steps. However, the short steps could be used to define a general direction and a more efficient method would be to move along such a direction. The Davey, Swann and Campey (DSC) [2] algorithm does this. In contrast to the univariate search the DSC algorithm takes advantage of the accumulating information about the function. Starting at the point $x^{(0)}$ one cycle of the univariate search determines the point $x^{(1)}$. The next search is along the line joining $x^{(0)}$ and $x^{(1)}$ which determines the point $x^{(2)}$, and then we search at right angles to the previous search direction to determine $x^{(3)}$. The next search direction is along the line joining $x^{(2)}$ and $x^{(3)}$, and so on. Figure 1(b) shows the iterations using the DSC algorithm. In this case far fewer steps and function evaluations are required.

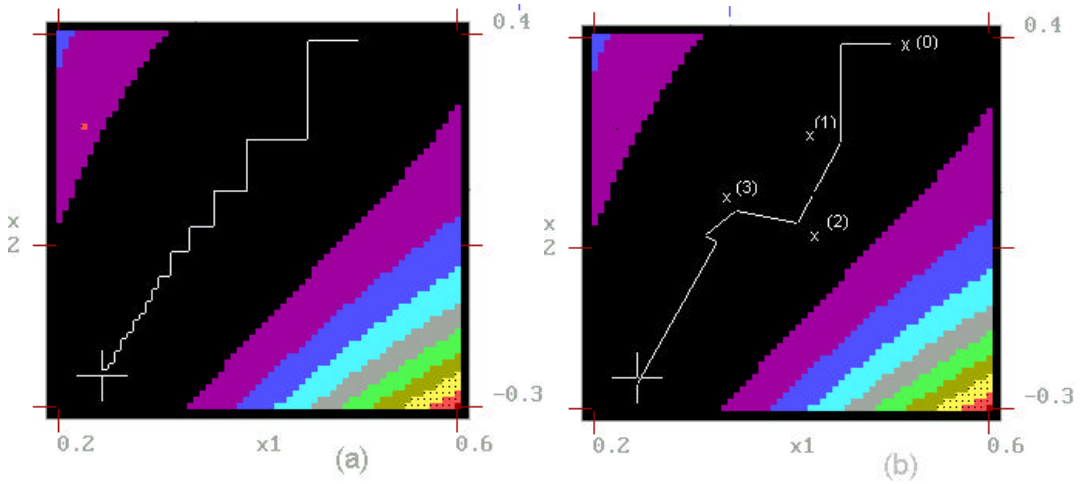


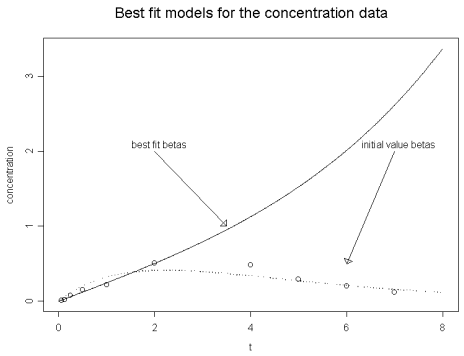

**Figure 1.** (a) The univariate search, locates the optimum, using 581 function evaluations, at (0.2442, -0.2402), with $f = 0.002$.

(b) The DSC algorithm uses 142 function evaluations to find the optimum at (0.2433, -0.2431). Both methods start at the point (0.5,0.39).

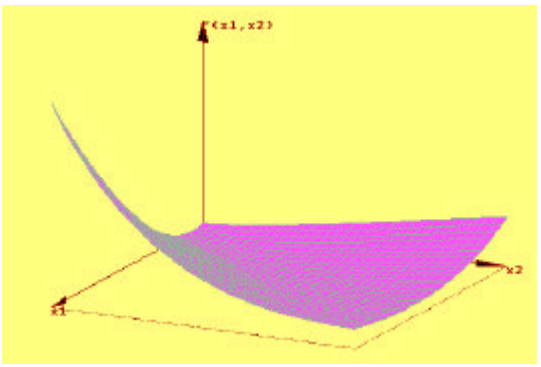
However, if our intrepid climber was allowed also to carry a spirit level, then he could use it to measure the lay of the land, and this extra information might lead him to choose his direction of search to be along the steepest descent. He might well find that such a strategy might produce a succession of large number of short steps similar to those of the univariate search. But being a smart climber he would realise that information about the gradients could be used, as in the DSC method, to determine a more efficient direction. This will lead him, no doubt, to discover the conjugate gradients method. Furthermore, having information about the gradients, he might consider gathering information about the curvature of the land, and using it might well develop Newton's type methods. It may well be that the terrain over which he is moving is very rocky - a noisy function - and therefore he may decide that he is much better off using the DSC strategy than the more elaborate methods which involve misleading gradient measurements.

All these strategies for numerical optimisation can readily be illustrated using graphs like those in Figure 1 and generalize to problems in more dimensions because the principles on which the methods are based are the same for two as for higher dimensions. The illustrations can easily be done using the software from McKeown et al. [2].

The function, specified with values for $b_1$ and $b_2$ which minimise $f$, fits the first six points of the data very well. There may be the temptation of assigning physical meaning to the estimated betas. However, when the rest of the observations are viewed, the fitted function is in complete disagreement with them. Any extrapolation using the fitted function, or a physical interpretation given to the parameters would have been unwise. On the other hand, it is simple to see that a set of values for $x_1$ and $x_2$ contained in the lowest contour of the sum of squares function are possible candidates for selection as values for the betas. For such a set there is not much change in the value of $f$. In particular, the pair of values at the start of the iteration fit the data almost as well as the ones that optimise $f$, and they happen to specify a function that gives reliable predictions for the extra data points. Figure 2 illustrates this.




(a)



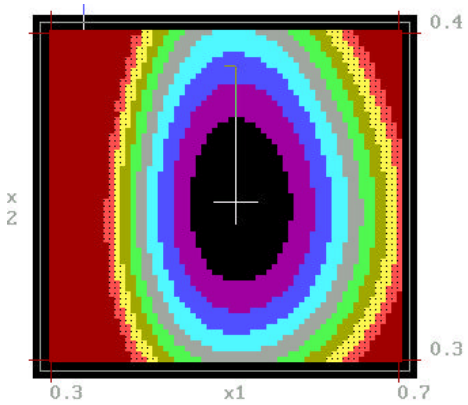

(b)

**Figure 2**. (a) Fitted function. (b) 3D Plot of $f(x_1, x_2)$.

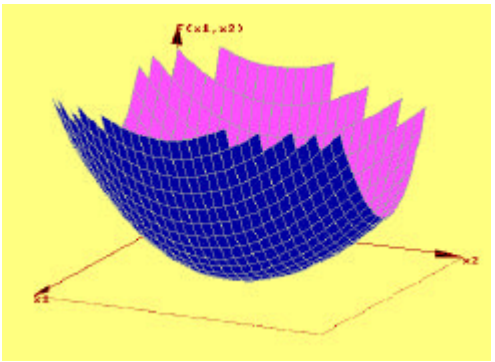
So, what is going on here?

The answer to the question lies in the fact that the function we are minimizing is insensitive to changes in $x_1$ and $x_2$. This is particularly visible in Figure 2 (b), which gives a 3D plot of $f$. The plot shows that $f$ is practically constant along the line joining the initial and optimal values of $x_1$ and $x_2$. Though we found a local minimum, its location is insensitive to changes along the ridge of $f$ shown in the figure. The problem is said to be ill-conditioned, and in such cases the fitted curve is only suitable for interpolation and no physical significance should be assigned to the estimated parameters. The data has forced us into a curve fitting problem and not a parameter extraction one.

By contrast when using the last six observed values to estimate the parameters we get the optimal values at $x_1 = 0.5153$, $x_2 = 0.3475$ and $f = 0.0363$. The contours of the new least squares function are given in Figure 3(a), they show that changes around the minimum lead to significant changes in $f$. The corresponding 3-D picture confirms that in this case there is no ill-conditioning.




(a)




(b)

**Figure 3.** (a) Contours of the sum of squares function for the last six data points. The steps of the univariate search are also illustrated from the starting point (0.5,0.39).

(b) The 3-D picture of the sum of squares function.

The plot of concentration against time in Figure 4 (a) shows that extrapolation is a lot less problematic when there is no ill-conditioning. Furthermore, a well-conditioned problem makes for a faster path to the optimum as illustrated in Figure 3 (a), showing the sequence of steps to the optimum when using the univariate search.

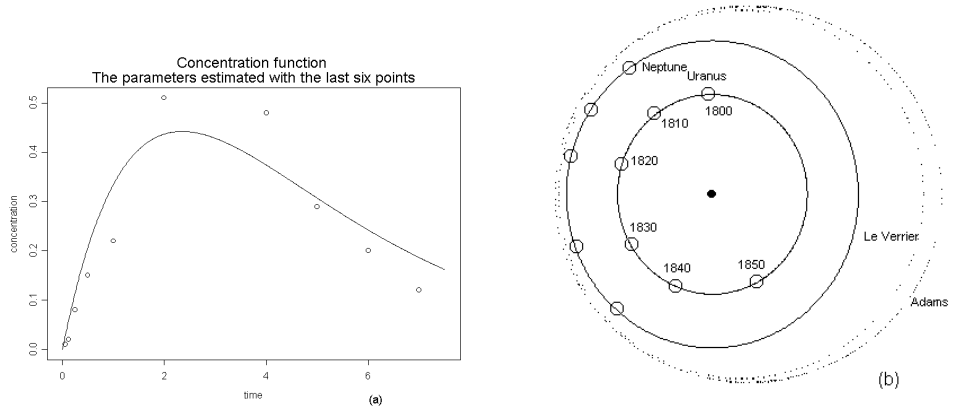



**Figure 4**. (a) Concentration against time fitted using the last six points.
(b) Orbits for Neptune - calculated and actual. The numbers on Uranus correspond to the year when its location was used to determine Neptune's orbit.

# A classical story of ill-conditioning

Recently the fascinating story of the discovery of the planet Neptune was published in a popularized form [3]. The story in the book contains a fair dose of human drama. It is exciting also because it is an example of a successful theoretical astronomical prediction. Using the discrepancies observed in the orbit of Uranus two mathematicians working independently, one French, Urbain Jean-Joseph Le Verrier, the other English, John Couch Adams, accounted for the discrepancies by predicting the existence of a new planet - Neptune–

These two mathematicians were breaking new ground. Newton's theory of gravitation had been used to calculate the effects of bodies on one another, but this was the first time that the theory was used to predict the position of a body from observations of the effects of its gravity on other bodies. However, not everyone was using the new planet explanation to try to account for the problems in Uranus' orbit. The Astronomer Royal George Airy supported the hypothesis that Newton's inverse square law did not apply over large distances. The perseverance of the two young mathematicians on the validity of their assumptions, against the pressures from a famous and established scientist are only part of the intricate drama that led to the discovery of Neptune. Their work not only helped in the discovery but it confirmed the universality of the gravitation law, and produced a model of work for the interaction between mathematicians and experimentalists.

Adams and Le Verrier were able to point out were in the sky to look for the planet. The astronomers duly found it in 1846. However, it is interesting that both mathematicians failed in determining with any accuracy the orbit of the planet for the region where there were no observations. Figure 4 (b) shows the theoretically proposed orbits and the actual one. Note that the maximum error in the predicted orbits is about half the radius of Uranus' orbit. This is

interesting to us, because it is an example of the consequences of ill-conditioning. To specify the orbit the mathematicians used the observations on the discrepancies observed in Uranus's orbit occurring during the first half of the 19th century. They were used to determine both, the position and the mass of Neptune. The mathematicians obtained a good fit to the data by overestimating the mass of the planet and the radius of the orbit. The errors compensated to give a fit acceptable in the region were the data was available but the calculated orbits were not suitable for extrapolation. The calculated orbits diverged more and more from Neptune's. Had the search for the planet taken place a few years earlier or later it would not have been found anywhere near the predicted location.

## Optimisation and mathematical education

Optimization is a decision-making problem: how to maximize or minimize the value of some quantity. In many cases this amounts to assigning values to certain quantities called the decision variables. We showed that optimization problems are common in science and engineering and that they usually cannot be solved by analytical methods and that computational methods must be used. There are two educational issues here, the first one is how to present a rationale for the numerical procedures for optimization. The second issue is to identify the applicability of the results of the optimization.

The analogy of 'hill-climbing' can be used as a powerful teaching tool to illuminate the ideas behind many of the numerical optimization methods. This is so because the algorithms for optimization can be illustrated with two-dimensional functions. We looked in particular at the idea behind the Davies Swann and Campey algorithm. From a simple description of the idea, the specification of the method – for any number of dimensions – seems a trivial generalization of the 'hill climbing' analogy. For example, we can state the DSC procedure for optimising a function of n variables as:

1 Set k = 1. Select an arbitrary starting point $\mathbf{x}^{(0)}$
2 Carry out one cycle of the univariate search algorithm to produce $\mathbf{x}^{(k)}$
3 Select $\mathbf{q} = \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}$ as a new search direction.
4 Generate n – 1 orthogonal directions and orthogonal to $\mathbf{q}$.
5 Search along $\mathbf{q}$ and each of the other n – 1 orthogonal directions to determine the new point $\mathbf{x}^{(k+1)}$. Each search begins at the end of the previous one.
6 If stopping criteria satisfied stop, else set k = k+1 and repeat from 3.

We used bold face to denote an n-dimensional vector. The algorithm above is a straightforward generalisation, to n dimensional functions, of the basic idea illustrated in Figure1 (b).

Further exploitation of the hill-climbing analogy might lead us to question the efficiency of obtaining exact determinations for the $\mathbf{x}^{(k)}$ s. It may be better not to find the optimum along a search direction but simply a better point from which to continue the search along a different direction. This policy may take more cycles, but overall, may require less use of the altimeter, and as changing direction involves no effort, a method with inexact line searches might be a more

efficient one. The educational possibilities when using sensible, imaginative ideas derived from the hill-climber analogy are boundless.

Optimisation is also taught as a procedure to fit equations to data. The objective, of course is to model a physical situation. However, the applicability of the fitted model is highly dependent on the conditioning of the problem. We illustrated that for two-dimensional problems ill conditioning implies a flatness, about the optimum, of the function we wish to optimize. Thus, the effect of ill-conditioning is to provide many possible, near optimal, but possibly dramatically different solutions. When this occurs, the only sensible use for the fitted model is for interpolation, which is not an unimportant outcome as the history of the location of Neptune testifies.

Though a mathematical treatment of ill-conditioning is an advance topic, the ideas and consequences of ill-conditioned problems can and should, as we have shown, be presented in more elementary courses in data analysis and optimization.

Finally, we feel that the teaching of numerical optimization should not be constrained by the use of 'analogies'. Their value is simply to provide another point of view, which might help to make the topic more interesting. We do not think that there is a unique solution to the teaching of the subject. It may well be that the problem of optimizing the teaching of mathematics is ill-conditioned, in the sense that there are many equally satisfactory solutions, and hence one should be careful to extrapolate from any of them.

# Concluding remarks

The analogy of hill-climbing has been shown to be useful for providing a motivation for numerical optimisation methods. The fundamental problem of using models, which are fitted to data, has been discussed. In particular we concentrated on the important distinction between data fitting and parameter extraction. We showed that when the problem is ill-conditioned, 'choosing the best part' can only be used for summarising the data and that no physical meaning should be associated to the parameters of the model. The discovery of the planet Neptune, during the middle of the 19th century, and the failure to specify its orbit was offered as an example of the effects of ill-conditioning. It would be an exciting project to investigate the conditioning of the problem using formal methods of analysis. There are, of course, such formal methods, McKeown and Sprevak [4] show how to use them in an application. It is not, however, the objective of this paper to deal with such formal methods but to offer a pictorial representation of ill-conditioning and of its consequences. We believe that everybody could profit by being aware that when fitting models to data, using optimization methods, the usefulness of the fitted model depends greatly on the conditioning of the problem. The moral of the lesson is: 'Optimam partem elegit', but be aware of its limitations.

**Acknowledgments** : We wish to thank Dr. J. J. McKeown for many interesting debates at the Thursday Seminars.

**Bibliography**

[1] Statistics for Experimenters. G.E.P.Box, W.G.Hunter and J.S.Hunter.
J.Wiley and Son, 1978.

[2] An Introduction to Unconstrained Optimization.
J.J.McKeown, D.Meegan and D.Sprevak. A Hilger, 1990.
[3] The Neptune File. T. Standage. Penguin Books, 2000.
[4] Parameter estimation versus curve fitting: new lamps for old.
J.J. McKeown, D. Sprevak. The Statistician, Vol. 41, 357-361, 1992.