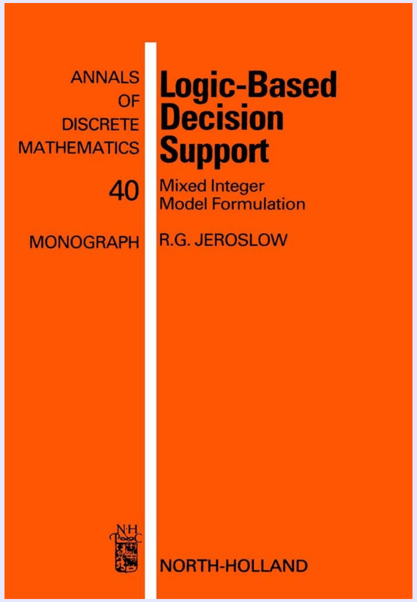
A Beautiful Paper

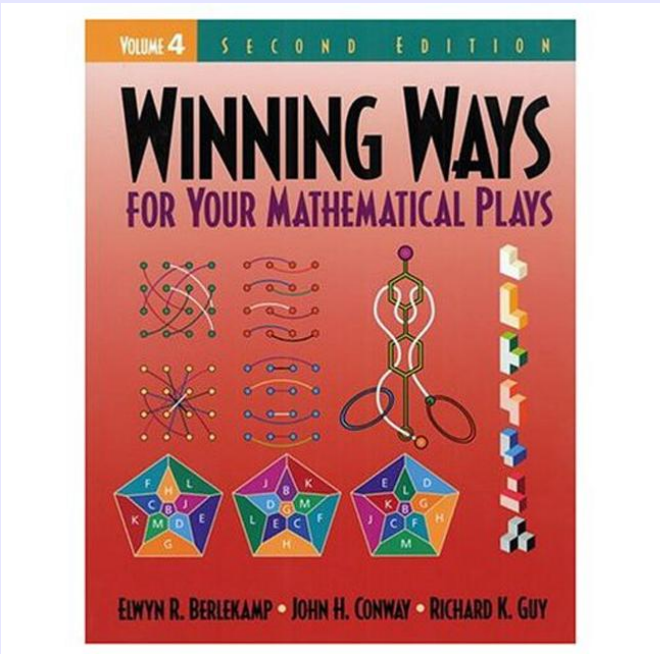
fabio.schoen@unifi.it
http://gol.dinfo.unifi.it

February 22, 2019

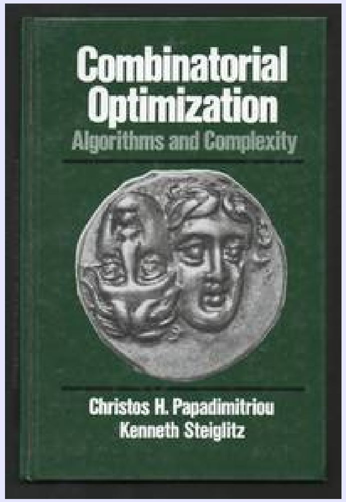
First attempts to find a topic: early passions



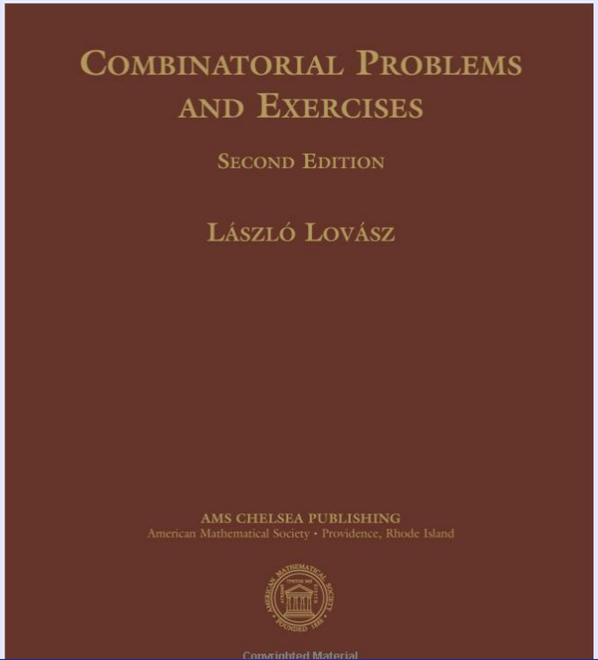
First attempts to find a topic: early passions



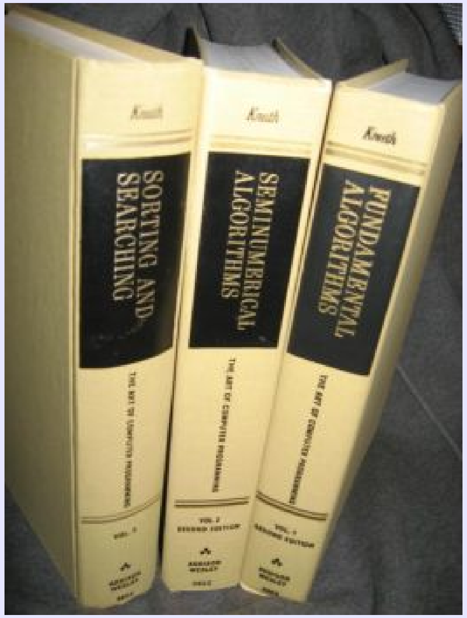
First attempts to find a topic: early passions



First attempts to find a topic: early passions



First attempts to find a topic: early passions



First attempts to find a topic: early passions

LINEAR AND NONLINEAR SEPARATION OF PATTERNS BY LINEAR PROGRAMMING

O. L. Mangasarian

Shell Development Company, Emeryville, California

(Received September, 1964)

A pattern separation problem is basically a problem of obtaining a criterion for distinguishing between the elements of two disjoint sets of patterns. The patterns are usually represented by points in a Euclidean space. One way to achieve separation is to construct a plane or a nonlinear surface such that one set of patterns lies on one side of the plane or the surface, and the other set of patterns on the other side. Recently, it has been shown that linear and ellipsoidal separation may be achieved by *nonlinear* programming. In this work it is shown that *both* linear and nonlinear separation may be achieved by *linear* programming.

A BASIC problem of pattern separation is this: Given two sets of patterns A and B, the set A consisting of m patterns, the set B of k patterns, where each pattern consists of n scalar observations, find a means of 'separating' the sets A and B, i.e., describing quantitatively whether a pattern belongs to the set A or the set B. An implementable and efficient solution of this problem is the key to the construction of pattern recognizing 'machines.' If the patterns are represented by points in an n-dimensional Euclidean space the separation problem then is to find a surface in this n-dimensional space such that all points representing

First attempts to find a topic: early passions

Mathematical Programming 39 (1987) 27–56 27
North-Holland

STOCHASTIC GLOBAL OPTIMIZATION METHODS PART I: CLUSTERING METHODS

A.H.G. RINNOOY KAN

Department of Industrial Engineering and Operations Research/Graduate School of Business Administration, University of California, Berkeley, CA, USA, and Econometric Institute, Erasmus University Rotterdam, The Netherlands

G.T. TIMMER

Econometric Institute, Erasmus University Rotterdam, The Netherlands, and ORTEC Consultants, Rotterdam, The Netherlands

Received 1 May 1985
Revised manuscript received 10 February 1987

In this stochastic approach to global optimization, clustering techniques are applied to identify local minima of a real valued objective function that are potentially global. Three different methods of this type are described; their accuracy and efficiency are analyzed in detail.

Then I found a picture (from the early 90's)

Big names in Global Optimization (GO) in the picture

- Panos Pardalos
- Reiner Horst
- Chris Floudas
- Manuel Bomze
- Marco Locatelli
- many many many others: Zelda, Gerardo, Eligius, Tibor, ...

and ...

Big names in GO in the picture

Don Jones 

who...? Donald Jones, General Motors

Document title	Authors	Year	Source	Cited by
Efficient Global Optimization of Expensive Black-Box Functions	Jones, D.R., Schonlau, M., Welch, W.J.	1998	Journal of Global Optimization 13(4), pp. 455-492	2631
View abstract ⌄ trova@unifi View at Publisher Related documents				
Lipschitzian optimization without the Lipschitz constant	Jones, D.R., Perttunen, C.D., Stuckman, B.E.	1993	Journal of Optimization Theory and Applications 79(1), pp. 157-181	1070
View abstract ⌄ trova@unifi View at Publisher Related documents				
A Taxonomy of Global Optimization Methods Based on Response Surfaces	Jones, D.R.	2001	Journal of Global Optimization 21(4), pp. 345-383	1011
View abstract ⌄ trova@unifi View at Publisher Related documents				

A beautiful GO paper

Journal of Global Optimization **13:** 455–492, 1998. 455

Efficient Global Optimization of Expensive Black-Box Functions

DONALD R. JONES[1], MATTHIAS SCHONLAU[2,*] and WILLIAM J. WELCH[3,**]
[1]*Operations Research Department, General Motors R&D Operations, Warren, MI, USA;* [2]*National Institute of Statistical Sciences, Research Triangle Park, NC, USA;* [3]*Department of Statistics and Actuarial Science and The Institute for Improvement in Quality and Productivity, University of Waterloo, Waterloo, Ontario, Canada*

(Accepted in final form 30 June 1998)

Abstract. In many engineering optimization problems, the number of function evaluations is severely limited by time or cost. These problems pose a special challenge to the field of global optimization, since existing methods often require more function evaluations than can be comfortably afforded. One way to address this challenge is to fit response surfaces to data collected by evaluating the objective and constraint functions at a few points. These surfaces can then be used for visualization, tradeoff analysis, and optimization. In this paper, we introduce the reader to a response surface methodology that is especially good at modeling the nonlinear, multimodal functions that often occur in engineering. We then show how these approximating functions can be used to construct an efficient global optimization algorithm with a credible stopping rule. The key to using response

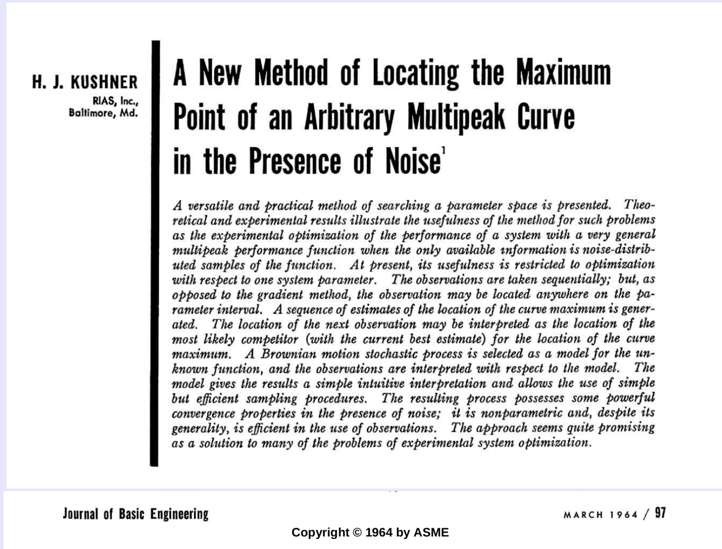
A beautiful GO paper

Why did I choose this paper? For many reasons:

- it is very well and very clearly written
- it received an impressive number of citations (not only from machine learning)
- it is a fascinating idea (although not his own and although subject to many improvements) - behind its complications the very essence of GO is contained: automatically mixing exploration and exploitation
- I did not mention Don Jones in my GO book (I just cited the origins of the method (see next slide)) and I took this opportunity to remedy. . .

Origin of the idea

Everything started and has its foundations in:



H. J. KUSHNER
RIAS, Inc.,
Baltimore, Md.

A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise[1]

A versatile and practical method of searching a parameter space is presented. Theoretical and experimental results illustrate the usefulness of the method for such problems as the experimental optimization of the performance of a system with a very general multipeak performance function when the only available information is noise-distributed samples of the function. At present, its usefulness is restricted to optimization with respect to one system parameter. The observations are taken sequentially; but, as opposed to the gradient method, the observation may be located anywhere on the parameter interval. A sequence of estimates of the location of the curve maximum is generated. The location of the next observation may be interpreted as the location of the most likely competitor (with the current best estimate) for the location of the curve maximum. A Brownian motion stochastic process is selected as a model for the unknown function, and the observations are interpreted with respect to the model. The model gives the results a simple intuitive interpretation and allows the use of simple but efficient sampling procedures. The resulting process possesses some powerful convergence properties in the presence of noise; it is nonparametric and, despite its generality, is efficient in the use of observations. The approach seems quite promising as a solution to many of the problems of experimental system optimization.

Journal of Basic Engineering MARCH 1964 / 97

Copyright © 1964 by ASME

and generalized by Jonas Mockus.

The problem

$$\min_{x \in S \subset \mathbb{R}^n} f(x)$$

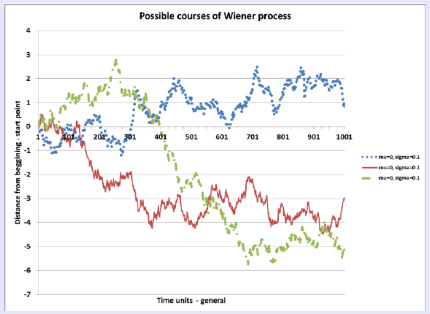
where

- we are interested in (an approximation of) the global optimum
- S is a "simple" feasible set (e.g., the unit box)
- evaluating f at any feasible x is extremely expensive

The model

Main idea: consider the objective function f as a realization of a stochastic process
Most frequently assumed model: a multidimensional Wiener process. In $\mathbb{R}^1$ a a stochastic process $W()$ is Wiener if:

- W has *independent increments*: for any t and $u \geq 0$ $W_{t+u} - W_t$ is independent of W_s for all $s < t$
- $W_{t+u} - W_t \sim \mathcal{N}(0, \sigma^2 u)$
- W is almost surely continuous



A stochastic model for the objective function

It is assumed that

$$f(x) = \mu + \varepsilon(x)$$

where:

- μ is an unknown constant
- $\varepsilon(x)$ is Gaussian with zero mean and variance σ^2
- $\mathrm{Corr}(\varepsilon(x^i), \varepsilon(x^j)) = \exp\left(-d(x^i, x^j)\right)$
- $d(x, y)$ is a generalized weighted Euclidean distance:

$$d(x, y) = \sum_{k=1}^{n} \theta_k |x_k^i - x_k^j|^{p_k}$$

where θ, p are parameters

Model estimation

Assume N observations are available $\{x^i, y^i\}_{i=1}^N$ with $y^i = f(x^i)$.
Let R be the Correlation matrix for these observations:
$R_{ij} = \mathrm{Corr}(x^i, x^j)$.
Then the likelihood turns out to be:

$$\mathscr{L}(\mu, \sigma^2, \theta, p) = \frac{1}{(2\pi\sigma^2)^{N/2}|R|^{1/2}} \exp\left(-(y - \mathbb{1}\mu)^T R^{-1}(y - \mathbb{1}\mu)/2\sigma^2\right)$$

and the maximum likelihood estimate of μ and σ^2 is:

$$\hat{\mu} = \mathbb{1}^T R^{-1} y / \mathbb{1}^T R^{-1} \mathbb{1}$$
$$\hat{\sigma}^2 = (y - \mathbb{1}\hat{\mu})^T R^{-1}(y - \mathbb{1}\hat{\mu})/N$$

while $\hat{\theta}, \hat{p}$ can be found maximining

$$\mathscr{L}(\hat{\mu}, \hat{\sigma}^2, \theta, p)$$

Prediction

Given x, where f has not been evaluated yet, the model can be used to predict $f(x)$.
Let r the correlation vector between x and the observations x^i.
Then the best unbiased estimator for $f(x)$ is

$$\hat{y}(x) = \hat{\mu} + r^T R^{-1}(y - \mathbb{1}\hat{\mu})$$
$$= \hat{\mu} + c^T r(x)$$

where

$$c = R^{-1}(y - \mathbb{1}\hat{\mu})$$
$$r_i(x) = \mathrm{Corr}(x, x^i)$$

while the variance of this estimate will be

$$s^2(x) = \hat{\sigma}^2 \left(\mathbb{1} - r^T R^{-1} r\right)$$

Prediction: Branin function

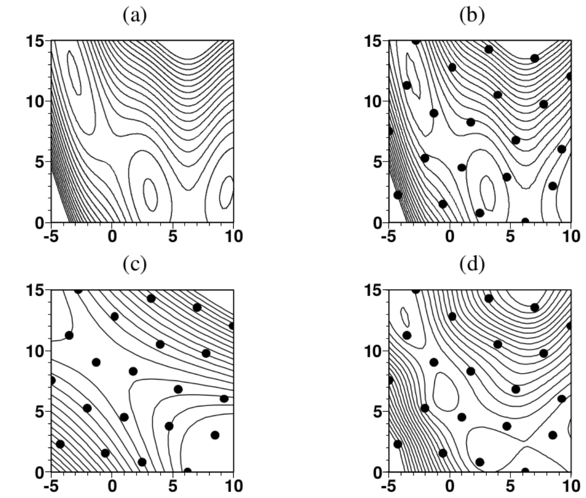


Figure 3. (a) Contours of the Branin test function; (b) contours of a DACE response surface based on the 21 sampled points shown as dots; (c) a quadratic surface fit to the 21 points; (d) a thin-plate spline fit to the 21 points.

Global Optimization

First idea: evaluate f at the global minimizer of the predictor.
Defects: greedy, stalling

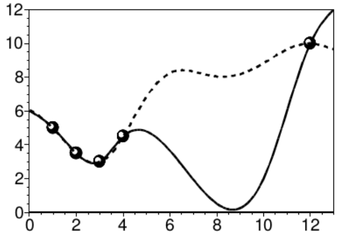


Figure 8. The solid line represents an objective function that has been sampled at the five points shown as dots. The dotted line is a DACE predictor fit to these points.

Global Optimization: variance

We should take variance into account

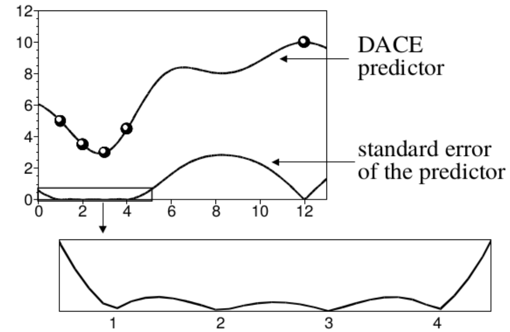
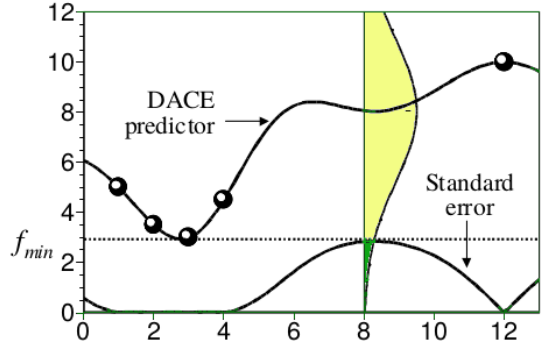


Figure 9. The DACE predictor and its standard error for a simple five-point data set.

Good choices should take into account good observations (exploitation) and unexplored regions (exploration)

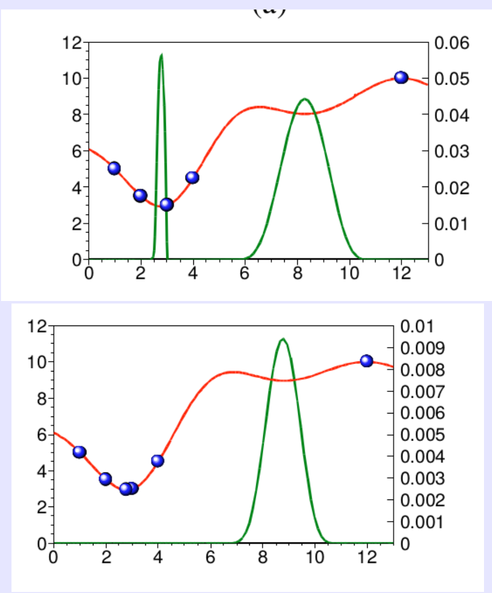
GO: expected improvement



$$E[I(x)] = E[\max\{0, f_{\min} - f(x)\}]$$
$$= (f_{\min} - \hat{y}(x))\Phi\left((f_{\min} - \hat{y}(x))/s\right) + s\phi\left((f_{\min} - \hat{y}(x))/s\right)$$

with Φ, ϕ: standard normal density and CDF.

Maximizing the expected improvement



A Beautiful Paper

fabio.schoen@unifi.it
http://gol.dinfo.unifi.it

February 22, 2019

Grazie Marco!