Global Optimization of Expensive Black-box Functions: a survey on model-based approaches

Fabio Schoen
coauthored by A. Cassioli

Global Optimization Laboratory “Gerardo Poggiali”
Dip. Sistemi e Informatica - Università di Firenze http://gol.dsi.unifi.it

Colloquia @ IASI - November 23, 2010

Global Optimization of expensive functions

$$\min_x f(x) \qquad x_j \in [\ell_j, u_j], j = 1, \ldots, n$$

Peculiarities

The analytical expression of the objective function is not available
Evaluating f is extremely expensive (hours or days of computation)

Examples: traffic simulation

Parameter calibration in traffic micro-simulation: let

- $p \in \mathbb{R}^n$: parameters of the simulation (e.g., parameters of drivers' behaviour, ...)
- $D \in \mathbb{R}^K$: data measured on a road network
- $O = O(p) \in \mathbb{R}^K$: data observed on a simulation run,

we wish to find

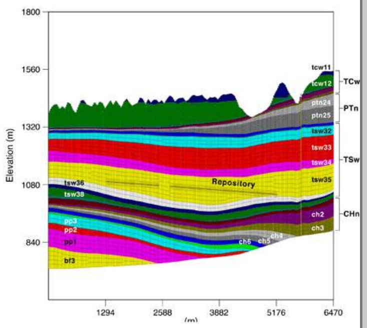
$$\min_p \|D - O(p)\|$$

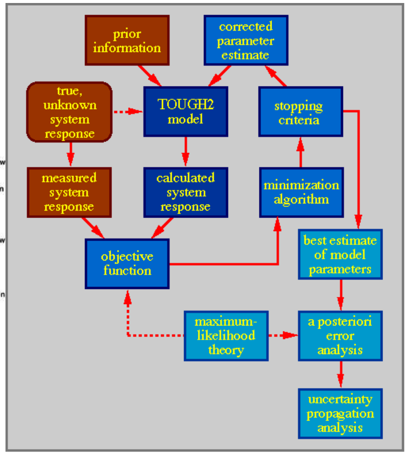
Examples of parameters:

- desired speed of vehicles
- acceleration and deceleration rates
- gap acceptance for road crossing
- gap acceptance for take over

Different usage:
Choose other parameters (e.g.: traffic light green phase durations) so that average queue lengths are minimized.

Example: inverse models (e.g.: Tough2)





Characteristics of black box optimization

- Objective function: extremely expensive
- No higher order information (e.g., gradients)
- No analytical expression
- Simple (lower and upper) bounds
- Multimodality
- Low dimension (usually less than 10 variables)
- possibly noisy

Surrogate based optimization models

The idea:

1. Choose an initial set of points $x^1, \dots, x^k$ and evaluate f
2. Build a surrogate model (interpolation or approximation) $s(x)$ of $f(x)$ based upon the current information $S^k = \{x^i, f(x^i)\}_{i=1}^k$
3. Choose the next evaluation point x^{k+1} through the surrogate model $s(x)$
4. Evaluate $f(x^{k+1})$ and update $k = k + 1$, $S^k = S^{k-1} \bigcup \{(x^k, f(x^k))\}$
5. Go to 2

Recipes

How to:

1. Choose an initial set of points $x^1, \dots, x^k$ through factorial design or through previously known function values (starting guesses)
2. Build a surrogate model $s(x)$ of $f(x)$ through Radial Basis function interpolation/regression
3. Choose the next evaluation point x^{k+1} through the optimization of a suitably defined *merit function* $\mathcal{M}(x)$

1. Choose an initial set of points $x^1, \ldots, x^k$
2. Build a surrogate model $s(x)$ of $f(x)$ through Radial Basis function interpolation/regression
3. Choose the next evaluation point x^{k+1} through the optimization of a suitably defined *merit function* $\mathcal{M}(x)$

Introduction to RBF interpolation

A *radial function* is defined as

$$s(x) = \sum_{j=1}^{h} \lambda_j \varphi(\|x - y^j\|)$$

where

- λ_j: coefficients
- $\varphi(\cdot)$: radial basis function
- y^j: *j–th* center of the radial function

Common choices for φ:

$$\begin{aligned} \text{cubic} &: \varphi(r) = r^3 \\ \text{thin plate spline} &: \varphi(r) = r^2 \log r \\ \text{gaussian} &: \varphi(r) = \exp -\gamma r^2 \end{aligned}$$

Existence of RBF interpolants

Given a sample of observed function values $S_k = \{x^i, f_i\}_{i=1}^k$ a solution to

$$s(x^i) = \sum_{j=1}^{h} \lambda_j \varphi(\|x^i - y^j\|) = f_i \qquad i = 1, \ldots, k$$

might not exists, even if we choose the centers y^j at the interpolation points. Example: if φ: thin plate spline, and $\{x^i\}$ are the vertices of a unit simplex, then $\varphi(\|x^i - x^j\|) = 0$ for all i, j.

Extension

In order to be able to guarantee interpolation of any set of scattered data, a polynomial is added to the RBF. Let $p_1, p_2, \ldots, p_{\hat{m}}$ be a basis for the space Π_m of polynomials of degree at most m.
Then we choose an interpolation of the form:

$$s(x) = \sum_{j=1}^{k} \lambda_j \varphi(\|x - y^j\|) + \sum_{\ell=1}^{\hat{m}} c_\ell p_\ell(x)$$

Let

$$\begin{aligned} \Phi_{ij} &= \varphi(\|x^i - y^j\|) \\ P_{i\ell} &= p_\ell(x^i) \end{aligned}$$

then the interpolation condition is

$$\Phi\lambda + Pc = f$$

Additional requirements

If the data can be fitted by a polynomial in Π_m, then this will be the unique interpolant found. I.e., if $\exists\, d$: $f = Pd$, then

$$\Phi\lambda + Pc = Pd$$

If we require that

$$Pc = 0 \Rightarrow c = 0$$
$$P^T\lambda = 0$$
$$\lambda \neq 0, P^T\lambda = 0 \Rightarrow \lambda^T\Phi\lambda > 0$$

then

$$\lambda^T\Phi\lambda + \lambda^T Pc = \lambda^T Pd \qquad \Rightarrow \lambda = 0$$
$$\Rightarrow Pc = Pd \qquad \Rightarrow c = d$$

Definition

A set of distinct points $x^1, \dots, x^k \in \mathbb{R}^n$ is Π_m–unisolvent if the unique polynomial $q \in \Pi_m$ (the space of polynomials of degree at most m) such that

$$q(x^j) = 0 \qquad \forall j$$

is the null polynomial.

Equivalently:

$$Pc = 0 \qquad \text{iff } c = 0$$

Example: for Π_1, a set is unisolvent iff it contains $n+1$ affinely independent points.

Conditional positiveness

A radial basis function φ is strictly conditional positive of order $m \geq 0$ if for every $x^1, \dots, x^k \in \mathbb{R}^n$ and every $\lambda \in \mathbb{R}^k$ with $\lambda \neq 0$, if

$$P^T\lambda = 0$$

then:

$$\lambda^T\Phi\lambda > 0$$

Existence and uniqeness

Let P be a basis for the polynomial space evaluated at sample points.

If φ is strictly positive definite of order $m \geq 0$ and $x^1, \dots, x^k$ is Π_{m-1}–unisolvent, then the linear system

$$\begin{bmatrix} \Phi & P \\ P^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ c \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}$$

admits a unique solution.

RBF interpolation

$$\begin{bmatrix} \Phi & P \\ P^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ c \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}$$

Strict positive definiteness is granted for gaussian RBF's without any need for an additional polynomial;
for thin plate spline RBF's, a linear polynomial is sufficient.

Connection with Natural Splines

In $\mathbb{R}^1$ it can be seen that an RBF based on a cubic radial basis, with the addition of a linear polynomial produces a natural cubic spline, i.e. an interpolation which is in $\mathscr{C}^2$ and with vanishing second derivative outside the interpolation interval.
This last condition is equivalent to $P^T\lambda = 0$
Natural cubic splines $s(\cdot)$ satisfy a *least curvature* property, i.e., they minimize

$$I(s) = \int_{-\infty}^{\infty} \left(s''(x)\right)^2 \, dx$$

among all interpolants.
It can be shown that:

$$I(s) = 12\,\lambda^T \Phi \lambda$$

Semi–norm

Let g and h be two RBF interpolants of the same points, but with different centers and possibly different polynomials:

$$g(x) = \sum_j \lambda_j \varphi(\|x - y^j\|) + p(x)$$
$$h(x) = \sum_i \mu_i \varphi(\|x - z^j\|) + q(x)$$

and define

$$\langle g, h \rangle = \sum_j \lambda_j h(x^j)$$

It can be shown that this is a semi–inner product, which induces a semi–norm.
Moreover

$$\langle g, g \rangle = (-1)^m \lambda^T \Phi \lambda$$

(m: order of the rbf base).

On the choice of centers

Under the assumptions which guarantee existence and uniqueness of the interpolant, the unique interpolant

$$g(x) = \sum_j \lambda_j \varphi(\|x - x^j\|) + \sum_\ell c_\ell p_\ell(x)$$

with

$$\begin{bmatrix} \Phi & P \\ P^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ c \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}$$

satisfies

$$\langle g, g \rangle \leq \langle h, h \rangle$$

for all interpolants h in the same family.

Connection with Natural Splines

In $\mathbb{R}^1$ an RBF based on a cubic radial basis, with the addition of a linear polynomial produces a natural cubic spline. Natural cubic splines $s(\cdot)$ satisfy a *least curvature* property, i.e., they minimize

$$I(s) = \int_{-\infty}^{\infty} \left(s''(x)\right)^2 \, dx$$

among all interpolants.
It can be shown that:

$$I(s) = 12\,\lambda^T \Phi \lambda$$

Thus there is a connection between natural splines and minimum semi–norm RBF interpolation. The measure $\lambda^T \Phi \lambda$ can be considered as a bumpiness[1] measure.

[1] Re:Gutmann, "A Radial Basis Function Method for Global Optimization" J.Glob.Opt., 19 (2001)

1. Choose an initial set of points $x^1, \ldots, x^k$
2. Build a surrogate model $s(x)$ of $f(x)$ through Radial Basis function interpolation/regression
3. Choose the next evaluation point x^{k+1} through the optimization of a suitably defined *merit function* $\mathcal{M}(x)$

Merit functions for RBF–based optimization

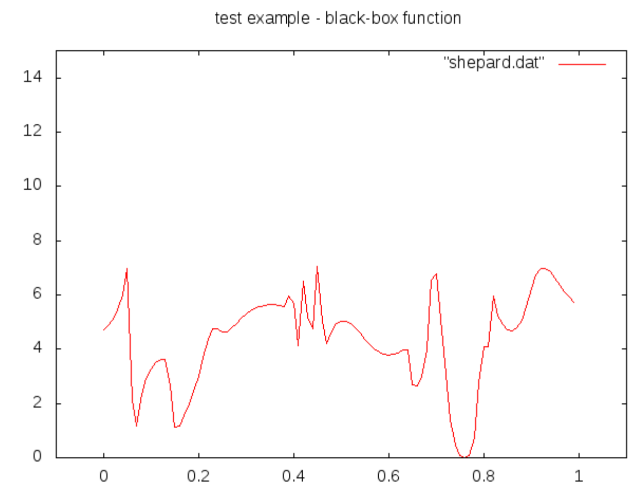
The simplest merit function (to be *minimized*):

$$s(x)$$

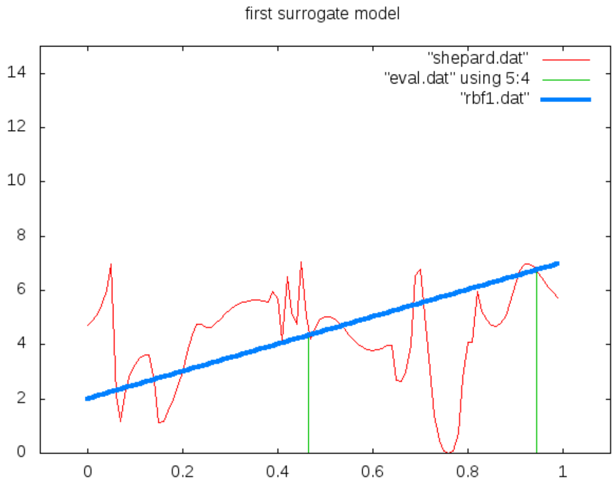
is the RBF interpolant itself.
Thus the new point at which $f(x)$ should be evaluated is the global minimizer of the interpolation $s(x)$, Defects:

- stalling (the same point found in different iterations)
- the model, based on few observations, is trusted as being correct $\Rightarrow$ the methods becomes too local

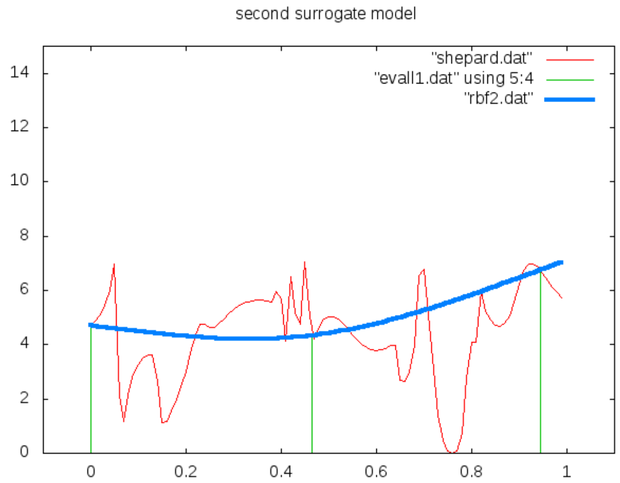
Example (random Shepard's function)



first interpolation



second interpolation



Methods based on an extended sample

Assume that a new point $\{\hat{x}\}$ is (symbolically) added to the sample $S = \{(x^j, f_j)\}_{j=1}^{k}$ and let $\hat{f}$ be (an estimate of) the objective function value at $\hat{x}$.
Assume a reasonable estimate of $\hat{f}$ is known or, otherwise, assume $\hat{f}$ to be an aspiration level.
Where is it "most likely" to find a point $\hat{x}$ at which the value $\hat{f}$ is attained?

Bumpiness as a merit function

Given and RBF interpolant, the quantity $\lambda^T \Phi \lambda$ is a bumpiness measure.
If a new observation is placed at $\bar{x}$ and it is expected that its value at $\hat{x}$ is $\hat{f}$, the new observation

$$x^{k+1} = \hat{x}$$

can be chosen so that the interpolation of f at $S^k = \{x^j, f(x^j)\} \bigcup \{\hat{x}, \hat{f}\}$ has minimum bumpiness.

Minimizing the bumpiness

$$\min_{\hat{x},\lambda} \lambda^T \Phi \lambda$$

$$\sum_{j=1}^{k+1} \lambda_j \varphi(\|x^i - x^j\|) + \sum_{\ell=1}^{\hat{m}} c_\ell p_\ell(x^i) = f_i \qquad \forall\, i$$

$$P^T \lambda = 0$$

where

$$x^{k+1} = \hat{x} \qquad f_{k+1} = \hat{f}$$

Minimizing the bumpiness

The problem can be shown to be equivalent to

$$\min_{\hat{x}} Bump(\hat{x}) \qquad \text{where}$$

$$Bump(\hat{x}) = (\hat{f} - s(\hat{x}))^2 g(\hat{x}) \qquad \text{and}$$

$$g(\hat{x}) = det \begin{bmatrix} \Phi & P \\ P^T & 0 \end{bmatrix} / \det \begin{bmatrix} \Phi & \phi_{k+1} & P \\ \phi_{k+1}^T & P^T & \pi_{k+1} \\ P^T & \pi_{k+1} & 0 \end{bmatrix}$$

where $\phi_{k+1}^T = [\phi(\|\hat{x} - x^1\|), \ldots, \phi(\|\hat{x} - x^k\|)]$ and π_{k+1} is the value of the polynomial basis evaluated at $\hat{x}$

Properties - 1

If the aspiraton level $\hat{f}$ is chosen in such a way that

$$\hat{f} < \min_x s(x|S_k)$$

then

$$\hat{x} \in \arg\min Bump(x) \Rightarrow \hat{x} \notin S_k$$

(i.e.: the new point is distinct from all points at which f has been evaluated)

Properties - 2

$$\lim_{x \to x^j} Bump(x) = +\infty$$

thus sample points have infinite bumpiness.

Properties - 3

If $\hat{f} \to -\infty$, then

$$\hat{x} \in \arg\min_{\hat{x}} g(x)$$

$$= \arg\min \det \begin{bmatrix} \Phi & P \\ P^T & 0 \end{bmatrix} / \det \begin{bmatrix} \Phi & \phi_{k+1} & P \\ \phi_{k+1}^T & P^T & \pi_{k+1} \\ P^T & \pi_{k+1} & 0 \end{bmatrix}$$

Choice of the aspiration level

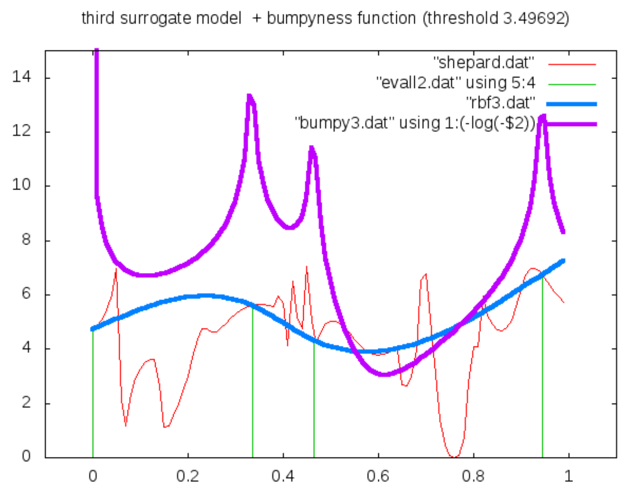
Many methods exists. Most of them alternate between

- $\hat{f} = s^\star - \varepsilon$ where $s^\star = \min s(x)$ (the interpolation is trusted)
- $\hat{f} = -\infty$ (a global exploration, no trust in the model)

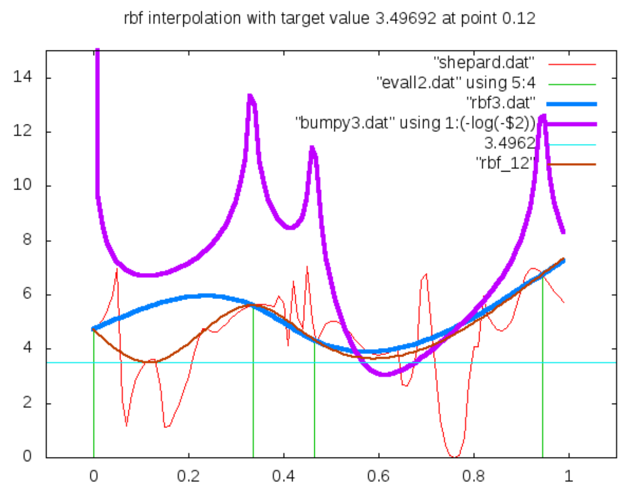
Other choices are possible (e.g., cyclic choice of values lower than $\hat{f}$)

Convergence property: convergence to the global optimum is guaranteed provided that the aspiration level $-\infty$ is chosen infinitely often (actually it is enough that $\hat{f}$ is sufficiently smaller than $s^\star$)

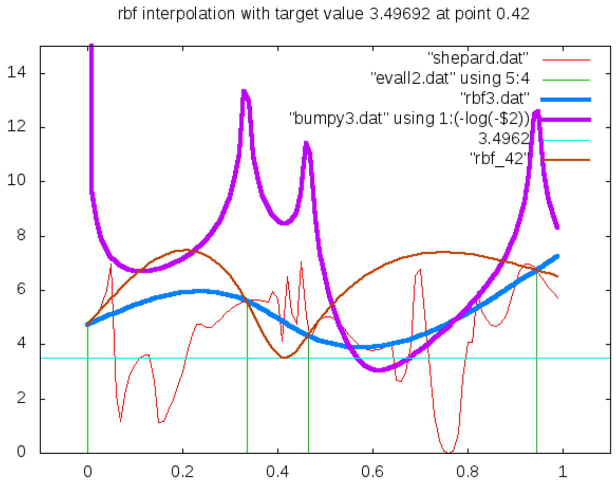
third interpolation



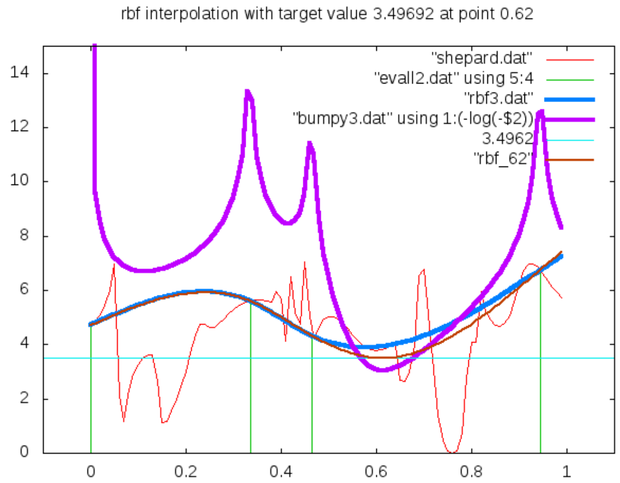
Bumpyness: target at $\hat{x} = 0.12$



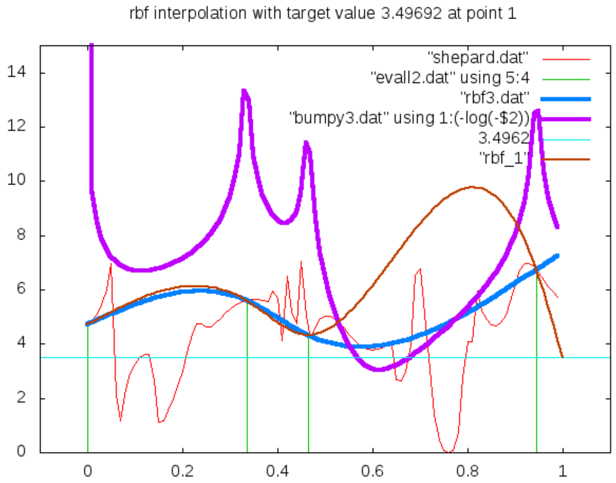
Bumpyness: target at $\hat{x} = 0.42$



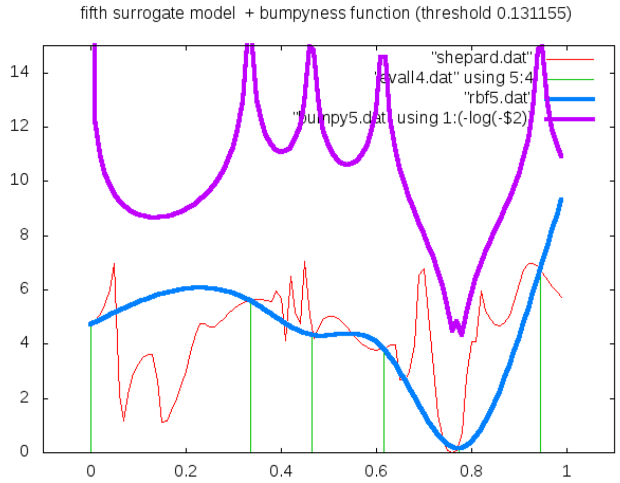
Bumpyness: target at $\hat{x} = 0.62$



Bumpyness: target at $\hat{x} = 1.0$



fourth interpolation



Extension: incorporating information in the model

Assume a lower bound $\underline{f}$ of f is known.
Any interpolation which attains a value lower than $\underline{f}$ has to be refused.
Model:

$$\begin{aligned} s(x^j) &= f_j & j &= 1, k \\ s(x) &\geq \underline{f} & &\forall x \end{aligned}$$

At each step of the algorithm the minimum of the interpolant $s^\star = \min s(x)$ is (approximately) found (by means of global optimization). Let $x_s^\star$ the global minimizer. Model: add a constraint

$$s(x_s^\star) \geq \underline{f}$$

to the interpolation rules.

Interpolation model with lower bounding

$$\begin{aligned} &\min \eta \\ &\sum_{j=1}^{k} \lambda_j \varphi(\|x^i - x^j\|) + \sum_{\ell=1}^{\hat{m}} c_\ell p_\ell(x^i) = f_i + \varepsilon_i & &\forall i \\ &\lambda^T P = 0 \\ &\sum_{j=1}^{k} \lambda_j \varphi(\|x_s^\star - x^j\|) + \sum_{\ell=1}^{\hat{m}} c_\ell p_\ell(x_s^\star) \geq \underline{f} \\ &\varepsilon_i \leq \eta & &\forall i \end{aligned}$$

(a linear program)with value 0

Next step...

$$\begin{aligned} &\min \lambda^T \Phi \lambda \\ &\sum_{j=1}^{k} \lambda_j \varphi(\|x^i - x^j\|) + \sum_{\ell=1}^{\hat{m}} c_\ell p_\ell(x^i) = f_i & &\forall i \\ &\lambda^T P = 0 \\ &\sum_{j=1}^{k} \lambda_j \varphi(\|x_s^\star - x^j\|) + \sum_{\ell=1}^{\hat{m}} c_\ell p_\ell(x_s^\star) \geq \underline{f} \end{aligned}$$

(a convex quadratic program)

Optimal choice

It can be proven that, given a pre-chosen interpolation point $x_s^\star$, bumpiness is a convex quadratic function in f_s (function value at $x_s^\star$) and, given the constraint

$$f_s \geq \underline{f}$$

the optimal (minimum bumpiness) interpolant is obtained by choosing

$$f_s = \max\{\min_x s(x|S_k), \underline{f}\}$$

Thus, including a lower bound for the interpolant at specific points (without observing the true objective function) can be easily accomplished simply by adding a new interpolation point to the RBF (to be removed in later stages).

A new globopt algorithm

1. Choose an initial set of points (forming an unisolvent set) $x^1, \ldots, x^k$ and evaluate f; let $S^k := \{x^i, f(x^i)\}_{i=1}^k$
2. Build a surrogate model $s(x|S^k) = \sum_j \lambda_j \varphi(\|x - x^j\|) + p(x)$ of $f(x)$ solving the linear system
$$\begin{bmatrix} \Phi & P \\ P^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ c \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}$$
3. find the global minimizer $s^\star$ of $s(x)$ at point $x_s^\star$
4. if $s^\star < \underline{f}$ then add the constraint $s(x_s^\star) >= \underline{f}$ to the model and go to 3
5. Otherwise: choose the next evaluation point x^{k+1} through the current surrogate model $s(x)$:
    1. Choose an aspiration level $\hat{f}$
    2. Let x^{k+1} be a global minimizer of the Bumpiness function
6. Evaluate $f(x^{k+1})$ and update $k = k + 1$, $S^k = S^{k-1}; \bigcup\{(x^k, f(x^k))\}$; remove all added points $x_s^\star$
7. Go to 2

Test environment

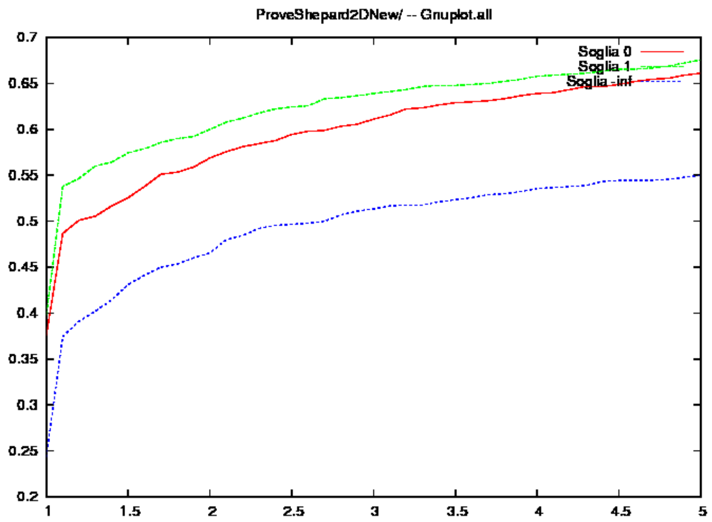
Test functions: "Shepard" with pre–chosen number of stationary points and range and uniform random location of stat points.
100 test functions for each combination of

range :	in $[0, 10], [0, 100], [0, 1000]$
number of stationary points	in 20, 50, 1000
dimension	in 2, 5

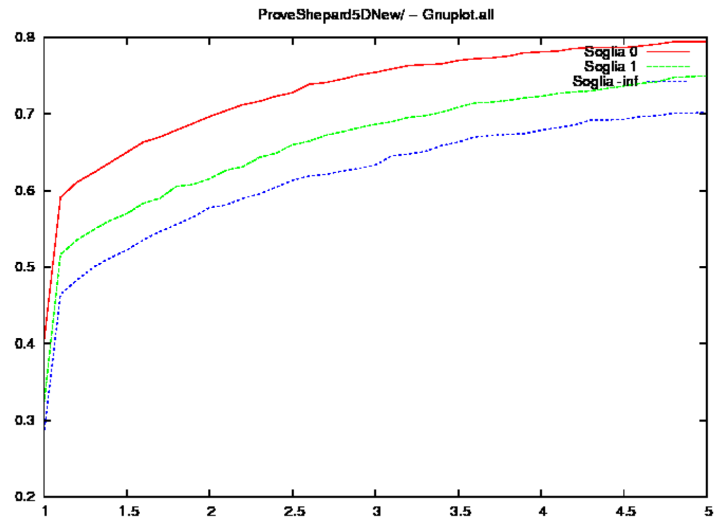
"Success" is defined as finding a point whose value is with 1% of the total range from the global optimum within 100 function evaluations.
Trials:

- our method with a correct lower bound threshold (equal to 0)
- our method with a strict lower bound threshold (equal to -1)
- the standard method (no lower bound, or threshold equal to $-\infty$)

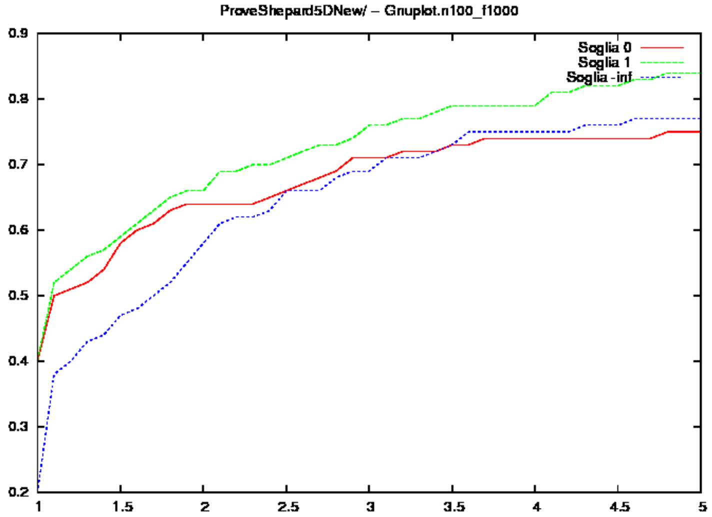
All (900) tests for 2D functions - minimum found



All (900) tests for 5D functions - minimum found



100 most difficult tests on 5D (100 stat.pts, range 0-1000)



Open problems

- what happens canceling the constraint $P^T\lambda = 0$? Existence of interpolant is guaranteed, uniqueness is not. However $\lambda^T\Phi\lambda$ is no more positive.
- We know how to find a minimum bumpiness interpolant in $\mathbb{R}^n$. But: how to find a minimum interpolant in a box?
- How to deal with more general constraints?
- How to choose threshold (aspiration) levels?
- How to deal with "Nan's" (function evaluation failures)
- How to avoid placing too many observations on the frontier?