Optimization of Noisy Functions: Application to Simulations

Geng Deng Michael C. Ferris

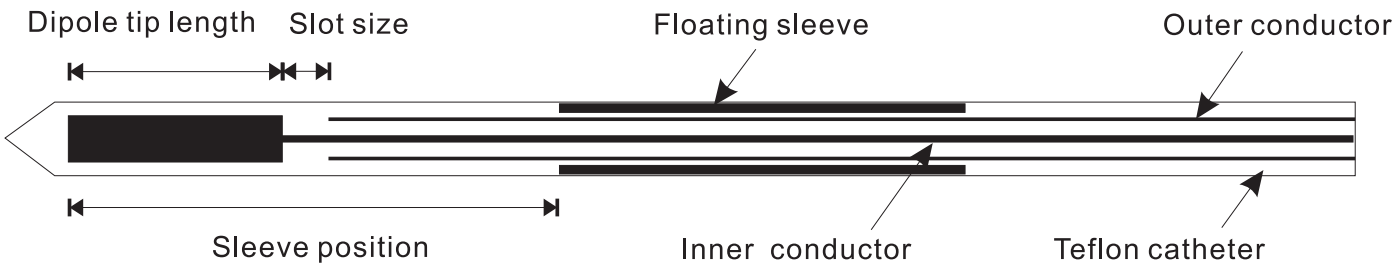
University of Wisconsin-Madison

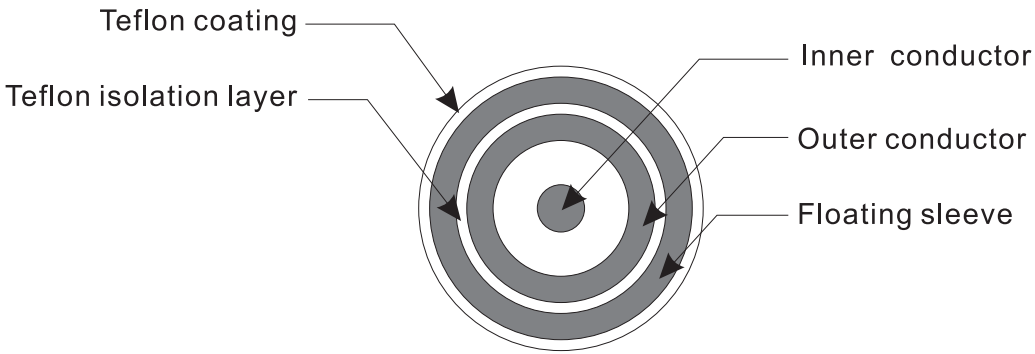
MIT, December 12, 2007

Simulation-based optimization problems

- Computer simulations are used as substitutes to evaluate complex real systems.
- Simulations are widely applied in epidemiology, engineering design, manufacturing, supply chain management, medical treatment and many other fields.
- The goal: Optimization finds the best values of the decision variables (design parameters or controls) that minimize some performance measure of the simulation.
- Other applications: calibration, SVM parameter tuning, inverse optimization, two-stage stochastic integer programming

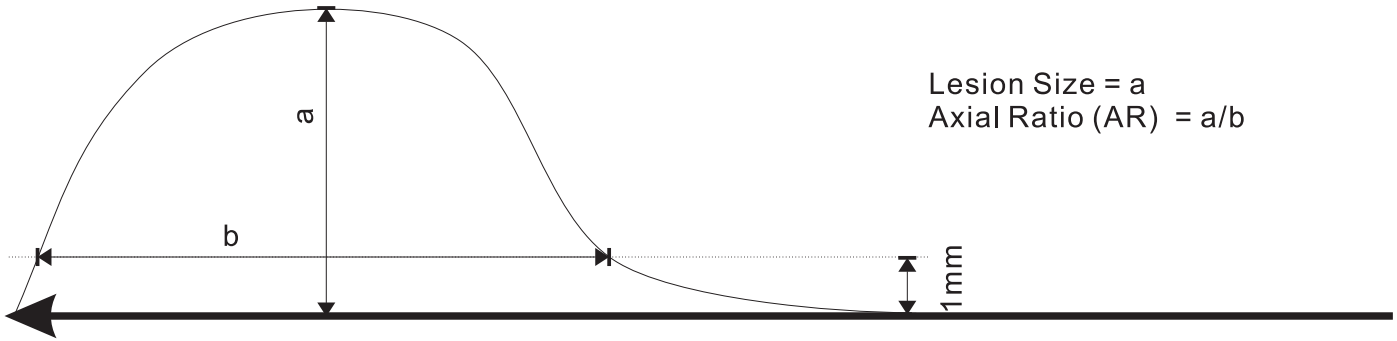
Design a coaxial antenna for hepatic tumor ablation





Simulation of the electromagnetic radiation profile

Finite element models (COMSOL MultiPhysics v3.2) are used to generate the electromagnetic (EM) radiation fields in liver given a particular design



Metric	Measure of	Goal
Lesion radius	Size of lesion in radial direction	Maximize
Axial ratio	Proximity of lesion shape to a sphere	Fit to 0.5
S_{11}	Tail reflection of antenna	Minimize

A general problem formulation

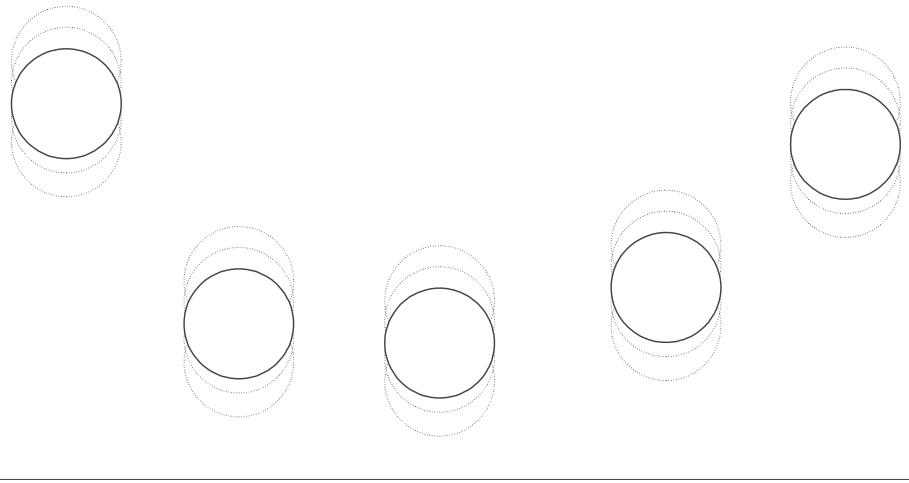
- We formulate the simulation-based optimization problem as

$$\min_{x \in \mathcal{S}} f(x) = \mathbb{E}[F(x, \xi(\omega))],$$

  $\xi(\omega)$ is a random component arising in the simulation process.
- The sample response function $F(x, \xi(\omega))$
  - typically does not have a closed form, thus cannot provide gradient or Hessian information
  - is normally computationally expensive
  - is affected by uncertain factors in simulation
- The underlying objective function $f(x)$ has to be estimated.

A simple discrete optimization case

- For example, test elasticity of a set of balls. Here $\mathcal{S} = \{1, 2, 3, 4, 5\}$ represents a set of 5 balls.



- Objective: Choose ball with the largest expected bounce height $f(x_i)$. $F(x_i, \xi_j)$ corresponds to a single measurement in an experiment.

How to select the best system

- Choose the maximum sample mean

$$\arg\max_{i \in \mathcal{S}} \bar{\mu}_i := \frac{1}{N_i} \sum_{j=1}^{N_i} F(x_i, \xi_j),$$

  where N_i is the number of experiments.
- Select the best system with high accuracy (PCS), while controlling the total amount of simulation runs.
- Two approaches
  - Indifference zone ranking and selection (Kim and Nelson, 2005)
  - Bayesian approach (Chick and Inoue, 2001a, 2001b)
- How to determine the replication number N_i?

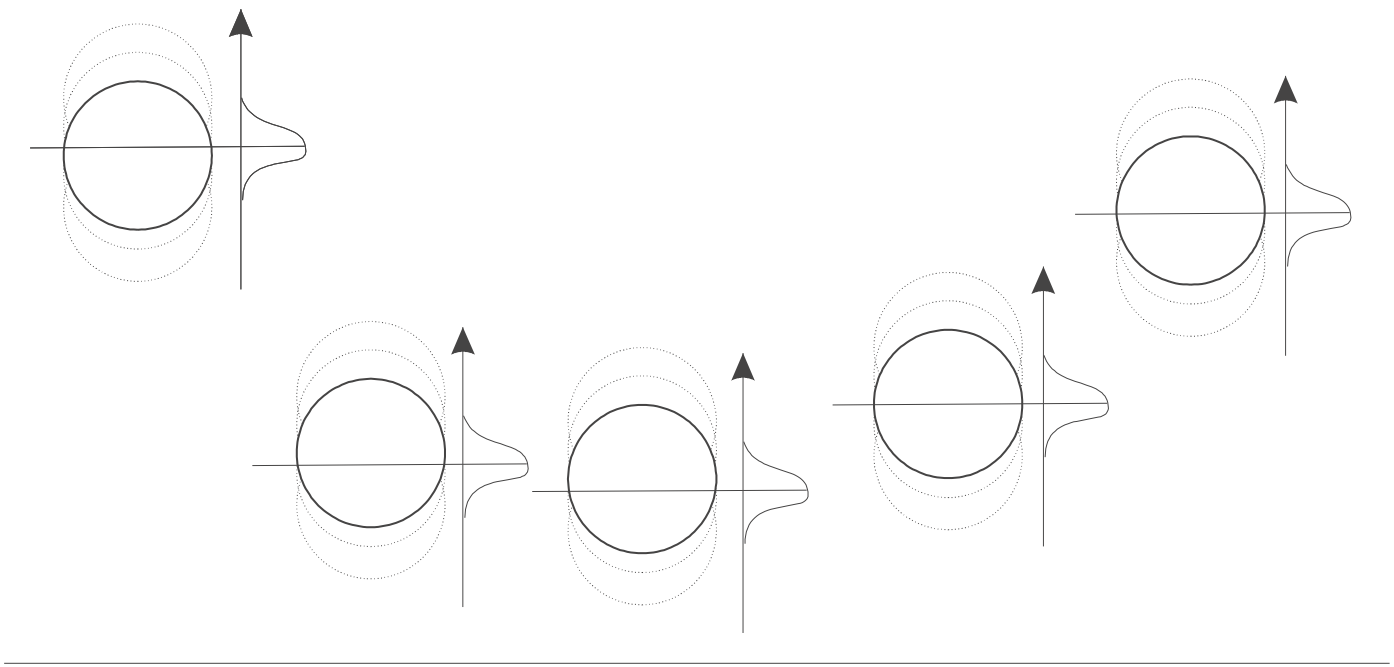
Bayesian approach

- Denote the mean of the simulation output for each system as $\mu_i = f(x_i) = \mathbb{E}[F(x_i, \xi(\omega))]$
- In a Bayesian perspective, the means are considered as Gaussian random variables whose posterior distributions can be estimated as

$$\mu_i | X \sim N(\bar{\mu}_i, \hat{\sigma}_i^2 / N_i),$$

where $\bar{\mu}_i$ is sample mean and $\hat{\sigma}_i^2$ is sample variance. The above formulation is one type of posterior distribution.

Posterior distributions facilitate comparison



Easy to compute the probability of correct selection (PCS).

Basic framework and tools

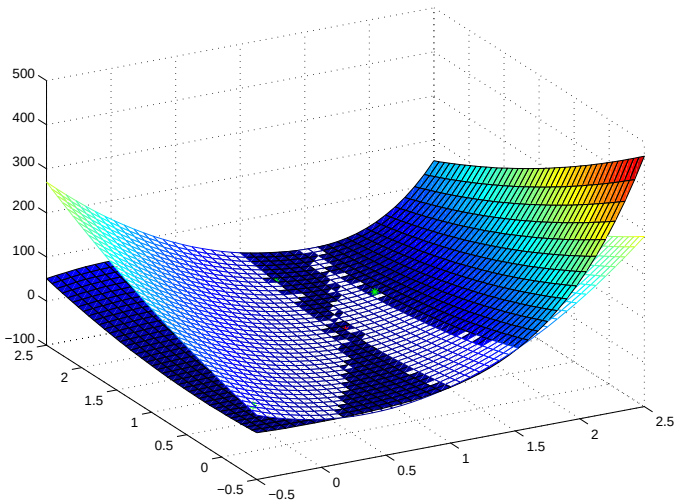
- Small scale x controls/design variables
- Simulation is refinable (replications, more samples in DES, finer discretization)

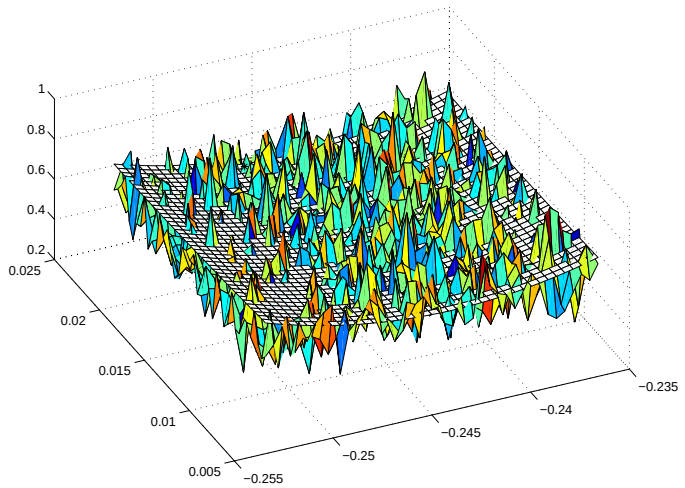
$$F(x) \simeq \frac{1}{N} \sum_{j=1}^{N} f(x, \omega_j)$$

- Bayesian approach
  - utilizes both mean and variance information
  - simple and direct to implement
  - flexible in choosing forms of posterior distributions
- Directly applicable to pattern search methods

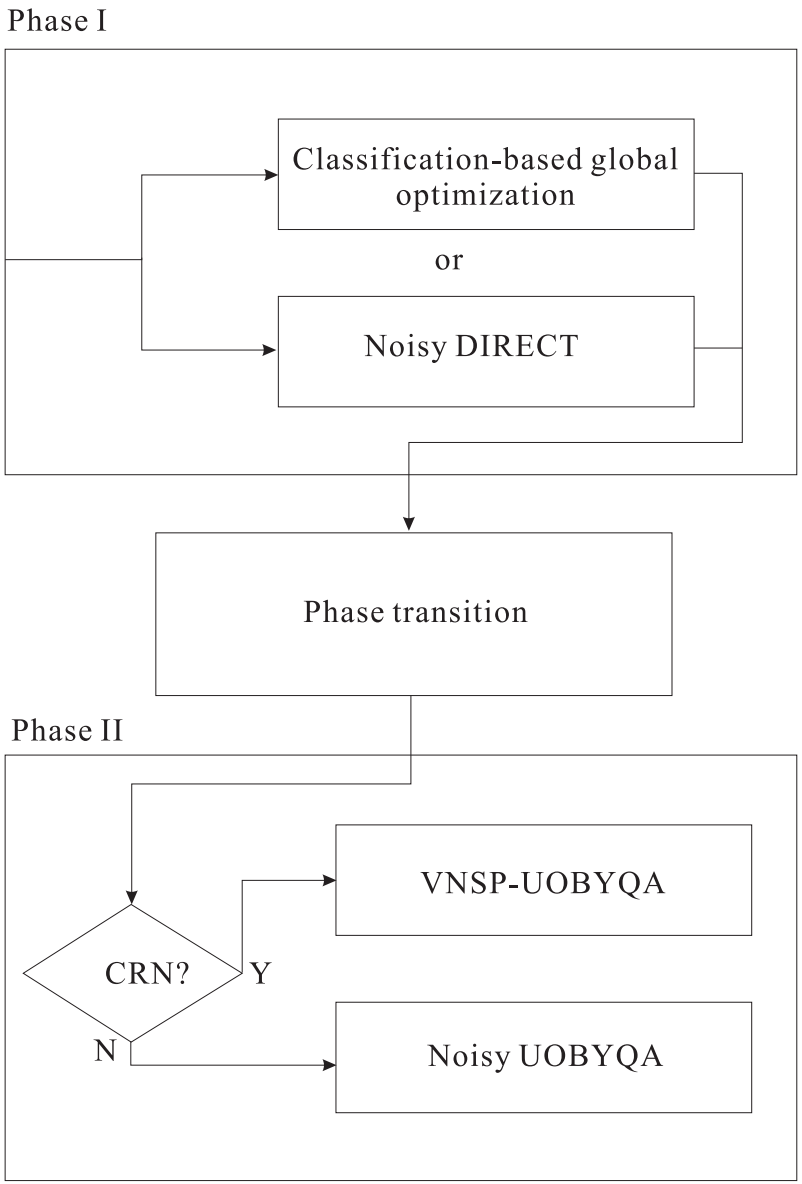
WISOPT two-phase optimization framework

1. **Phase I is a global exploration step.** The algorithm explores the entire domain and proceeds to determine potentially good subregions for future investigation.
2. **Phase II is a local exploitation step.** Local optimization algorithms are applied to determine the final solution.





The flow chart of WISOPT



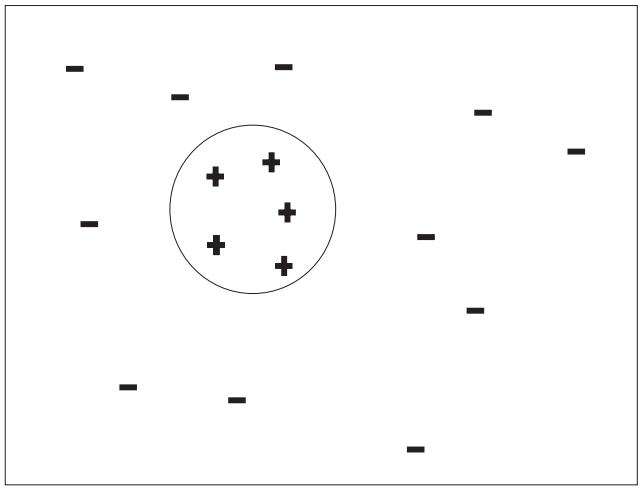
WISOPT Phase I: a classification based global search

- Classifier: surrogate for indicator function of the level set

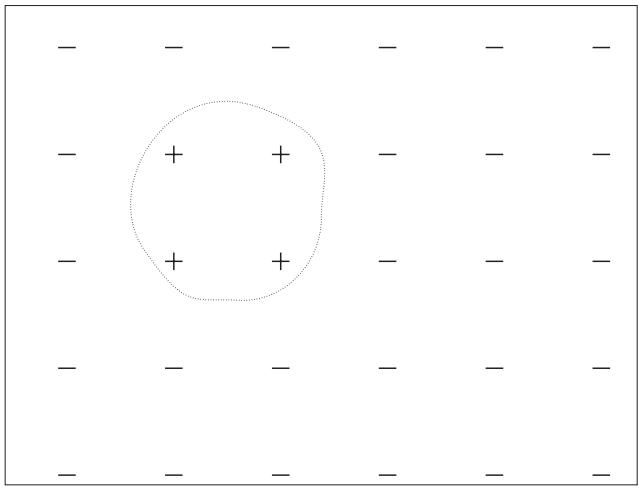
$$L(c) = \{x \,|\, f(x) \leq c\} \simeq \left\{ x \,\middle|\, \bar{f}(x) = \frac{1}{N} \sum_{j=1}^{N} F(x, \xi_j) \leq c \right\}$$

  c is a quantile point of the responses
- The level set corresponds to promising regions
- Training set: space-filling samples (points) from the whole domain (e.g. mesh grid; the Latin Hypercube Sampling)

Classifiers predict new refined samples as promising



(a) Training samples in $L(c)$ are classified as positive and others are negative. The solid circle represents estimated $L(c)$.



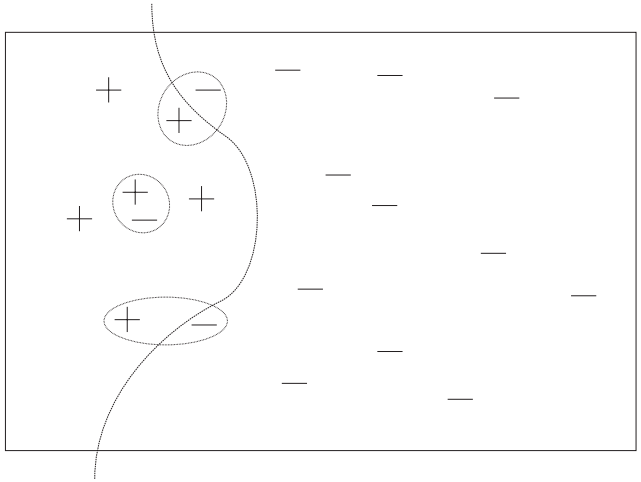
(b) Classify a set of more refined space-filling samples. Four points are predicted as positive and rest are negative. The classifier is refined.

Validate the subset of the identified promising points by performing additional simulations

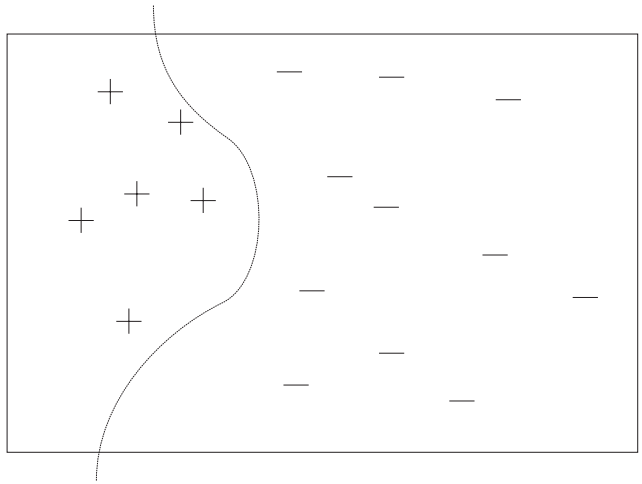
Imbalanced data

- To identify the top promising regions, the best 10% of the training samples are labeled as '+', and the rest are '-'
- The imbalance of the training set causes low classification accuracy, especially for positive members
- Balance the training data set
  - Under-sample of the negative class using one-sided selection
    - ★ Use 1-NN and retain only those negative samples needed to predict training set
    - ★ Clean the dataset with Tomek links
  - Over-sample of the positive class by duplicating positive samples
- Adjust the misclassification penalty

Cleaning the dataset with Tomek links

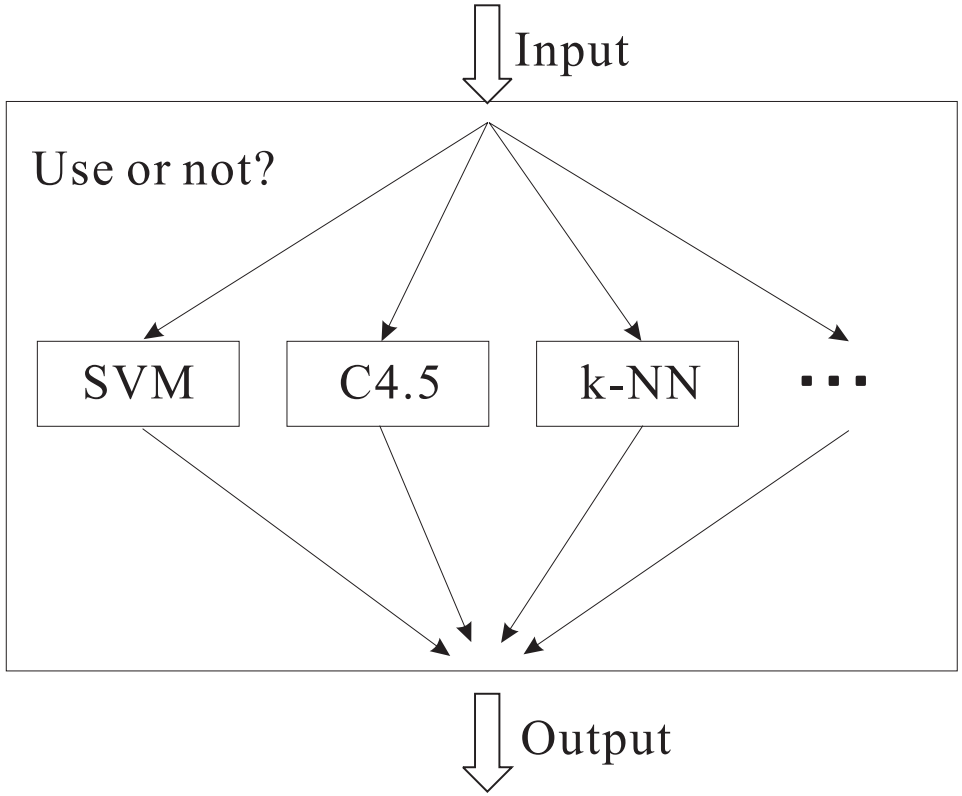


(c) Determine the pairs of Tomek links

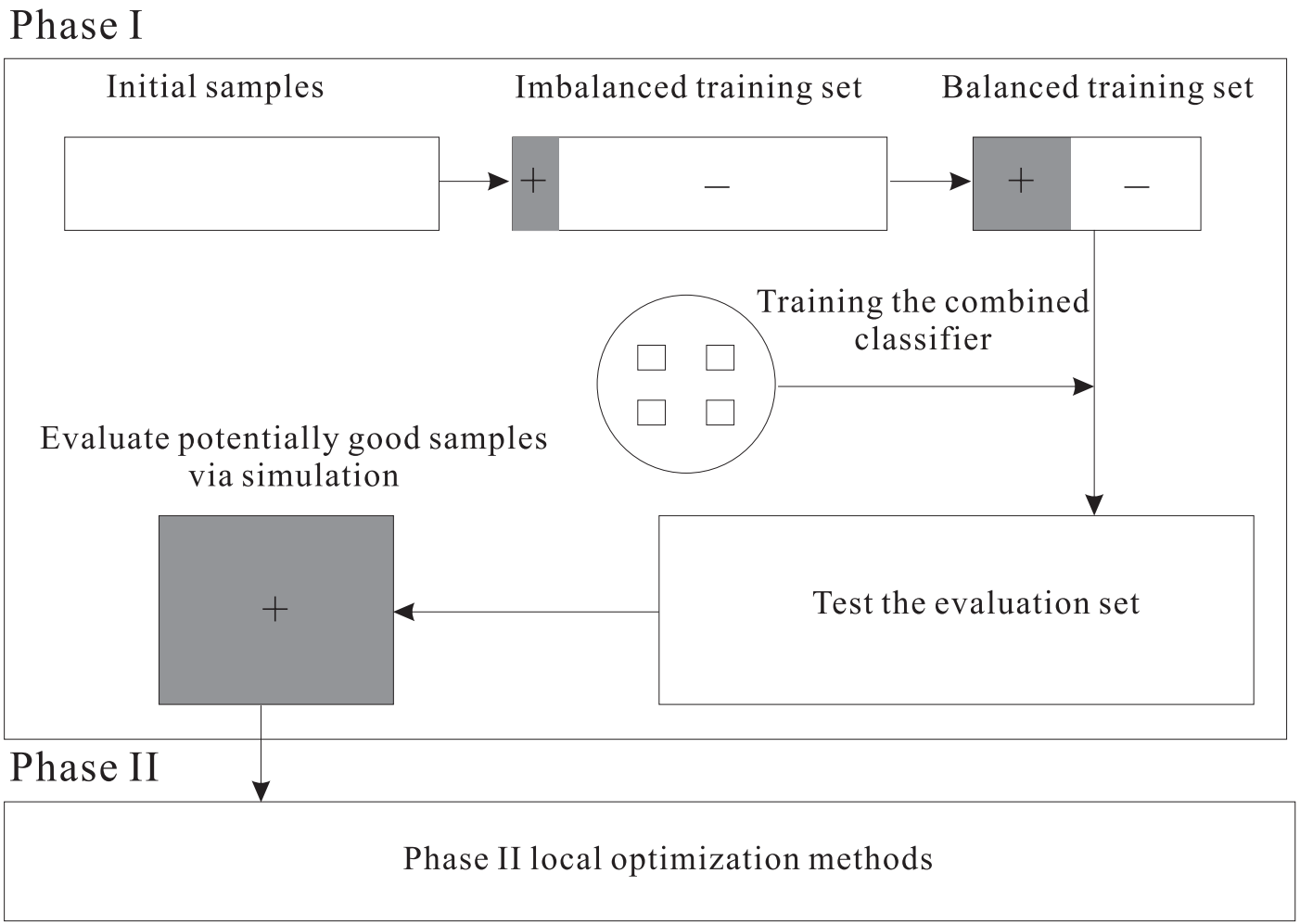


(d) Remove the negative samples participating as Tomek links

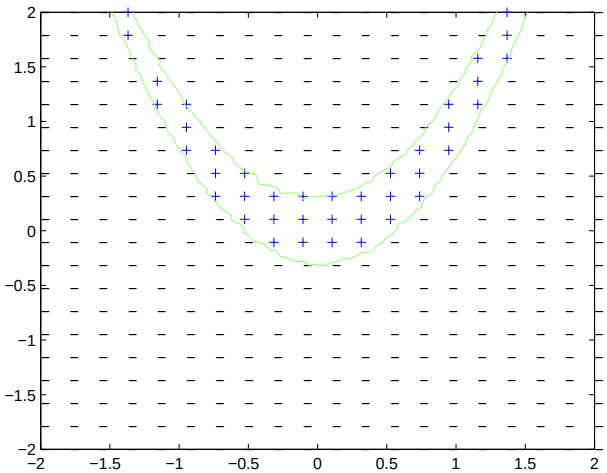
Assemble classifiers using a voting scheme

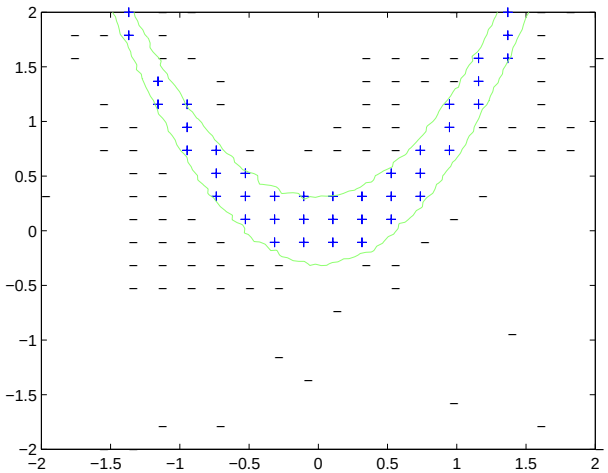


Classifier Phase I approach

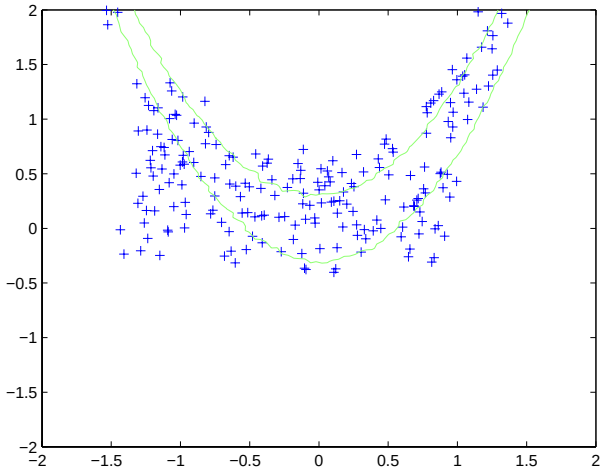


Banana example



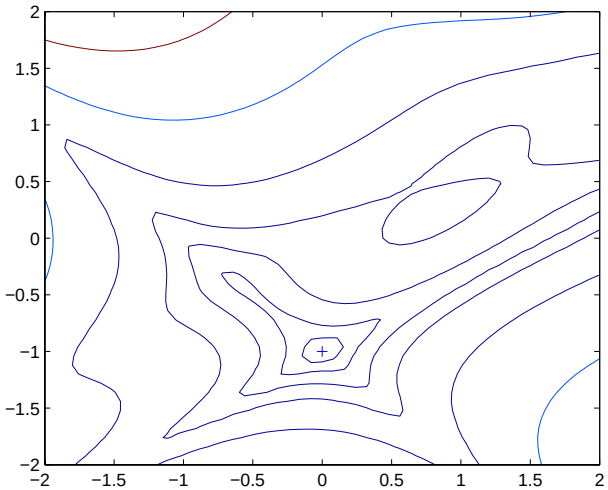


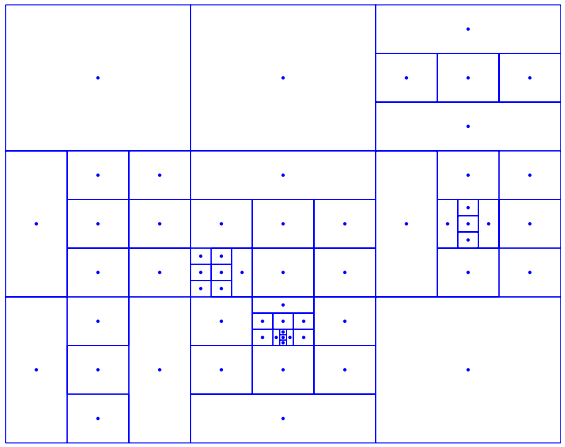
Original Predicted Training



WISOPT Phase I: the Noisy DIRECT (Jones et. al)

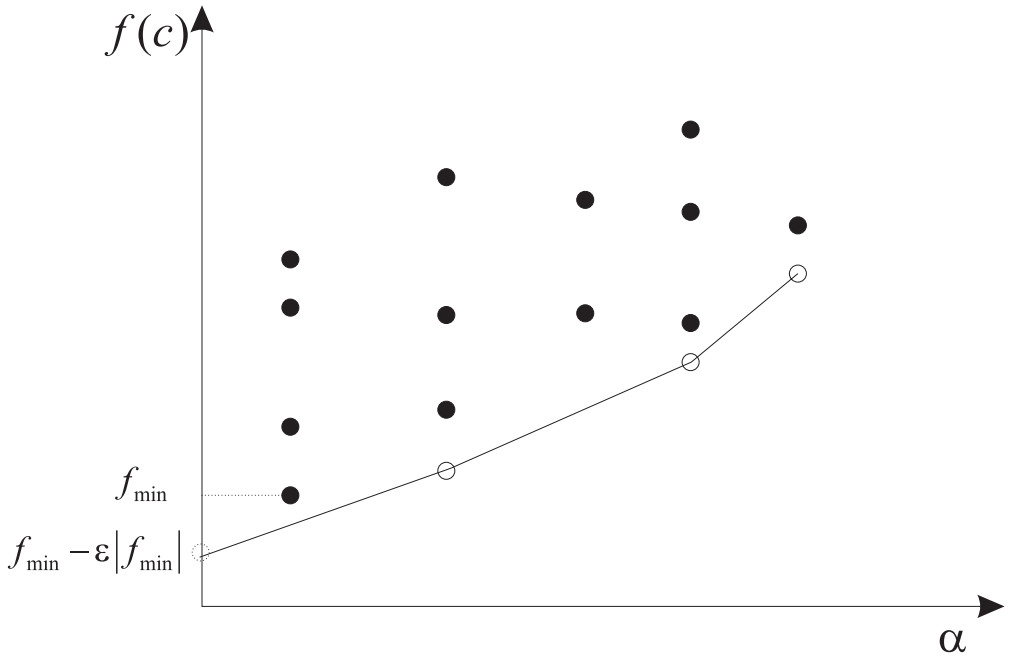
- At each iteration, trisect a collection of promising boxes (large box or small F)
- Evaluate F at center of newly generated boxes





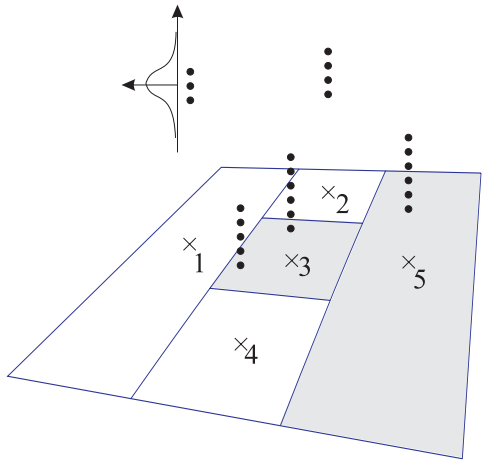
Partitioning hyperrectangles: DIRECT (DIviding RECTangles)

- Partitioning hyperrectangles
- Identifying potentially optimal hyperrectangles



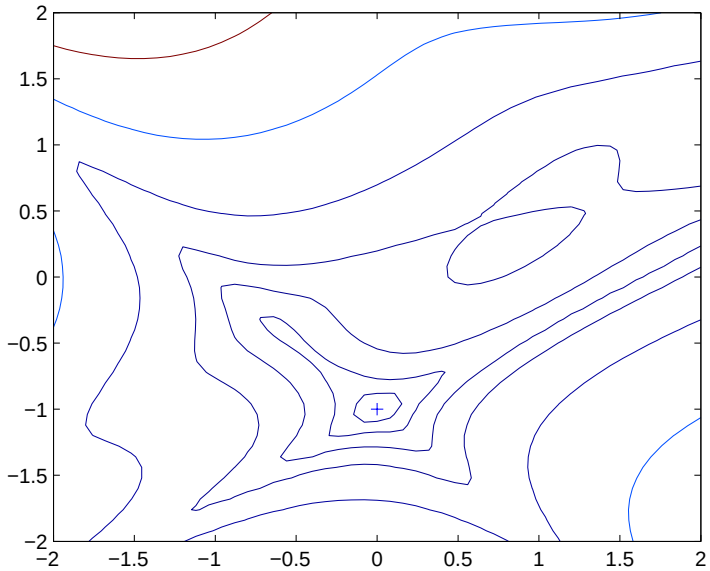
Noisy extension

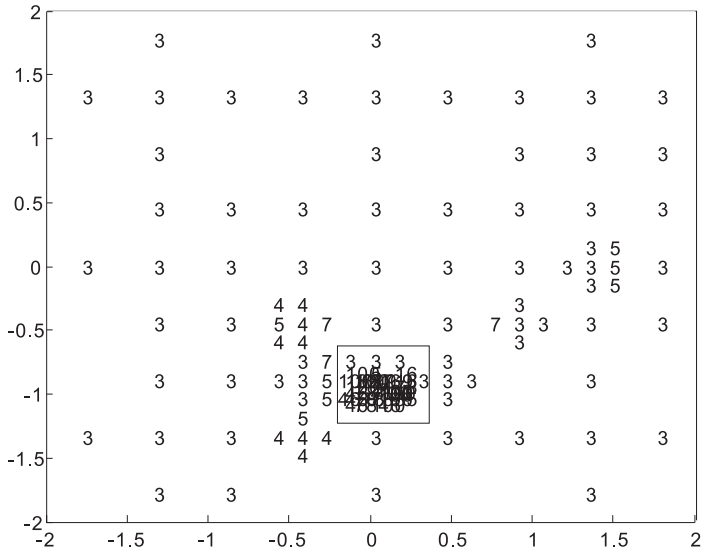
- Bayesian methods determine posterior distribution of "box center" F values

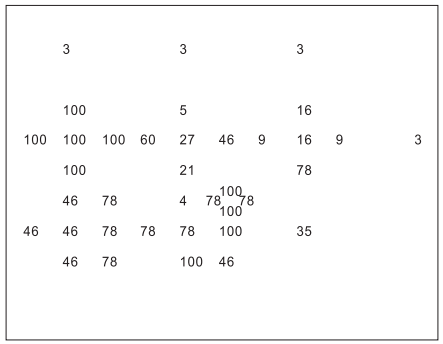


- Monte Carlo methods to generate "sampled" values for F; then use DIRECT to generate "trial" potential boxes
- Compare error rates against boxes generated from sample means
- When error rate large (sets of boxes chosen differ greatly), increase replications on those boxes that produce errors

Numerical results: Goldstein Price function







Phase I methods

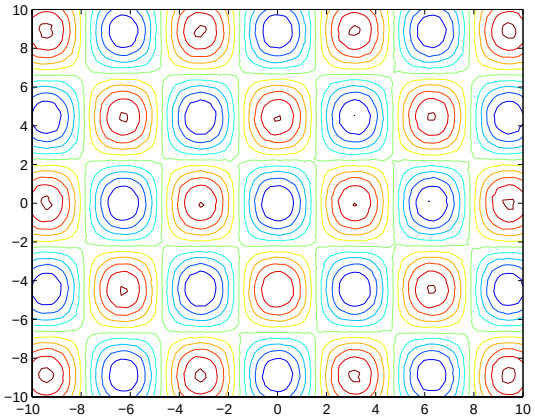
Properties of Phase I methods:

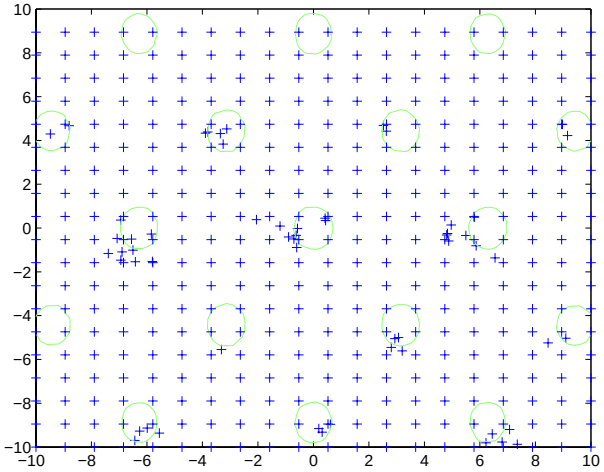
- Methods must provide a global view of function
- Yield samples densely distributed in promising subregions
- A nonparametric local regression procedure used to identify the promising regions

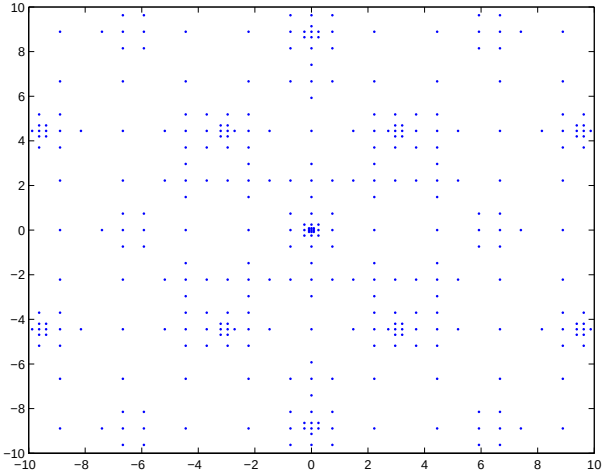
Comparisons of the classification-based search and the Noisy DIRECT method

- Robustness to noise
- Dimension of the problem
- Density of samples
- Implementation

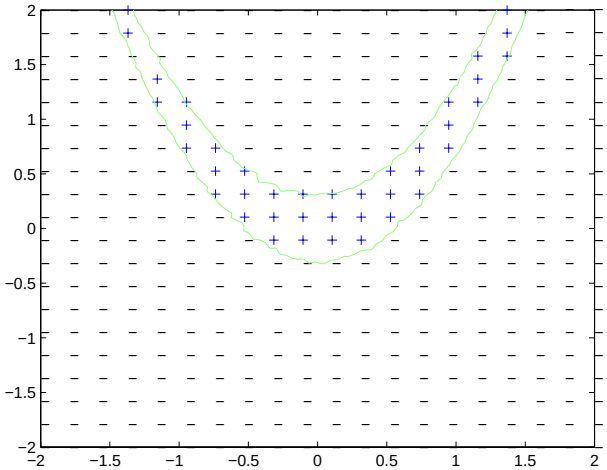
Classifier vs Direct (example Griewank)

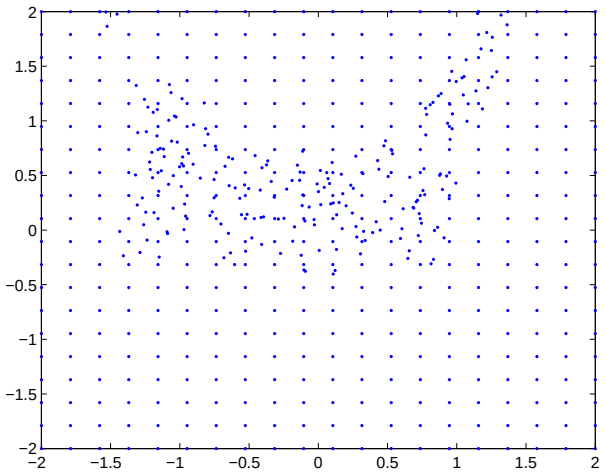






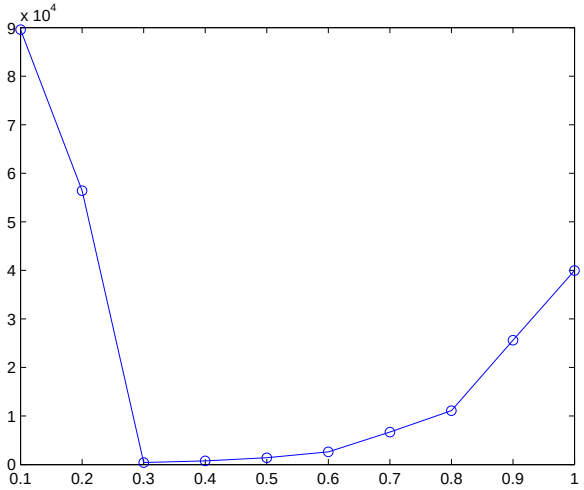
The non-parametric "linking" idea

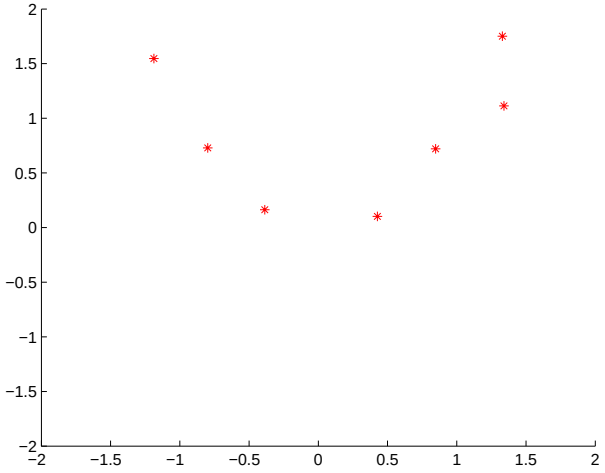




Original / $sse(h)$

Data / Result





Determine subregion radius by non-parametric regression

The idea is to determine the best 'window size' for non-parametric local quadratic regression

1. $\Delta \in \arg\min_h sse(h)$
2. $sse(h)$ is the sum of squared error of knock-one out prediction. Given a window-size h and a point y, the knock-one out predicted value is $Q_h^y(y)$, where $Q_h^y(x)$ is a quadratic regression function constructed using the data points within the ball $\{x | \|x - y\| \leq h\}/\{y\}$.

$$Q_h^y(x) = c + g^T(x - y) + \frac{1}{2}(x - y)^T H(x - y)$$

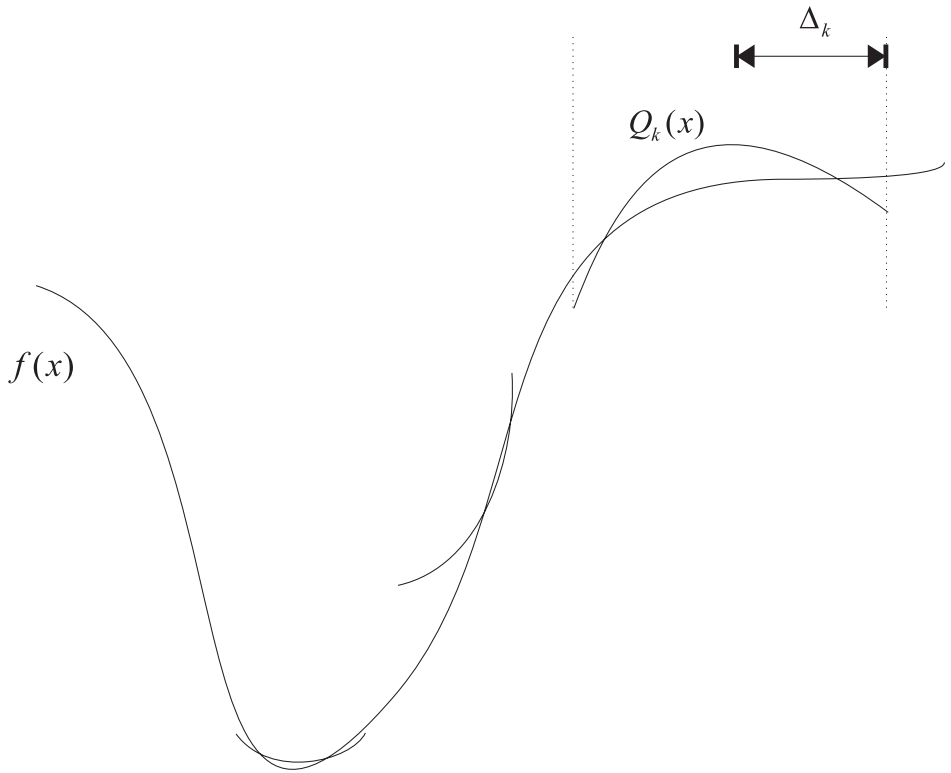
.

Phase II: refine solution

- Local optimization methods to handle noise
- Derivative-free methods
- Basic approach: reduce function uncertainty by averaging multiple samples per point.
- Potential difficulty:
  efficiency of algorithm vs number of simulation runs
- We apply Bayesian approach to determine appropriate number of samples per point, while simultaneously enhancing the algorithm efficiency

Phase II: Extensions of the UOBYQA algorithm

The base derivative free optimization algorithm: The UOBYQA (Unconstrained Optimization BY Quadratic Approximation) (Powell 2002) algorithm is based on a trust region method. It constructs a series of local quadratic approximation models of the underlying function.



Quadratic model construction and trust region subproblem solution

For iteration $k = 1, 2, \ldots,$

- $\ldots$
- Construct a quadratic model via interpolation

$$Q(x, \xi) = F(x_k, \xi) + g_Q^T(\xi)(x - x_k) + \frac{1}{2}(x - x_k)^T G_Q(\xi)(x - x_k)$$

  The model is unstable since interpolating noisy data

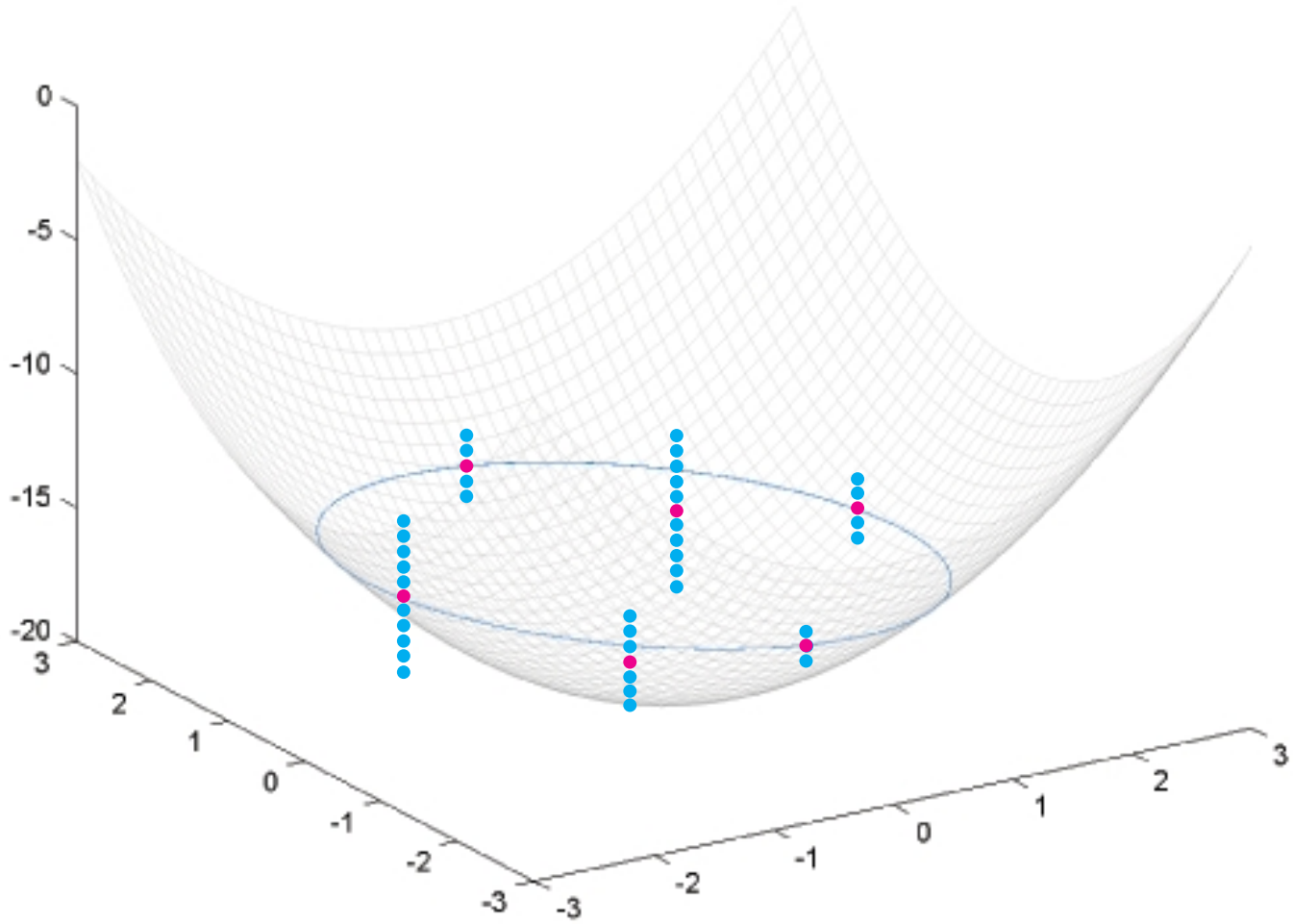
- Solve the trust region subproblem

$$\begin{aligned} s_k^*(\xi) = \arg\min_s \quad & Q(x_k + s, \xi) \\ s.t. \quad & \|s\|_2 \leq \Delta_k \end{aligned}$$

  The solution is thus unstable

- $\ldots$

Why is the quadratic model unstable?



How to stabilize the quadratic model?

Let $\{y^1, y^2, \ldots, y^L\}$ be the interpolation set.

- Quadratic interpolation model is a linear combination of Lagrange functions:

$$Q(x, \xi) = \sum_{j=1}^{L} \bar{f}(y^j, \xi) l_j(x).$$

- Each piece $l_j(x)$ is a quadratic polynomial, satisfying

$$l_j(y^i) = \delta_{ij}, i = 1, 2, \cdots, L.$$

- The coefficients of l_j are uniquely determined, independent of the random objective function.

Bayesian estimation of coefficients

In Bayesian approach, the mean of function output $\mu(y^j) := f(y^j) = \mathbb{E}[F(y^j, \xi(\omega))]$ is considered as a random variable:
Normal posterior distributions:

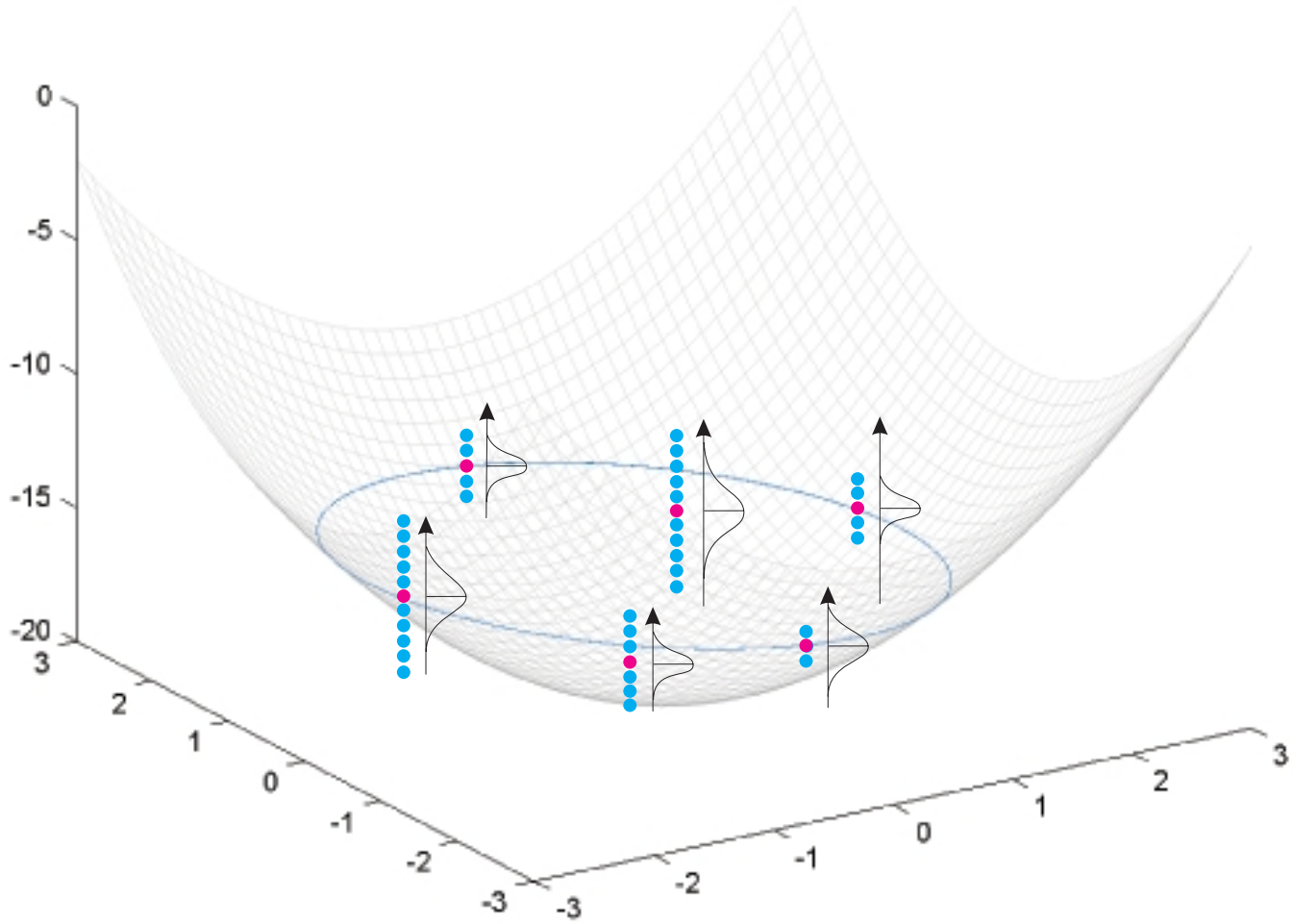
$$\mu(y^j)|X \sim N(\bar{\mu}(y^j), \hat{\sigma}^2(y^j)/N_j)$$

Thus the coefficients of the quadratic model Q^∞ are estimated as:

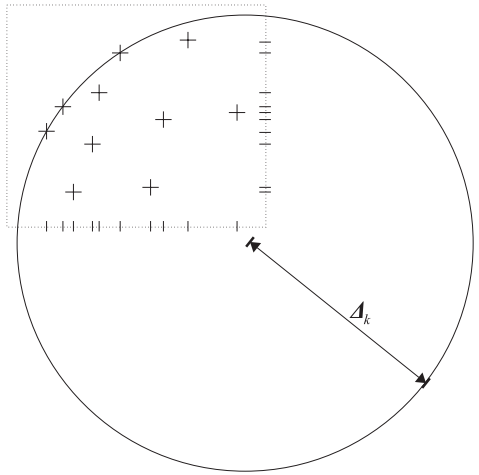
$$\begin{aligned} g_Q^\infty|X &= \sum_{j=1}^{L}(\mu(y^j)|X)g_j, \\ G_Q^\infty|X &= \sum_{j=1}^{L}(\mu(y^j)|X)G_j. \end{aligned}$$

- g_Q^∞, G_Q^∞ are coefficients of Q^∞
- g_j, G_j are coefficients of Lagrange functions l_j
- g_j, G_j are deterministic and determined by points y^j

Bayesian posterior distributions



Constraining the variance of solutions (Monte Carlo validation)



- Generate 'sample quadratic functions' that could arise given current function evaluations.
- Trial solutions are generated within a trust region. The standard deviation of the solutions are constrained.

$$\max_{i=1}^{n} std([s^{*(1)}(i), s^{*(2)}(i), \cdots, s^{*(M)}(i)]) \leq \beta \Delta_k.$$

Sufficient reduction criterion - CRN case

$$Pr\left(Q(x_k) - Q(x_k + s_k^*) \geq \kappa_{mdc} \|g_Q^\infty\| \min\left[\frac{\|g_Q^\infty\|}{\kappa_{Qh}}, \Delta_k \right] \right) \geq 1 - \alpha_k$$

- g_Q^∞ is the gradient of the quadratic model Q^∞
- α_k is a significance level.
- κ_{mdc} and κ_{Qh} are constants.
- $Q(x_k) - Q(x_k + s_k^*)$ is the observed model reduction. The criterion implies that g_Q^∞ is bounded by the model reduction.

Bayesian estimation

Given the posterior distribution $g_Q^\infty|X$, the probability value can be estimated:

$$Pr\left(Q(x_k) - Q(x_k + s_k^*) \geq \kappa_{mdc}\|g_Q^\infty|X\| \min\left[\frac{\|g_Q^\infty|X\|}{\kappa_{Qh}}, \Delta_k\right]\right) \geq 1 - \alpha_k$$

Lemma (Borel-Cantelli Lemma)

If $\sum_k Pr(E_k) < \infty$, then the probability that infinitely many E_k happen is 0 (implies that finitely many E_k happen with probability 1)

We require $\sum_k \alpha_k < \infty$ and let E_k be the event of failure to satisfy the sufficient reduction criterion. This lemma implies that there is a large enough index K, such that the sufficient reduction criterion is satisfied w.p.1 for $k \geq K$.

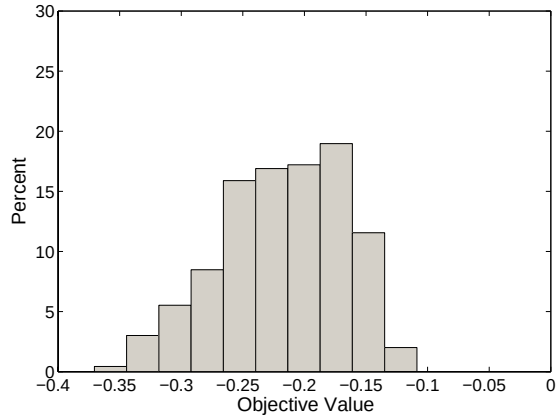
Noisy UOBYQA for Rosenbrock, $n = 2$ and $\sigma^2 = 0.01$

Iteration (k)	FN	$F(x_k)$	Δ_k
1	1	404	2
20	78	3.56	9.8×10^{-1}
40	140	0.75	1.2×10^{-1}
60	580	0.10	4.5×10^{-2}
80	786	0.0017	5.2×10^{-3}
✓ Stops with the new termination criterion			
100	1254	0.0019	2.8×10^{-4}
120	2003	0.0016	1.1×10^{-4}
✓ Stops with the termination criterion $\Delta_k \leq 10^{-4}$			

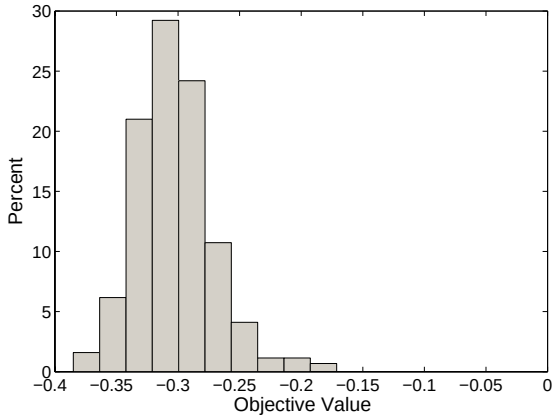
Two-phase approach to optimize antenna design parameters

- Uniform LHS to generate 2,000 design samples to evaluate with the FE simulation model (range [-0.3705, 3597])
- Histogram of objective values over interval [-0.3705, 0]
- $c = -0.2765$ the 10% quantile. $L(c)$ has 199 positive samples (1801 negative)
- Balancing procedure: 398 positive vs. 388 negative samples
- 5 (of 6 tested) classifiers in ensemble
- Refined data: 15,000 designs, 522 predicted by classifiers as positive, 74% correctly
- The best Phase I design has value -0.3850.

Coaxial antenna design



(e) First stage initial designs



(f) Designs predicted by classifiers

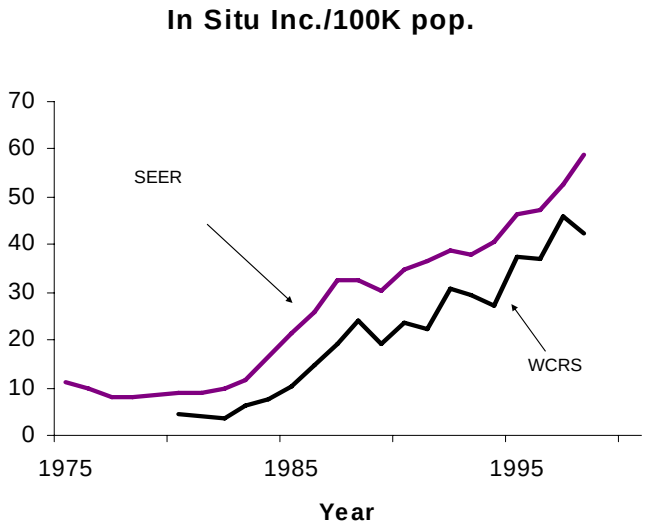
- (Modified) UOBYQA started from best point: (13.6 2.7 19.0 0.3 0.1) mm, value -0.3850.
- UOBYQA returned an optimal solution: (15.9 2.4 19.0 0.3 0.1) mm, value -0.4117.

Sample path extension: changing liver properties

- Common random numbers allow variance reduction, correlated noise.
- Extension of ideas to Variable-Number Sample-Path Optimization method.
- Application: Dielectric tissue properties varied within $\pm 10\%$ of average properties to simulate the individual variation.
- Bayesian VNSP algorithm yields an optimal design that is a 27.3% improvement over the original design and is more robust in terms of lesion shape and efficiency.

Simulation calibration

- Detailed individual-woman level discrete event simulation of Wisconsin Breast Cancer Incidence (using 4 processes):
  - Breast cancer natural history
  - Breast cancer detection
  - Breast cancer treatment
  - Non-breast cancer mortality among US women
- Replicate breast cancer surveillance data: 1975-2000

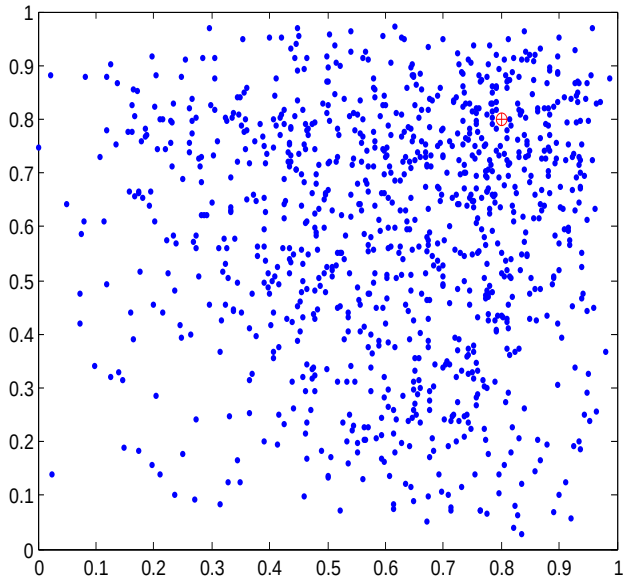


Application to WBCE

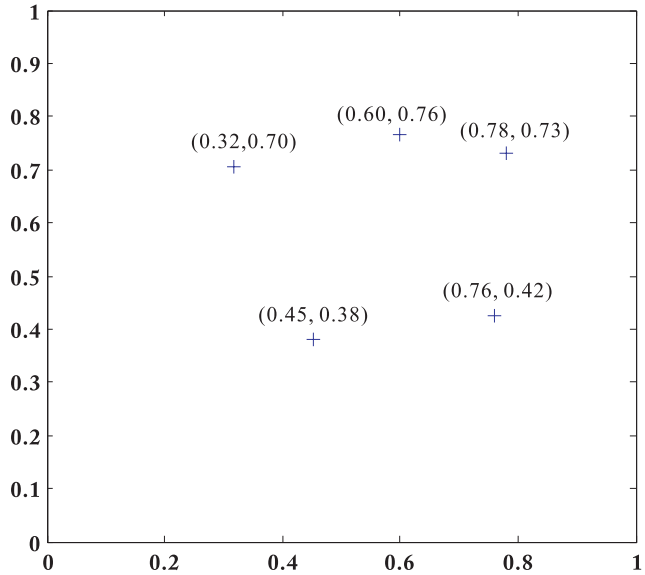
- 500,000 points x generated uniformly at random
- Using CONDOR (120 machines) can evaluate approximately 1000 per day $F(x, \xi)$ involves simulation of 3 million women
- 363 are in $L(10)$: "simulated points out of data envelope"
- Using Phase I: 10,000 points evaluated, 220 points suggested, 195 are in L(10)
- Phase I results in new points (all are good), but 2 of which seem better than the "experts" best solution
- Phase II: Using the idea of sample-path optimization. New dataset contains 10 replications at points
- Kriging models are constructed based on the dataset and optimization methods are applied to minimize the Kriging model.

Ambulance simulation

An ambulance is called when an emergency call occurs. Determine the locations of the ambulance bases such that the expected response time to emergency calls is minimized.



(g) The distribution of emergency calls



(h) The locations of ambulance bases

Conclusions and future work

- Coupling statistical and optimization techniques can effectively process noisy function optimizations
- Significant gains in system performance and robustness are possible
- WISOPT framework allows multiple methods to be "hooked" up

Future work:

- Problems with general constraints
- More optimization algorithms in both phases
- A phase transition module with variable radii