Bayesian Multiobjective Optimization with Gaussian Process Models

Foundations and Applications

Michael T. M. Emmerich
Johannes Kruisselbrink
Natural Computing Lab
LIACS, Leiden University
The Netherlands

April 14, 2022

Table of Contents

Introduction

Curse of dimensionality

Kriging

Theory

Hyperparameter Estimation

Global optimization

Conclusions

Multiobjective Optimization

Multiobjective Expected Improvement

Monotonicity Properties

Problem definition

Consider a (computer) experiment with deterministic, continuous, output:

$$f : \mathbb{R}^d \to \mathbb{R} \tag{1}$$

with expensive evaluation $y = f(\mathbf{x})$ of f (time, money).
We already evaluated $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}$ and the output was

$$y^{(1)} = f(\mathbf{x}^{(1)}), y^{(2)} = f(\mathbf{x}^{(2)}), \ldots, y^{(n)} = f(\mathbf{x}^{(n)})$$

.
Then for a new point $\mathbf{x}'$ we would like to predict $\hat{y}(\mathbf{x}') = f(\mathbf{x}')$.

Problem difficulty

A very general bound for function approximation can be stated for Hölder continuous functions.[1]:
The Hölder class of functions with parameters k and α is defined as

$$F = \{f : [a,b]^d \to \mathbb{R}| \quad |D^{\mathbf{r}} f(x) - D^{\mathbf{r}} f(y)| = ||x-y||^{\alpha}\}, \\ \forall \mathbf{r} \text{ with } |\mathbf{r}| = k, 0 < \alpha \leq 1. \qquad (2)$$

Here $D^{\mathbf{r}} := \frac{\partial^{|\mathbf{r}|}}{\partial x_1^{r_1} ... \partial x_d^{r_d}}$ (Schwartz derivative), $\mathbf{r} = (r_1, \ldots, r_d)$ is a n-tuple of non-negative integers with $|\mathbf{r}| = r_1 + \cdots + r_d$.

[1] Ritter, K., Wasilkowski, G., and Wozniakowski H.: On multivariate integration for stochastic processes. In: H. Brass and G. Hammerlin, Numerical Integration, Birkhäuser Verlag, Basel, 1993

Curse of dimensionality

Let $\mathcal{B} = [\mathbf{a}, \mathbf{b}]^d$ and define the error as

$$e_t = \max_{\mathbf{x} \in \mathcal{B}} (|y(\mathbf{x}) - \hat{y}_t(\mathbf{x})|)$$

, whereas $\hat{y}_t$ is the prediction based on t evaluations that are optimally placed in the search space, with regard to error minimization.
For the Hölder class of functions it is possible to prove the sharp lower bound for the maximal error e_t:

$$e_t \geq c n^{-(k+\alpha)/d} \text{ or } n = \left(\frac{c}{e_t} \right)^{d/(k+\alpha)} \tag{3}$$

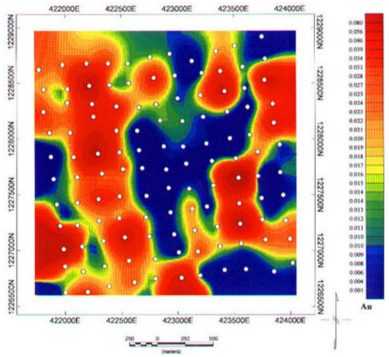
In the average case better approximations may be possible.

Some interpolation and prediction methods

- Regression methods (minimizing the squared residuals)
- Artificial neural networks (backpropagation of errors)
- Radial basis functions (minimizing the error at training points)
- Splines (functions maximizing smoothness, differentiability)
- *Statistical interpolation (random field interpretation of data)*

Kriging method from geostatistics[2]; mathematically developed by Matheron[3]

- The *kriging* interpolation method was proposed by mining engineer Krige
- Interpolation from an irregular grid of measurements

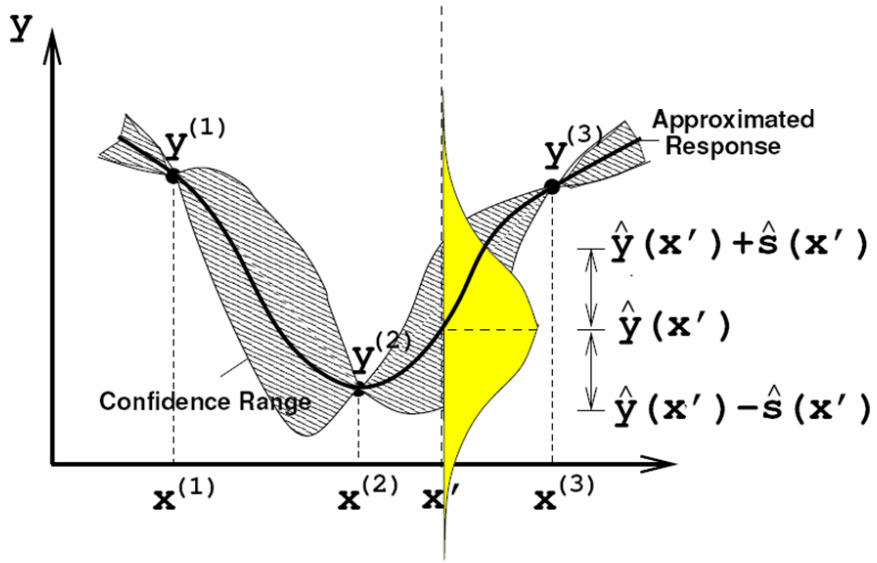


Goldfield exploration map. White points indicate measurement sites.

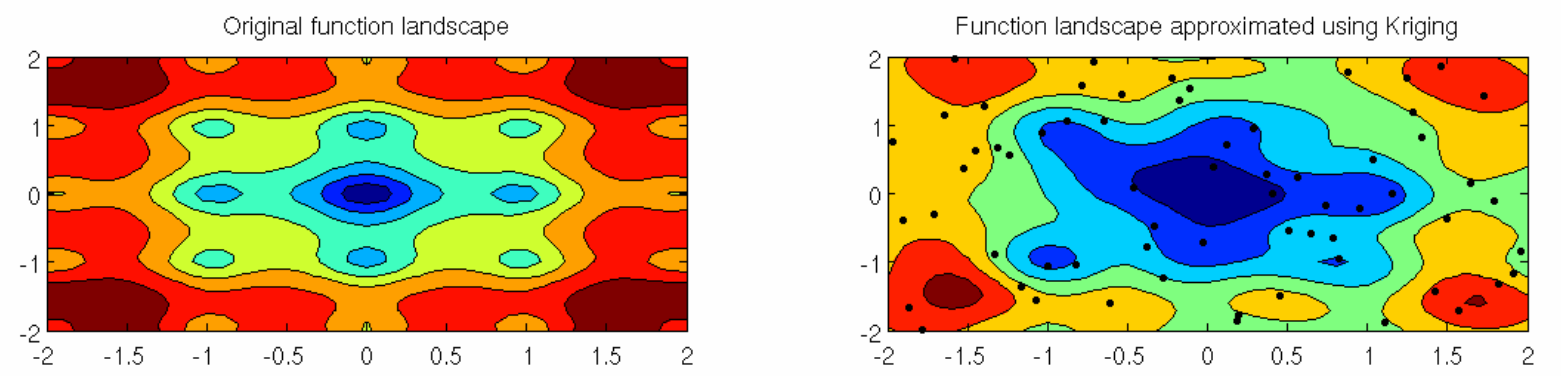
[2]Krige, Danie G. (1951) A statistical approach to some basic mine valuation problems on the Witwatersrand. J. of the Chem., Metal. and Mining Soc. of South Africa 52 (6): 119 - 139.

[3]Matheron, Georges (1962). Traité de géostatistique appliquée. Editions Technip.

1-D example of kriging



- Besides the prediction $\hat{y} \approx f(\mathbf{x}')$ also the uncertainty margin $\hat{s}$ is computed.
- A predictive (normal) distribution describes likelihood of different results at $\mathbf{x}'$.



An example of kriging in 2-D on Ackley's function[4] [5].

[4] Computed with kriging-matlab package by Johannes Kruisselbrink, LIACS, Leiden University.

[5] D. H. Ackley. A connectionist machine for genetic hillclimbing. Boston: Kluwer Academic Publishers, 1987.

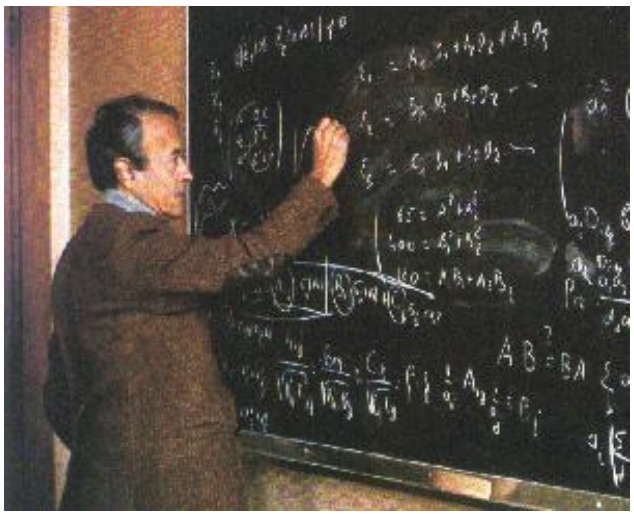
Kriging in computer experiments

- Sacks at al. (1989) suggested kriging for modeling from computer experiments[6] (input vectors $\hat{=}$ spatial coordinates)
- Exact interpolations are required at the data points in case of deterministic function interpolation
- Why to use a statistical approach for interpolating deterministic functions:
  - Data is interpreted as sample from realization of a stochastic process
  - Coherent approach to reason about model and prediction errors
  - Kriging can be viewed as regression with spatial correlation

[6] Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P., "Design and analysis of computer experiments." Statistical Science 4, pp. 409-435, 1989.

Theory of kriging

- Two ways to motivate gaussian processes and kriging:
- (A) Bayesian interpretation (gaussian processes)
- (B) Best linear unbiased prediction (kriging)
- In case of gaussian distributed random variables, both lead to equivalent or similar expressions

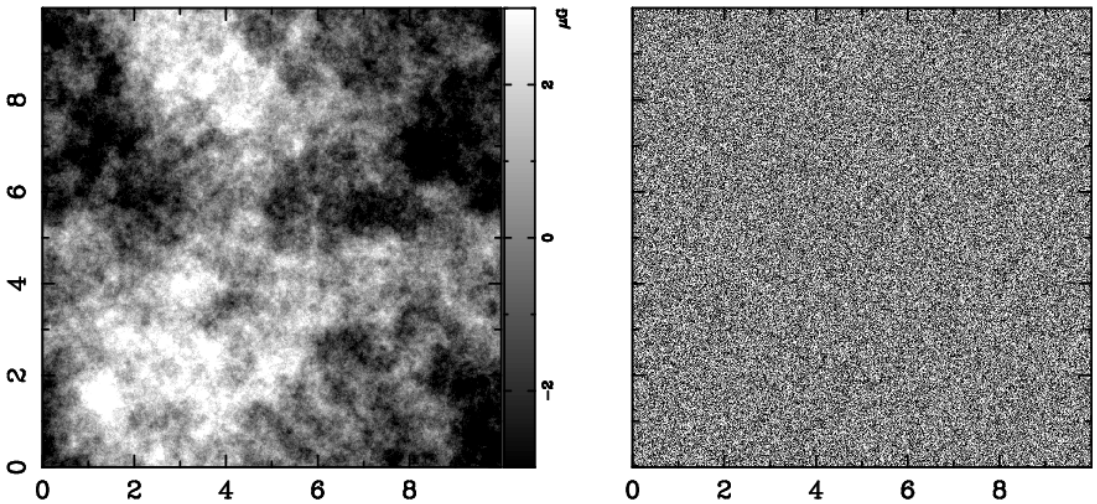


Georges F. P. M. Matheron (1930 August 7, 2000), French mathematician and geologist, known as the founder of geostatistics and a co-founder (together with Jean Serra) of mathematical morphology.

Examples for gaussian random field realizations

A gaussian random field (GRF) is a family of normal random variables indexed by space, say

$$\mathcal{F}_{\mathbf{x}}, \mathbf{x} \in \mathbb{R}^d, c(\mathcal{F}_{\mathbf{x}}, \mathcal{F}_{\mathbf{x}}) \equiv c(\mathbf{x}, \mathbf{x}')$$



Realizations of gaussian random fields

High correlation (left) and zero correlation (right, 'white noise').

Gaussian random fields

- A gaussian random field (GRF) is a family of 1-D gaussian random variables indexed by space, say

$$\mathcal{F}_{\mathbf{x}}, \mathbf{x} \in \mathbb{R}^d$$

- In (simple, ordinary) kriging we consider a random field that is homoscedastic and has a constant mean, i.e.

$$\forall \mathbf{x} \in \mathbb{R}^d : \mathcal{F}_{\mathbf{x}} \sim N(\mu, \sigma)$$

- The correlation between two $\mathcal{F}_{\mathbf{x}}$ and $\mathcal{F}_{\mathbf{x}'}$ is determined by the relative position of $\mathbf{x}$ and $\mathbf{x}'$ to each other:

$$c(\mathcal{F}_{\mathbf{x}}, \mathcal{F}_{\mathbf{x}'}) \equiv f(\mathbf{x}, \mathbf{x}')$$

- common are *stationary* correlation $c(\mathcal{F}_{\mathbf{x}}, \mathcal{F}_{\mathbf{x}'}) \equiv f(\mathbf{x} - \mathbf{x}')$, and even *isotropic* correlation $c(\mathcal{F}_{\mathbf{x}}, \mathcal{F}_{\mathbf{x}'}) \equiv f(||\mathbf{x} - \mathbf{x}'||)$.

Type of correlation

- The correlation decreases with distance between $\mathbf{x}$ and $\mathbf{x}'$.
- A typical example for a correlation function is the gaussian kernel:

$$c(\mathbf{x}, \mathbf{x}') = \exp^{-\theta||\mathbf{x}-\mathbf{x}'||^2}$$

or the product kernel:

$$c(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^{d} \exp^{-\theta_i||x_i-x_i'||^{q_i}}$$

with so-called *hyperparameters* θ_i and q_i, $i = 1, \ldots, d$.

- In geostatistics empirically fitted, polynomial kernels (covariogramm) are common.

Kriging as best linear prediction (BLP)

- One approach of Krige was to find the linear predictor $\lambda_0 + \sum_{i=1}^{n} \lambda_i y^{(i)}$ for the conditional distribution $\mathcal{F}_{\mathbf{x}'|\mathbf{X},\mathbf{y}}$ that minimizes the mean squared error (MSE):

$$\mathrm{MSE}(\mathbf{x}') = \mathbb{E}((\mathcal{F}_{\mathbf{x}'|\mathbf{X},\mathbf{y}} - \lambda_0 + \sum_{i=1}^{n} \lambda_i y^{(i)})^2).$$

- This *best linear predictor (BLP)* is given by weights

$$(\lambda_0^*, \boldsymbol{\lambda}^*) = \arg \min_{(\lambda_0,\boldsymbol{\lambda})\in\mathbb{R}\times\mathbb{R}^m} \mathbb{E}((\mathcal{F}_{\mathbf{x}'|\mathbf{X},\mathbf{y}} - \lambda_0 + \sum_{i=1}^{n} \lambda_i y^{(i)})^2)$$

Derivation of best linear predictor

$$\mathbb{E}(\{\mathcal{F}_{\mathbf{x}'|\mathbf{X},\mathbf{y}} - \lambda_0 - \boldsymbol{\lambda}^T\boldsymbol{\mu}\}^2) = \{\mu - \lambda_0 - \boldsymbol{\lambda}^T\boldsymbol{\mu}\}^2 + \sigma^2 + 2\boldsymbol{\lambda}^T\mathbf{k} + \boldsymbol{\lambda}^T\mathbf{K}\boldsymbol{\lambda}.$$

$(\mathbb{E}(Y^2) = ((\mathbb{E}(Y)))^2 + \mathrm{var}(Y), \quad \mathrm{var}(A+B) = \mathrm{var}(A) + 2\mathrm{cov}(A,B) + \mathrm{var}(B))$

To minimize the term we can always chose $\lambda_0^* = \mu + \boldsymbol{\lambda}^T\boldsymbol{\mu}$ and focus on the second (variance) term. For any $\boldsymbol{\lambda}, \boldsymbol{\nu} \in \mathbb{R}^m$:

$$\mathrm{var}(\mathcal{F}_{\mathbf{x}'|\mathbf{X},\mathbf{y}} - (\boldsymbol{\lambda} + \boldsymbol{\nu})^T\mathbf{y}) \tag{4}$$

$$= \quad \sigma^2 - 2\boldsymbol{\lambda}^T\mathbf{k} + \boldsymbol{\lambda}^T\mathbf{K}\boldsymbol{\lambda} + \boldsymbol{\nu}^T\mathbf{K}\boldsymbol{\nu} + 2(\mathbf{K}\boldsymbol{\lambda} - \mathbf{k})^T\boldsymbol{\nu}. \tag{5}$$

Now, we chose a $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$ with $\mathbf{K}\boldsymbol{\lambda}^* = \mathbf{k}$, which makes the last term zero. Then

$$\mathrm{var}(\mathcal{F}_{\mathbf{x}'|\mathbf{X},\mathbf{y}} - (\boldsymbol{\lambda} + \boldsymbol{\nu})^T\mathbf{y}) = \sigma^2 - 2\boldsymbol{\lambda}^T\mathbf{k} + \boldsymbol{\lambda}^T\mathbf{K}\boldsymbol{\lambda} + \boldsymbol{\nu}^T\mathbf{K}\boldsymbol{\nu}$$

$$\geq \quad \sigma^2 - 2\boldsymbol{\lambda}^T\mathbf{k} + \boldsymbol{\lambda}^T\mathbf{K}\boldsymbol{\lambda}. \tag{6}$$

Now, lower bound 6 is met for $\boldsymbol{\lambda}^* = \mathbf{K}^{-1}\mathbf{k}$, $\boldsymbol{\nu} = 0$.

Summary of best linear prediction

- The BLP is given by $\hat{y} = \lambda_0 + \mathbf{y}^T \mathbf{K}^{-1} \mathbf{k}$
- The computational effort for the first prediction (building the model) is $\mathcal{O}(md^2 + m^3)$.
- For any further prediction the computational effort reduces to $\mathcal{O}(dm)$
- The MSE of the BLP is given by $\sigma_0 - \mathbf{k}^T \mathbf{K}^{-1} \mathbf{k}$ and estimates the local prediction error
- The computational effort for the MSE scales with $\mathcal{O}(dm^2)$

Bayesian interpretation

In case of *gaussian* random fields the conditional distribution $\mathcal{F}_{\mathbf{x}'|\mathbf{X},\mathbf{y}}$ is given by[7]:

$$\mathcal{F}_{\mathbf{x}'|\mathbf{X},\mathbf{y}} \sim N(\lambda_0 + \boldsymbol{\lambda}^T\mathbf{y}, \sigma^2 - \mathbf{k}^T\mathbf{K}^{-1}\mathbf{k})$$

, where

$$\mathbf{K} = \sigma[c(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})]_{(i,j)=\{1,\ldots,m\}^2},$$

$$\mathbf{k} = \sigma[c(\mathbf{x}', \mathbf{x}^{(i)})]_{i=1,\ldots,m},$$

$$\boldsymbol{\lambda} = \mathbf{K}^{-1}\mathbf{k}, \lambda_0 = \mu - \mathbf{k}\mathbf{K}^{-1}\mathbf{y},$$

and $N(\mu, \sigma^2)$ is a 1-D normal distribution with mean μ and variance σ^2.

[7] Rao, C.R. (1973): Linear Statistical Inference and its Applications, second ed., Wiley, NY

Conditional gaussian distribution (derivation sketch)

Consider a multivariate gaussian distribution

$$\text{PDF}(\mathbf{x}) = \frac{1}{(2\pi)^{q/2}\det(\mathbf{K})^{1/2}} \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})).$$

Suppose the normal random vector is partitioned $\mathbf{X} = (\mathbf{X}_1^T\mathbf{X}_2^T)$.
Now, partition accordingly

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \mathbf{K} = \begin{pmatrix} \mathbf{K}_{1,1} & \mathbf{K}_{1,2} \\ \mathbf{K}_{2,1} & \mathbf{K}_{2,2} \end{pmatrix}$$

Then the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2 = \mathbf{x}_2$ is

$$N(\boldsymbol{\mu}_1 + \mathbf{K}_{1,2}\mathbf{K}_{2,2}^{-}(\mathbf{x}-\boldsymbol{\mu}_2), \mathbf{K}_{1,1} - \mathbf{K}_{1,2}\mathbf{K}_{2,2}^{-}\mathbf{K}_{2,1}),$$

where $\mathbf{K}_{2,2}^{-}$ is a generalized inverse of $\mathbf{K}_{2,2}$.

Bayesian approach vs. best linear predicton

- It turns out that in case of gaussian random variables
  - the best linear predictor (BLP) is equal to the conditional mean
  - the mean squared error (MSE) of the BLP is equal to the conditional variance
- For gaussian random fields the BLP is also the best nonlinear predictor
- The BLP derivation holds also for non-gaussian random fields
- In that case the BLP the equivalence to the conditional mean is not always given, and a non-linear predictor may lead to a better MSE[8]

[8] Michael L. Stein: Some theory of Kriging, Springer, 1999

Best Linear Unbiased Estimator (BLUP)

- Suppose we have the following model of a random field

$$\mathcal{F}_{\mathbf{x}} = \mathbf{m}(\mathbf{x})^T \beta + \mathcal{R}(\mathbf{x})$$

  where $\mathcal{R}$ is a random field with mean 0 and known covariance structure
- $\mathbf{m}$ is a known function $\mathbb{R}^d \to \mathbb{R}^p$ and β comprises p unknown coefficients.
- If β were known we could use the BLP:

$$\hat{y}(\mathbf{x}') = \mathbf{m}(\mathbf{x}')^T \beta + \mathbf{k}^T \mathbf{K}^{-1}(\mathbf{y} - \mathbf{m}\beta)$$

- If β is unknown, but all covariances are known, a natural approach is to replace β as the generalized least squares estimator

$$\hat{\beta} = (\mathbf{y}^T \mathbf{K} \mathbf{y})^{-1} \mathbf{y}^T \mathbf{K}^{-1} \mathbf{y}$$

Best Linear Unbiased Estimator (BLUP)

We may also minimize the mean squared error (MSE) among all predictors of the form $\lambda_0 + \lambda^T \mathbf{z}$ subject to the unbiasedness constraint $\mathbb{E}(\lambda_0 + \lambda^T \mathbf{y}) = \mathbb{E}(\mathcal{F}_{\mathbf{x}'})$ for all β, i. e.:

$$\lambda = \arg\min_{\lambda}(\mathbb{E}(\mathcal{F}'_{\mathbf{x}} - \lambda^T \mathbf{y})^2)$$

subject to

$$\lambda_0 = 0 \text{ and } [\mathbf{m}(\mathbf{x}^{(i)})]^T \lambda = \mathbf{m}(\mathbf{x}')$$

The solution we obtain is the same as if replacing the GLS estimator $\hat{\beta}$ for β in the BLP expression. The MSE is given as:

$$\text{MSE}(\mathbf{x}') = \sigma^2 - \mathbf{k}^T \mathbf{K}^{-1} \mathbf{k} + \gamma(\mathbf{y}^T \mathbf{K}^{-1} \mathbf{y})^{-1} \gamma$$

where $\gamma = \mathbf{m}(\mathbf{x}_0) - \mathbf{y}^T \mathbf{K}^{-1} \mathbf{y}$.

Kriging method - summary

The predicted value of kriging is the BLUP and its MSE:

$$\hat{y}(\mathbf{x}') = \mathbf{m}(\mathbf{x}')^T\hat{\beta} + \mathbf{k}^T\mathbf{K}^{-1}(\mathbf{y} - \mathbf{m}\hat{\beta}) \quad (7)$$

$$\hat{\beta} = (\mathbf{y}^T\mathbf{K}\mathbf{y})^{-1}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y} \quad (8)$$

$$\mathrm{MSE}(\mathbf{x}') = \sigma^2 - \mathbf{k}^T\mathbf{K}^{-1}\mathbf{k} + \gamma(\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y})^{-1}\gamma \quad (9)$$

$$\gamma = \mathbf{m}(\mathbf{x}_0) - \mathbf{y}^T\mathbf{K}^{-1}\mathbf{y} \quad (10)$$

Bayesian interpretation of BLUP (not discussed in detail)

- A *Bayesian interpretation* of the BLUP puts a prior on $\beta \sim N(\mu, \sigma^2)$ and then obtains the limiting expression for the conditional distribution $\mathcal{F}_{\mathbf{x}'|\mathbf{X},\mathbf{y}}$ for $\sigma^2 \to \infty$.
- It turns out that

$$\mathcal{F}_{\mathbf{x}'|\mathbf{X},\mathbf{y}} \sim N(\hat{y}(\mathbf{x}'), \mathrm{MSE}(\mathbf{x}'))$$

in case of gaussian random fields.[9].

[9] Omre H. (1987): Bayesian kriging - merging observations and qualified guesses in kriging, Math. Geol. 19 23-39

Three types of Kriging

Given

$$\mathcal{F}_{\mathbf{x}} = \mathbf{m}(\mathbf{x})^T \beta + \mathcal{R}(\mathbf{x})$$

the literature distinguishes three types of Kriging:

- Simple Kriging: $\mathbf{m}(\mathbf{x}) = 1$ and β is given as a prior
- Ordinary Kriging: $\mathbf{m}(\mathbf{x}) = 1$ and β is estimated from the data as $\hat{\beta}$
- Universal Kriging: $\mathbf{m}(\mathbf{x})$ is a vector valued function of $\mathbf{x}$, and $\boldsymbol{\beta}$ is estimated from the data as $\hat{\boldsymbol{\beta}}$.

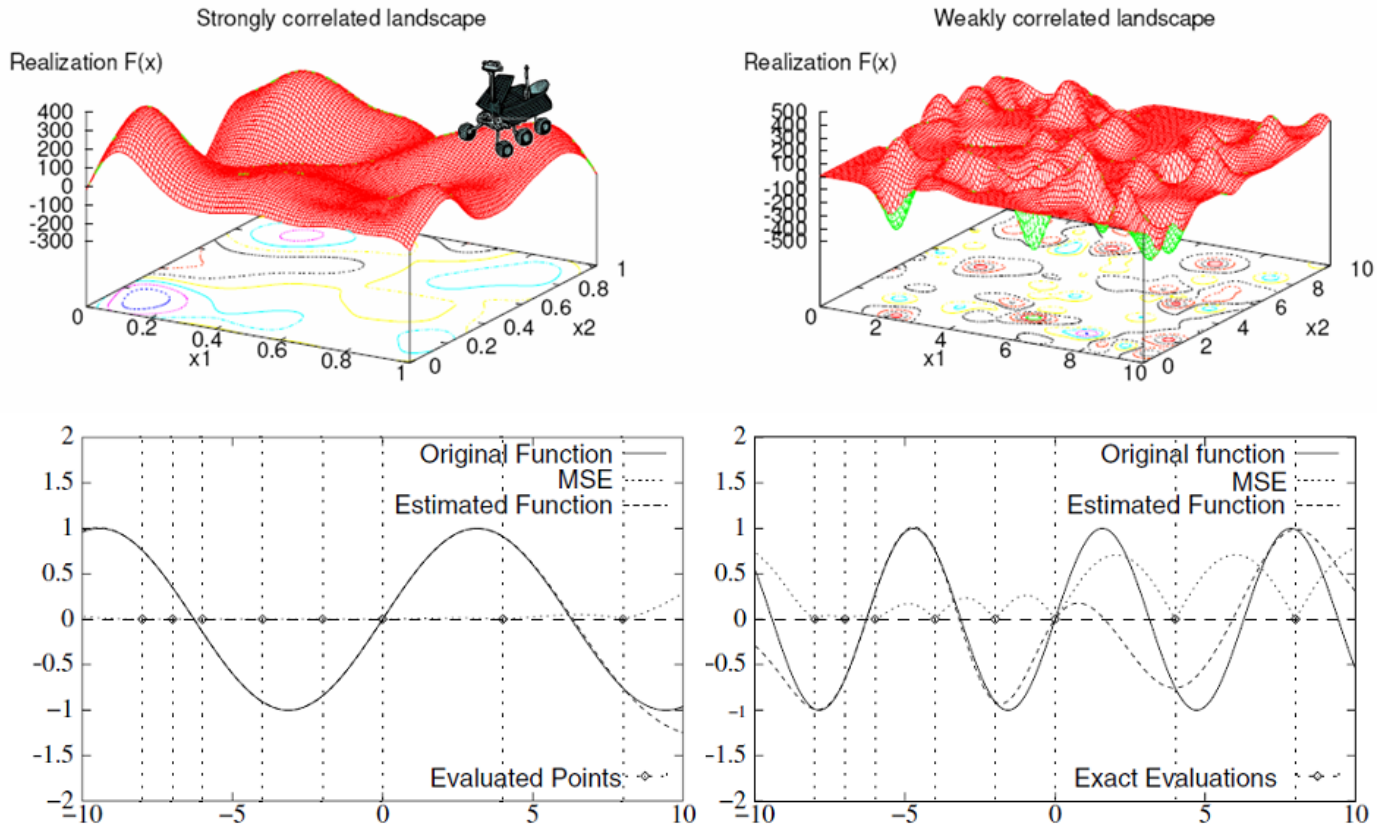
Comparison with regression, splines, and radial basis functions

- Universial Kriging ($\mathbf{m}(\mathbf{x})^T\beta + \mathcal{R}(\mathbf{x})$) corresponds to ordinary *regression models*, in case of zero correlation of $\mathcal{R}$.
- Simple Kriging is formally equivalent to *radial basis functions*, given the centers are chosen at the data points.For a derivation see Emmerich (2005).[10]
- For certain correlation functions, Kriging yields predictors that are equivalent to MARS and/or thin-plate splines[11]

[10] Michael Emmerich: Single- and multiobjective evolutionary design optimization assisted by Gaussian random field metamodels, ELDORADO, TU-Dortmund, 2005

[11] Koehler, J. R. and Owen, A. B., "Computer experiments." In: S. Ghosh and C. R. Rao (eds.), Handbook of Statistics. Elsevier Science, New York, 13, pp. 261-308, 1996.

Estimating the degree of autocorrelation of the landscape



- Error decreases with local density of points.
- The error estimate $\hat{s}$ increases faster in less correlated landscapes. How to estimate correlation parameters?

Estimation of hyperparameters

Suppose the autocorrelation structure, e.g. the hyperparameter θ of $c(\mathbf{x}, \mathbf{x}) = \exp(-\theta||\mathbf{x} - \mathbf{x}'||)$, of the landscape is unknown. The estimation of the hyperparameter can be accomplished by maximum likelihood estimation (MLE). Choose

$$\theta^* = \arg\max_{\theta \in \mathbb{R}^+} L_\theta \text{ with} \tag{11}$$

$$\begin{aligned} L_\theta &= \text{PDF}\{\mathcal{F}_{\mathbf{x}^{(1)}} = y^{(1)}, \mathcal{F}_{\mathbf{x}^{(2)}} = y^{(2)}, \ldots, \mathcal{F}_{\mathbf{x}^{(m)}} = y^{(m)}\} \\ &= \frac{1}{(2\pi)^{n/2}(\hat{\sigma}^2)^{n/2}\sqrt{\det(\mathbf{K}/\hat{\sigma})}} \exp\left[-\frac{(\mathbf{y} - \mathbf{1}\hat{\beta})^T(\frac{1}{\hat{\sigma}}\mathbf{K})^{-1}(\mathbf{y} - \mathbf{1}\hat{\beta})}{2\hat{\sigma}^2}\right] \end{aligned}$$

Estimation of hyperparameters

As L_θ is strictly positive we may instead maximize its logarithm and neglect some positive constants:

$$\theta^* = \arg\max_{\theta \in \mathbb{R}^+} L_\theta =$$

$$\arg\max_{\theta \in \mathbb{R}^+} \frac{1}{\hat{\beta}\log(\hat{s}) + \log\det(\mathbf{K}/\hat{\sigma})} =$$

$$\arg\min_{\theta \in \mathbb{R}^+} \hat{\beta}\log(\hat{\sigma}) + \log\det(\mathbf{K}/\hat{\sigma}).$$

Note, that L_{θ^*} is a relative indicator for how well the data is consistent with the gaussian random field interpretation.

Cross-validation

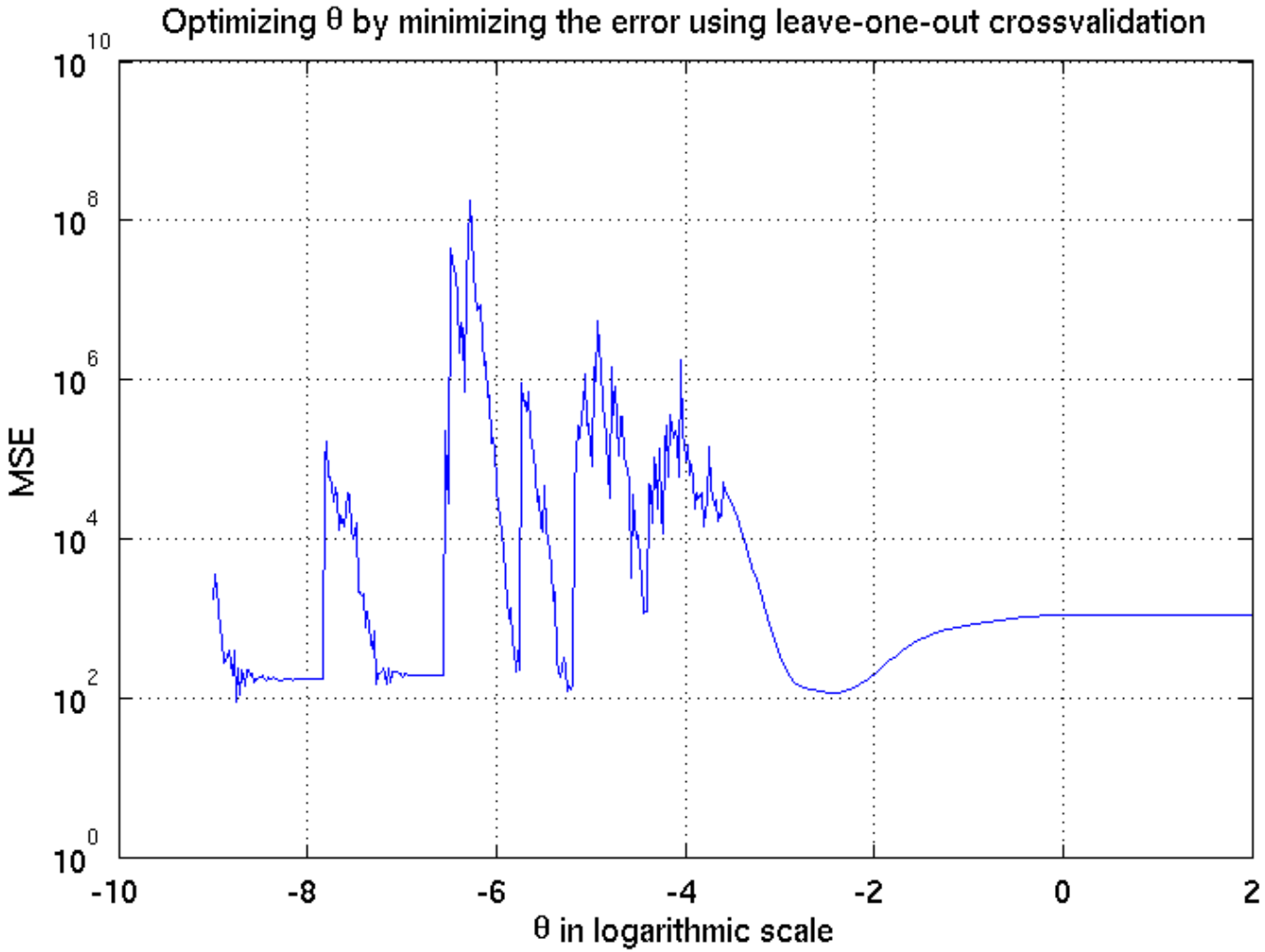
- cross-validation can be used as an alternative to maximum likelihood estimation
- for instance leave-one-out-cross validation minimizes the following root mean squared cross validation error:

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^+} \sqrt{\sum_{i=1}^{k} (y^{(i)} - \mathbb{E}(\mathcal{F}_{\mathbf{x}^{(i)}|\mathbf{X}\setminus\mathbf{x}^{(i)},\mathbf{y}\setminus y^{(i)}}))^2}$$

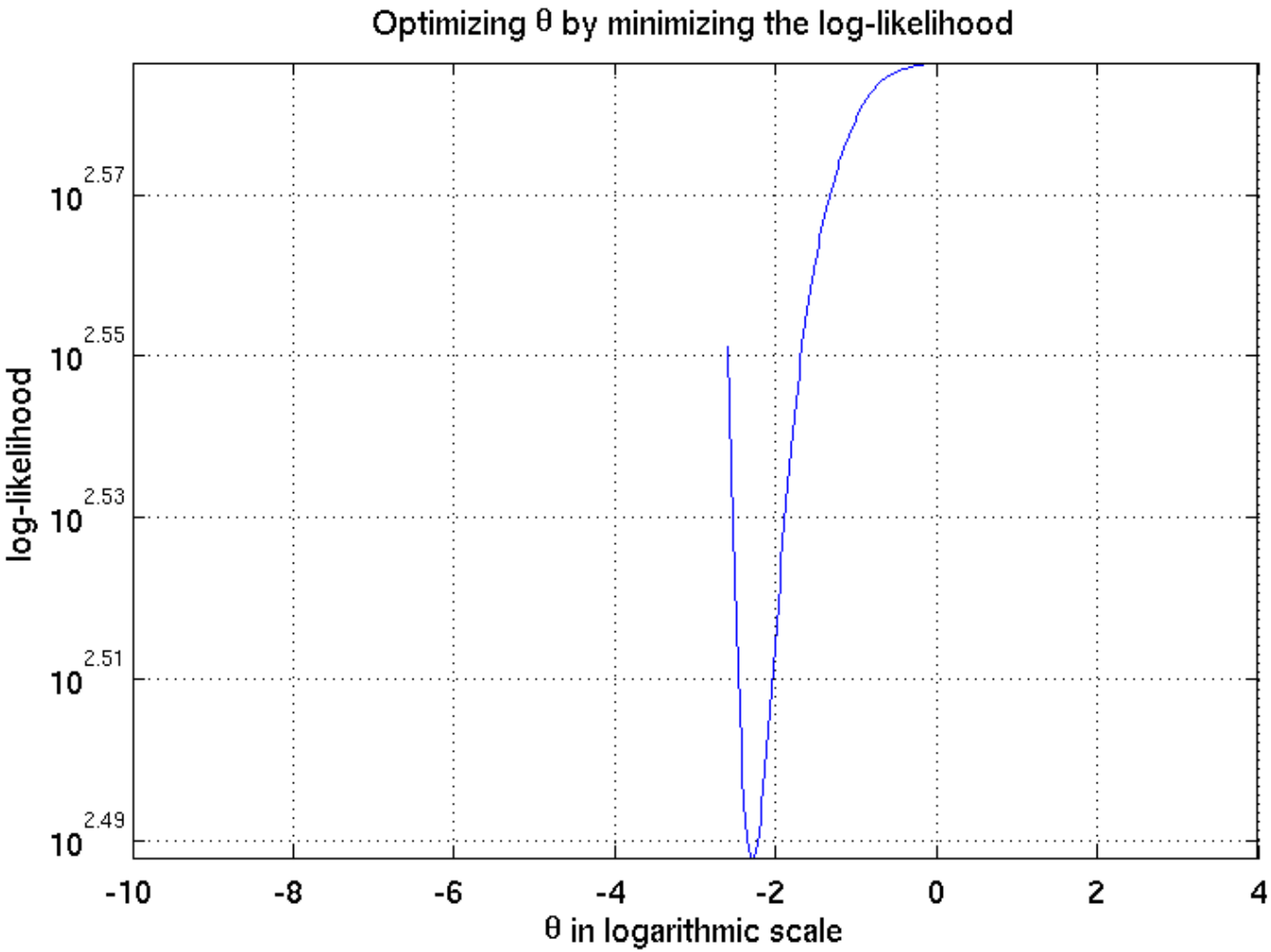
where $\mathbf{X} \setminus \mathbf{x}^{(i)}$ and $y \setminus y^{(i)}$ denotes the matrix after the i-th column or vector component gets deleted.

- a disadvantage of cross-validation could be, that it is more biased towards over-represented regions

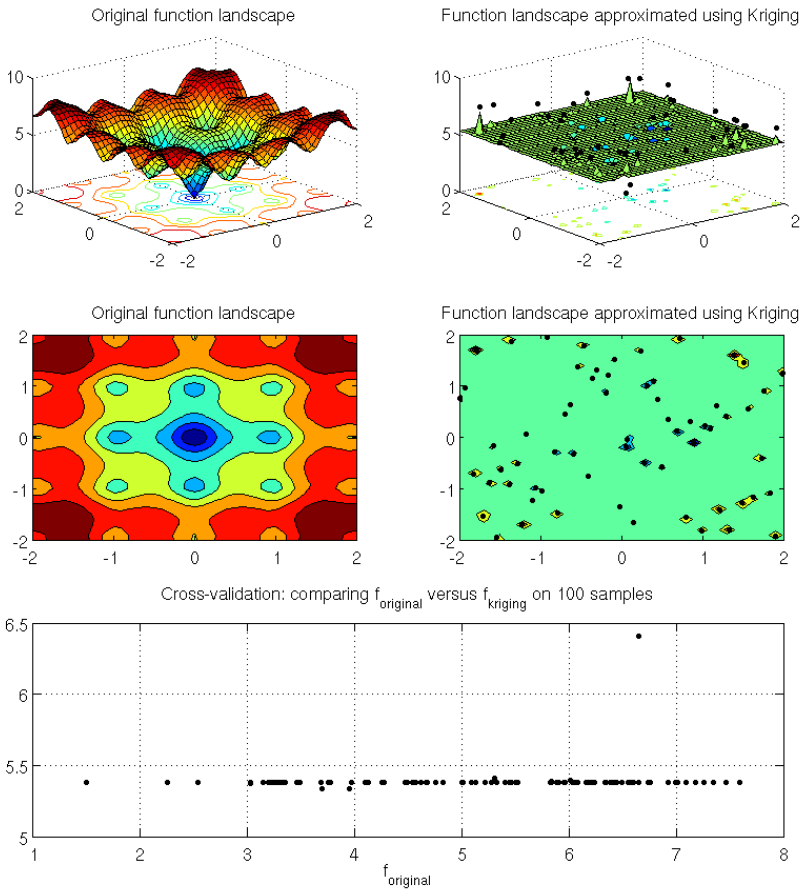
Leave-one-out cross validation error

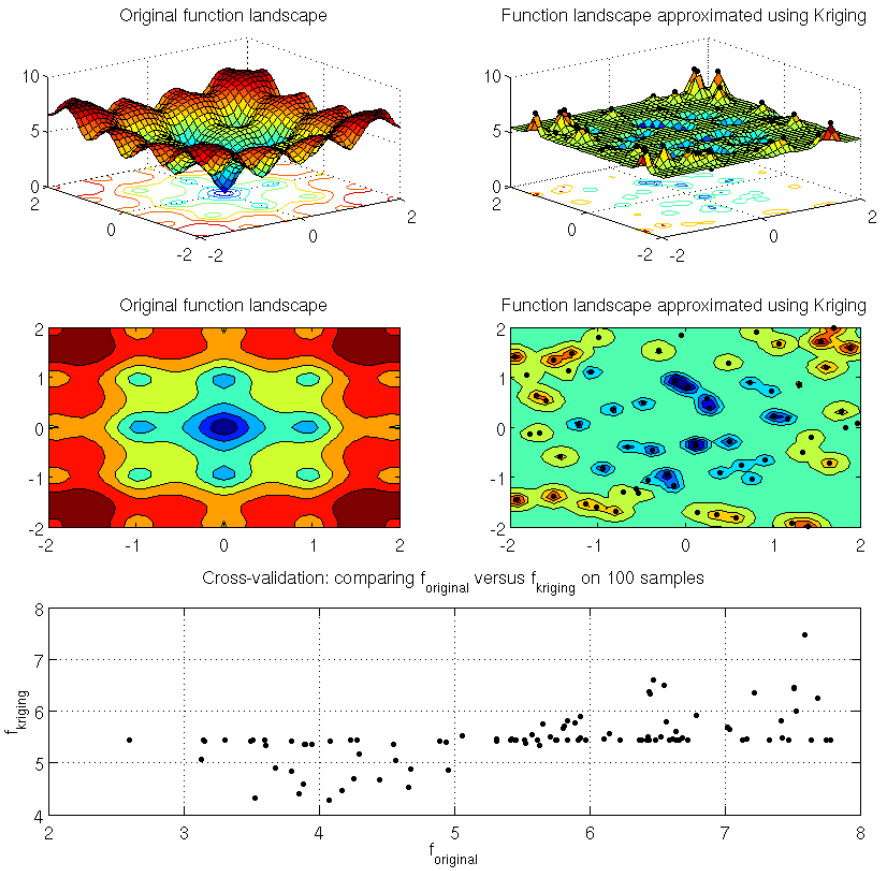


Likelihood for different θ

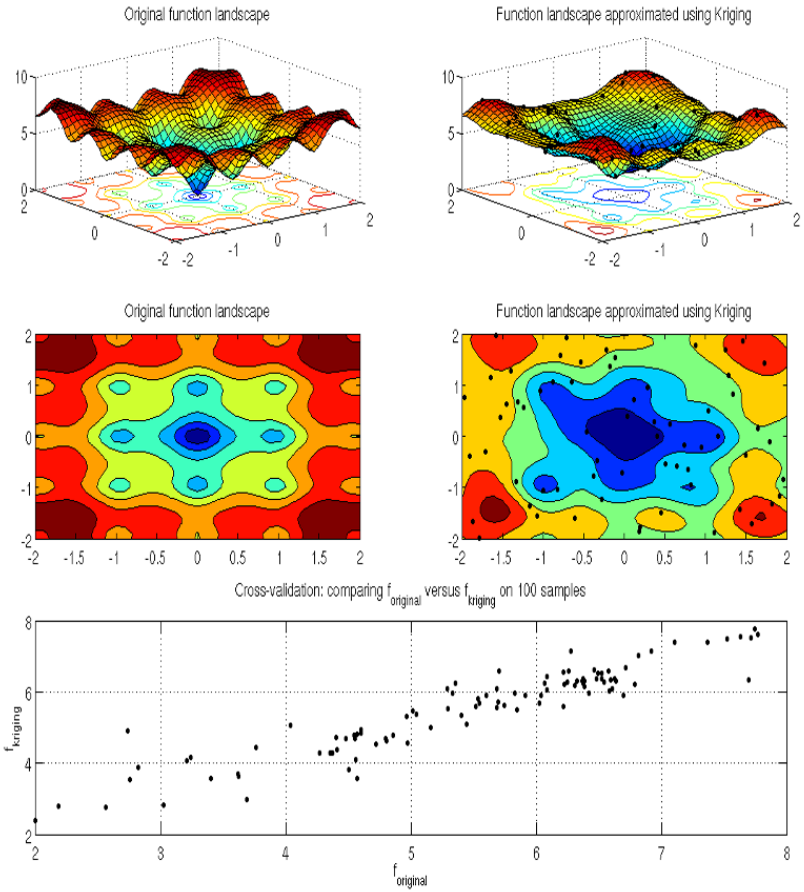


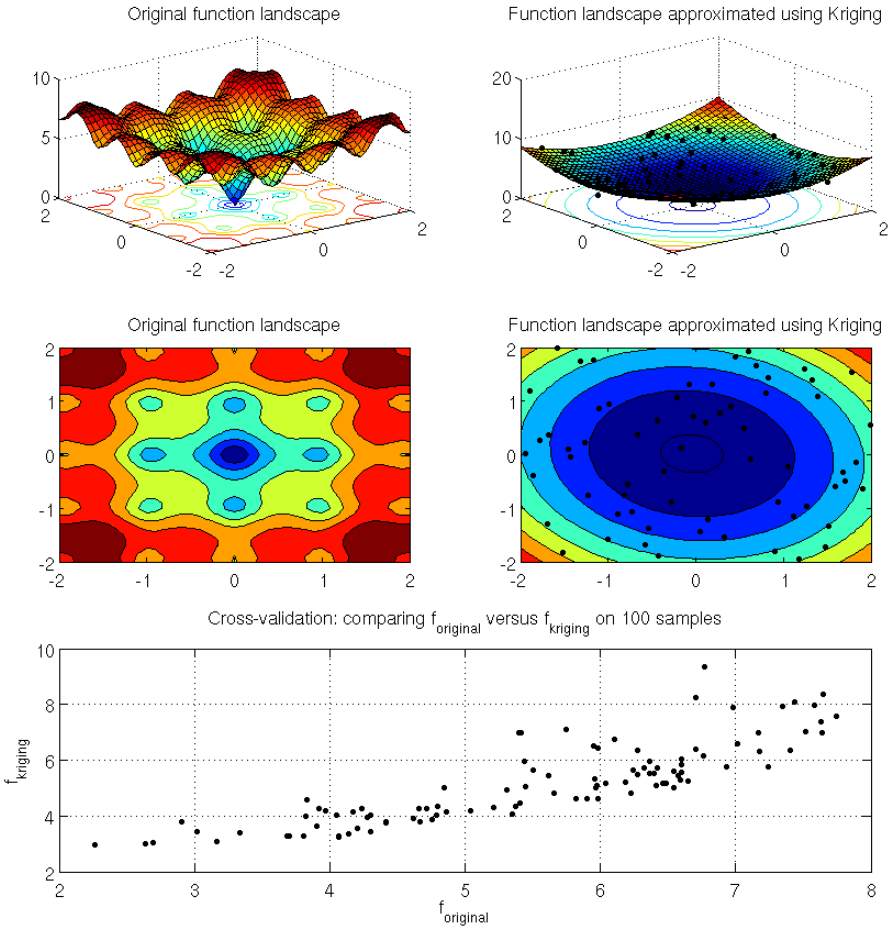
Very high and high θ



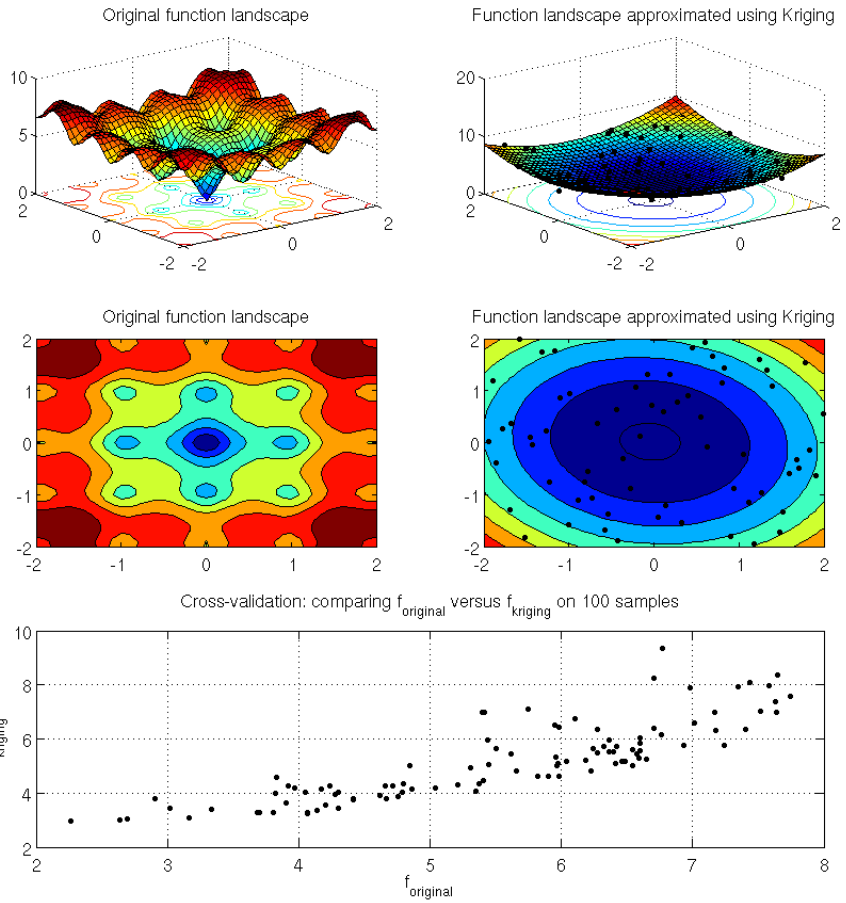


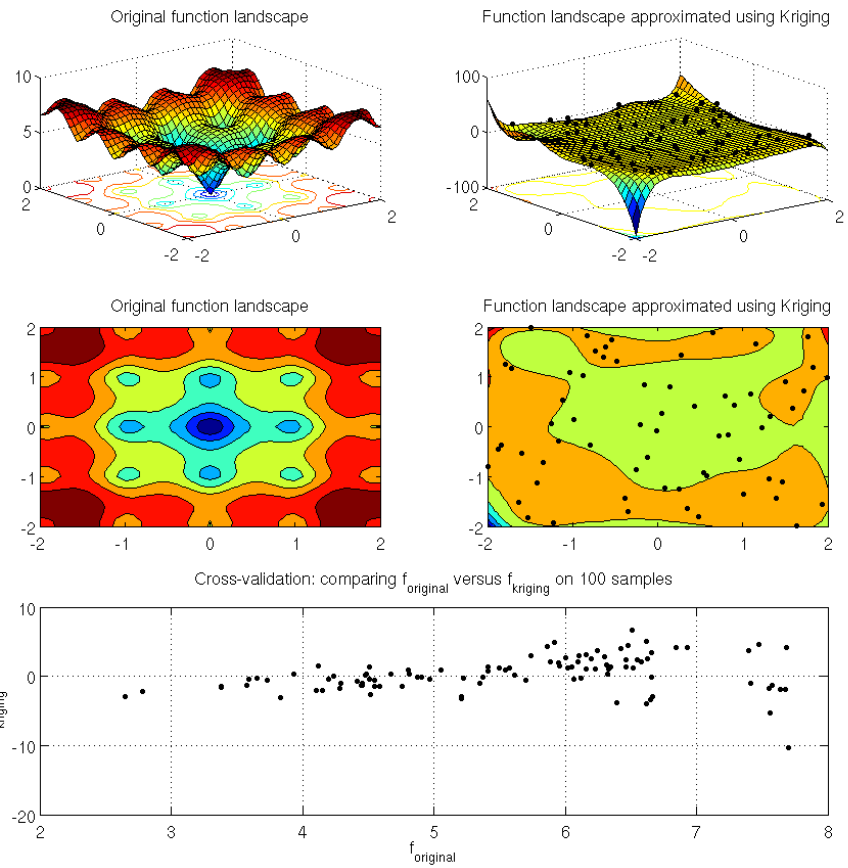
Medium and lower θ





Low and very low θ





Exploiting the Hessian

- Given we know the (approximate) Hessian matrix $\mathbf{H} \in \mathbb{R}^{d \times d}$ of a global quadratic trend (e.g. from FOCAL algorithm)
- we may then use it to adapt the metric in the correlation function[12]:
- For this measure the distance by means of the Mahalanobis distance:

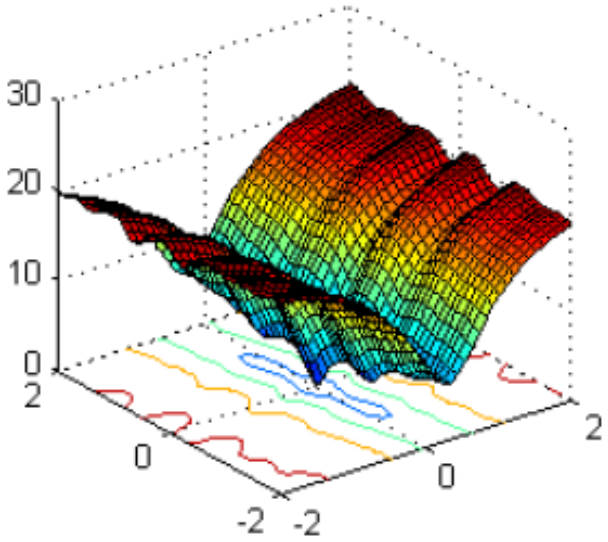
$$c(\mathbf{x}, \mathbf{x}) = \exp^{\theta M_{\mathbf{H}}(\mathbf{x}, \mathbf{x}')} \text{ with}$$

$$M_H = \sqrt{(\mathbf{x} - \mathbf{x}')^T \mathbf{H}^{-1} (\mathbf{x} - \mathbf{x}')}$$

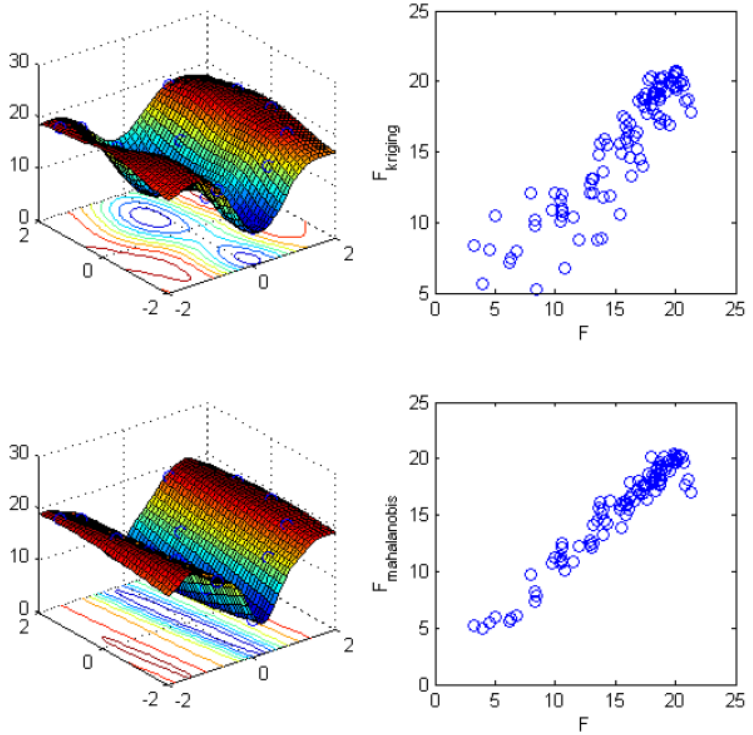
(Mahalanobis distance)

[12] Johannes Kruisselbrink, Michael Emmerich, Thomas Baeck: A Robust Optimization Approach Using Kriging Metamodels for Robustness Approximation in the CMA-ES, CEC 2010, Barcelona, IEEE-Press, 2010.

Exploiting the Hessian



Ackley function multiplied by matrix $f(\mathbf{H}^{-1}\mathbf{x})$.



Ordinary (top) and mahalanobis based (bottom) kriging.

Applications in optimization

- Robustness approximation in EA (Kruisselbrink et al. 2010)[13]
- Preselection in evolutionary algorithms (Emmerich et al. 2005)
- Optimization on the metamodel
  - Bayesian global optimization (e.g. Mockus, Mockus, and Mockus)[14]
  - Efficient global optimization (e.g. Jones et al. 1998)[15]

[13] Johannes Kruisselbrink, Michael Emmerich, Thomas Baeck: A Robust Optimization Approach Using Kriging Metamodels for Robustness Approximation in the CMA-ES, CEC 2010, Barcelona, IEEE-Press, 2010.

[14] A. Mockus, J. Mockus, and L. Mockus. Bayesian approach in stochastic and heuristic methods of global and discrete optimization. In Abstracts, Second World Meeting, International Society for Bayesian Analysis, pages 1011, Alicante, Spain, 1994.

[15] Jones, Schonlau, Welch: Efficient Global Optimization of Expensive Black-Box Functions,JOGO, 1994

Bayesian global optimization

$D_0 \leftarrow ((\mathbf{x}^{(1)}, f(\mathbf{x}^{(1)}), \ldots, \mathbf{x}^{(m_0)}, f(\mathbf{x}^{(0)})$ {Initialize database}
$t \leftarrow m_0$ {Initialize evaluation counter}
while $t < t_{eval,max}$ **do**
  Search for $\mathbf{x}_t^* = \text{argmin}_{\mathbf{x} \in \mathbb{S}} \hat{y}(D_t, \mathbf{x}) - \omega \hat{s}(D_t, \mathbf{x})$
  $y_t = f(\mathbf{x}_t^*)$
  if $y_t < y_{\text{min}}^t$ **then**
    $\mathbf{x}_{\text{min}}^t = \mathbf{x}_t^*$
    $y_{\text{min}}^t = y_t$
  end if
  $D_{t+1} = D_t \cup \{(\mathbf{x}_t^*, y_t)\}$
end while
return $y_{\text{min}}^t, \mathbf{x}_{\text{min}}^t$

Bayesian global optimization

Different criteria for the utility criterion have been suggested for pre-screening the search space by means of the metamodel (line 4). They all more or less refer to the trade-off already described by Kushner (1962)[16]:

> *The purpose of the utility function is to find trade-off between sampling in known promising regions versus sampling in under-explored regions or regions where the variation in function values is high.*

Prominent examples are $f_{sc} = \hat{y} - \omega\hat{s}$ and the expected improvement

$$\mathbb{E}(I(\mathcal{F}_{\mathbf{x}'|\mathbf{X},\mathbf{y}})), \quad I(y) = \max\{0, f_{min} - y\}$$

[16] H. J. Kushner. A versatile stochastic model of a function of unknown and time varying form. Journal of Math. Anal. Appl., pages 150167, 5 1962.

Metamodel-assisted optimization

- A global maximizer of the expected improvement can be found by using α Branch and Bound[17]
- Efficient global optimization (EGO) algorithm applies this[18]
- Generalizations of the expected improvement for hypervolume-based multiobjective optimization are compared by Wagner et al.[19]
- An overview on the use of metamodels in single-and multiobjective evolutionary optimization see Emmerich, 2005[20]

[17] Floudas, C.A., Deterministic Global Optimization, Kluwer, 2000

[18] Jones, Schonlau, Welch: Efficient Global Optimization of Expensive Black-Box Functions,JOGO, 1994

[19] Wagner, Emmerich, Deutz, Ponweiser: PPSN 2011

[20] Michael Emmerich: Single- and multiobjective evolutionary design optimization assisted by Gaussian random field metamodels, Dissertation, FB Informatik, Technical University Dortmund, Germany

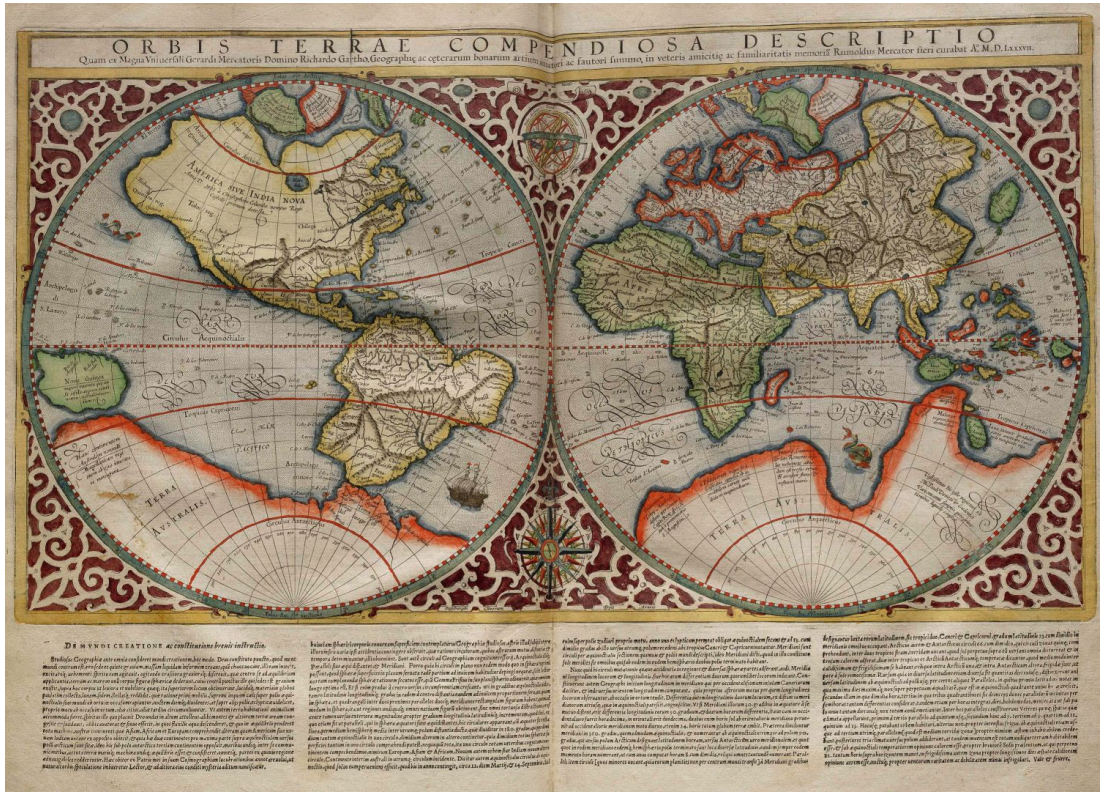
Conclusion

- Kriging is a versatile method for N-D interpolation with error assessment
- Different interpretations of predictors (Bayesian, linear prediction)
- Estimation of autocorrelation structure by maximum likelihood or cross validation
- Parameter estimation (model calibration) is main computational effort, predictions can be made in linear time
- Used in optimization with time consuming computer experimens

Outlook

- Estimation of correlation parameters and right choice of correlation function deserves further attention
- How does the point distribution influence the prediction quality?
- Interpretation of correlation parameters offers an interesting perspective in landscape analysis
- Finding critical points on prediction function can be useful for building optimization methods, in particular in the multiobjective case
- Kriging for non-standard search spaces (vector valued, discrete, stochastically perturbed functions)

Questions?



World map of Geradus Mercator, 17th century

Ackknowlegments:
Johannes Kruisselbrink (experimental results, MATLAB implementation of N-D Kriging)

Expected Improvement

- The expected improvement[21], measures for a given predictive distribution $\mathrm{PDF}_{\mathbf{x}}$ of an input point x, the expected gain in closeness to the global maximum:

Definition (Expected Improvement)

The expected improvement is defined as:

$$\mathrm{EI}(x, f_{min}) = \int_{y=-\infty}^{\infty} \mathrm{I}(y) \cdot \mathrm{PDF}_{\mathbf{x}}(y) dy$$

where $\mathrm{I}(y) = \min(y - f_{best})$ is the improvement with respect to f_{best}, the best found solution so far.

[21] Mockus, J.; Tiesis, V. and Zilinskas, A. (1978), The application of Bayesian methods for seeking the extremum, North-Holland.

Alternatve Criteria

- Mean Value [Giannakoglou et al.]

$$\hat{y}(\mathbf{x})$$

- Lower Confidence Bound (LBC_ω)

$$\hat{y}(\mathbf{x}) - \omega \cdot \hat{s}(\mathbf{x}), \omega > 0$$

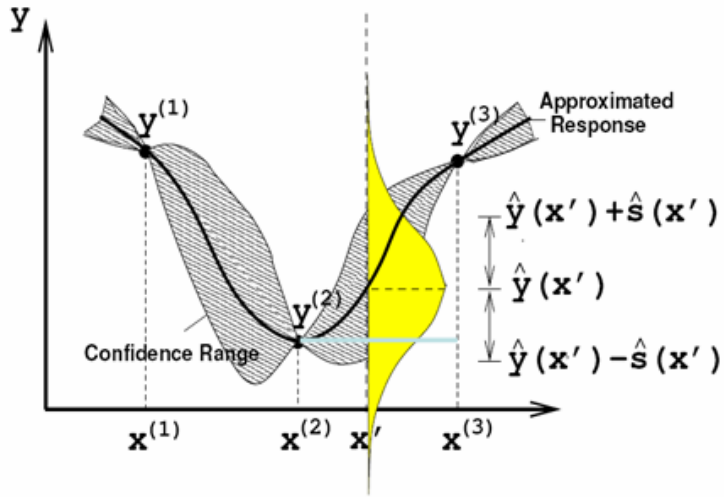
- Probability of Improvement (PoI)

$$\int_{-\infty}^{f_{best}^t} \mathrm{PDF}_{\mathbf{x}}(y)dy = \Phi(\frac{f_{best}^t - \hat{y}(\mathbf{x})}{\hat{s}(\mathbf{x})})$$

- Expected Improvement (ExI)

$$\int_{-\infty}^{f_{best}^t} \mathrm{I}(y) \cdot \varphi(\frac{f_{best}^t - \hat{y}(\mathbf{x})}{\hat{s}(\mathbf{x})})dy$$

$$\mathrm{I}(y(\mathbf{x})) = \max\{f_{best}^t - y(\mathbf{x}), 0\}$$



Multiobjective optimization

Multi-criterion problems $f_1(\mathbf{x}) \to \min, \ldots, f_m(\mathbf{x}) \to \min$, $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^n$, and the f_i are real-valued. Foremost we consider the cases $m = 2$ and $m = 3$.

Definition (Efficient Set, Pareto front)

The efficient set of a problem is the set of all solutions in $\mathcal{X}$ that are not dominated. Its image under $\mathbf{f}$ is the Pareto front.

Definition (Approximation set)

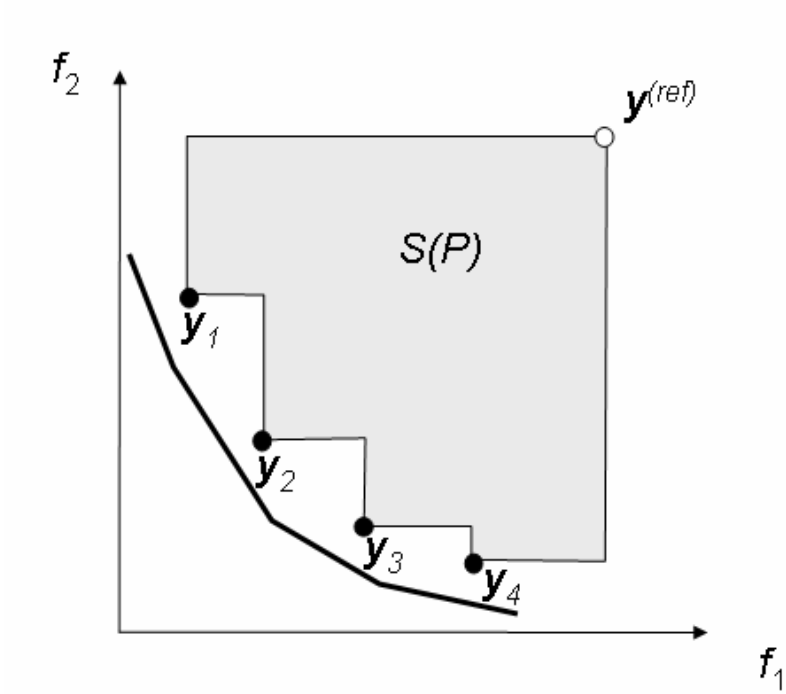
By $\prec$ we denote Pareto-dominance. Let $\mathcal{A}_m$ be defined as the set of all finite sets $\mathrm{A} \subset \mathbb{R}^m$ with $\forall \mathbf{a} \in \mathrm{A} : \nexists \mathbf{a}' \in \mathrm{A} : \mathbf{a}' \prec \mathbf{a}$. The elements of $\mathcal{A}_m$ are called *approximation sets*.

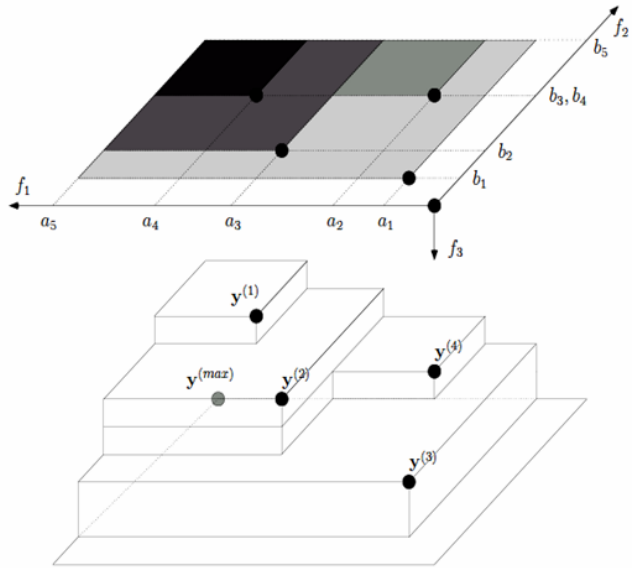
Hypervolume Indicator

The hypervolume indicator $\mathcal{H}$ (or: hypervolume) of an approximation set A is defined as the volume of the dominated subspace, bounded by a reference point $\mathbf{r}$ (a point which is dominated by each element of A):

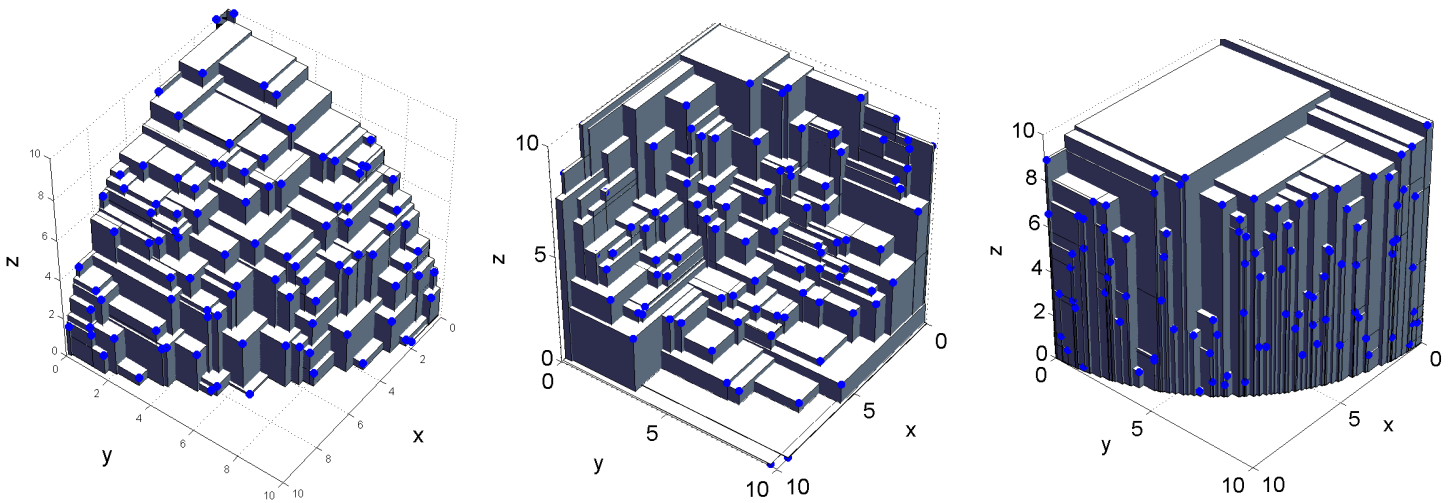
$$\mathcal{H}(\mathrm{A}) =$$
$$\mathrm{Vol}(\{\mathbf{y} \in \mathbb{R}^m | \mathbf{y} \text{ is dominated by some } \mathbf{y} \in \mathrm{A} \text{ and } \mathbf{y} \prec \mathbf{r}\})$$

Hypervolume examples





Hypervolume examples

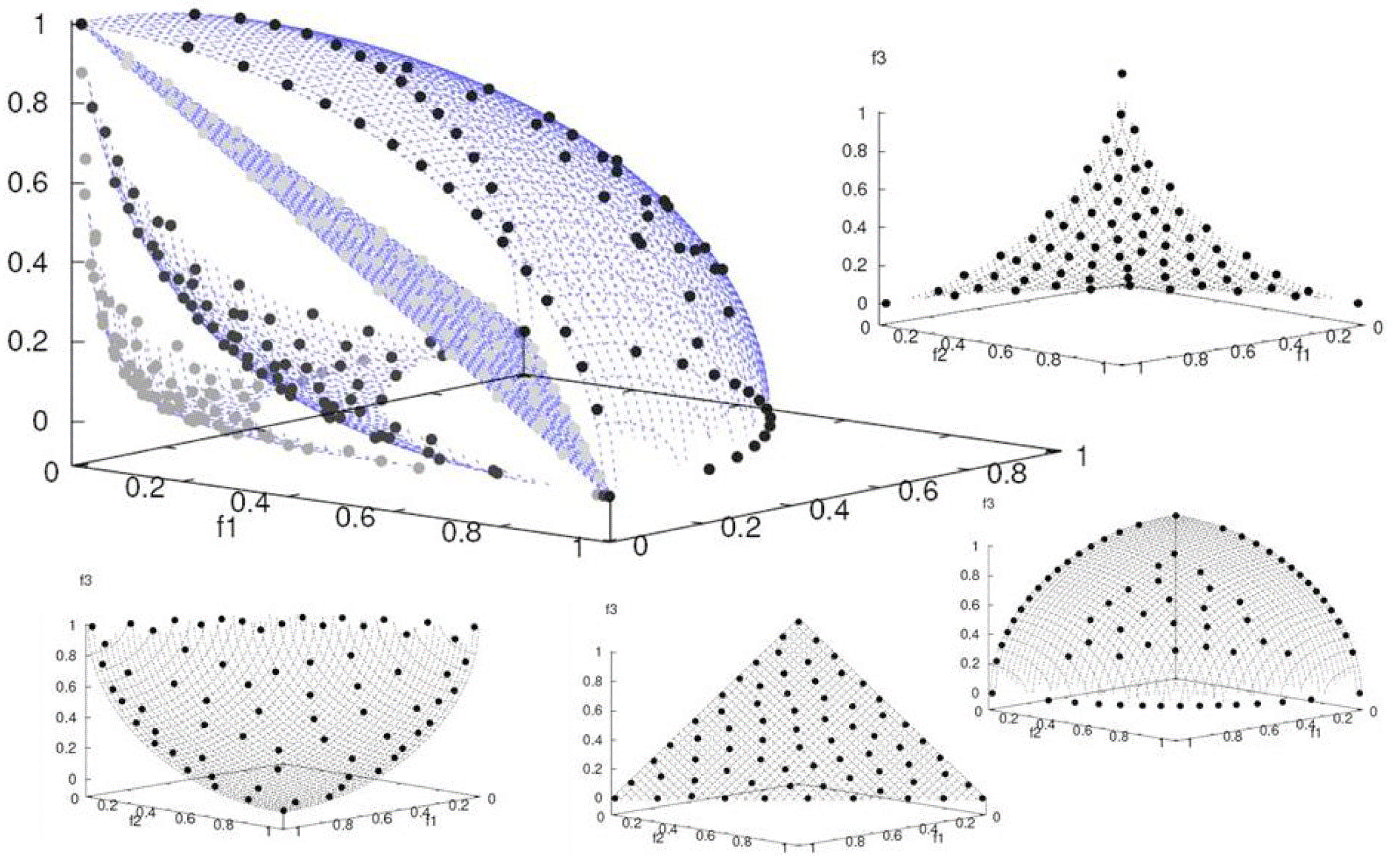


Complexity: 2-D and 3-D $\Theta(n \log n)$. 4-D $\mathcal{O}(n^2)$ (cf. [Guerreira, Fonseca, Emmerich; Canad. Conf. Comp. Geom. 2012])

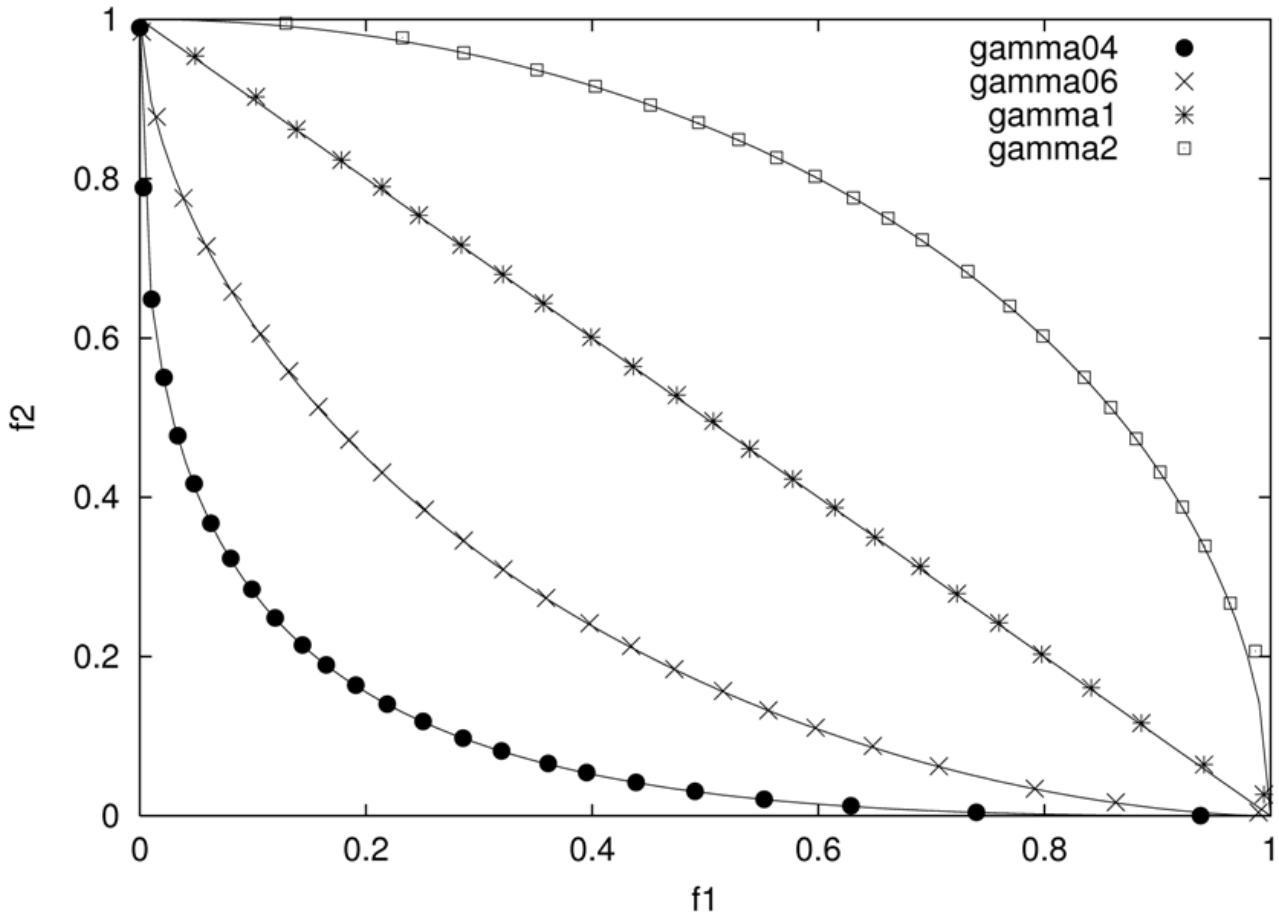
Hypervolume maximal distributions

$$\{(y_1, \ldots, y_m) \in \mathbb{R}_+^m \mid y_1^\gamma + \cdots + y_m^\gamma = 1\}$$

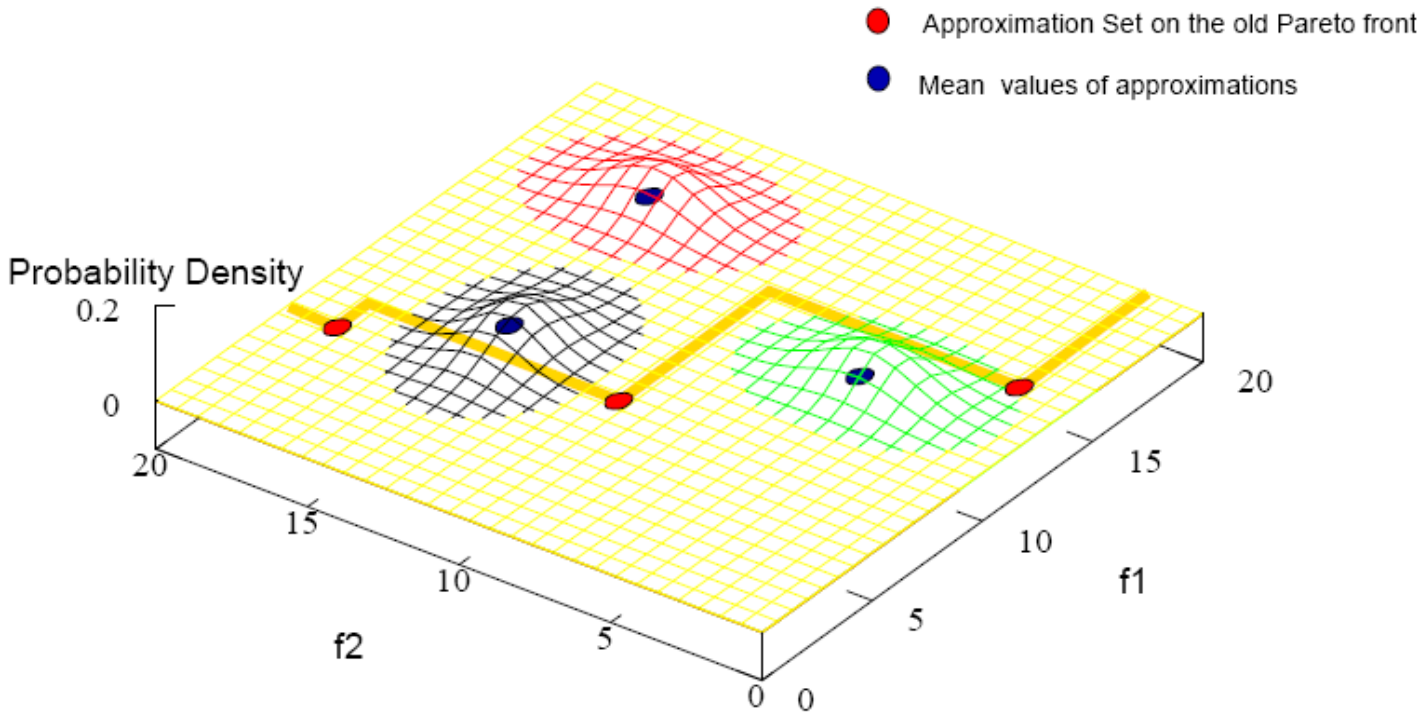
$\gamma = 0.4$
$\gamma = 0.6$
$\gamma = 1$
$\gamma = 2$



Hypervolume maximal distributions



Multiobjective Expected Improvement



Generalization to Multiobjective Optimization

Recently the expected improvement was generalized to multiobjective optimization in different ways:

Emmerich 2005 Expected Improvement of Hypervolume Indicator

Keane 2006 Momentum of Gaussian truncated over non-dominated space.

Knowles 2006 Expected Improvement of Tschebyscheff Scalarization (ParEGO).

Obayashi and Jeong 2007 Vector of single-objective expected improvements.

Ponweiser, Wagner, and Vinzce 2008 Upper confidence bound for increment of hypervolume (SMS-EGO).

Zhang et al. 2010

For an overview see Wagner, Emmerich and Deutz, PPSN 2011.

We will focus on the expected increment (improvement) in hypervolume[22]:

- direct computation procedure for the expected improvement in hypervolume
- new theorems on fundamental properties of the expected improvement in hypervolume, in particular monotonicity properties related to the variance of the predictive distribution
- first result on the behavior of the hypervolume-based expected improvement in the multiobjective generalization of Bayesian (or 'efficient') global optimization.

[22] Emmerich,M., Single- and Multiobjective Evolutionary Design Optimization Using Gaussian Random Field Metamodels, PhD Thesis, FB Informatik, TU Dortmund, 2005

Hypervolume Based Expected Improvement

Definition (Hypervolume-based Improvement function)

The hypervolume-based improvement function $\mathsf{I} : \mathbb{R}^m \times \mathcal{A}_m \to \mathbb{R}$ is defined as

$$\mathsf{I}(\mathbf{y}, \mathrm{A}) = \mathcal{H}(\mathrm{A} \cup \{\mathbf{y}\}) - \mathcal{H}(\mathrm{A}) \tag{12}$$

. Where $\mathbf{y} \in \mathbb{R}^m$ and $\mathrm{A} \in \mathcal{A}_m$.

Definition (Expected improvement)

The expected improvement at a point $\mathbf{x}$ with respect to the approximation set A, denoted by $\mathsf{EI}_{\mathcal{H}}(\mathbf{x}, \mathrm{A})$, is defined as follows.

$$\mathsf{EI}_{\mathcal{H}}(\mathbf{x}, \mathrm{A}) = \int_R \mathsf{I}(\mathbf{y}, \mathrm{A}) \cdot \mathsf{PDF}_{\mathbf{x}}(\mathbf{y}) d\mathbf{y}, \tag{13}$$

where the integration region R is $\mathbb{R}^m$.

Monotonicity of $\mathrm{EI}_{\mathcal{H}}$

Wagner et al.[23] stated desirable monotonicity properties for EI:

Definition

Let $\mathrm{A} \in \mathcal{A}_m$ denote an Pareto front approximation set:

- N1 If for two points $\mathbf{x}^+$ and $\mathbf{x}$ in the search space the predictive variances are the same, and for predictive means $\mathbf{y}^+$ and $\mathbf{y}$ it holds that $\mathbf{y}^+ > \mathbf{y}$, then $\mathrm{EI}_{\mathcal{H}}(\mathbf{x}^+, \mathrm{A}) > \mathrm{EI}_{\mathcal{H}}(\mathbf{x}, \mathrm{A})$.
- N2 If for two points $\mathbf{x}^+$ and $\mathbf{x}$ in the search space the predicted means are the same, and for the standard deviations $\mathbf{s}^+$ and $\mathbf{s}$ it holds that $\mathbf{s}^+ > \mathbf{s}$, then $\mathrm{EI}_{\mathcal{H}}(\mathbf{x}^+, \mathrm{A}) > \mathrm{EI}_{\mathcal{H}}(\mathbf{x}, \mathrm{A})$.

N1 was proven correct, for N2 empirical evidence was provided but an analytical proof for $m > 1$ remained open.

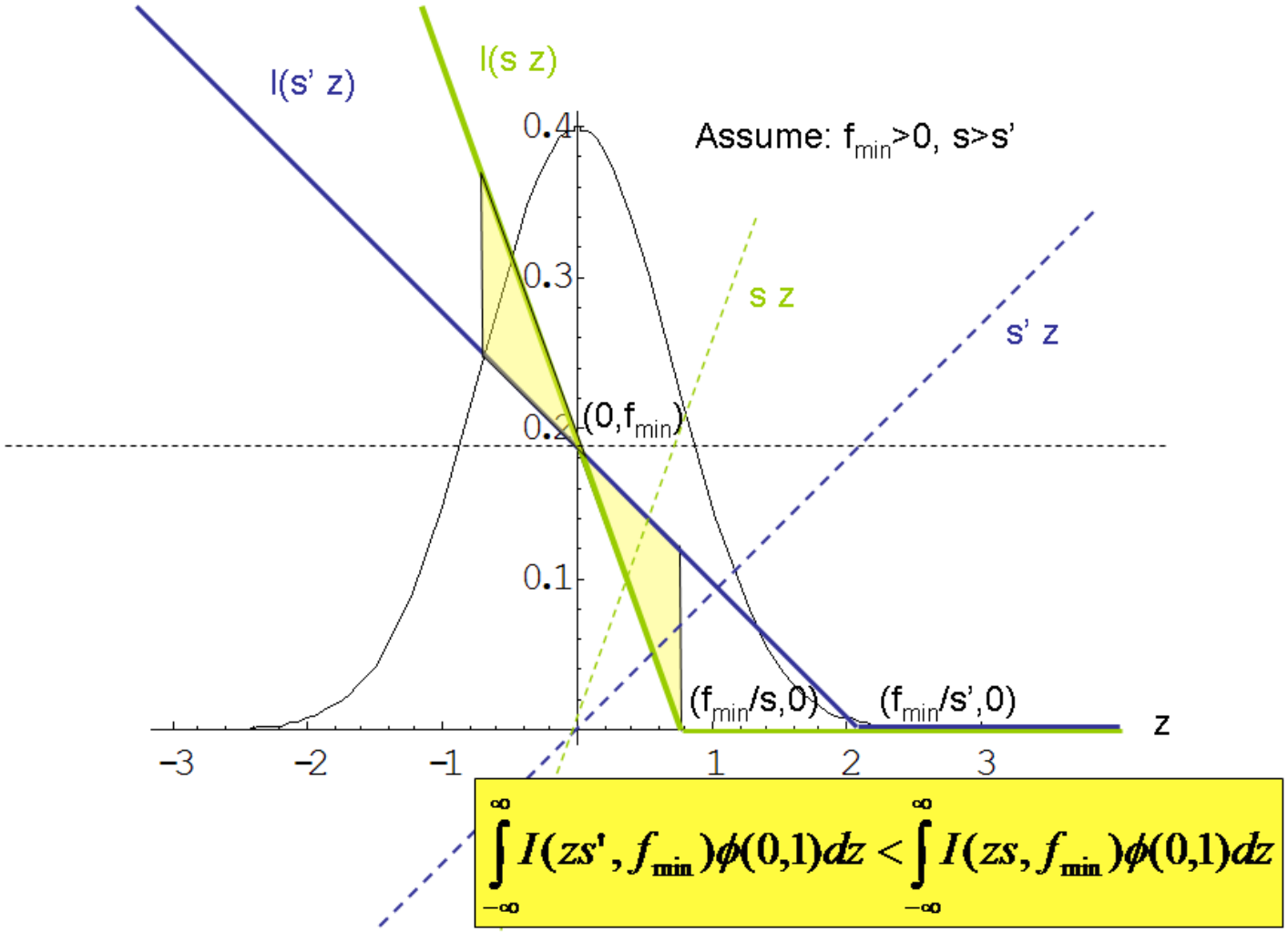
[23] Wagner, T.; Emmerich, M.; Deutz, A. and Ponweiser, W. (2010), On expected-improvement criteria for model-based multi-objective optimization, PPSN 2010, Springer.

Lemma

Given standard deviations s and s' and an arbitrary constant a, it holds:
If $s > s'$ then

$$\int_{-\infty}^{\infty} I(y, \{a\}) PDF_{0,s}(y) dy > \int_{-\infty}^{\infty} I(y, \{a\}) PDF_{0,s'}(y) dy \quad (14)$$

where $I(y, \{a\}) = \max\{0, a - y\}$.


I(s' z)
I(s z)
Assume: f_{min}>0, s>s'
s z
s' z
0.4
0.3
0.2
0.1
$(0,f_{min})$
$(f_{min}/s,0)$
$(f_{min}/s',0)$
z
−3
−2
−1
1
2
3
$\int_{-\infty}^{\infty} I(zs', f_{min})\phi(0,1)dz < \int_{-\infty}^{\infty} I(zs, f_{min})\phi(0,1)dz$

Algorithm: $\mathrm{EI}_{\mathcal{H}}$-based global multi-objective optimization

Generate initial sequence of points $X_k = (\mathbf{x}^{(1)}, ..., \mathbf{x}^{(k)})$

Evaluate initial points $Y_k = (\mathbf{y}^{(1)} = \mathbf{f}(\mathbf{x}^{(1)}), \ldots, \mathbf{y}^{(k)} = \mathbf{f}(\mathbf{x}^{(k)}))$

Set X_k^{nd}, Y_k^{nd} to the subsequence of non-dominated solutions among X_k, Y_k.

$t \leftarrow k$

while $t < t_{max}$ **do**

  $t \leftarrow t + 1$

  choose $\mathbf{x}^{(t)} \in \arg\min_{\mathbf{x} \in S} \mathrm{EI}_{\mathcal{H}}(\mathbf{x}, X_{t-1}, Y_{t-1}, Y_t^{nd})$

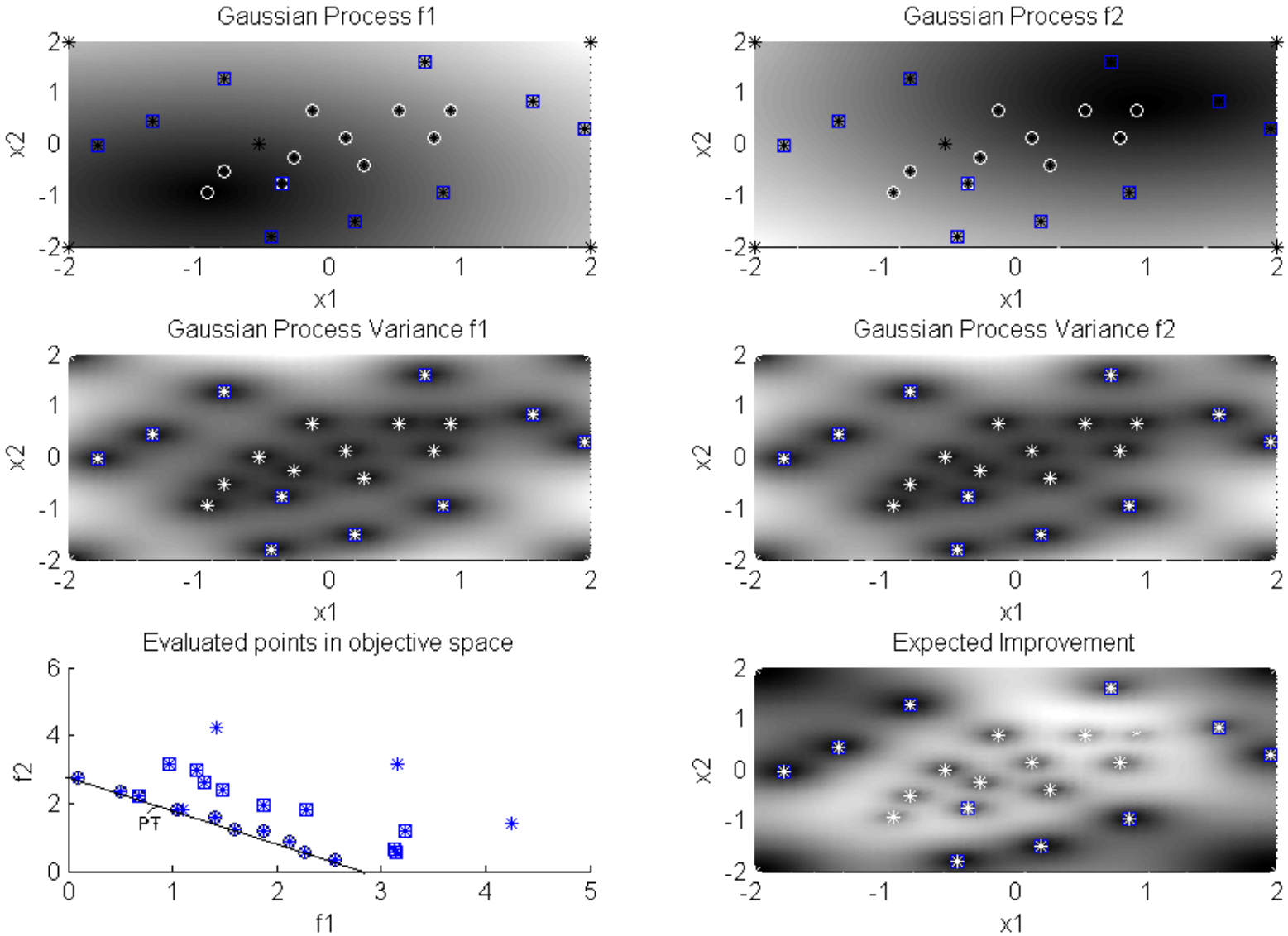
  $\mathbf{y}^{(t)} = \mathbf{f}(\mathbf{x}^{(t)})$

  $X_t = X_{t-1} \circ \mathbf{x}^{(t)}$; $Y_t = Y_{t-1} \circ \mathbf{y}^{(t)}$

  Set X_t^{nd}, Y_t^{nd} to the sequence of non-dominated solutions among X_t, Y_t.

end while

return X_t^{nd}, Y_t^{nd}



25 Evaluations, 10 Initial Points $\mathbb{R}^2 \rightarrow \mathbb{R}^2$

Conclusions

- The $\text{EI}_{\mathcal{H}}$ is compatible extension of the EI for multiple objectives.
- Like the single objective $\text{EI}_{\mathcal{H}}$ it is monotonous in its mean and variance.
- For positive variance it obtains strictly positive and finite values;
- For a multivariate Gaussian distribution with zero covariances the integral can be directly computed; (MATLAB Source code is made available at `http://natcomp.liacs.nl`
- EI guides Bayesian global optimization or MA-ES.
- In the example the Pareto front of a $\mathbb{R}^2 \to \mathbb{R}^2$ mapping was obtained in 25 evaluations.