**EngOpt 2012 - International Conference on Engineering Optimization**
Rio de Janeiro, Brazil, 1-5 July 2012.

# The asymptotic behaviour of the maximum likelihood function of Kriging approximations using the Gaussian correlation function

**Schalk Kok**

Advanced Mathematical Modelling, CSIR Modelling and Digital Science, Pretoria, 0001.
email: skok@csir.co.za

**Abstract**

This study reports on the asymptotic behavior of the maximum likelihood function, encountered when constructing Kriging approximations using the Gaussian correlation function. Of specific interest is a maximum likelihood function that decreases continuously as the correlation function hyper-parameters approach zero. Since the global minimizer of the maximum likelihood function is an asymptote in this case, it is unclear if maximum likelihood estimation (MLE) remains valid. Numerical ill-conditioning of the correlation matrix also occurs in this case. Analytical and numerical examples are presented that demonstrates the validity of MLE, provided that arbitrary precision arithmetic is used. A recent result that claims the MLE function always approaches infinity as the hyper-parameters approach zero is also disproved.

**Keywords:** Kriging, Maximum Likelihood Estimation, Gaussian correlation function, ill-conditioning.

## 1 Introduction

Kriging is a spatial data interpolation method developed by a South African mining engineer, D.G. Krige, in 1951 [1]. Kriging is used to construct a best linear unbiased predictor based on sampled data. In the context of optimization, Kriging is used to construct surrogate models, especially when optimization problems require expensive simulations using the finite element method (FEM) or computational fluid dynamics (CFD). The surrogate model is then optimized, rather than the original expensive problem.

In constructing a Kriging approximation, a spatial correlation function has to be chosen. This paper only considers the Gaussian correlation function. Although one of the most popular, it presents some notorious numerical challenges. A problem that occurs rather frequently when Kriging is used to approximate the response of computer experiments, is that the correlation function hyper-parameters cannot be determined robustly. The preferred technique to find the hyper-parameters is to solve the maximum likelihood optimization problem (assumed to be posed as a minimization problem in this paper). However, if the maximum likelihood estimation (MLE) function contains no local minimizer, the hyper-parameters need to be assigned values based on other criteria.

The case where the MLE function decreases monotonously as the hyper-parameters approach zero, or the case where the MLE function decreases monotonously as the hyper-parameters approach infinity, presents two specific cases where the MLE function contains no local minimizer. The behaviour of the MLE function as the hyper-parameters approach zero or infinity was considered by Zimmerman [2], and is referred to as the asymptotic behavior of the MLE function. The same problem is considered in this paper, but the main result of Zimmermann [2] is disproved.

## 2 Kriging fundamentals

A response $y(\boldsymbol{x})$ is considered to consist of a deterministic contribution $f(\boldsymbol{x})$ and a stochastic component $Z(\boldsymbol{x})$, i.e.

$$y(\boldsymbol{x}) = f(\boldsymbol{x}) + Z(\boldsymbol{x}). \tag{1}$$

The deterministic contribution $f(\boldsymbol{x})$ is often represented by a low order polynomial, with a constant trend being one of the most popular choices. The stochastic contribution $Z(\boldsymbol{x})$ is taken as a function with zero mean and covariance

$$\mathrm{Cov}[Z(\boldsymbol{x}^i), Z(\boldsymbol{x}^j)] = \sigma^2 \boldsymbol{\mathcal{R}}(R(\boldsymbol{x}^i, \boldsymbol{x}^j)), \tag{2}$$

where $\sigma^2$ is the process variance, $\boldsymbol{\mathcal{R}}$ is the correlation matrix and $R(\boldsymbol{x}^i, \boldsymbol{x}^j)$ is the correlation function between data points $\boldsymbol{x}^i$ and $\boldsymbol{x}^j$. The number of observations (data points) is equal to $n$, hence the correlation matrix is of size $n \times n$. The $(i, j)$ entry of the correlation matrix is given by

$$\mathcal{R}_{i,j} = R(\boldsymbol{x}^i, \boldsymbol{x}^j), \tag{3}$$

where a specific form of the correlation function $R$ still has to be selected. The resuling correlation matrix must be positive definite (all eigenvalues are positive) and is symmetric by definition. In computer experiment applications, the Gaussian correlation function is particularly popular. In this case, $R$ is given by

$$R(\boldsymbol{x}^i, \boldsymbol{x}^j) = \prod_{k=1}^{m} e^{-\theta_k |x_k^i - x_k^j|^2}, \tag{4}$$

where $m$ is the number of design variables (i.e. the dimension of the vector $\boldsymbol{x}$) and $\theta_k$ are adjustable hyper-parameters that parametrizes the correlation function. The adjustable hyper-parameters $\theta_k$ are often determined using Maximum Likelihood Estimation (MLE) [3], requiring the solution of a minimization problem:

$$\text{Minimize } \Phi(\boldsymbol{\theta}) = \frac{n \ln(\tilde{\sigma}^2(\boldsymbol{\theta})) + \ln(\det(\boldsymbol{\mathcal{R}}(\boldsymbol{\theta})))}{2}, \text{ subject to } \theta_k > 0, k = 1, 2, \ldots, m. \tag{5}$$

Here $\tilde{\sigma}^2(\boldsymbol{\theta})$ is an estimate of the variance, given by

$$\tilde{\sigma}^2(\boldsymbol{\theta}) = \frac{(\boldsymbol{y} - \boldsymbol{f}B(\boldsymbol{\theta}))^{\mathrm{T}} \boldsymbol{\mathcal{R}}^{-1}(\boldsymbol{\theta})(\boldsymbol{y} - \boldsymbol{f}B(\boldsymbol{\theta}))}{n}, \tag{6}$$

where $\boldsymbol{y}$ contains the available responses at the $n$ data points, $\boldsymbol{f}$ is a column vector of ones (since a constant trend $B$ is used) and $B(\boldsymbol{\theta})$ is given by

$$B(\boldsymbol{\theta}) = (\boldsymbol{f}^{\mathrm{T}} \boldsymbol{\mathcal{R}}^{-1}(\boldsymbol{\theta}) \boldsymbol{f})^{-1} \boldsymbol{f}^{\mathrm{T}} \boldsymbol{\mathcal{R}}^{-1}(\boldsymbol{\theta}) \boldsymbol{y}. \tag{7}$$

Finally, the estimated response $\tilde{y}(\boldsymbol{x})$ at a point $\boldsymbol{x}$ is given by

$$\tilde{y}(\boldsymbol{x}) = B + \boldsymbol{r}^{\mathrm{T}}(\boldsymbol{x}) \boldsymbol{\mathcal{R}}^{-1}(\boldsymbol{y} - \boldsymbol{f}B), \tag{8}$$

where $\boldsymbol{r}(\boldsymbol{x})$ is the correlation function vector between the point $\boldsymbol{x}$ and all the data points $\boldsymbol{x}^i$, $i = 1, 2, \ldots, n$. The terms $\boldsymbol{\mathcal{R}}^{-1}(\boldsymbol{y} - \boldsymbol{f}B)$ can be considered a weighting vector $\boldsymbol{w}$, with the estimated response given by a weighted sum of the correlation function vector:

$$\tilde{y}(\boldsymbol{x}) = B + \boldsymbol{r}^{\mathrm{T}}(\boldsymbol{x}) \boldsymbol{w}. \tag{9}$$

# 3 Difficulties in solving the MLE problem

If the hyper-parameters approach zero, the correlation matrix $\boldsymbol{\mathcal{R}}$ approaches the unit matrix (a matrix consisting of all ones, not be be confused with the identity matrix). The eigenvalues of an $n \times n$ unit matrix are $n$ (once) and zero ($n$-1 times). Hence, the condition number of $\boldsymbol{\mathcal{R}}$ approaches infinity as the hyper-parameters approach zero. This behaviour of the conditioning of $\boldsymbol{\mathcal{R}}$ as $\theta_k$ approaches zero is well-known, but the behaviour of the MLE function $\Phi(\boldsymbol{\theta})$ is of more interest. Zimmermann [2] claims to present analytical proof that "...eventually, when the condition number approaches infinity, so does the associated likelihood function". This proof however relies on a number of assumptions, which guarentees that the limit $\lim_{\|\boldsymbol{\theta}\| \to 0} B(\boldsymbol{\theta})$ exists. This paper will present an analytical example that disproves Zimmermann's result. A numerical example using arbitrary precision arithmetic provides additional experimental evidence.

The other case of interest is if the hyper-parameters approach infinity. The correlation matrix now becomes the identity matrix, with eigenvalue 1 (repeated $n$ times). The condition number is unity. No numerical problems are experienced in this case, but the resulting Kriging approximations reproduce

the deterministic contribution $f(\boldsymbol{x})$ and peak to interpolate the available data only at the locations of the observed response. A non-trivial example is presented by Zimmermann [2] which indicates $\theta_k \to \infty$. Conventional maximum likelihood estimation fails in this case, since no local minimum exists. In this paper it is demonstrated experimentally that problems of this type typically occur if very few sampling points are available. By adding sampling points, the MLE functions typically change behaviour such that local minimizers exist. Eventually, after adding a substantial number of sampling points, the MLE functions may change behaviour to the other extreme, where the hyper-parameters approach zero.

# 4 Examples

## 4.1 Analytical example

This first example constructs a Kriging fit to the function $y = x^2$, using $n$ evenly spaced points from 0 to 1. Matlab is used to compute the MLE function $\Phi(\theta)$ for $\theta$ varying between $10^{-4}$ and $10^3$, using Eq.(5). For $n = 5$ and 6, the condition number of the correlation matrix $\boldsymbol{\mathcal{R}}$ exceeds $10^{16}$ for small $\theta$, rendering the computations inaccurate. To circumvent this problem, analytical expressions for $\Phi(\theta)$ and $B(\theta)$ are found using the Matlab symbolic toolbox, for $n = 2$ to 6. The limits as $\theta$ approaches zero and infinity are also calculated analytically and included below.

$n = 2$

$$\Phi(\theta) = -2\,\ln(2) - \ln\left(1 - e^{-\theta}\right) + 1/2\,\ln\left(1 - e^{-2\,\theta}\right) \tag{10}$$

$$\lim_{\theta\to\infty} \Phi(\theta) = -2\ln(2) \tag{11}$$

$$\lim_{\theta\to 0} \Phi(\theta) = \infty \tag{12}$$

$$B = 1/2 \tag{13}$$

$\Phi(\theta)$ is a monotonically decreasing function as $\theta$ increases. Consequently, the MLE of $\theta$ is infinity (or some large upper bound). $\Phi(\theta)$ approaches infinity as $\theta$ approaches zero.

$n = 3$

$$\begin{aligned}\Phi(\theta) = -9/2\,\ln(2) - 3/2\,\ln(3) + 3/2\,\ln\left(13 - 3\,e^{-3/4\,\theta} - 3\,e^{-1/2\,\theta} - 3\,e^{-1/4\,\theta}\right) - \\ 3/2\ln\left(\left(-1 + e^{-\theta}\right)\left(-1 + e^{-1/4\,\theta}\right)\left(e^{-1/2\,\theta} + 2\,e^{-1/4\,\theta} + 3\right)\right) \\ + 1/2\,\ln\left(1 - 2\,e^{-1/2\,\theta} + 2\,e^{-3/2\,\theta} - e^{-2\,\theta}\right)\end{aligned} \tag{14}$$

$$\lim_{\theta\to\infty} \Phi(\theta) = -9/2\ln(2) - 3\ln(3) + 3/2\ln(13) \tag{15}$$

$$\lim_{\theta\to 0} \Phi(\theta) = \infty \tag{16}$$

$$B(\theta) = \frac{e^{-3/4\,\theta} + e^{-1/2\,\theta} + e^{-1/4\,\theta} - 5}{4\left(-1 + e^{-1/4\,\theta}\right)\left(e^{-1/2\,\theta} + 2\,e^{-1/4\,\theta} + 3\right)} \tag{17}$$

$\Phi(\theta)$ contains two critical points, a local minimizer at $\theta \approx 2.5881$ (marked with a red cross in Figure 1 (a)) and a local maximizer at $\theta \approx 9.1049$. $\Phi(\theta)$ and $B(\theta)$ approach infinity as $\theta$ approaches zero.

$n = 4$

$$\Phi(\theta) = -6\,\ln(2) - 8\,\ln(3) + 2\,\ln\Big(-9\,e^{-\frac{11}{9}\,\theta} - 27\,e^{-\frac{10}{9}\,\theta} - 54\,e^{-\theta} - 90\,e^{-\frac{8}{9}\,\theta} - 144e^{-\frac{7}{9}\,\theta} - 144\,e^{-2/3\,\theta} - 90\,e^{-5/9\,\theta} + 18\,e^{-4/9\,\theta} + 61\,e^{-1/3\,\theta} + 167\,e^{-2/9\,\theta} + 212\,e^{-1/9\,\theta} + 196\Big) - 2\ln\Big(e^{-4/3\,\theta} + 2\,e^{-\frac{11}{9}\,\theta} + 2\,e^{-\frac{10}{9}\,\theta} + e^{-\theta} - 2\,e^{-\frac{7}{9}\,\theta} - 4\,e^{-2/3\,\theta} - 4\,e^{-5/9\,\theta} - 3\,e^{-4/9\,\theta} + 2\,e^{-2/9\,\theta} + 3\,e^{-1/9\,\theta} + 2\Big) + 1/2\,\ln\Big(1 - 3\,e^{-2/9\,\theta} + 4\,e^{-2/3\,\theta} - 2\,e^{-\frac{8}{9}\,\theta} + e^{-4/9\,\theta} - 2\,e^{-4/3\,\theta} - 2\,e^{-\frac{10}{9}\,\theta} + 4\,e^{-\frac{14}{9}\,\theta} + e^{-\frac{16}{9}\,\theta} - 3\,e^{-2\,\theta} + e^{-\frac{20}{9}\,\theta}\Big) \quad (18)$$

$$\lim_{\theta\to\infty} \Phi(\theta) = -4\ln(2) - 8\ln(3) + 4\ln(7) \quad (19)$$

$$\lim_{\theta\to 0} \Phi(\theta) = \infty \quad (20)$$

$$B(\theta) = \frac{5\,e^{-5/9\,\theta} + 5\,e^{-1/9\,\theta} - 14}{18\left(e^{-1/9\,\theta} - 1\right)\left(e^{-4/9\,\theta} + e^{-1/3\,\theta} + e^{-2/9\,\theta} + e^{-1/9\,\theta} + 2\right)} \quad (21)$$

$\Phi(\theta)$ contains one critical point, a local minimizer at $\theta \approx 0.27659$ (marked with a green cross in Figure 1 (a)). $\Phi(\theta)$ and $B(\theta)$ approach infinity as $\theta$ approaches zero.

$n = 5$

$$\Phi(\theta) = -\frac{35}{2}\,\ln(2) - 5/2\,\ln(5) + 5/2\,\ln\Big(-\Big(15\,e^{-\frac{19}{16}\,\theta} + 60\,e^{-\frac{9}{8}\,\theta} + 150\,e^{-\frac{17}{16}\,\theta} + 251\,e^{-\theta} + 408\,e^{-\frac{15}{16}\,\theta} + 602\,e^{-\frac{7}{8}\,\theta} + 814\,e^{-\frac{13}{16}\,\theta} + 920\,e^{-3/4\,\theta} + 924\,e^{-\frac{11}{16}\,\theta} + 702\,e^{-5/8\,\theta} + 458\,e^{-\frac{9}{16}\,\theta} - 44\,e^{-1/2\,\theta} - 620\,e^{-\frac{7}{16}\,\theta} - 1214\,e^{-3/8\,\theta} - 1442\,e^{-\frac{5}{16}\,\theta} - 1660\,e^{-1/4\,\theta} - 1407\,e^{-3/16\,\theta} - 1118\,e^{-1/8\,\theta} - 724\,e^{-1/16\,\theta} - 435\Big)\Big) - 5/2\ln\Big(e^{-\frac{17}{16}\,\theta} + e^{-\theta} + e^{-\frac{15}{16}\,\theta} + e^{-\frac{7}{8}\,\theta} - e^{-\frac{11}{16}\,\theta} - e^{-5/8\,\theta} - 2\,e^{-\frac{9}{16}\,\theta} - 2\,e^{-1/2\,\theta} - e^{-\frac{7}{16}\,\theta} - e^{-3/8\,\theta} + e^{-3/16\,\theta} + e^{-1/8\,\theta} + e^{-1/16\,\theta} + 1\Big) - 5/2\ln\Big(e^{-3/8\,\theta} + 3\,e^{-\frac{5}{16}\,\theta} + 5\,e^{-1/4\,\theta} + 4\,e^{-3/16\,\theta} + 5\,e^{-1/8\,\theta} + 7\,e^{-1/16\,\theta} + 5\Big) + 1/2\,\ln\Big(1 + e^{-5/2\,\theta} + 3\,e^{-9/4\,\theta} + 3\,e^{-1/4\,\theta} + 6\,e^{-\theta} - 7\,e^{-1/2\,\theta} - 7\,e^{-2\,\theta} - 4\,e^{-1/8\,\theta} - 2\,e^{-5/8\,\theta} + 2\,e^{-5/4\,\theta} - 10\,e^{-\frac{9}{8}\,\theta} + 6\,e^{-3/8\,\theta} + 10\,e^{-\frac{7}{8}\,\theta} - 4\,e^{-3/4\,\theta} + 10\,e^{-\frac{13}{8}\,\theta} + 6\,e^{-3/2\,\theta} - 10\,e^{-\frac{11}{8}\,\theta} - 2\,e^{-\frac{15}{8}\,\theta} + 6\,e^{-\frac{17}{8}\,\theta} - 4\,e^{-\frac{19}{8}\,\theta} - 4\,e^{-7/4\,\theta}\Big) \quad (22)$$

$$\lim_{\theta\to\infty} \Phi(\theta) = -35/2\ln(2) - 5/2\ln(5) + 5/2\ln(3) + 5/2\ln(29) \quad (23)$$

$$\lim_{\theta\to 0} \Phi(\theta) = -15\ln(2) - 5/2\ln(5) + 5/2\ln(7) - 3/2\ln(3) \quad (24)$$

$$B(\theta) = 1/8\,\Big(2\,e^{-\frac{9}{16}\,\theta} + 4\,e^{-1/2\,\theta} + 6\,e^{-\frac{7}{16}\,\theta} + e^{-3/8\,\theta} + 5\,e^{-\frac{5}{16}\,\theta} - 5\,e^{-3/16\,\theta} - 10\,e^{-1/8\,\theta} - 8\,e^{-1/16\,\theta} - 15\Big)\Big(e^{-3/16\,\theta} - e^{-1/8\,\theta} + e^{-1/16\,\theta} - 1\Big)^{-1}\Big(e^{-3/8\,\theta} + 3\,e^{-\frac{5}{16}\,\theta} + 5\,e^{-1/4\,\theta} + 4\,e^{-3/16\,\theta} + 5\,e^{-1/8\,\theta} + 7\,e^{-1/16\,\theta} + 5\Big)^{-1} \quad (25)$$

$\Phi(\theta)$ contains no critical points, except in the limit as $\theta$ approaches infinity. As $\theta$ approaches zero, $\lim_{\theta\to 0} d\Phi/d\theta = 865/336$. Therefore, the maximimum likelihood estimate for $\theta$ is as small a number as the available numerical accuracy allows (since $\theta > 0$ is required for positive definiteness of the correlation matrix). $B(\theta)$ approaches infinity as $\theta$ approaches zero, but $\Phi(\theta)$ approaches a fixed value (approximately -11.2039).

$n = 6$

$$\Phi(\theta) = -6\,\ln(2) - 3\,\ln(3) - 12\,\ln(5) + 3\,\ln\Big(-\Big(2849 + 65746\,e^{-1/5\,\theta} + 9646\,e^{-1/25\,\theta} + 50721\,e^{-\frac{4}{25}\,\theta} + 78934\,e^{-\frac{9}{25}\,\theta} - 57855\,e^{-\frac{16}{25}\,\theta} - 10000\,e^{-\theta} - 61329\,e^{-\frac{19}{25}\,\theta} - 2850\,e^{-\frac{27}{25}\,\theta} + 85139\,e^{-\frac{7}{25}\,\theta} - 1275\,e^{-\frac{28}{25}\,\theta} + 34559\,e^{-\frac{3}{25}\,\theta} - 25\,e^{-\frac{31}{25}\,\theta} - 24896\,e^{-\frac{14}{25}\,\theta} - 150\,e^{-6/5\,\theta} + 20566\,e^{-\frac{2}{25}\,\theta} - 500\,e^{-\frac{29}{25}\,\theta} - 16050\,e^{-\frac{24}{25}\,\theta} + 45836\,e^{-\frac{11}{25}\,\theta} - 43776\,e^{-\frac{21}{25}\,\theta} - 1606\,e^{-\frac{13}{25}\,\theta} + 78204\,e^{-\frac{6}{25}\,\theta} - 65780\,e^{-\frac{18}{25}\,\theta} - 5650\,e^{-\frac{26}{25}\,\theta} + 85816\,e^{-\frac{8}{25}\,\theta} - 53445\,e^{-4/5\,\theta} + 65304\,e^{-2/5\,\theta} - 33470\,e^{-\frac{22}{25}\,\theta} - 65116\,e^{-\frac{17}{25}\,\theta} - 23959\,e^{-\frac{23}{25}\,\theta} - 44299\,e^{-3/5\,\theta} + 23111\,e^{-\frac{12}{25}\,\theta}\Big)\Big) - 3\ln\Big(e^{-2/5\,\theta} + 2\,e^{-\frac{9}{25}\,\theta} + 2\,e^{-\frac{8}{25}\,\theta} + 2\,e^{-\frac{7}{25}\,\theta} + 4\,e^{-\frac{6}{25}\,\theta} + 4\,e^{-1/5\,\theta} + 3\,e^{-\frac{4}{25}\,\theta} + 2\,e^{-\frac{3}{25}\,\theta} + 3\,e^{-\frac{2}{25}\,\theta} + 4\,e^{-1/25\,\theta} + 3\Big) - 3\ln\Big(e^{-\frac{3}{25}\,\theta} - e^{-\frac{2}{25}\,\theta} + e^{-1/25\,\theta} - 1\Big) - 3\ln\Big(-e^{-\frac{19}{25}\,\theta} - 4\,e^{-\frac{18}{25}\,\theta} - 8\,e^{-\frac{17}{25}\,\theta} - 12\,e^{-\frac{16}{25}\,\theta} - 16\,e^{-3/5\,\theta} - 19\,e^{-\frac{14}{25}\,\theta} - 19\,e^{-\frac{13}{25}\,\theta} - 16\,e^{-\frac{12}{25}\,\theta} - 11\,e^{-\frac{11}{25}\,\theta} - 4\,e^{-2/5\,\theta} + 4\,e^{-\frac{9}{25}\,\theta} + 11\,e^{-\frac{8}{25}\,\theta} + 16\,e^{-\frac{7}{25}\,\theta} + 19\,e^{-\frac{6}{25}\,\theta} + 19\,e^{-1/5\,\theta} + 16\,e^{-\frac{4}{25}\,\theta} + 12\,e^{-\frac{3}{25}\,\theta} + 8\,e^{-\frac{2}{25}\,\theta} + 4\,e^{-1/25\,\theta} + 1\Big) + 1/2\,\ln\Big(1 - 18\,e^{-\frac{32}{25}\,\theta} - 6\,e^{-\frac{66}{25}\,\theta} + 6\,e^{-\frac{4}{25}\,\theta} + 4\,e^{-\frac{16}{25}\,\theta} - 4\,e^{-\frac{28}{25}\,\theta} + 16\,e^{-\frac{14}{25}\,\theta} - 16\,e^{-6/5\,\theta} - 7\,e^{-\frac{64}{25}\,\theta} + 18\,e^{-\frac{38}{25}\,\theta} - 5\,e^{-\frac{2}{25}\,\theta} + 16\,e^{-\frac{24}{25}\,\theta} + 20\,e^{-\frac{48}{25}\,\theta} + 21\,e^{-\frac{52}{25}\,\theta} + 5\,e^{-\frac{68}{25}\,\theta} + 16\,e^{-8/5\,\theta} + 3\,e^{-\frac{34}{25}\,\theta} - 30\,e^{-\frac{44}{25}\,\theta} + 7\,e^{-\frac{6}{25}\,\theta} + 16\,e^{-\frac{62}{25}\,\theta} - 21\,e^{-\frac{18}{25}\,\theta} + 30\,e^{-\frac{26}{25}\,\theta} - 16\,e^{-\frac{8}{25}\,\theta} + 4\,e^{-\frac{42}{25}\,\theta} - 3\,e^{-\frac{36}{25}\,\theta} + 3\,e^{-4/5\,\theta} - 3\,e^{-2\,\theta} - 16\,e^{-\frac{56}{25}\,\theta} - 20\,e^{-\frac{22}{25}\,\theta} - 2\,e^{-\frac{58}{25}\,\theta} - 16\,e^{-\frac{46}{25}\,\theta} - 4\,e^{-\frac{54}{25}\,\theta} - e^{-\frac{14}{5}\,\theta} + 2\,e^{-\frac{12}{25}\,\theta}\Big) \tag{26}$$

$$\lim_{\theta\to\infty} \Phi(\theta) = -6\ln(2) - 6\ln(3) - 12\ln(5) + 3\ln(7) + 3\ln(11) + 3\ln(37) \tag{27}$$

$$\lim_{\theta\to 0} \Phi(\theta) = -\infty \tag{28}$$

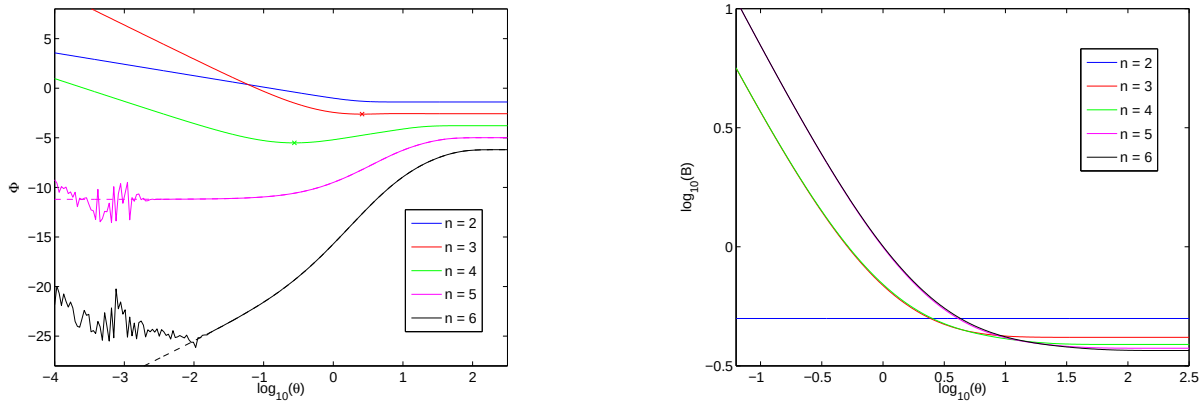
$$B(\theta) = \frac{1}{50}\,\Big(13\,e^{-\frac{13}{25}\,\theta} + 13\,e^{-\frac{12}{25}\,\theta} + 13\,e^{-\frac{11}{25}\,\theta} + 13\,e^{-2/5\,\theta} + 26\,e^{-\frac{9}{25}\,\theta} - 4\,e^{-\frac{8}{25}\,\theta} + 9\,e^{-\frac{7}{25}\,\theta} - 21\,e^{-\frac{6}{25}\,\theta} - 8\,e^{-1/5\,\theta} - 8\,e^{-\frac{4}{25}\,\theta} - 8\,e^{-\frac{3}{25}\,\theta} - 38\,e^{-\frac{2}{25}\,\theta} - 25\,e^{-1/25\,\theta} - 55\Big)\Big(e^{-2/5\,\theta} + 2\,e^{-\frac{9}{25}\,\theta} + 2\,e^{-\frac{8}{25}\,\theta} + 2\,e^{-\frac{7}{25}\,\theta} + 4\,e^{-\frac{6}{25}\,\theta} + 4\,e^{-1/5\,\theta} + 3\,e^{-\frac{4}{25}\,\theta} + 2\,e^{-\frac{3}{25}\,\theta} + 3\,e^{-\frac{2}{25}\,\theta} + 4\,e^{-1/25\,\theta} + 3\Big)^{-1}\Big(e^{-\frac{3}{25}\,\theta} - e^{-\frac{2}{25}\,\theta} + e^{-1/25\,\theta} - 1\Big)^{-1} \tag{29}$$

As with $n = 5$, $\Phi(\theta)$ contains no critical points, except in the limit as $\theta$ approaches infinity. As $\theta$ approaches zero, $\lim_{\theta\to 0}\Phi = -\infty$. Therefore, the maximimum likelihood estimate for $\theta$ is as small a number as the available numerical accuracy allows. $B(\theta)$ approaches infinity as $\theta$ approaches zero.

Even the analytical expressions for $\Phi(\theta)$ and $B(\theta)$ cannot be enumerated accurately using Matlab, hence the arbitrary precision Python module mpmath [4] was used to evaluate the analytical expressions, using 500 digits. $\Phi(\theta)$ and $B(\theta)$ are depicted graphically in Figure 1, for $n = 2$ to 6. The solid lines are obtained from the Matlab computations, while the dashed lines are the Python mpmath results.

This simple 1D example problem contains the complete spectrum of MLE function behaviour. In the case of $n = 2$, the MLE function decreases monotonically as $\theta$ increases and hence estimates $\theta$ to be infinity. Local minimizers exist for $n = 3$ and 4, while for $n = 5$ and 6 the MLE function estimates $\theta$ to be as small a positive number that the numerical accuracy allows.

The availability of analytical expressions also makes it possible to contradict the result of Zimmermann [2]. Clearly the MLE function does *not* always approach infinity as $\theta$ approaches zero (see $n = 5$




(a) (b)

Figure 1: Graphical depiction of (a) $\Phi$ and (b) B as a function of $\theta$. Solid lines are computed using Matlab, while the dashed lines are computed by making use of arbitrary precision arithmetic.

and 6) and clearly the limit $\lim_{\theta \to 0} B(\theta)$ does *not* always exist ($n = 3$ to 6). Therefore the statement that "... optimally trained nondegenerate spatial Gaussian processes cannot feature arbitrary ill-conditioned correlation matrices." [2] is incorrect.

## 4.2 Extrapolation

An alternative form of the Gaussian correlation function

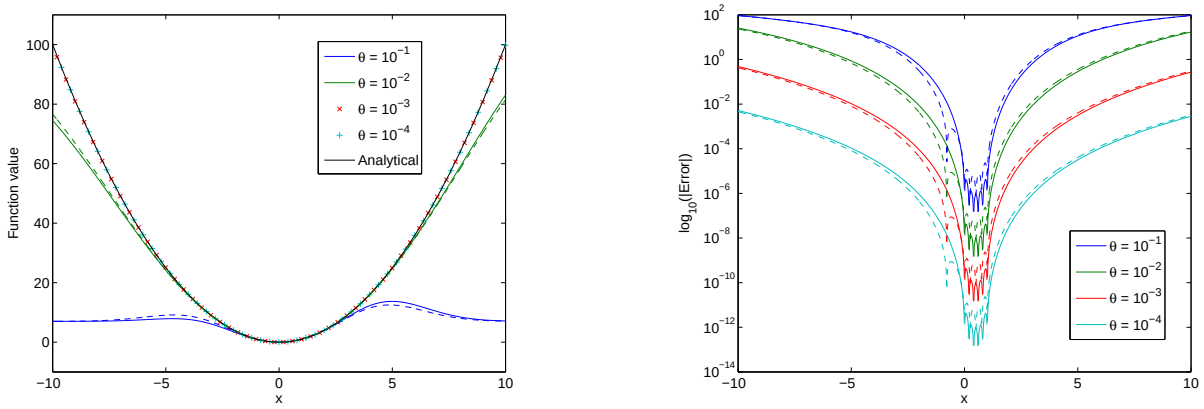
$$R(\boldsymbol{x}^i, \boldsymbol{x}^j) = \prod_{k=1}^{m} e^{-\left|\frac{x_k^i - x_k^j}{d_k}\right|^2}, \tag{30}$$

makes use of the range parameter $d_k$ instead of the hyper-parameter $\theta_k$. This allows an intuitive explanation of the range parameter, i.e. $d_k$ indicates the strength of influence of point $i$ on point $j$. At a distance $d_k$ apart, points $i$ and $j$ still has an influence of $e^{-2} = 0.1353$ or approximately 14% on each other. In terms of the hyper-parameters, $\theta_k \to 0$ corresponds to a range parameter approaching infinity and $\theta_k \to \infty$ corresponds to a range parameter of zero. Therefore $\theta_k \to 0$ occurs where all points affect one another equally strong (global approximation) while $\theta_k \to \infty$ occurs when each point is only affected by itself (local approximation). The analytical example demonstrated a progression from $\theta \to \infty$ (local approximation), to $\theta$ well-defined, to $\theta \to 0$ (global approximation) as the number of sampling points is increased, and this seems to be the general case, rather than an anomaly.

At first glance the cases where the MLE function contains no unique minimum ($n = 5$ and $n = 6$) presents an unsatisfactory result. The interpretation is often that maximum likelihood estimation is not valid for such a problem. However, since the case of $\theta \to 0$ indicates a global approximation (infinite range parameter), the resulting Kriging approximation may in fact not only be accurate over the range of construction, but also beyond.

Figure 2 compares the Kriging approximations to the analytical function $y = x^2$ for $n = 5$ (dashed) and $n = 6$ (solid). In particular, note the scale of the graph (-10 to 10). Although the Kriging approximation was only constructed between 0 and 1, the Kriging approximation is accurate well beyond this range. As $\theta$ is chosen smaller, the range over which the Kriging approximations are accurate becomes larger. This is further highlighted in Figure 3 for $n = 6$, where the approximations using $\theta = 10^{-3}$ and $\theta = 10^{-4}$ are plotted on an even larger domain (-400 to 400). This example demonstrates that at least in some cases where the MLE function continually decreases as $\theta$ decreases (i.e. the maximum likelihood estimate for $\theta$ is as small a positive number as the available numerical accuracy allows), the resulting Kriging approximations can be exceedingly accurate over a large domain.

To further demonstrate the ability of Kriging approximations to extrapolate accurately beyond the range of construction, consider the function $y = (x + a)^2$. This function has a unique minimizer at




(a) (b)

Figure 2: (a) Analytical function compared to Kriging approximations and (b) Absolute error between analytical function and Kriging approximations. $n = 6$: Solid lines, $n = 5$: dashed lines.

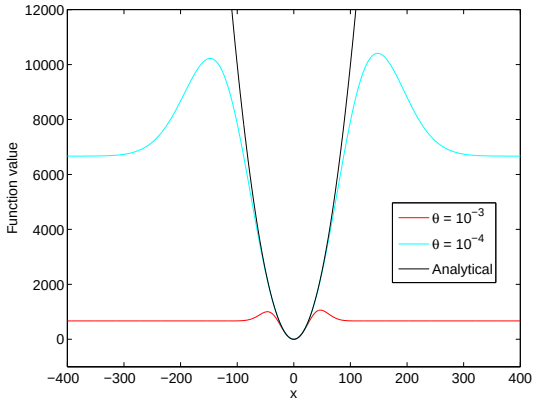



Figure 3: Analytical function $f(x) = x^2$ compared to Kriging approximations $\tilde{f}(x)$ for $n = 6$, using 50 digits for computation.

$x = -a$. This function is approximated using 6 evenly spaced points between 0 and 1. Independent of the choice of $a$, the MLE function contains no local minimum and estimates $\theta$ to be as small a positive number as the numerical accuracy allows. Specific small positive values are assigned to $\theta$, the Kriging approximation is constructed and then minimized. The minimizer $x^*$ of the Kriging approximation $\tilde{y}$, and this approximate function value $\tilde{y}(x^*)$ are compared to the exact solution in Table 1. Again notice that as $\theta$ is decreased, the accuracy of the approximation improves. Consider the specific case using $\theta = 10^{-5}$ and $a = 100$. The actual function values at the 6 points between 0 and 1 range from 10000 to 10201. The Kriging approximation estimates the minimum to occur at $x^* \approx -101.036$, and it estimates the function value to be approximately -35.4 at this point. MLE functions that indicate $\theta \to 0$ is not problematic (from the perspective if maximum likehood estimation is valid or not), rather it indicates that the approximation is sufficiently accurate to extrapolate well beyond the range of construction. This is especially noteworthy for the use of Kriging approximations in optimization, where updates outside the original range of construction is often desirable. It is however problematic from an ill-conditioning perspective, which is easily addressed by using arbitrary precision computation. Optimization problems requiring the solution of expensive finite element or computational fluid dynamics models can easily justify the use of arbitrary precision computations to construct and optimize the Kriging approximation.

Table 1: Minimizers $x^*$ of the Kriging approximation $\tilde{y}(x)$ to $y(x) = (x + a)^2$ using 6 evenly spaced points from 0 to 1 and 50 digits for the computations, for differenct choices of $\theta$.

| $a$ | $x^*$ | $\tilde{y}(x^*)$ | $y(x^*)$ |
|---|---|---|---|
| $\theta = 10^{-3}$ | | | |
| 1 | -1.0000079 | $-3.75 \times 10^{-6}$ | $6.24 \times 10^{-11}$ |
| 10 | -10.128 | $-4.54 \times 10^{-1}$ | $1.64 \times 10^{-2}$ |
| 100 | -37.777 | $4.77 \times 10^{3}$ | $3.87 \times 10^{3}$ |
| $\theta = 10^{-4}$ | | | |
| 1 | -1.0000000795 | $-3.76 \times 10^{-8}$ | $6.32 \times 10^{-15}$ |
| 10 | -10.00146 | $-5.11 \times 10^{-3}$ | $2.13 \times 10^{-6}$ |
| 100 | -105.067 | $-5.29 \times 10^{2}$ | $2.57 \times 10^{1}$ |
| $\theta = 10^{-5}$ | | | |
| 1 | -1.000000000796 | $-3.76 \times 10^{-10}$ | $6.34 \times 10^{-19}$ |
| 10 | -10.0000148 | $-5.19 \times 10^{-5}$ | $2.19 \times 10^{-10}$ |
| 100 | -101.036 | $-3.54 \times 10^{1}$ | $1.07 \times 10^{0}$ |

### 4.3 Two dimensional numerical example

A 2D numerical example is presented next. The six-hump Camelback function [3] is given by

$$y(x_1, x_2) = \left(4 - 2.1x_1^2 + \frac{x_1^4}{3}\right) x_1^2 + x_1 x_2 + (-4 + 4x_2^2)x_2^2. \tag{31}$$

The contours of the associated MLE function is plotted in Figure 4, computed by using an $n \times n$ full factorial DOE, for $n$ ranging from 2 to 7. Arbitrary precision arithmetic is used, and sufficient digits are used to ensure numerical accuracy (up to 120 digits for $n = 7$). The DOE occupies the design space [0:1] in each direction, and this is mapped to $x_1 \in [-2, 2]$, $x_2 \in [-1, 1]$ when evaluating the function.

The results do not demonstrate a simple progression from $\theta \to \infty$ to $\theta \to 0$ as the number of sample points increase (as with the 1D test problem). Nevertheless, the complete spectrum of hyper-parameter behaviour is again present in this problem. The hyper-parameters are estimated to be infinity for $n = 2$, and to be as small a positive number as available accuracy allows for $n = 4$ and $n = 6$. A unique internal local minimizer exists for $n = 5$. For $n = 3$ and $n = 7$, $\theta_2$ approaches zero faster than $\theta_1$. Of particular note is the cases where the MLE function decreases continuously as the hyper-parameters approach zero. To further highlight this behaviour, Figure 6 depicts $\Phi(\theta_1, \theta_2)$ for the $n = 6$ case, along the line $\theta_1 = \theta_2$ for $\theta_1$ between $10^{-20}$ and $10^2$. Arbitrary precision arithmetic, using 250 digits, was used to generate this result. Although this is not a formal proof, it does present convincing numerical evidence that MLE functions do not necessarily approach infinity as the hyper-parameters approach zero.

Using conventional double precision accuracy to perform the computations, the MLE function contours are displayed in Figure 5 (only real components are used in those cases where the calculations result in complex numbers). Note the smaller range on the hyper-parameters for $n = 5$ to 7. All the cases where MLE estimated $\theta \to 0$ ($n = 4$, $n = 6$ and $n = 7$) now contain spurious internal local minimizers, purely due to round-off error. This example could present an explanation for statements in the literature that local minimizers are often found "close to" ill-conditioned areas. In this problem these spurious local minimizers are actually well within the ill-conditioned area. If the MLE contours are only judged on appearance, the effect of ill-conditioning can be underestimated. Consider for example Figure 5 (e), where the MLE function contour appears sufficiently smooth that some analysts may incorrectly identify the spurious local minimizer as the actual solution. The condition number at this spurious minimizer is $10^{21}$. The MLE function along the line $\theta_1 = \theta_2$ for $n = 6$ is calculated using double precision, and the real component is superposed on the arbitrary precision calculation in Figure 6.

Finally, to demonstrate the accuracy of the Kriging approximation in cases where the hyper-parameters approach zero, the $n = 7$ Kriging approximation is constructed using $\theta_1 = 10^{-3.6}$ and $\theta_2 = 10^{-6}$. Figure 7 depicts the test function, Kriging approximation and the difference between them. Note that the condidion number for this case is $10^{75}$.

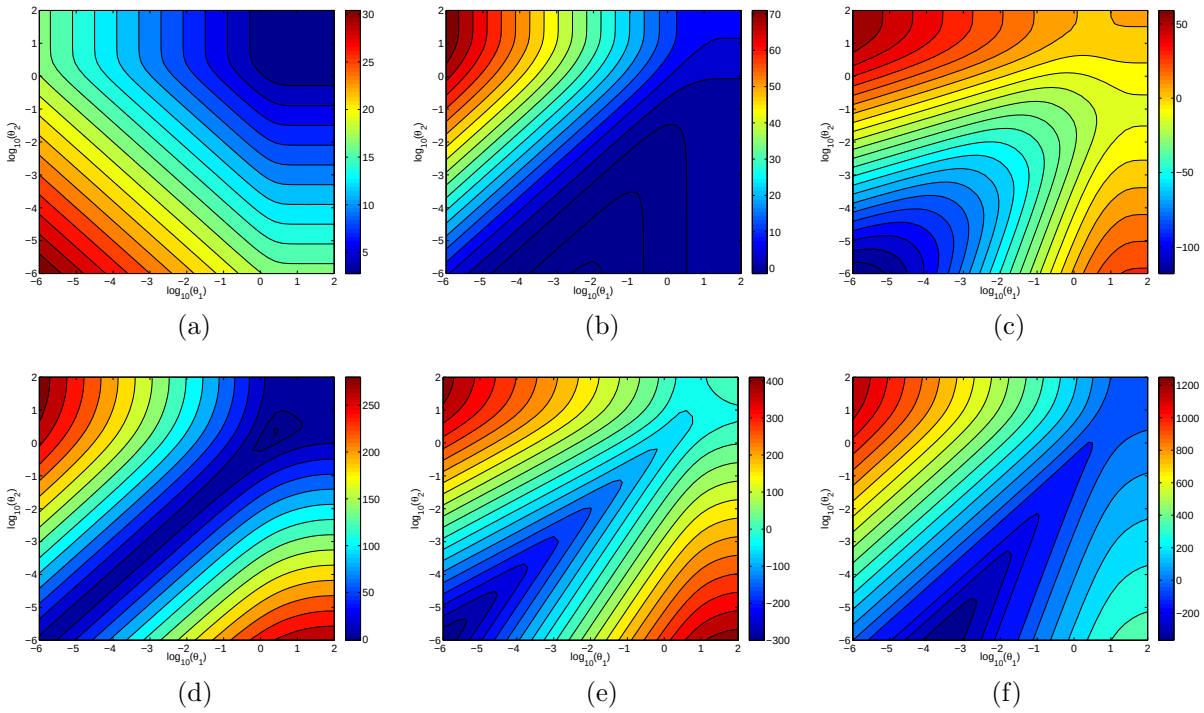



Figure 4: Contours of the MLE function of the six-hump Camelback function usign an $n \times n$ full factorial DOE for (a) $n = 2$ (b) $n = 3$, (c) $n = 4$, (d) $n = 5$, (e) $n = 6$ and (f) $n = 7$, using arbitrary precision computations.

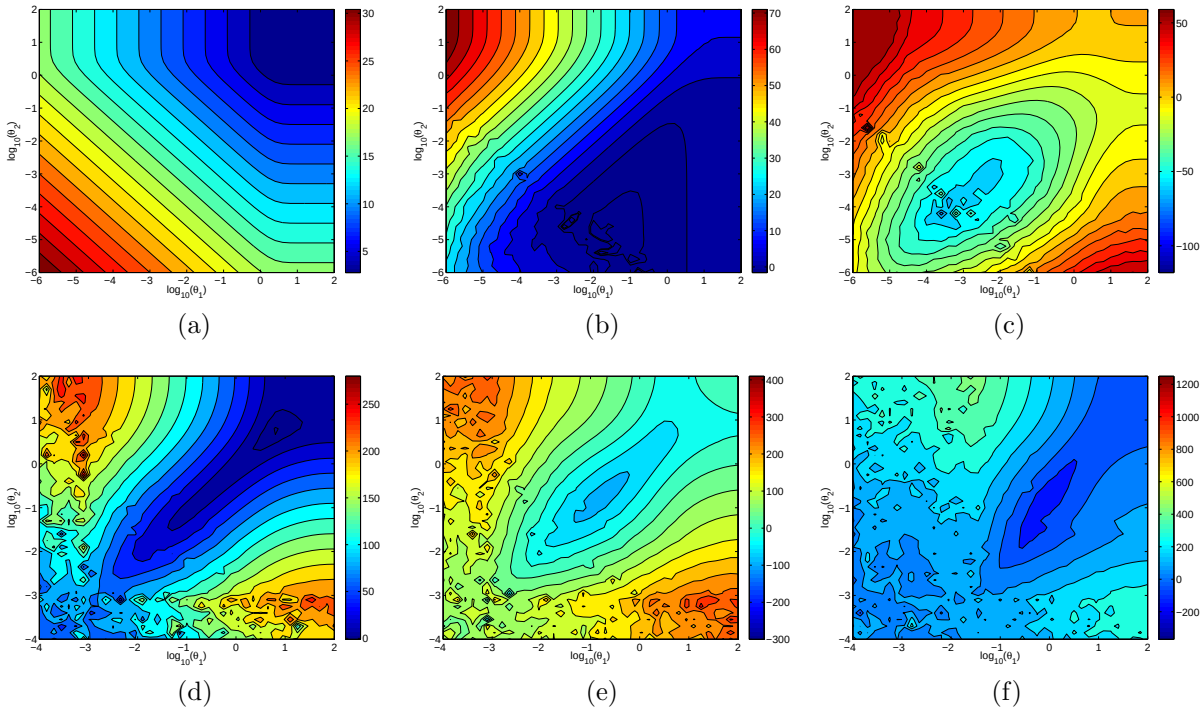



Figure 5: Contours of the MLE function of the six-hump Camelback function usign an $n \times n$ full factorial DOE for (a) $n = 2$ (b) $n = 3$, (c) $n = 4$, (d) $n = 5$, (e) $n = 6$ and (f) $n = 7$, using double precision computations.

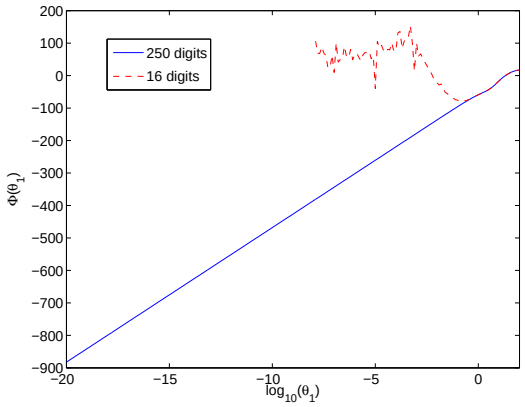



Figure 6: MLE function $\Phi(\theta_1, \theta_2)$ along the line $\theta_1 = \theta_2$ using 250 digits or 16 digits for the calculation, for the case $n = 6$.

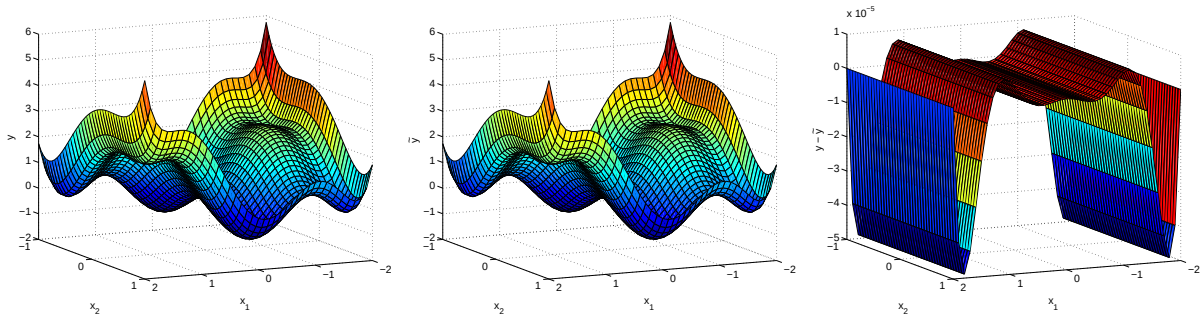



Figure 7: (a) Six-hump Camelback function $y(x_1, x_2)$, (b) Kriging approximation $\tilde{y}(x_1, x_2)$ and (c) the difference $y(x_1, x_2) - \tilde{y}(x_1, x_2)$.

# 5 Conclusions

The behaviour of the MLE function is investigated when the hyper-parameters approach zero or infinity. The main result is that some cases exist for which the MLE function decreases continuously as the hyper-parameters approach zero. This behaviour is typical if many sampling points are used. Ill-conditioning of the correlation matrix occurs in this case, which can produce spurious local minimizers in the MLE function if double precision arithmetic is used. The ill-conditioning can be addressed by using arbitrary precision computations, which is reasonable for optimization problems using expensive simulations.

# 6 References

[1] Krige D. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 1951, 52, 119–139.

[2] Zimmermann R. Asymptotic behavior of the likelihood function of covariance matrices of spatial Gaussian processes. *Journal of Applied Mathematics*, 2010, DOI:10.1155/2010/494070.

[3] Martin J.D., Simpson, T.W. On the use of Kriging models to approximate deterministic computer models. *Proceedings of DETC2004/DAC-57300*, 2004, Salt Lake City.

[4] Johansson F. *et al.*, mpmath: a Python library for arbitrary-precision floating-point arithmetic (version 0.14), February 2010. url: `http://code.google.com/p/mpmath/`.