

A Framework for Optimization under Limited Information

Tansu Alpcan
Technical University of Berlin
Deutsche Telekom Laboratories
Berlin, Germany
alpcan@sec.t-labs.tu-berlin.de

ABSTRACT

In many real world problems, optimization decisions have to be made often with limited information. The decision maker may have no a priori data about the (nonconvex) objective function except from on a limited number of points that are obtained over time through costly observations. This paper presents an optimization framework that takes into account the information collection (observation), estimation (regression), and optimization (maximization) aspects in a holistic and structured manner. Explicitly quantifying the information acquired at each optimization step using the entropy measure from information theory, the objective function to be optimized is modeled and estimated by adopting a Bayesian approach, specifically using Gaussian processes as a state-of-the-art regression method. The resulting iterative scheme allows the decision maker to solve the problem by expressing preferences for each aspect quantitatively and concurrently.

Categories and Subject Descriptors

G.1.6 [Optimization]: Global optimization; H.1.1 [Systems and Information Theory]: Value of information; I.2.6 [Learning]: Parameter learning—*Machine learning, statistical learning*

General Terms

Theory

Keywords

Optimization, information theory, learning, multi-objective optimization, decision making under limited information.

1. INTRODUCTION

In many real world problems, optimization decisions have to be made with limited information. Whether it is a static optimization or dynamic control problem, obtaining detailed and accurate information about the problem or system can often be a costly and time consuming process. In some cases, acquiring extensive information on system characteristics may be simply infeasible. In others, the observed system may be so nonstationary that by the

time the information is obtained, it is already outdated due to system's fast-changing nature. Therefore, the only option left to the decision-maker is to develop a strategy for collecting information efficiently, and choose a model to estimate the "missing portions" of the problem, while trying to reach the given objective.

To make the discussion more concrete, consider the problem of maximizing a (Lipschitz) continuous *nonconvex* objective function, which is unknown except from its value at only a small number of data points. The decision maker may have no a priori information about the function and start with zero data points. Furthermore, only a limited number of –possibly noisy– observations may be available before making a decision on the maximum value and its location. The function itself, however, remains unknown even after the decision is made. *What is the best strategy to address this problem?*

The decision making framework presented in this paper captures the posed problem by taking into account the information collection (observation), estimation (regression), and (multi-objective) optimization aspects in a holistic and structured manner. Hence, the framework enables the decision maker to solve the problem by expressing preferences for each aspect quantitatively and concurrently. It explicitly incorporates many concepts that have been implicitly considered by heuristic schemes, and builds upon many results from seemingly disjoint but relevant fields such as information theory, machine learning, and optimization and control theories.

Despite methods and approaches from machine (statistical) learning are heavily utilized in this framework, the problem at hand is very different from many classical machine learning ones, even in its learning aspect. In most classical application domains of machine learning such as data mining, computer vision, or image and voice recognition, the difficulty is often in handling significant amount of data in contrast to lack of it. Many methods such as Expectation-Maximization (EM) inherently make this assumption, except from "active learning" schemes [2]. Information theory plays an important role in evaluating scarce (and expensive) data and developing strategies for obtaining it. Interestingly, data scarcity converts at the same time the disadvantages of some methods into advantages, e.g. the scalability problem of Gaussian processes.

Decision making with limited information is related to search theory. The idea of using information (theory) in this context is hardly new as evidenced by the article "A New Look at the Relation Between Information Theory and Search Theory" from 1979 [17]. The subject is further studied in [7] and has been recently revis-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

VALUETOOLS 2011, May 16-20, Paris, France

Copyright © 2011 ICST 978-1-936968-09-1

DOI 10.4108/icst.valuetools.2011.245775

ited in [29].

The book [14] provides important and valuable insights into the relationship between information theory, inference, and learning. Measuring information content of experiments using Shannon information is explicitly mentioned and a slightly informal version of the bisection example in Subsection 3.2 is discussed. However, focusing mainly on more traditional coding, communication, and machine learning topics, the book does not discuss the type of optimization problems presented in this paper.

Learning plays an important role in the presented framework, especially *regression*, which is a classical machine (or statistical) learning method. A very good introduction to the subject can be found in [2]. A complementary and detailed discussion on kernel methods is in [21]. Another relevant topic is Bayesian inference [27, 14], which is in the foundation of the presented framework. In machine learning literature, Gaussian processes (GPs) are getting increasingly popular due to their various favorable characteristics. The book [18] presents a comprehensive treatment of GPs. Additional relevant works on the subject include [14, 21, 13], which also discuss GP regression.

Convex optimization is a well-known topic that is often easy to handle even if available information is limited. Optimizing nonconvex functions, however, is still a research subject [8]. It is interesting to note that the method known as *kriging* in global optimization is almost the same as GP regression in machine learning. The field *stochastic programming* focuses on optimization under uncertainty but assumes a certain amount of prior knowledge on the problem at hand and models the uncertainty probabilistically [20]. The popular heuristic method *simulated annealing* [19] is essentially based on iterative random search. Another popular heuristic scheme particle swarm optimization [9] is also based on random search but parallel in nature as a distinguishing characteristic rather than iterative.

Gaussian processes have been recently applied to the area of optimization and regression [3] as well as system identification [26]. While the latter of these recent works mentions active learning, neither work discusses explicit information quantification or builds a connection with Shannon information theory. The recent articles [11] that utilizes GP regression for optimization in a setting similar to the one in this paper and [28] for state-space inference and learning, respectively, do not consider information-theoretic aspects of the problem, either.

The area of active learning or experiment design focuses on data scarcity in machine learning and makes use of Shannon information theory among other criteria [23]. The paper [12] discusses objective functions which measure the expected informativeness of candidate measurements within a Bayesian learning framework. The subsequent study [22] investigates active learning for GP regression using variance as a (heuristic) confidence measure for test point rejection.

It is worth noting that the class of problems described here are much more frequently encountered in practice than it may first seem. For example, the class of black-box methods known as “kriging” [6] have been applied to such problems in geology and mining as well as to hydrology since mid-1960s. In addition, the solution framework proposed is applicable to a wide variety of fields due to its fundamental nature. One example is decentralized resource allocation decisions in networked and complex systems, e.g. wired and

wireless networks, where parameters change quickly and global information on network characteristics are not available at the local decision-making nodes. Another example is security-related decisions where opponents spend a conscious effort to hide their actions. A related area is security and information technology risk management in large-scale organizations, where acquiring information on individual subsystems and processes can be very costly. Yet another example application is in biological systems where individual organisms or subsystems operate autonomously (even if they are part of a larger system) under limited local information.

2. PROBLEM DEFINITION

Let $\mathcal{X} \subseteq \Psi \subset \mathbb{R}^d$ be a nonempty, convex, and compact (closed and bounded) subset of the original problem domain Ψ of d dimensions. The original domain Ψ does not have to be convex, compact, or even fully known. However, adopting a “divide and conquer” approach, the subset \mathcal{X} provides a reasonable starting point. Define next the objective function to be maximized

$$f : \mathcal{X} \rightarrow \mathbb{R},$$

which is unknown except from on a finite number of points (possibly imperfectly) observed. As a simplifying assumption, let f be Lipschitz continuous on \mathcal{X} . *The problem is the find the best search strategy that maximizes the function f .*

The simplest (both conceptually and computationally) strategy to solve Problem 1 is random search on the domain \mathcal{X} . As such no attempt is made to “learn” the properties of the function f . Unless, f is “algorithmically random” [10], which is rarely the case, this strategy wastes the information collected on f . A slightly more complicated and very popular set of strategies combine random search with simple modeling of the function through gradient methods. In this case, the collected information is used to model f rudimentarily using derived gradients to “define slopes” in a heuristic manner. Then, these slopes of f are explored step-by-step in the upwards direction to find a local maximum, after which the search algorithm randomly jumps to another location. It is also possible to randomize the gradient climbing scheme for additional flexibility [19].

The framework presented in this paper takes one further step and **explicitly** models the (entire) objective function f (on the set \mathcal{X}) using the information collected instead of heuristically describing only the slopes. The function \hat{f} , which models, approximates, and estimates f , belongs to a certain class functions such that $\hat{f} \in \mathcal{F}$. The selection and properties of this class is based on “a priori” information available and can be interpreted as the “world view” of the decision maker. These properties can often be expressed using meta-parameters which are then updated based on the observations through a separate optimization process. Likewise, a slower time-scale process can be used for model selection if processing capabilities permit a multi-model approach.

When doing random search on the domain \mathcal{X} , at each stage i.e. given the previous observations, each remaining candidate data point provides equivalent amount of information. However, this is not the case when doing model-based search. Depending on the model adopted and previous information collected, different unexplored points provide different amount of information. This information can be exactly quantified using the definition of entropy and information from the field of (Shannon) information theory. Accordingly, the scalar quantity $\mathcal{I}(\hat{f}, \Omega_n)$ denotes the aggregate information obtained from the set of observations Ω_n within the model rep-

represented by \hat{f} . A related issue is the reliability and possibly noisy nature of observations, which will be discussed in further detail in the next section.

Now, the problem can be formally stated as:

PROBLEM 1 (Model-based Search Problem). *Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a Lipschitz-continuous objective function on the d -dimensional nonempty, convex, and compact set $\mathcal{X} \subset \mathbb{R}^d$, which is unknown except from a finite number of observed data points. Further let $\hat{f}(x)$ be an estimate of the objective function obtained using an a priori model and observed data. What is the best search strategy $\Omega_N := \{x_1, \dots, x_N : x_i \in \mathcal{X} \forall i, N \geq 1\}$ that solves the multi-objective problem with the following components?*

- *Objective 1: $\max_{\Omega_N} f(x)$ given $\hat{f}(x)$*
- *Objective 2: $\arg \min_{\Omega_N} R(f(x), \hat{f}(x)), \hat{f} \in \mathcal{F}$*
- *Objective 3: $\max_{\Omega_N} \mathcal{I}(\hat{f}, \Omega_n)$*

Here, $R(\cdot, \cdot)$ is a risk or expected loss function quantifying the mismatch between actual and estimated functions on the observation data [18]. The scalar quantity \mathcal{I} is the aggregate information obtained from the set of observations Ω_N within the model represented by \hat{f} . The cardinality of Ω_N , N , can be either given, e.g. $N = 1$, or defined as an additional constraint $\sum_{x \in \Omega_n} c_o(x) \leq C$, where $c_o(x) : \mathcal{X} \rightarrow \mathbb{R}$ is the observation cost function, and the scalar C is the total “exploration budget”.

3. METHODOLOGY

This section presents the methods that are utilized within the framework which addresses the problem defined in the previous one. First, the regression model and Gaussian Processes (GP) are presented. Subsequently, modeling and measurement of information is discussed based on (Shannon) information theory.

3.1 Regression and Gaussian Processes (GP)

Problem 1 presented in the previous section involves inferring or learning the function f using the set of observed data points. This is known as the *regression* problem in machine learning and is a supervised learning method since the observed data constitutes at the same time the learning data set. This learning process involves selection of a “model”, where the learned function \hat{f} is, for example, expressed in terms of a set of parameters and specific basis functions, and at the same time minimization of an error measure between the functions f and \hat{f} on the learning data set. This paper focuses on Gaussian Process [18] as the chosen regression method within the framework developed without loss of any generality.

It is not possible to present here a comprehensive treatment of GP. Therefore, a very rudimentary overview is provided next within the context of the decision making problem. Consider a set of M data points $\mathcal{D} = \{x_1, \dots, x_M\}$, where each $x_i \in \mathcal{X}$ is a d -dimensional vector, and the corresponding vector of scalar values is $f(x_i)$, $i = 1, \dots, M$. Assume that the observations are distorted by a zero-mean Gaussian noise, n with variance $\sigma \sim \mathcal{N}(0, \sigma)$. Then, the resulting observations is a vector of Gaussian $y = f(x) + n \sim \mathcal{N}(f(x), \sigma)$.

A GP is formally defined as a collection of random variables, any finite number of which have a joint Gaussian distribution. It is completely specified by its mean function $m(x)$ and covariance function $C(x, \tilde{x})$, where

$$m(x) = E[\hat{f}(x)]$$

and

$$C(x, \tilde{x}) = E[(\hat{f}(x) - m(x))(\hat{f}(\tilde{x}) - m(\tilde{x}))], \forall x, \tilde{x} \in \mathcal{D}.$$

Let us for simplicity choose $m(x) = 0$. Then, the GP is characterized entirely by its covariance function $C(x, \tilde{x})$. Since the noise in observation vector y is also Gaussian, the covariance function can be defined as the sum of a *kernel function* $Q(x, \tilde{x})$ and the diagonal noise variance

$$C(x, \tilde{x}) = Q(x, \tilde{x}) + \sigma I, \forall x, \tilde{x} \in \mathcal{D}, \quad (1)$$

where I is the identity matrix. While it is possible to choose here any (positive definite) kernel $Q(\cdot, \cdot)$, one classical choice is

$$Q(x, \tilde{x}) = \exp\left[-\frac{1}{2}\|x - \tilde{x}\|^2\right]. \quad (2)$$

Note that GP makes use of the well-known *kernel trick* here by representing an infinite dimensional continuous function using a (finite) set of continuous basis functions and associated vector of real parameters in accordance with the *representer theorem* [21].

The (noisy)¹ training set (\mathcal{D}, y) is used to define the corresponding GP, $\mathcal{GP}(0, C(\mathcal{D}))$, through the $M \times M$ covariance function $C(\mathcal{D}) = Q + \sigma I$, where the conditional Gaussian distribution of any point outside the training set, $\tilde{y} \in \mathcal{X}, \tilde{y} \notin \mathcal{D}$, given the training data (\mathcal{D}, t) can be computed as follows. Define the vector

$$k(\tilde{x}) = [Q(x_1, \tilde{x}), \dots, Q(x_M, \tilde{x})] \quad (3)$$

and scalar

$$\kappa = Q(\tilde{x}, \tilde{x}) + \sigma. \quad (4)$$

Then, the conditional distribution $p(\tilde{y}|y)$ that characterizes the $\mathcal{GP}(0, C)$ is a Gaussian $\mathcal{N}(\hat{f}, v)$ with mean \hat{f} and variance v ,

$$\hat{f}(\tilde{x}) = k^T C^{-1} y \text{ and } v(\tilde{x}) = \kappa - k^T C^{-1} k. \quad (5)$$

This is a key result that defines GP regression as the mean function $\hat{f}(x)$ of the Gaussian distribution and provides a prediction of the objective function $f(x)$. At the same time, it belongs to the well-defined class $\hat{f} \in \mathcal{F}$, which is the set of all possible sample functions of the GP

$$\mathcal{F} := \{\hat{f}(x) : \mathcal{X} \rightarrow \mathbb{R} \text{ such that } \hat{f} \in \mathcal{GP}(0, C(\mathcal{D})), \forall \mathcal{D}\},$$

where $C(\mathcal{D})$ is defined in (1) and \mathcal{GP} through (3), (4), and (5), above. Furthermore, the variance function $v(x)$ can be used to measure the uncertainty level of the predictions with the mean value \hat{f} . This issue will be discussed next in detail.

3.2 Quantifying Information in Observations

In the framework presented, each observation provides a data point to the regression problem (estimating f by constructing \hat{f}) as discussed in the previous subsection. The *active learning* problem

¹The special case of perfect observation without noise is handled the same way as long as the kernel function $Q(\cdot, \cdot)$ is positive definite

defined in Section 2 requires addressing the question of “how to quantify information obtained and optimize the observation process?”. Making any decision on the next (set of) observations in a principled manner necessitates first *measuring the information obtained from each observation within the adopted model*. It is important to note that the information measure here is dependent on the chosen model. The same observation provides different amount of information to a random search model than a GP one.

Shannon information theory readily provides the necessary mathematical framework for measuring the information content of a variable. Let p be a probability distribution over the set of possible values of a discrete random variable A . The **entropy** of the random variable is given by $H(A) = \sum_i p_i \log_2(1/p_i)$, which quantifies the amount of uncertainty. Then, the information obtained from an observation on the variable, i.e. reduction in uncertainty, can be quantified simply by taking the difference of its initial and final entropy,

$$\mathcal{I} = H_0 - H_1.$$

It is important here to avoid the common conceptual pitfall of equating entropy to information itself as it is sometimes done in communication theory literature.² Within this framework, (Shannon) *information is defined as a measure of the decrease of uncertainty after (each) observation (within a given model)*. This can be best explained with the following simple example.

3.2.1 Example: Bisection

Choose a number between 1 and 64 randomly with uniform probability (prior). What is the best searching strategy for finding this number? Let the random variable A represent this number. In the beginning the entropy of A is

$$H_0(A) = \sum_{i=1}^{64} \frac{1}{64} \log_2 \left(\frac{1}{64} \right) = 6 \text{ (bits)}.$$

The information maximization problem is defined as

$$\max \mathcal{I} = \max H_0 - H_1 = \min H_1,$$

since H_0 , the entropy before the action (obtaining information) is constant. The entropy H_1 is the one after information is obtained, and hence is directly affected by the specific action chosen. Now, define the action as setting a threshold $1 < t < 64$ to check whether the chosen number is less or higher than this threshold t . To simplify the analysis, consider a continuous version of the problem by defining p as the probability of the chosen number being less than the threshold. Thus, in this *uniform prior case*, the problem simplifies to

$$\min_p H_1 = \min_p p \log(p) + (1-p) \log(1-p).$$

Clearly, the threshold $p^* = 0.5$ is the global minimum, which roughly corresponds to $t = 32$ (ignoring quantization and boundary effects). Thus, bisection from the middle is the optimal search strategy for the *uniform prior*. In this example, the number can be found in the worst-case in 6 steps, each providing one bit of information. Nonuniform probabilities (priors) can be handled in a similar way.

²Since this issue is not of great importance for the class of problems considered in communication theory, it is often ignored. However, the difference is of conceptual importance in this problem. See <http://www.ccrnp.ncifcrf.gov/~toms/information.is.not.uncertainty.html> for a detailed discussion.

If this search process (bisection) is repeatedly applied without any feedback, then it results in the optimal quantization of the search space both in the uniform case above and for the nonuniform probabilities. If feedback is available, i.e. one learns after each bisection whether the number is larger or less than the boundary, then this is as shown the best search strategy.

4. MODEL

The model adopted in the framework for decision making with limited information consists of three main parts: observation, update of GP for regression, and optimization to determine next action. These three steps, shown in Figure 1 are taken iteratively to achieve the objectives in Problem 1. As a result of its iterative nature, this approach can be considered in a sense similar to the well-known Expectation-Maximization algorithm [2].

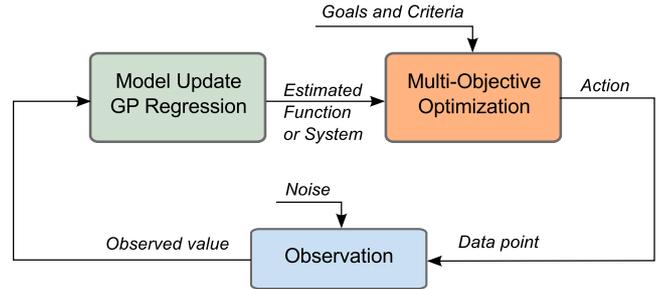


Figure 1: The main parts of the underlying model of the decision making framework.

Observations, given that they are a scarce resource in the class of problems considered, play an important role in the model. Uncertainties in the observed quantities can be modeled as additive noise. Likewise, properties (variance or bias) of additive noise can be used to model the reliability of (and bias in) the data points observed. GPs provide a straightforward mathematical structure for incorporating these aspects to the model under some simplifying assumptions.

The set of observations collected provide the (supervised) training data for GP regression in order to estimate the characteristics of the function or system at hand. This process relies on the GP methods described in Subsection 3.1. Thus, at each iteration an up-to-date description of the function or system is obtained based on the latest observations. Specifically, \hat{f} provides an estimate of the original function f .³ Assuming an additive Gaussian noise model, the noise variance σ can be used to model uncertainties, e.g. older and noisy data resulting in higher σ values.

When making a decision on the next action through multi-objective optimization, there are (infinitely) many candidate points. A pragmatic solution to the problem of finding solution candidates is to (adaptively) sample the problem domain \mathcal{X} to obtain the set

$$\Theta := \{x_1, \dots, x_T : x_i \in \mathcal{X}, x_i \notin \mathcal{D}, \forall i\}$$

that does not overlap with known points. In low (one or two) dimensions, this can be easily achieved through grid sampling methods. In higher dimensions, (Quasi) Monte Carlo schemes can be utilized. For large problem domains, the current domain of interest

³See [18, Chap 7.2] for a discussion on asymptotic analysis of GP regression. It should not be noted, however, that asymptotic properties are of little relevance to the problem at hand.

\mathcal{X} can be defined around the last or most promising observation in such a way that such a sampling is computationally feasible. Likewise, multi-resolution schemes can also be deployed to increase computational efficiency.

A natural question that arises is: whether and under what conditions does such a sampling method give satisfactory results. The following result from [24, 25] provides an answer to this question in terms of number of samples required.

THEOREM 1. *Define a multivariate function $f(x)$ on the convex, compact set \mathcal{X} , which admits the maximum*

$$x^* = \arg \max_{x \in \mathcal{X}} f(x).$$

Based on a set of N random samples $\Theta = \{x_1, \dots, x_N : x_i \in \mathcal{X} \forall i\}$ from the entire set \mathcal{X} , let $\hat{x} := \arg \max_{x \in \Theta} f(x)$ be an estimate of the maximum x^ .*

Given an $\varepsilon > 0$ and $\delta > 0$, the minimum number of random samples N which guarantees that

$$Pr(Pr[f(x^*) > f(\hat{x}) \leq \varepsilon] \geq 1 - \delta),$$

i.e. the probability that 'the probability of the real maximum surpassing the estimated one being less than ε ' is larger than $1 - \delta$, is

$$N \geq \frac{\ln 1/\delta}{1/(1-\varepsilon)}.$$

Furthermore, this bound is tight if the function f is continuous on \mathcal{X} .

It is interesting and important to note that this bound is independent of the sampling distribution used (as long as it covers the whole set \mathcal{X} with nonzero probability), the function f itself, as well as the properties and dimension of the set \mathcal{X} .

4.1 Quantifying Information in GP

The information measurement and GP approaches in Section 3 can be directly combined. Let the zero-mean multivariate Gaussian (normal) probability distribution be denoted as

$$p(x) = \frac{1}{\sqrt{2\pi|C_p(x)|}} \exp\left(-\frac{1}{2}[x - m]^T |C_p(x)|^{-1} [x - m]\right), \quad (6)$$

where $x \in \mathcal{X}$, $|\cdot|$ is the determinant, m is the mean (vector) as defined in (5), and $C_p(x)$ is the covariance matrix as a function of the newly observed point $x \in \mathcal{X}$ given by

$$C_p(x) = \begin{bmatrix} C(\mathcal{D}) & k(x)^T \\ k(x) & \kappa \end{bmatrix}. \quad (7)$$

Here, the vector $k(x)$ is defined in (3) and κ in (4), respectively. The matrix $C(\mathcal{D})$ is the covariance matrix based on the training data \mathcal{D} as defined in (1).

The entropy of the multivariate Gaussian distribution (6) is [1]

$$H(x) = \frac{d}{2} + \frac{d}{2} \ln(2\pi) + \frac{1}{2} \ln |C_p(x)|,$$

where d is the dimension. Note that, this is the entropy of the GP estimate at the point x based on the available data \mathcal{D} . The aggregate entropy of the function on the region \mathcal{X} is given by $H := \int_{x \in \mathcal{X}} \frac{1}{2} \ln |C_p(x)| dx$.

The problem of choosing a new data point \hat{x} such that the information obtained from it within the GP regression model is maximized can be formulated in a way similar to the one in the bisection example:

$$\begin{aligned} \hat{x} &= \arg \max_{\tilde{x}} \mathcal{I} = \arg \max_{\tilde{x}} \int_{x \in \mathcal{X}} [H_0 - H_1] dx \\ &= \arg \min_{\tilde{x}} \int_{x \in \mathcal{X}} \frac{1}{2} \ln |C_q(x, \tilde{x})| dx, \end{aligned} \quad (8)$$

where the integral is computed over all $x \in \mathcal{X}$, and the covariance matrix $C_q(x, \tilde{x})$ is defined as

$$C_q(x, \tilde{x}) = \begin{bmatrix} C(\mathcal{D}) & k^T(\tilde{x}) & k^T(x) \\ k(\tilde{x}) & \tilde{\kappa} & Q(x, \tilde{x}) \\ k(x) & Q(x, \tilde{x}) & \kappa \end{bmatrix}, \quad (9)$$

and $\tilde{\kappa} = Q(\tilde{x}, \tilde{x}) + \sigma$. Here, $C(\mathcal{D})$ is a $M \times M$ matrix and C_q is a $(M+2) \times (M+2)$ one, whereas κ and $Q(x, \tilde{x})$ are scalars and k is a $M \times 1$ vector. This result is summarized in the following proposition.

PROPOSITION 1. *As a maximum information data collection strategy for a Gaussian Process with a covariance matrix $C(\mathcal{D})$, the next observation \hat{x} should be chosen in such a way that*

$$\hat{x} = \arg \max_{\tilde{x}} \mathcal{I} = \arg \min_{\tilde{x}} \int_{x \in \mathcal{X}} \ln |C_q(x, \tilde{x})| dx,$$

where $C_q(x, \tilde{x})$ is defined in (9).

An Approximate Solution to Information Maximization

Given a set of (candidate) points Θ sampled from \mathcal{X} , the result in Proposition 1 can be revisited. The problem in (8) is then approximated [25] by

$$\begin{aligned} \max_{\tilde{x}} \mathcal{I} &\approx \min_{\tilde{x}} \sum_{x \in \Theta} \ln |C_q(x, \tilde{x})| \\ \Rightarrow \hat{x} &= \arg \min_{\tilde{x} \in \Theta} \prod_{x \in \Theta} |C_q(x, \tilde{x})|, \end{aligned} \quad (10)$$

using monotonicity property of the natural logarithm and the fact that the determinant of a covariance matrix is non-negative. Thus, the following counterpart of Proposition 1 is obtained:

PROPOSITION 2. *As an approximately maximum information data collection strategy for a Gaussian Process with a covariance matrix $C(\mathcal{D})$ and given a collection of candidate points Θ , the next observation $\hat{x} \in \Theta$ should be chosen in such a way that*

$$\hat{x} = \arg \min_{\tilde{x} \in \Theta} \prod_{x \in \Theta} |C_q(x, \tilde{x})| \approx \arg \max_{\tilde{x} \in \Theta} \mathcal{I},$$

where $C_q(x, \tilde{x})$ is given in (9).

Although it is an approximation, finding a solution to the optimization problem in Proposition 2 can still be computationally costly.

Therefore, a greedy algorithm is proposed as a computationally simpler alternative. Let, $x^* \in \Theta$ be defined as

$$x^* := \arg \max_{x \in \Theta} |C_p(x)| = |C(\mathcal{D})| |\kappa(x) - k(x)C^{-1}(\mathcal{D})k^T(x)|,$$

where the matrix C_p is given by (7) [16]. The first term above, $|C(\mathcal{D})|$ is fixed and the second one,

$$|\kappa(x) - k(x)C^{-1}(\mathcal{D})k^T(x)|,$$

is the same as the GP variance $v(x)$ in (5). Hence, the sample x^* is one of those with the maximum variance in the set Θ , given current data \mathcal{D} .

It follows from (9) and basic matrix theory that if $\tilde{x} = x$ for a given x then $|C_q(x, \tilde{x})|$ is minimized. As a simplification, ignore the dependencies between $C_q(x, \tilde{x})$ matrices for different $x \in \Theta$. Then, choosing the maximum variance \hat{x} as

$$\hat{x} = \arg \max_{\tilde{x} \in \Theta} v(\tilde{x}) \approx \arg \min_{\tilde{x} \in \Theta} \prod_{x \in \Theta} |C_q(x, \tilde{x})|,$$

leads to a large (possibly largest) reduction in $\prod_{x \in \Theta} |C_q(x, \hat{x})|$, and hence provides a rough approximate solution to (10) and to the result in Proposition 1. This result is consistent with widely-known heuristics such as ‘‘maximum entropy’’ or ‘‘minimum variance’’ methods [23] and a variant has been discussed in [12].

PROPOSITION 3. *Given a Gaussian Process with a covariance matrix $C(\mathcal{D})$ and a collection of candidate points Θ , an approximate solution to the maximum information data collection problem defined in Proposition 1 is to choose the sample point(s) \tilde{x} in such a way that it has (they have) the maximum variance within the set Θ .*

5. OPTIMIZING WITH LIMITED INFORMATION

The static optimization Problem 1 introduced in Section 2 is formulated next as a multi-objective optimization problem within the model of Section 4.

The first and basic objective is the maximization of the function $f(x)$ on $x \in \mathcal{X}$. As a simplification, observations are assumed to be sequential, one at a time. Since f is basically unknown, this problem has to be formulated as

$$\max_{\tilde{x} \in \mathcal{X}} F_1(\tilde{x}) = \hat{f}(\tilde{x}),$$

where \hat{f} is the best estimate obtained through GP regression (5) using the current data set \mathcal{D} . Data uncertainty (observation errors) is modeled through additive Gaussian noise with variance σ as a first approximation.

The second objective is to minimize the difference (estimation error) between \hat{f} and f . Define $e(x) = \hat{f}(x) - f(x)$, $\forall x \in \mathcal{X}$. Given the set of noisy observations

$$\mathcal{O} = \{f(x_i) + n(x_i) : x \in \mathcal{D}, \forall i\},$$

where $n \sim \mathcal{N}(0, \sigma)$ denotes zero mean Gaussian noise, it is possible to use another GP regression (5) to estimate this error function, $\hat{e}(\mathcal{D}, x)$, on the entire set \mathcal{X} . Thus, the second objective is to ensure that the next observation \tilde{x} solves

$$\min_{\tilde{x} \in \mathcal{X}} F_2(\tilde{x}) = \int_{\tau \in \mathcal{X}} |\hat{e}(\tilde{x}, \mathcal{D}, \tau)| d\tau.$$

Note that, F_2 here corresponds to a risk or loss estimate function.

The third objective is to maximize the amount of information obtained with each observation \tilde{x} , or

$$\max_{\tilde{x} \in \mathcal{X}} F_3(\tilde{x}) = \mathcal{I}(\tilde{x}, \hat{f}) = \int_{x \in \mathcal{X}} \ln |C_q(x, \tilde{x})| dx,$$

given the best estimate of the original function, \hat{f} . This objective has already been discussed in Section 3.2 in detail.

The values of the three objectives, F_1 , F_2 , F_3 , cannot be evaluated numerically on the entire set \mathcal{X} . Therefore, a sampling method is used as described in Section 4 to obtain a set of solution candidates Θ , which replaces \mathcal{X} in the maximization and minimization problems above. Next, specific problem formulations are presented based on such a sampling of solution candidates. The overall structure of the framework is visualized in Figure 2.

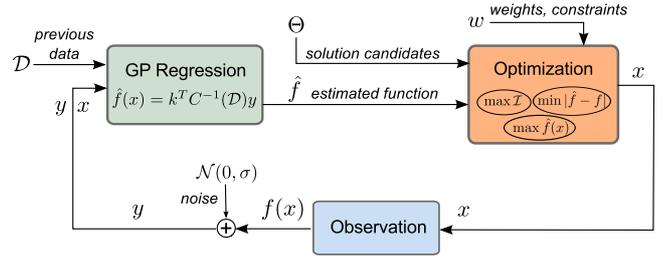


Figure 2: The decision making framework for static optimization with limited information.

5.1 Solution Approaches

The most common approach to multi-objective optimization is the **weighted sum method** [15, 5]. The three objectives discussed above can be combined to obtain a single objective using the respective weights $[w_1, w_2, w_3]$, $\sum_{i=1}^3 w_i = 1$, $0 \leq w_i \leq 1 \forall i$. Assuming a single data point is chosen from and observed among the candidates Θ at each step, i.e. $\tilde{x} = \Omega_1$, a specific weighted sum formulation to address Problem 1 is obtained.

PROPOSITION 4. *The solution, $\tilde{x} \in \Theta$, to the optimization problem*

$$\max_{\tilde{x} \in \Theta} w_1 \hat{f}(\tilde{x}) - w_2 \frac{1}{N} \sum_{\tau \in \Theta} |\hat{e}(\tilde{x}, \mathcal{D}, \tau)| + w_3 \mathcal{I}(\tilde{x}, \hat{f}), \quad (11)$$

constitutes the best search strategy for this weighted sum formulation of Problem 1.

As discussed in Subsection 3.2 and stated in Proposition 2, the information objective, F_3 , in (11) can be approximated by substituting it with GP variance $v(x)$ in (5) to decrease computational load. Thus, an approximation to the solution in Proposition 4 is:

PROPOSITION 5. *The solution, $\tilde{x} \in \Theta$, to the optimization problem*

$$\max_{\tilde{x} \in \Theta} w_1 \hat{f}(\tilde{x}) - w_2 \frac{1}{N} \sum_{\tau \in \Theta} |\hat{e}(\tilde{x}, \mathcal{D}, \tau)| + w_3 v(\tilde{x}), \quad (12)$$

where $v(\tilde{x})$ is defined in (5), approximates the search strategy in Proposition 4.

The weighting scheme described is only meaningful if the three objectives are of the same order of magnitude. Therefore, the original objective functions, F_i , $i = 1, 2, 3$, have to be transformed or “normalized”. There are many different approaches to perform such a transformation [15, 5]. The most common one, which coincidentally is known as normalization, aims to map each objective function to a predefined interval, e.g. $[0, 1]$. To do this, estimate first an upper F_i^U and lower F_i^L bound on each individual objective $F_i(x)$. Then, the i^{th} normalized objective is

$$F_i^N(x) = \frac{F_i(x) - F_i^L}{F_i^U - F_i^L}.$$

The main issue in normalization is to determine the appropriate upper and lower bounds, which is a very problem-dependent one. In the case of Proposition 5, the estimated functions \hat{f} and \hat{e} on the set Θ as well as the existing observations \mathcal{D} , can be utilized to obtain these values. The specific bounds for the respective objectives

$$\begin{aligned} F_1^U &= \max_{x \in \Theta} \hat{f}(x), F_1^L = \min_{x \in \Theta} \hat{f}(x), \\ F_2^U &= \max_{x \in \Theta} |\hat{e}(x, \mathcal{D})|, F_2^L = 0, \\ F_3^U &= \max_{x \in \Theta} \kappa(x), \text{ and } F_3^L = 0 \end{aligned}$$

provide a suitable starting estimate and can be combined with a prior domain knowledge if necessary. Thus, a normalized version of the formulation in Proposition 5 is obtained.

PROPOSITION 6. *The solution, $\tilde{x} \in \Theta$, to the optimization problem*

$$\max_{\tilde{x} \in \Theta} \frac{w_1}{\Delta_1} \left(\hat{f}(x) - F_1^L \right) - \frac{w_2}{\Delta_2} \frac{1}{N} \sum_{\tau \in \Theta} |\hat{e}(\tilde{x}, \mathcal{D}, \tau)| + \frac{w_3}{\Delta_3} v(\tilde{x}), \quad (13)$$

where $\Delta_i = F_i^U - F_i^L$ $i = 1, 2, 3$, provides an approximation to the best search strategy for solving the normalized weighted-sum formulation of Problem 1.

The **bounded objective function** method provides a suitable alternative to the weighted sum formulation above in addressing the multi-objective problem defined. The bounded objective function method minimizes the single most important objective, in this case $F_1(x)$, while the other two objective functions $F_2(x)$ and $F_3(x)$ are converted to form additional constraints. Such constraints are in a sense similar to QoS ones that naturally exist in many real life problems. As an advantage, in the bounded objective formulation there is no need for normalization.

The bounded objective counterpart of the result in Proposition 5 is as follows.

PROPOSITION 7. *The solution, $\tilde{x} \in \Theta$, to the constrained optimization problem*

$$\begin{aligned} &\max_{\tilde{x} \in \Theta} \hat{f}(x) \quad (14) \\ \text{such that } &0 \leq F_2(\tilde{x}) = \frac{1}{N} \sum_{\tau \in \Theta} |\hat{e}(\tilde{x}, \mathcal{D}, \tau)| \leq b_1, \\ &\text{and } 0 \leq F_3(\tilde{x}) = v(\tilde{x}) \leq b_2, \end{aligned}$$

where b_1 and b_2 are given (predetermined) scalar bounds on F_2 and F_3 , respectively, provides an approximate best search strategy for a bounded-objective formulation of Problem 1.

The advantage of the bounded objective function method is that it provides a bound on the information collection and estimation objectives while maximizing the estimated function. This leads in practice to an initial emphasis on information collection and correct estimation of the objective function. In large-scale or high-dimensional problems, however, the space to explore for satisfying any bound on information is simply immense. Therefore, one does not have the luxury of identifying the function first to maximize it later as it would take too many samples to do this. In such cases, it makes more sense to deploy the weighted sum method, possibly along with a “cooling” scheme to modify the weights as part of a cooling scheme to balance depth-first vs. breadth-first search.

5.2 Algorithm and Numerical Examples

An algorithmic summary of the solution approaches discussed above for a specific set of choices is provided by Algorithm 1, which describes both weighted-sum and bounded objective variants.

Algorithm 1 Optimization with Limited Information

- 1: **Input:** Function domain, \mathcal{X} , GP meta-parameters, objective weights $[w_1, w_2, w_3]$ or bounds b_1, b_2 , initial data set (\mathcal{D}, y) .
 - 2: Use GP with a Gaussian kernel and specific expected error variances for function \hat{f} and error function \hat{e} estimation.
 - 3: **while** Search budget available, $1 \leq n \leq N_{max}$. **do**
 - 4: Sample domain \mathcal{X} to obtain $\Theta(n)$. In some cases, $\Theta(n) = \Theta \forall n$.
 - 5: Estimate \hat{f} and \hat{e} based on observed data (\mathcal{D}, y) on $\Theta(n)$ using GPs.
 - 6: Compute variance, $v(x)$, of \hat{f} (5) on $\Theta(n)$ as an estimate of $\mathcal{I}(\hat{f})$.
 - 7: **if** Weighted-sum method **then**
 - 8: Next action maximizes a normalized and weighted sum of objectives $\sum_{i=1}^3 F_i^N$ as stated in Proposition 6.
 - 9: **else if** Bounded objective method **then**
 - 10: Next action is solution to the constrained problem in Proposition 7.
 - 11: **end if**
 - 12: Update the observed data (\mathcal{D}, y) .
 - 13: **end while**
-

Example 1

The first numerical example aims to visualize the presented framework and algorithm. Hence, the chosen function is only one dimensional, $f(x) = \sin(5x)/x$ on the interval $\mathcal{X} = [0.1, 3.9]$. The interval is linearly sampled to obtain a grid with a distance of 0.01 between points, i.e. $\Theta = \{x_i \in \mathcal{X} \forall i : x_1 = 0.1, x_2 = 0.11, \dots, x_N = 3.9\}$. A Gaussian kernel with variance 0.1 is chosen for estimating both \hat{f} and \hat{e} . The weights are equal to one, $w = [1, 1, 1]$, in the weighted-sum method. The bounds are $b_1 = 0.5$ for the error bound and $b_2 = 0.2$ for the bound on maximum variance estimate in the bounded objective method. The initial data consists of a single point, $x = 0.1$.

Figure 3 shows the results based on the normalized weighted-sum method in Proposition 6 after 5 iterations (6 samples in total, together with the initial data point). The variance here is $v(x)$ of the estimated function \hat{f} using data points \mathcal{D} . Clearly, the estimated peak is not the one of the real function f .

Next, Figure 4 shows that after 11 iterations (12 data points in \mathcal{D}), the function and the location of its peak is estimated correctly. The

sequence of points selected during the iteration process are:

$$\mathcal{D} = \{0.47, 3.22, 1.17, 1.66, 2.43, 2.06, 3.9, 2.83, 3.6, 0.82, 1.42\}.$$

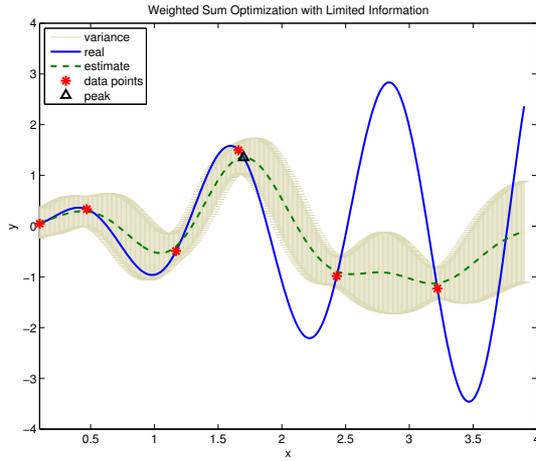


Figure 3: Optimization result using the weighted-sum method with 6 data points.

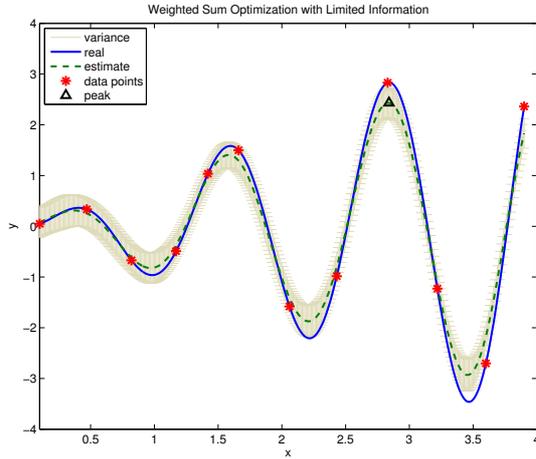


Figure 4: Optimization result using the weighted-sum method with 12 data points.

Finally, Figure 5 depicts the results of the bounded objective method with the given bounds. The number of iterations is 11 as before, which gives an opportunity of direct comparison with the weighted-sum method. The sequence of points selected during the iteration process are:

$$\mathcal{D} = \{0.47, 3.22, 1.17, 1.66, 2.43, 2.06, 3.9, 2.83, 3.6, 0.82, 1.42\}.$$

Example 2

The objective function in the second numerical example is the Goldstein&Price function [4], which is shown in Figure 6 in its inverted form to ensure consistency with the maximization formulation in this paper. The problem domain consists of the two dimensional rectangular region $\mathcal{X} = [-2, 2] \times [-2, 2]$, which is linearly sampled to obtain a uniform grid with a 0.05 interval between sample points. A Gaussian kernel with variance 0.5 and 0.1 is chosen

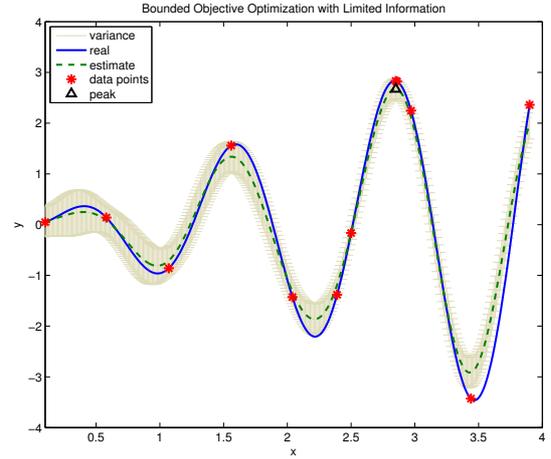


Figure 5: Optimization result using the bounded objective method with 12 data points.

for estimating \hat{f} and \hat{e} , respectively. The weighted-sum method is utilized in Algorithm 1 with the weights $w = [4, 2, 3]$. The search budget is chosen as 50 before stopping the algorithm (for the search space of approx. 6400 samples in the grid). The real global minimum (peak) of the (inverted) Goldstein&Price function is at $(0, -1)$ and the location found by the algorithm using the 50 data points is $(-0.15, -1.05)$. Figure 7 depicts the estimated function, the data points as well as the optimum found. Although the real optimum value is -3 (in the inverted version) while the obtained one is -9.75 , the result is still very satisfactory considering that the simple sampling scheme used and the Goldstein&Price function takes values in a range of 1 million, i.e. the error is less than 0.001 percent of the range. Finally, Figure 8 depicts the mean variance v and entropy \mathcal{I} of the estimated function \hat{f} on Θ at each iteration step. It is observed that the two quantities are closely correlated in accordance with the results in Section 4.

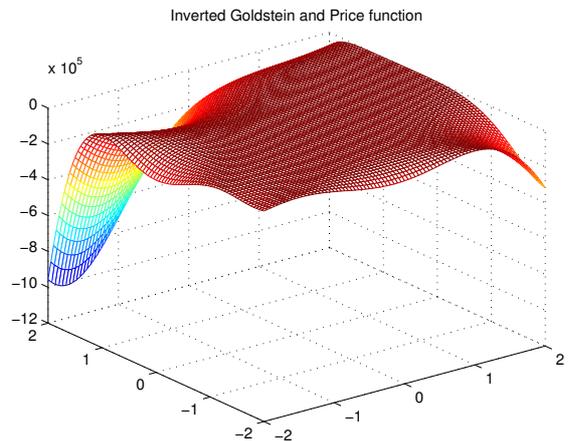


Figure 6: The inverted Goldstein&Price function [4].

6. CONCLUSIONS

The decision making framework presented in this paper addresses the problem of decision making under limited information by tak-

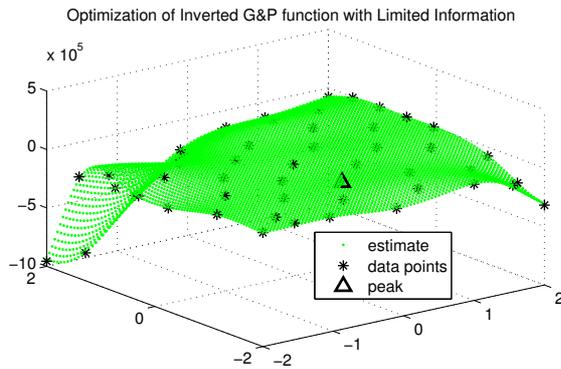


Figure 7: Optimization of the inverted Goldstein&Price function [4] using the weighted-sum method with 50 data points.

ing into account the information collection (observation), estimation (regression), and (multi-objective) optimization aspects in a holistic and structured manner. The methodology is based on Gaussian processes and active learning. Various issues such as quantifying information content of new data points using information theory, the relationship between information and GP variance as well as related approximation and multi-objective optimization schemes are discussed. The framework is demonstrated with multiple numerical examples.

The presented framework should be considered mainly as an initial step. Future research directions are abundant and include further investigation of the exploration-exploitation trade-off, adaptive weighting parameters, and random sampling methods for problems in higher dimensional spaces. Additional research topics are the relationship of the framework with genetic/evolutionary methods, dynamic control problems, and multi-person decision making, i.e. game theory.

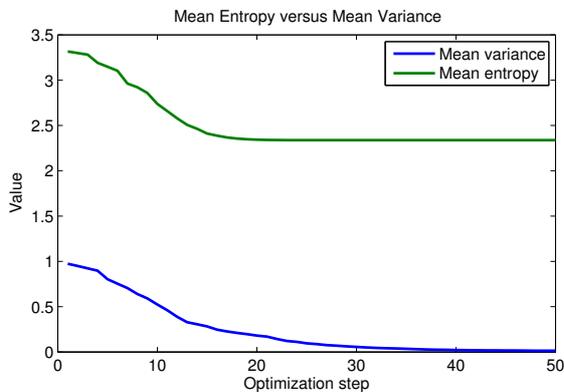


Figure 8: Mean variance v and entropy \mathcal{I} on Θ at each iteration step.

7. ACKNOWLEDGEMENTS

This work is supported by Deutsche Telekom Laboratories. The author wishes to thank Holger Boche, Kivanc Mihcak, Ozgur Erceetin, Lacra Pavel, and Slawomir Stanczak for stimulating discussions on the subject.

8. REFERENCES

- [1] N. Ahmed and D. Gokhale. Entropy expressions and their estimators for multivariate distributions. *IEEE Transactions on Information Theory*, 35(3):688–692, May 1989.
- [2] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
- [3] P. Boyle. *Gaussian Processes for Regression and Optimisation*. PhD thesis, Victoria University of Wellington, Wellington, New Zealand, 2007.
- [4] A. A. Goldstein and J. F. Price. On descent from local minima. *Mathematics of Computation*, 25(115):569–574, July 1971.
- [5] O. Grodzevich and O. Romanko. Normalization and other topics in multi-objective optimization. Proceedings of the Fields–MITACS Industrial Problems Workshop, 2006.
- [6] D. Huang, T. Allen, W. Notz, and N. Zeng. Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of Global Optimization*, 34:441–466, 2006.
- [7] E. T. Jaynes. Entropy and search-theory. In C. R. Smith and J. W. T. Grandy, editors, *Maximum-Entropy and Bayesian Methods in Inverse Problems*, page 443. Springer, 1985.
- [8] D. R. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21:345–383, December 2001.
- [9] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proc. of IEEE Intl. Conf. on Neural Networks*, volume 4, pages 1942–1948, November 1995.
- [10] M. Li and P. Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Texts in Computer Science. Springer, New York, NY, USA, 2nd edition, 1997.
- [11] D. Lizotte, T. Wang, M. Bowling, and D. Schuurmans. Gaussian process regression for optimization. in NIPS 2005 Workshop on Value of Information in Inference, Learning and Decision-Making, 2005.
- [12] D. J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.
- [13] D. J. C. MacKay. Introduction to Gaussian processes. In C. M. Bishop, editor, *Neural Networks and Machine Learning*, NATO ASI Series, pages 133–166. Kluwer Academic Press, 1998.
- [14] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
- [15] R. T. Marler and J. S. Arora. Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optimization*, 26(6):369–395, 2004.
- [16] K. B. Petersen and M. S. Pedersen. *The matrix cookbook*, October 2008.
- [17] J. G. Pierce. A new look at the relation between information theory and search theory. Technical report, Office of Naval Research, Arlington, VA, USA, June 1978.
- [18] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.
- [19] R. Rutenbar. Simulated annealing algorithms: an overview. *IEEE Circuits and Devices Magazine*, 5(1):19–26, January 1989.
- [20] N. V. Sahinidis. Optimization under uncertainty: state-of-the-art and opportunities. *Computers & Chemical Engineering*, 28(6-7):971–983, June 2004. FOCAPO 2003

Special issue.

- [21] B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.
- [22] S. Seo, M. Wallat, T. Graepel, and K. Obermayer. Gaussian process regression: active data selection and test point rejection. In *Proc. of IEEE-INNS-ENNS Intl. Joint Conf. on Neural Networks IJCNN 2000*, volume 3, pages 241–246, July 2000.
- [23] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.
- [24] R. Tempo, E. W. Bai, and F. Dabbene. Probabilistic robustness analysis: Explicit bounds for the minimum number of samples. *Systems & Control Letters*, 30(5):237–242, 1997.
- [25] R. Tempo, G. Calafiore, and F. Dabbene. *Randomized Algorithms for Analysis and Control of Uncertain Systems*. Springer-Verlag, London, UK, 2005.
- [26] K. R. Thompson. *Implementation of gaussian process models for non-linear system identification*. PhD thesis, University of Glasgow, Glasgow, Scotland, 2009.
- [27] M. E. Tipping. Bayesian inference: An introduction to principles and practice in machine learning. In *Advanced Lectures on Machine Learning*, pages 41–62, 2003.
- [28] R. Turner, M. P. Deisenroth, and C. E. Rasmussen. State-space inference and learning with gaussian processes. In *Proc. of 13th Intl. Conf. on Artificial Intelligence and Statistics (AISTATS)*, Chia Laguna Resort, Sardinia, Italy, May 2010.
- [29] Q. Zhu and J. Oommen. On the optimal search problem: the case when the target distribution is unknown. In *Proc. of XVII Intl. Conf. of Chilean Computer Science Society*, pages 268–277, 1997.