A RISK-BASED APPROACH TO OPTIMISATION UNDER LIMITED INFORMATION

TANSU ALPCAN*

Abstract. A risk-based black box optimisation approach is presented that addresses a class of nonconvex multi-variable optimisation problems often encountered in wired and wireless network resource allocation. Information limitations both in terms of search budget and lack of knowledge on the objective function play a defining role in such problems. A novel algorithm is introduced that reduces a risk-like metric in a greedy manner. This risk metric strikes a balance between the conflicting information acquisition and function maximisation objectives. Hence, it brings a novel perspective to the problem. Initial numerical analysis indicates that the presented algorithm is more robust than the alternatives in [2] possibly due to its emphasis on risk reduction.

Key words. Global optimisation, information theory, Gaussian Process Regression.

AMS subject classifications. 90C26, 94A17, 60G15, 91B30.

1. Introduction.

The Problem and Motivation. This paper studies a class of black box optimisation problems, where the objective function is unknown to the optimiser. The optimiser conducts a search on the given problem domain, which is assumed to be known, in order to find the maximum of the objective function. However, the number of points that can be evaluated is limited either due to high cost of information acquisition or time limitations, which may be a result of the problems transient nature. Furthermore, even after the search the objective function remains unknown except from the set of data points explored.

The problem considered is encountered in practice much more frequently than it may first seem. Examples include decentralised resource allocation in wired and wireless networks, security-related decisions, biological systems, and management decisions in large-scale organisations. Black-box methods known as "kriging" [3] have been applied to similar problems in geology, mining, and hydrology since mid-1960s. In wired and wireless networks, system parameters often change quickly and global information on network characteristics are not available at the local decision-making nodes. In many security-related decisions the opponents spend a conscious effort to hide their actions. In large-scale organisations acquiring information on individual subsystems and processes can be very costly, which profoundly affects management decisions. In biological systems, individual subsystems often operate autonomously under limited local information.

The Approach and Contribution. The simplest method (both conceptually and computationally) to solve the problem defined is to conduct a random search on the problem domain. As such no attempt is made to "learn" the properties of the objective function. Unless the function is "algorithmically random" [6] this strategy wastes the information collected. A slightly more complicated and popular set of strategies, e.g. simulated annealing, combine random search with simple modelling of the objective function [11] through identification of local gradients or "slopes".

The random search approaches are not applicable when the number of search points is limited. Therefore, this paper adopts a Bayesian learning approach [8].

*Department of Electrical and Electronic Engineering, The University of Melbourne, VIC 3010 Australia (tansu.alpcan@unimelb.edu.au).

A best estimate of the objective function is derived using the observed data within a selected model. The model choice reflects prior information on the general class the optimisation problem belongs to and can be interpreted as the "world view" of the optimiser. Hence, the estimated objective function is essentially a model-based interpolation of the observations; a combination of the data and the model. Following the Bayesian principles, each new data point is used to refine the estimation in an iterative learning process. The same data can also be used for refining the model itself, e.g. choosing meta parameters in a slower time-scale. However, model selection, which is a separate problem, will not be addressed in this paper.

The learning process here is fundamentally intertwined with information acquisition. Information plays a crucial role due to limitations on the search budget. Hence, the problem is partly one of *active learning* or *experiment design*. Within the chosen Bayesian model each data points provides a different amount of information. Using Shannon Information Theory, this information is quantified as the difference between the entropy values that capture estimation uncertainty before and after observation.

The search problem can be formulated as a weighted sum of two objectives. First is maximising the estimated objective function and the second one is acquiring information in the most efficient manner. This formulation, presented in [2], allows to address the exploration versus exploitation trade-off explicitly. However, if the maximum value of the objective function at the optimal point is known or can be estimated accurately, it is possible to develop a risk-based alternative formulation.

The **main contribution** of this paper is the risk-based approach as a way of combining the exploration and optimisation objectives. A standard definition of risk is combined with an information theoretic approach to obtain a robust optimisation formulation to address the problem. A risk-based algorithm for addressing general (nonconvex) optimisation problems is presented along with a well-defined stopping criterion. The algorithm is demonstrated with a numerical example.

The rest of the paper is organised as follows. The problem and the underlying model used to address it are formulated in the next section along with a brief overview of Gaussian process regression and an entropy-based metric for quantifying Shannon information. Section 3 presents the risk-based algorithm developed. A numerical example is provided in Section 4. The paper concludes with brief remarks in Section 5.

2. Problem Analysis and Model. Let $\mathcal{X} \subset \mathbb{R}^d$ be the nonempty, compact, and known problem domain. The original objective function to be maximised $f_o : \mathcal{X} \to \mathbb{R}$, is unknown except from on a finite number of points observed. As a special but broad case, let f_o belong to the L^p space, $1 \leq p < \infty$. Then, given $\epsilon > 0$, there is a continuous function f such that $||f_o - f|| < \epsilon$. It immediately follows from compactness of $\mathcal{X}$ that f is bounded and assumes its maximum and minimum [10]. Based on this approximation, the focus is on maximisation of the continuous real valued function f on $\mathcal{X}$.

One of the main distinguishing characteristics of this problem is the limitations on set of observations $\Omega_n := \{x_1, \ldots, x_n \ : \ x_i \in \mathcal{X} \ \forall i, \ n \geq 1\}$, due to cost of obtaining information or non-stationarity of the underlying system. In many cases these observations may also be noisy. Accordingly, a basic search problem is defined:

PROBLEM 1 (*Search Problem*). *Consider a continuous objective function $f : \mathcal{X} \to \mathbb{R}$ on the d-dimensional nonempty and compact set $\mathcal{X} \subset \mathbb{R}^d$. The function is and will be unknown except from on a finite number of observed data points. What is the best search strategy $\Omega_N := \{x_1, \ldots, x_N \ : x_i \in \mathcal{X} \ \forall i, \ N \geq 1\}$ to find $x^* = \arg\max_x f(x)$ such that $x^* \in \Omega_N$?*

The number of observations, $N \geq 1$, in Problem 1 may be imposed by the nature of the specific application domain. In a nonstationary problem, there is opportunity for making only a certain number of observations, N, in a given time window. Alternatively, the problem may be stationary but there is an observation cost $c_o(x) : \mathcal{X} \to \mathbb{R}$ and exploration budget C such that $\sum_{x \in \Omega_n} c_o(x) \leq C$.

The Problem 1 involves two distinct objectives. First one is the estimation of the objective function. A method for this task is presented next. The second one is information acquisition taking into account the limited exploration budget, which is discussed in Subsection 2.2.

2.1. Gaussian Process (GP) Regression. This paper uses Gaussian Process (GP) based regression [9] for learning the function $\hat{f}$ that estimates the objective function f on the set $\mathcal{X}$ using the information collected. 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

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

Consider a set of M data (observation) points $\mathcal{D} = \{x_1, \ldots, x_M\}$, where each $x_i \in \mathcal{X}$ is a $d-$dimensional vector, and the corresponding vector of scalar values is $f(x_i), \ i = 1, \ldots, M$. 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 $y = f(x) + n \sim \mathcal{N}(f(x), \sigma)$.

Let us for simplicity choose $m(x) = 0$. Then, the GP is characterised 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

$$C(x, \tilde{x}) = Q(x, \tilde{x}) + \sigma I, \ \forall \, x, \tilde{x} \in \mathcal{D}, \tag{2.1}$$

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

$$Q(x, \tilde{x}) = \exp\left[-\frac{1}{2} \left\|x - \tilde{x}\right\|^2\right]. \tag{2.2}$$

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* [12].

The (noisy)[1] training set $(\mathcal{D}, y)$ is used to define the corresponding GP 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, $\bar{y} \in \mathcal{X}, \bar{y} \notin \mathcal{D}$, given the training data $(\mathcal{D}, t)$ can be computed as follows. Define

$$k(\bar{x}) = [Q(x_1, \bar{x}), \ldots Q(x_M, \bar{x})], \ \ \kappa = Q(\bar{x}, \bar{x}) + \sigma. \tag{2.3}$$

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

$$\hat{f}(\bar{x}) = k^T C^{-1} y \text{ and } v(\bar{x}) = \kappa - k^T C^{-1} k. \tag{2.4}$$

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

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 $f(x)$. Furthermore, the variance function $v(x)$ can be used to measure the uncertainty level of the predictions provided by $\hat{f}$, as it will be discussed in the next subsection.

2.2. Quantifying Information. An important aspect of Problem 1 is to maximise the amount of information obtained with each new observation $\tilde{x}$. 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).*

Define Θ as a discrete search domain obtained by sampling $\mathcal{X}$ [2, 14]. The problem of choosing the optimal new data point $\hat{x} \in \Theta$ such that the information obtained from it within the GP regression model is maximised can be formulated as

$$\hat{x} = \arg\max_{\tilde{x}} \mathcal{I} = \arg\max_{\tilde{x}} \left[H_0 - H_1(\tilde{x})\right]. \tag{2.5}$$

While the uncertainty (entropy) before observation, H_0 is fixed, the uncertainty after the observation is a function of the observation $\tilde{x}$.

The entropy of a multivariate Gaussian distribution is $H(x) = 0.5d(\ln(2\pi) + 1) + 0.5 \ln |C_{\mathcal{D}}(x)|$, where d is the dimension, and $C_{\mathcal{D}}$ is the covariance matrix based on the data set $\mathcal{D}$. The aggregate entropy of the estimated function is given by

$$H^{agg} := \frac{1}{2} \sum_{x \in \Theta} \ln |C_{\mathcal{D}}(x)| + \frac{d}{2} \ln(2\pi e). \tag{2.6}$$

Thus, the information obtained by the new observation $\tilde{x}$ is

$$\mathcal{I} = \frac{1}{2} \sum_{x \in \Theta} \ln \left(\frac{|C_{\mathcal{D}}(x)|}{|C_{\mathcal{D} \cup \tilde{x}}(x)|} \right) \tag{2.7}$$

Although the optimisation problem in (2.5) is not analytically tractable (see e.g. [5] for an interesting discussion), if $\tilde{x}$ is chosen such that the variance is maximised, then this leads to a large (possibly largest) reduction in the denominator of (2.7), and hence provides a rough approximate solution. This result corresponds to the widely-known heuristics such as "maximum entropy" or "minimum variance" methods [13] and a variant has been discussed in [7]. It is also quite intuitive: the maximum amount of information is obtained if the search is conducted away from known data points in empty parts of the search space.

[2] The often ignored difference is of conceptual importance in this problem. See `http://www.ccrnp.ncifcrf.gov/~toms/information.is.not.uncertainty.html` for a detailed discussion.

3. Risk-Based Algorithm. The optimisation problem analysed in the previous section can be formulated as a weighted sum of two objectives. First one is maximising the estimated objective function and the second is acquiring information in the most efficient manner This formulation, presented in [2], allows to address the exploration versus exploitation trade-off explicitly. However, if the maximum value of the objective function f^* can be estimated, it is possible to develop a **risk-based alternative formulation**, which is the main contribution of this paper.

Risk in laymen terms means "something bad could happen". A longer definition is "the probability and magnitude of a loss, disaster, or other undesirable event" [4]. In this case, the magnitude of loss is $|f^* - f(\tilde{x})|$, where $\tilde{x} = \arg\max_{x \in \mathcal{D}} f(x)$ is the best point found after the search. The probability of loss is denoted by $0 \leq p_{risk} \leq 1$. Thus, the formal definition of risk given the set of observations, $\mathcal{D}$, in the context of the optimisation problem analysed is:

$$R(\mathcal{D}) = (f^* - f(\tilde{x}))\ p_{risk}. \tag{3.1}$$

It is assumed that the probability of loss (PrL), p_{risk}, has the following properties: (i) If no data (observation) is available on the objective function, then PrL is maximum, $p_{risk} = 1$. (ii) If the problem space is exhaustively searched, i.e. the objective function is completely known, then PrL is minimum, $p_{risk} = 0$. (iii) The PrL is non-increasing in the number of observation made, i.e. the learning model is accurate enough such that each new data point $\tilde{x}$ improves the quality of the estimates, $p_{risk}(\mathcal{D} \cup \tilde{x}) \leq p_{risk}(\mathcal{D})$.

A metric satisfying these properties is *normalised uncertainty* defined as the current aggregate entropy H_c given in (2.6), divided by the initial entropy, H_0. Thus, the following **risk-based optimisation problem** is obtained:

$$\min_{\tilde{x} \in \Theta} R = (f^* - f(\tilde{x}))\ \frac{H_c(\tilde{x})}{H_0}. \tag{3.2}$$

Note that, it is possible to solve (3.2) iteratively by adopting a greedy algorithm, which is equivalent to addressing Problem 1. Let H_{prev} be the entropy before the observation is made and the information provided by data point $\tilde{x}$ be $\mathcal{I} = H_{prev} - H_c(\tilde{x})$. Then, the risk after the observation is

$$R = \frac{f^* H_{prev}}{H_0} - \mathcal{I}\frac{f^*}{H_0} - f\frac{H_{prev}}{H_0} + f\,\mathcal{I}\frac{1}{H_0}.$$

Since f^*, H_0, and H_{prev} are constant, (3.2) can be re-written as

$$\max_{\tilde{x} \in \Theta}\ \mathcal{I}(\tilde{x}) f^* + f(\tilde{x})\, H_{prev} - f(\tilde{x})\, \mathcal{I}, \tag{3.3}$$

where $f(\tilde{x}) = \max(f(\tilde{x}), f(\mathcal{D}))$. As in [2], it is possible to use variance as an approximation of information leading to $\mathcal{I}(\tilde{x}) \approx v(\tilde{x}) / |\Theta|$. The resulting greedy risk-based optimisation scheme is summarised in Algorithm 1.

4. Numerical Example. The Algorithm 1 is illustrated on the two-dimensional inverted Ackley function [1]. A uniformly random sampling of the domain $\mathcal{X} = [-1, 1]^2$ with 5000 points defines the search space Θ. The Gaussian kernel in (2.2) with variance 0.1 is chosen for estimating $\hat{f}$. The stopping criterion (risk threshold) is chosen as $T = 0.1$. The initial data point is chosen as $x = (1, 1)$. The algorithm

Algorithm 1 Greedy Risk-Based Algorithm

1: **Input:** Function domain, $\mathcal{X}$, GP meta-parameters, risk threshold T, initial data set $(\mathcal{D}, y)$.
2: Use GP with a Gaussian kernel and specific expected error variances for function $\hat{f}$ estimation.
3: **while** Risk $R > T$ **do**
4: Sample domain $\mathcal{X}$ to obtain $\Theta(n)$ (or as a simplification, $\Theta(n) = \Theta\ \forall n$).
5: Estimate $\hat{f}$ based on observed data $(\mathcal{D}, y)$ on $\Theta(n)$ using GP regression.
6: Compute variance, $v(x)$, of $\hat{f}$ (2.4) on $\Theta(n)$ as an estimate of $\mathcal{I}$.
7: Choose the point that solves (3.3) as the next observation, $y(\tilde{x})$.
8: Update the observed data $(\mathcal{D}, y)$.
9: **end while**

reaches the threshold in less than 25 observations. The results are depicted in Figures 4.1 and 4.3. The search path on the search space, which starts at $x = (1, 1)$ is shown in Figure 4.2. This and similar numerical experiments indicate that the algorithm is more conservative yet more robust when compared to the fine-tuned weighted-sum scheme in [2].

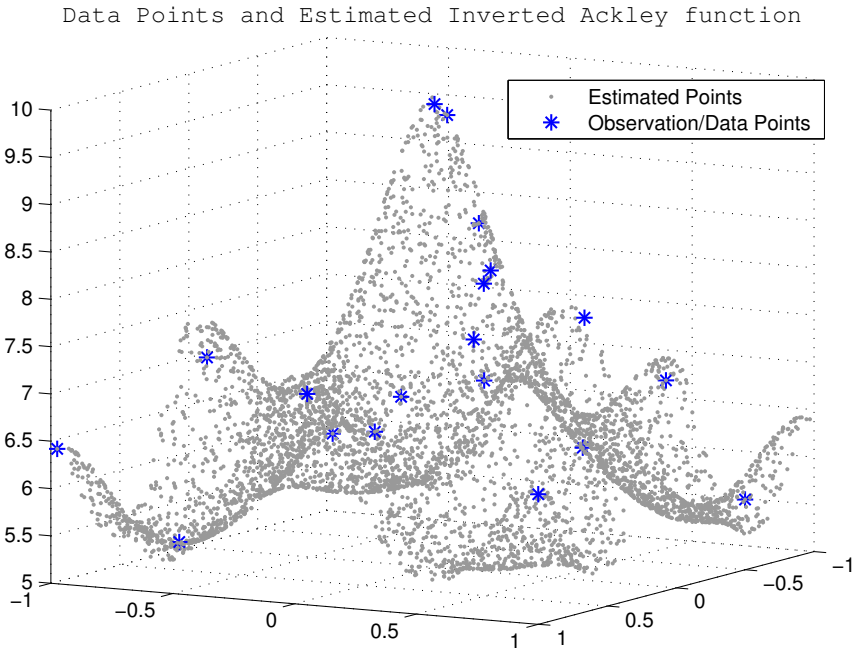

FIG. 4.1. *Optimisation of the inverted inverted Ackley function.*

5. Conclusion. A risk-based black box optimisation algorithm is developed to address nonconvex multi-variable optimisation problems often encountered in wired and wireless network resource allocation. Information limitations both in terms of search budget and objective function define this class of problems. Unlike the results

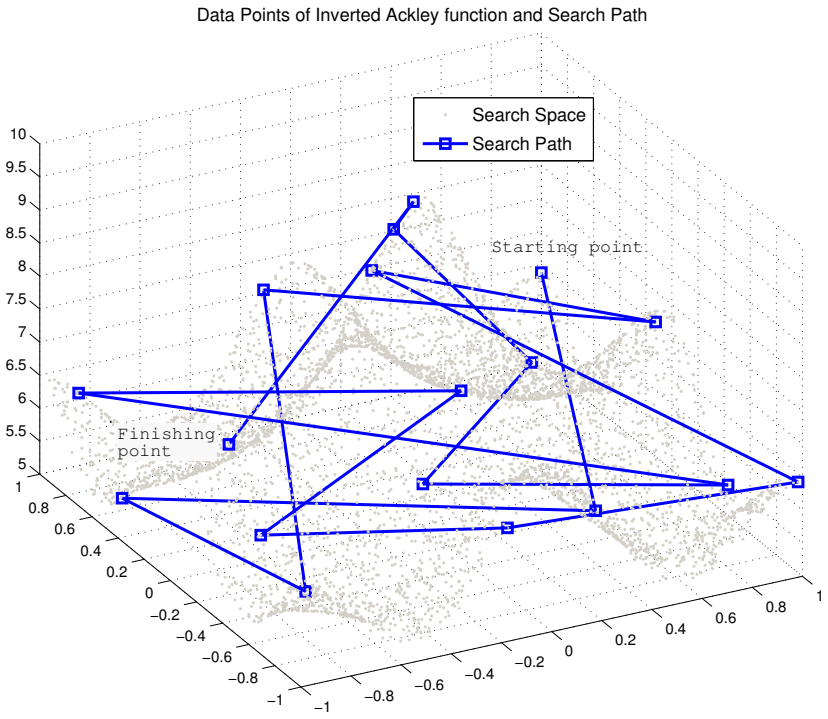

FIG. 4.2. *The search path from starting point* $(1, 1)$ *until the point when risk threshold is reached.*

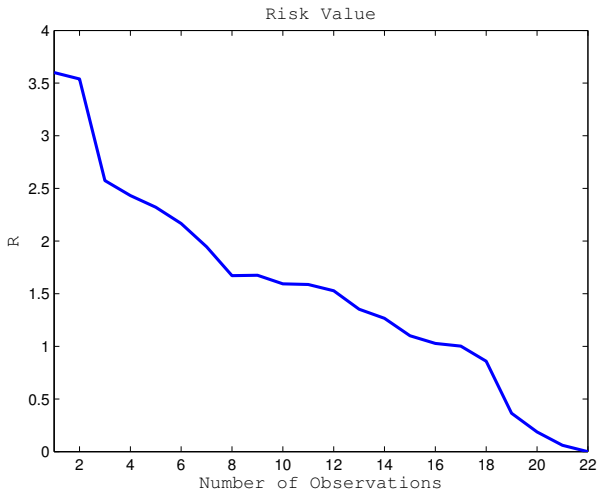

FIG. 4.3. *Evolution of the risk value over observations.*

in [2], the algorithm here aims to reduce risk in a greedy manner instead of striking a balance of information and maximisation or emphasising first search and then maximisation. Hence, it brings a novel perspective to the problem. In addition, it is observed to be more robust than alternatives in [2] due to its emphasis on "risk reduction".

REFERENCES

[1] D.H. Ackley, *A connectionist machine for genetic hillclimbing*, Kluwer Boston Inc., Hingham, MA, 1987.

[2] T. Alpcan, *A framework for optimization under limited information*, in 5th Intl. Conf. on Performance Evaluation Methodologies and Tools (ValueTools), ENS, Cachan, France, May 2011.

[3] 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 (2006), pp. 441–466.

[4] D. W. Hubbard, *The Failure of Risk Management: Why It's Broken and How to Fix It*, John Wiley and Sons, Hoboken, New Jersey, USA, 2009.

[5] N.D. Lawrence, M. Seeger, and R. Herbrich, *Fast sparse Gaussian process methods: The informative vector machine*, Advances in neural information processing systems, 15 (2002), pp. 609–616.

[6] Ming Li and Paul Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications*, Texts in Computer Science, Springer, New York, NY, USA, 2nd ed., 1997.

[7] David J. C. MacKay, *Information-based objective functions for active data selection*, Neural Computation, 4 (1992), pp. 590–604.

[8] ———, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.

[9] Carl Edward Rasmussen and Christopher K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press, 2005.

[10] H. L. Royden, *Real Analysis*, Prentice-Hall, New Jersey, USA, 3rd ed., 1988.

[11] R.A. Rutenbar, *Simulated annealing algorithms: an overview*, IEEE Circuits and Devices Magazine, 5 (1989), pp. 19–26.

[12] Bernhard Scholkopf and Alexander J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, USA, 2001.

[13] Burr Settles, *Active learning literature survey*, Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[14] R. Tempo, G. Calafiore, and F. Dabbene, *Randomized Algorithms for Analysis and Control of Uncertain Systems*, Springer-Verlag, London, UK, 2005.