5-2024

Selected Topics on Sequential Designs for Decision Making

Caroline Kerfonta
ckerfon@clemson.edu

Selected Topics on Sequential Designs for Decision Making

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Mathematical Sciences

by
Caroline M. Kerfonta
May 2024

Accepted by:
Dr. Qiong Zhang, Committee Chair
Dr. Christopher McMahan
Dr. Patrick Gerard
Dr. Deborah Kunkel

Abstract

This dissertation is comprised of three parts. The first proposes a sequential approach to determine the experimental setting with the minimum variance (Kerfonta et al., 2024). Two acquisition functions are developed to assist developing the approach. Theoretical results along with a case study using data from crystallization experiments is conducted to show the ability of the proposed method to correctly select the experiment with the minimum variance. The second and third parts propose adaptations to the Bayesian optimization algorithm using transformed additive Gaussian processes (TAG) as the surrogate model. The goal of using the TAG framework is to decompose the optimization problem into multiple one-dimensional optimization problems. The second part of this dissertation proposes a Bayesian optimization algorithm for single objective optimization using TAG as the surrogate model and a modified expected improvement acquisition function. To demonstrate the advantages of the proposed method, it will be compared to Bayesian optimization with a Gaussian process surrogate model using the expected improvement acquisition function. The final part of this dissertation proposes a bi-objective Bayesian optimization algorithm that uses TAG as the surrogate model and a modified expected hypervolume improvement acquisition function. This approach is compared to classical bi-objective Bayesian optimization using a Gaussian process surrogate model. Functions from existing bi-objective optimization literature are used to demonstrate the advantages of the proposed method.

Dedication

For my parents Robert and Verna Kerfonta

And for my friend Fr. Marcin Zahuta

And for my fiancé Nathaniel Nemire

Acknowledgments

I am deeply grateful to my PhD advisor Dr. Qiong Zhang. Through her support and guidance, I have learned to conduct statistical research on interesting problems. I am thankful for all of the opportunities that she has given me to work with many different collaborators on a variety of applications of statistics. Without her, I would not have grown to be the statistician that I am today.

I would also like to thank my committee members Dr. Chris McMahan, Dr. Patrick Gerard, and Dr. Deborah Kunkel, their support and comments have helped to better develop this dissertation. I would like to thank Dr. McMahan for his help in navigating through the PhD program and its requirements. I am grateful to Dr. Gerard for his encouraging words and careful edits. I would like to thank Dr. Kunkel for being my first year advisor at Clemson, she ensured that I started the program with the best classes and information.

Throughout my undergraduate and graduate school career, I have had the opportunity to study and work with amazing faculty, staff, and colleagues. I would like to thank all of the faculty and staff in the School of Mathematical and Statistical Sciences at Clemson University. I am grateful to my fellow Clemson students for their support and friendship. I would also like to thank all of the faculty and staff in the Department of Statistics at the University of South Carolina. I am grateful for the opportunities I have had to intern with the Boeing Co. while in school. I would like to thank Kelsea Cox, Dr. Stephen Jones, and the entire Boeing Applied Math team for their support and encouragement. I would also like to thank the Clemson VIPR-GS center for allowing me to be a part of their groundbreaking research. I am very thankful for my high school teachers, Nancy Pate and Katie Durham who encouraged me to continue studying mathematics.

I am deeply grateful to have the support of my family and friends through this process. First, I would like to thank my sister Maggie, the first person I ever told I was considering a PhD,

without her support then and today I would not have started or finished this degree. There are not enough words to thank my parents for their constant support. I am very grateful for their encouragement and all of the lessons they have taught me and continue to teach me. I would like to thank my brothers: Matt, for always sending me the recipes and support I need, and Andy, for all of the late night texts and encouragement. I would also like to thank my siblings in-law: Rachel, Dana, and James, and my niece and nephews: James, Henry, John, Audrey, and Jake for their support. I am lucky to have friends like Becca and Fr. Marcin who are able to make me laugh, no matter the occasion.

Finally, I would like to thank my fiancé Nathan for his never-ending support and encouragement. I am excited to be able to finish this chapter and start our next phase of life together.

Table of Contents

Title Page . . . . **i**

Abstract . . . . **ii**

Dedication . . . . **iii**

Acknowledgments . . . . **iv**

List of Tables . . . . **viii**

List of Figures . . . . **ix**

1 Introduction **1**

2 Sequential Selection for Minimizing the Variance **6**
2.1 Introduction 6
2.2 Problem Description 9
2.3 Bayesian Sequential Update for the Variance 11
2.4 Acquisition Functions for Minimizing the Variance 12
2.5 Numerical Study 17
2.6 Case Study 22
2.7 Conclusion 26

3 Bayesian Optimization with Transformed Additive Gaussian Processes **27**
3.1 Introduction 27
3.2 Bayesian Optimization 29
3.3 Proposed Method 32
3.4 Numerical Results 36
3.5 Conclusion 39

4 Bi-Objective Bayesian Optimization with Transformed Additive Gaussian Processes **41**
4.1 Introduction 41
4.2 Background 43
4.3 Proposed Method 46
4.4 Numerical Results 50
4.5 Conclusion 53

Appendices **55**
A Proof of Results in Chapter 2 56
B Proofs of Results in Chapter 3 61
C Proofs of results in Chapter 4 63

Bibliography **67**

List of Tables

2.1 Experimental Settings of Crystallization 23

3.1 Table of parameter settings of (3.18) for the three cases in Figure 3.1 37
3.2 Table of parameter settings of (3.18) for the three cases in Figure 3.2 38
3.3 Table of parameter settings of (3.18) for the three cases in Figure 3.3 39

4.1 Table of parameter settings of (4.18) for the three cases in Figure 4.1. 51
4.2 Table of parameter settings of (4.18) for the three cases in Figure 4.2 52

List of Figures

1.1 Flowchart depicting the Bayesian optimization algorithm for blackbox functions. . . 3

2.1 A microscope image of manganese oxalate hydrate crystals synthesized from the reaction crystallization. 7
2.2 Prior Setting (1): $\hat{P}(t)$ and $\hat{C}(t)$ of the five methods with different m and K. 19
2.3 Prior Setting (2): $\hat{P}(t)$ and $\hat{C}(t)$ of the five methods with different m and K. 20
2.4 Prior Setting (3): $\hat{P}(t)$ and $\hat{C}(t)$ of the five methods with different m and K. 21
2.5 Prior Setting (4): $\hat{P}(t)$ and $\hat{C}(t)$ of the five methods with different m and K. 22
2.6 Case Study with Simulator I: $\hat{P}(t)$ and $\hat{C}(t)$ of the five methods with different m and K. 25
2.7 Case Study with Simulator II: $\hat{P}(t)$ and $\hat{C}(t)$ of the five methods with different K. . 25

3.1 The mean, 10-th, and 90-th quantiles of cost over 30 replications and 20 steps for optimizing the objective functions in (3.18) with the parameter settings in Table 3.1. 37
3.2 The mean, 10-th and 90-th quantiles of cost over 30 replications and 20 steps for optimizing the objective functions in (3.18) with the parameter settings in Table 3.2. 38
3.3 The mean, 10-th and 90-th quantiles of cost over 30 replications and 20 steps for optimizing the objective functions in (3.18) with the parameter settings in Table 3.3. 39

4.1 The mean, 10-th and 90-th quantiles of approximated hypervolume over 30 replications and 20 steps for optimizing the objective functions in (4.18) with the parameter settings in Table 4.1. 51
4.2 The mean, 10-th and 90-th quantiles of approximated hypervolume over 30 replications and 20 steps for optimizing the objective functions in (4.18) with the parameter settings in Table 4.2. 52
4.3 The mean, 10-th and 90-th quantiles of approximated hypervolume over 30 replications and 20 steps for the two problems in (4.19) and (4.20). 53
4 The resulting designs of KG-IG based on simulator I in Section 2.6 for $K = 6$ (left), $K = 12$ (middle), and $K = 30$ (right) over 60 experiments. The red dot represents the true optimal. The size of the dots is relative to the number of replications on each design point. 60
5 Illustration of dominated area for a sample EHI calculation. 66

Chapter 1

Introduction

This dissertation proposes several adaptations to the Bayesian optimization algorithm. Chapter 2 proposes a proposes a Bayesian optimization algorithm to find the experimental setting with the smallest variance. This chapter has been published as Kerfonta et al. (2024). Chapter 3 proposes a Bayesian optimization algorithm for single objective optimization problems using transformed additive Gaussian processes as the surrogate model and a modified expected improvement acquisition function. Chapter 4 also uses transformed additive Gaussian processes as the surrogate model in the Bayesian optimization algorithm, but for the bi-objective case, and a modified expected hypervolume improvement acquisition function. This chapter has been submitted for publication to the proceedings of Winter Simulation Conference 2024.

There are many different techniques for sequential designs for decision making. Sequential problems can have a discrete or continuous decision space. When the decision space is discrete decision making is often called ranking and selection. A commonly used technique when the decision space is continuous decision making is Bayesian optimization (Powell and Ryzhov, 2012). There are also frequentist approaches to sequential decision making (Xu and Zeevi, 2023).

A review of ranking and selection problems, applications, and advances is given by Hong et al. (2021). Ranking and selection methods are are classified as either fixed budget procedures or fixed precision procedures (Hunter and Nelson, 2017). For fixed budget procedures, there is a set number of experiments to be conducted and a design is created to sample alternatives so that the design is optimized (Frazier et al., 2008). When using a fixed precision procedure, designs are sampled until a certain probability of correct selection is reached (Paulson, 1964). Ranking and selection methods can be computationally expensive. In order to handle problems with large numbers of alternative designs, parallel ranking and selection procedures have been developed in the literature (Hong et al., 2022; Zhong and Hong, 2022). There are many problems that may have multiple attributes that measure performance; Butler et al. (2001) develops a ranking and selection procedure to handle this type of problem. Many ranking and selection problems evaluate a single problem, rather than performing a pairwise comparison. Xiao et al. (2023) develops a technique to perform pairwise comparisons.

A survey on recent developments to Bayesian optimization techniques is provided in Wang et al. (2023). One weakness of Bayesian optimization is that the technique does not perform well when the dimension of the decision space is large (Kandasamy et al., 2015). Chapter 3 and Chapter 4 will propose methods for Bayesian optimization with a high dimensional decision space. Spagnol et al. (2019) proposed a method to reduce the decision space through variable selection with sensitivity analysis. Bayesian optimization has also been used to solve combinatorial problems (Wang et al., 2023). This is done by handling a problem with a discrete decision space as having a continuous decision space (Garrido-Merchán and Hernández-Lobato, 2020). Another area of interest in Bayesian optimization is developing robust techniques for data with outliers, Martinez-Cantin et al. (2018) develops an algorithm using Student's T distribution to classify outliers.

We summarize the generic Bayesian optimization algorithm in a flowchart and an algorithm. Figure 1.1 shows the optimization process that is continued until a budget or a maximum number of steps is exhausted. The generic Bayesian optimization algorithm is shown in Algorithm 1 (Frazier, 2018). The surrogate model is developed in step 3 and the acquisition function is optimized in step 4. Step 5 updates the dataset and step updates the prior and posterior distributions. Chapter 4 proposes a bi-objective Bayesian optimization algorithm. In bi-objective Bayesian optimization, a surrogate model is fit to each of the objective functions and then an acquisition function that is able to handle a higher dimension is optimized (Loka et al., 2022).

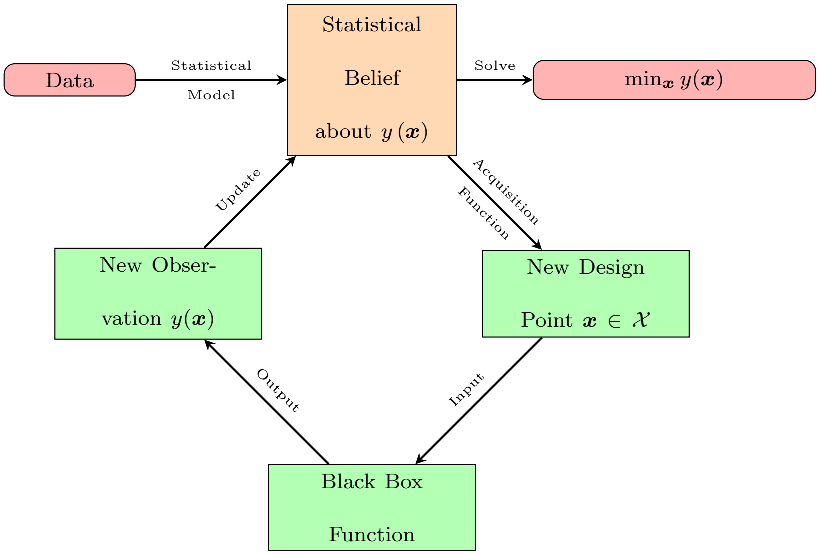
Data
Statistical
Model
Statistical
Belief
about $y(\boldsymbol{x})$
Solve
$\min_{\boldsymbol{x}} y(\boldsymbol{x})$
Update
Acquisition
Function
New Observation $y(\boldsymbol{x})$
New Design
Point $\boldsymbol{x} \in \mathcal{X}$
Output
Input
Black Box
Function

Figure 1.1: Flowchart depicting the Bayesian optimization algorithm for blackbox functions.

Algorithm 1 Generic Bayesian optimization Algorithm (Frazier, 2018)

```
Place a prior distribution on $f(\boldsymbol{x})$
for $n = 0, 1, \ldots, N-1$ do
    Compute acquisition function based on the prior distribution of $y(\boldsymbol{x})$
    Find $\boldsymbol{x}_{n+1}$ that maximizes the acquisition over $\boldsymbol{x} \in \mathcal{X}$
    Collect a new data point by evaluating $y(\boldsymbol{x})$ at $\boldsymbol{x}_{n+1}$,
    Update the prior distribution of $y(\boldsymbol{x})$ as the posterior distribution of $y(\boldsymbol{x})$ given all existing data.
end for
Return the point with the largest posterior mean.
```

Chapter 2 focuses on developing a sequential selection procedure to select the experiment with the smallest variance. That is to say it focuses on the problem

$$x^* \in \text{argmin}_{x \in \mathcal{X}} \sigma_x^2,$$

where σ_x^2's are unknown variance parameters for each of the experimental settings and there is a finite candidate set $\mathcal{X} = \{1, \ldots, K\}$. The size uniformity of crystals used in manufacturing processes can be critical quality attributes, and influential factors for quality attributes (Mascia et al., 2013; Abolhasani and Jensen, 2016; Hadiwinoto et al., 2019; McDonald et al., 2021; Mou et al., 2022). The reliable quantification of actual crystal size often still have to rely on time-consuming manual counting, for samples with overlapping crystals and shape/size difference. Therefore, it is desired to evaluate and optimize experimental conditions with a small number of experiments (and measurements). Using these crystals as motivation, the sequential selection procedure will be developed. To develop the approach, first a Bayesian adaptive method to incorporate new uniformity measurements in each step based on the conjugacy property of chi-squared and inverse gamma distributions is derived. Second, acquisition functions based on knowledge gradient as well as expected improvement are designed. A case study utilizing data from crystallization experiments will be used to show the advantages of the proposed procedure. Through numerical simulations and the case study, it is shown that the knowledge gradient and expected improvement acquisition functions for the inverse gamma prior that are proposed in this chapter are the most robust amongst the existing methods to which they are compared.

Chapters 3 and 4 introduce Bayesian optimization algorithms using transformed additive Gaussian processes as the surrogate model. Bayesian optimization is a commonly used technique to optimize complex equations (Frazier, 2018). This technique is comprised of two important parts: the surrogate model to fit the data, and the acquisition function which is optimized to determine the next location to sample. These chapters will use the Transformed Additive Gaussian Process (TAG) framework as described in Lin and Joseph (2020) to model the data and decompose the objective function(s) into one-dimensional additive pieces. They will also modify commonly used acquisition functions to accommodate the TAG framework as the surrogate model.

Chapter 3 will use the TAG framework as a surrogate model for single objective optimization problems. The problem that this chapter seeks to solve, through Bayesian optimization is, the minimization problem, which is defined as:

$$\begin{aligned} \min \quad & y(\boldsymbol{x}) \\ \text{s.t.} \quad & x \in \mathcal{X} \end{aligned}$$

where $\mathcal{X} \subset \mathcal{R}^n$ is the feasible set. Although the Bayesian optimization is derivative free, as the dimension of the problem increases, so does the computational expense and time needed to run the algorithm (Malu et al., 2021). To simplify the problem, and decrease the computational complexity, we propose fitting a TAG model to the data as the surrogate model. This framework will decompose the objective function into one-dimensional pieces that can be optimized separately. Expected Improvement is a commonly used acquisition function in Bayesian optimization (Jones et al., 1998) and we will modify it to be able to optimize the decomposed objective functions. A numerical study is conducted to show the advantages of the proposed method over Bayesian optimization with a traditional Gaussian process surrogate model and expected improvement as the acquisition function. It is shown that as the dimension of the objective function increases as does the advantage of the proposed method. The method also performs well when the objective function is transformed to be additive. If there is an interaction or the method cannot be transformed to be additive the method does not out perform traditional methods.

To further the work done in Chapter 3, Chapter 4 will use the TAG framework as a surrogate model in bi-objective Bayesian optimization. In this chapter, the problem that we seek to solve is:

$$\begin{aligned} \min \quad & \boldsymbol{y}(\boldsymbol{x}) = [y_1(\boldsymbol{x}), y_2(\boldsymbol{x})] \\ & \text{s.t.} \quad \boldsymbol{x} \in \mathcal{X} \end{aligned}$$

where $\mathcal{X} \subset \mathcal{R}^d$ is a feasible set. Just as in the single objective case, as the dimension of the objective functions increases, as does the difficulty to optimize them. A TAG model will be fit to each objective function as the surrogate model. The problem will then be decomposed into one-dimensional bi-objective problems that can be optimized separately. A commonly used acquisition function in bi-objective Bayesian optimization is expected hypervolume improvement (Emmerich et al., 2011). The proposed method will modify the expected hypervolume improvement calculation to be able to optimize the many one-dimensional bi-objective optimization problems. A numerical study using simple objective functions as well as an example from existing bi-objective optimization literature are used to demonstrate the advantages of the proposed method. As the dimension of the decision space increases, so does the advantage of the proposed method over classical bi-objective Bayesian optimization. The proposed method performs well when the objective function can be transformed to be additive, rather than having interactions or have an additive structure without being transformed.

Chapter 2

Sequential Selection for Minimizing the Variance

2.1 Introduction

Many important products, such as pharmaceuticals and batteries, contain crystals in the final products or involve crystals as manufacturing intermediates (Jiang et al., 2014; Jiang and Braatz, 2016; McDonald et al., 2021; Zambrano and Jiang, 2023). Besides the chemical composition, physical aspects of crystals are also important for product quality and consistency, such as the crystal size and the size variance (Variankaval et al., 2008; Adamo et al., 2016; Mou et al., 2022; McDonald et al., 2021). Starting from solution, crystals form from phase transition, which involves multiple possible physical and chemical phenomena (e.g., nucleation, growth, attrition, aggregation). Often, the size distribution of outcome crystals are sensitive to experimental conditions such as temperature and concentration. It is of great interest to identify the experimental conditions that consistently lead to the smallest variance of the crystal sizes, which indicates the best uniformity of crystals.

Microscope images (e.g., Figure 2.1) have a fast and convenient data format to determine crystal size quality using microscopes. However, reliably measurements of crystal sizes from microscope images often rely on time-consuming manual counting of all crystals, especially when images contain overlapping crystals and/or too many crystals. Therefore, it is still desired to optimize the experimental condition using as few trials (and microscope measurements) as possible. Additionally, reducing the number of experiments reduces materials, reliance on advanced equipment, and human operation errors under tight schedules. Thus, it is preferable to conduct the experiments in a sequential manner such that the resources can be strategically allocated to more promising experimental settings based on the information collected along the process.

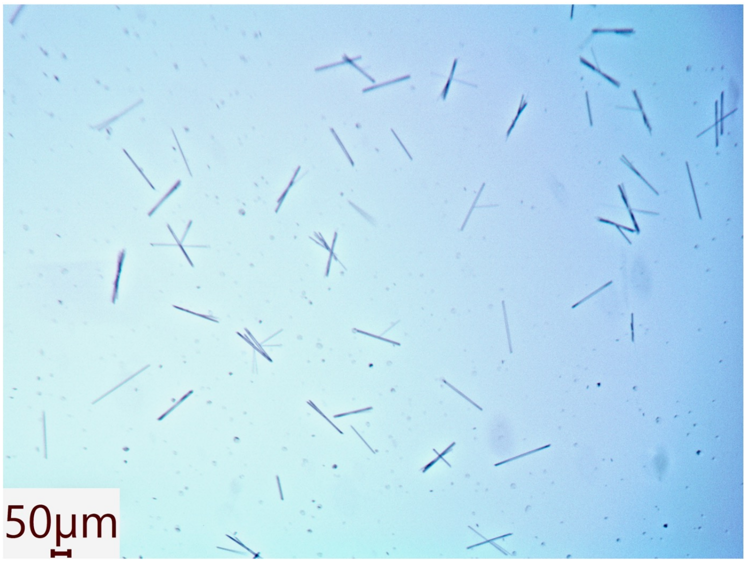

Figure 2.1: A microscope image of manganese oxalate hydrate crystals synthesized from the reaction crystallization.

Motivated by the application of crystallization experiments, our aim is to develop a sequential selection approach to find the optimal experimental condition that minimizes the variation of crystal sizes. The proposed sequential selection approach contains (1) a Bayesian adaptive method to incorporate new uniformity measurements in each step based on the conjugacy property of chi-squared and inverse gamma distributions, and (2) design acquisition functions based on knowledge gradient as well as expected improvement to improve the selection of the most promising experimental setting for minimizing the variance of crystal size. The proposed selection approach is considered a ranking and selection (R&S) problem, which is a classic mathematical framework for identifying the optimal alternative from multiple alternatives. We review the R&S literature and point out the distinction of our method.

The R&S problem can date back as early as Bechhofer (1954). There have been various schools of thought on designing budget allocation strategies in the literature of R&S. One particular school of thought is to use sequential allocation strategies under Bayesian models, for which the experimenter first spends part of the budget, collects information to update the prior belief, then decides how to allocate the remaining budget such that the expected one-step-ahead (or multiple-step-ahead) gain under certain criteria can be maximized based on the latest belief. Some examples of sequential allocation strategies following this school of thought include but not limited to expected improvement (EI; Jones et al. 1998) and knowledge gradient (KG; Gupta and Miescke 1996). The EI-type methods are broadly applied due to their computational efficiency and practical performance; for example, see Tesch et al. (2011), Wang et al. (2021) and Chen et al. (2022) for some applications of EI-type methods. Generally speaking, the EI-type methods evaluate the expected improvement over the current best estimate that would have been obtained if an alternative were selected to sample, and select the one that possesses the largest expected improvement to sample. There are plenty of developments of EI-type methods in recent years, such as knowledge gradient (Frazier et al., 2008), the expected value of information (Chick et al., 2010), and complete expected improvement (Salemi et al., 2019). The performance of the EI-type methods has also been thoroughly studied recently; for example, Ryzhov (2016) analyzes the convergence rate of classic EI and other related methods, and Chen and Ryzhov (2019) gives the first tuning-free sequential approach that is able to achieve the optimal allocation asymptotically under normal sampling distributions.

The computational tractability and efficiency of the above EI-type methods are usually achieved by assuming a certain sampling distribution (mostly, normal). Then, with normal prior (conjugate prior), the posterior distribution of the unknown parameter is also in the normal distribution family, which allows for closed-form expressions for selection criteria and sequential updates for estimators. Therefore, to efficiently identify the optimal experimental setting that produces crystals of the best uniformity, we propose an R&S framework by constructing a new EI-type method, which is a computationally-tractable sequential selection procedure with a closed-form knowledge-gradient-based criterion or an expected-improvement-based criterion to sample the alternatives. Our approach significantly distinguishes from the existing literature of R&S. First, the performance of the alternatives in our problem is the uniformity of the crystals produced and is estimated by the sample variances of the collected observations. By contrast, most existing works on R&S take the population means of the sampling distribution as the optimization objective, while only a few focus on optimizing the variance of the alternatives, such as Trailovic and Pao (2004) and Hunter and McClosky (2016). Second, we model the sample variances of the alternatives by chi-squared distributions and derive closed-form sequential updates for the estimators by assuming conjugate inverse-

Gamma prior. In fact, R&S under sampling distributions other than normal is gaining momentum recently; for example, Zhang et al. (2020) consider R&S under exponential sampling distributions, and Chen and Ryzhov (2022) propose a tuning-free sequential budget allocation strategy that can achieve the large-deviation-based optimal allocation with regard to maximizing the probability of correct selection (PCS) under general sampling distributions (Glynn and Juneja, 2004). Hence, our methodology also enriches the literature of R&S under non-normal sampling distributions.

This chapter is organized as follows. Section 2.2 will describe the problem and detail the framework of the proposed sequential selection method. Section 2.3 will describe a Bayesian adaptive method to incorporate the collected uniformity measurements in a sequential manner. Section 2.4 will provide design acquisition functions based on knowledge gradient as well as expected improvement for minimizing the variance. Section 2.5 will provide a numerical study comparing the proposed method to existing methods, and Section 2.6 will provide a case study using data from crystallization experiments. A conclusion will be given in Section 2.7.

2.2 Problem Description

Consider the problem of minimizing the variance of experimental settings within a finite candidate set $\mathcal{X} = \{1, \ldots, K\}$, i.e.,

$$x^* \in \operatorname{argmin}_{x \in \mathcal{X}} \sigma_x^2, \tag{2.1}$$

where σ_x^2's are unknown variance parameters. By conducting an experiment under a setting x, we collect a sample variance S_x^2 with a fixed sample size m to estimate the variance σ_x^2. Denote by m the number of crystals obtained in each crystallization experiment. Note that K is finite and S_x^2's are i.i.d. noisy observations of σ_x^2 for each x. Given a fixed budget, the total number of experiments that may be conducted under all settings is also fixed, thus sampling one alternative not only collects information about this alternative but also reduces the opportunity to learn other alternatives. Consequently, not every alternative can be sampled sufficiently to obtain an accurate estimator of σ_x^2. Moreover, some alternatives may not be even worth sampling many times if their performances are too poor. Taking the Bayesian perspective, we view the unknown variance parameters σ_x^2's as random variables. Our initial belief about the variances of different experimental settings is specified by their prior distributions. In particular, we assume that $\sigma_1^2, \ldots, \sigma_K^2$ are independent random variables following the inverse gamma (denoted by IG) distribution, i.e.,

$$\sigma_x^2 \sim \mathrm{IG}\left(a_x^{(0)}, b_x^{(0)}\right), \quad \text{for all} \;\; x \in \mathcal{X}, \tag{2.2}$$

where $a_x^{(0)} > 1$ and $b_x^{(0)} > 0$ are the initial prior parameters before data collection. Assume that the sample variances S_x^2's from different experimental settings are independent. Given the variance parameter σ_x^2, we model the sample variance S_x^2 by

$$(m-1)S_x^2 \mid \sigma_x^2 \sim \sigma_x^2 \chi_{m-1}^2, \tag{2.3}$$

where χ_{m-1}^2 represents a chi-squared random variable with $m-1$ degrees of freedom. Furthermore, suppose that the prior distributions and the sample variances from different experimental settings are independent. Note that the model assumption for S_x^2 holds approximately if the sample variance S_x^2 is computed based on a random sample of a fixed size m.

Let N be the total number of experiments. Denote by $x^{(t+1)}$ and $S_{x^{(t+1)}}^2$ the experimental setting selected and the sample variance collected at time t, respectively. Based on the prior distribution in (2.2) and data model in (2.3), we develop a fully sequential selection procedure to solve the target problem (2.1). In each step of the sequential procedure, we choose an experimental setting x from $\{1, \ldots, K\}$, collect the sample variance S_x^2 through experiments, and update the posterior distribution. Then, a data-driven solution to (2.1) at time t can be recursively given by

$$x^*(t) \in \mathrm{argmin}_{x \in \mathcal{X}} \mathbb{E}\left[\sigma_x^2 | S_{x^{(j+1)}}^2, j \in \mathcal{A}_x(t)\right], \tag{2.4}$$

where $\mathcal{A}_x(t) = \{j \leq t : x^{(j+1)} = x\}$ is an index set indicating the stages at which x is sampled up to time t. We say "correct selection" occurs at time t if $x^*(t)$ correctly finds x^*. We show that (2.4) can be efficiently solved by a sequential selection procedure, which consists of (1) a Bayesian update scheme for the posterior distribution based on sequential data acquisition (Section 2.3); (2) an acquisition function based on the collected information for measuring the potential improvement to the decision-making by sampling a new experimental setting (Section 2.4). To give a clear overview, we summarize the whole procedure of sequential selection in Algorithm 2, in which the proposed Bayesian update and acquisition function are applied at Step 3 and Step 5, respectively.

Algorithm 2 An Algorithm on Sequential Selection for Minimizing the Variance

```
Specify the initial prior parameters, $a_x^{(0)}$ and $b_x^{(0)}$ for all $x \in \mathcal{X}$.
for $0 \le t \le N-1$ do
    Evaluate the acquisition function for each $x \in \mathcal{X}$ given the prior parameters $a_x^{(t)}$ and $b_x^{(t)}$, and choose the experimental setting $x^{(t+1)}$ that maximizes the acquisition function over $x \in \mathcal{X}$;
    Conduct an experiment under $x^{(t+1)}$ to obtain an observation $S^2_{x^{(t+1)}}$;
    Update the prior parameters to $a_x^{(t+1)}$ and $b_x^{(t+1)}$ for all $x \in \mathcal{X}$ using $S^2_{x^{(t+1)}}$.
end for
Return the experimental setting that minimizes the expected variance at $t = N$ according to (2.4).
```

2.3 Bayesian Sequential Update for the Variance

This section constructs a Bayesian update to the posterior distribution in each step of the proposed sequential procedure. Typically, a Bayesian update is based on the conjugacy (or approximate conjugacy) of the prior and the posterior distribution, i.e., the posterior distribution belongs to the same distribution family as the prior. Lemma 1 proves the conjugacy of the inverse gamma prior for the χ^2 data model.

Lemma 1 *Assume that*

$$(m-1)S^2 \mid \sigma^2 \sim \sigma^2 \chi^2_{m-1} \quad \text{and} \quad \sigma^2 \sim \mathrm{IG}(a,b).$$

Then, the posterior distribution of σ^2 given S^2 is

$$\sigma^2 \mid S^2 \sim \mathrm{IG}\left(a + \frac{m-1}{2}, b + \frac{(m-1)S^2}{2}\right). \tag{2.5}$$

Lemma 1 implies that, for all $x \in \mathcal{X}$, the posterior distribution of σ_x^2 at time t can be denoted by $\mathrm{IG}(a_x^{(t+1)}, b_x^{(t+1)})$, where

$$a_x^{(t+1)} = a_x^{(t)} + \frac{m-1}{2}, \quad b_x^{(t+1)} = b_x^{(t)} + \frac{(m-1)S^2_{x^{(t+1)}}}{2} \quad \text{if} \quad x = x^{(t+1)} \tag{2.6a}$$

$$a_x^{(t+1)} = a_x^{(t)}, \quad b_x^{(t+1)} = b_x^{(t)} \quad \text{if} \quad x \neq x^{(t+1)}. \tag{2.6b}$$

Note that $\sigma_x^2 \sim \mathrm{IG}\left(a_x^{(t+1)}, b_x^{(t+1)}\right)$ is also the prior distribution at time $t+1$. In summary, the posterior distributions of σ_x^2 can be efficiently updated through only updating two parameters $a_x^{(t)}$ and $b_x^{(t)}$ whenever a new observation $S^2_{x(t+1)}$ is available. Consequently, (2.4) becomes

$$x^*(t) \in \mathrm{argmin}_{x \in \mathcal{X}} \left\{ \frac{b_x^{(t+1)}}{a_x^{(t+1)} - 1} \right\}, \tag{2.7}$$

where $\frac{b_x^{(t+1)}}{a_x^{(t+1)}-1}$ is simply the expected value of the posterior distribution of σ_x^2 at time t. Based on this efficient Bayesian learning model, Section 2.4 will develop acquisition functions for selecting the most promising experimental setting to sample at each step.

Remark 1 *In this framework, each sample S^2 is a noisy observation of the unknown population variance σ^2, and is computed using the length measurements of the crystals in Figure 2.1. Denote by $Y_1, \ldots, Y_m$ the lengths of m crystals and assume that they are i.i.d. samples under a normal-inverse-gamma prior model (e.g., Koch 2007):*

$$\begin{aligned} Y_i|\mu, \sigma^2 &\overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2) \quad \text{for} \quad i = 1, \ldots, m, \\ \mu|\sigma^2 &\sim \mathcal{N}(\theta, \sigma^2/\tau) \quad \text{and} \quad \sigma^2 \sim \mathrm{IG}(a, b). \end{aligned}$$

Then, the posterior marginal distribution of σ^2 will also be an inverse gamma distribution with parameters updated by

$$a \leftarrow a + \frac{m-1}{2} \quad \text{and} \quad b \leftarrow b + \frac{\sum_{i=1}^m (Y_i - \bar{Y})^2}{2} \quad \text{with} \quad \bar{Y} = \frac{1}{m}\sum_{i=1}^m Y_i.$$

Therefore, the above Bayesian model is equivalent to our model in (2.5)-(2.6) *if the variance measurement S^2 is computed as the sample variance of $Y_1, \ldots, Y_m$. Also, under the normality assumption of Y_i's, the χ^2 distribution assumption in Lemma 1 holds automatically.*

2.4 Acquisition Functions for Minimizing the Variance

In this section, we propose two acquisition functions for sequentially choosing an experimental setting to sample at each step based on the Bayesian update scheme in Section 2.3.

2.4.1 Knowledge Gradient

We first develop a knowledge-gradient-based (Frazier et al., 2008) acquisition function based on our Bayesian model. Recall that our aim is to solve problem (2.1). In each step, we obtain a data-driven solution from (2.7) by minimizing the posterior expectation of σ_x^2. Therefore, the improvement we achieve from step t to step $t+1$ by solving the target problem is

$$\min_{x\in\mathcal{X}}\left\{\frac{b_x^{(t)}}{a_x^{(t)}-1}\right\}-\min_{x\in\mathcal{X}}\left\{\frac{b_x^{(t+1)}}{a_x^{(t+1)}-1}\right\}. \tag{2.8}$$

According to the Bayesian updating equations in (2.6), at the $(t+1)$-st step, $b_x^{(t+1)}$ is a random variable depending on the incoming data $S^2_{x^{(t+1)}}$. Thus, this improvement can be assessed by taking expectations with respect to the marginal distribution of $S^2_{x^{(t+1)}}$, i.e., the distribution of $S^2_{x^{(t+1)}}$ without conditioning on σ_x^2, i.e.,

$$\mathrm{KG}^{(t+1)}(x)=\min_{x'\in\mathcal{X}}\left\{\frac{b_{x'}^{(t)}}{a_{x'}^{(t)}-1}\right\}-\mathbb{E}^{(t+1)}\left\{\min_{x'\in\mathcal{X}}\left\{\frac{b_{x'}^{(t+1)}}{a_{x'}^{(t+1)}-1}\right\}\middle| x^{(t+1)}=x\right\}, \tag{2.9}$$

where $\mathbb{E}^{(t+1)}$ denotes the expectation taken with respect to the marginal distribution of $S^2_{x^{(t+1)}}$ at step $t+1$. Then, we will sample the experimental setting that maximizes the expected improvement, i.e.,

$$x^{(t+1)}\in\mathrm{argmax}_{x\in\mathcal{X}}\mathrm{KG}^{(t+1)}(x).$$

Therefore, a closed-form expression of $\mathrm{KG}^{(t+1)}(x)$ is necessary for determining $x^{(t+1)}$ conveniently.

Now we show how to derive a closed-form expression for the selection criterion in (2.9) following the idea in Frazier et al. (2008). According to the updating equations in (2.6), given that $x^{(t+1)}=x$, we have

$$\min_{x'\in\mathcal{X}}\left\{\frac{b_{x'}^{(t+1)}}{a_{x'}^{(t+1)}-1}\right\}=\begin{cases}\frac{b_x^{(t+1)}}{a_x^{(t+1)}-1} & \text{if } \frac{b_x^{(t+1)}}{a_x^{(t+1)}-1}<\min_{x'\neq x}\left\{\frac{b_{x'}^{(t)}}{a_{x'}^{(t)}-1}\right\}\\ \min_{x'\neq x}\left\{\frac{b_{x'}^{(t)}}{a_{x'}^{(t)}-1}\right\} & \text{o.w.}\end{cases}. \tag{2.10}$$

Then, we compute the expectation in (2.9) based on the truncated expectation of the marginal distribution of $S^2_{x^{(t+1)}}$. First, we prove the following results regarding the marginal distribution of $S^2_{x^{(t+1)}}$ and its truncated expectation.

Lemma 2 *Suppose the assumptions in Lemma 1 hold. Then, the marginal distribution of S^2 follows a modified Beta Prime distribution, i.e.,*

$$\frac{m-1}{2b}S^2 \sim \mathrm{BetaPrime}\left(\frac{m-1}{2}, a\right).$$

The probability density function (PDF) of $\mathrm{BetaPrime}(\alpha, \beta)$ *is*

$$f(z) = \frac{z^{\alpha-1}(1+z)^{-\alpha-\beta}}{B(\alpha,\beta)} \quad \text{for } z > 0,\ \alpha > 0 \text{ and } \beta > 0,$$

where $B(\alpha, \beta)$ is the beta function.

Proposition 1 *Assume that*

$$Z \sim \mathrm{BetaPrime}(\alpha, \beta),$$

and denote its cumulative distribution function (CDF) by $F_{\alpha,\beta}(z)$. Then, for any fixed constant c, we have that

$$\mathbb{E}[Z \mid Z \leq c] = \frac{\alpha+1}{\beta} \cdot \frac{F_{\alpha+1,\beta-1}(c)}{F_{\alpha,\beta}(c)}. \tag{2.11}$$

Based on the updating equations and the truncated expectation in Proposition 1, we can derive a closed-form expression for (2.9) in Proposition 2, which allows us to efficiently compute KG for each experimental setting and choose the experimental setting that maximizes KG in each step.

Proposition 2 *Suppose the model assumptions in* (2.2) *and* (2.3) *hold. Then, the knowledge gradient defined in* (2.9) *can be expressed as*

$$\begin{aligned} \mathrm{KG}^{(t+1)}(x) = \min\nolimits_{x' \in \mathcal{X}} \left\{ \frac{b_{x'}^{(t)}}{a_{x'}^{(t)} - 1} \right\} - C_x^{(t)} \left[1 - F_{\frac{m-1}{2}, a_x^{(t)}} \left(\tilde{C}_x^{(t)} \right) \right] \\ - \frac{b_x^{(t)}}{a_x^{(t)} + \frac{m-1}{2} - 1} \left\{ \frac{m+1}{2a_x^{(t)}} \cdot F_{\frac{m+1}{2}, a_x^{(t)} - 1} \left(\tilde{C}_x^{(t)} \right) + F_{\frac{m-1}{2}, a_x^{(t)}} \left(\tilde{C}_x^{(t)} \right) \right\}, \end{aligned} \tag{2.12}$$

where

$$C_x^{(t)} = \min\nolimits_{x' \neq x} \left\{ \frac{b_{x'}^{(t)}}{a_{x'}^{(t)} - 1} \right\} \quad \text{and} \quad \tilde{C}_x^{(t)} = \frac{\left(a_x^{(t)} + \frac{m-1}{2} - 1 \right) C_x^{(t)}}{b_x^{(t)}} - 1.$$

Remark 2 *In our problem, the value of KG can actually be non-positive. For example, if $\tilde{C}_x^{(t)} \leq 0$,* *then both* $F_{\frac{m-1}{2},a_x^{(t)}}$ *and* $F_{\frac{m+1}{2},a_x^{(t)}-1}$ *will be zero, which leads to*

$$\mathrm{KG}^{(t+1)}(x) = \min_{x' \in \mathcal{X}} \left\{ \frac{b_{x'}^{(t)}}{a_{x'}^{(t)} - 1} \right\} - \min_{x' \neq x} \left\{ \frac{b_{x'}^{(t)}}{a_{x'}^{(t)} - 1} \right\} \leq 0.$$

Essentially, the occurrence of a non-positive KG is due to the non-negative support of the chi-squared distribution, because the probability that a single new sample can further improve (decrease) the current best (smallest) posterior mean can get close to 0 if the current best posterior mean itself is close to 0. For example, suppose the current best alternative $x^*(t)$ *is unique. From the definition of KG in* (2.9)*, whether there will be a negative KG for* $x^*(t)$ *depends on how likely a new sample of* $x^*(t)$ *will exceed its current posterior mean, i.e.,* $\frac{b_{x^*(t)}^{(t)}}{a_{x^*(t)}^{(t)}-1}$*. Due to the non-negative support of the chi-squared distribution, if* $\frac{b_{x^*(t)}^{(t)}}{a_{x^*(t)}^{(t)}-1}$ *is small and close to 0, then the probability for a new sample of* $x^*(t)$ *to exceed* $\frac{b_{x^*(t)}^{(t)}}{a_{x^*(t)}^{(t)}-1}$ *will be quite large and get close to 1. In this case, from* (2.10)*, the probability that* $\min_{x' \in \mathcal{X}} \left\{ \frac{b_{x'}^{(t+1)}}{a_{x'}^{(t+1)}-1} \right\}$ *is larger than* $\frac{b_{x^*(t)}^{(t)}}{a_{x^*(t)}^{(t)}-1}$ *will also be close to 1. Consequently, this will make the second term on the RHS of* (2.9) *larger than* $\frac{b_{x^*(t)}^{(t)}}{a_{x^*(t)}^{(t)}-1}$*, which is exactly the first term on the RHS of* (2.9)*. Then, the KG value of* $x^*(t)$ *will be negative. A negative KG in this situation means that no positive improvement may be expected by sampling* x*, then simply taking* $x^{(t+1)} \in \mathrm{argmax}_{x \in \mathcal{X}} \mathrm{KG}^{(t+1)}(x)$ *may potentially make the sequential selection procedure fail to sample all alternatives infinitely often as* $t \to \infty$ *(equivalent to* $N \to \infty$*), which is necessary for guaranteeing PCS* $\to 1$*. Therefore, we make a modification to* $\mathrm{KG}^{(t+1)}(x)$ *in* (2.12) *to solve this issue: (1) if* $\tilde{C}_x^{(t)} \leq 0$ *for some but not all* x*, then such* x *will not be selected as* $x^{(t+1)}$*; and (2) if* $\tilde{C}_x^{(t)} \leq 0$ *for all* $x \in \mathcal{X}$*, we will randomly select one alternative from* $\mathcal{X}$ *as* $x^{(t+1)}$*.*

2.4.2 Weighted Expected Improvement

The drawback of the knowledge gradient criterion is that its value is not guaranteed to be non-negative. Thus, we provide an alternative acquisition function that is always non-negative and guarantees to sample all alternatives infinitely often. We design this acquisition function based on the Expected Improvement (EI) criterion from Jones et al. (1998). First, the non-negative improvement that can be obtained by sampling x at time t is given by

$$\max\left\{\min_{x\in\mathcal{X}}\left(\frac{b_x^{(t)}}{a_x^{(t)}-1}\right)-S_x^2,0\right\}.$$

In other words, we will only factor in the improvement from a new sample when it is positive; otherwise, we will just consider it zero. Taking expectations with respect to the marginal distribution of S_x^2, the EI criterion is given by

$$\mathrm{EI}^{(t+1)}(x)=\mathbb{E}\left[\max\left\{\min_{x\in\mathcal{X}}\left(\frac{b_x^{(t)}}{a_x^{(t)}-1}\right)-S_x^2,0\right\}\right]. \tag{2.13}$$

Based on Lemma 2 and Proposition 1, we can also obtain a closed-form expression of $\mathrm{EI}^{(t+1)}(x)$.

Proposition 3 *Suppose the model assumptions in* (2.2) *and* (2.3) *hold. Then, the expected improvement defined in* (2.13) *can be expressed as*

$$\begin{aligned}\mathrm{EI}^{(t+1)}(x)=&\min_{x\in\mathcal{X}}\left\{\frac{b_x^{(t)}}{a_x^{(t)}-1}\right\}F_{\frac{m-1}{2},a_x^{(t)}}\left(\frac{m-1}{2b_x^{(t)}}\min_{x\in\mathcal{X}}\left\{\frac{b_x^{(t)}}{a_x^{(t)}-1}\right\}\right)\\&-\frac{(m+1)b_x^{(t)}}{(m-1)a_x^{(t)}}F_{\frac{m+1}{2},a_x^{(t)}-1}\left(\frac{m-1}{2b_x^{(t)}}\min_{x\in\mathcal{X}}\left\{\frac{b_x^{(t)}}{a_x^{(t)}-1}\right\}\right),\end{aligned}$$

where $F_{\frac{m-1}{2},a}$ is the CDF of the Beta Prime distribution with parameters $\frac{m-1}{2}$ and a.

Although the above EI criterion can properly measure the non-negative improvement from a sample, it does not account for the variance of the collected samples. Therefore, following the framework in Gramacy (2020), we scale (2.13) by the posterior variance of σ_x^2 at time t and call the resulting criterion weighted Expected Improvement (wEI):

$$\mathrm{wEI}^{(t+1)}(x)=\mathrm{Var}\left[\sigma_x^2|S^2_{x^{(j+1)}},j\in\mathcal{A}_x(t)\right]\cdot\mathrm{EI}^{(t+1)}(x), \tag{2.14}$$

where $\mathrm{Var}\left[\sigma_x^2|S^2_{x^{(j+1)}},j\in\mathcal{A}_x(t)\right]=\left(b_x^{(t)}\right)^2\left(a_x^{(t)}-1\right)^{-2}\left(a_x^{(t)}-2\right)^{-1}$ due to the inverse gamma posterior. Then, we will sample the experimental setting that maximizes (2.14), i.e.,$x^{(t+1)}\in$ $\mathrm{argmax}_{x\in\mathcal{X}}\mathrm{wEI}^{(t+1)}(x)$.

2.5 Numerical Study

We conduct several numerical experiments to compare the empirical performance of our approach with three other sequential selection approaches from the existing literature for minimizing the variance:

1. Knowledge Gradient with Inverse Gamma Prior (KG-IG): The proposed method in this chapter. The priors are updated using (2.5), and the acquisition function for sequential selection is given by Proposition 2 in Section 2.4.1.

2. Weighted Expected Improvement with Inverse Gamma Prior (wEI-IG): The proposed method in this chapter. The acquisition function for sequential selection is given by (2.14) in Section 2.4.2.

3. Knowledge Gradient with Normal Prior (KG-Normal): This method is conducted using the knowledge gradient for the normal distribution. As described in Frazier et al. (2008), the prior distribution is assumed to be normal:

   $$\log(\sigma_x^2) \sim \mathcal{N}\left(\theta_x^{(0)}, \delta_x^{2,(0)}\right),$$

   and the distribution of new measurements is assumed to be

   $$\log\left((m-1)S_x^2\right)|\sigma_x^2 \sim \mathcal{N}\left(\log(\sigma_x^2), \lambda_x^2\right).$$

   Note that the first order Taylor's expansion of $\log(\sigma_x^2)$ at the mean of σ_x^2 gives

   $$\log(\sigma_x^2) \approx \log\left(\frac{b_x^{(0)}}{a_x^{(0)}-1}\right) + \frac{a_x^{(0)}-1}{b_x^{(0)}}\left(\sigma_x^2 - \frac{b_x^{(0)}}{a_x^{(0)}-1}\right).$$

   Thus, the prior parameters are set below as the mean and variance of the first-order Taylor's expansion of $\log(\sigma_x^2)$:

   $$\theta_x^{(0)} = \log\left(\frac{b_x^{(0)}}{a_x^{(0)}-1}\right) \quad \text{and} \quad \delta_x^{2,(0)} = \frac{1}{a_x^{(0)}-2}.$$

   Also, λ_x^2 is set as the variance of the log-transformed random samples from the χ_{m-1}^2 distribution.

4. Random selection with Inverse Gamma prior (Rand-IG): We use the proposed Bayesian update under the inverse gamma prior described in (2.6). The selection of experimental settings in each step is random. To ensure a balanced design, we replicate each experimental setting N/K times and randomize the order of the N experiments.

5. Thompson Sampling (Russo et al., 2018) with Inverse Gamma prior (TS-IG): We use the proposed Bayesian update under the inverse gamma prior as described in (2.6). In each step, we generate one random sample from each inverse gamma prior in (2.2), i.e.,

$$\sigma_x^{2,(t)} \sim \mathrm{IG}(a_x^{(t)}, b_x^{(t)}), \quad \text{for all} \quad x \in \mathcal{X},$$

   and choose the experimental setting that minimizes these random samples to sample next, i.e., $x^{(t+1)} \in \mathrm{argmin}_{x\in\mathcal{X}} \sigma_x^{2,(t)}$.

The true variances σ_x^2's are generated as independent uniform variables from $(0, 2)$. We run the sequential selection procedure for $N = 60$ steps and compare the performances of these approaches for different values of m and K. For each sequential selection method, we repeatedly use its acquisition approach to determine $x^{(t+1)}$, generate a new sample from the distribution $\sigma_x^2\chi_{m-1}^2$ and update the posterior distribution, until the budget of N experiments is exhausted. We run R macro-replications with true variances regenerated in every replication. For clarity, let $x_i^*(t)$ denote the estimate $x^*(t)$ at time t in the i-th replication, and define $\mathbb{I}_i(t)$ as a binary indicator that is equal to 1 if $x_i^*(t) = x^*$ and 0 otherwise. The performances of different approaches are evaluated in terms of the average opportunity cost $\hat{C}(t)$ and the estimated probability of correct selection $\hat{P}(t)$ over R macro-replications, where

$$\hat{P}(t) = \frac{1}{R}\sum_{i=1}^{R} \mathbb{I}_i(t) \quad \text{and} \quad \hat{C}(t) = \frac{1}{R}\sum_{i=1}^{R} \left(\sigma^2_{x_i^*(t)} - \sigma^2_{x^*}\right). \tag{2.15}$$

Note that $\hat{C}(t)$ is always non-negative. Intuitively, $\hat{P}(t)$ measures how likely $x_i^*(t)$ successfully identifies x^* and $\hat{C}(t)$ measures how close the variance of $x_i^*(t)$ is to that of x^*.

We compare four prior settings: (1) $a_x^{(0)} = 2.1$ and $b_x^{(0)}$'s are i.i.d. random variables generated from $U(0.5, 1)$; (2) $a_x^{(0)} = 2.1$ and $b_x^{(0)}$'s are i.i.d. random variables generated uniformly from $(2.5, 5)$; (3) $a_x^{(0)} = 2.1$ and $b_x^{(0)}$'s are i.i.d. random variables generated uniformly from $(0.25, 0.5)$; (4) $a_x^{(0)} = m$ and $b_x^{(0)} = \sigma_x^2 G_x$, where G_x's are i.i.d. random variables generated from the distribution χ_{m-1}^2.

Priors (1)-(3) are non-informative prior, whereas prior (4) carries the information from a single sample for each alternative.

In Figures 2.2-2.5, we show $\hat{P}(t)$ (left) and $\hat{C}(t)$ (right) for $t = 1, \ldots, 60$ over R macro-replications for the four prior settings. We use $R = 100$ for prior settings (1)-(3) and $R = 1000$ for prior setting (4), since prior setting (4) has higher noise level and thus more replications are needed to make the estimates $\hat{P}(t)$ and $\hat{C}(t)$ converge. Also, to exhibit $\hat{C}(t)$ clearly, we depict its 4th root $(\hat{C}(t))^{1/4}$ instead of its original value. Several immediate observations follow. First, the proposed methods, KG-IG and wEI-IG, have the best overall performance. Specifically, both methods have the most robust performance across different prior settings. Second, KG-Normal performs slightly better than TS-IG, and both of them have inferior performance in contrast to Rand-IG under prior settings (1) and (2). Third, cases with $K = 30$ are more challenging, and the performance of different approaches can be better distinguished. Lastly, among the three non-informative priors, the best overall performance for all methods is observed in prior setting (3), while KG-IG, TS-IG, KG-normal and wEI-IG perform similarly well under the informative prior setting (4).

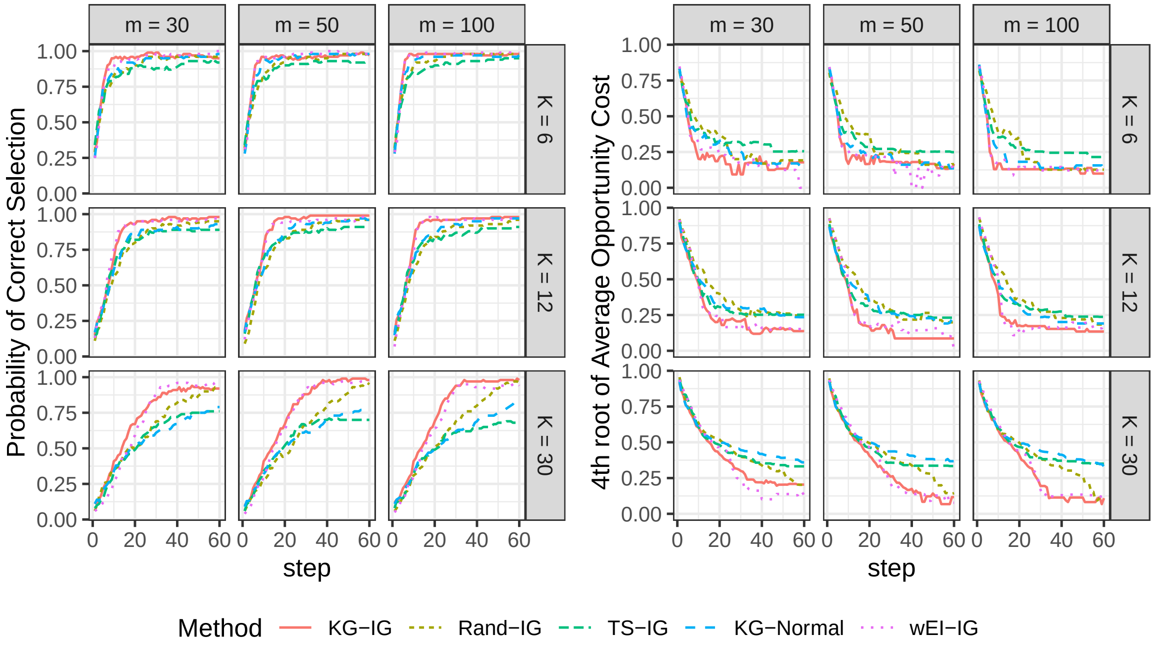

Figure 2.2: Prior Setting (1): $\hat{P}(t)$ and $\hat{C}(t)$ of the five methods with different m and K.

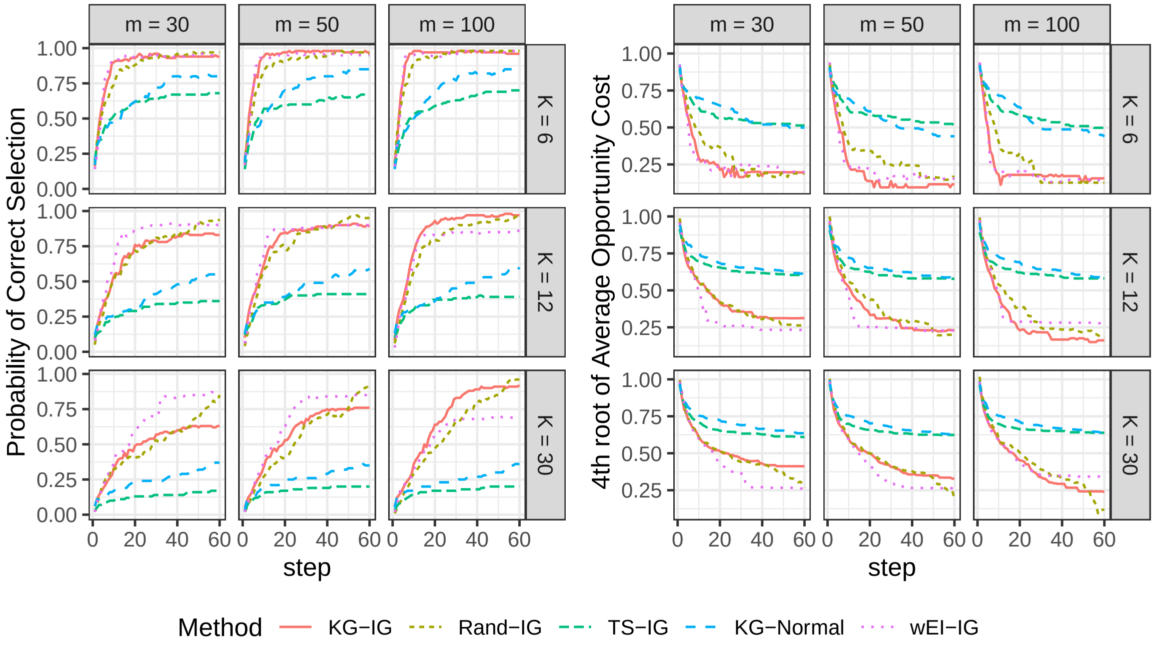

Figure 2.3: Prior Setting (2): $\hat{P}(t)$ and $\hat{C}(t)$ of the five methods with different m and K.

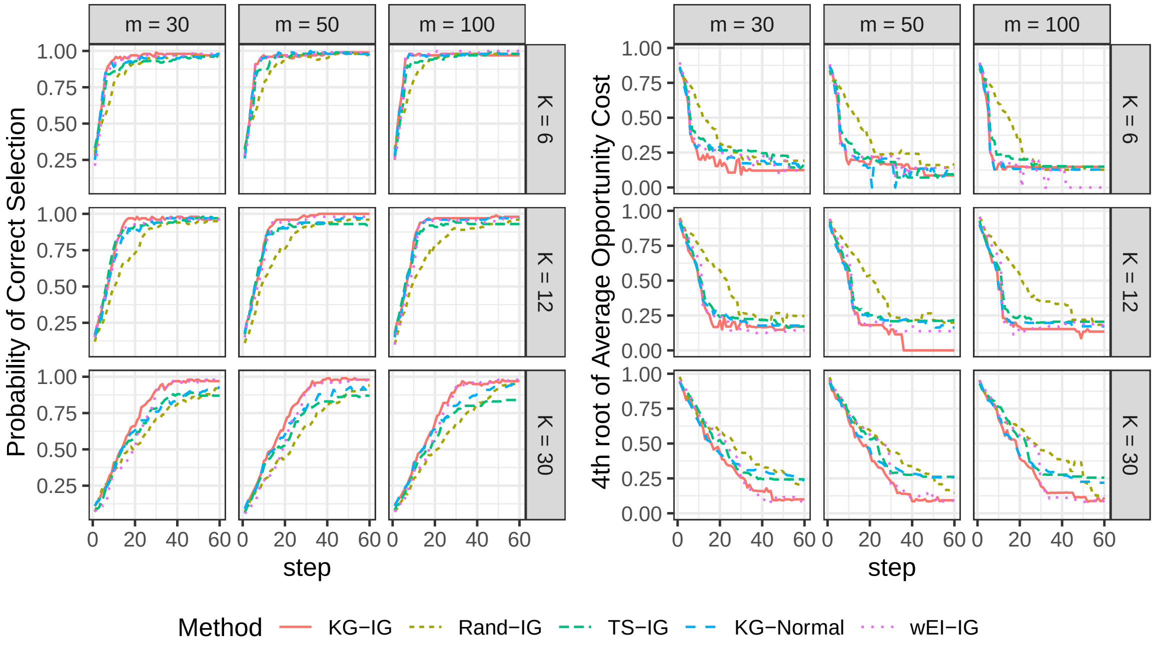

Figure 2.4: Prior Setting (3): $\hat{P}(t)$ and $\hat{C}(t)$ of the five methods with different m and K.

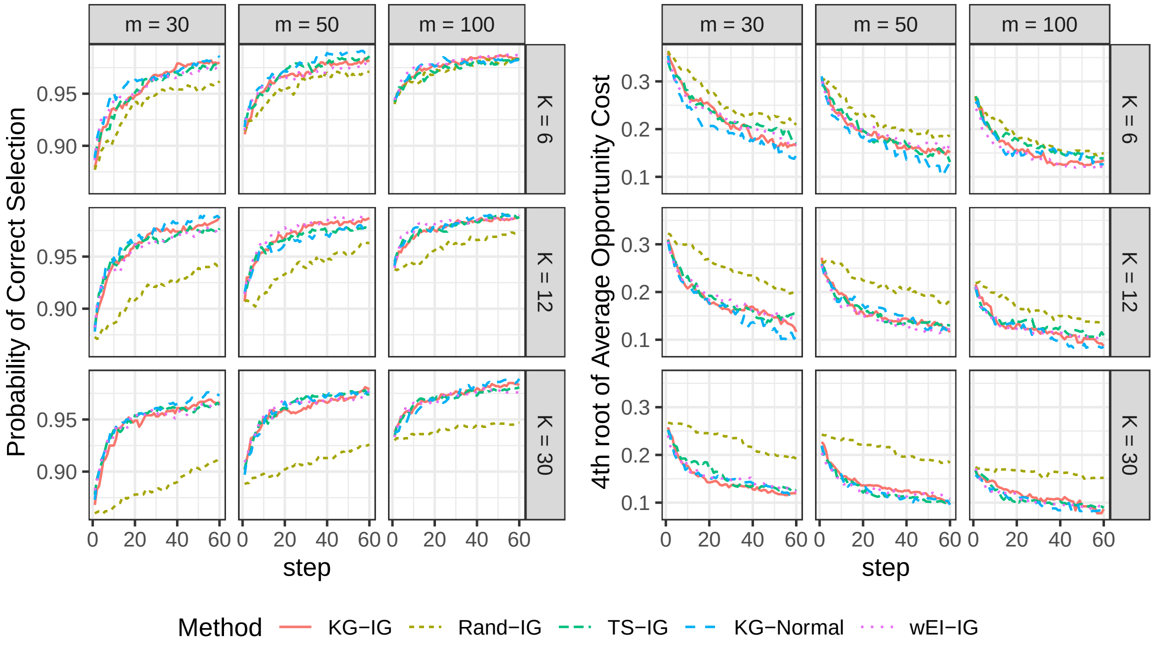

Figure 2.5: Prior Setting (4): $\hat{P}(t)$ and $\hat{C}(t)$ of the five methods with different m and K.

2.6 Case Study

We consider a case study about the crystallization experiment described in Section 2.1. The experimental settings are given in Table 2.1. The total number of unique experimental settings is 21, and each time one trial under a selected experimental setting is conducted. The total number of level combinations of the two discrete factors Temp and Conc is six due to experimental constraints and limitations. In each trial, we obtain an image of needle-like crystals (usually 30-100 crystals in each image, as shown in Figure 2.1), measure the sizes of the crystals and compute the sample variance as the uniformity score.

Table 2.1: Experimental Settings of Crystallization

Experimental Setting	Explanation	Levels or Ranges
Temp. (C)	Temperature	20, 30, 40, 50
Conc. (M)	Concentration in molarity	0.025, 0.03, 0.035
R.T. (min)	Resident time for crystallization	8-40
G.F. (ml/min)	Gas Flow Rate	1.5-15

As shown in Section 2.5, a large number of macro-replications, i.e., R in (2.15), is typically required to make valid assessments across different methods. However, it is infeasible to replicate the whole experiment under different settings sufficiently in real-world experiments. Therefore, we build two different types of pseudo-simulators based on a real dataset to mimic the real-world experiments. The simulators are used to sequentially generate new measurements of uniformity or variance based on a selected experimental setting. For both simulators, we assume m is a constant in all experiments. In practice, a fixed value of m such as in (2.9) can be approximated beforehand using the average number of crystals when computing the acquisition function and deciding the new experimental setting, and once the observations are collected, the actual value of m can be used in parameter update (2.6). We develop two simulators in this chapter. Simulator I is based on the predictions from the fitted Gaussian process of the real data. This simulator can simulate pseudo experiments for settings that are not available in real data. The use of Gaussian process may smooth the original distribution of real data. Simulator II is based on resampling of real observations. Then this simulator can only generate pseudo experiments for the alternatives that are available in the real data. More details about the two simulators are given as follows.

Simulator I: This simulator is developed by fitting a Gaussian process (Roustant et al., 2015) to the real dataset. First, we scale the range of experimental conditions (Temperature, Concentration, R.T., and G.F.) to be between 0 and 1. Second, we fit a Gaussian process (GP) model with the scaled experimental conditions as the input and the logarithm of the uniformity measure (sample variance divided by the square of sample mean) of the experiments as the output. We specify the experimental settings of the K (taking value from $\{6, 12, 30\}$ for three cases) candidates as follows: we first generate a $K \times 2$ Latin Hypercube design as the values for R.T. and G.F., and then replicate the six-level combinations of temperature and concentration for $K/6$ times. The resulting

K experimental settings are given by a $K \times 4$ matrix. We use the fitted GP simulator to predict the output σ_x^2 of the candidate settings and generate a new sample by $\sigma_x^2 \chi_{m-1}^2$ in each step as a simulated random uniformity measure.

Simulator II: This simulator is developed by resampling the crystals obtained from each experiment. We fix $K = 21$ or 10 (a subset of the 21 original experimental settings) as the experimental alternatives. If one experimental setting is selected, we will randomly sample m crystals with replacement (e.g., sample $m = 30$ crystals from Figure 2.1), and use the sample variance of the lengths of these m crystals as the uniformity measure of this experiment.

We run $R = 100$ macro-replications as in Section 2.5 to obtain $\hat{C}(t)$ and $\hat{P}(t)$ according to (2.15). We set non-informative initial priors with $a_x^{(0)} = 2.1$ and $b_x^{(0)}$ generated independently from $Q \times U(0.5, 1)$, where $Q = 0.2$ and 20 for simulators I and II respectively and $U(0.5, 1)$ represents the uniform distribution on $(0.5, 1)$. The value of Q is specified based on the scale of the uniformity measures from the two simulators. Using different m and K as in Section 2.5, Figure 2.6 and 2.7 exhibit $\hat{P}(t)$ and $\hat{C}(t)$ for all five methods with simulators I and II, respectively. We give an example of the resulting design given by the proposed method in Appendix A.6. The results show that KG-IG and wEI-IG are the most robust among the five methods in both situations.

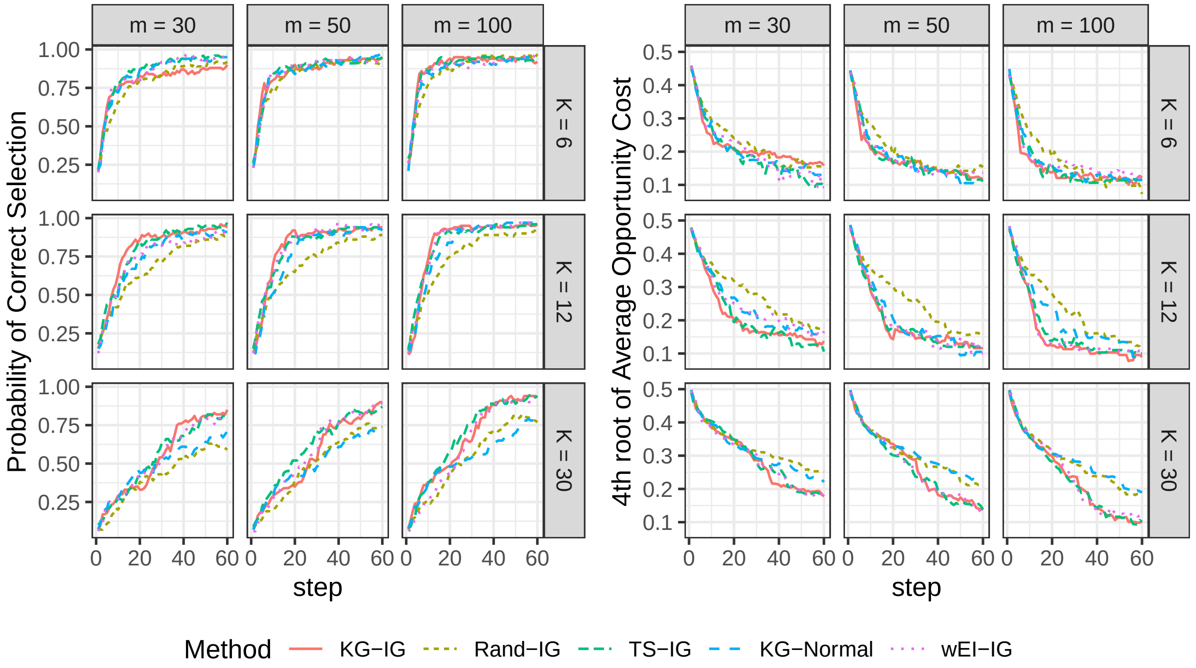

Figure 2.6: Case Study with Simulator I: $\hat{P}(t)$ and $\hat{C}(t)$ of the five methods with different m and K.

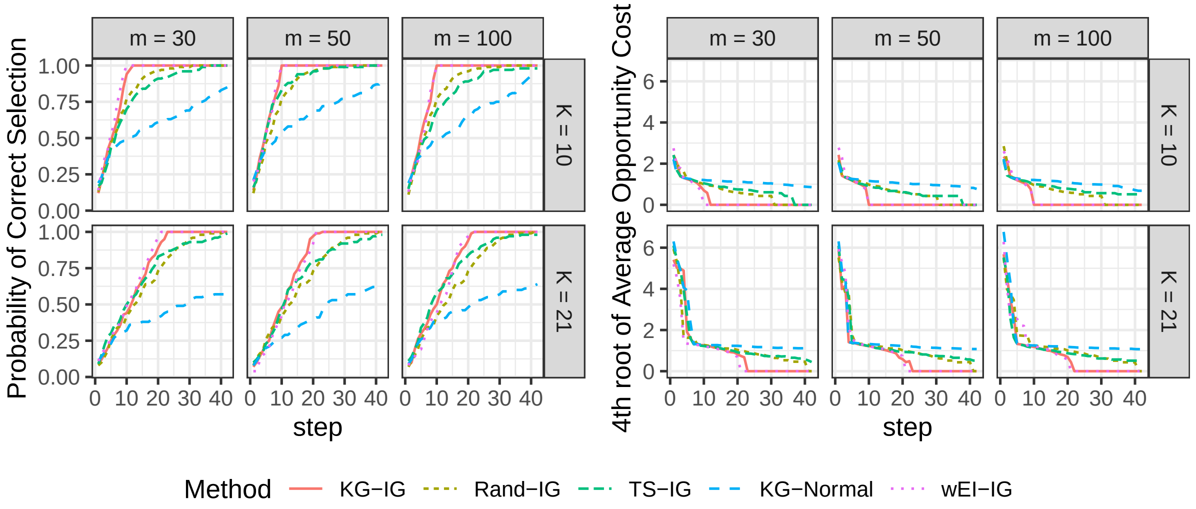

Figure 2.7: Case Study with Simulator II: $\hat{P}(t)$ and $\hat{C}(t)$ of the five methods with different K.

2.7 Conclusion

This chapter has developed a new methodology for selecting experimental settings to minimize the variance. Though improving the uniformity (minimizing the variability) is of great interest in many situations such as crystallization experiments, it has been relatively overlooked in the existing R&S literature. Our approach has constructed an efficient Bayesian sequential updating framework for learning the variances under the conjugate inverse gamma prior. Furthermore, we have proposed two sequential selection procedures based on knowledge gradient as well as expected improvement for selecting the experimental settings, and have derived closed-form expressions for both proposed selection criteria. Through numerical experiments and case studies, we demonstrate that the proposed approach is able to identify the experimental setting with the minimum variance and exhibits competitive empirical performance compared with other approaches.

There are several promising future directions. First, our framework may be extended to solve multiobjective optimization, especially under nonnormal sampling distributions that may arise in applications such as crystallization, reaction engineering and materials engineering. Furthermore, R&S under nonnormal sampling distributions is a positive research area, and our approach specifically provides a new idea for designing efficient budget allocation strategies under chi-squared sampling distributions, which we believe can inspire many future works in this area.

Chapter 3

Bayesian Optimization with Transformed Additive Gaussian Processes

3.1 Introduction

Tradespace decomposition can introduce many complex and interesting optimization problems. The objective functions of these optimization problems are often blackbox functions. A black box function is a function without a known closed form expression; where the relationship between the response and the decision space is unknown Jones et al. (1998). Black box functions can be complex and difficult to solve (Alarie et al., 2021). Some black-box functions may not have a closed form solution and approximation techniques may be required in order to approximate a solution (Vu et al., 2017). One way that these problems are simplified is through the use of Bayesian optimization (Wang et al., 2023). Bayesian optimization has two main components: the surrogate model and the acquisition function. The surrogate model is used to model the data. Gaussian processes (GP) are commonly used as surrogate models. The acquisition function is optimized to determine the next location to sample a point from. A commonly used acquisition function is expected improvement. One important and useful property of Bayesian optimization is that it is derivative free. This means that in order to optimize the acquisition function (optimizing function), we do not need to use a higher order approximation method (Frazier, 2018; Shahriari et al., 2016). This decreases the time and computational complexity of the optimization problem. Another useful property of this method of optimization is that it can handle noisy data (Brochu et al., 2010). Bayesian optimization works quite well for these functions, however it begins to struggle when the dimensionality is increased. Although the method is derivative free, the complexity and computation expense and time is greatly increased in higher dimensional problems (Malu et al., 2021). Methods to lower the dimensionality of an objective function to simplify Bayesian optimization have been proposed. These methods include using non-linear feature mapping and reconstruction mapping to lower the dimensionality (Moriconi et al., 2020), using variable dropout methods (Li et al., 2018), and ensemble Bayesian optimization (Wang et al., 2018).

A Gaussian process is a distribution of collection of random variables such that any finite subset of the collection follow a multivariate normal distribution (Schulz et al., 2018). Gaussian Processes are often used because they include noise and randomness in the model. They utilize a correlation function to determine the strength of the correlation between two points (Jäkel et al., 2007). A commonly used correlation function, called the Gaussian correlation function produces smooth local trends (MacKay et al., 1998). The process of Bayesian optimization can be used to optimize objective functions that are modeled by Gaussian Processes. Using this method, the prior distribution that is placed on the function y, will be a Gaussian Process. Gaussian Processes are very useful models in optimization because they can account for noise and randomness (Zhang and Notz, 2015). Using Gaussian processes for optimization can be very useful because they allow for uncertainty quantification (McLeod et al., 2018). When working with Gaussian processes, we are able to write the conditional distributions to be written in a closed form (Snoek et al., 2012). This property is particularly useful because the joint distributions are multivariate normal an the conditional distributions are also Gaussian processes.

Acquisition functions can be complex or computationally inexpensive. When using Bayesian optimization, there are many different acquisition functions that can be used in order to decrease complexity and minimize computation (Wilson et al., 2018a). A commonly used acquisition function in Bayesian optimization is expected improvement. We define expected improvement as in Jones et al. (1998). This acquisition function measures the expected improvement over the current minimum when a new point is added.

In this chapter, we propose an approach to optimizing complex black-box functions with an additive structure. This is done by using the Transformed Additive Gaussian process (TAG) framework proposed in (Lin and Joseph, 2020) as the surrogate model. We will then propose a modification to the expected improvement acquisition function to be able to accomodate the TAG framework as the surrogate model.

This chapter will be organized as follows. Section 3.2 will describe the Bayesian optimization framework that will be used and adapted. Section 3.3 will describe the proposed framework using transformed additive Gaussian processes with the expected improvement acquisition function. Section 3.4 will use a numerical study to compare the proposed method to the classical Bayesian optimization framework. A conclusion and statement of future work will be given in Section 3.5.

3.2 Bayesian Optimization

Bayesian optimization is comprised of two components, the surrogate model to represent the data and the acquisition function which is optimized to determine where to sample next. In this section, we will discuss the components of traditional Bayesian optimization with a Gaussian process surrogate model and Expected Improvement acquisition function. The minimization problem is defined as:

$$\begin{aligned} \min \quad & y(\boldsymbol{x}) \\ \text{s.t.} \quad & \boldsymbol{x} \in \mathcal{X} \end{aligned} \tag{3.1}$$

where $\mathcal{X} \subset \mathcal{R}^d$ is the feasible set. The image of the feasible set, $\mathcal{Y} := \{y(\boldsymbol{x}) \in \mathcal{R} : \boldsymbol{x} \in \mathcal{X}\}$, is referred to as the outcome set (Boyd and Vandenberghe, 2004). The generic Bayesian algorithm is shown in Algorithm 3. A prior distribution for the data is given in step 1. Then, the acquisition function is used in step 2 to find the location to sample next. In step 3, the point is collected and the surrogate model is updated.

3.2.1 Gaussian Process

When implementing Bayesian optimization, the surrogate model is used to represent the data, which could be obtained from a blackbox function. Gaussian processes are one of the most

Algorithm 3 Generic Bayesian Optimization Algorithm

```
Provide a prior distribution of $y(\boldsymbol{x})$
for $n = 0, 1, \ldots, N-1$ do
    Compute acquisition function based on the distribution of $y(\boldsymbol{x})$ and find $\boldsymbol{x}_{n+1}$ that maximizes the acquisition over $\boldsymbol{x} \in \mathcal{X}$
    Collect a new data point by evaluating $y(\boldsymbol{x})$ at $\boldsymbol{x}_{n+1}$ and update the distribution of $y(\boldsymbol{x})$ as the posterior distribution of $y(\boldsymbol{x})$ given all existing data.
end for
Return the input point with the smallest posterior mean as the optimal decision.
```

used surrogate models Sacks et al. (1989). Let

$$Z(\boldsymbol{x}) \sim \mathrm{GP}(0, \sigma^2 R(\cdot)),$$

be a mean zero Gaussian process with variance σ^2 and correlation function $R(\cdot)$. Next, let $Y(\boldsymbol{x})$ be a stochastic process and $y(\boldsymbol{x})$, be a realization of that stochastic process, then:

$$Y(\boldsymbol{x}) = \mu + Z(\boldsymbol{x}), \tag{3.2}$$

where μ is the unknown deterministic mean value. The Gaussian correlation function (Sacks et al., 1989) is a very popular choice when working with Gaussian processes. The correlation between two inputs points $\boldsymbol{x}$ and $\boldsymbol{x}'$ is given by

$$R\left(\boldsymbol{x} - \boldsymbol{x}'\right) = \exp\left(-\sum_{i=1}^{d} \theta_i (x_i - x_i')^2\right), \tag{3.3}$$

where θ_i's are the unknown correlation parameters. The unknown correlation parameter represents the correlation between each dimension of $\boldsymbol{x}$. When a Gaussian process is fit to data, these parameters need to be estimated. Under Gaussian processes assumptions with the Gaussian correlation function, given two input points $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, the model can be represented as

$$Y(\boldsymbol{x}) \sim N(\mu, \sigma^2) \quad \text{and} \quad \mathrm{Cov}(Y(\boldsymbol{x}), Y(\boldsymbol{x}')) = \sigma^2 R(\boldsymbol{x} - \boldsymbol{x}').$$

Following these functions, if they are evaluated at n input points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the outputs $\boldsymbol{y} = (y(\boldsymbol{x}_1) \ldots, y(\boldsymbol{x}_n))^\top$ are produced. Another useful aspect of Gaussian processes, is that at a new input point $\boldsymbol{x}$, the conditional distribution of $Y(\boldsymbol{x})$ given $\boldsymbol{y}_n$ follows the normal distribution (Jones et al., 1998). This distribution is given by:

$$Y(\boldsymbol{x}) \mid \boldsymbol{y} \sim N(\hat{y}(\boldsymbol{x}), s^2(\boldsymbol{x})), \tag{3.4}$$

with mean:

$$\hat{y}(\boldsymbol{x}) = \hat{\mu} - \boldsymbol{r}_n(\boldsymbol{x})\boldsymbol{R}_n^{-1}(\boldsymbol{y}_n - \mathbf{1}\hat{\mu}) \tag{3.5}$$

and variance:

$$s^2(\boldsymbol{x}) = \sigma^2\left(1 - \boldsymbol{r}_n(\boldsymbol{x})^T\boldsymbol{R}_n^{-1}\boldsymbol{r}_n(\boldsymbol{x})\right) - \frac{(1 - \mathbf{1}^\top\boldsymbol{R}_n^{-1}\boldsymbol{r}_n(\boldsymbol{x}))^2}{\mathbf{1}^\top\boldsymbol{R}_n\mathbf{1}}, \tag{3.6}$$

where $\mathbf{1}$ is a vector of ones of size n, the mean is estimated by the unbiased estimator $\hat{\mu} = \frac{\mathbf{1}^\top\boldsymbol{R}_n\boldsymbol{y}_n}{\mathbf{1}^\top\boldsymbol{R}_n\mathbf{1}}$, $\boldsymbol{r}(\boldsymbol{x}) = (R(\boldsymbol{x} - \boldsymbol{x}_1), \ldots, R(\boldsymbol{x} - \boldsymbol{x}_n))^\top$, and $\boldsymbol{R}$ be the $n \times n$ correlation matrix with the (i, j)-th element $R(\boldsymbol{x}_i - \boldsymbol{x}_j)$.

3.2.2 Expected Improvement

Acquisition functions can be complex or computationally inexpensive (Wilson et al., 2018b). When using Bayesian optimization, there are many different acquisition functions that can be used in order to decrease complexity and minimize computation (Wilson et al., 2018a). One of the most commonly used acquisition functions that can be used in Bayesian optimization, is expected improvement. We define expected improvement as in (Jones et al., 1998)

$$EI\left(\boldsymbol{x}\right) = \mathrm{E}\left[\max\left(y^{min} - \hat{y}(\boldsymbol{x}), 0\right)\right], \tag{3.7}$$

where y^{min} is the current minimum and $\hat{y}(x)$ and $\hat{s}(\boldsymbol{x})$ are the values generated at the current iteration. Then when using Gaussian processes, and taking the expectation with respect to the conditional distribution in (3.4), expected improvement can be rewritten in the closed form as

$$\mathrm{EI}\left(\boldsymbol{x}\right) = \left(y^{min} - \hat{y}(\boldsymbol{x})\right)\Phi\left(\frac{y^{min} - \hat{y}(\boldsymbol{x})}{\hat{s}(\boldsymbol{x})}\right) + \hat{s}(\boldsymbol{x})\phi\left(\frac{y^{min} - \hat{y}(\boldsymbol{x})}{\hat{s}(\boldsymbol{x})}\right), \tag{3.8}$$

where $\Phi(\cdot)$ is the cumulative probability distribution of the normal distribution and $\phi(\cdot)$ is the probability density function of the normal distribution.

3.3 Proposed Method

In this section, we propose an adaptation to the Bayesian Optimization algorithm using Transformed Additive Gaussian Processes as the surrogate model and a modified expected improvement acquisition function. The data will be modeled using the transformed additive Gaussian process framework, it will be broken down into one-dimensional pieces. These pieces will then be optimized using a modification of the expected improvement acquisition function given in 3.7.

3.3.1 Surrogate Modeling with Transformed Additive Gaussian Processes

The goal of Transformed Additive Gaussian processes (TAG) is to take a high dimensional response and use a transformation function to make the response the addition of lower or single dimensional functions Lin and Joseph (2020). Assume that $y(\boldsymbol{x})$ is a realization of a transformed additive Gaussian process that is used as a surrogate model to the optimization problem in 3.2. Then the optimization problem can be written as

$$y = g_\lambda^{-1}\left(\mu + z_1(x_1) + \ldots + z_d(x_d)\right),$$

where λ is the transformation parameter of the Box-Cox transformation Box and Cox (1964), which is given by :

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases} \tag{3.9}$$

Then, the TAG model can be written as:

$$g_\lambda(y) = \mu + z\left(\boldsymbol{x}\right) + \epsilon\left(\boldsymbol{x}\right), \tag{3.10}$$

where $z(\boldsymbol{x})$ follows a GP with a mean of 0 and covariance $\tau^2 R(\cdot)$, that is to say: $z\left(\boldsymbol{x}\right) \sim GP\left(0, \tau^2 R(\cdot)\right)$. Then, $z_k\left(x_k\right) \sim GP\left(0, \tau_k^2 R_k(\cdot)\right)$ is the prior distribution placed on all $z_k(\cdot)$. Let the correlation between $g_\lambda(y(\boldsymbol{x}))$ and $g_\lambda(y(\boldsymbol{x'}))$ be

$$R(\boldsymbol{x} - \boldsymbol{x'}) = \sum_{k=1}^{p} \omega_k R_k\left(x_k - x'_k\right),$$

where $\tau^2 = \sum_{k=1}^{d} \tau_k^2$ and $\omega_k = \tau_k^2 \tau^2$. The Gaussian correlation function in 3.3, could be used as the correlation function in this model. Assume that the additive noise $\epsilon(\boldsymbol{x}) \sim N(0, \sigma^2)$ for this model is independent from the distributions of the $z_k(x_k)'s$. The unknown parameters, λ, τ, ω, and σ are estimated through empirical Bayes estimation and non-linear optimization as noted in Lin and Joseph (2020).

3.3.2 Conditional Distribution of Transformed Additive Gaussian Processes

In order to optimize each objective function, the conditional distribution of each dimension given the data, $z_k(x_k)|\boldsymbol{y}$, can be optimized individually and then combined together to generate a new point. The conditional distribution of each $z_k(x_k)$ given the data is needed to calculate the expected improvement for each of the p one-dimensional additive pieces.

Proposition 4 *Assume that data has been collected from the evaluated objective functions, which can be written as $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$. Let $\boldsymbol{r}_k(x_k)$ be the vector of correlations: $(R(x_k - x_{1k}), \ldots R(x_k - x_{nk}))$ and $\boldsymbol{R}$ be the correlation matrix with the ijth element being $R_k(\boldsymbol{x}_i - \boldsymbol{x}_j)$. Then under the TAG framework (Lin and Joseph, 2020),*

$$\begin{pmatrix} g_\lambda\left(y(\boldsymbol{x}) - \mu\mathbf{1}\right) \\ z_k(x_k) \end{pmatrix} \sim N_{n+1}\left[\begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2\boldsymbol{R} + \sigma^2\boldsymbol{I} & \tau_k^2\boldsymbol{r}_k(x_k) \\ \tau_k^2\boldsymbol{r}_k(x_k)^T & \tau_k^2 \end{pmatrix}\right] \tag{3.11}$$

for $k = 1, \ldots, p$, where $\mathbf{1}$ is the vector containing all 1's. Then, the conditional distribution of $z_k(x_k)$ given $\boldsymbol{y}$ can be obtained through the conditional distribution for the normal distribution from Eaton (2007). The conditional distribution is given by

$$z_k(x_k)|\boldsymbol{y} \sim N\left(\hat{z}_k(x_k), s_k^2(x_k)\right) \tag{3.12}$$

with

$$\hat{z}_k(x_k) = \omega_k \boldsymbol{r}_k(x_k)^\top \left(\boldsymbol{R} + \delta\boldsymbol{I}\right)^{-1}\left(g_\lambda(\boldsymbol{y}) - \mu\right)$$

and

$$s_k^2(x_k) = \tau^2\omega_k\left(1 - \boldsymbol{r}_k(x_k)^\top\left(\boldsymbol{R} + \delta\boldsymbol{I}\right)^{-1}\omega_k\boldsymbol{r}_k(x_k)\right),$$

where $\delta = \sigma^2/\tau^2$.

3.3.3 Additive Expected Improvement

Now that we have the conditional distributions of each dimension of the surrogate model, we need an acquisition function with which we can optimize the objective function and generate new points. Our proposed method uses transformed additive Gaussian processes to decompose the objective function into one-dimensional pieces and then use a modified expected improvement acquisition function to optimize each dimension separately. The expected improvement acquisition function as described in 3.2.2 will be utilized in each dimension. Now, using the TAG model, the objective function can be rewritten as

$$g_\lambda\left(y(\boldsymbol{x})\right) \approx \mu + \sum_{k=1}^{d} z_k\left(x_k\right). \tag{3.13}$$

The optimization problem in (3.1) can be simplified to

$$\min_{x_k \in \mathcal{X}_k} \quad \boldsymbol{z}_k(x_k) \quad \text{for} \quad k = 1, \ldots, d, \tag{3.14}$$

where $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$. Then, the modified expected improvement acquisition function for the optimization problem in (3.14) can be defined as

$$\begin{aligned} \mathrm{EI}^k\left(x_k\right) &= \mathrm{E}\left[\max\left(y_k^{min} - \hat{y}(x_k), 0\right)\right] & (3.15)\\ &= \left(y_k^{min} - \hat{y}(x_k)\right)\Phi\left(\frac{y_k^{min} - \hat{y}(x_k)}{s_k(x_k)}\right) + s_k(x_k)\phi\left(\frac{y_k^{min} - \hat{y}(x_k)}{s_k(x_k)}\right). & (3.16)\end{aligned}$$

To generate a new point, $\mathrm{EI}^k\left(x_k\right)$ is optimized for each dimension the objective function. A new design point $\boldsymbol{x}_{new} = (x_1^*, \ldots, x_d^*)$ is selected by optimizing

$$x_k^* = \max\nolimits_{x \in \mathcal{X}_k} \mathrm{EI}^k\left(x_k\right) \quad \text{for} \quad i = 1, \ldots, d.$$

In the next section of this chapter, we propose an algorithm to implement the TAG surrogate model and the proposed additive EI acquisition function.

3.3.4 The Proposed Algorithm

For the proposed method, the surrogate model is a TAG model from (3.13) evaluated using the TAG R package (Lin and Joseph, 2021). For each iteration of the algorithm, a new set of candidate points will be generated using a Latin Hypercube design in the `lhs` R package (Carnell, 2022) and a TAG model will be fit to the data, $\mathcal{X}_m$. The mean and variance for each $z_k(x_k)$ will be calculated using the conditional distribution of each z_k. These values will then be used to calculate the expected improvement for each dimension of the input space for each candidate point. The value that maximizes expected improvement for each dimension of the input space x_k^* will be combined to generate the new data point: $\boldsymbol{x}^{(m+1)} = (x_1^*, \ldots, x_d^*)$. The new data point will be added to the existing dataset and the process will continue until the budget is exhausted.

This process is summarized in Algorithm 3.3.4. The surrogate model is fit in step 3 and the acquisition function is optimized in step 6. The data is updates in step 10. When implementing the algorithm, it was seen that it is not necessary to regenerate the TAG models in every iteration of the algorithm. This can be computationally expensive. In our numerical study, we will regenerate the TAG models after every 10 iterations of the algorithm. This will be done to save computational cost and time.

Algorithm 4 Additive Bayesian Optimization with TAG

```
1:  Given the data $\mathcal{X}_n$, we:
2:  for $m = n, n+1, \ldots, N$ do
3:      Fit the TAG model based on $\mathcal{X}$ and find $z_k(x_k)$ for $k = 1, \ldots, d$
4:      Generate a $q \times d$ matrix $X$ as a random Latin Hypercube design with $ij$-th entries $X_{ij}$.
5:      for $k = 1, \ldots, d$ do
6:          Evaluate EI for $\hat{\boldsymbol{x}}$
7:          Select point $X_k^*$ that maximizes EI for $\hat{\boldsymbol{x}}$
8:      end for
9:      Set new input point $\boldsymbol{x}^{(m+1)} = (X_1^*, \ldots, X_d^*)$ and obtain $\boldsymbol{y}(\boldsymbol{x}^{(m+1)})$
10:     Update $\mathcal{X}_m \to \mathcal{X}_{m+1}$
11: end for
```

3.4 Numerical Results

In this section, we will compare the proposed method with the classical Bayesian optimization methods described in (Roustant et al., 2012). This method uses Gaussian process based optimization using the approach described in (Jones et al., 1998) and (Mockus, 2005), we will call this the "EGO" approach. In each step of this classical Gaussian process approach a Gaussian process model is fit to the data using all of the design points generated until that iteration. The models will be fit using GP kriging models as the surrogate models, evaluated using the DiceKriging R package (Roustant et al., 2012). Next, a new point is chosen by maximizing the expected improvement acquisition function. The process will be run using the functions in the `DiceOptim` R package (Picheny et al., 2021).

In our simulation, the candidate points in each iteration, for both methods will be generated using a random Latin hypercube design as described in the `lhs` R package (Stein, 1987). We run the procedure for $N = 20$ steps, varying the number of initial data points in order to compare different aspects of the method's performance. For each step, we use expected improvement to select $x^{(t+1)}$. For each setting we run $R = 30$ macro replications using the same initial dataset for both methods in each replication. We update the distributions until our budget N steps is reached.

Let $y\left(x_i^{(t)}\right)$ be the estimate of y at step t, for replication i and let x^*, be the value that produces the true minimum value of the objective function. The metric that will be used for comparison is cost and is defined as

$$\hat{C}(t) = \frac{1}{R}\sum_{i=1}^{R}\left(y\left(x_i^*(t)\right) - y\left(x^*\right)\right). \tag{3.17}$$

This value will always be non-negative and measures how close to the true minimum the method's approach is as the algorithm runs. The confidence intervals will be constructed using the 10^{th} and 90^{th} quantiles of the cost for each iteration. The default range of λ in (3.9) in the `TAG` package is -2 to 2, we use this default range. To analyze the performance of the proposed method, we will vary the number of points in the initial dataset, the dimension of the objective functions, and the degree to which the objective functions are additive. The objective function that we will use for our numerical study is

$$y(\boldsymbol{x}) = \exp\left(\Sigma_{k=1}^{d}\left(x_k - a\right)^2 + cx_1x_2\right). \tag{3.18}$$

This objective function is easily transformed to be additive using the Box-Cox transformation (3.9), when $c = 0$ and $\lambda = 0$.

First, we compare the performance of the proposed method when the amount of data points in the initial dataset varies. The setting of the objective function for this comparison is found in Table 3.1 and the results are shown in Figure 3.1. It can be seen that our method has an advantage for the EGO approach for each number of initial points. However, the proposed method has a greater advantage when there is a smaller number of initial points.

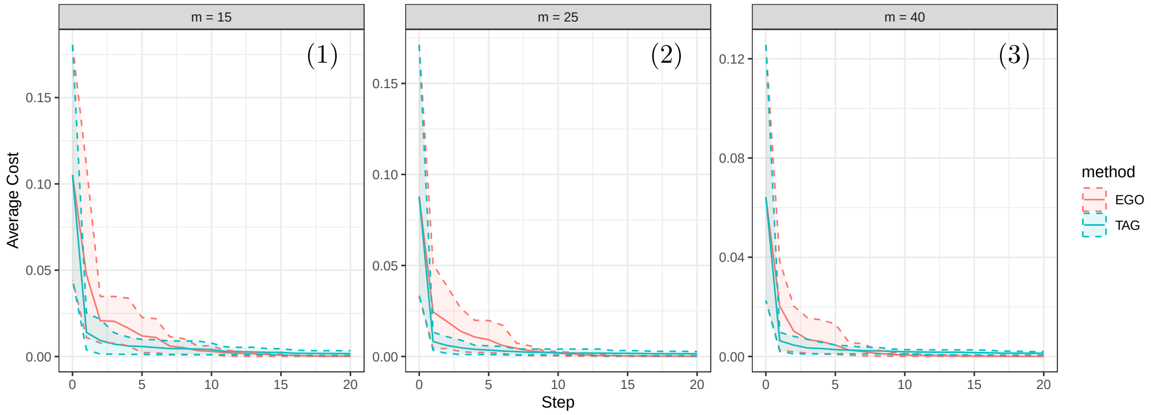

Figure 3.1: The mean, 10-th, and 90-th quantiles of cost over 30 replications and 20 steps for optimizing the objective functions in (3.18) with the parameter settings in Table 3.1.

Table 3.1: Table of parameter settings of (3.18) for the three cases in Figure 3.1

(1)				(2)				(3)			
d	a	c	m	d	a	c	m	d	a	c	m
4	0.3	0	15	4	0.3	0	25	4	0.3	0	40

.

Second, we compare the performance of the proposed method over different dimensions of the objective function. The setting of the objective function for this comparison is found in Table 3.2 and the results are shown in Figure 3.2. It can be seen that as the dimension of the objective increases, as does the advantage of the proposed method over the EGO approach.

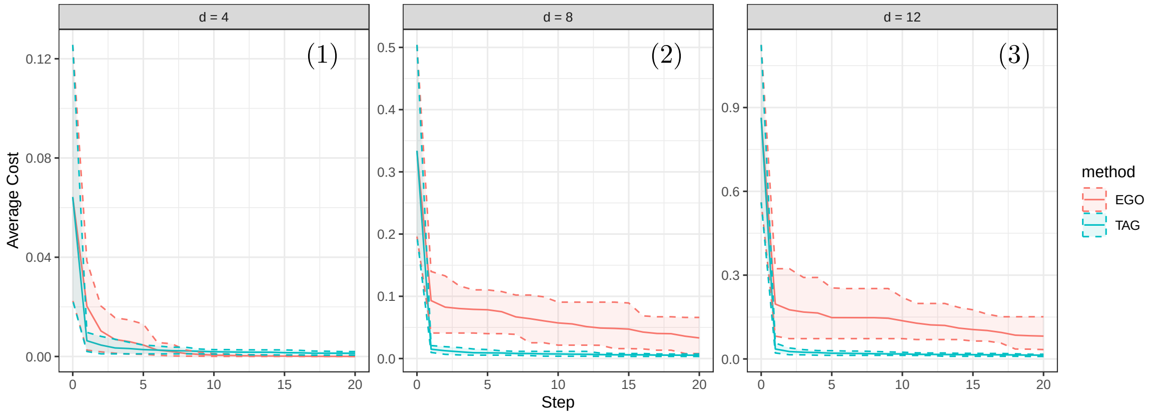

Figure 3.2: The mean, 10-th and 90-th quantiles of cost over 30 replications and 20 steps for optimizing the objective functions in (3.18) with the parameter settings in Table 3.2.

Table 3.2: Table of parameter settings of (3.18) for the three cases in Figure 3.2

(1)				(2)				(3)			
d	a	c	m	d	a	c	m	d	a	c	m
4	0.3	0	40	8	0.3	0	40	12	0.3	0	40

Finally, we compare the performance of the proposed method when the objective function can not exactly be transformed to be additive, this is done by adding an interaction term to the objective function. The setting of the objective function for this comparison is found in Table 3.3 and the results are shown in Figure 3.3. It can be seen that as the contribution of the interaction term increases, the proposed TAG method does not have an advantage over the EGO approach.

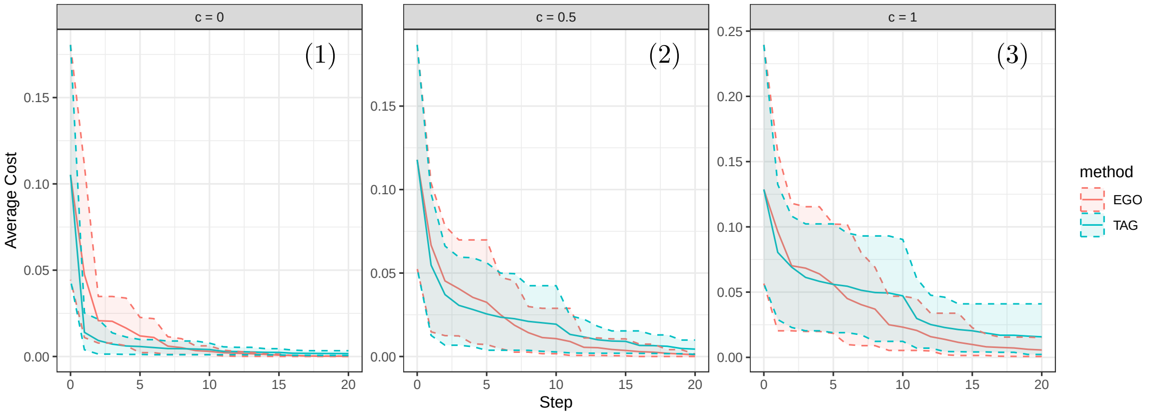

Figure 3.3: The mean, 10-th and 90-th quantiles of cost over 30 replications and 20 steps for optimizing the objective functions in (3.18) with the parameter settings in Table 3.3.

Table 3.3: Table of parameter settings of (3.18) for the three cases in Figure 3.3

(1)				(2)				(3)			
d	a	c	m	d	a	c	m	d	a	c	m
4	0.3	0	15	4	0.3	0.5	15	4	0.3	1	15

.

3.5 Conclusion

This chapter has proposed a new adaptation to the Bayesian optimization algorithm. It utilizes the TAG framework to decompose the objective function into smaller pieces. It also proposes a modification to the expected improvement acquisition function to optimize these additive pieces. We have conducted a numerical study to demonstrate the advantages of the proposed method as compared to the EGO approach. The method has a greater advantage when the amount of initial data is lower. As the dimension of the decision space increases, so does the advantage of the proposed method. When there is an interaction in the objective function, that is to say when the objective function is not transformed to be additive, the proposed method does not have an advantage over the EGO approach. Future and parallel work include expanding to the bi-objective and multi-objective cases.

Acknowledgements

This work was supported by Clemson University's Virtual Prototyping of Autonomy Enabled Ground Systems (VIPR-GS), under Cooperative Agreement W56HZV-21-2-0001 with the US Army DEVCOM Ground Vehicle Systems Center (GVSC).

Disclaimers

All opinions, conclusions and findings wherein are those of the authors and may not be those of the affiliated institutions DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited. OPSEC # 8425.

Chapter 4

Bi-Objective Bayesian Optimization with Transformed Additive Gaussian Processes

4.1 Introduction

Simulation is a popular approach used to optimize engineering and scientific design (Lee et al., 2007; Zhang, 2008; Fliege and Xu, 2011; Hunter and McClosky, 2016; Hunter et al., 2019; de Castro et al., 2022). For complex engineering design problems, running the corresponding simulation codes is often time-consuming. Also, the input and output relationship of the simulation model may not be expressed in a closed form to enable gradient-based optimization approaches. Thus, Bayesian optimization is often used to solve these problems by efficiently guiding new experiments through running the simulation models (Frazier, 2018). Bayesian optimization approaches contain two key components: a computationally inexpensive surrogate model of the simulation and an acquisition function to select input points for the follow-up experiments. In terms of surrogate models, Gaussian process is a common choice (e.g., Jones et al. (1998)). A popular choice of the acquisition function is the expected improvement (EI, e.g, Qin et al. (2017)), which is the expected absolute difference between the current optimal objective value and the optimal objective value computed after adding a new simulation evaluation (i.e., the improvement with respect to the target optimization problem). Then the input point for the new evaluation is given as a maximizer of EI over the input space.

Some engineering design problems involve two or more objectives. For multi-objective problems, the solution improvement can not be simply defined as for the single objective problems. In the literature of multi-objective optimization, the quality of the solution given by an optimization approach can be quantified by a variety of performance indicators, see various concepts given by (Audet et al., 2021). Among them, the hypervolume indicator is a popular performance indicator to measure the quality of the solutions to multi-objective optimization problems (Knowles et al., 2003). Hence, in the literature on multi-objective Bayesian optimization (Hunter et al., 2019), a popular choice for the acquisition function is the expected hypervolume improvement (EHI), which is the expected difference between the hypervolume indicators computed after and before adding a new simulation evaluation (Emmerich et al., 2011). Even though EHI is widely used, there are some computational challenges to maximizing this acquisition function. When the dimension of the decision space increases, the optimization of EHI becomes much more complex, requires longer computing time, and takes more memory space to complete (Hupkens et al., 2015). There are a number of existing approaches proposed to reduce the complexities of maximizing EHI. Yang et al. (2019) consider bi-objective Bayesian optimization and provide an algorithm to evaluate the gradient function of EHI to enhance the optimization of EHI for searching new input points. A truncated form of the EHI calculation to optimize computing time is also proposed by Yang et al. (2016). This approach uses prior knowledge about the objective function values to more efficiently compute EHI. Since the gradient of EHI can be extremely complex and difficult to compute, Daulton et al. (2020) provide a method for computing EHI by using the gradients of the Monte Carlo estimator via auto-differentiation. This method simplifies the calculation of EHI by using first-order and quasi-second-order methods.

In this chapter, we provide an algorithm for bi-objective Bayesian optimization to solve the computational challenges of maximizing EHI. In contrast to perspectives in current literature, we approximate the objective functions to additive functions for each dimension of the decision space using the transformed additive Gaussian processes (Lin and Joseph, 2020). Then the maximization of EHI over the entire decision space becomes the maximization of multiple one-variable EHI functions, i.e., the number of EHI functions is equal to the number of dimensions of the input space. Maximizing EHI with a single input can be efficiently done by enumerating on a set of discrete points. We compare the numerical performance of the proposed approach with the bi-objective Bayesian optimization under the classical Gaussian process model.

This chapter is organized as follows. Section 4.2 gives a review of related background, including single-objective Bayesian optimization with Gaussian processes and EI, bi-objective optimization and the hypervolume indicator, and bi-objective Bayesian optimization with Gaussian processes and EHI. Section 4.3 proposes our method, which uses transformed additive Gaussian process (TAG) models as the surrogate models and derives the EHI expression under this model assumption. Section 4.4 provides a numerical study that demonstrates the advantages of the proposed method. Section 4.5 concludes the chapter and points out directions for future improvement of the proposed method.

4.2 Background

In this section, we review related background approaches including bi-objective optimization, the hypervolume indicator, and bi-objective Bayesian optimization with classical Gaussian processes and EHI.

4.2.1 Bi-objective Optimization

Let $y_1(\boldsymbol{x})$ and $y_2(\boldsymbol{x})$ be two scalar-valued deterministic black-box functions with input $\boldsymbol{x} \in \mathcal{R}^d$. The bi-objective optimization problem is defined by

$$\begin{aligned} \min \quad & \boldsymbol{y}(\boldsymbol{x}) = [y_1(\boldsymbol{x}), y_2(\boldsymbol{x})] \\ \text{s.t.} \quad & \boldsymbol{x} \in \mathcal{X} \end{aligned} \tag{4.1}$$

where $\mathcal{X} \subset \mathcal{R}^d$ is a feasible set. Euclidean spaces $\mathcal{R}^d$ and $\mathcal{R}^2$ are respectively referred to as the decision space and the objective (outcome) space. Given two input points $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, the outcome $\boldsymbol{y}(\boldsymbol{x})$ is said to dominate the outcome $\boldsymbol{y}(\boldsymbol{x}')$, denoted as $\boldsymbol{y}(\boldsymbol{x}) \prec \boldsymbol{y}(\boldsymbol{x}')$, if and only if $y_l(\boldsymbol{x}) \leq y_l(\boldsymbol{x}')$ for $l = 1, 2$ and $y_l(\boldsymbol{x}) < y_l(\boldsymbol{x}')$ for $l = 1$ or $l = 2$. If $y_l(\boldsymbol{x}) \leq y_l(\boldsymbol{x}')$ for $l = 1, 2$, the outcome $\boldsymbol{y}(\boldsymbol{x})$ is said to weakly dominate the outcome $\boldsymbol{y}(\boldsymbol{x}')$, denoted as $\boldsymbol{y}(\boldsymbol{x}) \preceq \boldsymbol{y}(\boldsymbol{x}')$. An outcome point is said to be nondominated if there does not exist another outcome point dominating it. To solve problem (4.1), we apply the classical concept of Pareto-optimality, that is to identify the set of all input points whose images are nondominated outcomes Ehrgott (2005). The Pareto set is defined by

$$\mathcal{P} = \{\boldsymbol{y}(\boldsymbol{x}) \in R^2 : \boldsymbol{x} \in \mathcal{X} \mid \nexists\, \boldsymbol{x}' \in \mathcal{X} \quad \text{s.t.} \quad \boldsymbol{y}(\boldsymbol{x}') \prec \boldsymbol{y}(\boldsymbol{x})\}. \tag{4.2}$$

Since the Pareto set is typically not available in a closed form even if the objective and constraint functions are available, various computational approaches and algorithms have been developed to provide approximation of different types (Ruzika and Wiecek, 2005; Herzel et al., 2021). Due to the diversity of these methods and the resulting sets being exactly or only approximating the Pareto set, many ways to measure their quality have been proposed (Faulkenberg and Wiecek, 2010; Audet et al., 2021). One way to measure the quality of the computed Pareto set is the hypervolume indicator (e.g., Knowles et al. (2003); Guerreiro et al. (2021)). Given a reference point $\boldsymbol{t} = (t_1, t_2) \in \mathcal{R}^2$, the hypervolume indicator $\mathrm{H}(\mathcal{P}, \boldsymbol{t})$ is the two-dimensional Lebesgue measure Λ of the region weakly dominated by $\mathcal{P}$ and bounded above by the reference point $\boldsymbol{t}$, i.e.,

$$\mathrm{H}(\mathcal{P}, \boldsymbol{t}) = \Lambda\left(\{\boldsymbol{v} \in R^2 \mid \nexists\, \boldsymbol{p} \in \mathcal{P} : \boldsymbol{p} \preceq \boldsymbol{v} \text{ and } \boldsymbol{v} \preceq \boldsymbol{r}\}\right). \tag{4.3}$$

The hypervolume improvement (HVI) Hupkens et al. (2015) can be used to assess the effect of evaluating the vector-valued objective function $\boldsymbol{y}(\boldsymbol{x})$ at a new input point $\boldsymbol{x} \in \mathcal{X}$:

$$\mathrm{HVI}\left(\boldsymbol{y}(\boldsymbol{x}), \mathcal{P}, \boldsymbol{t}\right) = \mathrm{H}\left(\mathcal{P} \cup \{\boldsymbol{y}(\boldsymbol{x})\}, \boldsymbol{t}\right) - \mathrm{H}\left(\mathcal{P}, \boldsymbol{t}\right). \tag{4.4}$$

4.2.2 Bi-objective Bayesian Optimization with EHI

In bi-objective Bayesian optimization, the expected value of the hypervolume improvement (EHI) in (4.4) is a commonly used acquisition function Emmerich et al. (2011). Given that the two objectives $y_1(\boldsymbol{x})$ and $y_2(\boldsymbol{x})$ are modeled by two independent Gaussian processes $Y_1(\boldsymbol{x})$ and $Y_2(\boldsymbol{x})$ as in (3.2), we provide the expression of EHI.

Assume that we evaluated the two objectives at n input points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, and collected the outputs $\boldsymbol{y}_{n,1} = (y_1(\boldsymbol{x}_1) \ldots, y_1(\boldsymbol{x}_n))^\top$ and $\boldsymbol{y}_{n,2} = (y_2(\boldsymbol{x}_1) \ldots, y_2(\boldsymbol{x}_n))^\top$. Following (3.4), we have

$$Y_1(\boldsymbol{x})|\boldsymbol{y}_{n,1} \sim N(\hat{y}_1(\boldsymbol{x}), \hat{s}_1^2(\boldsymbol{x})) \quad \text{and} \quad Y_2(\boldsymbol{x})|\boldsymbol{y}_{n,2} \sim N(\hat{y}_2(\boldsymbol{x}), \hat{s}_2^2(\boldsymbol{x})), \tag{4.5}$$

where $\hat{y}_l(\boldsymbol{x})$ and $\hat{s}_l^2(\boldsymbol{x})$ for $l = 1, 2$ are the conditional means and variances, which can be calculated using (3.5) and (3.6). Under the assumption that $Y_1(\boldsymbol{x})$ and $Y_2(\boldsymbol{x})$ are independent, $Y_1(\boldsymbol{x})|\boldsymbol{y}_{n,1}$ and $Y_2(\boldsymbol{x})|\boldsymbol{y}_{n,2}$ are also independent. Note that the two Gaussian processes $Y_1(\boldsymbol{x})$ and $Y_2(\boldsymbol{x})$ can have different parameters and correlation functions. Given a reference point $\boldsymbol{t}$, EHI is defined by

$$\mathrm{EHI}_n\left(\boldsymbol{x};\mathcal{P},\boldsymbol{t}\right)=\mathrm{E}\left\{\mathrm{HVI}\left((Y_1(\boldsymbol{x}),Y_2(\boldsymbol{x})),\mathcal{P},\boldsymbol{t}\right)\right\}, \tag{4.6}$$

where HVI is given by (4.4) with the new point $(Y_1(\boldsymbol{x}),Y_2(\boldsymbol{x}))$ and the expectation is taken with respect to the conditional distributions in (4.5). By maximizing $\mathrm{EHI}_n\left(\boldsymbol{x};\mathcal{P},\boldsymbol{t}\right)$ over $\boldsymbol{x}\in\mathcal{X}$, we find an input point for the new function evaluation. Given an approximation of the Pareto set $\mathcal{P}$ by a set of nondominated points, EHI can be calculated empirically with a closed-form expression Hupkens et al. (2015). We denote the nondominated points by $\boldsymbol{y}^j=(y_1^j,y_2^j)=(y_1(\boldsymbol{x}_{i_j}),y_2(\boldsymbol{x}_{i_j}))$ for $j=1,\ldots,p$ with $\{i_1,\ldots,i_p\}\subset\{1,\ldots,n\}$. Without loss of generality, we assume that those nondominated points are ordered based on the values of the first objective, i.e., $y_1(\boldsymbol{x}_{i_1})<y_1(\boldsymbol{x}_{i_2})<\ldots<y_1(\boldsymbol{x}_{i_p})$. Then the approximation of the Pareto set $\mathcal{P}$ is denoted by

$$\mathcal{P}^{(n)}=\left\{\boldsymbol{y}^0,\boldsymbol{y}^1,\ldots,\boldsymbol{y}^p,\boldsymbol{y}^{p+1}\right\}, \tag{4.7}$$

where $\boldsymbol{y}^0=(-\infty,t_2)$ and $\boldsymbol{y}^{p+1}=(t_1,-\infty)$ with t_1 and t_2 being the two coordinates of the reference point $\boldsymbol{t}=(t_1,t_2)$. Following Emmerich et al. (2011), EHI in (4.6) can be calculated empirically based on the conditional distributions in (4.5) as

$$\begin{aligned}\mathrm{EHI}_n\left(\boldsymbol{x};\mathcal{P}^{(n)},\boldsymbol{t}\right)&=\mathrm{E}\left\{\mathrm{HVI}\left((Y_1(\boldsymbol{x}),Y_2(\boldsymbol{x})),\mathcal{P}^{(n)},\boldsymbol{t}\right)\right\}\\&=\sum_{j=1}^{p+1}\left(y_1^{j-1}-y_1^j\right)\cdot\Phi\left(\frac{y_1^j-\hat{y}_1(\boldsymbol{x})}{\hat{s}_1(\boldsymbol{x})}\right)\cdot\Psi\left(y_2^j,y_2^j,\hat{y}_2(\boldsymbol{x}),\hat{s}_2(\boldsymbol{x})\right)\\&+\sum_{j=1}^{p+1}\left(\Psi\left(y_1^{j-1},y_1^{j-1},\hat{y}_1(\boldsymbol{x}),\hat{s}_1(\boldsymbol{x})\right)-\Psi\left(y_1^{j-1},y_1^j,\hat{y}_1(\boldsymbol{x}),\hat{s}_1(\boldsymbol{x})\right)\right)\\&\cdot\Psi\left(y_2^j,y_2^j,\hat{y}_2(\boldsymbol{x}),\hat{s}_2(\boldsymbol{x})\right),\end{aligned} \tag{4.8}$$

where $\Psi\left(a,b,\mu,\sigma\right)=\sigma\phi\left(\frac{b-\mu}{\sigma}\right)+\left(a-\mu\right)\Phi\left(\frac{b-\mu}{\sigma}\right)$.

4.3 Proposed Method

We propose a bi-objective Bayesian optimization algorithm using transformed additive Gaussian processes (TAG) (Lin and Joseph, 2020) as the surrogate model. TAG approximates each of the two objective functions by additive functions, each additive term associated with one-dimensional decision variable. Under this model assumption, maximizing the acquisition function EHI given by (4.6) can be simplified to maximize the EHI's for each decision variable. In this section, we introduce using TAG as a surrogate model for bi-objective problems, develop the EHI functions under the TAG model assumption, and summarize our algorithm implementation.

4.3.1 Bi-objective Surrogate with Transformed Additive Gaussian Processes

Transformed additive Gaussian processes aim to simplify a complex response to additive functions. Consider the bi-objective problem in (4.1). As noted in Section 4.2.2, we assume that $y_1(\boldsymbol{x})$ and $y_2(\boldsymbol{x})$ are the realizations of two independent transformed additive Gaussian processes $Y_1(\boldsymbol{x})$ and $Y_2(\boldsymbol{x})$ with $\boldsymbol{x} \in \mathcal{X} \subset R^d$. Equivalently, following Lin and Joseph (2020), we have

$$g_{\lambda_l}(Y_l(\boldsymbol{x})) = \mu_l + \sum_{k=1}^{d} Z_{lk}(x_k) + \epsilon_l, \quad \text{for} \quad l = 1, 2. \tag{4.9}$$

We explain the notation in turn as follows. The function $g_{\lambda_l}(\cdot)$ is the Box-Cox transformation (Box and Cox, 1964),

$$g_{\lambda_l}(y) = \begin{cases} \frac{y^{\lambda_l}-1}{\lambda_l} & \lambda_l \neq 0 \\ \log y & \lambda_l = 0, \end{cases} , \quad \text{for} \quad l = 1, 2, \tag{4.10}$$

where λ_l is the parameter of the transformation. The deterministic means of TAG are denoted by μ_l's. For $k = 1, \ldots, d$ and $l = 1, 2$, $Z_{lk}(x_k)$'s are mutually independent mean-zero Gaussian processes with variances τ_{lk}^2's, and correlation functions $R_{lk}(\cdot)$'s. The additive noise $\epsilon_l \sim N(0, \sigma_l^2)$ is independent of $Z_{lk}(x_k)$'s.

Under this model assumption, the correlation of $g_{\lambda_l}(Y_l(\boldsymbol{x}))$ and $g_{\lambda_l}(Y_l(\boldsymbol{x}'))$ for $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ is

$$R_l(\boldsymbol{x} - \boldsymbol{x}') = \sum_{k=1}^{d} \omega_{lk} R_{lk}\left(x_k - x_k'\right), \tag{4.11}$$

where $\tau_l^2 = \sum_{k=1}^d \tau_{lk}^2$ and $\omega_{lk} = \tau_{lk}^2/\tau_l^2$. An example of the correlation function is given by (3.3), which is associated with some unknown correlation parameters. In Lin and Joseph (2020), empirical Bayes estimation and non-linear optimization toolboxes are used to estimate those unknown parameters, including τ_l, ω_{lk}, λ_l and correlation parameters, which are referred to as hyperparameters of TAG.

Assume that we collected outputs $\boldsymbol{y}_{n,1} = (y_1(\boldsymbol{x}_1)\ldots,y_1(\boldsymbol{x}_n))^\top$ and $\boldsymbol{y}_{n,2} = (y_2(\boldsymbol{x}_1)\ldots,y_2(\boldsymbol{x}_n))^\top$. For $l = 1,2$ and $k = 1,\ldots,d$, we have

$$\begin{pmatrix} g_{\lambda_l}(\boldsymbol{y}_{n,l}) - \mu_l\mathbf{1} \\ Z_{lk}(x_k) \end{pmatrix} \sim N_{n+1}\left[\begin{pmatrix} \mathbf{0}_n \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_l^2\boldsymbol{R}_l + \sigma_l^2\boldsymbol{I} & \tau_{lk}^2\boldsymbol{r}_{lk}(x_k) \\ \tau_{lk}^2\boldsymbol{r}_{lk}(x_k)^\top & \tau_{lk}^2, \end{pmatrix}\right], \tag{4.12}$$

where $\mathbf{0}_n$ is the $n\times 1$ vector of zeros, $\boldsymbol{I}$ is the $n\times n$ identify matrix, $\boldsymbol{r}_{lk}(x_k) = (R_{lk}(x_k - x_{1k}),\ldots,R_{lk}(x_k - x_{nk}))^\top$ and $\boldsymbol{R}_l$ is the $n\times n$ correlation matrix with the (i,j)-th element $R_l(\boldsymbol{x}_i - \boldsymbol{x}_j)$ given by (4.11). Using the conditional distribution for the normal distribution (e.g., (Eaton, 2007)), we can derive the conditional distribution of $Z_{lk}(x_k)$ given $\boldsymbol{y}_{n,l}$ as

$$Z_{lk}(x_k)|\boldsymbol{y}_{n,l} \sim N\left(\hat{z}_{lk}(x_k), s_{lk}^2(x_k)\right) \quad \text{for} \quad l = 1,2 \tag{4.13}$$

with

$$\hat{z}_{lk}(x_k) = \omega_{lk}\boldsymbol{r}_{lk}(x_{lk})^\top\left(\boldsymbol{R}_l + \delta_l\boldsymbol{I}\right)^{-1}\left(g_{\lambda_l}(\boldsymbol{y}_{n,l}) - \mu_l\right)$$

and

$$s_{lk}^2(x_k) = \tau_l^2\omega_{lk}\left(1 - \boldsymbol{r}_{lk}(x_{lk})^\top\left(\boldsymbol{R}_l + \delta_l\boldsymbol{I}\right)^{-1}\omega_{lk}\boldsymbol{r}_{lk}(x_{lk})\right),$$

where $\delta_l = \sigma_l^2/\tau_l^2$ and μ_l can be replaced by $\hat{\mu}_l = \mathbf{1}^\top\left(\boldsymbol{R}_l + \delta_l\boldsymbol{I}\right)^{-1}g_{\lambda_l}(\boldsymbol{y}_{n,l})/\mathbf{1}^\top\left(\boldsymbol{R}_l + \delta_l\boldsymbol{I}\right)^{-1}\mathbf{1}$ as a plug-in estimator. The conditional distribution of $Z_{lk}(x_k)$ will then be used to develop EHIs for each dimension of the decision variable.

4.3.2 Expected Hypervolume Improvement with Transformed Additive Gaussian Processes

Based on the conditional distributions in (4.13), we simplify the EHI in (4.6) to the expected hypervolume improvements for the bi-objective problems for each dimension of the decision variable.

Provided by the approximation of TAG, we have

$$g_{\lambda_l}\left(y_l(\boldsymbol{x})\right) \approx \mu_l + \sum_{k=1}^{d} z_{lk}\left(x_k\right) \quad \text{for} \quad l = 1, 2, \tag{4.14}$$

where $z_{lk}(x_k)$ is a realization of $Z_{lk}(x_k)$. Assume that the decision space $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$. Then the original bi-objective optimization problem in (4.1) is simplified into d one-dimensional bi-objective problems

$$\min_{x_k \in \mathcal{X}_k} \quad \boldsymbol{z}_k(x_k) = \left[z_{1,k}\left(x_k\right), z_{2,k}\left(x_k\right)\right] \quad \text{for} \quad k = 1, \ldots, d \tag{4.15}$$

Let $\mathcal{P}_k$ be the Pareto set of (4.15), i.e.,

$$\mathcal{P}_k = \left\{\boldsymbol{z}_k(x_k) \in \mathcal{R}^2 \mid \nexists\, \boldsymbol{z}_k(x_k') \in \mathcal{R}^2 : \boldsymbol{z}_k(x_k') \prec \boldsymbol{z}_k(x_k)\right\} \tag{4.16}$$

for $k = 1, \ldots, d$. Under TAG, we can define the EHI for the problem in (4.15) as

$$\mathrm{EHI}_{n,k}\left(x_k; \mathcal{P}_k, \boldsymbol{t}_k\right) = \mathrm{E}\left\{\mathrm{HVI}\left((Z_{1,k}(x_k), Z_{2,k}(x_k)), \mathcal{P}_k, \boldsymbol{t}\right)\right\}, \tag{4.17}$$

where the expectation is taken with respect to the conditional distribution of $Z_{lk}(x_k)$ in (4.13). As a result, instead of maximizing $\mathrm{EHI}_n\left(\boldsymbol{x}; \mathcal{P}, \boldsymbol{t}\right)$ over $\boldsymbol{x} \in \mathcal{X}$, a new input point is given by maximizing $\mathrm{EHI}_{n,k}\left(x_k; \mathcal{P}_k, \boldsymbol{t}\right)$ with one dimensional decision variable $x_k \in \mathcal{X}_k$ for $k = 1, \ldots, d$. To be specific, the new design point is generated by selecting $\boldsymbol{x}_{new} = (x_1^*, \ldots, x_d^*)$ with

$$x_k^* \in \max_{x \in \mathcal{X}_k} \mathrm{EHI}_{n,k}\left(x_k; \mathcal{P}_k, \boldsymbol{t}\right) \quad \text{for} \quad k = 1, \ldots, d.$$

Given an approximation of the Pareto set $\mathcal{P}_k$, EHI can attain a close form expression as in (4.8) using the conditional distribution in (4.13). In the next subsection, we describe our detailed implementation of the proposed algorithm.

4.3.3 The Proposed Algorithm

We detail our implementation of the proposed bi-objective Bayesian optimization algorithm in this subsection. The surrogate model is a TAG model in (4.9) implemented by the R package `TAG` Lin and Joseph (2021). The calculation of EHI depends on the approximation of the Pareto set in (4.7). We use the R package `ecr` Bossek (2017) to approximate Pareto set. Given responses $\boldsymbol{y}_{m,1}$ and $\boldsymbol{y}_{m,2}$ of size m, the approximations of the Pareto sets of $\mathcal{P}$ in (4.2) and $\mathcal{P}_k$ in (4.16) are denoted by $\mathcal{P}^{(m)}$ and $\mathcal{P}_k^{(m)}$, respectively. We summarize our implementation in Algorithm 5. We first generate an initial dataset of size n (line 1 in Algorithm 5), and then add one more data point at each step until we exhaust the budget of function evaluation N (line 2-10 in Algorithm 5). In each step, we fit a TAG model for each objective function. For simplicity, the maximization of $\mathrm{EHI}_{n,k}(x_k; \mathcal{P}_k, \boldsymbol{t})$ is done by evaluating q random candidate points in $\mathcal{X}_k$ (line 4-9 in Algorithm 5). In our implementation, we set $q = 10$ and generate a $q \times d$ random Latin hypercube design using the R package `lhs` Carnell (2022) with k-th column being the candidate points in $\mathcal{X}_k$ (line 4 in Algorithm 5). At each step, we compute the hypervolume indicator in (4.3) based on the approximated Pareto set $\mathcal{P}^{(m)}$ (line 10 in Algorithm 5).

Algorithm 5 Bi-objective Bayesian Optimization with TAG

```
1:  Given an initial dataset with inputs $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ and the outputs $\boldsymbol{y}_{n,1}$ and $\boldsymbol{y}_{n,2}$, compute an initial
    Pareto set approximation $\mathcal{P}^{(n)}$ and compute $\mathrm{H}(\mathcal{P}^{(n)}, \boldsymbol{t})$
2:  for $m = n, n+1, \dots, N$ do
3:      Fit the TAG models based on $\boldsymbol{y}_{m,1}$ and $\boldsymbol{y}_{m,2}$ and $\boldsymbol{x}$, find $\mathcal{P}_k^{(m)}$ for $k = 1, \dots, d$.
4:      Generate a $q \times d$ matrix $D$ as a random Latin Hypercube design with $ij$-th entries $D_{ij}$.
5:      for $k = 1, \dots, d$ do
6:          Evaluate $\mathrm{EHI}_{n,k}\left(D_{ik}; \mathcal{P}_k^{(m)}, \boldsymbol{t}\right)$
7:          Select point $D_k^*$ that maximizes $\mathrm{EHI}_{n,k}\left(D_{ik}; \mathcal{P}_k^{(m)}, \boldsymbol{t}\right)$ over $i = 1, \dots, q$
8:      end for
9:      Set new input point $\boldsymbol{x}^{(m+1)} = (D_1^*, \dots, D_d^*)$ and obtain $\boldsymbol{y}(\boldsymbol{x}^{(m+1)})$
10:     By including the new response, update $\boldsymbol{y}_{m,1} \to \boldsymbol{y}_{m+1,1}$, $\boldsymbol{y}_{m,2} \to \boldsymbol{y}_{m+1,2}$, and $\mathcal{P}^{(m)} \to \mathcal{P}^{(m+1)}$.
        Compute $\mathrm{H}(\mathcal{P}^{(m+1)}, \boldsymbol{t})$
11: end for
    Return $\mathrm{H}(\mathcal{P}^{(m)}, \boldsymbol{t})$ for $m = n, n+1, \dots, N$.
```

It is worth noting that the hyperparameters (such as λ_l's and τ_l's etc) in TAG are not updated in every step in our implementation. To save computation time and resources, we will refit those hyperparameters in every ten steps.

4.4 Numerical Results

In this section, we compare the proposed algorithm as described in Section 4.3.3 with the bi-objective Bayesian optimization procedure implemented with the function `GParetoptim` in the R package `GPareto` Binois and Picheny (2019). This algorithm uses independent Gaussian processes as a surrogate model for the two objectives, and uses EHI in (4.8) as the acquisition function. To distinguish between the two approaches, our method is denoted by "Transformed Additive GP", and the method in `GPareto` is referred to as "Classical GP".

First, we generate initial data using the R package `lhs` Carnell (2022). Next, we fit an initial surrogate model to the data. For our method, the surrogate model is a TAG model as in (4.9). The classical GP approach will use the Gaussian processes as the surrogate models using the R package Roustant et al. (2012) `DiceKriging`. The reference point $\boldsymbol{t}$ is chosen as a point that bounds the outcome space of the bi-objective problem when $0 \leq x_k \leq 1$ for $k = 1, \dots, d$. We sequentially add new points one by one using EHI for 20 steps. In each step, we compute $\mathrm{H}(\mathcal{P}, \boldsymbol{t})$ in (4.3) based on approximations of $\mathcal{P}$ following Fonseca et al. (2006).

We replicate the whole procedure 30 times (with 30 different random initial datasets) for both methods. For the returned hypervolume values of each method in each step, we compute their average and use 10^{th} and 90^{th} quantiles to construct a confidence band. For our method, in the R package `TAG`, the default range of λ_l in (4.10) is -2 to 2, we use this default range for the examples unless otherwise noted.

Example 1

We construct bi-objective functions as follows:

$$y_l(\boldsymbol{x}) = \exp\left(\Sigma_{k=1}^{d} \left(x_k - a_l\right)^2 + c_l x_1 x_2\right) \quad \text{for} \quad l = 1, 2, \tag{4.18}$$

which can be transformed to additive functions using the Box-Cox transformation (4.10) for $\lambda_l = 0$ and $c_l = 0$.

We first study the impact of the dimension of the decision space on the proposed method as compared to the classical GP approach. We consider the performance of the proposed algorithm under different sizes of d. We specify three cases with parameters of (4.18) in Table 4.1, where the dimension of input spaces changes from 4 to 8. To ensure there are a sufficient number of initial points for $d = 8$, we use 25 initial data points in this study. The results are depicted in Figure 4.1. As the dimension of the decision space increases, the advantage of the proposed method is greater.

Table 4.1: Table of parameter settings of (4.18) for the three cases in Figure 4.1.

Cases	(1)			(2)			(3)		
l	d	a_l	c_l	d	a_l	c_l	d	a_l	c_l
1	4	0.3	0	6	0.3	0	8	0.3	0
2	4	0.4	0	6	0.4	0	8	0.4	0

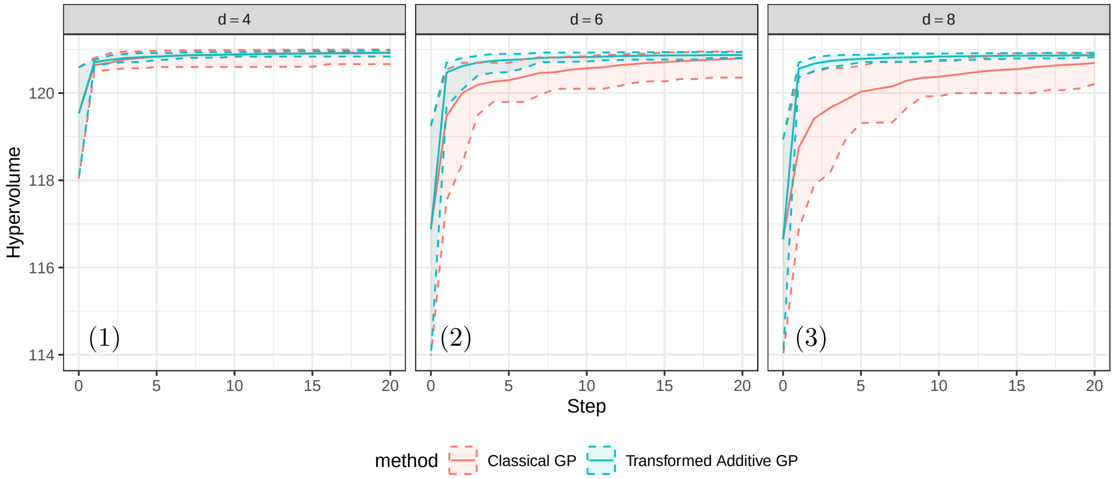

Figure 4.1: The mean, 10-th and 90-th quantiles of approximated hypervolume over 30 replications and 20 steps for optimizing the objective functions in (4.18) with the parameter settings in Table 4.1.

Next, we study the performance of the proposed method when the objective functions have an interaction term and are not strictly transformed to additive functions with one-dimensional input. We consider the performance of the proposed algorithm under different values of c_l. We specify three cases with parameters of (4.18) in Table 4.2, where the value of c_2 changes from 0 to 0.5. For this example, 15 initial data points are used. The results are depicted in Figure 4.2. The results show that when the interaction term is negligible as $c_2 = 0$ or 0.01, our method is more robust than Classical GP in early stages. However, when the interaction term increases to $c_2 = 0.5$, the performances of the two approaches are comparable.

Table 4.2: Table of parameter settings of (4.18) for the three cases in Figure 4.2

Cases	(1)			(2)			(3)		
l	d	a_l	c_1	d	a_l	c_1	d	a_l	c_1
1	4	0.3	0	4	0.3	0	4	0.3	0
2	4	0.4	0	4	0.4	0.01	4	0.4	0.5

.

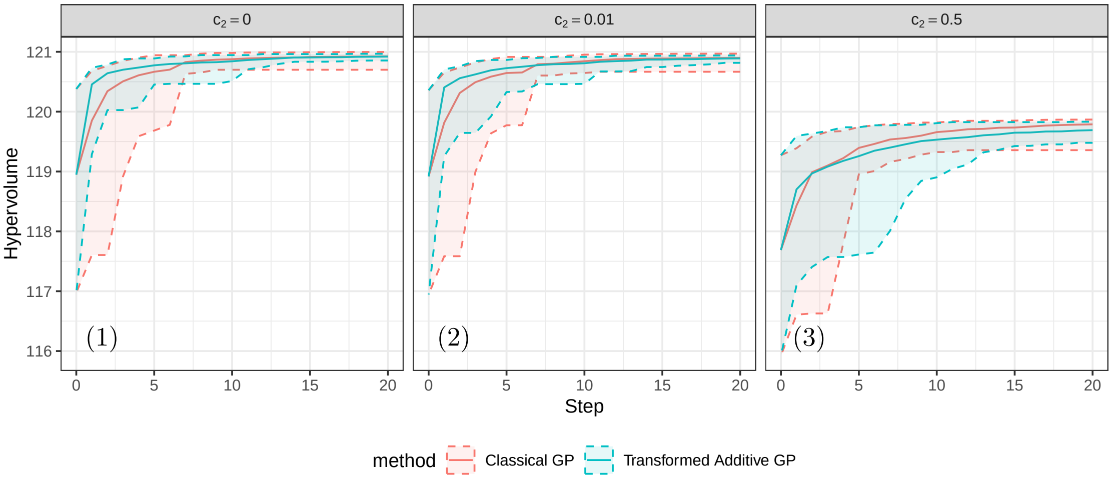

Figure 4.2: The mean, 10-th and 90-th quantiles of approximated hypervolume over 30 replications and 20 steps for optimizing the objective functions in (4.18) with the parameter settings in Table 4.2.

Example 2

We use the example FES1 from the literature of bi-objective optimization Fieldsend et al. (2003) to demonstrate the performance of the proposed method. The original problem of FES1 is given by in (4.19). For a comparison purpose, we construct a modified version of FES1 in (4.20) by taking the exponential values of the original objectives. FES1 is an additive function, whereas FES1-modified can be transformed into an additive function. For FES1, we restrict λ_l for $l = 1, 2$ in TAG to be between 0.5 and 1.5 since the existing procedure does not estimate it properly. For this example, 20 initial data points are used. The results are depicted in Figure 4.3. Our method is comparable to the classical GP method when the function is originally additive without transformation. However, when the additive objective functions are not originally additive (for FES1-modified), our method outperforms the classical GP method in the early stages.

$$\text{FES1} = \begin{cases} y_1(\boldsymbol{x}) = \sum_{i=1}^{4} |x_i - \frac{\exp\{\left(\frac{i}{4}\right)^2\}}{3}|^{0.5} \\ y_2(\boldsymbol{x}) = \sum_{i=1}^{4} \left(x_i - 0.5\cos\left(\frac{10\pi i}{4}\right) - 0.5\right)^2 \end{cases} \tag{4.19}$$

$$\text{FES1-modified} = \begin{cases} y_1(\boldsymbol{x}) = \exp\{\sum_{i=1}^{4} |x_i - \frac{\exp\{\left(\frac{i}{4}\right)^2\}}{3}|^{0.5}\} \\ y_2(\boldsymbol{x}) = \exp\{\sum_{i=1}^{4} \left(x_i - 0.5\cos\left(\frac{10\pi i}{4}\right) - 0.5\right)^2\} \end{cases} \tag{4.20}$$

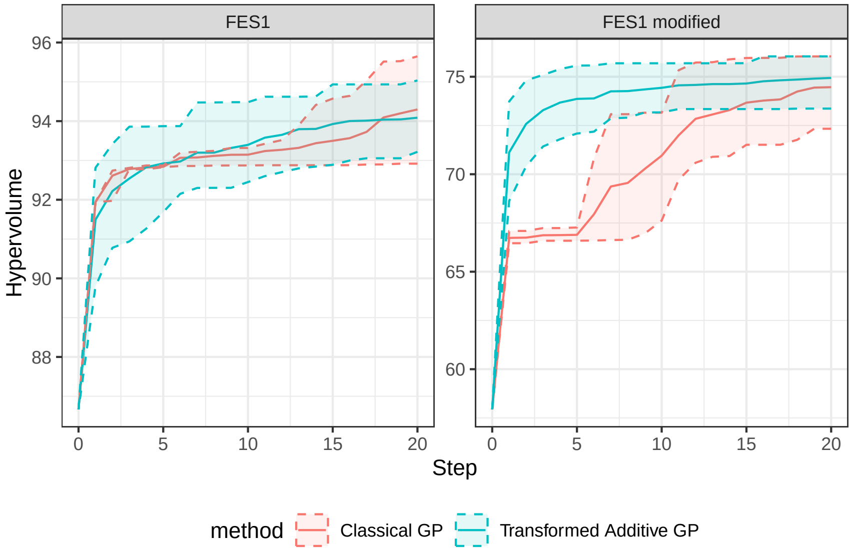

Figure 4.3: The mean, 10-th and 90-th quantiles of approximated hypervolume over 30 replications and 20 steps for the two problems in (4.19) and (4.20).

4.5 Conclusion

We have proposed a method for bi-objective Bayesian optimization using TAG as the surrogate model to simplify the optimization of EHI by decomposing the decision space into one-dimensional additive subspaces. Through a numerical study, we have demonstrated the advantage of the proposed method over classical bi-objective Bayesian optimization. Future directions of this work include improving the parameter tuning in the TAG framework, extending the method to the multi-objective case, and integrating other existing approaches into the proposed algorithm to optimize EHI.

Acknowledgements

This work was supported by Clemson University's Virtual Prototyping of Autonomy Enabled Ground Systems (VIPR-GS), under Cooperative Agreement W56HZV-21-2-0001 with the US Army DEVCOM Ground Vehicle Systems Center (GVSC).

Disclaimers

All opinions, conclusions and findings wherein are those of the authors and may not be those of the affiliated institutions DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited. OPSEC # 8425.

Appendices

Appendix A Proof of Results in Chapter 2

A.1 Proof of Lemma 1

Note that the posterior density function of σ^2 given S^2 satisfies

$$p(\sigma^2|S^2) \propto p\left((m-1)S^2|\sigma^2\right)\pi(\sigma^2),$$

where $\pi(\sigma^2)$ is the PDF of $IG(a,b)$, and $p\left((m-1)S^2|\sigma^2\right)$ is the PDF of $\sigma^2\chi^2_{m-1}$. Thus, we have

$$\begin{aligned} p(\sigma^2|S^2) &\propto \frac{1}{2^{\frac{m-1}{2}}\Gamma(\frac{m-1}{2})}\left(\frac{(m-1)}{\sigma^2}S^2\right)^{\frac{m-1}{2}-1}\exp\left(-\frac{m-1}{\sigma^2}\frac{S^2}{2}\right)\frac{m-1}{\sigma^2}\frac{b^a}{\sigma^{2(a+1)}\Gamma(a)}\exp\left(-\frac{b}{\sigma^2}\right) \\ &\propto \left(\sigma^2\right)^{-(a+\frac{m-1}{2})-1}\exp\left(-\frac{1}{\sigma^2}\left[\frac{(m-1)S^2}{2}+b\right]\right), \end{aligned}$$

which implies that the posterior distribution of σ^2 given S^2 is an inverse gamma distribution with parameters $a+\frac{m-1}{2}$ and $b+\frac{S^2(m-1)}{2}$.

A.2 Proof of Lemma 2

The marginal density function of S^2 based on the Bayesian model in Lemma 1 is given by

$$p(S^2) = \int p(S^2|\sigma^2)\pi(\sigma^2)d\sigma^2.$$

Following the proof of Lemma 1, we have that

$$p(S^2|\sigma^2)\pi(\sigma^2) = \frac{b^a\left(S^2\right)^{\frac{m-1}{2}-1}(m-1)^{\frac{m-1}{2}}}{2^{\frac{m-1}{2}}\Gamma(\frac{m-1}{2})\Gamma(a)}\cdot\frac{\Gamma(a+\frac{m+1}{2})}{\left(b+\frac{m-1}{2}S^2\right)^{\frac{m+1}{2}+a}}\cdot g(\sigma^2,S^2),$$

where

$$g(\sigma^2,S^2) = \frac{\left(b+\frac{m-1}{2}S^2\right)^{\frac{m+1}{2}+a}}{\Gamma\left(\frac{m+1}{2}+a\right)}(\sigma^2)^{-(a+\frac{m-1}{2})-1}\exp\left(-\frac{1}{\sigma^2}\left[b+\frac{(m-1)S^2}{2}\right]\right)$$

Noting that $g(\sigma^2, S^2)$ is exactly the PDF of an inverse gamma distribution for σ^2, we have

$$\begin{aligned}
p(S^2) &= \frac{b^a \left(S^2\right)^{\frac{m-1}{2}-1} (m-1)^{\frac{m-1}{2}}}{2^{\frac{m-1}{2}} \Gamma(\frac{m-1}{2}) \Gamma(a)} \cdot \frac{\Gamma(a + \frac{m+1}{2})}{\left(b + \frac{m-1}{2} S^2\right)^{\frac{m+1}{2}+a}} \cdot \int g(\sigma^2, S^2) d\sigma^2 \\
&= \frac{b^a \left(S^2\right)^{\frac{m-1}{2}-1} (m-1)^{\frac{m-1}{2}}}{2^{\frac{m-1}{2}} \Gamma(\frac{m-1}{2}) \Gamma(a)} \cdot \frac{\Gamma(a + \frac{m+1}{2})}{\left(b + \frac{m-1}{2} S^2\right)^{\frac{m+1}{2}+a}} \\
&= \frac{1}{B(\frac{m-1}{2}, a)} \cdot \frac{\left(\frac{m-1}{2b} S^2\right)^{\frac{m-1}{2}-1} \frac{m-1}{2b}}{\left(\frac{m-1}{2b} S^2 + 1\right)^{\frac{m-1}{2}+a}},
\end{aligned}$$

which is proportional to the PDF of the beta prime distribution (Johnson et al., 1995). Therefore, we have

$$\frac{m-1}{2b} S^2 \sim \text{BetaPrime}\left(\frac{m-1}{2}, a\right).$$

A.3 Proof of Proposition 1

Letting $B(\alpha, \beta)$ denote the beta function, the truncated expectation of the beta prime distribution is given by

$$\mathbb{E}[Z|Z \leq c] = \frac{\int_{-\infty}^{c} z f_{\alpha,\beta}(z) dz}{F_{\alpha,\beta}(c) - F_{\alpha,\beta}(-\infty)},$$

where $f_{\alpha,\beta}(z) = \frac{z^{\alpha}(1+z)^{-\alpha-\beta}}{B(\alpha,\beta)}$ and $F_{\alpha,\beta}$ are the PDF and CDF of a beta prime distribution with parameters α and β, respectively. Then, we have

$$\begin{aligned}
\mathbb{E}[Z|Z \leq c] &= \frac{1}{F_{\alpha,\beta}(c)} \int_{-\infty}^{c} \frac{z \left(z^{\alpha} (1+z)^{-\alpha-\beta}\right)}{B(\alpha, \beta)} dz \\
&= \frac{1}{F_{\alpha,\beta}(c)} \int_{-\infty}^{c} \frac{z^{\alpha+1} (1+z)^{-(\alpha+1)-(\beta-1)} \Gamma((\alpha+1) + (\beta-1))}{\Gamma(\alpha)\Gamma(\beta)} dz \\
&= \frac{\alpha+1}{\beta} \frac{1}{F_{\alpha,\beta}(c)} \int_{-\infty}^{c} \frac{z^{\alpha+1} (1+z)^{-(\alpha+1)-(\beta-1)} \Gamma((\alpha+1) + (\beta-1))}{\Gamma(\alpha+1)\Gamma(\beta-1)} dz \\
&= \frac{\alpha+1}{\beta} \cdot \frac{F_{\alpha+1,\beta-1}(c)}{F_{\alpha,\beta}(c)}.
\end{aligned}$$

A.4 Proof of Proposition 2

Denoting $C_x^{(t)} = \min_{x' \neq x} \left\{ \frac{b_{x'}^{(t)}}{a_{x'}^{(t)} - 1} \right\}$, (2.10) becomes:

$$\min_{x' \in \mathcal{X}} \left\{ \frac{b_{x'}^{(t+1)}}{a_{x'}^{(t+1)} - 1} \right\} = \begin{cases} \frac{b_x^{(t+1)}}{a_x^{(t+1)} - 1} & \text{if} \quad \frac{b_x^{(t+1)}}{a_x^{(t+1)} - 1} \leq C_x^{(t)} \\ C_x^{(t)} & \text{o.w.} \end{cases} . \tag{21}$$

Then, we have

$$\begin{aligned} \mathbb{E}\left[\min_{x' \in X} \frac{b_{x'}^{(t+1)}}{a_{x'}^{(t+1)} - 1}\right] &= \mathbb{E}\left[\frac{b_x^{(t+1)}}{a_x^{(t+1)} - 1} \middle| \frac{b_x^{(t+1)}}{a_x^{(t+1)} - 1} \leq C_x^{(t)} \right] \mathbb{P}\left(\frac{b_x^{(t+1)}}{a_x^{(t+1)} - 1} \leq C_x^{(t)} \right) \\ &\quad + C_x^{(t)} \mathbb{P}\left(\frac{b_x^{(t+1)}}{a_x^{(t+1)} - 1} > C_x^{(t)} \right) \end{aligned} \tag{22}$$

From (2.6), we can see that

$$\frac{b_x^{(t+1)}}{a_x^{(t+1)} - 1} = \frac{b_x^{(t)}}{a_x^{(t)} + \frac{m-1}{2} - 1} \left(1 + \frac{(m-1) S_x^{2,(t+1)}}{2 b_x^{(t)}} \right) \leq C_x^{(t)} \tag{23}$$

is equivalent to

$$\frac{m-1}{2 b_x^{(t)}} S_x^{2,(t+1)} \leq \tilde{C}_x^{(t)},$$

where $\tilde{C}_x^{(t)} = \frac{\left(a_x^{(t)} + \frac{m-1}{2} - 1\right) C_x^{(t)}}{b_x^{(t)}} - 1$. Consequently, we have

$$\mathbb{P}\left(\frac{b_x^{(t+1)}}{a_x^{(t+1)} - 1} \leq C_x^{(t)} \right) = \mathbb{P}\left(\frac{m-1}{2 b_x^{(t)}} S_x^{2,(t+1)} \leq \tilde{C}_x^{(t)} \right) = F_{\frac{m-1}{2}, a_x^{(t)}} \left(\tilde{C}_x^{(t)} \right), \tag{24}$$

$$\begin{aligned} \mathbb{E}\left[\frac{b_x^{(t+1)}}{a_x^{(t+1)} - 1} \middle| \frac{b_x^{(t+1)}}{a_x^{(t+1)} - 1} \leq C_x^{(t)} \right] &= \mathbb{E}\left[\frac{b_x^{(t+1)}}{a_x^{(t+1)} - 1} \middle| \frac{m-1}{2 b_x^{(t)}} S_x^{2,(t+1)} \leq \tilde{C}_x^{(t)} \right] \\ &= \frac{b_x^{(t)}}{a_x^{(t)} + \frac{m-1}{2} - 1} \mathbb{E}\left[\frac{m-1}{2 b_x^{(t)}} S_x^{2,(t+1)} + 1 \middle| \frac{m-1}{2 b_x^{(t)}} S_x^{2,(t+1)} \leq \tilde{C}_x^{(t)} \right] \end{aligned} \tag{25}$$

$$= \frac{b_x^{(t)}}{a_x^{(t)} + \frac{m-1}{2} - 1} \left[\frac{m+1}{2 a_x^{(t)}} \cdot \frac{F_{\frac{m+1}{2}, a_x^{(t)} - 1}(\tilde{C}_x^{(t)})}{F_{\frac{m-1}{2}, a_x^{(t)}}(\tilde{C}_x^{(t)})} + 1 \right], \tag{26}$$

where (24) holds since $\frac{m-1}{2b_x^{(t)}} S_x^{2,(t+1)} \sim \text{BetaPrime}\left(\frac{m-1}{2}, a_x^{(t)}\right)$ by Lemma 2, (25) holds due to (23), and (26) holds by Proposition 1. Finally, applying (24)-(26) to (2.9) and (22), we have

$$
\begin{aligned}
\mathbb{E}\left[\min_{x'\in X} \frac{b_{x'}^{(t+1)}}{a_{x'}^{(t+1)}-1}\right] &= \frac{b_x^{(t)}}{a_x^{(t)}+\frac{m-1}{2}-1}\left[\frac{m+1}{2a_x^{(t)}}\cdot\frac{F_{\frac{m+1}{2},a_x^{(t)}-1}(\tilde{C}_x^{(t)})}{F_{\frac{m-1}{2},a_x^{(t)}}(\tilde{C}_x^{(t)})}+1\right] F_{\frac{m-1}{2},a_x^{(t)}}(\tilde{C}_x^{(t)}) \\
&\quad + C_x^{(t)}\left[1-F_{\frac{m-1}{2},a_x^{(t)}}(\tilde{C}_x^{(t)})\right], \\
\text{KG}^{(t+1)}(x) &= \min_{x'\in\mathcal{X}}\left\{\frac{b_{x'}^{(t)}}{a_{x'}^{(t)}-1}\right\} - C_x^{(t)}\left[1-F_{\frac{m-1}{2},a_x^{(t)}}(\tilde{C}_x^{(t)})\right] \\
&\quad -\frac{b_x^{(t)}}{a_x^{(t)}+\frac{m-1}{2}-1}\left[\frac{m+1}{2a_x^{(t)}}\cdot\frac{F_{\frac{m+1}{2},a_x^{(t)}-1}(\tilde{C}_x^{(t)})}{F_{\frac{m-1}{2},a_x^{(t)}}(\tilde{C}_x^{(t)})}+1\right] F_{\frac{m-1}{2},a_x^{(t)}}(\tilde{C}_x^{(t)}).
\end{aligned}
$$

A.5 Proof of Proposition 3

First, we have that

$$
\begin{aligned}
\text{EI}^{(t+1)}(x) &= \mathbb{E}\left[\max\left\{\min_{x\in\mathcal{X}}\left(\frac{b_x^{(t)}}{a_x^{(t)}-1}\right)-S_x^2, 0\right\}\right] \\
&= \frac{2b_x^{(t)}}{m-1}\mathbb{E}\left[\max\left\{\frac{m-1}{2b_x^{(t)}}\min_{x\in\mathcal{X}}\left(\frac{b_x^{(t)}}{a_x^{(t)}-1}\right)-\frac{m-1}{2b_x^{(t)}}S_x^2, 0\right\}\right],
\end{aligned}
$$

thus the expectation can be taken with respect to the BetaPrime distribution of $\frac{m-1}{2b_x^{(t)}}S_x^2$.

Suppose that Z follows the Beta Prime distribution with parameters α and β, Lemma 2 gives

$$
\begin{aligned}
\mathbb{E}\left[\max\left(u-Z,0\right)\right] &= \mathbb{P}\left(Z\le u\right)\mathbb{E}\left[u-Z|Z\le u\right] \\
&= uF_{\alpha,\beta}(u) - F_{\alpha,\beta}(u)\mathbb{E}\left[Z|Z\le u\right] \\
&= uF_{\alpha,\beta}(u) - \frac{\alpha+1}{\beta}F_{\alpha+1,\beta-1}(u),
\end{aligned}
$$

which directly leads to the result in Proposition 3 by letting $u = \frac{m-1}{2b_x^{(t)}}\min_{x\in\mathcal{X}}\left\{\frac{b_x^{(t)}}{a_x^{(t)}-1}\right\}$, $\alpha = (m-1)/2$ and $\beta = a_x^{(t)}$.

A.6 Examples of Resulting Designs

Figure 4 visualizes the resulting designs of the proposed KG-IG method based on simulator I in Section 2.6. We show the designs for $K = 6, 12$ and 30; for example, the left figure ($K = 6$)

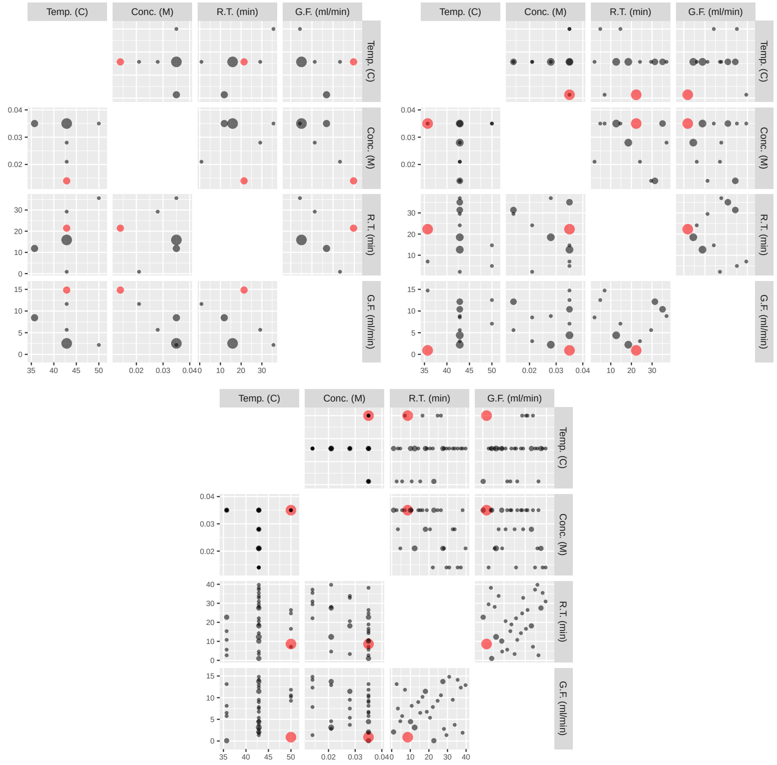

Figure 4: The resulting designs of KG-IG based on simulator I in Section 2.6 for $K = 6$ (left), $K = 12$ (middle), and $K = 30$ (right) over 60 experiments. The red dot represents the true optimal. The size of the dots is relative to the number of replications on each design point.

contains exactly six dots that correspond to the six experimental settings.

Appendix B Proofs of Results in Chapter 3

B.1 Proof of Proposition 4

First, we have the TAG framework Lin and Joseph (2020):

$$g(y) = \mu + z(\mathbf{x}) + \epsilon(\mathbf{x}),$$

where $z(\mathbf{x}) \sim GP(0, \tau^2 R(\cdot))$ and $\epsilon(\mathbf{x}) \sim N\left(0, \sigma^2\right)$. Then,

$$z_k(x_k) \sim GP(0, \tau_k^2 R_k(\cdot)),$$

for $k = 1, \ldots, d$. Let the correlation function for the Gaussian process be written as:

$$R(\boldsymbol{x} - \boldsymbol{x}') = \sum_{k=1}^{d} \omega_{lk} R_k\left(x_k - x_k'\right), \tag{27}$$

.

In order to find the marginal distribution, first the join distribution of $z_k(x_k)$ and $\mathbf{y}$ is needed. The new data can be written as $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

$$\begin{pmatrix} \mathbf{y} \\ z_k(x_k) \end{pmatrix} \sim N_{n+1}\left[\begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2\mathbf{R} + \sigma^2\mathbf{I} & \tau_k^2\mathbf{r}_k(x_k) \\ \tau_k^2\mathbf{r}_k(x_k)^T & \tau_k^2 \end{pmatrix}\right] \tag{28}$$

Where $\mathbf{r}_k(x_k)$ is the vector of correlations: $(R(x_k - x_{1k}), \ldots R(x_k - x_{nk}))$, $\mathbf{R}$ is the correlation matrix with the ijth element being $R_k(\mathbf{x}_i - \mathbf{x}_j)$. R and R_k are defined above.

Now let

$$\begin{bmatrix} A & \mathbf{b} \\ \mathbf{b}^T & c \end{bmatrix} = \begin{bmatrix} \tau^2\mathbf{R} + \sigma^2\mathbf{I} & \tau_k^2\mathbf{r}_k(x_k) \\ \tau_k^2\mathbf{r}_k(x_k)^T & \tau_k^2 \end{bmatrix}^{-1} \tag{29}$$

Then,

$$\begin{aligned}
z_k(x_k) &\propto \exp\left(-\frac{1}{2}\begin{bmatrix}\mathbf{y}\\ z_k(x_k)\end{bmatrix}^T\begin{bmatrix}A & \mathbf{b}\\ \mathbf{b}^T & c\end{bmatrix}\begin{bmatrix}\mathbf{y}\\ z_k(x_k)\end{bmatrix}+\frac{1}{2}z_k(x_k)^T\left(\tau^2\mathbf{R}+\sigma^2\mathbf{I}\right)^{-1}z_k(x_k)\right)\\
&= \exp\left(-\frac{1}{2}\mathbf{y}^2+2\left(\mathbf{b}^T z_k(x_k)\mathbf{y}+z_k(x_k)^T A z_k(x_k)\right)+\frac{1}{2}z_k(x_k)^T\left(\tau^2\mathbf{R}+\sigma^2\mathbf{I}\right)^{-1}z_k(x_k)\right)\\
&\propto \exp\left(-\frac{1}{2}\left(\mathbf{y}+\frac{\mathbf{b}^T z_k(x_k)}{c}\right)^2\right)
\end{aligned}$$

From 29 we can see that: $\left(\tau^2\mathbf{R}+\sigma^2\mathbf{I}\right)\mathbf{b}+c\tau_k^2\mathbf{r}_k(x_k)=0$ and $\tau_k^2\mathbf{r}_k(x_k)^T\mathbf{b}+c\tau_k^2=1$

Then, we can also see that:

$$\begin{aligned}
\mathbf{b} &= \frac{\left(\tau^2\mathbf{R}+\sigma^2\mathbf{I}\right)^{-1}\tau_k^2\mathbf{r}_k(x_k)}{\tau_k^2+\sigma_k^2-\left(\tau_k^2\mathbf{r}_k(x_k)\right)^T\left(\tau^2\mathbf{R}+\sigma^2\mathbf{I}\right)^{-1}\tau_k^2\mathbf{r}_k(x_k)}\\
c &= \frac{1}{\tau_k^2-\left(\tau_k^2\mathbf{r}_k(x_k)\right)^T\left(\tau^2\mathbf{R}+\sigma^2\mathbf{I}\right)^{-1}\tau_k^2\mathbf{r}_k(x_k)}
\end{aligned}$$

Now combining the above, we get:

$$z_k(x_k) \propto \exp\left(-\frac{\left(\mathbf{y}-\left(\tau_k^2\mathbf{r}_k(x_k)\right)^T\left(\tau^2\mathbf{R}+\sigma^2\mathbf{I}\right)^{-1}z_k(x_k)\right)^2}{2\left(\tau_k^2-\left(\tau_k^2\mathbf{r}_k(x_k)\right)^T\left(\tau^2\mathbf{R}+\sigma^2\mathbf{I}\right)^{-1}\tau_k^2\mathbf{r}_k(x_k)\right)}\right)$$

Therefore:

$$z_k(x_k) \sim N\left(\tau_k^2\mathbf{r}_k(x_k)^T\left(\tau^2\mathbf{R}+\sigma^2\mathbf{I}\right)^{-1}(g_\lambda(\mathbf{y})-\mu), \tau_k^2-\left(\tau_k^2\mathbf{r}_k(x_k)\right)^T\left(\tau^2\mathbf{R}+\sigma^2\mathbf{I}\right)^{-1}\tau_k^2\mathbf{r}_k(x_k)\right). \tag{30}$$

Which can be rewritten as:

$$z_k(x_k)|\boldsymbol{y} \sim N\left(\hat{z}_k(x_k), s_k^2(x_k)\right) \tag{31}$$

with

$$\hat{z}_k(x_k)=\omega_k\boldsymbol{r}_k(x_k)^\top\left(\boldsymbol{R}+\delta\boldsymbol{I}\right)^{-1}\left(g_\lambda(\boldsymbol{y})-\mu\right)$$

and

$$s_k^2(x_k)=\tau^2\omega_k\left(1-\boldsymbol{r}_k(x_k)^\top\left(\boldsymbol{R}+\delta\boldsymbol{I}\right)^{-1}\omega_k\boldsymbol{r}_k(x_k)\right),$$

where $\delta=\sigma^2/\tau^2$.

Appendix C Proofs of results in Chapter 4

C.1 Proof of Closed Form Expression of Expected Hypervolume Improvement

Given the equations for the hypervolume indicator and hypervolume improvement in (4.3) and (4.4), we can derive the closed for expression of expected hypervolume improvement found in (4.6). For this proof, let $\pi_{\hat{y},\hat{s}}\left(\mathbf{y}\right)$, be the bi-variate normal distribution that is th joint distribution of the conditional distributions of the objective functions and let $\phi_{\hat{y}_1,\hat{s}_1}\left(y_1\right)$ and $\phi_{\hat{y}_2,\hat{s}_2}\left(y_2\right)$ be the independent normal distributions that are the conditional distributions or each of the objective functions.

$$
\begin{aligned}
\mathrm{EHI}_n\left(\boldsymbol{x};\mathcal{P},\boldsymbol{t}\right) &= \mathrm{E}\left\{\mathrm{HVI}\left((Y_1(\boldsymbol{x}),Y_2(\boldsymbol{x})),\mathcal{P},\boldsymbol{t}\right)\right\}\\
&= \int_R \mathrm{HVI}\left((Y_1(\boldsymbol{x}),Y_2(\boldsymbol{x})),\mathcal{P},\boldsymbol{t}\right)\pi_{\hat{y},\hat{s}}\left(\mathbf{y}\right)d\mathbf{y}\\
&= \int_{y_1=-\infty}^{\infty}\int_{y_2=-\infty}^{\infty}\sum_{i=1}^{n+1}\Lambda_2\left[S_i\cap\Delta\left(y_1,y_2\right)\right]\cdot\pi_{\hat{y},\hat{s}}\left(y_1,y_2\right)dy_1dy_2\\
&= \sum_{i=1}^{p+1}\int_{y_1=-\infty}^{y_1^{i-1}}\int_{y_2=-\infty}^{y_2^{i}}\Lambda_2\left[S_i\cap\Delta\left(y_1,y_2\right)\right]\cdot\pi_{\hat{y},\hat{s}}\left(y_1,y_2\right)dy_1dy_2\\
&= \sum_{i=1}^{p+1}\int_{y_1=-\infty}^{y_1^{i-1}}\int_{y_2=-\infty}^{y_2^{i}}\Lambda_2\left[S_i\cap\Delta\left(y_1,y_2\right)\right]\cdot\phi_{\hat{y}_1,\hat{s}_1}\left(y_1\right)\cdot\phi_{\hat{y}_2,\hat{s}_2}\left(y_2\right)dy_1dy_2\\
&= \sum_{i=1}^{p+1}\int_{y_1-\infty}^{y_1^{i}}\left(y_1^{i-1}-y_1^{i}\right)\phi_{\hat{y}_1,\hat{s}_1}\left(y_1\right)dy_1\cdot\int_{y_2=-\infty}^{y_2^{i}}\left(y_2^{i}-y_2\right)\cdot\phi_{\hat{y}_2,\hat{s}_2}(y_2)dy_2\\
&+\sum_{i=1}^{p+1}\int_{y_1^{i}}^{y_1^{i-1}}\left(y_1^{i-1}-y_1\right)\phi_{\hat{y}_1,\hat{s}_1}\left(y_1\right)dy_1\cdot\int_{y_2=-\infty}^{y_2^{i}}\left(y_2^{i}-y_2\right)\cdot\phi_{\hat{y}_2,\hat{s}_2}(y_2)dy_2\\
&= \sum_{i=1}^{p+1}\int_{y_1=-\infty}^{y_1^{i}}\left(y_1^{i-1}-y_1^{i}\right)\phi_{\hat{y}_1,\hat{s}_1}\left(y_1\right)dy_1\int_{y_2=-\infty}^{y_2^{i}}\left(y_2^{i}-y_2\right)\frac{1}{\hat{s}_2(\boldsymbol{x})}\phi\left(\frac{y_2-\hat{y}_2(\boldsymbol{x})}{\hat{s}_2(\boldsymbol{x})}\right)dy_2\\
&+\sum_{i=1}^{p}\int_{y_1^{i}}^{y_1^{(i-1)}}\left(y_1^{i-1}-y_1\right)\frac{1}{\hat{s}_1}\phi\left(\frac{y_1-\hat{y}_1(\boldsymbol{x})}{\hat{s}_1}\right)\int_{y_2=-\infty}^{y_2^{i}}\left(y_2^{i}-y_2\right)\frac{1}{\hat{s}_2(\boldsymbol{x})}\phi\left(\frac{y_2-\hat{y}_2(\boldsymbol{x})}{\hat{s}_2(\boldsymbol{x})}\right)dy_2\\
&= \sum_{i=1}^{p+1}\left(y_1^{i-1}-y_1^{i}\right)\cdot\Phi\left(\frac{y_1^{i}-\hat{y}_1(\boldsymbol{x})}{\hat{s}_1(\boldsymbol{x})}\right)\cdot\Psi\left(y_2^{i},y_2^{i},\hat{y}_2(\boldsymbol{x}),\hat{s}_2(\boldsymbol{x})\right)\\
&+\sum_{i=1}^{p+1}\left(\Psi\left(y_1^{i-1},y_1^{i-1},\hat{y}_1(\boldsymbol{x}),\hat{s}_1(\boldsymbol{x})\right)-\Psi\left(y_1^{i-1},y_1^{i},\hat{y}_1(\boldsymbol{x}),\hat{s}_1(\boldsymbol{x})\right)\right)\\
&\cdot\Psi\left(y_2^{i},y_2^{i},\hat{y}_2(\boldsymbol{x}),\hat{s}_2(\boldsymbol{x})\right)
\end{aligned}
$$

where $\Lambda_2\left[S_i\cap\Delta\left(y_1,y_2\right)\right]$ is the Lebesgue measure on R^2, S_i are the sections of the dominated area bounded by $S_i=\left(\left(y_1^i,-\infty\right)^\top,\left(y_1^{i-1},y_2^i\right)^\top\right)$, and $\Delta\left(y_1,y_2\right)$ is the dominated area represented by the new point.

C.2 Example of Expected Hypervolume Improvement Calculation

This example will be a two-dimension numerical calculation of the equation derived above. It will show the calculation of the EHI for a set of three two-dimensional points. The goal of the EHI calculations in this example will be to maximize the the objective functions and find the Pareto set.

For this example, let the Pareto set be:

$$\mathcal{P} = \begin{bmatrix} 3 & 1 \\ 2 & 1.5 \\ 1 & 2.5 \end{bmatrix}$$

Let the data follow the form $y_i \sim N(\hat{y}_i, \hat{s}_i)$. Then, the values for this example are given by: $y_1 \sim N(2, 0.7)$ and $y_2 \sim N(1.5, 0.6)$. Let the reference point be $\boldsymbol{t} = (4, 4)^\top$. Lastly, let the point from which the EHi will be calculate be new point $\mathbf{y}^{new} = (1.50, 0.50)$ which is a rounded value selected randomly from the independent normal distribution based on the values of μ and σ above. This point is draw in the figure to illustrate the area of improvement and the original Pareto set. The EHI calculation finds the average improvement over all possible nondominated points in the area bounded by the reference point.

C.2.1 Calculations

Table of values:

	y_1	y_2
y^0	4	$-\infty$
y^1	3	1
y^2	2	1.5
y^3	1	2.5
y^4	$-\infty$	4

Table of Calculations:

i	Calculation	Value
1	$(4-3)\Phi\left(\frac{3-2}{0.7}\right)\Psi(1,1,1.5,0.6) + (\Psi(4,4,2,0.7) - \Psi(4,3,2,0.7))\,\Psi(1,1,1.5,0.6)$	0.0663
2	$(3-2)\Phi\left(\frac{2-2}{0.7}\right)\Psi(1.5,1.5,1.5,0.6) + (\Psi(3,3,2,0.7) - \Psi(3,2,2,0.7))\,\Psi(1.5,1.5,1.5,0.6)$	0.1783
3	$(2-1)\Phi\left(\frac{1-2}{0.7}\right)\Psi(2.5,2.5,1.5,0.6) + (\Psi(2,2,2,0.7) - \Psi(2,1,2,0.7))\,\Psi(2.5,2.5,1.5,0.6)$	0.2582
4	$(1--\infty)\Phi\left(\frac{-\infty-2}{0.7}\right)\Psi(4,4,1.5,0.6) + (\Psi(1,1,2,0.7) - \Psi(1,-\infty,2,0.7))\,\Psi(4,4,1.5,0.6)$	0.0602

$$
\begin{aligned}
\mathrm{EHI}_n\left(\boldsymbol{x};\mathcal{P},\boldsymbol{t}\right) &= \sum_{i=1}^{p+1}\left(y_1^{i-1}-y_1^{i}\right)\cdot\Phi\left(\frac{y_1^{i}-\hat{y}_1(\boldsymbol{x})}{\hat{s}_1(\boldsymbol{x})}\right)\cdot\Psi\left(y_2^{i},y_2^{i},\hat{y}_2(\boldsymbol{x}),\hat{s}_2(\boldsymbol{x})\right)\\
&+\sum_{i=1}^{p+1}\left(\Psi\left(y_1^{i-1},y_1^{i-1},\hat{y}_1(\boldsymbol{x}),\hat{s}_1(\boldsymbol{x})\right)-\Psi\left(y_1^{i-1},y_1^{i},\hat{y}_1(\boldsymbol{x}),\hat{s}_1(\boldsymbol{x})\right)\right)\\
&\cdot\Psi\left(y_2^{i},y_2^{i},\hat{y}_2(\boldsymbol{x}),\hat{s}_2(\boldsymbol{x})\right)\\
&= 0.563
\end{aligned}
$$

The expected hypervolume improvement for the Pareto set described above is 0.563. This value was equal to that found used the Matlab code given by Michael Emmerich and Andre Deutz, LIACS, Leiden University, 2010. Hupkens et al. (2015). This is shown in Figure 5, where the red area in the plot is the improvement if the point y^{new} is added andthe blue area is the hypervolume dominated by the current Pareto set.

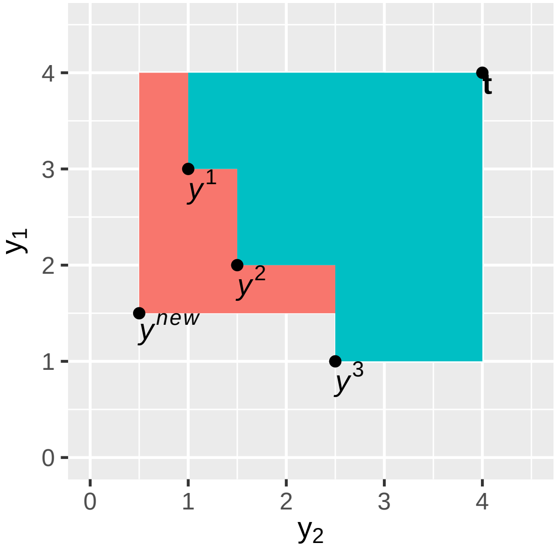

Figure 5: Illustration of dominated area for a sample EHI calculation.

Bibliography

M. Abolhasani and K. F. Jensen. Oscillatory multiphase flow strategy for chemistry and biology. *Lab on a Chip*, 16(15):2775–2784, 2016.

A. Adamo, R. L. Beingessner, M. Behnam, J. Chen, T. F. Jamison, K. F. Jensen, J.-C. M. Monbaliu, A. S. Myerson, E. M. Revalor, D. R. Snead, et al. On-demand continuous-flow production of pharmaceuticals in a compact, reconfigurable system. *Science*, 352(6281):61–67, 2016.

S. Alarie, C. Audet, A. E. Gheribi, M. Kokkolaras, and S. Le Digabel. Two decades of blackbox optimization applications. *EURO Journal on Computational Optimization*, 9:100011, 2021. ISSN 2192-4406. doi: https://doi.org/10.1016/j.ejco.2021.100011. URL `https://www.sciencedirect.com/science/article/pii/S2192440621001386`.

C. Audet, J. Bigeon, D. Cartier, and S. L. Digabel. Performance indicators in multiobjective optimization. *European Journal of Operational Research*, 292:397–422, 2021.

R. Bechhofer. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics*, 25(1):16–39, 1954.

M. Binois and V. Picheny. GPareto: An R package for gaussian-process-based multi-objective optimization and analysis. *Journal of Statistical Software*, 89(8):1–30, 2019. doi: 10.18637/jss.v089.i08.

J. Bossek. ecr 2.0: A modular framework for evolutionary computation in r. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, GECCO '17, pages 1187–1193, 2017. ISBN 978-1-4503-4939-0.

G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964. ISSN 00359246. URL `http://www.jstor.org/stable/2984418`.

S. P. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

E. Brochu, V. M. Cora, and N. De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

J. Butler, D. J. Morrice, and P. W. Mullarkey. A multiple attribute utility theory approach to ranking and selection. *Management Science*, 47(6):800–816, 2001.

R. Carnell. *lhs: Latin Hypercube Samples*, 2022. URL `https://CRAN.R-project.org/package=lhs`. R package version 1.1.6.

Y. Chen and I. O. Ryzhov. Complete expected improvement converges to an optimal budget allocation. *Advances in Applied Probability*, 51(1):209–235, 2019. doi: 10.1017/apr.2019.9.

Y. Chen and I. O. Ryzhov. Balancing optimal large deviations in sequential selection. *Management Science*, 2022.

Y. Chen, Q. Zhang, M. Li, and W. Cai. Sequential selection for accelerated life testing via approximate bayesian inference. *Naval Research Logistics (NRL)*, 69(2):336–351, 2022.

S. E. Chick, J. Branke, and C. Schmidt. Sequential Sampling to Myopically Maximize the Expected Value of Information. *INFORMS Journal on Computing*, 22(1):71–80, 2010.

S. Daulton, M. Balandat, and E. Bakshy. Differentiable expected hypervolume improvement for parallel multi-objective bayesian optimization. *Advances in Neural Information Processing Systems*, 33:9851–9864, 2020.

P. de Castro, H. Stewart, C. Turner, M. Wiecek, G. Hartman, D. Rizzo, D. Gorsich, A. Skowronska, and R. Agusti. Decomposition and coordination to support tradespace analysis for ground vehicle systems. Technical report, SAE Technical Paper, 2022.

M. L. Eaton. Multivariate statistics: A vector space approach. *Lecture Notes-Monograph Series*, 53: i–512, 2007. ISSN 07492170. URL http://www.jstor.org/stable/20461449.

M. Ehrgott. *Multicriteria optimization*, volume 491. Springer Science & Business Media, 2005.

M. T. Emmerich, A. H. Deutz, and J. W. Klinkenberg. Hypervolume-based expected improvement: Monotonicity properties and exact computation. In *2011 IEEE Congress of Evolutionary Computation (CEC)*, pages 2147–2154. IEEE, 2011.

S. L. Faulkenberg and M. M. Wiecek. On the quality of discrete representations in multiple objective programming. *Optimization and Engineering*, 11:423–440, 2010.

J. E. Fieldsend, R. M. Everson, and S. Singh. Using unconstrained elite archives for multiobjective optimization. *IEEE Transactions on Evolutionary Computation*, 7(3):305–323, 2003.

J. Fliege and H. Xu. Stochastic multiobjective optimization: sample average approximation and applications. *Journal of optimization theory and applications*, 151:135–162, 2011.

C. M. Fonseca, L. Paquete, and M. López-Ibánez. An improved dimension-sweep algorithm for the hypervolume indicator. In *2006 IEEE international conference on evolutionary computation*, pages 1157–1163. IEEE, 2006.

P. I. Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

P. I. Frazier, W. B. Powell, and S. Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.

E. C. Garrido-Merchán and D. Hernández-Lobato. Dealing with categorical and integer-valued variables in bayesian optimization with gaussian processes. *Neurocomputing*, 380:20–35, 2020.

P. W. Glynn and S. Juneja. A large deviations perspective on ordinal optimization. In R. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, editors, *Proceedings of the 2004 Winter Simulation Conference*, pages 577–585, 2004.

R. B. Gramacy. *Surrogates: Gaussian Process Modeling, Design and Optimization for the Applied Sciences*. Chapman Hall/CRC, Boca Raton, Florida, 2020. http://bobby.gramacy.com/surrogates/.

A. P. Guerreiro, C. M. Fonseca, and L. Paquete. The hypervolume indicator: Computational problems and algorithms. *ACM Computing Surveys (CSUR)*, 54(6):1–42, 2021.

S. Gupta and K. Miescke. Bayesian look ahead one-stage sampling allocations for selection of the best population. *Journal of Statistical Planning and Inference*, 54(2):229–244, 1996.

G. D. Hadiwinoto, P. C. Kwok, H. H. Tong, S. N. Wong, S. F. Chow, and R. Lakerveld. Integrated continuous plug-flow crystallization and spray drying of pharmaceuticals for dry powder inhalation. *Industrial & Engineering Chemistry Research*, 58(36):16843–16857, 2019.

A. Herzel, S. Ruzika, and C. Thielen. Appoximation methods for multiobjective optimization problems: A survey. *INFORMS Journal on Computing*, 33(4):1284–1299, 2021.

L. J. Hong, W. Fan, and J. Luo. Review on ranking and selection: A new perspective. *Frontiers of Engineering Management*, 8(3):321–343, 2021.

L. J. Hong, G. Jiang, and Y. Zhong. Solving large-scale fixed-budget ranking and selection problems. *INFORMS Journal on Computing*, 34(6):2930–2949, 2022.

S. R. Hunter and B. McClosky. Maximizing quantitative traits in the mating design problem via simulation-based pareto estimation. *IIE Transactions*, 48(6):565–578, 2016.

S. R. Hunter and B. L. Nelson. Parallel ranking and selection. In *Advances in Modeling and Simulation: Seminal Research from 50 Years of Winter Simulation Conferences*, pages 249–275. Springer, 2017.

S. R. Hunter, E. A. Applegate, V. Arora, B. Chong, K. Cooper, O. Rincón-Guevara, and C. Vivas-Valencia. An introduction to multiobjective simulation optimization. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 29(1):1–36, 2019.

I. Hupkens, A. H. Deutz, K. Yang, and M. T. Emmerich. Faster exact algorithms for computing expected hypervolume improvement. In *EMO (2)*, pages 65–79, 2015.

F. Jäkel, B. Schölkopf, and F. A. Wichmann. A tutorial on kernel methods for categorization. *Journal of Mathematical Psychology*, 51(6):343–358, 2007.

M. Jiang and R. Braatz. Integrated control of continuous (bio) pharmaceutical manufacturing. *American Pharmaceutical Review*, 19(6):110–115, 2016.

M. Jiang, Z. Zhu, E. Jimenez, C. D. Papageorgiou, J. Waetzig, A. Hardy, M. Langston, and R. D. Braatz. Continuous-flow tubular crystallization in slugs spontaneously induced by hydrodynamics. *Crystal growth & design*, 14(2):851–860, 2014.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions, volume 2*, volume 289. John wiley & sons, 1995.

D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492, 1998.

K. Kandasamy, J. Schneider, and B. Póczos. High dimensional bayesian optimisation and bandits via additive models. In *International conference on machine learning*, pages 295–304. PMLR, 2015.

C. M. Kerfonta, S. Kim, Y. Chen, Q. Zhang, and M. Jiang. Sequential selection for minimizing the variance with application to crystallization experiments. *The American Statistician*, (just-accepted):1–16, 2024.

J. D. Knowles, D. W. Corne, and M. Fleischer. Bounded archiving using the lebesgue measure. In *The 2003 Congress on Evolutionary Computation, 2003. CEC'03.*, volume 4, pages 2490–2497. IEEE, 2003.

K.-R. Koch. *Introduction to Bayesian statistics*. Springer Science & Business Media, 2007.

L. H. Lee, C. U. Lee, and Y. P. Tan. A multi-objective genetic algorithm for robust flight scheduling using simulation. *European Journal of Operational Research*, 177(3):1948–1968, 2007.

C. Li, S. Gupta, S. Rana, V. Nguyen, S. Venkatesh, and A. Shilton. High dimensional bayesian optimization using dropout. *arXiv preprint arXiv:1802.05400*, 2018.

L.-H. Lin and V. R. Joseph. Transformation and additivity in gaussian processes. *Technometrics*, 62(4):525–535, 2020.

L.-H. Lin and V. R. Joseph. *TAG: Transformed Additive Gaussian Processes*, 2021. URL `https://CRAN.R-project.org/package=TAG`. R package version 0.5.1.

N. Loka, I. Couckuyt, F. Garbuglia, D. Spina, I. Van Nieuwenhuyse, and T. Dhaene. Bi-objective bayesian optimization of engineering problems with cheap and expensive cost functions. *Engineering with Computers*, pages 1–11, 2022.

D. J. MacKay et al. Introduction to gaussian processes. *NATO ASI series F computer and systems sciences*, 168:133–166, 1998.

M. Malu, G. Dasarathy, and A. Spanias. Bayesian optimization in high-dimensional spaces: A brief survey. In *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pages 1–8. IEEE, 2021.

R. Martinez-Cantin, K. Tee, and M. McCourt. Practical bayesian optimization in the presence of outliers. In *International conference on artificial intelligence and statistics*, pages 1722–1731. PMLR, 2018.

S. Mascia, P. L. Heider, H. Zhang, R. Lakerveld, B. Benyahia, P. I. Barton, R. D. Braatz, C. L. Cooney, J. M. Evans, T. F. Jamison, et al. End-to-end continuous manufacturing of pharmaceuticals: integrated synthesis, purification, and final dosage formation. *Angewandte Chemie*, 125(47): 12585–12589, 2013.

M. A. McDonald, H. Salami, P. R. Harris, C. E. Lagerman, X. Yang, A. S. Bommarius, M. A. Grover, and R. W. Rousseau. Reactive crystallization: a review. *Reaction Chemistry & Engineering*, 6 (3):364–400, 2021.

M. McLeod, S. Roberts, and M. A. Osborne. Optimization, fast and slow: optimally switching between local and bayesian optimization. In *International Conference on Machine Learning*, pages 3443–3452. PMLR, 2018.

J. Mockus. The bayesian approach to global optimization. In *System Modeling and Optimization: Proceedings of the 10th IFIP Conference New York City, USA, August 31–September 4, 1981*, pages 473–481. Springer, 2005.

R. Moriconi, M. P. Deisenroth, and K. Sesh Kumar. High-dimensional bayesian optimization using low-dimensional feature spaces. *Machine Learning*, 109:1925–1943, 2020.

M. Mou, A. Patel, S. Mallick, B. P. Thapaliya, M. P. Paranthaman, J. H. Mugumya, M. L. Rasche, R. B. Gupta, S. Saleh, S. Kothe, et al. Scalable advanced li (ni0. 8co0. 1mn0. 1) o2 cathode materials from a slug flow continuous process. *ACS omega*, 7(46):42408–42417, 2022.

E. Paulson. A sequential procedure for selecting the population with the largest mean from k normal populations. *The Annals of Mathematical Statistics*, pages 174–180, 1964.

V. Picheny, D. G. Green, and O. Roustant. *DiceOptim: Kriging-Based Optimization for Computer Experiments*, 2021. URL `https://CRAN.R-project.org/package=DiceOptim`. R package version 2.1.1.

W. Powell and I. Ryzhov. *Optimal Learning*. Wiley Series in Probability and Statistics. Wiley, 2012. ISBN 9780470596692. URL `https://books.google.com/books?id=hnsVMbx5HOAC`.

C. Qin, D. Klabjan, and D. Russo. Improving the expected improvement algorithm. *Advances in Neural Information Processing Systems*, 30, 2017.

O. Roustant, D. Ginsbourger, and Y. Deville. Dicekriging, diceoptim: Two r packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of Statistical Software*, 51:1–55, 2012.

O. Roustant, D. Ginsbourger, Y. D. Contributors, and M. O. Roustant. Package 'dicekriging', 2015.

D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

S. Ruzika and M. M. Wiecek. Approximation methods in multiobjective programming. *Journal of optimization theory and applications*, 126(3):473–501, 2005.

I. O. Ryzhov. On the convergence rates of expected improvement methods. *Operations Research*, 64(6):1515–1528, 2016.

J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and Analysis of Computer Experiments. *Statistical Science*, 4(4):409 – 423, 1989. doi: 10.1214/ss/1177012413. URL `https://doi.org/10.1214/ss/1177012413`.

P. L. Salemi, E. Song, B. L. Nelson, and J. Staum. Gaussian markov random fields for discrete optimization via simulation: Framework and algorithms. *Operations Research*, 67(1):250–266, 2019. doi: 10.1287/opre.2018.1778.

E. Schulz, M. Speekenbrink, and A. Krause. A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 85:1–16, 2018.

B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016. doi: 10.1109/JPROC.2015.2494218.

J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

A. Spagnol, R. L. Riche, and S. D. Veiga. Global sensitivity analysis for optimization with variable selection. *SIAM/ASA Journal on uncertainty quantification*, 7(2):417–443, 2019.

M. Stein. Large sample properties of simulations using latin hypercube sampling. *Technometrics*, 29(2):143–151, 1987.

M. Tesch, J. Schneider, and H. Choset. Using response surfaces and expected improvement to optimize snake robot gait parameters. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1069–1074, 2011.

L. Trailovic and L. Y. Pao. Computing budget allocation for efficient ranking and selection of variances with application to target tracking algorithms. *IEEE Transactions on Automatic Control*, 49(1):58–67, 2004.

N. Variankaval, A. S. Cote, and M. F. Doherty. From form to function: Crystallization of active pharmaceutical ingredients. *AIChE journal*, 54(7):1682–1688, 2008.

K. K. Vu, C. d'Ambrosio, Y. Hamadi, and L. Liberti. Surrogate-based methods for black-box optimization. *International Transactions in Operational Research*, 24(3):393–424, 2017.

T. Wang, J. Xu, J.-Q. Hu, and C.-H. Chen. Optimal computing budget allocation for regression with gradient information. *Automatica*, 134:109927, 2021.

X. Wang, Y. Jin, S. Schmitt, and M. Olhofer. Recent advances in bayesian optimization. *ACM Computing Surveys*, 55(13s):1–36, 2023.

Z. Wang, C. Gehring, P. Kohli, and S. Jegelka. Batched large-scale bayesian optimization in high-dimensional spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 745–754. PMLR, 2018.

J. Wilson, F. Hutter, and M. Deisenroth. Maximizing acquisition functions for bayesian optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018a. URL `https://proceedings.neurips.cc/paper_files/paper/2018/file/498f2c21688f6451d9f5fd09d53edda7-Paper.pdf`.

J. Wilson, F. Hutter, and M. Deisenroth. Maximizing acquisition functions for bayesian optimization. *Advances in neural information processing systems*, 31, 2018b.

H. Xiao, Y. Zhang, G. Kou, S. Zhang, and J. Branke. Ranking and selection for pairwise comparison. *Naval Research Logistics (NRL)*, 70(3):284–302, 2023.

Y. Xu and A. Zeevi. Bayesian design principles for frequentist sequential learning. In *International Conference on Machine Learning*, pages 38768–38800. PMLR, 2023.

K. Yang, A. Deutz, Z. Yang, T. Back, and M. Emmerich. Truncated expected hypervolume improvement: Exact computation and application. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 4350–4357. IEEE, 2016.

K. Yang, M. Emmerich, A. Deutz, and T. Bäck. Multi-objective bayesian global optimization using expected hypervolume improvement gradient. *Swarm and Evolutionary Computation*, 44:945–956, 2019. ISSN 2210-6502.

C. D. P. V. Zambrano and M. Jiang. l-glutamic acid crystals of pure form and uniform size distribution from continuous non-seeded reaction crystallization in slug flow. *CrystEngComm*, 25: 2227–2236, 2023. doi: 10.1039/D2CE01528E.

G. Zhang, H. Li, and Y. Peng. Sequential sampling for a ranking and selection problem with exponential sampling distributions. In *2020 Winter Simulation Conference (WSC)*, pages 2984–2995, 2020.

H. Zhang. Multi-objective simulation-optimization for earthmoving operations. *Automation in construction*, 18(1):79–86, 2008.

Y. Zhang and W. I. Notz. Computer experiments with qualitative and quantitative variables: A review and reexamination. *Quality Engineering*, 27(1):2–13, 2015. doi: 10.1080/08982112.2015.968039. URL `https://doi.org/10.1080/08982112.2015.968039`.

Y. Zhong and L. J. Hong. Knockout-tournament procedures for large-scale ranking and selection in parallel computing environments. *Operations Research*, 70(1):432–453, 2022.