Discrete-event Simulation and Optimization to Improve the Performance of a Healthcare System

Hao Huang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Zelda B. Zabinsky, Chair

Joseph A. Heim

Paul A. Fishman

Program Authorized to Offer Degree:
Industrial & Systems Engineering

University of Washington

Abstract

Discrete-event Simulation and Optimization to Improve the Performance of a Healthcare System

Hao Huang

Chair of the Supervisory Committee:
Professor Zelda B. Zabinsky
Industrial & Systems Engineering

Healthcare systems have attracted the attention of management and analysis due to their high percentage of the gross domestic product (GDP) and increasing rate of growth of expenditures. Within the various types of healthcare problems, this dissertation focuses on resource allocation decisions because they can significantly improve the performance of care delivery systems. Since most healthcare systems have significant uncertainty and complexity, and mathematical closed form models may not exist, discrete-event simulation is considered here as a suitable approach to model a healthcare system. Also, multiple objectives are a natural consideration in healthcare problems where at least cost and health outcome are usually considered. This research has two contributions: developing decision support models for three healthcare systems; and developing algorithms for simulation optimization that can be used to provide insights into healthcare problems.

The three healthcare resource allocation problems in this dissertation are: (i) design an occupational health campaign in which two departments in a hospital must coordinate their activities in order to provide a high level of service with intensive service demands in a compressed time period; (ii) investigate how to optimize hepatitis C screening and treatment allocation strategies; (iii) evaluate cost and health quality trade-offs when allocating portable and console type ultrasound instruments for orthopedic clinics.

These healthcare systems are modeled with discrete-event simulation and optimization.

The occupational health campaign was designed with limited resources from a clinic and a laboratory to meet high demands for influenza immunization and tuberculosis (TB) screening. All design configurations of the campaign were simulated to understand the system performance. Hepatitis C screening and treatment allocation strategies involving budget constraints were simulated to maximize the total discounted health utility gain of a cohort over its lifetime. The number of strategies was too large to be exhaustively investigated, so the first simulation optimization algorithm was applied to explore sensitivity of solutions. The portable ultrasound machine allocation problem considered portable ultrasound machines in orthopedic clinics as a viable alternative to a centralized MRI service, regarding the trade-off between cost and health utility loss. This problem also involved a large number of designs as well as a high degree of randomness and multiple objectives. The use of the second simulation optimization technique played an important role in this situation.

The second contribution in this dissertation is to develop two simulation optimization algorithms that can be applied to a general black-box simulation with randomness, such as those used in the hepatitis C screening and treatment strategy and the portable ultrasound instrument allocation problem. The first algorithm focuses on single objective problems and approximates a target level set of solutions bounded by a quantile. The approximated level set provides insights through the shape of the level set and allows decision makers to consider issues out of the simulation model. The second algorithm is designed for multiple objective problems and provides an approximated Pareto optimal set and efficient frontier for decision makers to investigate the trade-offs between objective functions. The algorithms include several important characteristics. First, they can process mixed integer and continuous variables. Second, the objective function can be a black-box simulation with noise or uncertainty. Third, statistical qualities of solutions provided by the algorithms are derived to support decision making. The developed algorithms are evaluated with test functions and applied to the hepatitis C screening and treatment strategy and the ultrasound instrument allocation problem.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I wish to express my sincere appreciation to Professor Zelda B. Zabinsky and Dr. Joseph Heim for their patient guidance and support through my research. They provided endless encouragement and advising throughout my research on both theoretical and practical application perspectives. This dissertation exists only with their help and expertise. I would like to extend my thanks to Professor Paul Fishman and Professor Archis Ghate, who help me shape the dissertation and their willingness to be in my committee. I would also like to thank Professor Abraham Flaxman to be a member of my committee. I am also thankful to Professor Shan Liu for bringing the opportunity to apply my algorithm to a healthcare problem. I would also wish to thank the great colleagues I have worked with at Seattle Children's Hospital. My gratitude also goes to all the friends and the administrative staff at the Industrial and Systems Engineering department at the University of Washington for their support and friendship throughout my doctoral program. The happy hours, BBQ and other gatherings truly delighted my experience at University of Washington. Certainly, I express my deepest gratitude to my parents who have supported me endlessly in my life and have faith that I can accomplish this milestone. I would also like to extend my gratitude to the dear friends of mine, Ching-Wei and Jerome, who accompanied me through the journey with their friendship. Also, my thank also goes to all my friends were part of the Anime Discovery Project club, where we shared similar habits and developed friendships.

The research presented in this dissertation was also supported in part by the Department of Laboratory Medicine at Seattle Childrens Hospital and National Science Foundation under grant CMMI-1235484.

Chapter 1

INTRODUCTION

1.1 Motivation

According to the World Health Organization (WHO), 17 percent of the United States' Gross Domestic Product in 2012 was healthcare expenditure [90], which is the highest ratio in the world. Also, annual personal health care expenditures increased from 1.2 trillion dollars in 2000 to 2.2 trillion dollars in 2010 [19]. The high ratio and significant increasing rate of growth of expenditures indicates that there are opportunities to provide better healthcare through a more cost-effective system.

Discrete-event simulation is a useful analysis tool for understanding and improving healthcare systems [34, 41, 52]. Also, optimization techniques can be used to further investigate healthcare systems and aid decision makers in designing an efficient and effective system. This dissertation considers discrete-event simulation as a primary approach to modeling a healthcare system, in the situation where a closed-form mathematical model may not exist because of the extreme complexity of the target system or requiring too many impractical assumptions. Furthermore, discrete-event simulation is able to represent the stochastic nature of healthcare systems conveniently and evaluate performance with detailed processes [76]. Another major advantage of discrete-event simulation is its extremely low cost and short running time compared to real experiments.

This dissertation considers optimization techniques as a valuable approach to find not only the optimal design but also a set of good solutions to provide insights for decision makers. Since the simulation model may also ignore some practical considerations, a set of near-optimal solutions is useful to provide a form of sensitivity analysis. However, even though simulation requires less resources than real experiments, simulation could still need significant computation time, especially as the system becomes more complicated and the number of potential designs becomes large. Therefore, complete enumeration of all potential

designs may be impractical. Simulation optimization algorithms can provide solutions that have certain quality (near-optimal) without running all combinations of designs.

A wide range of approaches have been proposed to find the best configuration and/or to investigate performance of a complex system with limited computational resources. For example, design of experiments is a widely used approach, which systematically selects decision variables' levels for evaluation in the simulation of a target system. Within these approaches, simulation optimization combines optimization methodologies with statistics, to efficiently evaluate complicated systems. Many simulation optimization methodologies aim to provide a comprehensive understanding of the target system, with a global optimal design, a ranking of solutions, or a family of solutions. However, only a limited number of simulation optimization methodologies addresses multiple objectives [35]. Most optimization software embedded in simulation softwares are meta-heuristic methods such as genetic algorithms [35], and the software Simio (used in this dissertation) includes two simulation optimization add-ins for finding the best design, OptQuest and Select Best Scenario [67].

One way of gaining insights is through sensitivity analysis, which quantifies the impact on the objective function of small changes in the decision variables. However, it is very difficult to perform a sensitivity analysis when the objective function is a black-box simulation, except empirically sampling the neighbor of the approximated global optimum. We consider that the shape of an approximate level set provides insights for decision makers like a type of sensitivity analysis. Furthermore, an approximate level set allows the decision makers to incorporate not quantifiable considerations, which are not included in the simulation model, by selecting a good enough solution from the level set.

A major challenge when optimizing these healthcare system designs is multiple objectives. In practice, there are usually multiple metrics, such as cost, patient health outcome, and patient satisfaction, that are important in a good system. Integrating these into one metric, such as weighted cost or weighted utility is a common approach. However, instead of a optimal solution to a single objective function, a set of designs that considers the trade-offs between the multiple metrics could give decision makers more insights into the system. Identifying the whole Pareto optimal solution set can provide insights into the system and suggest combinations of design variables that performs well.

This dissertation addresses the challenges of simulating and optimizing the performance of healthcare systems by introducing two new algorithms and applying them to specific healthcare resource allocation problems.

1.2 Research Contributions

The three different healthcare resource allocation challenges investigated in this dissertation are: staffing and scheduling an occupational health campaign; hepatitis C screening and treatment allocation strategies; and ultrasound machine allocation across multiple orthopaedic clinics. The occupational health campaign demonstrates the value of discrete-event simulation models in decision making. The number of potential designs of the campaign is small enough to be enumerated within reasonable assumptions. The second problem is to determine budget resources for either screening and treatment in a hepatitis C cohort through the lifetime gain in health utility. We treat the system dynamic model (as in [58]) as a black box simulation and find a set of good designs with statistical assurance of the performance. The first new algorithm that approximates a level set provided insights into screening and treatment policies. The third problem is an ultrasound machine allocation problem with MRI capacity allocation and multiple objectives, cost and health utility loss. A simulation of the system including patient wait time and potential mis-diagnosis is developed. For this problem, the second new algorithm is used to provide a Pareto set of solutions along the efficient frontier to illustrate trade-offs between cost and health utility loss. In summary, the three applications to healthcare systems are:

- Staffing resource allocation of an occupational health campaign:
  An occupational health campaign was designed that required resources from a clinic and a laboratory to meet high demands for influenza immunization and tuberculosis (TB) screening. A discrete-event simulation model was developed to analyze the campaign under a variety of scenarios and designs. An adaptive procedure was implemented to refine the simulation model as the campaign unfolded and allow reallocation of resources to achieve performance goals. With the discrete-event simulation model, the clinic and laboratory managers were able to coordinate their operations and de-

sign a campaign that was concluded as short as possible while returning the TB test results within 14 days.

- Hepatitis C screening and treatment allocation strategy:
  Optimizing the budget allocation between treatment and screening of hepatitis C is an important and challenging problem. The decision variables are fractions of the budget used for screening in each decision time period, with the remainder of the budget used for treatment. The objective is to maximize the total discounted health utility gain over the lifetime of each health cohort. By using the simulation optimization algorithm developed in this dissertation, ranges on decision variables providing the top 10% performance are provided to gain insights in the problem.

- Portable ultrasound machine allocation for orthopedic clinics:
  Portable ultrasound machines in clinics have become a viable alternative to a centralized MRI service. Since decisions need to be made at each clinic and imaging center, regarding the number of portable ultrasound machines and the amount of MRI capacity to reserve, a simulation model was developed and used to evaluate performances. The trade-offs between two objective functions, cost and health utility loss, are analyzed for the decision makers with the multiple objective simulation optimization algorithm developed in this dissertation.

In order to aid in decision making for these, and other healthcare resource allocation problems, two simulation optimization algorithms have been developed. The algorithms are able to handle uncertainties that are common to these problems by evaluating performance with a black-box discrete-event simulation with noise, mixed integer/continuous design variables, and multiple objectives.

The first algorithm, called Probabilistic Branch and Bound (PBnB) for level set approximation, aims to approximate a set of solutions associated with the best 100δ percent of solutions, i.e., the level set. A level set allows decision makers to select from near-optimal solutions considering factors out of the model. Also, the shape of the level set provides insights on the sensitivities of each decision variable. The algorithm uses statistical estimation

and a partitioning scheme to categorize subregions as being inside the level set, outside the level set, or undecided. Probability bounds are derived to statistically justify the quality of solutions that are provided by the algorithm. The quality of solutions provide guidance for decision makers to understand their confidence in the decisions.

The second algorithm, called Multiple Objective Probabilistic Branch and Bound (MOPBnB), is a multi-objective version that approximates the Pareto optimal set of solutions to allow decision makers to examine trade-offs between objective functions. It is also capable of handling uncertainties via a discrete-event simulation model with mixed integer/continuous variables. A probability bound also is derived regarding the quality of the solutions.

In summary, this dissertation provides solutions and insights into healthcare resource allocation problems with discrete-event simulation models and two simulation optimization algorithms.

Chapter 2

BACKGROUND AND LITERATURE SURVEY

2.1 Allocation of Healthcare System Resources

Petrou and Wolstenholme [72] discussed several possible approaches to allocating resources in both market-based and publicly funded healthcare systems. Ethical considerations are important to provide fair services to all patients in a public healthcare system. Similar to Petrou and Wolstenholme's review [72], a series of studies focus on the methodology of healthcare design assessment. For a healthcare system design, cost is not the only metric. The quality of the service or health states of the community may be equally important. Quality of life assessment is an approach to evaluate the health effectiveness of each design on resource allocation problems [31, 32], with quality adjusted life years and life years gained, the most commonly used metric. Economic evaluation is a popular method which provides information of the efficiency of alternatives and compares costs and consequences such as quality adjusted life years of at least two alternatives to support decision makers [16, 70].

Resource allocation is studied for various healthcare systems, physician's practices [82], scarce medical interventions such as organs, resources for genetic tests [71, 78], operating rooms and staffs [17, 29, 61], and outpatient services including clinics and emergency departments. For healthcare policy designs, many approaches are used to support the planning for different diseases. Rauner et al. [75] proposes a generic dynamic resource allocation model for chronic diseases and uses breast cancer as a case study to demonstrate the model. Deo et al. [30] focuses on the planning of HIV screening and care providing with a non-linear dynamic programming model for the Greater Los Angeles Veterans Health Administration station. Hepatitis C studies usually use a cohort approach because there is a specific high risk cohort in the U.S. [86]. These cohort studies propose a model to simulate the health and disease progression of the target cohort and use the model to evaluate the cost effectiveness of each policy design. Liu et al. [62] model the cohort with a decision-analytic

Markov model. Najafzadeh et al. [66] use discrete-event simulation and Chhatwal et al. [25] simulate the cohort by HCV infection natural history microsimulation.

In this dissertation, the problems' metrics are evaluated by simulation models which include the quality of life concept. Also, different metrics for designs of the system are provided to the decision makers.

2.2 Application of Discrete-event Simulation in Healthcare Systems

Currently, more decision support tools, which are popular in other fields, are being introduced to healthcare management problems, including discrete-event simulation. Banks [8] provided a general guidance for applying simulation in healthcare, focusing on patient services. Also, Jun et al. [52] discussed the use of discrete-event simulation on modeling a healthcare system with its outstanding flexibility. Guenal and Pidd [41] categorized the application of discrete-event simulation to healthcare into five categories: accident and emergency (A&E) departments, inpatient facilities, outpatient facilities, other hospital units, and whole hospital simulations. In these categories, previous studies focused on outpatient facilities and emergency departments because of their patient-intensive nature [33, 41, 52]. Weng and Houshmand [89] used simulation for clinic planning. For the emergency department, Garcia et al. [39] reduced patient waiting time for emergency medicine by introducing a simulation model. Brailsford et al. [15] also used simulation modeling to evaluate accident and emergency departments in Nottingham, England, which plays an important role in the "total conceptional map model". In addition, Ahmed and Alkhamis [3] not only modeled the emergency department but also optimize the system. For inpatient facilities, van Berkel and Blake analysed waiting lists for surgery in Nova Scotia, Canada [12]. In addition to patient services, laboratory operations were also analyzed using discrete-event simulation models. Bodtker et al. [14] and Berchtold et al. [10] analyzed the application of simulation modeling for clinical laboratories' management. Marshall et al. [63] discussed the trend of modeling patient flow through different phases of care in a hospital. Shechter et al. [79] created a model for the patient flow of end-stage liver disease and organ allocation. In other healthcare problems, discrete-event simulation is also considered to design the vaccine allocation for the outbreak of influenza [1, 42].

2.3 Simulation Optimization Algorithms

There are three common challenges or characteristics encountered in optimizing extremely complex problems: lack of structure; presence of large uncertainties; and large search spaces [44]. These characteristics also result in difficult or costly performance evaluation. Stochastic global optimization becomes a powerful approach for these complex problems with estimations of the performance from realizations of a black-box simulations or direct observations, which is called simulation optimization. Simulation optimization provides a routine procedure that is computationally efficient [35].

There are five primary categories of simulation optimization, ranking and selection/ordinal optimization, stochastic approximation, response surface methodology, sample path optimization, and random search [35, 36, 37]. Fu [35] reviewed recent studies and the relationship with some commercial software. Most commercial software's core algorithms are random search algorithms. Ranking and selection and response surface methodology are also applied to many problems such as air traffic management [21], but they are not the popular choices for software packages. Ranking and selection and ordinal optimization are categorized together because they focus on a comparison theme rather than search algorithms [35]. Ranking and selection is a widely used approach based on multiple comparison procedures for candidate solutions, where optimal computing budget allocation (OCBA) is a series of ranking and selection algorithms which focus on efficiently using limited computing resources [22, 23, 24]. Two primary advantages for ordinal optimization are exponential convergence and goal softening [44, 45]. Stochastic approximation uses gradient search methodology as the core algorithm in convex optimization. The advantage of stochastic approximation is the statistical validity of proving convergence to the solution [38]. Response surface methodology samples systematically in order to approximate the relationship function for input variables and performance value, called metamodel. In some simulation optimization algorithms, a local surrogate response surface is generated to determine the searching strategy in each searching iteration [7]. Sample path optimization algorithms approximate the objective function deterministically with multiple sample paths [8], which is an approach to handle complicated constraints. Random search algorithms move from

the current solution to a neighborhood following a certain procedure. Most random search simulation optimization algorithms concentrate on discrete variables [6, 38], however, there are also continuous random search algorithms such as stochastic branch-and-bound [68, 92] and nested partitions [69, 80, 81].

The existing simulation optimization algorithms are mostly designed for single objective problems [38]. However, decision makers usually consider multiple objectives in the real world [51]. There are a number of multiple objective optimization studies [13], particularly evolutionary algorithms such as the genetic algorithm [51], but only limited studies used these algorithms to develop simulation optimization algorithms [55, 56, 57, 61, 85].

The probabilistic branch and bound algorithm (PBnB) was first developed by Prasetio [74] for both continuous and discrete optimization problems. It is a random search algorithm, and similar to a nested partitioning optimization framework, using partitioning and random sampling. In PBnB and the nested partition algorithm [80], the feasible region S is partitioned into M disjointed subregions. In the nested partition algorithm, the most promising subregion in iteration k, σ_k, is further sampled and partitioned in the next iteration until the stopping criteria is met. In order to calculate the promising subregion, each subregion is sampled by a certain random sampling scheme. In PBnB, Prasetio [74] introduced order statistics to determine the quality of each subregion for pruning and branching decisions. The algorithm dynamically determines the required number of samples and replications in each contending subregions. Eventually, the algorithm returns the incumbent solution and a small subregion with a probabilistic bound that ensures the probability of the subregion containing a near-optimal solution. The algorithm dynamically determines the required number of samples and replications in each current subregion. Unlike other partition-based algorithms, PBnB stops not only further partitioning but also further sampling to increase the efficiency while maintaining a probability bound on the chance of missing the desirable region.

Wang [88, 94] proposed a modified version of PBnB (MPBnB). MPBnB provides a probabilistic bound of containing a qualified solution which is simpler then the one provided by Prasetio [74]. MPBnB has a different focus from PBnB which is the goal softening. Specifically, MPBnB intends to identify the solutions in a certain level set such as the top

10 percent solutions. A main modification of MPBnB is the evaluation of pruning steps with the concentration of desired solutions in each contending subregions. The concentration of contending subregions is guaranteed to be improved as MPBnB progresses. As a result, the remaining subregions can approximate the target level-set, but MPBnB only guarantees that the remaining subregions contain a part of the target level set. Wang [88] also proposed a framework of further extension to PBnB with an empirical quantile estimation and a new pruning and maintaining strategy. This extended PBnB demonstrates the capability of approximating a full level set empirically and the potential of decision support with a level set [88]. However, it does not provide statistical analysis of the quality of solutions. In this dissertation, further extensions of PBnB and a statistical analysis are presented, not only for level set approximation, but also for multiple objectives.

Chapter 3

STAFFING RESOURCE ALLOCATION OF AN OCCUPATIONAL HEALTH CAMPAIGN

This chapter describes the first healthcare application; using discrete-event simulation to design an occupational health campaign and determine nurse staffing and laboratory testing schedules [43]. The occupational health campaign is a concurrent influenza immunization and tuberculosis (TB) screening campaign, which a hospital wanted to complete in a compressed time frame prior to the start of the next flu season. The concurrent campaign reduces health care workers (HCWs) clinic visits and immunizes them against influenza as soon as possible. We modeled the clinic and laboratory as a discrete-event simulation model and used the model to perform analysis with all possible system design combinations. The design and implementation procedure has been used through 2011 to 2013, where new data were collected during the campaign and used for refining the simulation model. Most of this chapter appeared in [43].

3.1 Introduction

Designing a campaign of concurrent influenza immunization and TB screening in a compressed time frame involves careful planning and coordination. The compressed time frame of less than 90 days is motivated by the desire to immunize health care workers (HCWs) as soon as possible, once the vaccines are available, before the start of flu-season. HCWs are also required to have an annual TB screening. The main reason to combine immunization and TB screening in a single campaign is to reduce the number of clinic visits for HCWs. An effective and efficient concurrent campaign requires detailed planning by Occupational Health Services (OHS) to provide the necessary immunization, blood draw and registration resources as well as coordination with the Laboratory Medicine Department (LMD) to assure the resources are available to provide timely tuberculosis screening results. The combined format of this campaign as well as the mixture of populations to be served (i.e.,

patients and HCWs) increased the work load and introduced process complexity for OHS and LMD. There were two primary goals for the performance of the campaign: 1) return the results of the TB screening to the HCW within 14 days; and 2) provide a good experience for the HCWs and patients at OHS (i.e., short waiting time). The OHS manager had to decide how many additional staff are needed to achieve the goal performance. The main concern of LMD was the arrival of large batches of specimens to be processed and tested in a timely manner. The design of the compressed campaign required decisions regarding campaign length, OHS staffing and LMD staffing. The discrete-event simulation model presented here considers the relationship between clinical procedures, laboratory services, and the detailed operations of each during the concurrent campaign. The resources and interactions of both departments are captured in the simulation model. The simulation model effectively enables planning for OHS and LMD to coordinate and execute an efficient campaign.

3.2 System Modeling and Alternative Scenarios

Seattle Children's Hospital (Seattle, Washington) is a pediatric care center and a teaching hospital associated with the University of Washington School of Medicine. At the time of the study, there were approximately 5,000 employees in the hospital system. The Department of Laboratories performed more than 600 different clinical laboratory tests and processed more than one million requisitions per year. Tuberculosis testing is performed in the chemistry laboratory, which was staffed daily from 8-5, but only one staff member on Saturdays and Sundays.

The Seattle Children's Hospital flu/TB campaign had three primary components: OHS (Occupational Health Services), a department located within the hospital; Roosevelt Clinic, an external facility located on another campus some distance from the hospital; and LMD (Laboratory Medicine Department), the clinical laboratory in the hospital. Two populations were served by OHS during the campaign: HCWs (paid employees, students, and volunteers) and patients (hospital in-patients, out-patients, and immediate family members).

We constructed a discrete-event simulation model to evaluate the performance of the campaign. The simulation model supported decision making in two phases: a planning phase

and an operational phase (Figure 3.1). In the planning phase, all alternative configurations are evaluated in the simulation to find the best design. During the operational phase, the simulation model was periodically refined as new information becomes available (direct observation and data collection). Our objective during the operational phase was to improve performance during the remaining campaign time by making adjustments to OHS and LMD (e.g., staffing, delivery schedules, and instrument cycles). At the end of the campaign, an improved simulation model was ready for subsequent planning phases (e.g., next year).

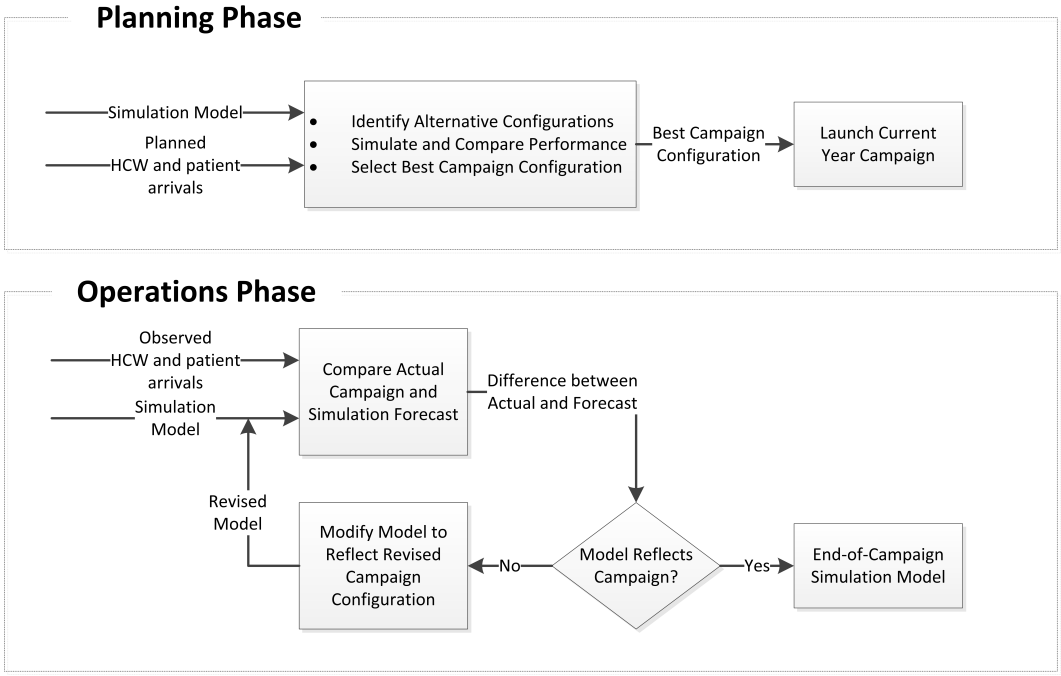

Figure 3.1: Process of decision making in the planning phase and operational phase with simulation models

The planning and operational decision process was used in 2011, 2012 and 2013. In the early autumn of 2011, the simulation model was used to compare alternative configurations, answer what-if questions, and select the best campaign design. Once the campaign started, the process enters the operational phase. During the operational phase, data such as the arrival patterns of HCWs and patients and input parameters were collected, the models were modified to incorporate the new information, and adjustments to the responses were made

periodically (every 2 to 4 weeks). At the end of the 2011 campaign, the simulation model was modified to incorporate lessons learned, such as greater variance in certain activities and emergent processes. The simulation model was used again to plan for the 2012 and 2013 campaign and updated during the operation of 2012 and 2013.

Figure 3.2 illustrates the campaign system with the three primary components. HCWs may go to OHS or Roosevelt Clinic to have their blood drawn for the TB screening and receive their influenza immunization. Patients must go to OHS for their influenza immunization. Some HCWs may decline the influenza immunization if they complete certain documents, and some HCWs may have reason not to provide a blood specimen for testing.

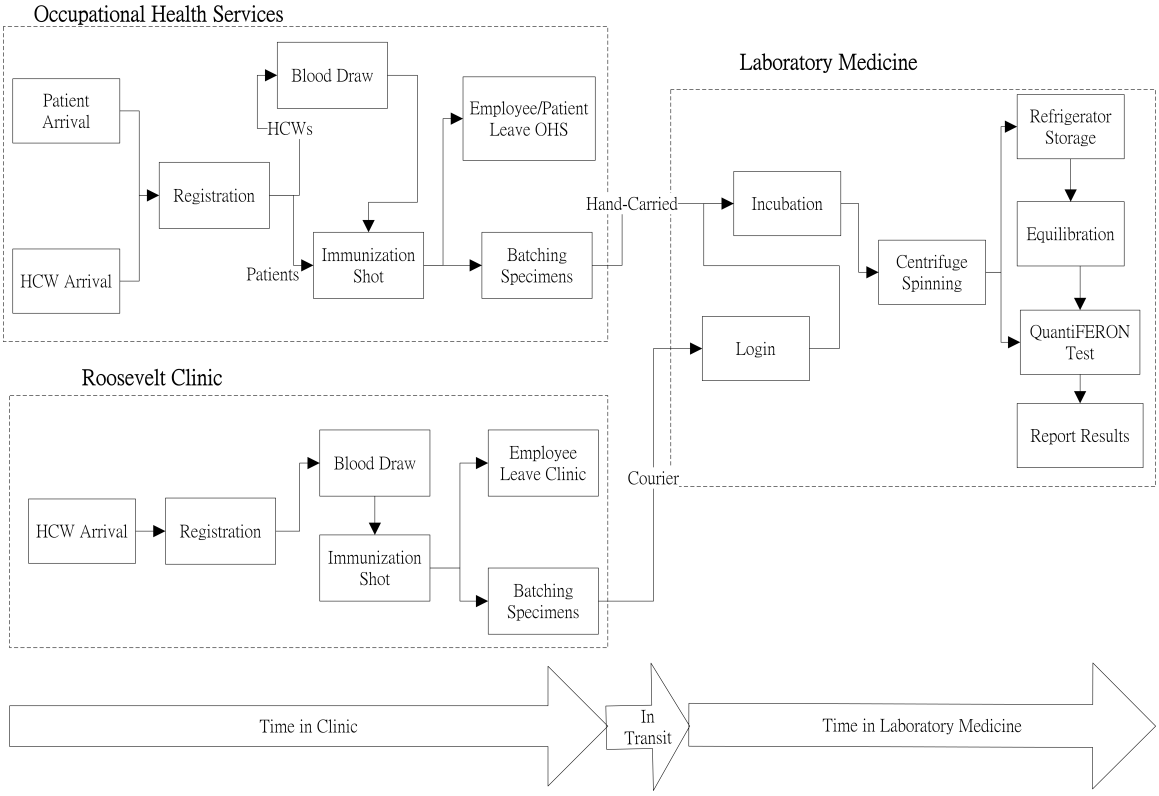

Figure 3.2: Overview of the employee healthcare services system

The TB screening process at LMD is illustrated on the right side of Figure 3.2. Specifically, the preparation processes for all specimens are incubation and centrifuge. However, the specimens arriving from Roosevelt Clinic were logged into the LMD information sys-

tem, by Central Processing Area (CPA) lab technicians. After those processes were completed, specimens were loaded and analyzed for a QuantiFERON test by the Dynex DS2 ELISA workstation (DS2). The lab processes specimens according to a strict first-in-first-out (FIFO) policy. If the DS2 is not available (i.e., currently in use), and there are enough specimens already in queue to fill the next testing batch, then the next batch of specimens will be placed in the refrigerator (in FIFO sequence) after the centrifuge spinning step in Figure 3.2. When specimens are taken from the refrigerator they must reach room temperature (equilibrate) before they can be placed in the DS2 for testing.

A QuantiFERON test at DS2 included sampling and testing. The DS2 first takes samples from the blood specimens on a plate, at most, 29 employee specimens, and automatically begins the TB testing process. Two plates can be processed at the same time, but the sampling process must be staggered, that is illustrated in Figure 3.3. The following discussion focuses on 2011 campaign to illustrate the procedure.

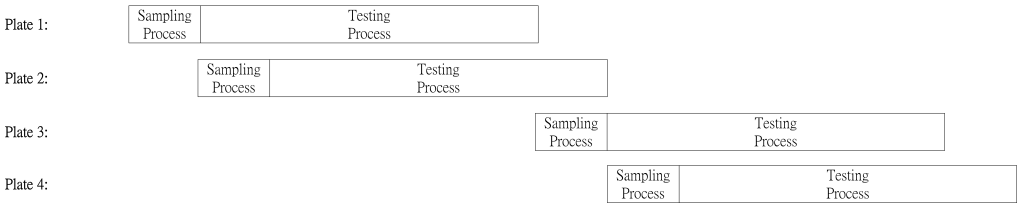

Figure 3.3: Sampling and testing processes in the QuantiFeron instrument

The three primary decisions for the compressed campaign in 2011 were: campaign length, OHS staffing, and LMD staffing, led to 240 campaign configurations in 2011. The campaign length has six scenarios, illustrated in Figure 3.4. The campaign length affects the arrival pattern of HCWs to OHS. In additional to considering one, two and three month campaigns, a distribution were made between a historical pattern of arrivals and a prescribed level-loaded pattern (obtained by appointments).

The number of contract nurses and phlebotomists assigned to the campaign was a critical OHS staffing decision. Health care workers required both venipuncture for TB screening and injections for flu immunization while patients required only immunization. Phlebotomists may draw blood but cannot give injections while nurses can do both. Contract nurses are

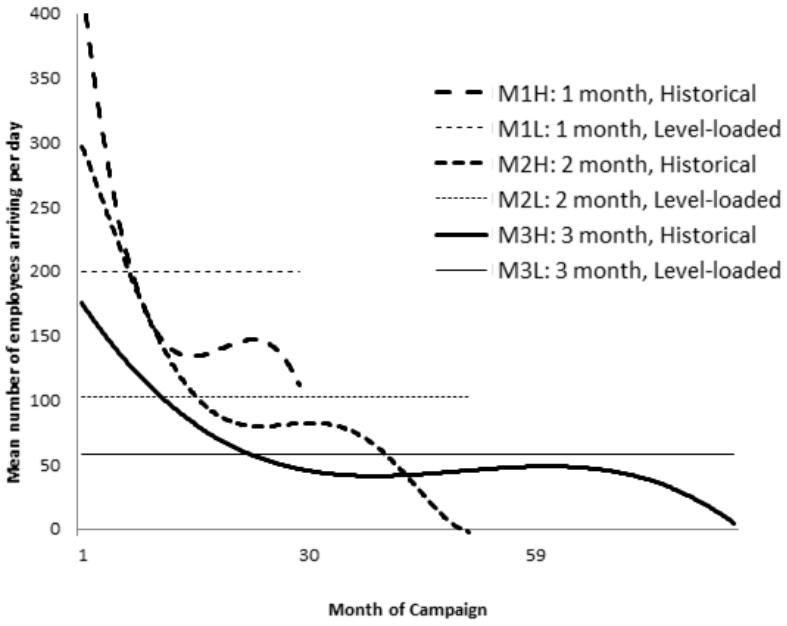

Figure 3.4: Comparison of arrival patterns of HCWs to OHS in 2011

more expensive than phlebotomists, so the number and mix of nurses and phlebotomists is an important resource planning decision. The OHS staffing scenarios are shown in Table 3.1, and labeled as S1 to S4. There are five stations at OHS, so the total number of nurses and phlebotomists combined was five. In additional to deciding the number of nurses and phlebotomists, OHS also had to decide between two scheduling scenarios. As shown in Table 3.1, scenario O1 provided 12 hours per day on Monday, Tuesday and Wednesday and 8 hours on Thursday and Friday; scenario O2 simply assigned 12 hours to each of the five weekdays. No weekend hours were scheduled in 2011. Combining staffing and scheduling decisions, the OHS concluded 8 alternatives.

An important design decision at LMD is the number of dedicated hours of the lab technologists to process TB screening tasks. In 2011, we used number of dedicated hours of LMD staff to create five alternative scenarios (See Table 3.2, Q1 to Q5). All together, the number of design configurations totaled 240; six scenarios associated with campaign length,

eight OHS staffing and scheduling scenarios and five LMD scheduling scenarios.

Table 3.1: OHS staffing and scheduling scenarios for 2011 planning phase

OHS Staff Levels	Week 1 through 13	
Scenario	Nurses	Phlebotomists
S1	1	4
S2	2	3
S3	3	2
S4	4	1
OHS Service Times	Week 1 through 13	
Scenario		
O1	12 hour day on Mon, Tue, Wed, 8 hour day on Thur, Fri(12/3+8/2)	
O2	12 hour a day, 5 days a week (12/5)	

Two criteria drove the selection of design configurations: the time to report TB screening results for each HCW with a goal of less than 14 days, and the wait time for services at the OHS clinic. Other considerations of the campaign length, such as timely influenza immunization, and flexibility for HCW schedules, were outside the scope of the simulation model.

As we monitored the progress during the operational phase of the campaign, we periodically revised the simulation model to reflect changes in arrival patterns as well as revised laboratory processing schedules and staff availability. After each update, the simulation model was rerun to discover if further changes in staffing or scheduling was needed to achieve performance goals. Five updates were performed in 2011 and four in 2012. At the end of the campaign, the refined arrival patterns and simulation model helps inform the planning process for the next year's campaign.

Table 3.2: LMD scheduling scenarios for 2011 planning phase

LMD Dedicated Hour	Week 1 through 13
Scenario	
O1	8 hour a day, 5 days a week (8/5)
O2	12 hour a day, 5 days a week (12/5)
O3	13 hour a day, 5 days a week (13/5)
O4	14 hour a day, 5 days a week (14/5)
Q5	12 hour a day, 5 days a week and 8 hours on Sat (12/5+8/1)

3.3 Results

The analysis during the planning phase evaluated the system performance with design decisions for: campaign length, OHS staffing, and LMD staffing. The goal of providing TB screening results within 14 days was achieved in 2011, and reduced to four days in 2012 and 2013. The wait time for HCWs in the clinic was considered satisfactory, even though the number of arrivals doubled from 2010. Also the campaign was compressed from 100 days in 2010, to 90 days in 2011, and 75 days in 2012 and 2013. The operational phase was useful in refining the resource allocation as the situation was observed, and the updated data was used to set up the model for the following year. The detailed analysis presented in this section focuses on 2011 as an example to demonstrate the decision procedure.

3.3.1 Determine Appropriate Campaign Length and OHS Staffing When Considering Time HCWs Spend in the OHS Clinic

In 2011, we considered one, two, and three months (historical and level-loaded) as alternatives for campaign length, and eight OHS staffing scenarios as defined in Table 3.1. The simulation model indicated that compressing the campaign to one month results in extremely long waiting times for HCWs at OHS, as might be expected. Therefore, a one month campaign was ruled out as a viable alternative. And the scenario in which only one

nurse is hired (S1) is eliminated because the simulation model showed that the other staffing scenarios performed better.

Although the level-loaded scenarios perform well with regard to HCW waiting time, we decided to eliminate level-loaded scenarios due to the following concerns. From the viewpoint of immunization, the earlier the immunizations are performed, the lower the risk to HCWs and hospital patients they encounter. Also, HCWs prefer flexibility on choosing their time of visit instead of adhering to a scheduled or assigned appointment time.

In Table 3.3, the performance of the remaining configurations are listed. The historical two month (M2H) scenarios performed much worse than the historical three month scenarios (M3H). Also, the scenarios with a mixed schedule (S2O1, S3O1, S4O1) result in two to four minutes longer waiting times than 12 hour days (S2O2, S3O2, S4O2). This information allowed the OHS manager to decide to keep the OHS open for 12 hours a day, five days a week (Monday to Friday) in 2011. The differences among S2O2, S3O2 and S4O2 configurations for the historical three month scenario are relatively small. Because of earlier decisions in 2011, OHS manager preferred to implement the scenario of 4 nurses, 1 phlebotomist and 12 hour days (S4O2) for the entire three month campaign.

Table 3.3: Time in OHS in 2011: details of 2 and 3 months historical arrival patterns

(minutes)	S2O1	S2O2	S3O1	S3O2	S4O1	S4O2
M2H	460.30	411.98	254.86	251.24	258.39	248.67
M3H	26.01	22.16	23.39	20.91	22.89	20.11

3.3.2 Refining the Simulation Planning Model during the Operational Phase of the Campaign

Using an approach similar to the evaluation of the OHS clinic, an analysis of clinical laboratory participation in the 2011, 2012, and 2013 compressed campaigns was undertaken. In 2011, as a consequence of the OHS analysis, except the eliminated one month campaign length and level-loaded arrivals, an historical arrival pattern with a two month or three

month campaign duration were analyzed for 2011. When HCW arrival follows an historical pattern, scenarios Q2, Q3, Q4, and Q5 perform better than Q1, as shown in Table 3.4. Even the two month campaign with Q1, the time to report results is averagely 12.32 days, which achieved the goal performance 14 days. However, recall that the waiting time at OHS were significantly longer than three month.

In the end, the lab manager concurred that the 2011 campaign length would be three months. Although the Q1 staffing scenario (8 hour days) does not perform as well as others, it still achieves the 14 day goal. The lab manager chose Q1 as the lab staffing scenario, so the current staffing schedule would not be disrupted.

Table 3.4: Details of time in LMD until results reported in 2011

(days)	Q1	Q2	Q3	Q4	Q5
M2H	12.32	3.74	2.38	2.25	2.21
M3H	4.17	2.22	2.18	2.14	1.81

3.3.3 Refining the Simulation Planning Model during the Operational Phase of the Campaign

After the campaign was underway, we periodically compared the actual parameters and conduct of the laboratory operations to the operating policies to have been employed. The objective was to incorporate new information and data in the simulation model and support operational planning for the time remaining in the campaign. Further refinements to the simulation model would also improve the ability of the model to predict performance of future campaigns with alternative resource configurations and levels of demand. The refining in 2011 will be described to illustrate the procedure.

Several changes were made to the campaign and were subsequently included in the simulation model. The major change was an increase in the scale of the campaign: approximately 1,000 hospital volunteer workers would require flu immunization and TB screening. Also, the number of times that the specimens were transported from OHS to LMD was increased

from 2 times per day to 3 times a day in an effort to level-load the flow of specimens into the lab.

We had initially assumed that a full-time lab technologist would be assigned to the specimen processing centrifuge operation. During the campaign, however, the clinical labs had to change the centrifuge resource allocation due to staff shortages. Using the simulation model to examine staffing alternatives, we found that if the centrifuge was staffed for TB screening the first 2 hours of the day, the TB results could still be reported within the 14-day goal. Another unexpected change on lab operations was the addition of a login process which the labs had to complete on receipt of specimens from the OHS clinic.

The QuantiFERON test schedule (the number of plates run per day) changed several times throughout the campaign even with the determined dedicated hour scenarios, because the lab technologists often needed to address routine lab tasks and could not follow a precise plate schedule. The actual average of 3.26 plates per day, rather than the maximum of 5 plates presented in scenario Q1, changed the time to report TB screening results. The consequences for predicting system performance were significant enough that the simulation model was revised to incorporate a planned number-of-plates-per-day parameter, which would better reflect the number of plates run.

Data reflecting the actual arrival pattern of the employees to OHS was collected throughout the campaign and the input model was refined accordingly. Although not anticipated, initial data revealed that a number of employees had declined the TB blood draw; therefore, the potential for declination was added to the model during the campaign. The actual pattern of HCWs arrival to OHS is shown in Figure 3.5.

Table 3.5 summarizes the results of the operational phase at different periods of time by comparing observations from the actual system with simulation model forecasts using two key metrics: the number of pending specimens waiting to be processed by the lab and the number of processed specimens.

At the end of the 2011 campaign, to validate our refined simulation model, we ran the final modified version of the model and compared the results of the actual system to model forecasts using the actual arrival pattern of employees and patients. Table 3.6 summarizes these results using the two metrics, number of pending specimens and number of processed

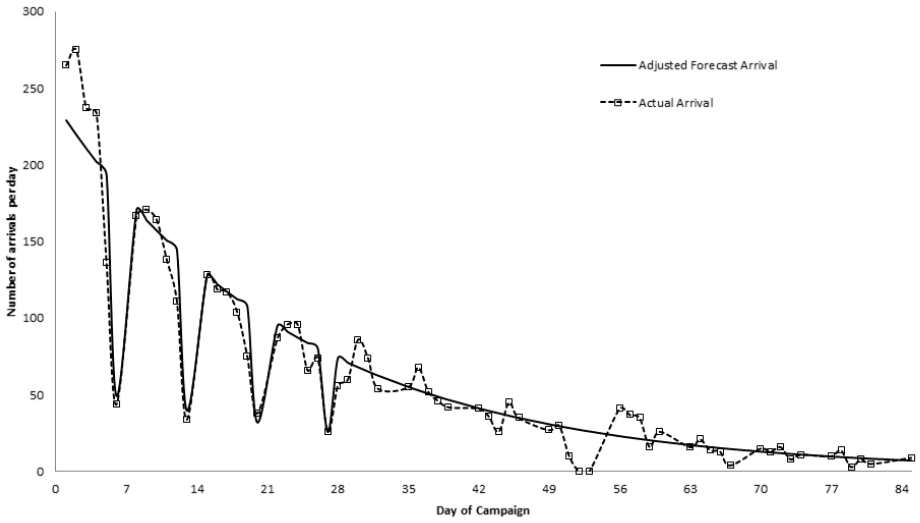

Figure 3.5: Comparison of forecasted HCWs arrivals and actual HCWs arrivals in 2011

Table 3.5: Results of the operational phase

Pending specimens:	10/12/2011	10/27/2011	11/08/2011	11/28/2011	12/28/2011
Actual	1092	900	600	101	0
Model estimated	267	812	623	177	0
Processed specimens:	10/12/2011	10/27/2011	11/08/2011	11/28/2011	12/28/2011
Actual	899	1830	2953	3650	4019
Model estimated	1124	1793	2561	3765	4016

specimens.

As shown in the results, the final simulation model, based on the selected metrics, provides a good approximation of the actual systems performance. The average time to report TB screening results for the campaign was 10.11 days; the model accurately estimated the turn-around-time as 10.42 days. Furthermore, the arrival data for 2011 (see Figure 3.5) were used to create the arrival forecast model for a compressed 2.5 month period without

Table 3.6: Final model estimation results

Pending specimens:	10/12/2011	10/27/2011	11/08/2011	11/28/2011	12/28/2011
Actual	1092	900	600	101	0
Model estimated	785	972	762	13	0
Processed specimens:	10/12/2011	10/27/2011	11/08/2011	11/28/2011	12/28/2011
Actual	899	1830	2953	3650	4019
Model estimated	871	1838	2529	3731	4016

increasing the starting peak (a consequence of OHS capacity limitations).

3.4 Discussion

With the implementation of the simulation model, the design of the campaign can adapt to varying circumstances for each year, and further improve the performance using experience from previous years. Specifically, the decisions made in the planning phase and the following adjustments in 2011 and 2012 were able to achieve the goal of reporting the TB screening results within 14 days, even though the HCW arrivals increased from 2,600 in 2010 to 5,000 in 2011. Furthermore, based on their experience in 2011, the LMD was able to manage the schedule of plates to reduce the report time of TB screening to less than 5 days in 2012 and 2013. Also, the length of the campaign was compressed from 100 days to 75 days to provide the immunization to all HCWs earlier. The hospital used a mandatory policy that required every HCW to receive immunization and TB screening during the campaign, therefore the percentage of HCWs receiving these services did not change with the campaign implementation. The compressed campaign did ensure that HCWs received immunization before the outset of the flu season.

In the planning phase of the campaign, the goal is to find the best configuration that achieves the performance goals. A three month campaign was implemented in 2011. In 2012 and 2013, the campaign length was reduced to 75 days. A video camera system was installed to encourage level loading of demand during the OHS open time. The camera system web page displayed the waiting queue outside the OHS clinic and an expected waiting time.

Health care workers used the queue length and waiting time information to decide when they would visit the clinic. The camera system was primarily intended to reduce waiting time in the OHS queue, but it may have distributed arrivals to approximate level-loaded demand throughout the day.

The resource allocation in 2012 and 2013 was also a factor in reducing campaign length and providing good performance. In 2011, the OHS clinic staff was one phlebotomist and four nurses, but in 2012 and 2013, five nurses were scheduled. The analysis suggested a more efficient plate running scenario at LMD for 2012 and 2013. Specifically, more flexibility in running plates and integrating them into routine work resulted in better TB screening result time.

The operational phase of the campaign used the simulation model to understand the impact of in-process adjustments to the scheduling, staffing and service times of the immunization and TB screening system; these changes would also improve the performance of the model when used in subsequent campaigns. Although most of information and decisions needed for a new campaign are assembled during the planning phase, changes in demand, availability of resources and inaccurate data can significantly impact a complex and manually managed system. In this situation, discrete-event simulation models were used to evaluate the performance of a complex system based upon an extensive set of initial input parameters and operating policies. When the parameter values no longer reflected the operating characteristics of the campaign being modeled, the parameter values, and sometime model structure/logic, were revised to realign the model so that it continued to adequately represent the campaign immunization and screening system. Following those revisions, the model provided a new forecast of system performance which was used to adjust resource levels as well as mitigate laboratory technician anxiety concerning the number of pending specimens for TB screening.

Table 3.7 summarizes the campaign performance from 2010 (before simulation modeling) through 2013 (with simulation modeling). The number of HCW arrivals nearly doubled from 2010, but the objectives of reducing campaign length and achieving a 14 day TB screening results were accomplished.

The simulation model presented in this section was developed to support resource plan-

Table 3.7: Final model estimation results

Year	Approximate Number of HCW Arrivals	Campaign Length (days)	Average Time of TB Screening (days)
2010	2600	100	3.90
2011	5000	90	10.11
2012	5000	75	4.43
2013	6000	75	3.60

ning for a compressed medical campaign conducting concurrent flu immunization and tuberculosis screening that requires coordination between OHS and LMD. Coordination between departments (in this case, OHS and LMD) is an issue in many situations, however there are few tools available to assist with coordinated resource planning. Some studies involving coordination focus on introducing computer systems to facilitate sharing of information [11, 83]. In [65], screening methods were evaluated to coordinate admission and surgery, and in [18], and physicians and pharmacists collaborate to improve blood pressure control. However, these methodologies are not extendable to resource planning, whereas our simulation methodology may be adapted to assist other departments in coordinated planning and operations.

One of the benefits of using a simulation model is the ability to support early planning decisions and evaluate trade-offs between different configurations of number of nurses, phlebotomists and lab-technicians that would be needed to meet campaign performance requirements before the campaign was launched. Another benefit is the information provided by the simulation model during operations to allow reconfiguration of the campaign when changes occur. A drawback is the need to invest time in creating an accurate discrete-event simulation model with valid data to support the model. Once the simulation model is developed, it is maintainable at low cost, and can update data collected during operations to prepare the simulation model for subsequent use. For the compressed concurrent campaign, the benefits of the simulation model overshadowed the cost of initial development, as it has been used over three years. The model has helped the decision-makers in OHS and

LMD evaluate the need for mid-campaign configuration adjustments and increased their confidence in completing the campaign on time with the resources budgeted.

Chapter 4

PROBABILISTIC BRANCH AND BOUND FOR SET APPROXIMATION

Chapter 4 contains the primary methodological results of the dissertation, two variations of a partition-based simulation optimization algorithm, Probabilistic Branch and Bound (PBnB) [74, 88]. Using the characteristic of partition-based algorithm, PBnB can provide a set of solutions by the partitioned solution space. Taking a set of solutions as an advantage, the first variation focuses on approximating a level set of solutions instead of the global optimal [46]. The second variation considers problems with multiple objectives and develops an algorithm for the Pareto optimal set approximation [47]. For both variations, we derive the statistical quality of solutions by probability bounds and perform numerical analysis with benchmark test functions. The PBnB for level set approximation is described in Section 4.1 and the multiple objective version is illustrated in Section 4.2. Most of the material in Section 4.1 appears in [48].

4.1 Probabilistic Branch and Bound for Level Set Approximation

4.1.1 Introduction

Real world problems are often extremely complicated and involve uncertainty, making decision-making a challenge. For these complex problems, a closed form performance function may not exist. Simulation is considered to be a useful approach to estimate the performance of complex systems. In order to optimize a complex system, many approaches using simulation optimization have been developed [35, 36, 38]. Most of the simulation optimization algorithms focus on approximating a local or global optimal solution. Instead of just providing a single solution, our approach is to provide a set of solutions that achieves a target quantile (e.g., best 10%), allowing decision makers to make trade-offs between simulated performance and other issues. Others have also looked at methods to relax the objective of

finding a single optimal solution. Ordinal optimization [44, 45] introduced the concept of "goal softening" which finds at least one solution out of the best m solutions. Also, instead of finding the best design, indifference-zone procedures focused on getting a design with a "good" performance [53, 67, 77, 81].

Our approach, called Probabilistic Branch and Bound (PBnB), operates on subsets and is applicable with continuous and/or integer variables. The algorithm iteratively updates its confidence interval on the value of the target quantile, and seeks the target level set, that is, the set of solutions within the target quantile (e.g., all solutions in the best 10%). As the algorithm proceeds, each subregion is statistically identified as: 1) contained within the target level set; 2) no intersection with the target level set; 3) there is not sufficient evidence to decide. Probability bounds are derived on the maximum volume of incorrectly determined regions at each iteration, providing useful information to the user on the quality of solutions provided. The algorithm and analysis in this paper builds upon previous results in [88] and [46].

The benefit of estimating a level set in simulation optimization in contrast to a single solution is that decision makers can consider information that is not evaluated in the simulation model. For instance, flexible threshold design problems [73] can use a level set solution to ensure that a threshold tolerance is met. When there is noise in the performance measure (this is typically the case in discrete-event simulation), decision makers may be indifferent to small differences in the estimate of the objective function. By providing a set of solutions, PBnB enables decision makers to explore other aspects of the solutions that are in the target level set. Csendes and Pintér [27] considered a level set as an approach to sensitivity analysis.

A partition based approach is used here for approximating a level set since it can provide a collection of subregions forming the approximation. Partition based algorithms have been used for finding global optimal solutions. The nested partition algorithm [69, 80, 81] partitions solution sets into multiple subregions and finds a promising subregion with a high likelihood of containing the global optimal solution. In contrast to nested partition, the algorithm proposed in this dissertation does not narrow in on the global optimum, but creates a set of subregions that are considered "maintained" because there is statistical

confidence that the maintained region is contained in the target level set at every iteration.

In addition, our algorithm considers subregions when determining whether implementing further samples is computationally warranted. A similar concept is used in ranking and selection algorithms and Optimal Computing Budget Allocation (OCBA), where a fixed number of simulation runs is given, and the algorithm evaluates the objective function at different solutions [21]. Xu et al. [91] also group the solutions and perform ranking and selection procedures for the sample size. Xu and Nelson [92] developed a partition based algorithm, named empirical stochastic branch and bound (ESBnB), with empirical bounds for each subregion on future sampled performance.

Simulation optimization algorithms often provide asymptotic convergence results to a local or global solution, and maximize the probability of correct selection or minimize a computation budget. Also, stochastic approximation, nested partitioning and model-based algorithms ensure convergence asymptotically [37, 69]. However, in practice, decision makers may want to run the algorithm for awhile, and decide to continue or terminate. In contrast to asymptotic results, we derive bounds on the performance of the algorithm at any iteration. Finite time results help the decision makers interpret the quality of the reported solutions, and decide whether to continue the algorithm or terminate.

The remainder of the chapter is organized as follows. An overview of the algorithm for level set approximation is introduced in Section 4.1.2, and Section 4.1.3 covers the details of the algorithm. Section 4.1.4 presents several theorems that describe the quality of the level set approximation provided by the algorithm with probability bounds. Numerical results with test functions are presented in Section 4.1.5.

4.1.2 Framework of PBnB for Level Set Approximation

The proposed algorithm aims to approximate a level set with respect to a performance function $f(x)$ for a black-box simulation model, where the optimization problem is

$$(\mathcal{P}) \min_{x \in S} f(x) \tag{4.1}$$

where $f(x) = E_{\Xi}[g(x, \xi_x)]$ is the expected value of a black-box simulation or a noisy function $g(x, \xi_x)$, and ξ_x is the random variable representing the noise. Typically, ξ_x is assumed to be normally distributed since the sample mean is often used to estimate $f(x)$, and the Central Limit Theorem can be applied ([21]). The design variable x is a vector in n dimensions, and the values may be integer or real-valued. Typically, the feasible set S is defined by upper and lower bounds on x, however any bounded set is possible as long as it is easily partitioned.

We are interested in the δ quantile associated with $f(x)$, denoted $y(\delta, S)$, which is defined as

$$y(\delta, S) = \underset{y \in \{f(x): x \in S\}}{\arg\min} \{P(f(x) \leq y | x \in S) \geq \delta\}, \text{ for } 0 < \delta < 1, \tag{4.2}$$

and we let $L(\delta, S)$ be our desired set of best δ-quantile solutions, where

$$L(\delta, S) = \{x \in S : f(x) \leq y(\delta, S)\}, \text{ for } 0 < \delta < 1. \tag{4.3}$$

We recognize that $y(\delta, S)$ can also be expressed in terms of probability as

$$P(f(x) \leq y(\delta, S)) = \frac{v(\{x : f(x) \leq y(\delta, S)\})}{v(S)} \geq \delta \tag{4.4}$$

$$P(f(x) < y(\delta, S)) = \frac{v(\{x : f(x) < y(\delta, S)\})}{v(S)} \leq \delta \tag{4.5}$$

when x is uniformly distributed in S and $v(S)$ denotes the n-dimensional volume of set S.

The PBnB algorithm for level set approximation has two primary components, as illustrated in Figure 4.1. The first component has three steps to develop a confidence interval for the target quantile $y(\delta, S)$. In Step 1, the algorithm samples over all current subregions. The number of sample points and replications depends on the desired confidence level and the size of the current subregion. In order to rank the samples correctly under noise, we apply a two-stage procedure based on Bechhofer et al. [9] to update the replication number, in Step 2. We estimate the quantile with confidence intervals in Step 3, using the ordered sample function values.

The second component focuses on processing subregions of the solution space, by maintaining, pruning, or branching. Based on the confidence interval, Step 4 finds the elite and

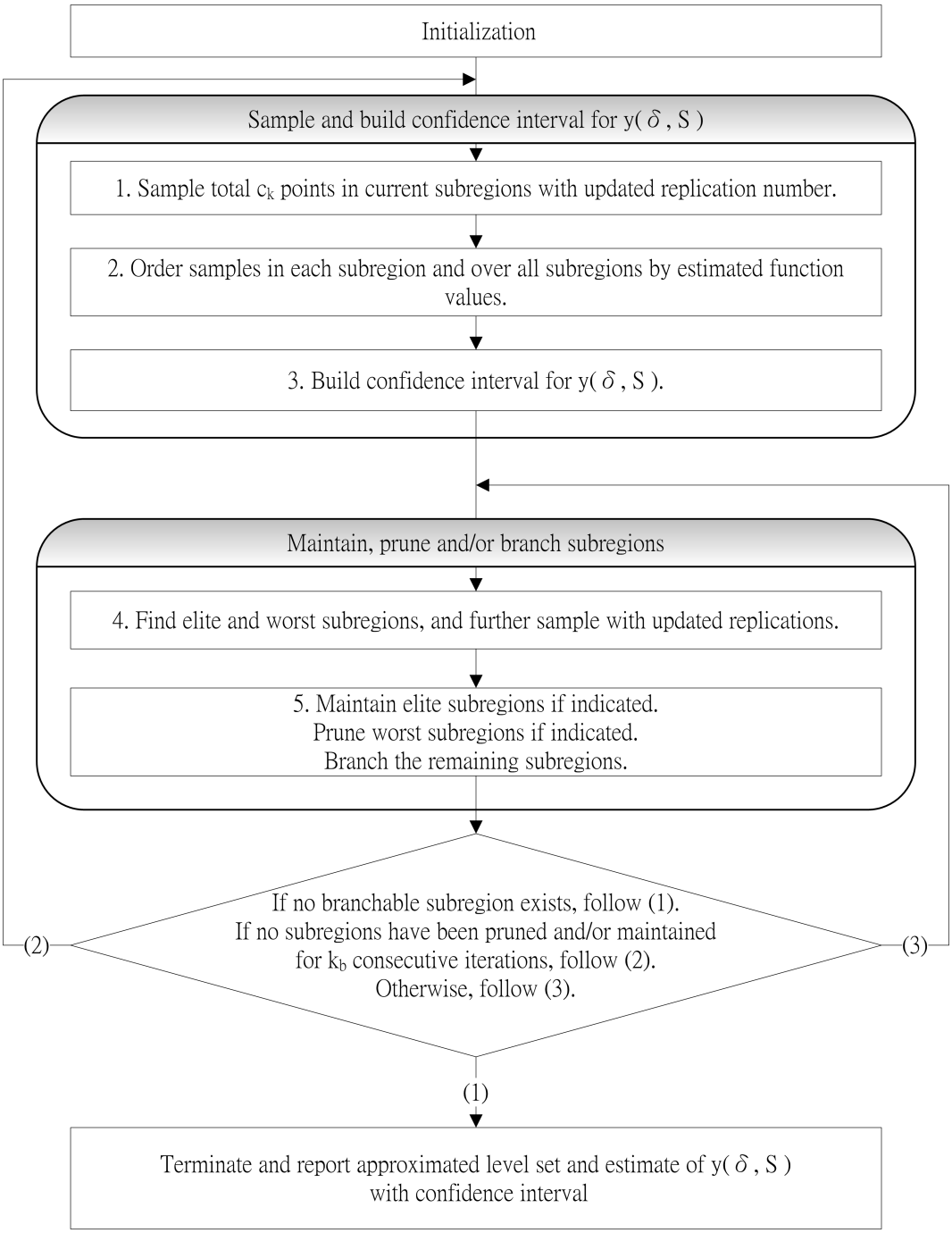

Figure 4.1: Procedure of PBnB for level set approximation.

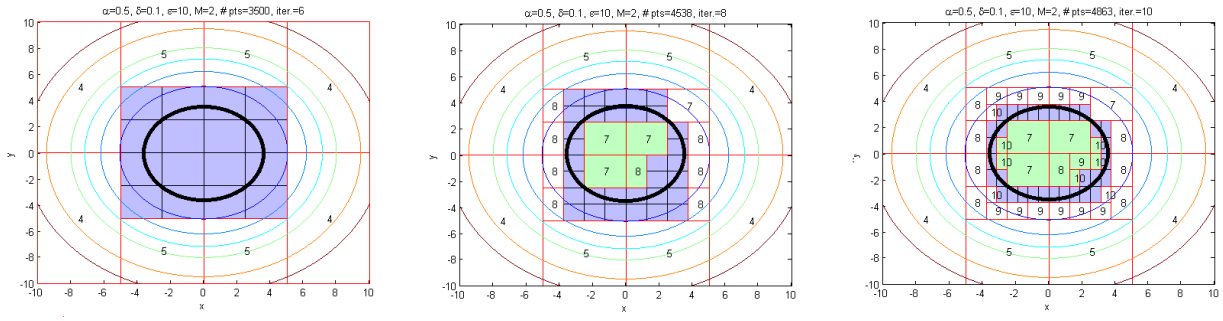

Figure 4.2: PBnB for level set approximation on a two dimensional sphere function, at iterations 6, 8 and 10.

worst subregions which have a potential to be maintained or pruned, and implements further sampling in the elite and worst subregions to statistically confirm the maintaining and pruning. In Step 5, the algorithm updates the maintained subregions, the pruned subregions and branches the rest of the subregions to create the current set of subregions.

PBnB for level set approximation has a straight forward stopping condition, route (1) in Figure 4.1. It proceeds until all subregions are either maintained, pruned, or reach a user-defined size. The size of an "unbranchable" hyper-rectangle impacts the precision of the level set approximation as well as the computation. In high dimensional problems, it could be difficult to achieve a tight approximation, and the user could terminate the algorithm by other methods, such as the volume of maintained subregions and/or a number of iterations.

If the algorithm does not terminate, it decides whether to update the quantile estimation, route (2) in Figure 4.1, or to continue branching and sampling with the current quantile estimation, route (3) in Figure 4.1, based on the number of consecutive iterations with unsuccessful maintaining and pruning. This balances the computation between reducing the number of current subregions (by maintaining or pruning), and narrowing the confidence interval of the quantile estimation for future maintaining and pruning.

Consider a two-dimensional sphere function as an example. The function details are described in (4.50) within Section 5. Figure 4.2 illustrates the progress of the algorithm with maintained, pruned, and branched subregions, overlaid on the contours of the sphere objective function, at iterations 6, 8, and 10. The target quantile is set to 10%, and the

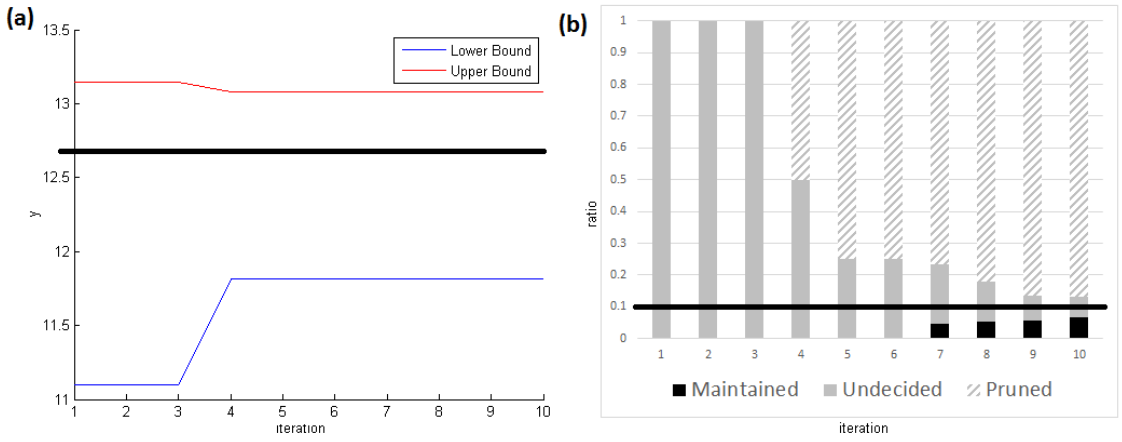

Figure 4.3: (a) Quantile interval estimation (upper and lower bounds) on $y(\delta, S)$ for iterations $1, 2, \ldots, 10$. The bold line in (a) indicates the true value of $y(\delta, S)$. (b) The ratios of volumes of the maintained, pruned and undecided subregions to the volume of S. The bold line in (b) indicates the target 10% on the two dimensional sphere function.

associated contour is bold in Figure 4.2. The white rectangles indicate the pruned subregions, and the numbers in the rectangles indicate the iteration at which they were pruned. The light gray (green) rectangles represent the maintained subregions that are contained in the target level set, and the dark gray (blue) rectangles are the current undecided subregions. As the algorithm progresses, the undecided subregions form the boundary of the level set. In the example, at the sixth iteration, PBnB already pruned the outer part of the solution space. By iteration 8, four subregions in the center of the level set are maintained, and at iteration 10, several smaller subregions are maintained, and the current undecided subregions trace the level set boundary. Figure 4.3a shows the shrinking quantile interval estimation (upper and lower bounds on $y(\delta, S)$), and the bold line indicates the true quantile value $y(\delta, S)$ of the sphere function. Figure 4.3b shows the fractional volume of the maintained, pruned, and undecided subregions, for iterations 1 to 10 of the algorithm on the sphere function. Notice the volume of the undecided region decreases while the volume of the maintained and pruned regions increases with iteration.

4.1.3 Algorithm Details of PBnB for Level Set Approximation

The input parameters defined by the user include: δ, α, ϵ, k_b, B, c, and R^0. The parameter δ is used to define a δ-quantile for the target level set. For example, the user may be interested in the set of solutions in the best 10%, in which case $\delta = 0.1$. The user does not have to specify a range on the objective function. If a specific value is known a priori, the algorithm can be modified to accommodate the information.

The following two parameters, α and ϵ, are used to determine the quality of the level set approximation. The parameter α is used in the confidence level of the estimation of $y(\delta, S)$ and in the probabilities of incorrectly pruning or maintaining. The approximation can be wrong in two ways, it could prune a portion of the level set or it could maintain some area that is not in the level set. The parameter ϵ is the volume of solutions that can be tolerated to be categorized incorrectly. The main results of the analysis are stated in Corollaries 9, 10 and 11, giving probability bounds on the quantile estimation, on incorrectly pruning a volume of size ϵ, and on incorrectly maintaining a volume of size ϵ, respectively.

The parameter k_b is the maximum number of consecutive iterations without maintaining or pruning before further overall sampling for quantile estimation. The branching scheme is defined by B, which is the number of evenly-sized subregions to create when a subregion is branched, c is the incremental sample size for each estimation of quantile when route (2) in Figure 4.1 is taken, and R^o is the initial replication number at iteration 1.

Probabilistic Branch and Bound (PBnB) for Level Set Approximation

Step 0. Initialization: Input user-defined parameters, δ, α, ϵ, k_b, B, c, and R^0. Also, initialize the maintain, prune, and current subregion collections and iterative counters as $\Sigma_1 = \{S\}$, $\widetilde{\Sigma}_1^C = S$, $\widetilde{\Sigma}_1^M = \phi$, $\widetilde{\Sigma}_1^P = \phi$, $\delta_1 = \delta$, $\alpha_1 = \frac{\alpha}{B}$, $\epsilon_1 = \frac{\epsilon}{B}$, $R_0 = R^o$, and $k = 1, k_c = k_b, c_1 = c$.

Step 1. Sample total c_k points in current subregions with updated replication number: For current undecided subregion $\widetilde{\Sigma}_k^C$, uniformly sample additional points such that the total number of points in $\widetilde{\Sigma}_k^C$ is c_k. For each subregion $\sigma_i \in \Sigma_k$, denote the sample points as $x_{i,j} \in \sigma_i$, for $j = 1, \ldots, N_k^i$ and $i = 1, \ldots, ||\Sigma_k||$. Note that

$\sum_{i=1}^{||\Sigma_k||} N_k^i = c_k$. For notational convenience, let $N_k = c_k$. If $f(x)$ is noisy, PBnB evaluates $g(x,\xi)$ with R_{k-1} replications to estimate mean and variance of each sample. Specifically, for each $x_{i,j} \in \sigma_i$, $j = 1,\ldots,N_k^i$ and $i = 1,\ldots,||\Sigma_k||$, perform R_{k-1} replications of $g(x,\xi_x^r)$, and evaluate the sample mean and sample variance,

$$\hat{f}(x_{i,j}) = \frac{\sum_{r=1}^{R_{k-1}} g(x_{i,j},\xi_x^r)}{R_{k-1}} \text{ and } S_{\hat{f}}^2(x_{i,j}) = \frac{1}{(R_{k-1}-1)} \sum_{r=1}^{R_{k-1}} (g(x_{i,j},\xi_x^r) - \hat{f}(x_{i,j}))^2. \tag{4.6}$$

Step 2. Order samples in each subregion and over all subregions by estimated function values: For each subregion i, $i = 1,\ldots,||\Sigma_k||$, order the sampled points, $x_{i,(1)},\ldots,x_{i,(N_k^i)}$, by their estimated function value so that

$$\hat{f}(x_{i,(1)}) \le \hat{f}(x_{i,(2)}) \le \cdots \le \hat{f}(x_{i,(N_k^i)}).$$

Similarly, order all sampled points, $z_{(1)},\ldots,z_{(N_k)}$, in all current subregions in Σ_k by their function values, so that

$$\hat{f}(z_{(1)}) \le \hat{f}(z_{(2)}) \le \cdots \le \hat{f}(z_{(N_k)}).$$

If $f(x)$ is noisy, we need to check the ordering with further replications calculated as follows.

(2.A) Calculate the differences between ordered samples, let $d_{i,j} = \hat{f}(x_{i,(j+1)}) - \hat{f}(x_{i,(j)})$, where $i = 1,\ldots,||\Sigma_k||$ and $j = 1,\ldots,N_k^i - 1$.

Determine $d^* = \min_{i=1,\ldots,||\Sigma_k||, j=1,\ldots,N_k^i-1} d_{i,j}$ and $S^{*2} = \max_{i=1,\ldots,||\Sigma_k||, j=1,\ldots,N_k^i} S_{\hat{f}}^2(x_{i,(j)})$.

(2.B) Calculate the updated replication number $R_k = \max\left\{R_{k-1}, \left(\frac{z_{\alpha_k/2} S^*}{d^*/2}\right)^2\right\}$, where $z_{\alpha_k/2}$ is the $1-\alpha_k/2$ quantile of the standard normal distribution. Perform $R_k - R_{k-1}$ more replications for each sample point. Re-estimate the performance of each sample point with R_k replications by $\hat{f}(x_{i,j}) = \frac{\sum_{r=1}^{R_k} g(x_{i,j},\xi_{x_{i,j}}^r)}{R_k}$.

Within each subregion $\sigma_i \in \Sigma_k$, rank all the sample points $x_{i,j}$ as $x_{i,(j)}$ representing the jth best point in subregion, according to the estimated function value, and also

update the entire order of all current samples with updated replications, so that

$$\hat{f}(x_{i,(1)}) \leq \hat{f}(x_{i,(2)}) \leq \cdots \leq \hat{f}(x_{i,(N_k^i)}), \text{ and } \hat{f}(z_{(1)}) \leq \hat{f}(z_{(2)}) \leq \cdots \leq \hat{f}(z_{(N_k)}).$$

Step 3. Build confidence interval for $y(\delta, S)$***:*** To build the confidence interval of quantile $y(\delta, S)$, first, calculate the lower and upper bounds of δ_k as

$$\delta_{kl} = \delta_k - \frac{v(\widetilde{\Sigma}_k^P)\epsilon}{v(S)v(\widetilde{\Sigma}_k^C)} \text{ and } \delta_{ku} = \delta_k + \frac{v(\widetilde{\Sigma}_k^M)\epsilon}{v(S)v(\widetilde{\Sigma}_k^C)}. \tag{4.7}$$

Then calculate the confidence interval lower bound $CI_l = \hat{f}(z_{(r)})$ and the upper bound $CI_u = \hat{f}(z_{(s)})$, where r and s are selected by

$$\max r : \sum_{i=0}^{r-1} \binom{N_k}{i} (\delta_{kl})^i (1-\delta_{kl})^{N_k - i} \leq \frac{\alpha_k}{2} \text{ and} \tag{4.8}$$

$$\min s : \sum_{i=0}^{s-1} \binom{N_k}{i} (\delta_{ku})^i (1-\delta_{ku})^{N_k - i} \geq 1 - \frac{\alpha_k}{2}. \tag{4.9}$$

Step 4. Find elite and worst subregions, and further sample with updated replications: Step 4 identifies the indices of elite and worst subregions as e and w, representing the subregions that are likely to be maintained or pruned. The sets e and w are defined with the quantile confidence interval as

$$e = \{i | \hat{f}(x_{i,(N_k^i)}) < CI_l, \text{ for } i \in 1, \ldots, ||\Sigma_k||\} \tag{4.10}$$

$$w = \{i | \hat{f}(x_{i,(1)}) > CI_u, \text{ for } i \in 1, \ldots, ||\Sigma_k||\}. \tag{4.11}$$

Statistically confirm maintaining and pruning for each elite or worst subregion by sampling points up to N_k^n, where

$$N_k^n = \left\lceil \frac{\ln \alpha_k}{\ln\left(1 - \frac{\epsilon}{v(S)}\right)} \right\rceil, \text{for all } i \in \{e \cup w\}. \tag{4.12}$$

For each new sample, perform R_k replications, and evaluate the sample mean and sample variances as in (4.6). Reorder the sampled points in each subregion σ_i, and update d^* and S^{*2} as in (2.A). As in (2.B), calculate the updated replication number $R_k^n = \max\left\{R_k, \left(\frac{z_{\alpha_k/2} S^*}{d^*/2}\right)^2\right\}$, where $z_{\alpha_k/2}$ is the $1 - \alpha_k/2$ quantile of the standard normal distribution.

Perform $R_k^n - R_k$ more replications for each new sample point. Update $\hat{f}(x_{i,(1)}) = \min_{x_{i,j} \in \sigma_i} \hat{f}(x_{i,j})$ and $\hat{f}(x_{i,(N_k^i)}) = \max_{x_{i,j} \in \sigma_i} \hat{f}(x_{i,j})$ for $i \in \{e \cup w\}$.

Step 5. Maintain, Prune, and Branch: Update the maintaining indicator functions M_i, for $i \in e$, and the pruning indicator functions P_i, for $i \in w$, as

$$M_i = \begin{cases} 1, \text{ if } \hat{f}(x_{i,(N_k^i)}) < CI_l \\ 0, \text{ otherwise} \end{cases} \quad \text{and } P_i = \begin{cases} 1, \text{ if } \hat{f}(x_{i,(1)}) > CI_u \\ 0, \text{ otherwise.} \end{cases} \tag{4.13}$$

Update the maintained set $\widetilde{\Sigma}_{k+1}^M$ and the pruned set $\widetilde{\Sigma}_{k+1}^P$ as

$$\widetilde{\Sigma}_{k+1}^M = \widetilde{\Sigma}_k^M \bigcup_{i \in e: M_i = 1} \sigma_i \quad \text{and} \quad \widetilde{\Sigma}_{k+1}^P = \widetilde{\Sigma}_k^P \bigcup_{i \in w: P_i = 1} \sigma_i,$$

and branch the remaining current subregions in the following manner.

Determine δ_{k+1} by

$$\delta_{k+1} = \frac{\delta_k v(\widetilde{\Sigma}_k^C) - \sum_{i:M_i=1} v(\sigma_i)}{v(\widetilde{\Sigma}_k^C) - \sum_{i:P_i=1} v(\sigma_i) - \sum_{i:M_i=1} v(\sigma_i)}. \tag{4.14}$$

Determine next step by follows:

(1) If all subregions $\sigma_i \in \Sigma_k$ are not branchable, terminate the algorithm.

Otherwise, if σ_i is branchable, and If σ_i, $i = 1, \ldots, ||\Sigma_k||$, has not been maintained or pruned, then partition σ_i to $\bar{\sigma}_i^1, \ldots, \bar{\sigma}_i^B$ and update the current set of subregions

$$\Sigma_{k+1}^C = \{\bar{\sigma}_i^j : \forall i \text{ to be branched}, j = 1, \ldots, B\} \text{ and } \widetilde{\Sigma}_{k+1}^C = \bigcup_{i \text{ to be branched}} \left(\bigcup_{j=1}^{B} \bar{\sigma}_i^j \right).$$

Set

$$k_c = \begin{cases} k_c + 1 & , \text{ if } \sum_{i \in e} M_i + \sum_{i \in w} P_i = 0 \\ 0 & , \text{ otherwise.} \end{cases} \tag{4.15}$$

Set $\alpha_{k+1} = \frac{\alpha_k}{B}$, $\epsilon_{k+1} = \frac{\epsilon_k}{B}$, $c_{k+1} = c_k + c$, and increment $k \leftarrow k + 1$.

(2) If $k_c \geq k_b$, set $k_c = 1$ and go to Step 1.

(3) If $k_c < k_b$, go to Step 4.

4.1.4 Performance Analysis of PBnB for Level Set Approximation

In this section, we analyze the performance of PBnB for level set approximation by deriving confidence intervals on $y(\delta, S)$ and probability bounds on the relationship of the pruned set, $\widetilde{\Sigma}_k^P$, and the maintained set, $\widetilde{\Sigma}_k^M$, with the desired level set, $L(\delta, S)$, at every iteration $k \geq 1$.

First, we analyze the performance assuming the objective function can be evaluated exactly, with no noise. In Theorem 1, we derive the interval estimation of the target quantile $y(\delta, S)$ while shifting the δ quantile over S to the δ_k quantile over the current region, $\widetilde{\Sigma}_k^C$, at iteration k. Specifically, Theorem 1 develops bounds on $y(\delta, S)$ using $y(\delta_{kl}, \widetilde{\Sigma}_k^C)$ and $y(\delta_{ku}, \widetilde{\Sigma}_k^C)$ assuming upper bounds on the incorrect maintaining and pruning volumes and then provides a confidence interval on the target quantile, $y(\delta, S)$, using the volume of subregions maintained and pruned. We provide the quality of the level set approximation by deriving probability bounds on the volume of incorrect pruning and maintaining in Theorems 2 and 3, for a single iteration. Theorem 4 considers the *sequence* of iterations from 1 to k, and provides quantile lower and upper bounds for one-sided interval estimation, while accounting for the impact of the volume of incorrect pruning and maintaining. Theorems 5 and 6 provide probability bounds on incorrect pruning and maintaining regions of maximum volume ϵ, for the *sequence* of iterations 1 to k.

The latter section discusses the impact of *noise* on the performance analysis. In Theorem 7, we derive the impact of noise on the probability of correctly ordering the samples on a single iteration and Theorem 8 extends the probability bounds to all the iterations from 1 to k. By incorporating the probability of correctly ordering the samples, Corollary 9, corresponding to Theorem 4, derives the probability that the sequence of interval quantile estimation up to iteration k captures the true target quantile when the function is noisy. Similarly, Corollaries 10 and 11 propose noisy versions of Theorems 5 and 6.

Analysis on a Deterministic Function

In Theorem 1, we let $\epsilon_k^M = v(\widetilde{\Sigma}_k^M) - v(L(\delta, S) \cap \widetilde{\Sigma}_k^M)$ denote the volume of $\widetilde{\Sigma}_k^M$ that is incorrectly maintained, and let $\epsilon_k^P = v(L(\delta, S) \cap \widetilde{\Sigma}_k^P)$ denote the volume of $\widetilde{\Sigma}_k^P$ that is incorrectly pruned at iteration k. Since volume is always non-negative, zero is a natural

lower bound on ϵ_k^M and ϵ_k^P. Upper bounds on ϵ_k^M and ϵ_k^P are used in Theorem 1 to obtain upper and lower bounds on the target quantile and provide an interval estimation.

Theorem 1. *For any iteration $k \geq 1$, suppose there is no noise and $0 \leq \epsilon_k^P \leq \frac{\epsilon v(\widetilde{\Sigma}_k^P)}{v(S)}$ and $0 \leq \epsilon_k^M \leq \frac{\epsilon v(\widetilde{\Sigma}_k^M)}{v(S)}$. Then the bounds of the target quantile are*

$$y(\delta_{kl}, \widetilde{\Sigma}_k^C) \leq y(\delta, S) \leq y(\delta_{ku}, \widetilde{\Sigma}_k^C) \tag{4.16}$$

where δ_{kl} and δ_{ku} are from (4.7). Therefore, an interval estimate of the quantile is

$$P\left(f(z_{(r)}) \leq y(\delta, S) \leq f(z_{(s)})\right) \geq 1 - \alpha_k, \tag{4.17}$$

where $z_{(1)}, \ldots, z_{(N_k)}$ are the N_k uniform samples from the current region $\widetilde{\Sigma}_k^C$ at iteration k, ordered by function values (as in Step 2), $0 < \alpha_k < 1$, and r and s satisfy (4.23) and (4.24).

Proof. We consider the iterative effect on δ_k as subregions are pruned or maintained. We use the superscript k to denote the iteration that subregions are pruned $\{\sigma_i^k : P_i = 1\}$ or maintained $\{\sigma_i^k : M_i = 1\}$. By (4.14),

$$\begin{aligned}
\delta_k &= \frac{\delta_{k-1} v(\widetilde{\Sigma}_{k-1}^C) - \sum_{i:M_i=1} v(\sigma_i^{k-1})}{v(\widetilde{\Sigma}_{k-1}^C) - \sum_{i:P_i=1} v(\sigma_i^{k-1}) - \sum_{i:M_i^{k-1}=1} v(\sigma_i^{k-1})} = \frac{\delta_{k-1} v(\widetilde{\Sigma}_{k-1}^C) - \sum_{i:M_i=1} v(\sigma_i^{k-1})}{v(\widetilde{\Sigma}_k^C)} \\
&= \frac{\frac{\delta_{k-2} v(\widetilde{\Sigma}_{k-2}^C) - \sum_{i:M_i=1} v(\sigma_i^{k-2})}{v(\widetilde{\Sigma}_{k-1}^C)} v(\widetilde{\Sigma}_{k-1}^C) - \sum_{i:M_i=1} v(\sigma_i^{k-1})}{v(\widetilde{\Sigma}_k^C)} \\
&= \frac{\delta_{k-2} v(\widetilde{\Sigma}_{k-2}^C) - \sum_{l=k-2}^{k-1} \sum_{i:M_i=1} v(\sigma_i^l)}{v(\widetilde{\Sigma}_k^C)} \\
&\vdots \\
&= \frac{\delta_1 v(\widetilde{\Sigma}_1^C) - \sum_{l=1}^{k-1} \sum_{i:M_i=1} v(\sigma_i^l)}{v(\widetilde{\Sigma}_k^C)}
\end{aligned}$$

and by the initial setting of δ_1 and $\widetilde{\Sigma}_1^C$,

$$= \frac{\delta v(S) - \sum_{l=1}^{k-1} \sum_{i:M_i=1} v(\sigma_i^l)}{v(\widetilde{\Sigma}_k^C)}$$

and $\sum_{l=1}^{k-1} \sum_{i:M_i=1} v(\sigma_i^l) = v(\widetilde{\Sigma}_k^M)$ since it denotes the volume of all maintained subregions at the end of the $k-1$ iteration, therefore,

$$= \frac{\delta v(S) - v(\widetilde{\Sigma}_k^M)}{v(\widetilde{\Sigma}_k^C)}. \tag{4.18}$$

Based on the definition of quantile, and when x is uniformly sampled, we have

$$y(\delta, S) = \underset{y \in \{f(x): x \in S\}}{\arg\min} \{P(f(x) \leq y | x \in S) \geq \delta\} = \underset{y \in \{f(x): x \in S\}}{\arg\min} \left\{ \frac{v(\{x \in S : f(x) \leq y\})}{v(S)} \geq \delta \right\}$$

and subtracting $\frac{v(\widetilde{\Sigma}_k^M) - \epsilon_k^M + \epsilon_k^P}{v(S)}$ from both sides and multiplying $\frac{v(S)}{v(\widetilde{\Sigma}_k^C)}$ on both sides,

$$= \underset{y \in \{f(x): x \in S\}}{\arg\min} \left\{ \frac{v(\{x \in S : f(x) \leq y\}) - v(\widetilde{\Sigma}_k^M) + \epsilon_k^M - \epsilon_k^P}{v(\widetilde{\Sigma}_k^C)} \geq \frac{\delta v(S) - v(\widetilde{\Sigma}_k^M) + \epsilon_k^M - \epsilon_k^P}{v(\widetilde{\Sigma}_k^C)} \right\}$$

and by (4.18), also $v(\{x \in \widetilde{\Sigma}_k^P : f(x) < y\}) = \epsilon_k^P$ and $v(\{x \in \widetilde{\Sigma}_k^M : f(x) < y\}) = v(\widetilde{\Sigma}_k^M) - \epsilon_k^M$,

$$= \underset{y \in \{f(x): x \in S\}}{\arg\min} \left\{ \frac{v(\{x \in S \setminus \{\widetilde{\Sigma}_k^P \cup \widetilde{\Sigma}_k^M\} : f(x) \leq y\})}{v(\widetilde{\Sigma}_k^C)} \geq \delta_k + \frac{\epsilon_k^M}{v(\widetilde{\Sigma}_k^C)} - \frac{\epsilon_k^P}{v(\widetilde{\Sigma}_k^C)} \right\}$$

and since $\widetilde{\Sigma}_k^C = S \setminus \{\widetilde{\Sigma}_k^P \cup \widetilde{\Sigma}_k^M\}$, and x is uniformly distributed in $\widetilde{\Sigma}_k^C$ and $x \in \widetilde{\Sigma}_k^C \subset S$,

$$= \underset{y \in \{f(x): x \in \widetilde{\Sigma}_k^C\}}{\arg\min} \left\{ P(f(x) \leq y | x \in \widetilde{\Sigma}_k^C) \geq \delta_k + \frac{\epsilon_k^M}{v(\widetilde{\Sigma}_k^C)} - \frac{\epsilon_k^P}{v(\widetilde{\Sigma}_k^C)} \right\}$$

$$= y\left(\delta_k + \frac{\epsilon_k^M}{v(\widetilde{\Sigma}_k^C)} - \frac{\epsilon_k^P}{v(\widetilde{\Sigma}_k^C)}, \widetilde{\Sigma}_k^C \right).$$

Since $0 \leq \epsilon_k^M \leq \frac{\epsilon \widetilde{\Sigma}_k^M}{v(S)}$ and $0 \leq \epsilon_k^P \leq \frac{\epsilon \widetilde{\Sigma}_k^P}{v(S)}$, an upper bound of $y\left(\delta_k + \frac{\epsilon_k^M}{v(\widetilde{\Sigma}_k^C)} - \frac{\epsilon_k^P}{v(\widetilde{\Sigma}_k^C)}, \widetilde{\Sigma}_k^C \right)$ can be achieved when $\epsilon_k^P = 0$ and $\epsilon_k^M = \frac{\epsilon v(\widetilde{\Sigma}_k^M)}{v(S)}$, yielding

$$y(\delta, S) \leq y\left(\delta_k + \frac{\epsilon v(\widetilde{\Sigma}_k^M)}{v(S) v(\widetilde{\Sigma}_k^C)}, \widetilde{\Sigma}_k^C \right) = y(\delta_{ku}, \widetilde{\Sigma}_k^C). \tag{4.19}$$

Similarly, we have a lower bound when $\epsilon_k^M = 0$ and $\epsilon_k^P = \frac{\epsilon v(\widetilde{\Sigma}_k^P)}{v(S)}$, yielding

$$y(\delta, S) \geq y\left(\delta_k - \frac{\epsilon v(\widetilde{\Sigma}_k^P)}{v(S) v(\widetilde{\Sigma}_k^C)}, \widetilde{\Sigma}_k^C \right) = y(\delta_{kl}, \widetilde{\Sigma}_k^C). \tag{4.20}$$

Note, if $\epsilon_k^M = 0$ and $\epsilon_k^P = 0$, $y(\delta_k, \widetilde{\Sigma}_k^C) = y(\delta, S)$.

At the beginning of any iteration k, the current set $\widetilde{\Sigma}_k^C$ is uniformly sampled for $N_k = c_k$ samples. Therefore, we can build the $1 - \alpha_k$ quantile confidence interval as $f(z_{(r)}) \leq y(\delta, S) \leq f(z_{(s)})$ [26], where $f(z_{(r)})$ and $f(z_{(s)})$ are the rth and sth order samples that have

the following properties

$$P(f(z_{(r)}) > y(\delta_{kl}, \widetilde{\Sigma}_k^C)) \leq \sum_{i=0}^{r-1} \binom{N_k}{i} (\delta_{kl})^i (1-\delta_{kl})^{N_k-i} \tag{4.21}$$

$$P(f(z_{(s)}) \geq y(\delta_{ku}, \widetilde{\Sigma}_k^C)) \geq \sum_{i=0}^{s-1} \binom{N_k}{i} (\delta_{ku})^i (1-\delta_{ku})^{N_k-i}. \tag{4.22}$$

The $1-\alpha_k$ confidence interval can be approximated by two one-sided intervals. Therefore, find the maximum r for which (4.21) is less than or equal to $\frac{\alpha_k}{2}$ and the minimum s for which (4.22) is greater than or equal to $1-\frac{\alpha_k}{2}$, that is

$$\max r : \sum_{i=0}^{r-1} \binom{N_k}{i} (\delta_{kl})^i (1-\delta_{kl})^{N_k-i} \leq \frac{\alpha_k}{2} \text{ and} \tag{4.23}$$

$$\min s : \sum_{i=0}^{s-1} \binom{N_k}{i} (\delta_{ku})^i (1-\delta_{ku})^{N_k-i} \geq 1 - \frac{\alpha_k}{2}. \tag{4.24}$$

Based on [26], we have

$$P\left(f(z_{(r)}^k) \leq y(\delta_{kl}, \widetilde{\Sigma}_k^C) \leq y(\delta_{ku}, \widetilde{\Sigma}_k^C) \leq f(z_{(s)}^k)\right) \geq 1-\alpha_k. \tag{4.25}$$

When there is no noise, $0 \leq \epsilon_k^P \leq \frac{\epsilon v(\widetilde{\Sigma}_k^P)}{v(S)}$ and $0 \leq \epsilon_k^M \leq \frac{\epsilon v(\widetilde{\Sigma}_k^M)}{v(S)}$, the $1-\alpha_k$ confidence interval of $y(\delta, S)$ is given by $[f(z_{(r)}), f(z_{(s)})]$ based on (4.19), (4.20) and (4.25), that is

$$P\left(f(z_{(r)}) \leq y(\delta, S) \leq f(z_{(s)})\right) \geq 1-\alpha_k.$$

□

Theorem 1 analyzes the impact of incorrect pruning and maintaining on shifting the δ_k quantile to correspond to the current set, $\widetilde{\Sigma}_k^C$. Theorem 1 assumes that ϵ_k^M is bounded by $\frac{\epsilon \widetilde{\Sigma}_k^M}{v(S)}$, which we denote as event $A_k^M = \left\{ v(L(\delta, S) \cap \widetilde{\Sigma}_k^M) \geq v(\widetilde{\Sigma}_k^M) - \frac{\epsilon v(\widetilde{\Sigma}_k^M)}{v(S)} \right\}$, and that ϵ_k^P is bounded by $\frac{\epsilon \widetilde{\Sigma}_k^P}{v(S)}$, which is represented by event $A_k^P = \left\{ v(L(\delta, S) \cap \widetilde{\Sigma}_k^P) \leq \frac{\epsilon v(\widetilde{\Sigma}_k^P)}{v(S)} \right\}$. The event that $y(\delta_{kl}, \widetilde{\Sigma}_k^C)$ and $y(\delta_{ku} \widetilde{\Sigma}_k^C)$ are bounded correctly is denoted as

$$A_k^{CI} = \left\{ f(z_{(r)}^k) \leq y(\delta_{kl}, \widetilde{\Sigma}_k^C) \leq y(\delta_{ku}, \widetilde{\Sigma}_k^C) \leq f(z_{(s)}^k) \right\}.$$

All three events, denoted

$$A_k = \left\{ A_k^M \cap A_k^P \cap A_k^{CI} \right\}, \tag{4.26}$$

ensure that $y(\delta, S)$ is bounded correctly on iteration k, i.e., $f(z^k_{(r)}) \leq y(\delta_{kl}, \widetilde{\Sigma}^C_k) \leq y(\delta, S) \leq y(\delta_{ku}, \widetilde{\Sigma}^C_k) \leq f(z^k_{(s)})$.

We next analyze the quality of pruned subregions and maintained subregions for a single iteration in Theorems 2 and 3 assuming event A_k. Probability bounds on interval estimations of the target quantile from iteration 1 to k are derived in Theorem 4. The final performances of the algorithm are derived in Theorems 5 and 6.

Theorem 2. *Consider any iteration k of PBnB on Problem $(\mathcal{P})$ where there is no noise, and suppose $\hat{\sigma}^k_p = \bigcup_{i \in w: \mathrm{P}_i = 1} \sigma_i$ has been pruned on the k^{th} iteration. Also, suppose A_k in (4.26) is true. Then, the volume of the incorrectly pruned region, i.e., $v(L(\delta, S) \cap \hat{\sigma}^k_p)$, is less than or equal to $D^P_k \epsilon_k$, where D^P_k is the number of subregions pruned at iteration k, with probability at least $1 - \alpha_k$, that is*

$$P(v(L(\delta, S) \cap \hat{\sigma}^k_p) \leq D^P_k \epsilon_k | A_k) \geq 1 - \alpha_k. \tag{4.27}$$

Proof. The probability that the volume of the incorrectly pruned region is less than or equal to $D^P_k \epsilon_k$ can be expressed as

$$P(v(L(\delta, S) \cap \hat{\sigma}^k_p) \leq D^P_k \epsilon_k | A_k) = P(v(\{x : f(x) \leq y(\delta, S), x \in \hat{\sigma}^k_p\}) \leq D^P_k \epsilon_k | A_k) \tag{4.28}$$

Consider the probability expression of quantile in (4.5), and let $\delta_p = \frac{D^P_k \epsilon_k}{v(\hat{\sigma}^k_p)}$. When x is uniformly sampled in $\hat{\sigma}^k_p$, we have

$$P(f(x) < y(\delta_p, \hat{\sigma}^k_p)) = \frac{v(\{x : f(x) < y(\delta_p, \hat{\sigma}^k_p), x \in \hat{\sigma}^k_p\})}{v(\hat{\sigma}^k_p)} \leq \delta_p = \frac{D^P_k \epsilon_k}{v(\hat{\sigma}^k_p)},$$

then multiplying $v(\hat{\sigma}^k_p)$ on both sides, we have

$$v\left(\{x : f(x) < y\left(\delta_p, \hat{\sigma}^k_p\right), x \in \hat{\sigma}^k_p\}\right) \leq D^P_k \epsilon_k. \tag{4.29}$$

Applying (4.29) to (4.28), we get a new bound on the probability as

$$\begin{aligned} &P(v(\{x : f(x) \leq y(\delta, S), x \in \hat{\sigma}^k_p\}) \leq D^P_k \epsilon_k | A_k) \\ &\quad \geq P\left(v(\{x : f(x) \leq y(\delta, S), x \in \hat{\sigma}^k_p\}) \leq v\left(\{x : f(x) < y\left(\delta_p, \hat{\sigma}^k_p\right), x \in \hat{\sigma}^k_p\}\right) \middle| A_k\right) \end{aligned}$$

and manipulating the probability to account for a possible discontinuity in $f(x)$, and associating event $\left\{v(\{x : f(x) = y(\delta, S)\}) = v(\{x : f(x) = y(\delta_p, \hat{\sigma}^k_p)\})\right\}$ to $\{v(\{x : f(x) \leq y(\delta, S)\})$

$= v(\{x : f(x) \leq y(\delta_p, \hat{\sigma}_p^k)\})\big\}$, we get

$$= P\left(v(\{x : f(x) \leq y(\delta, S), x \in \hat{\sigma}_p^k\}) \leq v\left(\{x : f(x) \leq y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\}\right) \middle| A_k\right)$$
$$-P\left(v(\{x : f(x) \leq y(\delta, S), x \in \hat{\sigma}_p^k\}) = v\left(\{x : f(x) \leq y\left(\delta_p, \hat{\sigma}_p^k\right), x \in \hat{\sigma}_p^k\}\right) \middle| A_k\right),$$

and considering the association of the volume of a level set with its quantile value, we know that $v(\{x : f(x) \leq y(\delta, S)\})$ is associated with $y(\delta, S)$, similarly, $v(\{x : f(x) \leq y(\delta_p, \hat{\sigma}_p^k), x \in \hat{\sigma}_p^k\})$ is associated with $y(\delta_p, \hat{\sigma}_p^k)$, hence, we have

$$= P\left(y(\delta, S) \leq y\left(\delta_p, \hat{\sigma}_p^k\right) \middle| A_k\right) - P\left(y(\delta, S) = y\left(\delta_p, \hat{\sigma}_p^k\right) \middle| A_k\right)$$
$$= P\left(y(\delta, S) < y\left(\delta_p, \hat{\sigma}_p^k\right) \middle| A_k\right),$$

and by the condition A_k and the pruned assumption, we have $y(\delta, S) \leq y(\delta_{ku}, \widetilde{\Sigma}_k^C) \leq f(z_{(s)}) < f(x_{(p),(1)})$, where $x_{(p),(1)}$ is the best sample in $\hat{\sigma}_p^k$, therefore,

$$\geq P\left(f(x_{(p),(1)}) \leq y\left(\delta_p, \hat{\sigma}_p^k\right) \middle| A_k\right)$$
$$= 1 - P\left(f(x_{(p),(1)}) > y\left(\delta_p, \hat{\sigma}_p^k\right) \middle| A_k\right),$$

and since we uniformly sample $D_k^P N_k^p$ independent samples in $\hat{\sigma}_p^k$,

$$\geq 1 - (1 - \delta_p)^{D_k^P N_k^p}. \tag{4.30}$$

Since we desire the probability to be greater than $1 - \alpha_k$, we let

$$\begin{aligned} 1 - (1 - \delta_p)^{D_k^P N_k^p} &\geq 1 - \alpha_k \\ \Rightarrow (1 - \delta_p)^{D_k^P N_k^p} &\leq \alpha_k \\ \Rightarrow D_k^P N_k^p &\geq \frac{\ln \alpha_k}{\ln (1 - \delta_p)}. \end{aligned} \tag{4.31}$$

Since $N_k^p = \left\lceil \frac{\ln \alpha_k}{\ln (1 - \frac{\epsilon_k}{v(\sigma_i)})} \right\rceil$ in Step 4, and $\delta_p = \frac{D_k^P \epsilon_k}{v(\hat{\sigma}_p)} = \frac{D_k^P \epsilon_k}{D_k^P v(\sigma_i)} = \frac{\epsilon_k}{v(\sigma_i)}$, where σ_i is a subregion pruned at the kth iteration, and $D_k^P \geq 1$, then (4.31) holds and consequently (4.27) holds.

□

Theorem 3. *Consider any iteration k of PBnB on Problem $(\mathcal{P})$ where there is no noise, and suppose $\hat{\sigma}_m^k = \bigcup_{i \in w : \mathrm{M}_i = 1} \sigma_i$ has been maintained on the kth iteration. Also, suppose A_k in (4.26) is true. Then, the volume of the correctly maintained region is greater than*

or equal to $v(\hat{\sigma}_m^k) - D_k^M \epsilon_k$*, where* D_k^M *is the number of subregions pruned at iteration* k*, with probability at least* $1-\alpha_k$*, or, in other words, the volume of the incorrectly maintained region is less than or equal to* $D_k^M \epsilon_k$ *with probability* $1-\alpha_k$*,*

$$P(v(\hat{\sigma}_m^k \setminus L(\delta, S)) \leq D_k^M \epsilon_k | A_k) \geq 1 - \alpha_k. \tag{4.32}$$

Proof. The proof is similar to Theorem 2. □

Theorem 1 assumes $0 \leq \epsilon_k^M \leq \frac{\epsilon v(\widetilde{\Sigma}_k^M)}{v(S)}$ and $0 \leq \epsilon_k^P \leq \frac{\epsilon v(\widetilde{\Sigma}_k^P)}{v(S)}$, and Theorems 2 and 3 assume A_k. Now, in Theorem 4, we remove these conditions.

Theorem 4. *For any iteration* $k \geq 1$*, suppose there is no noise and use* δ_{kl} *and* δ_{ku} *from (4.7) to estimate target quantile* $y(\delta, S)$*. The probability that all interval estimates from iteration* 1 *to* k *capture the original quantile* $y(\delta, S)$ *is bounded as*

$$P\left(\cap_{i=1}^k f(z_{(r)}^i) \leq y(\delta, S) \leq f(z_{(s)}^i)\right) \geq (1-\alpha)^3 \tag{4.33}$$

where $z_{(r)}^i$ and $z_{(s)}^i$ are from Step 3, using (4.8) and (4.9) at iteration i.

Proof. Without loss of generality, suppose that the algorithm builds a confidence interval every iteration, that is $k_b = 1$.

Because we estimate the target quantile using $y(\delta_{ku}, \widetilde{\Sigma}_k^C)$ and $y(\delta_{kl}, \widetilde{\Sigma}_k^C)$ as upper and lower bounds in Step 3, we consider a lower bound for a single estimation at iteration k as

$$\begin{aligned}
&P\left(f(z_{(r)}^k) \leq y(\delta, S) \leq f(z_{(s)}^k)\right) \\
&\quad \geq P\left(\left\{f(z_{(r)}^k) \leq y(\delta_{kl}, \widetilde{\Sigma}_k^C) \leq y(\delta, S) \leq y(\delta_{ku}, \widetilde{\Sigma}_k^C) \leq f(z_{(s)}^k)\right\}\right)
\end{aligned}$$

and when the conditions of Theorem 1 are satisfied, that is, the incorrect maintained volume is less than or equal to $\frac{\epsilon v(\widetilde{\Sigma}_k^M)}{v(S)}$, and the incorrect pruned volume is less than or equal to $\frac{\epsilon v(\widetilde{\Sigma}_k^P)}{v(S)}$, then $y(\delta_{kl}, \widetilde{\Sigma}_k^C) \leq y(\delta, S) \leq y(\delta_{ku}, \widetilde{\Sigma}_k^C)$, so we get

$$\begin{aligned}
&\geq P\left(\left\{v(L(\delta, S) \cap \widetilde{\Sigma}_k^M) \geq v(\widetilde{\Sigma}_k^M) - \frac{\epsilon v(\widetilde{\Sigma}_k^M)}{v(S)}\right\} \bigcap \left\{v(L(\delta, S) \cap \widetilde{\Sigma}_k^P) \leq \frac{\epsilon v(\widetilde{\Sigma}_k^P)}{v(S)}\right\}\right. \\
&\qquad \left.\bigcap \left\{f(z_{(r)}^k) \leq y(\delta_{kl}, \widetilde{\Sigma}_k^C) \leq y(\delta_{ku}, \widetilde{\Sigma}_k^C) \leq f(z_{(s)}^k)\right\}\right) \\
&= P\left(\left\{A_k^M \cap A_k^P \cap A_k^{CI}\right\}\right). \qquad (4.34)
\end{aligned}$$

Now, working towards (4.33), we derive a lower bound on the probability that the quantile interval estimate captures the target quantile at every iteration from 1 to k, using (4.34), that is

$$P\left(\cap_{i=1}^{k} f(z_{(r)}^{i}) \leq y(\delta, S) \leq f(z_{(s)}^{i})\right) \geq P(\cap_{i=1}^{k} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}). \tag{4.35}$$

We start proving (4.35) by using mathematical induction to show that

$$P(\cap_{i=1}^{k} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}) \geq \prod_{i=1}^{k-1}(1-\alpha_i) \prod_{i=1}^{k-1}(1-\alpha_i) \prod_{i=1}^{k}(1-\alpha_i). \tag{4.36}$$

For $k = 1$, $P(A_1^M \cap A_1^P \cap A_1^{CI}) = P(\left\{A_1^M \cap A_1^P | A_1^{CI}\right\})P(A_1^{CI}) \geq 1 \cdot (1-\alpha_1) \geq (1-\alpha_1)$, where $P(A_1^{CI}) > (1-\alpha_1)$ because the bounds are built with (4.23) and (4.24), and $P(\left\{A_1^M \cap A_1^P | A_1^{CI}\right\}) = 1$ since $\widetilde{\Sigma}_1^M$ and $\widetilde{\Sigma}_1^P$ are $\{\phi\}$.

Suppose $k = j$ holds, we have

$$P(\cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}) \geq \prod_{i=1}^{j-1}(1-\alpha_i) \prod_{i=1}^{j-1}(1-\alpha_i) \prod_{i=1}^{j}(1-\alpha_i). \tag{4.37}$$

For $k = j+1$,

$$\begin{aligned}
&P(\cap_{i=1}^{j+1} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}) \\
&\quad = P\left(\left\{A_{j+1}^M \cap A_{j+1}^P \cap A_{j+1}^{CI}\right\} | \cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}\right) P(\cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}) \\
&\quad = P\left(\left\{A_{j+1}^M \cap A_{j+1}^P\right\} | \cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\} \cap A_{j+1}^{CI}\right) \\
&\qquad \cdot P\left(A_{j+1}^{CI} | \cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}\right) P(\cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}).
\end{aligned}$$

Note, A_{j+1}^{CI} is $\left\{f(z_{(r)}^{j+1}) \leq y(\delta_{j+1l}, \widetilde{\Sigma}_{j+1}^C) \leq y(\delta_{j+1u}, \widetilde{\Sigma}_{j+1}^C) \leq f(z_{(s)}^{j+1})\right\}$, which is selecting bounds for $y(\delta_{j+1l}, \widetilde{\Sigma}_{j+1}^C)$ and $y(\delta_{j+1u}, \widetilde{\Sigma}_{j+1}^C)$ with a predetermined δ_{j+1l} and δ_{j+1u}. Therefore, all A_i^M and A_i^P, for $i \leq j$, only affect the development of δ_{j+1l} and δ_{j+1u} but do not have any impact on A_{j+1}^{CI}. Furthermore, since the $j+1$ iteration reuses samples from previous iterations, we have $P\left(A_{j+1}^{CI} | \cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}\right) > P\left(A_{j+1}^{CI}\right)$. Therefore,

$$P(\cap_{i=1}^{j+1}\left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\})$$
$$\geq P\left(\left\{A_{j+1}^M \cap A_{j+1}^P\right\} | \cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\} \cap A_{j+1}^{CI}\right)$$
$$\cdot P\left(A_{j+1}^{CI}\right) P(\cap_{i=1}^{j}\left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}),$$

and since the elite and worse subregions are mutually exclusive and the samples used for pruning and maintaining are sampled from seperate subregions, A_{j+1}^M and A_{j+1}^P are independent, thus we have

$$\geq P\left(A_{j+1}^M | \cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\} \cap A_{j+1}^{CI}\right) P\left(A_{j+1}^P | \cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\} \cap A_{j+1}^{CI}\right)$$
$$\cdot P\left(A_{j+1}^{CI}\right) P(\cap_{i=1}^{j}\left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\})$$

and by separating the subregions that have been maintained and pruned on iteration $j+1$ from previous iterations 1 to j $\left(\widetilde{\Sigma}_{j+1}^M = \hat{\sigma}_m^j \cap \widetilde{\Sigma}_j^M \text{ and } \widetilde{\Sigma}_{j+1}^P = \hat{\sigma}_p^j \cap \widetilde{\Sigma}_j^P\right)$ and since the probabilities are conditioned on A_j^M and A_j^P, we get a lower bound by considering the volumes of incorrect pruning and maintaining only on the last iteration,

$$\geq P\left(v(\hat{\sigma}_m^j \setminus L(\delta, S)) \leq D_j^M \epsilon_j | \cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\} \cap A_{j+1}^{CI}\right)$$
$$\cdot P\left(v(L(\delta, S) \cap \hat{\sigma}_p^j) \leq D_j^P \epsilon_j | \cap_{i=1}^{j} \left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\} \cap A_{j+1}^{CI}\right)$$
$$\cdot P\left(A_{j+1}^{CI}\right) P(\cap_{i=1}^{j}\left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}),$$

and since removing the conditioning on the event A_{j+1}^{CI} and the events $\left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\}$ for prior iterations, $i < j$, can allow for more incorrect pruning and maintaining on the current iteration, we have

$$\geq P\left(v(\hat{\sigma}_m^j \setminus L(\delta, S)) \leq D_j^M \epsilon_j | \left\{A_j^M \cap A_j^P \cap A_j^{CI}\right\}\right)$$
$$\cdot P\left(v(L(\delta, S) \cap \hat{\sigma}_p^j) \leq D_j^P \epsilon_j | \left\{A_j^M \cap A_j^P \cap A_j^{CI}\right\}\right)$$
$$\cdot P\left(A_{j+1}^{CI}\right) P(\cap_{i=1}^{j}\left\{A_i^M \cap A_i^P \cap A_i^{CI}\right\})$$

and using Theorems 2 and 3, (4.25) and (4.37), we have

$$\geq (1-\alpha_j)(1-\alpha_j)(1-\alpha_{j+1})\prod_{i=1}^{j-1}(1-\alpha_i)\prod_{i=1}^{j-1}(1-\alpha_i)\prod_{i=1}^{j}(1-\alpha_i)$$
$$=\prod_{i=1}^{j}(1-\alpha_i)\prod_{i=1}^{j}(1-\alpha_i)\prod_{i=1}^{j+1}(1-\alpha_i).$$

Hence, (4.37) holds for $k=j+1$, and by mathematical induction (4.36) holds.

Using (4.36) to lower bound (4.35), yields

$$P\left(\cap_{i=1}^{k} f(z^i_{(r)}) \leq y(\delta,S) \leq f(z^i_{(s)})\right) \geq \prod_{i=1}^{k-1}(1-\alpha_i)\prod_{i=1}^{k-1}(1-\alpha_i)\prod_{i=1}^{k}(1-\alpha_i) = (1-\alpha_k)\prod_{i=1}^{k-1}(1-\alpha_i)^3$$

and in order to establish a pattern, and since $(1-\alpha_k)<1$, we write

$$\geq (1-\alpha_k)^B(1-\alpha_k)^B(1-\alpha_k)^B\prod_{i=1}^{k-1}(1-\alpha_i)^3$$

and by applying Bernoulli's inequality, and since $\alpha_k=\frac{\alpha_{k-1}}{B}$, we get

$$\geq (1-\alpha_{k-1})(1-\alpha_{k-1})(1-\alpha_{k-1})\prod_{i=1}^{k-1}(1-\alpha_i)^3 = (1-\alpha_{k-1})^6\prod_{i=1}^{k-2}(1-\alpha_i)^3$$

and now, since $B\geq 2$, the pattern starts to repeat as,

$$\geq (1-\alpha_{k-1})^B(1-\alpha_{k-1})^B(1-\alpha_{k-1})^B\prod_{i=1}^{k-2}(1-\alpha_i)^3$$

and by repeatedly applying Bernoulli's inequality,

$$\geq (1-\alpha)^3. \tag{4.38}$$

□

Theorem 5. *For any iteration $k\geq 0$ of PBnB on $(\mathcal{P})$ when there is no noise, the volume of the incorrectly pruned region is at most ϵ with probability at least $(1-\alpha)^4$, that is*

$$P(v(L(\delta,S)\cap\widetilde{\Sigma}^P_{k+1})\leq\epsilon)\geq(1-\alpha)^4. \tag{4.39}$$

Proof. First consider the conditional probability of the incorrectly pruned volume, given all confidence intervals capture the true quantile by A_i for all $i=1,\ldots,k$, and by the definition

of $\widetilde{\Sigma}^P_{k+1}$, $v(L(\delta,S)\cap\widetilde{\Sigma}^P_{k+1}) = v(\cup_{i=1}^k L(\delta,S)\cap\hat{\sigma}^i_p)$, we have

$$\begin{aligned} &P(v(L(\delta,S)\cap\widetilde{\Sigma}^P_{k+1})\leq\epsilon|\cap_{i=1}^k A_i),\\ &= P\left(v\left(\cup_{i=1}^k L(\delta,S)\cap\hat{\sigma}^i_p\right)\leq\epsilon\middle|\cap_{i=1}^k A_i\right) \end{aligned} \tag{4.40}$$

and at each iteration, at D_i^P subregions are pruned, yielding,

$$\geq P\left(\cap_{i=1}^k v\left(L(\delta,S)\cap\hat{\sigma}^i_p\right)\leq D_i^P\epsilon_i\middle|\cap_{l=1}^k A_i\right)$$

considering a lower bound that every subregion is pruned with probability bounds as given in Theorem 2,

$$\geq\prod_{i=1}^k(1-\alpha_i)=\prod_{i=1}^k\left(1-\frac{\alpha}{B^i}\right)\geq\left(1-\frac{\alpha}{B^k}\right)^B\prod_{i=1}^{k-1}\left(1-\frac{\alpha}{B^i}\right),$$

and by applying Bernoulli's inequality, $\left(1-\frac{\alpha}{B^k}\right)^B\geq\left(1-B\frac{\alpha}{B^k}\right)$, repeatedly

$$\geq\left(1-\frac{\alpha}{B}\right)^B\geq 1-\alpha. \tag{4.41}$$

Combining (4.33) from Theorem 4 with (4.41), we get

$$P(v(L(\delta,S)\cap\widetilde{\Sigma}^P_{k+1})\leq\epsilon)\geq(1-\alpha)^4. \tag{4.42}$$

□

Theorem 6. *For any iteration $k\geq 0$ of PBnB on $(\mathcal{P})$ when there is no noise, the volume of the incorrectly maintained region is at most ϵ with probability at least $(1-\alpha)^4$, that is*

$$P(v(\widetilde{\Sigma}^M_{k+1}\setminus L(\delta,S))\leq\epsilon)\geq(1-\alpha)^4. \tag{4.43}$$

Proof. The proof is similar to Theorem 5. □

Analysis on a Noisy Function

In the previous analysis, we assume the objective function could be evaluated without noise, and in this section we account for a noisy function. Theorem 7 provides probability bounds for correctly ordering the estimated function values. Theorem 8 combines all iterations from 1 to k and gives a probability bound of $1-\alpha$ on the correct ordering. Corollaries 9, 10 and 11 provide noisy versions of Theorems 4, 5 and 6.

We use the analysis of a two-stage replication approach, by [9], in Theorem 7, which has following assumption:

(A1) The noisy function is normally distributed with an unknown common variance σ^2, and at each solution $z_i \in S$, the variance can be expressed as $a_i\sigma^2$ where a_i is a known constant for each i.

In PBnB, the constants a_i are not known, hence, we cannot exactly implement their two-stage replication approach. Our modified two-stage replication approach is described after Theorem 7.

Theorem 7. *(Based on [9]) With Assumption (A1), the probability of correctly ordering all samples in the current region at iteration k is*

$$P(\hat{f}(z^k_{(1)}) \leq \hat{f}(z^k_{(2)}) \leq \cdots \leq \hat{f}(z^k_{(N_l)}) | f(z^k_{(1)}) \leq f(z^k_{(2)}) \leq \cdots \leq f(z^k_{(N_k)})) \geq 1 - \alpha_k, \quad (4.44)$$

given that we have a_iR_0 as the first stage replication number for each sample point to estimate the common variance by $S_0^2 = \frac{1}{N_k}\sum_{i=1}^{N_k} \frac{1}{a_i} S^2_{\hat{f}}(z_{(i)})$ and set up the minimum difference desired to be separated as d^*, and sampling $R_k = \max\{a_iR_0, 2(\frac{hS_0^*}{d^*/2})^2\}$ in the second stage of the procedure, where h is a value in the H c.d.f. of a multivariate student t's distribution with $N_k - 1$ dimensions such that $H(h) = 1 - \alpha_k$.

Proof. See [9]. □

For any iteration k, we take a conservative approach to achieve the $1-\alpha_k$ probability of correct ordering by separating each estimated performance mean $\hat{f}(z_{(i)})$ with its neighbor by the smallest difference $d^* = \min_{i=1,\ldots,N_k-1} d_i$ of any two neighbors. We also use the largest variance $S^{*2} = \max_{i=1,\ldots,N_k} S^2_{\hat{f}}(z_{(i)})$ so that all ordering is conservative. Specifically, the implemented two-stage procedure for any iteration k is in (2.A) and (2.B) in Steps 2 and 4. For the implemented two-stage procedure, we drop the assumption of a common variance with $a_i\sigma^2$ and only assume the noise is normally distributed. In Step 2 of the implemented two-stage procedure we use $z_{\alpha_k/2}$, the $1-\alpha_k/2$ quantile of the standard normal distribution, together with d^* and S^* so that the correct ordering of two function values separated by d^* with S^{*2} variance is achieved with probability at least $1 - \alpha_k$.

The following Theorem 8 considers Assumption (A1) and Theorem 7 to derive a bound on the probability of correctly ordering the estimated function values for iteration 1 to k.

Theorem 8. *With Assumption (A1), the probability of correct ordering from iteration* 1 *to iteration* k *is*

$$P\left(\bigcap_{l=1}^{k}\left(\hat{f}(z^l_{(1)}) \leq \hat{f}(z^l_{(2)}) \leq \cdots \leq \hat{f}(z^l_{(N_l)}) | f(z^l_{(1)}) \leq f(z^l_{(2)}) \leq \cdots \leq f(z^l_{(N_l)})\right)\right) \geq 1-\alpha, \tag{4.45}$$

where $z^l_{(j)}$ *is the* j^{th} *ordered sampled point at iteration* l*.*

Proof. The probability of correct ordering from iteration 1 to iteration k can be expressed using conditional probabilities as

$$P\left(\bigcap_{l=1}^{k}\left(\hat{f}(z^l_{(1)}) \leq \hat{f}(z^l_{(2)}) \leq \cdots \leq \hat{f}(z^l_{(N_l)}) | f(z^l_{(1)}) \leq f(z^l_{(2)}) \leq \cdots \leq f(z^l_{(N_l)})\right)\right)$$

where R_k is non-decreasing and R_k is chosen so that Theorem 7 is satisfied, hence the probability the ordering is correct on iteration l given that the ordering was correct on the previous iterations is greater than or equal to the unconditioned probability that the ordering is correct on iteration l, therefore,

$$\geq \prod_{l=1}^{k} P(\hat{f}(z^l_{(1)}) \leq \hat{f}(z^l_{(2)}) \leq \cdots \leq \hat{f}(z^l_{(N_l)}) | f(z^l_{(1)}) \leq f(z^l_{(2)}) \leq \cdots \leq f(z^l_{(N_l)}))$$

and by Theorem 7 and applying Bernoulli's inequality repeatedly

$$\geq \prod_{l=1}^{k}\left(1-\frac{\alpha}{B^l}\right) \geq 1-\alpha. \tag{4.46}$$

□

With the probability bound of correct ordering, we next derive the probability that the sequence of interval estimation is correct in Corollary 9. In Corollaries 10 and 11, we derive versions of Theorem 5 and Theorem 6 with noise.

Corollary 9. *Consider any iteration* $k \geq 1$ *of PBnB on* $(\mathcal{P})$ *where (A1) is assumed. The probability that all one-sided interval estimates from iteration* 1 *to* k *capture the original*

quantile $y(\delta, S)$ *is bounded by*

$$P\left(\cap_{i=1}^{k} f(z_{(r)}^{i}) \leq y(\delta, S) \leq f(z_{(s)}^{i})\right) \geq (1-\alpha)^4 \tag{4.47}$$

where $z_{(s)}^{i}$ *and* $z_{(r)}^{i}$ *are selected as in Step 3 for iteration* $i = 1, \ldots, k$.

Proof. Theorem 4 considers the probability of a sequence of interval estimations until iteration k with no noise in the objective function. Theorem 4 holds under the condition that each iteration's ordering of samples is correct, however when the objective function is noisy, the order of samples may be incorrect. Therefore, the probability bounds for a noisy function should include the probability of correct ordering, as in Theorem 8, which requires including a $(1-\alpha)$ probability term in the original bound. □

Corollary 10. *Consider any iteration* $k \geq 1$ *of PBnB on* $(\mathcal{P})$ *where (A1) is assumed. The probability of incorrectly pruning a volume of at most* ϵ *is bounded by*

$$P(v(L(\delta, S) \cap \widetilde{\Sigma}_{k+1}^{P}) \leq \epsilon) \geq (1-\alpha)^5. \tag{4.48}$$

Corollary 11. *Consider any iteration* $k \geq 1$ *of PBnB on* $(\mathcal{P})$ *where (A1) is assumed. The probability of incorrectly maintaining a volume of at most* ϵ *is bounded by*

$$P(v(\widetilde{\Sigma}_{k+1}^{M} \setminus L(\delta, S)) \leq \epsilon) \geq (1-\alpha)^5. \tag{4.49}$$

4.1.5 Numerical Results

In this section, we test PBnB for level set approximation with several test functions to emphasize PBnB's capability and quality of the level set approximation. The test functions are the sphere function, Rosenbrock's function, the centered sinusoidal function, and the shifted sinusoidal function [4]. The definition of test functions are listed below:

- Sphere function $(-10 \leq x_i \leq 10, i = 1, \ldots, n)$

$$g_0(x) = \sum_{i=1}^{n} x_i^2. \tag{4.50}$$

  The global optimum is located at $x_* = (0, \ldots, 0)$ with $g_1(x_*) = 0$.

- Rosenbrock's function $(-2 \leq x_i \leq 2, i = 1, \ldots, n)$

$$g_1(x) = \sum_{i=1}^{n-1} [(1 - x_i)^2 + 100(x_{i+1} - x_i^2)^2]. \tag{4.51}$$

  The global optimum is located at $x_* = (1, \ldots, 1)$ with $g_1(x_*) = 0$.

- Centered sinusoidal function $(0 \leq x_i \leq 180, i = 1, \ldots, n)$

$$g_3(x) = -2.5 \prod_{i=1}^{n} \sin\left(\frac{\pi x_i}{180}\right) - \prod_{i=1}^{n} \sin\left(\frac{\pi x_i}{36}\right). \tag{4.52}$$

  The global optimum is located at $x_* = (90, \ldots, 90)$ with $g_2(x_*) = -3.5$.

- Shifted sinusoidal function $(0 \leq x_i \leq 180, i = 1, \ldots, n)$

$$g_3(x) = -2.5 \prod_{i=1}^{n} \sin\left(\frac{\pi(x_i + 60)}{180}\right) - \prod_{i=1}^{n} \sin\left(\frac{\pi(x_i + 60)}{36}\right). \tag{4.53}$$

  The global optimum is located at $x_* = (30, \ldots, 30)$ with $g_4(x_*) = -3.5$.

First we consider the sphere function, because we can identify the true level set and then evaluate the quality of the level set approximation with theoretical results. The other three test functions are tested in the no noise setting with dimensions ranging up to ten. Following no noise functions, we focus on the influence of a normal noise added to each test function. We also demonstrate how the algorithm performs with integer variables. The parameters of the algorithm are set as follows, $\delta = 0.1$, $B = 2$, $\alpha = 0.05$, $\epsilon = 0.025v(S)$, $k_b = 2$, $c_k = 1000$, and $R_o = 20$ when the function is noisy. Also, we apply an upper bound of the sample size for each subregion by the density of sample points as $\frac{N_k^i}{v(\sigma_i)} \leq \frac{100^n}{v(S)}$.

Sphere Function with Normal Noise

The true level set of a sphere function can be easily calculated for two and three dimensional problems and used to compare the numerical performance with the theoretical analysis of Theorems 5, 6, Corollaries 10 and 11. The algorithm was run 100 times under no noise condition, and no run incorrectly pruned or maintained a region of volume larger than ϵ. In fact, the volume of incorrectly pruned or maintained for all 100 runs was zero - the result was

Table 4.1: Solution quality for sphere function with 100 runs.

n	number of runs incorrect maintain > 0	number of runs incorrect maintain $> \epsilon$	number of runs incorrect prune > 0	number of runs incorrect prune $> \epsilon$
2	0	0	5	0
3	0	0	0	0

perfect. Therefore, we focus on the function with $N(0, 1)$ noise. Table 4.1 lists the number of runs (out of 100) that had incorrectly maintained and pruned volumes greater than 0 and ϵ. For all 100 runs of PBnB for level set approximation on the sphere function with $N(0, 1)$ noise, no run's incorrect maintained volume exceeded the user-defined parameter, $\epsilon = 0.025v(S)$. For $n = 2$, only 5 runs have an incorrectly pruned volume greater than 0. This suggests that the sample size used in the algorithm is conservative and the bounds in Corollaries 10 and 11 are not tight.

Continuous Test Functions with No Noise

In this section, we illustrate the pruning and maintaining subregions in Figure 4.4 for the 10th iteration of PBnB on a single run on the Rosenbrock's function, the centered sinusoidal function, and the shifted sinusoidal function in two dimensions without adding noise. In Figure 4.4, the dark gray (blue) boxes are the current undecided subregions, the white boxes represent the pruned subregions, the light gray (green) boxes are the maintained subregions, and the bold line represents the target level set (as in Figure 4.2). The maintained subregions are clearly contained in the target level set, and the current subregions form the boundary of the level set. Practically speaking, there is a potential that some pruned subregions contain part of the target level set. However, the algorithm ensures that the volume is bounded by the user-defined parameter ϵ with probability bounds in Corollaries 10 and 11. From the two dimensional problems, we can observe that the interaction of the test function's level set and the partition scheme affects the volume maintained and confirmed as a part of the

level set at iteration 10 in Figure 4.4.

Figure 4.5 illustrates the updates of interval quantile estimations for the two dimensional Rosenbrock's function, centered sinusoidal function, and shifted sinusoidal function. This interval estimation, $[\hat{f}(z_{((r)}), \hat{f}(z_{((s)})]$, represents the performance bound of the target level set as a reference for decision makers. The estimation updates for Rosenbrock's functions narrow the interval width to help pruning and maintaining. For the shifted sinusoidal function, there are subregions pruned or maintained for every iteration from $k = 3$. Hence, the quantile estimation does not update with new estimations. The centered sinusoidal function's interval estimation shrinks at the first update but loosens at the second update because more pruned and maintained subregions widen the δ_{ku} and δ_{kl}. Currently, c_k increases linearly with iterations. However, for difficult functions, a non-linear increasing of c_k may provide tighter confidence intervals considering the use of the δ_{ku} and δ_{kl}.

Figure 4.6 demonstrates the relative volume of solutions pruned, maintained, and undecided for iteration 1 to 10. In two dimensions, almost the entire target level set is captured as illustrated in Figure 4.4. Three and five dimensional test functions, $n = 3, 5$, are also used to test the algorithm, and the summary results are shown in Table 4.2. For all test functions, the algorithm still approximates the target 10% quantile level set and achieves between 4 and 6% maintained volume before terminating. Also, we can observe that the shifted sinusoidal function requires fewer samples for 2 and 3 dimensional problems. It is possible that the structure of the function allows the algorithm to prune a large portion of the solution space earlier than other two functions, as shown in Figure 4.6.

In higher dimensional problems ($n = 7$ and 10), approximating the entire target level set may be computationally expensive. PBnB for level set approximation allows early termination to capture a part of the target level set. As in Figure 4.6, the volume of maintained subregions increases with more iterations and sample points. In Table 4.3, the number of iterations and sample points are shown for the first subregion maintained for each test dimensions, $n = 2, 3, 5, 7$, and 10. The result indicates that PBnB can maintain the first subregion fairly quickly. Specifically, the centered sinusoidal function takes more iterations and sample points to first maintain a subregion. The shifted sinusoidal function consistently maintains a subregion before the other two test functions. It shows the speed

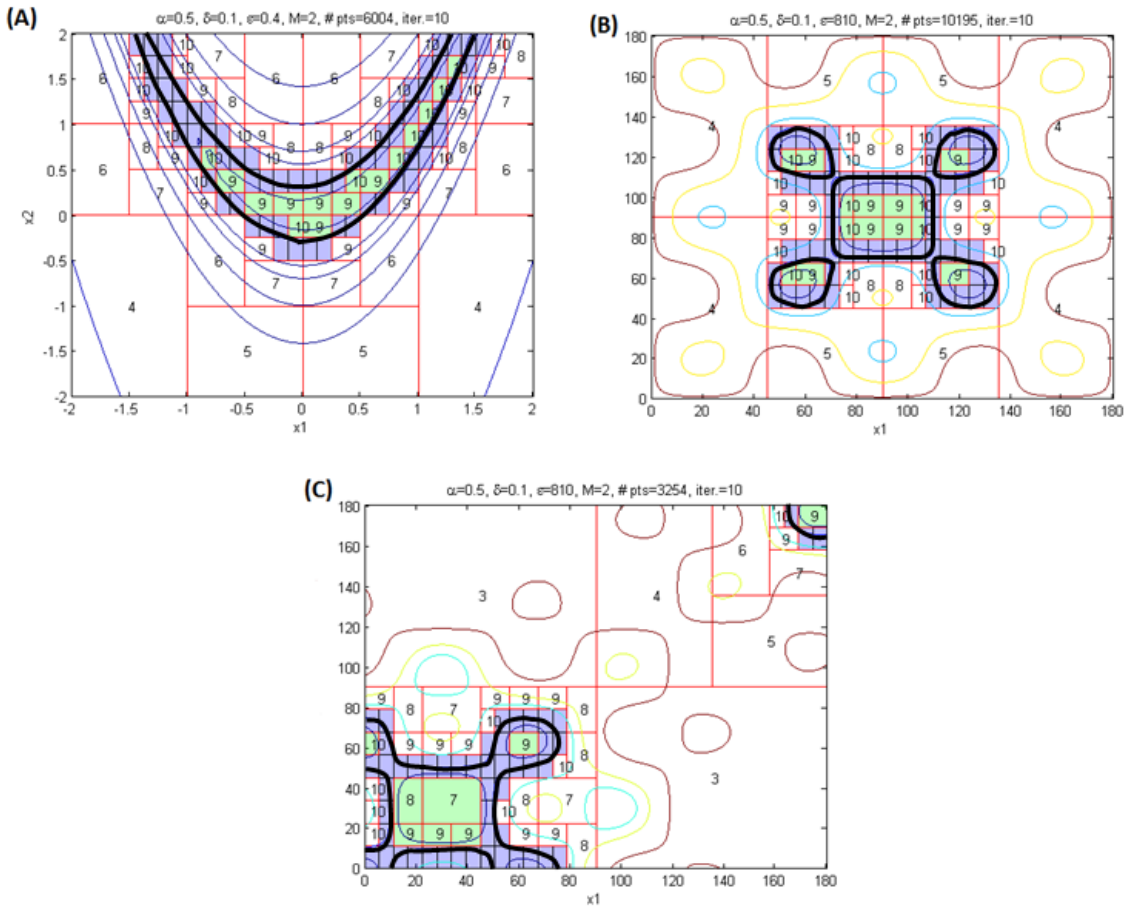

Figure 4.4: Approximating the level set bounded by the 0.1 quantile on the tenth iteration of PBnB for two dimensional (A) Rosenbrock's function, (B) centered sinusoidal function, and (C) shifted sinusoidal function by the maintained green (light gray) subregions.

Table 4.2: Comparison of Rosenbrock, centered and shifted sinusoidal functions with number of samples and ratio of volume maintained (RVM).

Test function	Rosenbrock		Centered Sinusoidal		Shifted Sinusoidal	
Dimension	# sampled	RVM	# sampled	RVM	# sampled	RVM
$n = 2$	6,004	4.30%	10,195	4.29%	3,254	4.49%
$n = 3$	159,655	5.83%	212,563	4.87%	104,825	5.38%
$n = 5$	71,529,641	5.77%	126,702,225	4.14%	130,595,965	4.67%

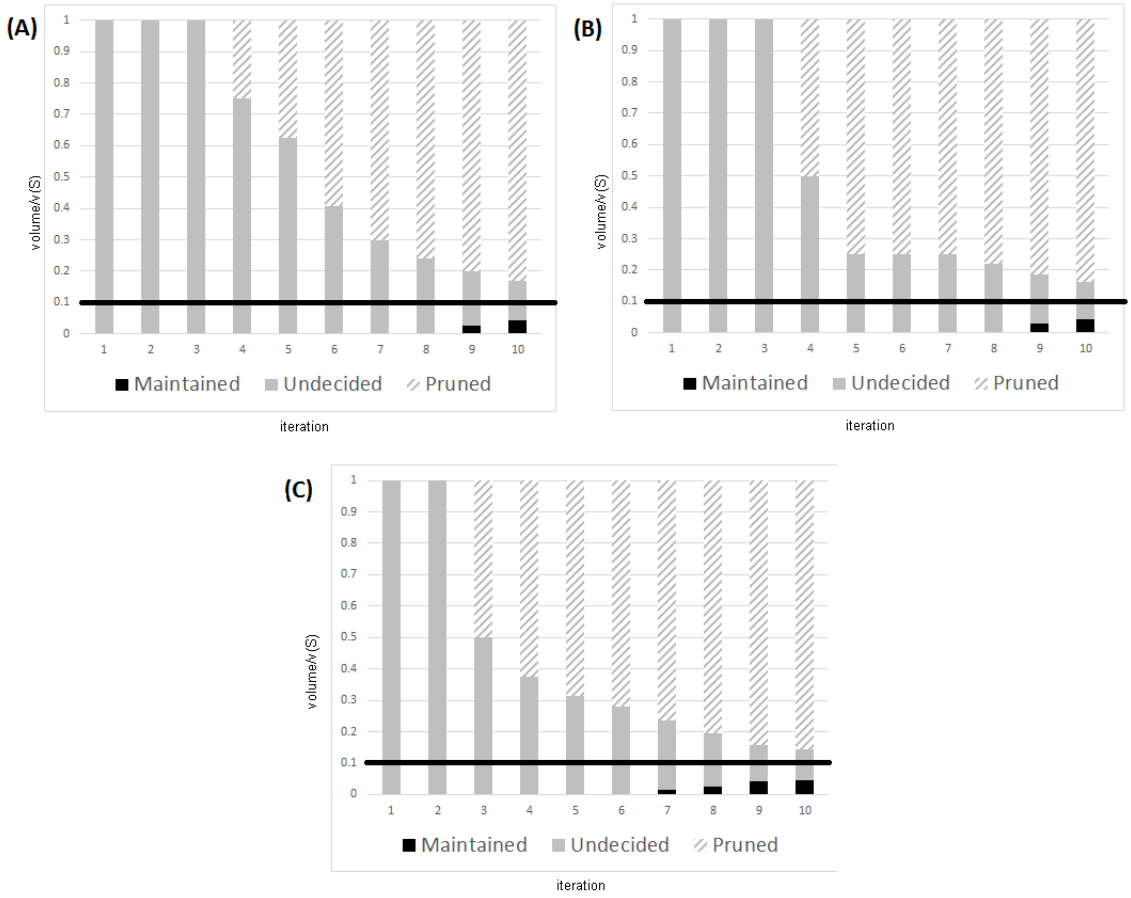

Figure 4.5: Volume maintained and pruned by PBnB approximating the level set bounded by the 0.1 quantile for two dimensional (A) Rosenbrock's function, (B) centered sinusoidal function, and (C) shifted sinusoidal function.

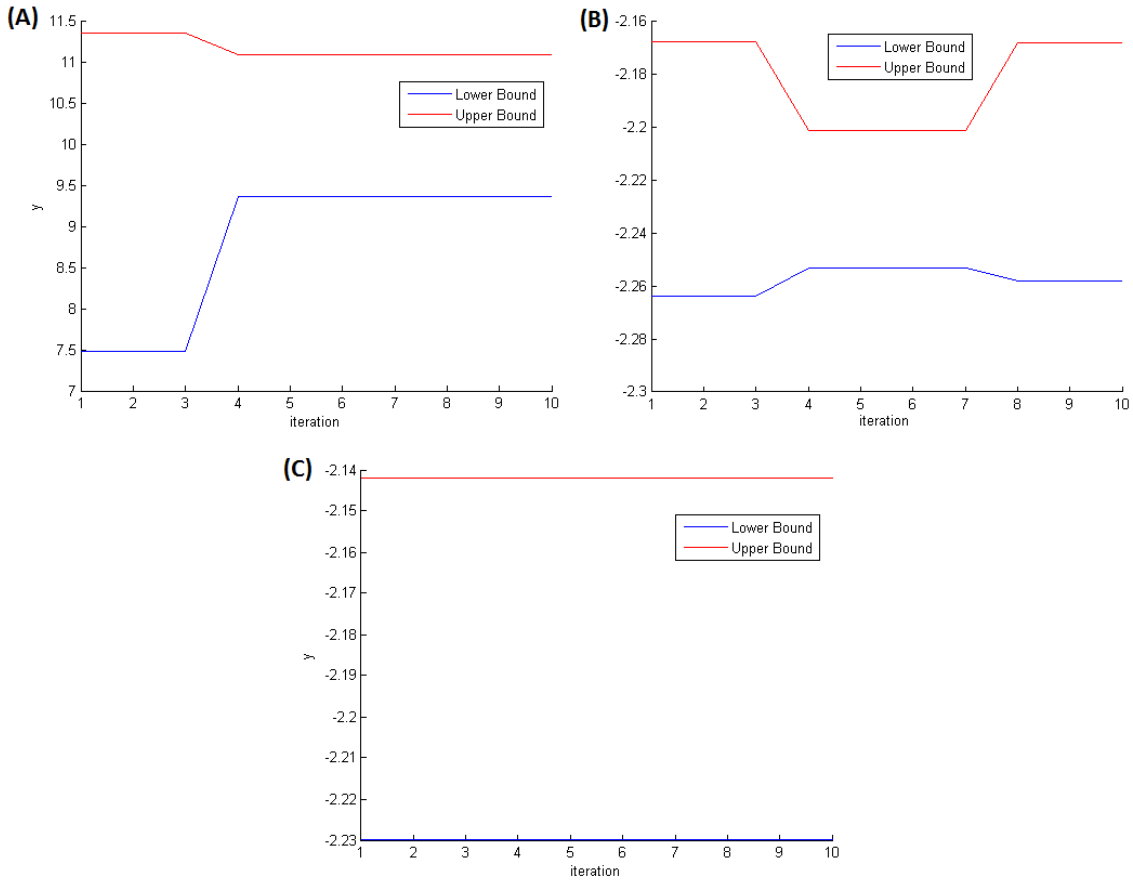

Figure 4.6: Volume maintained and pruned by PBnB approximating the level set bounded by the 0.1 quantile for two dimensional (A) Rosenbrock's function, (B) centered sinusoidal function, and (C) shifted sinusoidal function.

Table 4.3: Comparing the iteration for first maintaining a subregion for Rosenbrock, centered and shifted sinusoidal functions, with number of samples and iteration of first maintained subregion.

Test function	Rosenbrock		Centered Sinusoidal		Shifted Sinusoidal	
Dimension	# sampled	Iteration	# sampled	Iteration	# sampled	Iteration
$n = 2$	5,707	9	11,009	9	1,972	7
$n = 3$	25,531	10	33,315	10	3,057	7
$n = 5$	101,880	11	473,366	13	10,970	8
$n = 7$	1,228,908	14	5,589,142	16	20,962	9
$n = 10$	92,448,129	20	229,773,645	21	108,073	11

of maintaining a subregion is related to the shape of the target level set of the function.

Continuous Test Functions with Normal Noise

In order to illustrate the impact of noise, we apply $N(0, 1)$ noise to each test function. Figure 4.7 illustrates the pruned and maintained subregions in a two-dimensional solution space on the tenth iteration. The undetermined subregions for the noisy functions in Figure 4.7 are larger than the non-noisy counterparts in Figure 4.4. Although the shifted function performed well in the non-noisy condition, PBnB for level set approximation only captures a small part of the shifted sinusoidal function's target level set with noise, because $N(0, 1)$ noise is relatively large for the function value of the sinusoidal function. However, it is possible to maintain more subregions with further iterations that have smaller partitioned subregions and larger number of sample points.

Integer Test Functions

The PBnB algorithm for level set approximation can handle both integer and real-valued decision variables. The partitioning scheme on a discrete set must be adapted so that any discrete point belongs to only one subregion. The experiment in this section discretized the two dimensional centered sinusoidal function at two levels of discretization. The centered sinusoidal function is motivated by an optimal composite structure design problem

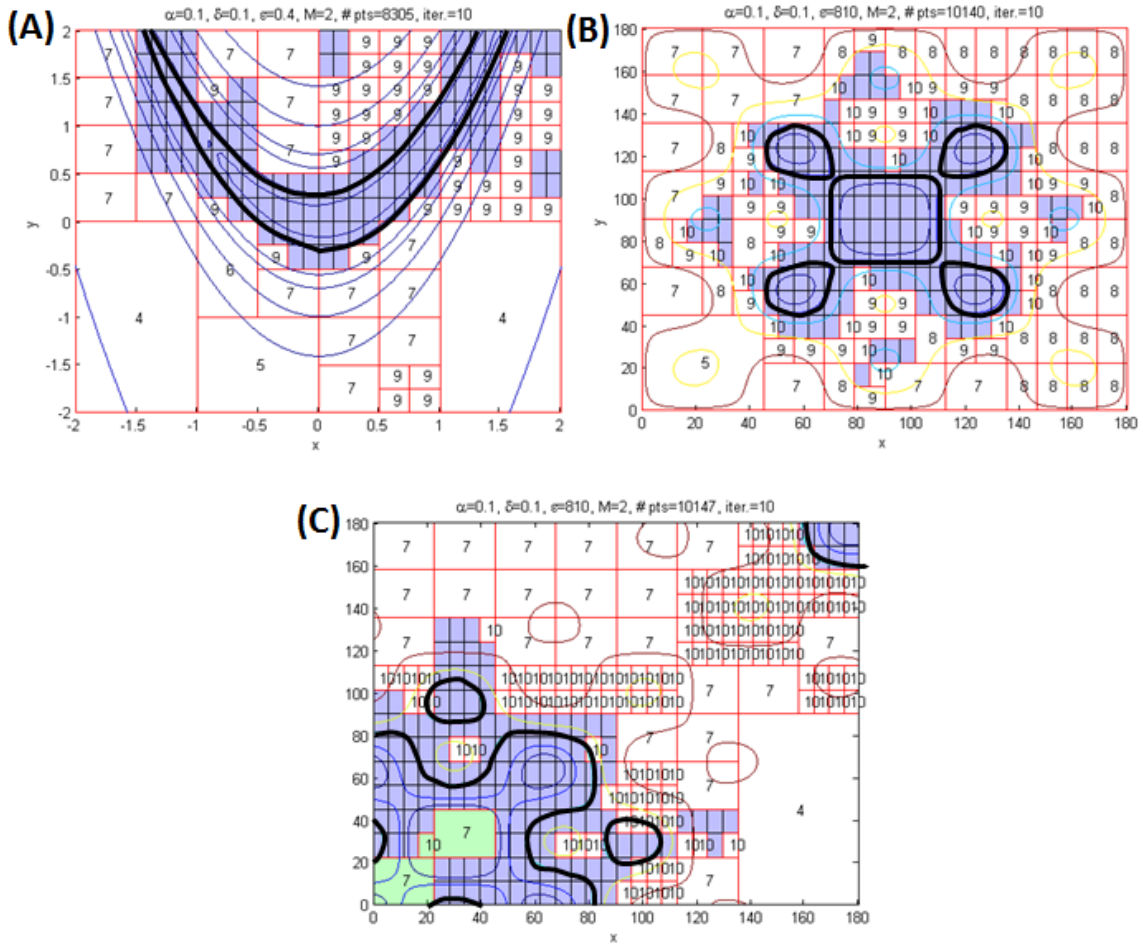

Figure 4.7: Approximating the 0.1 level set on the tenth iteration of PBnB for two dimensional (A) Rosenbrock function, (B) centered sinusoidal function, and (C) shifted sinusoidal function with N(0,1) noise by the maintained green (light gray) subregions.

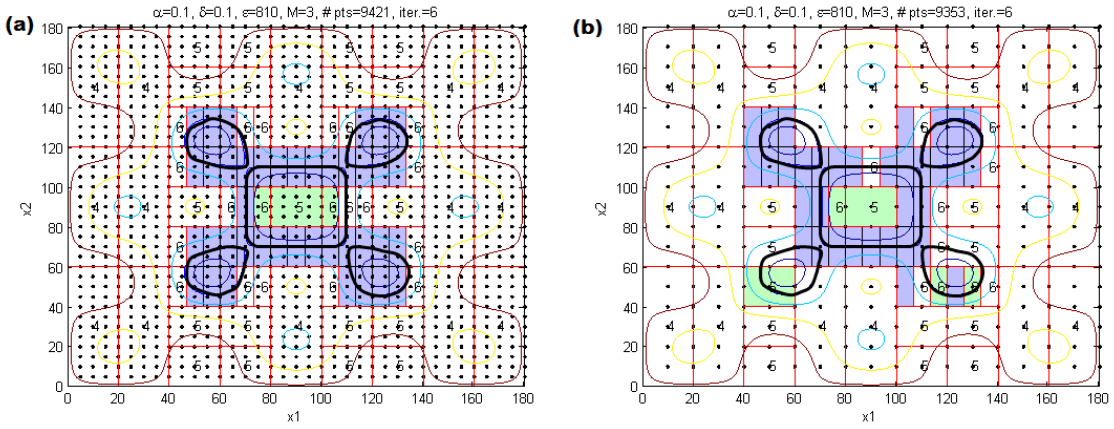

Figure 4.8: Results of discretized sinusoidal function on the sixth iteration of PBnB with (a) 5 differences between points and (b) 10 differences between points.

[5, 93] that is interested in different levels of discretization. Specifically, we consider two discretization scenarios; the difference between points is 5 and 10 (as shown in Figure 4.8). For the discretized sinusoidal function, we perform the same PBnB for level set approximation with a modified definition of how to branch the subregions in order to ensure that any discrete point belongs to exactly one subregion. The subregion in this discrete setting is $(l, u]$, where l and u are the lower and upper points of the box. Based on this definition, the results for two dimensional centered sinusoidal function are shown in Figure 4.8. Although the maintained subregions in Figure 4.8 may seem to contain some regions outside of the level set, the discrete points in the maintained subregions actually are contained in the level set. Therefore, the PBnB algorithm is still viable for discretized problems. However, the partitioning scheme needs to be designed carefully to ensure one discrete point belongs to one subregion.

4.2 Multiple Objective Probabilistic Branch and Bound

4.2.1 Introduction

Decision makers are commonly faced with multiple objectives when evaluating real world systems [51]. Instead of aggregating the multiple objectives to a single objective with

weights, decision makers want to understand the trade-offs between different objectives. The Pareto optimal set provides a family of solutions with objective function values that are not dominated by other solutions. The set of Pareto optimal solutions provides important information that allows the decision makers to determine the final decision. There are many multiple objective optimization studies [13], especially using evolutionary algorithms such as the genetic algorithm [28, 51], but only limited studies used these algorithms to develop interactive simulation optimization algorithms [85]. Some of the multiple objective simulation optimization algorithms were meta-heuristic based [61]. Meta-heuristic based multiple objective algorithms could be applied to large scale problems and different types of variables depending on the meta-heuristic algorithm. However, meta-heuristic based multiple objective algorithms were mostly numerically tested without statistical analysis. For our proposed algorithm, we derive probability bounds on the quality of solutions. Our algorithm is designed to provide a set of solutions that is defined by hyper-rectangles in the solution space, in order to provide more insights than obtaining only a few sampled solutions.

Another group of studies, categorized as ranking and selection, can efficiently identify the non-dominated solutions in a finite set [55, 56, 57]. The multiple objective ranking and selection algorithms provided rigorous statistical quality of the solutions by the probability of correct selection. Ranking and selection is an effective approach when the number of alternatives is relatively small. However, it is not efficient when there are continuous variables or a very large number of alternatives. Therefore, our proposed algorithm focuses on mixed-integer/continuous problems.

The probabilistic branch and bound (PBnB) algorithm was first presented in the Winter Simulation Conference 2011 [94], and extended to approximate a desired level-set for single objective problems [46]. The algorithm presented here extends PBnB by considering multiple objectives and approximating the set of Pareto optimal solutions. Hence, we refer to the new algorithm as multiple objective probabilistic branch and bound (MOPBnB).

A simulation often has several performance measures that are output from a single run. Our algorithm considers a black box simulation that provides m measures as multiple objectives. When we specify r replications, we are intending that the simulation will be

repeated r times, and the simulation provides m measures for each replication.

In Section 4.2.2, we describe the details of MOPBnB. In Section 4.2.3, we derive probability bounds for the performance of the algorithm. Several numerical results are presented in Section 4.2.4.

4.2.2 Details of MOPBnB

The goal of the algorithm is to find an approximation to the efficient frontier in the objective space and the associated Pareto optimal set in the solution space. The multiple objective optimization problem is in the following form:

$$(\mathcal{P}) \ \min_{x \in S} \{f_1(x), \ldots, f_m(x)\}, \tag{4.54}$$

where the objective functions can only be estimated by $f_l(x) = E_\Xi[g_l(x, \xi_x)], l = 1, \ldots, m$ and $g_l(x, \xi_x)$ is a noisy function and ξ_x is the noise term. We assume the distribution of the noise term ξ_x may be different for each x, but is normally distributed. The feasible set S is assumed to be a hyper-rectangle in n dimensions, to simplify the partition scheme. The variables may be real-valued or integer valued in the hyper-rectangle, with upper and lower bounds, $l_i \leq x_i \leq u_i$ for a continuous variable, and $x_i \in \{l_i, l_i + 1, \ldots, u_i\}$ for an integer variable.

In a single objective problem, we can define the quantile of the objective, as in [94] and [46], to represent the desirable performance. However, with multiple objectives, the definition of the desirable performance is more difficult to characterize. Therefore, we define an indicator $D(x)$, the minimum distance to the true efficient frontier in the objective space, for every x as follows,

$$D(x) = \inf\{||f(x) - f(x^*)|| \ | x^* \in S_e, f(x^*) \text{ dominates } f(x)\}, \tag{4.55}$$

where S_e is the Pareto set and $||f(x) - f(x^*)||$ is the Euclidean distance in the objective space. Other distance measures besides Euclidean distance may be used in the definition of $D(x)$ and the theoretical analysis still holds. Recognizing that $D(x)$ is not able to be evaluated, we only use it in the analysis to define the desirable performance.

We use $D(x)$ to define a target set $L(\delta, S)$, which can be interpreted as a relaxation of the true Pareto optimal set, for $0 < \delta < 1$. The target set $L(\delta, S)$ is defined in terms of the associated quantile $y(\delta, S)$. We also use $y(p, \sigma)$ and $L(p, \sigma)$ in the analysis of the algorithm for $0 < p < 1$ and $\sigma \subseteq S$. Let $y(p, \sigma)$ be the p quantile of value $D(x)$ where $x \in \sigma$, defined as

$$y(p, \sigma) = \underset{y \in \{D(x): x \in \sigma\}}{\arg\min} \{P(D(X) \leq y | X \in \sigma) \geq p\}, \text{ for } 0 < p < 1, \tag{4.56}$$

where X is uniformly distributed in σ. Furthermore, let $L(p, \sigma)$ be a set of best p solutions, defined as

$$L(p, \sigma) = \{x \in \sigma : D(x) \leq y(p, \sigma)\}, \text{ for } 0 < p < 1. \tag{4.57}$$

MOPBnB provides two types of results. First, the set of non-dominated points that have been sampled from S by the algorithm and their associated estimated objective function values are provided and stored in NS_k at iteration k. Second, a collection of hyper-rectangles that approximates the Pareto optimal set is provided, and called the retained region, denoted Σ_k, at iteration k. In Section 3, we derive probability bounds to represent the quality of the solutions. Next, we present the MOPBnB algorithm, incorporating the sample size from [94] and the ordering method from [46].

Multiple Objective Probabilistic Branch and Bound (MOPBnB)

Step 0. Initialize: Set user-defined parameters $\delta, \alpha, R^o, \epsilon$, and $B \geq 2$. Partition S into B subregions, $\sigma_1, \ldots, \sigma_B$, and set $\Sigma_1 = \{\sigma_1, \ldots, \sigma_B\}$, $\widetilde{\Sigma}_1 = S$, $\alpha_1 = \frac{\alpha}{B}$, $NS_k = \phi$, $R_0 = R^o$, and $k = 1$.

Step 1. Sample: For each subregion $\sigma_i \in \Sigma_k$, $i = 1, \ldots, ||\Sigma_k||$, uniformly sample additional points such that the total number of points in σ_i is

$$N_k = \left\lceil \frac{\ln \alpha_k}{\ln (1 - \delta)} \right\rceil. \tag{4.58}$$

Denote the sample points in σ_i by $x_{i,j}$, for $j = 1, \ldots, N_k$ and $i = 1, \ldots, ||\Sigma_k||$. For each $x_{i,j} \in \sigma_i$, $j = 1, \ldots, N_k$ and $i = 1, \ldots, ||\Sigma_k||$, perform R_{k-1} replications to obtain

$g_l(x, \xi_x^r)$ for $l = 1, \ldots, m$, $r = 1, \ldots, R_{k-1}$, and evaluate the sample mean and sample variance for each objective l,

$$\hat{f}_l(x_{i,j}) = \frac{\sum_{r=1}^{R_{k-1}} g_l(x_{i,j}, \xi_x^r)}{R_{k-1}} \text{ and}$$
$$S^2_{\hat{f}_l}(x_{i,j}) = \frac{1}{(R_{k-1}-1)} \sum_{r=1}^{R_{k-1}} (g_l(x_{i,j}, \xi_x^r) - \hat{f}_n(x_{i,j}))^2 \text{for } l = 1, \ldots, m.$$

Step 2. Calculate additional replications: For $j = 1, \ldots, N_k$ and $i = 1, \ldots, ||\Sigma_k||$ order each objective function l for $l = 1, \ldots, m$ so that

$$\hat{f}_l(z_{(1)}) \leq \hat{f}_l(z_{(2)}) \leq \cdots \leq \hat{f}_l(z_{(||\Sigma_k||N_k)}), \text{ where } z_{(q)} = \underset{x_{i,j} \in \widetilde{\Sigma}_k \setminus \bigcup_{l=1}^{q-1} z_{(l)}}{\arg\min} \hat{f}_l(x_{i,j}).$$

Let $d_{i,l} = \hat{f}_l(z_{(i+1)}) - \hat{f}_l(z_{(i)})$, where $i = 1, \ldots, N_k - 1$. Determine

$$d^* = \min_{i=1,\ldots,N_k-1, l=1,\ldots,m} d_{i,l} \text{ and } S^{*2} = \max_{i=1,\ldots,N_k, l=1,\ldots,m} S^2_{\hat{f}_l}(z_{(i)}).$$

Calculate the updated replication number $R_k = \max\left\{R_{k-1}, \left(\frac{\tilde{z}(\alpha_k/2)S^*}{d^*/2}\right)^2\right\}$, where $\tilde{z}(\alpha_k/2)$ is the $1 - \alpha_k/2$ quantile of the standard normal distribution. Perform $R_k - R_{k-1}$ more replications for each sample point. Re-estimate the performance of objective l of each sample point with R_k replications by $\hat{f}_l(x_{i,j}) = \frac{\sum_{r=1}^{R_k} g_l(x_{i,j}, \xi^r_{x_{i,j}})}{R_k}$.

Step 3. Identify current non-dominated set: Implement an existing comparison algorithm [54] to find the non-dominated sample points by $\hat{f}_l(x_{i,j})$ and update NS_k with all non-dominated points and their estimated function values $\hat{f}_l$.

Step 4. Update: For each i, $i = 1, \ldots, ||\Sigma_k||$, update the pruning indicator functions P_i,

$$P_i = \begin{cases} 1, \text{ if } \sigma_i \cap NS_k = \phi \\ 0, \text{ otherwise.} \end{cases} \tag{4.59}$$

Also, set $\alpha_{k+1} = \frac{\alpha_k}{B}$, $\epsilon_{k+1} = \frac{\epsilon_k}{B}$. If $P_i = 0$, and if σ_i is branchable, then partition σ_i to $\bar{\sigma}_i^1, \ldots, \bar{\sigma}_i^B$ and update the current set of subregions

$$\Sigma_{k+1} = \{\bar{\sigma}_i^j : P_i = 0, j = 1, \ldots, B\} \text{ and } \widetilde{\Sigma}_{k+1} = \bigcup_{i:P_i=0} \left(\bigcup_{j=1}^{B} \bar{\sigma}_i^j \right).$$

A subregion is unbranchable when the longest Euclidean distance of a subregion is less than ϵ.

Step 5. Stopping Condition: If all subregions $\sigma_i \in \Sigma_k$ are not branchable, terminate the algorithm.
Otherwise, $k = k + 1$ and return to **Step 1.**

Users need to determine several parameters: δ, $0 < \delta < 1$, the desired level of quantile; α, $0 < \delta < 1$ to determine the probability bounds involving $1 - \alpha$ on the quality of solutions; R^o as a minimal number of replications, $R^o \geq 1$; $\epsilon > 0$, to determine the longest Euclidian distance of a subregion that is considered unbranchable, and B, $B \geq 2$ to determine the partition scheme.

In Step 1 and Step 2, MOPBnB samples uniformly in the current subregions and orders the samples with their estimated function values. A two-stage procedure is applied that is based on a method from [9] and [40] to rank the samples correctly under noise. The two-stage procedure to determine the replications is extended from [46] for multiple objectives. The first-stage replications R_{k-1} are evaluated in Step 1. The original ranking procedure assumes that there is a common variance σ^2 but we do not. Step 2 uses the first-stage replications to estimate these differences, d_i for $i = 1, \ldots, N_k - 1$. We use the maximum sample variance to update the replication number for the second-stage. Consequently, additional replications may be necessary.

In Step 3, the algorithm applies any existing method to compare the m objective function estimates for all samples in the retained subregions to eliminate dominated solutions. In this section, we use the widely used Kung's algorithm [54]. The method will find the non-dominated samples from the retained subregions.

In Step 4, the algorithm updates each subregion to be pruned or further branched into smaller subregions. For each subregion σ_i, the dimension with longest length is chosen for partitioning. For a continuous variable, the length is defined as the difference of the bounds. For a discrete variable, the length is defined as the number of elements. For a continuous variable, partition the associated dimension into even sized B subregions. For a dimension

associated with a discrete variable, if it is not possible to partition evenly, then partition into unequal sized subregions ordered by size. For example, suppose the hyper-rectangle for σ_i in two dimensions has a continuous variable x_1, $x_1 \in [1, 10]$, and an integer variable x_2, $x_2 \in \{1, 2, 3, 4, 5\}$. The length of σ_i on dimension 1 is nine, and on dimension 2 is five, so dimension 1 will be partitioned. And for $B = 2$, the resulting partition on dimension 1 is $[1, 5.5]$ and $[5.5, 10]$. If dimension 2 is partitioned with $B = 2$, the result is $\{1, 2\}$ and $\{3, 4, 5\}$.

MOPnB follows the same stopping condition as [94] and [46]. This stopping condition assures the algorithm will terminate in a finite number of iterations, however users may wish to terminate earlier.

4.2.3 Performance Analysis of MOPBnB

In this section, we analyze the performance of MOPBnB by deriving probability bounds to indicate the quality of solutions. First, we consider the deterministic case without the noise term. Theorem 12 is analogous to Lemma 1 in [94] where the single objective function is replaced with the distance function $D(x)$, representing the closeness of the estimated function values of x to the true efficient frontier. The extension is allowed because our use of order statistics has minimal assumptions on the function. In Theorem 13, we derive the quality of the retained region Σ_k at iteration k with no noise. The impact of the noisy function is discussed in Theorems 14 and 8. These are used in Corollary 9 and Theorem 17.

The main results providing probability bounds on the quality of solutions are stated in Corollary 9 and Theorem 17. Corollary 9 is the noisy version of Theorem 13 and gives probability bounds that we correctly retain subregions in our approximation of the Pareto optimal set because they intersect the true Pareto optimal set. Theorem 17 shows that the concentration of points with objective function values close to the true efficient frontier increases with iteration, with probability bounds.

Theorem 12. *(c.f. [94]) Let $y(\delta, \sigma_m)$ be the δ quantile as defined in (4.56) for some subregion $\sigma_m \in \Sigma_k$ and $0 < \delta < 1$. Suppose $N^* = \left\lceil \frac{\ln \alpha_k}{\ln(1-\delta)} \right\rceil$ and $X_{m,1}, \ldots, X_{m,N^*}$ are sampled*

i.i.d. uniformly from σ_m. *Let* $D(X_{m,(1)})$ *be the minimum of* $D(X_{m,1}), \ldots, D(X_{m,N^*})$, *then*

$$P(D(X_{m,(1)}) < y(\delta, \sigma_m)) \geq 1 - \alpha. \tag{4.60}$$

Using (4.60), we derive the quality of the approximate Pareto optimal set Σ_k in Theorem 13.

Theorem 13. *Suppose MOPBnB has progressed to the current* k^{th} *iteration on Problem* $(\mathcal{P})$ *where there is no noise, with* $\alpha_1 < \alpha$. *At iteration* k *of MOPBnB, let the number of sample points be given as in Step 1,*

$$N_k = \left\lceil \frac{\ln \alpha_k}{\ln (1 - \delta)} \right\rceil, \tag{4.61}$$

and let σ_p^k *denote the pruned region at iteration* k, *then*

$$P\left(\left(S \setminus \bigcup_{h=1}^{k} \sigma_p^h \right) \bigcap L(\delta, S) \neq \phi \right) \geq 1 - \alpha. \tag{4.62}$$

Proof. Initially, PBnB branches S into M subregions $\sigma_1, \ldots, \sigma_M$ and samples $N_1 = \left\lceil \frac{\ln \alpha_1}{\ln (1-\delta)} \right\rceil$ on each subregion. Therefore, the total number of sample points from all subregions at first iteration is $M \left\lceil \frac{\ln \alpha_1}{\ln (1-\delta)} \right\rceil$. Within the $M \left\lceil \frac{\ln \alpha_1}{\ln (1-\delta)} \right\rceil$ samples, there is a sample, $X_{S,(1)}$, which performs best with regard to our distance metric, $D(x)$. Note

$$M \left\lceil \frac{\ln \alpha_1}{\ln (1 - \delta)} \right\rceil \geq M \left\lceil \frac{\ln \alpha}{\ln (1 - \delta)} \right\rceil \geq \left\lceil \frac{\ln \alpha}{\ln (1 - \delta)} \right\rceil, \tag{4.63}$$

thus we know

$$P(D(X_{S,(1)}) < y(\delta, S)) \geq 1 - \alpha \tag{4.64}$$

by applying Theorem 12 to S. This indicates that $L(\delta, S)$ is sampled by at least one of the sample points in the first iteration with $1 - \alpha$ probability. Furthermore, later iterations could have sampled points that dominate the best point of the first iteration, but the non-dominated points will also be in $L(\delta, S)$ with at least $1 - \alpha$ probability. Also, non-dominated points will not be pruned by the algorithm. Hence, the intersection of the retaining subregions and $L(\delta, S)$ is non-empty with probability at least $1 - \alpha$, and the proof is completed. □

The previous analysis assumes the objective functions $f_l(x)$ are not noisy. In Theorem 14, we consider the objectives that can only be estimated by $\hat{f}_l(x)$. Therefore, in order to achieve the quality in Theorem 13, we seek a correct ordering of the different objective function values to eliminate the dominated solutions, and subregions. Theorem 14 provides rigorous probability bounds for correctly ordering at iteration k, where Theorem 8 combines all iterations from 1 to k and gives a probability bound of $1-\alpha$ on the correct ordering.

We use the analysis of a two-stage replication approach, by Bechhofer et al. [9], in Theorem 14, which has following assumption:

(A1) The noisy function is normally distributed with an unknown common variance σ^2, and at each solution $z_j \in S$, the variance can be expressed as $V(z_j) = a_j\sigma^2$ where a_j is a known constant for each j.

In MOPBnB, the constants a_j are not known, hence, we implement a modified two-stage replication approach (described after Theorem 14, c.f. [46]).

Theorem 14. *(Based on [9]) With Assumption (A1), the probability of correct ordering all samples in the current region at iteration k is*

$$P(\hat{f}_l(Z^k_{(1)}) \leq \hat{f}_l(Z^k_{(2)}) \leq \cdots \leq \hat{f}_l(Z^k_{(N_k)}) | f_l(Z^k_{(1)}) \leq f_l(Z^k_{(2)}) \leq \cdots \leq f_l(Z^k_{(N_k)})) \geq 1-\alpha_k, \tag{4.65}$$

given that we have a_jR_0 as the first stage replication number for each sample point to estimate the common variance by $S_0^2 = \frac{1}{N_k}\sum_{i=1}^{N_k}\frac{1}{a_j}S^2_{\hat{f}}(z_{(j)})$ and set up the minimum difference desired to be separated as d^*, and sampling $R_k = \max\{a_jR_0, 2(\frac{hS_0^*}{d^*/2})^2\}$ in the second stage of the procedure, where h is a value in the H c.d.f. of a multivariate student t's distribution with $N_k - 1$ dimensions such that $H(h) = 1-\alpha_k$.

The algorithm intends to find the non-dominated points by correctly ordering points by their objective function values. Although the objective function values could be correlated, the comparison of the function values is separated for each objective function. Therefore, the probability of ordering m objectives correctly for a current iterations is greater than the probability of correct ordering of each objective separately, bounded by $(1-\alpha_k)^m \geq 1-m\alpha_k$. For any iteration k, we take a conservative approach (c.f. [46]) to achieve the $1-m\alpha_k$

probability of correct ordering by separating each estimated performance mean $\hat{f}(z_{(i)})$ with its neighbor by the smallest difference $d^* = \min_{i=1,\ldots,N_k-1,l=1,\ldots,n} d_{i,l}$ of any two neighbors. We also use the largest variance $S^{*2} = \max_{i=1,\ldots,N_k,l=1,\ldots,n} S^2_{\hat{f}_l}(z_{(i)})$ so that all ordering is conservative. The implemented two-stage procedure for any iteration k is

1. Implement R_{k-1} replications to estimate the function value $\hat{f}(x)$ and sample variance $S^2_{\hat{f}}(x)$ and calculate the differences $d_{i,l}$ between ordered samples $\hat{f}_l(z_{(i)})$ and $\hat{f}_l(z_{(i+1)})$ for $i = 1, \ldots, N_k - 1$ and $l = 1, \ldots, n$.
2. Implement additional replications $R_k - R_{k-1}$ where $R_k = \max\{R_{k-1}, (\frac{\tilde{z}(\alpha_k/2)S^*}{d^*/2})^2\}$.

The following Theorem 15 considers the assumptions and rigorous situation in Theorem 14 to derive a rigorous bound for iteration 1 to k.

Theorem 15. *With Assumption (A1), the probability of correct ordering from iteration* 1 *to iteration* k *is*

$$P\left(\bigcap_{q=1,\ldots,k,l=1,\ldots,m}\left(\hat{f}_l(Z^q_{(1)}) \le \hat{f}_l(Z^q_{(2)}) \le \cdots \le \hat{f}_l(Z^q_{(N_q)}) | f_l(Z^l_{(1)}) \le f_l(Z^q_{(2)}) \le \cdots \le f_l(Z^q_{(N_q)})\right)\right)$$
$$\ge 1 - m\alpha, \tag{4.66}$$

where $f_l(Z^q_{(j)})$ *is the* j^{th} *ordered sampled point at iteration* q *by objective* l.

Proof. The probability of ordering m objective functions correctly through iteration k is bounded by the individual probability of correct ordering for each iteration and each objective function, given by

$$P\left(\bigcap_{q=1,\ldots,k,l=1,\ldots,m}\left(\hat{f}_l(Z^q_{(1)}) \le \hat{f}_l(Z^q_{(2)}) \le \cdots \le \hat{f}_l(Z^q_{(N_q)}) | f_l(Z^q_{(1)}) \le f_l(Z^q_{(2)}) \le \cdots \le f_l(Z^q_{(N_q)})\right)\right)$$
$$\ge \prod_{q=1}^{k}\left(1 - m\frac{\alpha}{B^q}\right), \tag{4.67}$$

and then apply Bernoulli's inequality repeatedly to yield the final result. □

Theorem 13 holds in the noisy condition with (A1) assumption if the orderings of function values of each objective at iteration 1 to k are correct. Therefore, the quality of solution with noise is quantified as Corollary 16, which is the noisy version of Theorem 13.

Corollary 16. *Consider the K^{th} iteration of MOPBnB on $(\mathcal{P})$ where (A1) is assumed. The intersection of the approximating set and $L(\delta, S)$ is non-empty with*

$$P\left(\left(S \setminus \bigcup_{k=1}^{K} \sigma_p^k\right) \bigcap L(\delta, S) \neq \phi\right) \geq (1-\alpha)(1-m\alpha). \tag{4.68}$$

Under the noisy condition, Corollary 16 indicates that the intersection of the approximated Pareto optimal set Σ_k with the target set $L(\delta, S)$ is non-empty with at least $(1-\alpha)(1-m\alpha)$ probability.

In order to evaluate the incremental improvement of the algorithm, we consider a metric defined by Zabinsky et al. [94], the *concentration* of the retained unpruned region. The concentration considers the proportion of desirable region in the remaining unpruned region, and is defined as

$$\mathcal{C}_k = \frac{v(\{S \setminus \cup_{i=1}^{k} \sigma_p^i\} \cap L(\delta, S))}{v(S \setminus \cup_{i=1}^{k} \sigma_p^i)}, \tag{4.69}$$

where $v()$ is the volume of a region of solutions. When a dimension is associated with a discrete variable, the volume is calculated as counting the elements in the region.

Theorem 17. *Assuming the algorithm progresses to K^{th} iteration and the best distance of the sampled points has not reached the desirable set, i.e., $D(X_{NS_k,(1)}) > y(\delta, S)$ for $k = 1, \ldots, K-1$, then*

$$P(\mathcal{C}_{k+1} \geq \mathcal{C}_k) \geq (1-\alpha)(1-m\alpha). \tag{4.70}$$

Proof. (Based on [88]) For any $k = 1, \ldots, K-1$, we rewrite the concentration of desired solutions $\mathcal{C}_{k+1}$ after the iteration $k+1$ as follows,

$$\frac{v(\{S \setminus \cup_{i=1}^{k+1} \sigma_p^i\} \cap L(\delta, S))}{v(S \setminus \cup_{i=1}^{k+1} \sigma_p^i)} = \frac{v(\{S \setminus \cup_{i=1}^{k} \sigma_p^i\} \cap L(\delta, S)) - v(\sigma_p^{k+1} \cap L(\delta, S))}{v(S \setminus \cup_{i=1}^{k} \sigma_p^i) - v(\sigma_p^{k+1})} \equiv \frac{A-B}{C-D}. \tag{4.71}$$

Thus (4.70) is equivalent to $P\left(\frac{A-B}{C-D} \geq \frac{A}{C}\right) \geq (1-\alpha)(1-m\alpha)$. Note that with $A > B, C > D$ and $A, B, C, D > 0$,

$$\frac{A-B}{C-D} \geq \frac{A}{C} \Leftrightarrow AC - BC \geq AC - AD \Leftrightarrow AD \geq BC \Leftrightarrow \frac{A}{C} \geq \frac{B}{D}. \tag{4.72}$$

For each iteration, if $||\sigma_p^k|| = 0$, then the proof is trivial. Therefore, consider $||\sigma_p^k|| \geq 1$ First, consider the $\frac{B}{D}$ part. Let D_p^k be the best performance associated to a sample point from all points in the pruned regions at iteration k, and let $D_{NS,p}^k$ be the distance performance associated to a sample from the non-dominated set which dominates the sample point associated to D_p^k considering all m objectives. From Theorem 14, under (A1) assumption, we know that the order will be correct with $1 - \alpha_k$ probability for one objective. Hence, with $1 - m\alpha_k$ probability, $D_{NS,p}^k < D_p^k$. Let σ_p^k be partitioned into the following two disjoint sets

$$\left\{x \in \sigma_p^k : D(x) < D_p^k\right\} \text{and} \left\{x \in \sigma_p^k : D(x) \geq D_p^k\right\}.$$

Also, $D_{NS,p}^k \geq D(X_{NS,(1)}) > y(\delta, S)$ by the assumption. Therefore, there is $1 - m\alpha_k$ probability that

$$\left\{x \in \sigma_p^k : D(x) \geq D_p^k\right\} \bigcap L(\delta, S) = \phi.$$

Hence, with $1 - m\alpha_k$ probability, we can say

$$\begin{aligned} v\left(\sigma_p^k \bigcap L(\delta, S)\right) &= v\left(\left\{x \in \sigma_p^k : D(x) < D_p^k\right\} \bigcap L(\delta, S)\right) + v\left(\left\{x \in \sigma_p^k : D(x) \geq D_p^k\right\} \bigcap L(\delta, S)\right) \\ &= v\left(\left\{x \in \sigma_p^k : D(x) < D_p^k\right\} \bigcap L(\delta, S)\right) \\ &\leq v\left(\left\{x \in \sigma_p^k : D(x) < D_p^k\right\}\right). \end{aligned} \tag{4.73}$$

By Theorem 12, it holds that

$$P(D_p^k < y(\delta, \sigma_p^k)) \geq 1 - \alpha_k. \tag{4.74}$$

Combining (4.73) and (4.74), with at least $(1 - \alpha_k)(1 - m\alpha_k)$ probability

$$v\left(\sigma_p^k \bigcap L(\delta, S)\right) \leq v\left(\left\{x \in \sigma_p^k : D(x) < y(\delta, \sigma_p^k)\right\}\right). \tag{4.75}$$

By the definition of quantile, we have

$$\frac{v\left(\left\{x \in \sigma_p^k : D(x) < y(\delta, \sigma_p^k)\right\}\right)}{v\left(\sigma_p^k\right)} \leq \delta. \tag{4.76}$$

Combining (4.75) and (4.76), we have at least $(1 - \alpha_k)(1 - m\alpha_k)$ probability that

$$\frac{B}{D} = \frac{v\left(\sigma_p^k \bigcap L(\delta, S)\right)}{v\left(\sigma_p^k\right)} \leq \delta. \tag{4.77}$$

Replacing k with $k+1$ since (4.77) is for any iteration, we have

$$\frac{B}{D} = \frac{v\left(\sigma_p^{k+1} \bigcap L(\delta, S)\right)}{v\left(\sigma_p^{k+1}\right)} \leq \delta. \tag{4.78}$$

For the $\frac{A}{C}$ part,because $S \bigcap L(\delta, S) = L(\delta, S)$ and $v(L(\delta, S)) \geq \delta v(S)$ by quantile definition, we can derive

$$\begin{aligned} v\left(\left\{S \setminus \cup_{i=1}^{k} \sigma_p^i\right\} \bigcap L(\delta, S)\right) &= v\left(S \bigcap L(\delta, S)\right) - \sum_{i=1}^{k} v\left(\sigma_p^i \bigcap L(\delta, S)\right) \\ &\geq \delta v(S) - \sum_{i=1}^{k} v\left(\sigma_p^i \bigcap L(\delta, S)\right) \\ &\geq \delta v(S) - \delta v(\sigma_p^1) - \sum_{i=2}^{k} v\left(\sigma_p^i \bigcap L(\delta, S)\right) \end{aligned}$$

with at least $(1-\alpha_k)(1-m\alpha_k)$ probability by (4.77). Then repeating the step for $i = 2, \ldots, k$, with at least $\prod_{i=1}^{k}(1-\alpha_i)(1-m\alpha_i)$ probability, the previous statement becomes

$$\geq \delta v(S) - \delta \sum_{i=1}^{k} v(\sigma_p^i) = \delta v\left(S \setminus \cup_{i=1}^{k} \sigma_p^i\right). \tag{4.79}$$

Dividing both sides by the volume term on the right hand side, we have, with at least $\prod_{i=1}^{k} 1-\alpha_i)(1-m\alpha_i)$ probability,

$$\frac{A}{C} = \frac{v\left(\left\{S \setminus \cup_{i=1}^{k} \sigma_p^i\right\} \bigcap L(\delta, S)\right)}{v\left(S \setminus \cup_{i=1}^{k} \sigma_p^i\right)} \geq \delta. \tag{4.80}$$

Combining (4.78) and (4.80), we have that $\frac{A}{C} \geq \frac{B}{D}$ with at least $\prod_{i=1}^{k+1} 1(-\alpha_i)(1-m\alpha_i)$ probability, which is strictly greater than $(1-\alpha)(1-m\alpha)$ by Bernoulli's inequality and the proof of (4.70) is completed. □

4.2.4 Numerical Results

Two test functions, he Fonseca and Fleming function and the Kursawe function [28], are used in this section to illustrate the performance of MOPBnB, both test functions have two objective functions and are box-constrained. The Fonseca and Fleming function has a two-dimensional feasible region S, so we can visually illustrate the algorithm on the solution space. The Kursawe function is in three dimensional space, so we only illustrate the efficient frontier.

- Fonseca and Fleming function ($n = 2, -4 \leq x_i \leq 4, i = 1, \ldots, n$)

$$\text{Minimize} = \begin{cases} g_1(x) = 1 - \exp\left(-\sum_{i=1}^{n}\left(x_i - \frac{1}{\sqrt{n}}\right)^2\right) \\ g_2(x) = 1 - \exp\left(-\sum_{i=1}^{n}\left(x_i + \frac{1}{\sqrt{n}}\right)^2\right) \end{cases} \tag{4.81}$$

- Kursawe function ($n = 3$ and $-5 \leq x_i \leq 5, i = 1, \ldots, n$)

$$\text{Minimize} = \begin{cases} g_1(x) = \sum_{i=1}^{2}\left[-10\exp\left(-0.2\sqrt{x_i^2 + x_{i+1}^2}\right)\right] \\ g_2(x) = \sum_{i=1}^{3}\left[|x_i|^0.8 + 5\sin(x_i^3)\right] \end{cases} \tag{4.82}$$

The parameters were set as $\delta = 0.1$, $\alpha = 0.05$, and $B = 2$. The ϵ parameter was set at one percent of the longest Euclidean distance of S. For both test functions, we tested with no noise, $\xi_{x,l} = 0$, and with noise, $\xi_{x,l} = N(0, 0.3)$ for Fonseca and Fleming function and $\xi_{x,l} = N(0, 1)$ for Kursawe function. The noise is additive, $g_l(x) + \xi_{x,l}$. The variability of the noise was determined by considering of the range of the efficient frontier's function value.

Figure 4.9 illustrates the efficient frontier and the approximated Pareto optimal set for the Fonseca and Fleming function. Panels (A) and (C) of Figure 4.9 are with no noise, and panels (B) and (D) are with noise. Panels (A) and (B) illustrate the efficient frontier in the objective function space and panels (C) and (D) illustrate the approximated Pareto optimal set in the solution space. In panel (A) of Figure 4.9, the stars (blue) represent the empirically estimated true efficient frontier and were obtained by performing grid search for 100x100 points. The dots (red) are the estimated objective function values from NS_k in the MOPBnB algorithm at the final iteration ($k = 14$), there are 1,589 non-dominated points in NS_k, out of 11,643 total sampled points. As illustrated, our approximation of the efficient frontier shows excellent closeness and spreading. Panel (B) of Figure 4.9 includes noise, $\xi_{x,l} = N(0, 0.3)$. The stars (blue) are the same as in panel (A), and the dots (red) are the estimated objective function values $\hat{f}_l$ from the MOPBnB algorithm, under the noisy condition. It is interesting that the red dots appear "better" than the blue stars, so we also plotted the true function values with no noise, f_l (blue circles) associated with the non-dominated points. The blue circles are actually overlapping or very close to the true efficient frontier. Furthermore, in the noisy condition, there were 22 non-dominated points in NS_k, out of 8,742 total sampled points.

Panel (C) of Figure 4.9 illustrates the progress of the MOPBnB algorithm in the solution space, with the final subregions approximating the Pareto optimal set. Each rectangle represents a subregion in the algorithm, where the number in a subregion represents the iteration that the subregion is pruned. Similarly, panel (D) in Figure 4.9 shows the results of Fonseca and Fleming function with $\xi_{x,l} = N(0, 0.3)$.

For the Kursawe function, only the approximate efficient frontiers (no noise and $N(0, 1)$ noise) are shown in Figure 4.10, because the solution space in three dimensions is difficult to visualize. The approximated efficient frontier illustrates a good approximation to the true efficient frontier with regard to both closeness and spread. Similarly, the noise version in panel (B) of Figure 4.10 shows the bias of the estimated efficient frontier (red dots), but the corresponding true function values (blue circles) estimate the efficient frontier well. For no noise, the algorithm found 415 non-dominated solutions out of 31,104 points sampled, and, with noise, the algorithm found 56 non-dominated solutions from total 17,967 samples.

Both test functions show that the influence of noise decreases the spread and the size of the approximated Pareto optimal set. A possible reason is that the MOPBnB is currently designed by only considering the closeness of the solution. The sample size and pruning step do not consider the spread and the potential loss of Pareto optimal set. Hence, a good future research direction is to incorporate the spread and the size of approximated Pareto optimal set to the algorithm.

4.3 Summary and Discussion

We developed the PBnB algorithms to approximate a target set of solutions for stochastic global optimization problems. A set of solutions allows decision makers to have further understanding of their problem. The algorithms are capable to solve problems with mixed continuous and discrete variables. Two variations are developed, PBnB for level set approximation and MOPBnB.

PBnB for level set approximation iteratively partitions the solution space into subregions, and prunes or maintains subregions based on the interval estimation of the target quantile. With the dynamically allocated computational resources, the algorithm provides an interval estimation of the desired quantile $[\hat{f}(z_{((r)}), \hat{f}(z_{((s)})]$, a pruned set $\widetilde{\Sigma}^P$ and a main-

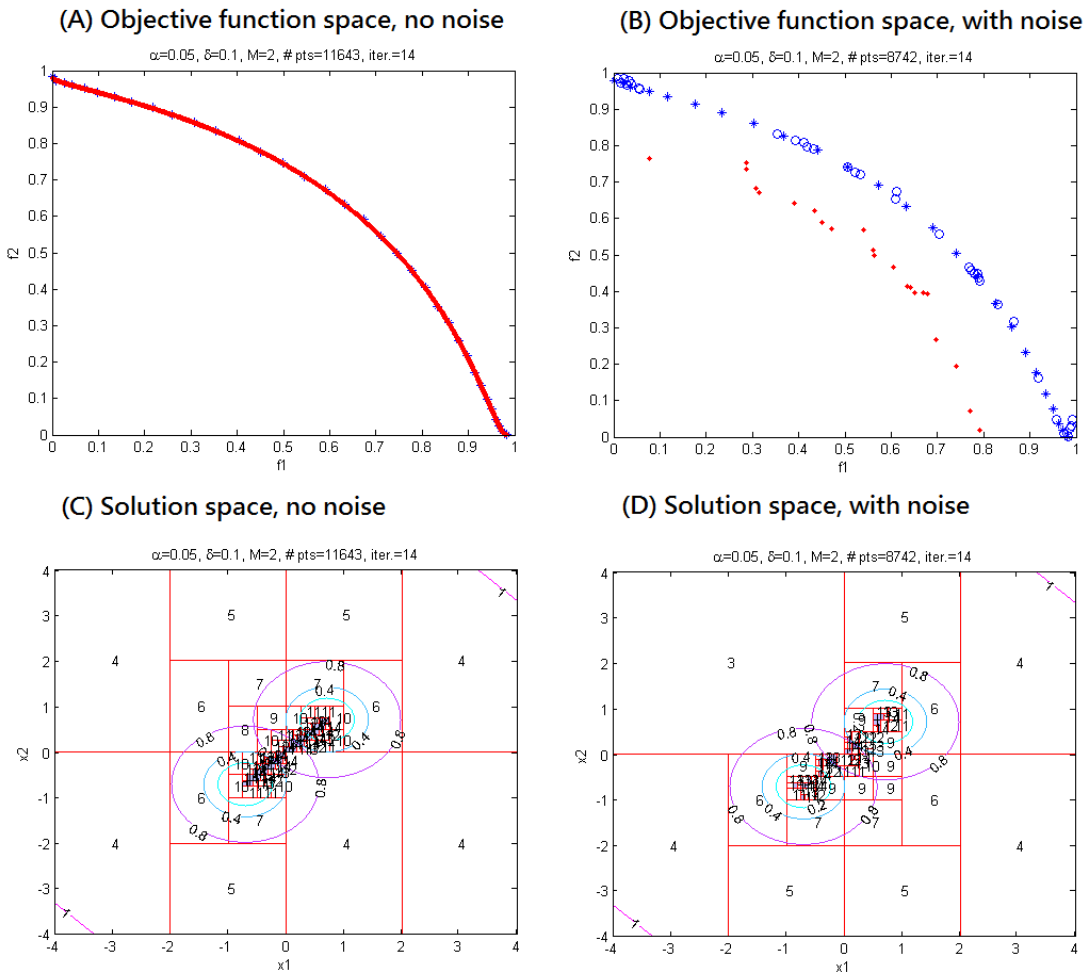

Figure 4.9: Fonseca and Fleming function. Panels (A) and (B) plot the two objective functions to illustrate the efficient frontier, (A) no noise, $\xi_x = 0$, and (B) with noise, $\xi_x = N(0, 1)$. Panels (C) and (D) plot the approximated Pareto optimal set, (C) no noise, $\xi_x = 0$, and (D) with noise, $\xi_x = N(0, 1)$.

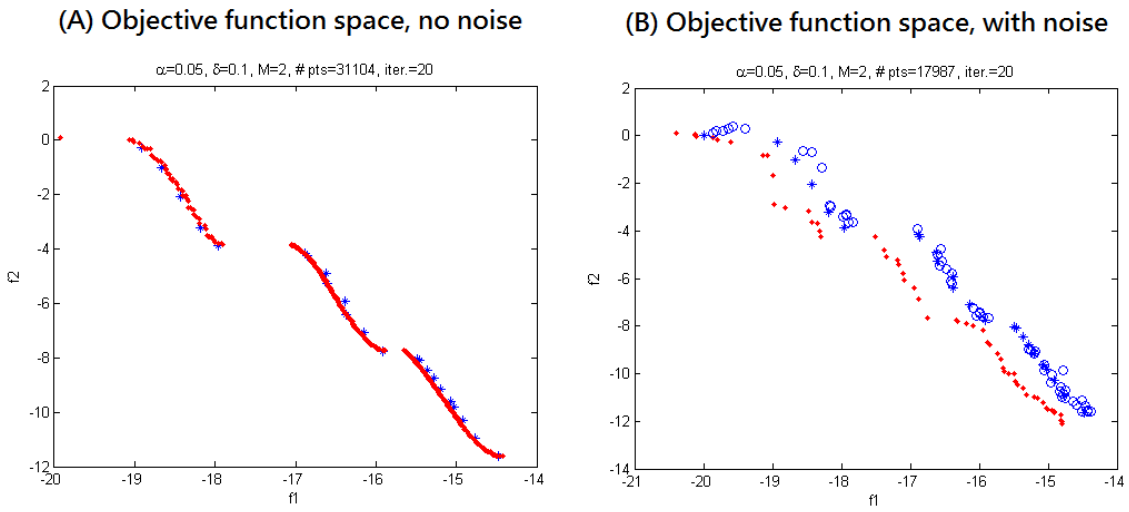

Figure 4.10: Kursawe function. Panels (A) and (B) plot the two objective functions to illustrate the efficient frontier, (A) no noise, $\xi_x = 0$ and (B) with noise, $\xi_x = N(0,1)$.

tained set $\widetilde{\Sigma}^M$ with probability bounds and tolerable loss. The analysis of the probability bounds are provided in Section 4.1.4, where Theorems 5 and 6 assume the objective function is deterministic, and Corollaries 10 and 11 involve normal noise to the objective function. The numerical results demonstrate the capability of PBnB to approximate a target level set, and the required sample size is less than performing a grid search especially when the dimension increases.

The MOPBnB algorithm is developed to provide the approximated efficient frontier by the objective function values of the sample points in NS_k in the objective space and the approximated Pareto optimal set Σ_k in the solution space for multiple objective stochastic global optimization problems. Similar to PBnB for level set approximation, the framework of MOPBnB is to partition the solution space iteratively and determine whether each sub-region should be pruned or further branched depending on the observed non-dominated samples. The algorithm provides the quality of the approximated Pareto optimal set Σ_k by considering the distance to the true efficient frontier in the objective space. We prove that, with probability $(1-\alpha)(1-m\alpha)$, the approximated Pareto optimal set intersects the relaxed Pareto set $L(\delta, S)$. We also prove that the concentration of "good" solutions in the

retained region increases with iteration.

Chapter 5

HEALTHCARE DECISION SUPPORT WITH PROBABILISTIC BRANCH AND BOUND

This chapter includes the two healthcare resource allocation applications that were modeled with simulation and optimized with PBnB. Section 5.1 describes a hepatitis C screening and treatment allocation strategy problem [50], which is modeled as a Markov decision process to evaluate a cohort's health utility gain [58]. We apply PBnB for level set approximation to provide not only a best strategy but also a set of strategies with statistical assurance of the performance to investigate further insights. In Section 5.2, a portable ultrasound machine allocation problem for orthopedic clinic diagnosis is modeled as a discrete-event simulation model [49]. Two performance metrics are considered in this problem: total system cost and health utility loss. In order to find the Pareto optimal set of this multiple objective problem, we use MOPBnB to approximate an Pareto optimal set and analyze the trade-offs by the estimated efficient frontier.

5.1 Hepatitis C Screening and Treatment Allocation Strategy

5.1.1 Introduction

Chronic hepatitis C virus (HCV) is an important public health problem in the U.S. [20, 62]. Li [58], Li et al. [59], and Li et al. [60] modeled the disease progression as a Markov model to evaluate health utility gain using quality-adjusted life years (QALYs) for different screening and treatment budget allocation by birth-cohort. Li [58] performed a grid search over ten years and found a near optimal strategy for each target cohort and total budget scenario. However, the results are empirically searched without statistical understanding of the solution's quality. More importantly, it does not provide further understanding on the impact of each decision period on the system performance. PBnB for level set approximation is the approach used to analyze this problem. A level set contains a set of viable solutions

rather than a single solution. Furthermore, it provides an understanding of the decision variables by providing ranges of values. For this HCV application, our approach provides an understanding of how the screening and treatment decisions by time period impacts the health utility gain.

5.1.2 Hepatitis C Markov Model

Li et al. [60] developed a compartmental simulation model using a Markov decision process to mimic the HCV transmission, progression, screening, treatment, and death in the health-care system. The model allocates budget shares across a ten year period, with a two-year budget planning cycle. The decision is the percentage of the budget to be used for screening and treatment in each of five two-year time periods, by HCV birth-cohort (born 1945-1975), to maximize health outcomes from a U.S. societal perspective.

An implementation strategy from Li and Li et al. [58, 60] is considered in this paper: allocate a percentage of the budget to screening, then treat patients with the remainder giving priority to the sickest patients. After the planning horizon, the model continues to simulate the progression of people in the cohort until end-of-life (additional 60 years).

The population in each cohort is categorized into three major groups: (A) HCV status is unknown, which is the target screening population, (B) HCV+, identified treatment candidates who are waiting for treatment, and (C) HCV-, individuals who are not HCV infected, determined through screening or cleared infections by successful treatments or spontaneous viral clearance.

The large boxes in Figure 5.1 represent the three population groups and illustrate how people transition between groups. Screening could identify the individuals in group (A), and move them to group (B) or group (C) depending on whether they test HCV positive or negative. Also, HCV+ individuals could be successfully treated and become HCV-. However, individuals in group (C), HCV-, could be reinfected and rejoin group (A). Figure 5.1 also describes the natural history model within the boxes, which is a Markov model representing the health state and lifelong disease progression of chronic HCV patients with possible transitions occurring every three months. Combining the transitions between groups and the

disease progression, a full Markov model (one for each birth-cohort) is developed, including transitions impacted by screening and treatment. The details of the model and parameters can be found in Li et al. [60].

Using the Markov model, Li et al. [60] calculated the number of people in each health state of each group. Li et al. [60] evaluate each strategy by the total discounted health utility gain (QALYs),

$$\sum_t \sum_g \sum_i utility^i (A^i_{t,g} + B^i_{t,g} + C^i_{t,g}) \cdot (1 - \delta_q)^{-t} \tag{5.1}$$

where t is the discrete time period in quarters (3 month period), g represents gender, and i represents health state. The population groups $A^i_{t,g}$, $B^i_{t,g}$ and $C^i_{t,g}$ are the states of the Markov chain, and represent the number of individuals in the birth-cohort in gender g and health state i at time t. The $utility^i$ function maps the states at time t to a health utility gain, and δ_q is the discount factor for each quarter time period.

The decision variables include $S_{t,g}$ and $T^i_{t,g}$, where $S_{t,g}$ is the fraction of candidates in group A and gender g during time period t who are offered HCV screening tests, and $T^i_{t,g}$ is the fraction of patients in group B and gender g and health state i during time period t who are treated. The fraction of candidates is based on the budget allocation for each intervention and the implementation policy.

The progression of the disease in the Markov model, and how $A^i_{t,g}$, $B^i_{t,g}$ and $C^i_{t,g}$ are updated is based on Li et al. [60]. For group A, the number of individuals for each health state i on the next time period is based on the unscreened individuals from the last period and reinfected candidates from the last period. Specifically,

$$A^i_{t+1,g} = \sum_j \theta^{ji}_{g,waiting}(1 - \alpha S_{t,g}\beta)A^j_{t,g} + C^i_{t,g}\gamma, \tag{5.2}$$

where $\theta^{ji}_{g,waiting}$ is the health states transition probability from health state i to j for untreated individuals. The coefficients, α and β, represent the proportion of the cohort that goes to a healthcare provider and accept screening, respectively. The coefficient γ represents the proportion of the cohort that may become reinfected. Similarly, the number of individuals in group B for the next time period is formed with new patients identified through

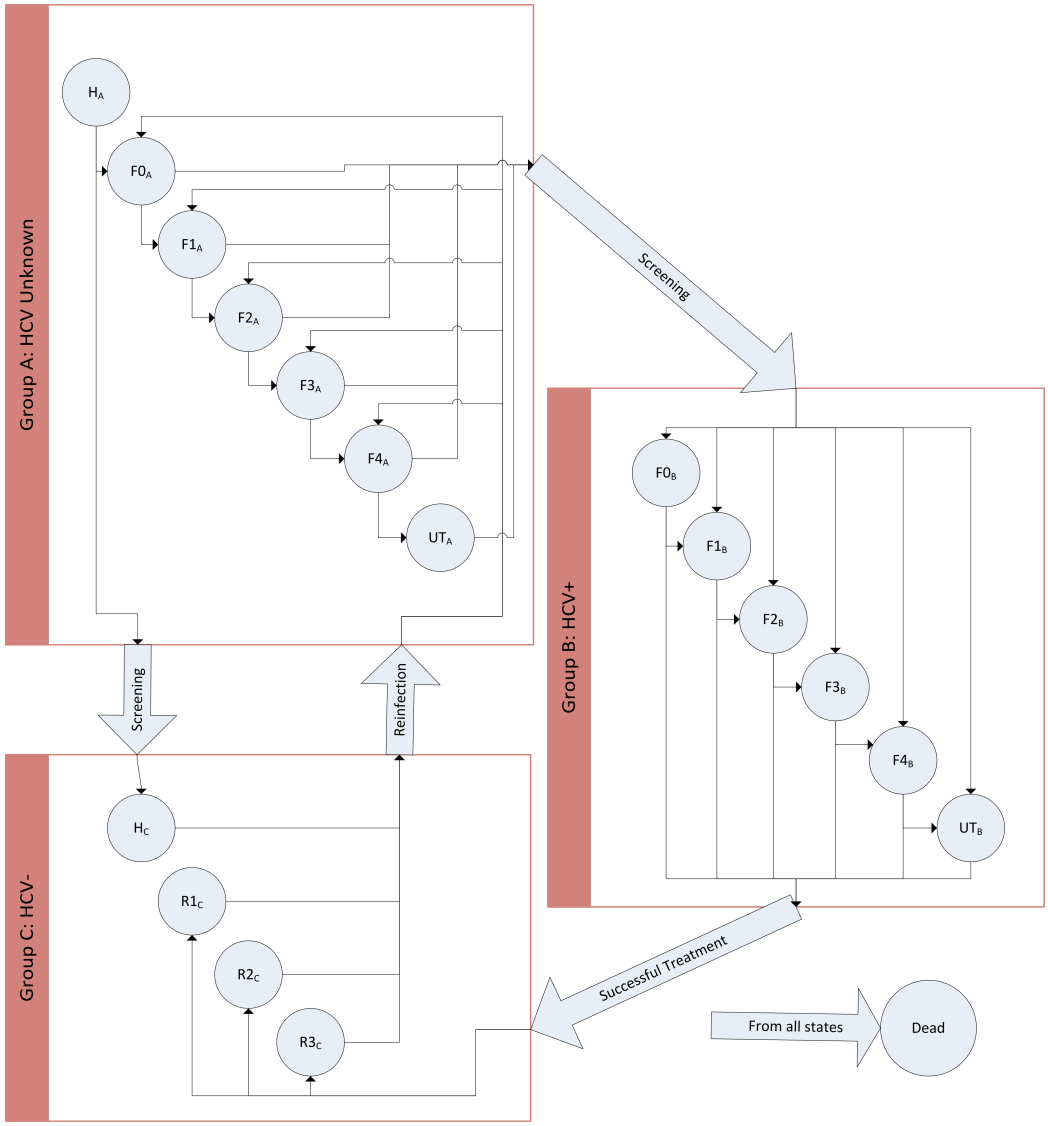

Figure 5.1: Hepatitis C Markov model, combining three groups, (A) HCV status unknown, (B) HCV+ and (C) HCV-, and the natural history model. The natural history model has the following states: (H) Healthy, (F0) no fibrosis, (F1) portal fibrosis with no septa, (F2) portal fibrosis with few septa, (F3) numerous septa without cirrhosis, (F4) compensated cirrhosis, (UT) untreatable, (R1) recovered with history of mild fibrosis, (R2) recovered with history of moderate fibrosis, (R3) recovered with history of advanced fibrosis, and Dead state. (F0) and (F1) are considered mild fibrosis, (F2) and (F3) are moderate fibrosis, and (F4) is advanced fibrosis.

screening, patients waiting for treatment, and patients with unsuccessful treatment. The function is described as follows,

$$B^i_{t+1,g} = \sum_j \theta^{ji}_{g,waiting} \alpha S_{t,g} \beta A^j_{t,g} + \sum_j \theta^{ji}_{g,waiting}(1 - T_{t,g}) B^j_{t,g} + \sum_j \theta^{ji}_{g,ongoing} T_{t,g} B^j_{t,g} \quad (5.3)$$

and $\theta^{ji}_{g,ongoing}$ is the transition probability from health state i to j for individuals undergoing treatment. The number of individuals in group C for the next time period is related to four sources: individuals screened with negative result, spontaneous viral clearance patients, successfully treated patients, and original C group individuals without reinfection. The equation is

$$\begin{aligned} C^i_{t+1,g} = & \sum_j \theta^{ji}_{g,waiting} \alpha S_{t,g} \beta A^j_{t,g} + \sum_j \theta^{ji}_{g,waiting}(1 - T_{t,g}) B^j_{t,g} \\ & + \sum_j \theta^{ji}_{g,ongoing} T_{t,g} B^j_{t,g} + C^i_{t,g}(1 - \gamma). \end{aligned} \quad (5.4)$$

The model also includes a budget constraint with discounting, where $budget(\tau)$ represents the budget function for $\tau = 1, \ldots, 10$, and each τ represents a one-year period. Specifically, the constraint is

$$\sum_{i,t \in Q_\tau, g} CS^i S_{t,g} A^i_{t,g} (1 - \delta_q)^{-t} + \sum_{i,t \in Q_\tau, g} CT^i T^i_{t,g} B^i_{t,g} (1 - \delta_q)^{-t} \leq budget(\tau)(1 - \delta)^{-\tau}, \quad (5.5)$$

where Q_τ is the quarter time periods in decision period τ and CS^i and CT^i are the costs for screening and treating individuals with health state i.

Solving this optimization problem is not trivial due to the large number of variables and their interactions in the dynamic equations. Therefore, we redefined the decision variables to be the percentages of each annual budget used for screening in five decision periods, denoted as $x_1, \ldots, x_5 \in [0, 1]$, where each decision period is two years. Given a two year budget, x_k, the associated screening and treatment strategies $S_{t,g}$ and $T^i_{t,g}$ are determined in the optimization model. The budget constraint in (5.5) is revised to allow disaggregation of the two-year budget into screening and treatment decisions quarterly, as follows

$$\sum_{i,t \in Q_\tau, g} CS^i S_{t,g} A^i_{t,g} (1 - \delta_q)^{-t} \leq x_k budget(\tau)(1 - \delta)^{-\tau}, \quad (5.6)$$

$$\sum_{i,t \in Q_\tau, g} CT^i T^i_{t,g} B^i_{t,g} (1 - \delta_q)^{-t} \leq (1 - x_k) budget(\tau)(1 - \delta)^{-\tau}. \quad (5.7)$$

Table 5.1: Summarized grid search results from [60]

Abbreviation	Cohort Age Group	Total Budget	Grid Search Budget Screening Ratio $[x_1, x_2, x_3, x_4, x_5]$	Total QALYs (discounted)
40-Hb	40-49	50 billion	[**0.6**, **0**, 0, 0, 0]	3,842,081,631
50-Hb	50-59	50 billion	[0, **0.2**, **0.2**, 0, 0]	3,186,339,442
60-Hb	60-69	50 billion	[0, **0.2**, 0, **0.2**, 0]	1,802,296,572
40-Lb	40-49	10 billion	[**0**, **0**, 0, 0, 0]	3,839,336,212
50-Lb	50-59	10 billion	[**0**, **0**, 0, 0, 0]	3,182,871,728
60-Lb	60-69	10 billion	[**0**, **0**, 0, 0, 0]	1,800,940,994

The formulation determines the number of individuals screened and treated based on the implementation policy, and uses all of the available budgets in each two-year decision period.

5.1.3 Optimization of the Hepatitis C Markov Model

We analyze strategy decisions for three age groups: 40 to 49, 50 to 59, and 60 to 69. Two budget scenarios are considered in this paper: a high budget scenario with 50 billion dollars for the ten-year period (Hb), and a low budget scenario with 10 billion dollars (Lb). The age groups and budget scenarios makes six scenarios. The initial population size of each age cohort is given by: 40 to 49, 43,599,555; 50 to 59, 41,962,930; 60 to 49, 29,253,187.

We first used grid search on each scenario to understand the major decision period impacts on the performance, measured by health utility gain in QALYs, and summarized in Table 5.1. The decisions $(x_1, \ldots, x_5)$ in Table 5.1 represent the percentage of budget allocated for screening in the five decision periods.

When the budget is low (10 billion), the entire budget is allocated to treatment and there is no screening. When the budget is high (50 billion), we observe that at most only two decision periods are suggested to have budget allocated for screening. In order to understand the sensitivity of total health utility gain to the proportion of budget allocated

to screening, we apply the Probabilistic Branch and Bound (PBnB) algorithm [46, 48] to approximate the set of the top 10% solutions, called the level set. The shape of the level set indicates the sensitivity of each decision period, similar to performing a sensitivity analysis. Also, the decision makers understand the trade-offs for selecting a non-optimal but good enough solution because some considerations are not easily quantifiable.

Two types of results are provided by PBnB. The first result is an approximation of the δ level set formed by maintained subregions $\widetilde{\Sigma}^M$ that are statistically confident in the level set, and bounded by $\widetilde{\Sigma}^C$, undecided subregions. The pruned region, $\widetilde{\Sigma}^P$, are outside the level set with $(1-\alpha)$ statistical confidence. Second, an interval estimate of the quantile (upper and lower bounds) is provided for decision makers on the performance of the level set. For analyzing the HCV screening and treatment strategies, we use $\delta = 0.1$, $\epsilon = 0.025$, and $\alpha = 0.25$.

The shape of a five dimensional level set produced from PBnB is difficult to visualize. So we identified two dominant decision periods for each scenario, indicated in bold in Table 5.1. For the 40-Hb scenario, the first two decision periods x_1, x_2 are considered to have dominant impact and are selected as the major decision periods and x_3, x_4 and x_5 are set to zero. For the 50-Hb scenario, x_2 and x_3 are considered dominant, and for the 60-Hb scenario, x_2 and x_4 are dominant and the rest of the decision periods are set to zero. Similarly, 40-Lb, 50-Lb and 60-Lb scenarios also use x_1, x_2 as the major decision periods and set x_3, x_4 and x_5 to zero. We perform a PBnB search on all five decision periods to confirm the time periods with screening budget allocation. For instance, the PBnB five dimensional search result is $[0.4632, 0.5242, 0.0041, 0.0002, 0.0005]$ for the scenario 40-Hb representing the 40-49 age group with a high budget of 50 billion dollars. The search result shows that the algorithm finds low values on time periods 3, 4, and 5, but is not able to reach zero since the algorithm samples uniformly between 0 to 1.

5.1.4 Results

The PBnB numerical results for the six scenarios are presented in Figures 5.2-5.7. The light gray (green) boxes are the maintained subregions and the dark gray (blue) boxes represent

the undecided subregions. The white boxes are the pruned subregions. The numbers in the boxes indicate the iteration of the algorithm running when the subregion is pruned or maintained. The black star is the grid search solution and the red star is the PBnB approximated best strategy. We optimized all six scenarios with a PBnB level set analysis and observed several patterns of level set within these six scenarios.

The shape of the level sets for the low budget scenarios, 40-Lb, 50-Lb, and 60-Lb in Figures 5.2, 5.3, and 5.4, is triangular, indicating that it is possible to allocate some budget screening in early years (x_1, x_2) without degrading the health utility gain too much. The shape of the level set in Figure 5.2 shows that x_1 is more sensitive than x_2, ranging up to 0.3 inside the level set, whereas x_2 ranges up to 0.5. The interval estimation of the 10% quantile is [3,839,318,552, 3,839,320,071]. The theory analysis of PBnB ensures that points in the level set have a health utility gain better than 3,839,318,552 QALYs with $1 - \alpha$ confidence. Hence, the solutions in the approximated 10% level set of scenario 40-10b perform at least 3,839,318,552 QALYs, and the difference between the best grid search performance, 3,839,336,212, and the lower bound is 17,660 QALYs, which is 0.00045% of the optimal strategy performance.

For scenarios 50-Lb and 60-Lb, the lower bounds on the 10% quantile are 3,182,760,942 and 1,800,895,621. The triangular patterns are similar to scenario 40-Lb, as shown in Figures 5.3 and 5.4, however the range of x_1 goes up to 0.4 as opposed to 0.3 for 40-Lb. Also, their differences between the grid search optimal strategy, 3,182,871,728 and 1,800,940,994, and the 10% quantile lower bounds, are 110,786 and 45,373 respectively, that are 0.0034% and 0.0025% of their optimal strategy performance.

Figure 5.5 shows the PBnB results for scenario 40-Hb, where the two decision variables are x_1 and x_2, while setting $x_3, x_4, x_5 = 0$. The maintained region in Figure 5.5 indicates that x_1 can range between 0.4 to 0.85, which means x_1 is not very sensitive if we set x_2 close to zero. Also, $0 < x_2 < 0.12$ is also a viable choice when x_1 is close to 0.45. Some part of the approximated level set appears as a vertical undecided region when x_1 is close to 0.5. This may be due to the difficulty that PBnB has in correctly determining the level set in this scenario, because 40-Hb is a relatively young cohort where the health utility gain difference between different screening budget settings is small.

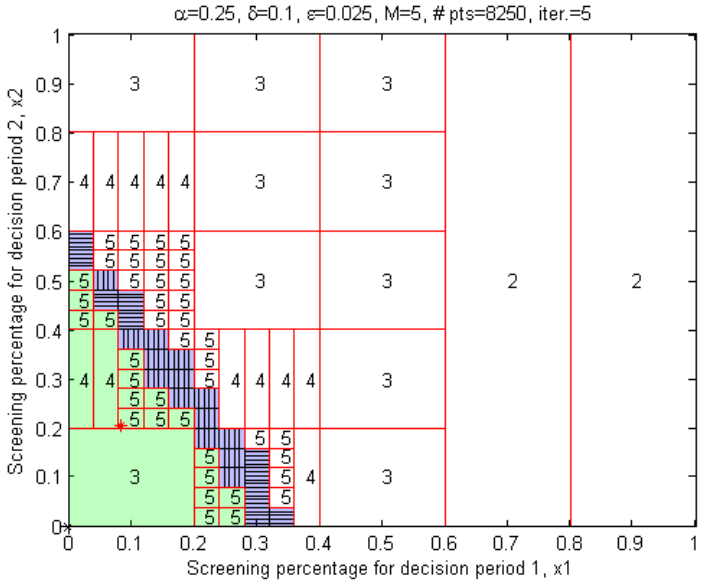

Figure 5.2: Approximated 0.1 level set for scenario 40-Lb with $x_3, x_4, x_5 = 0$. The grid search solution (black star) is (0, 0), and the PBnB approximated best strategy (red star) is (0.08, 0.21).

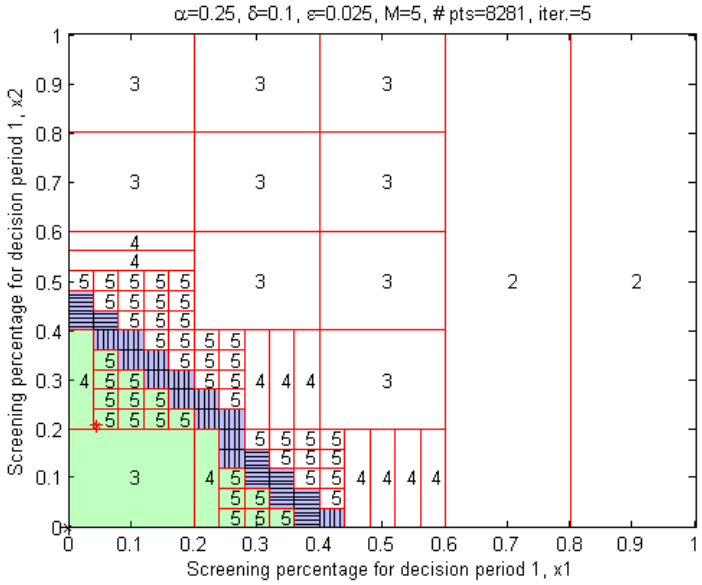

Figure 5.3: Approximated 0.1 level set for scenario 50-Lb with $x_3, x_4, x_5 = 0$ The grid search solution (black star) is (0, 0), and the PBnB approximated best strategy (red star) is (0.04, 0.21).

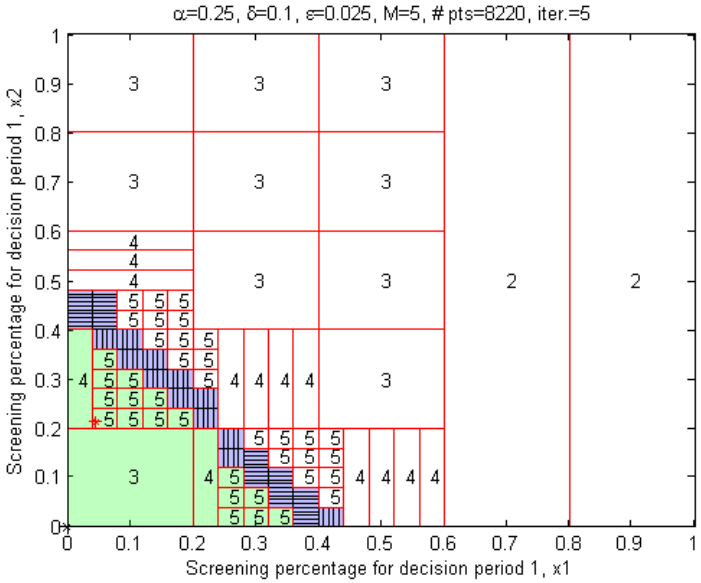

Figure 5.4: Approximated 0.1 level set for scenario 60-Lb with $x_3, x_4, x_5 = 0$ The grid search solution (black star) is (0, 0), and the PBnB approximated best strategy (red star) is (0.04, 0.21).

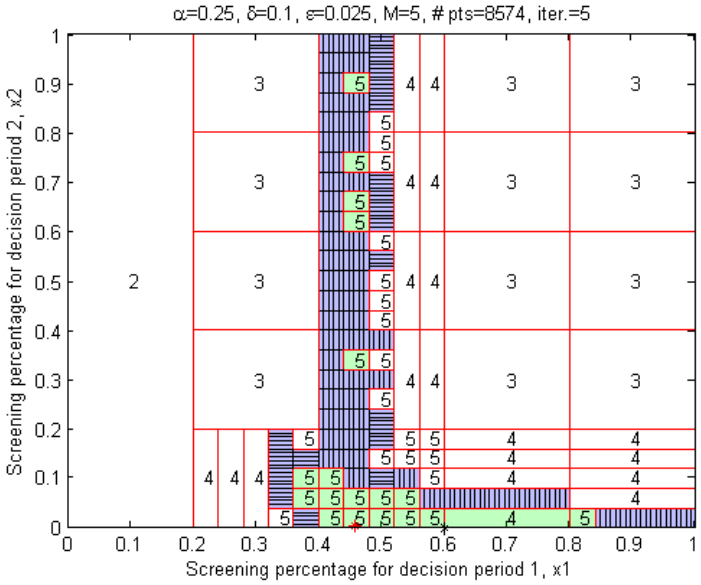

Figure 5.5: Approximated 0.1 level set for scenario 40-Hb with $x_3, x_4, x_5 = 0$. The grid search solution (black star) is (0.6, 0), and the PBnB approximated best strategy (red star) is (0.45, 0.004).

The algorithm also provides an interval estimation on the 10% quantile as between 3,842,041,405 and 3,842,041,540, indicating that the strategies in the maintained boxes statistically have at least 3,842,041,405 QALYs gain with $1 - \alpha$ confidence. Comparing to the PBnB approximated best strategy with 3,842,081,631 QALYs gain, choosing one in the level set may have 40,226 less QALYs (about 0.0014% of the best strategy performance). However, it could still be a good enough strategy while considering some other preferences that are not easily quantified and modeled.

The 50-Hb scenario has different major decision periods, x_2 and x_3, and the pattern of the level set is shown in Figure 5.6. The pattern of the 50-Hb scenario is a heavily stretched shape with a more flexible combination of x_2 and x_3. Figure 5.6 shows that x_2 and x_3 have a strong correlation to sum up to about 0.4. The performance of the maintained strategies is bounded by the 10% quantile lower bound, 3,186,265,607, where the PBnB approximated optimal strategy is 3,186,349,671. The difference between 10% quantile lower bound and the PBnB approximated optimal is 84,064 QALYs, which is 0.0026% of the best strategy performance and slightly greater than scenario 40-Hb. It may indicate the 50-Hb scenario is more sensitive to the screening and treatment budget allocation since the older cohort may have more immediate treatment requirements.

Figure 5.7 shows the level set for scenario 60-Hb, with a similar pattern as that for the scenario 50-Hb but with different major decision periods x_2 and x_4 instead of x_2 and x_3 for 50-Hb. The relationship between x_2 and x_4 in scenario 60-Hb has a steeper slope compared to 50-Hb, which shows that x_2 has a stronger impact in the 60-69 cohort. The health utility gain of the 60-Hb maintained strategies is bounded by 1,802,259,934, where the PBnB approximated optimal is 1,802,298,104 and the difference is 0.0021% of the best strategy performance.

In the lower budget scenarios, we observe that the grid search solution of (0, 0) is better than the PBnB approximated solution for two reasons. First, the box including (0, 0) is maintained in an early iteration of PBnB. Once a box is maintained, the algorithm does not continue to sample there refining the optimal solution. The second reason is that PBnB performs uniform sampling in the box, and it is practically impossible to hit (0, 0). For the high budget scenarios, the grid search optimum is not at (0, 0), and hence, PBnB is able to

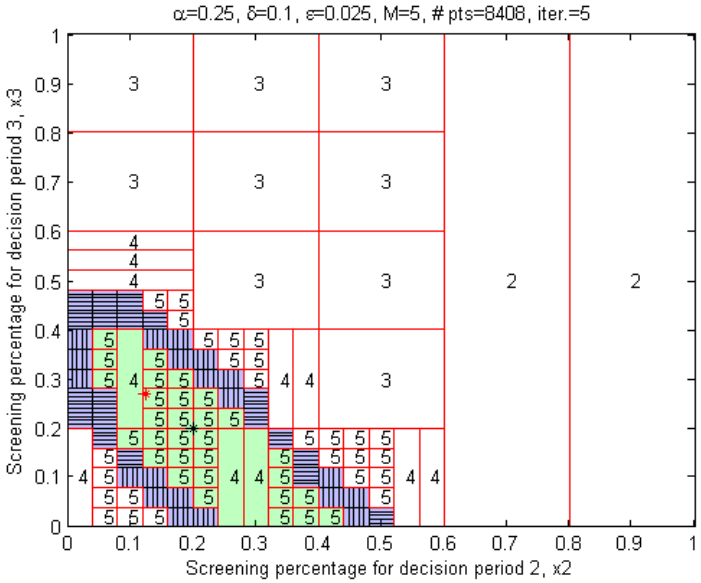

Figure 5.6: Approximated 0.1 level set for scenario 50-Hb with $x_1, x_4, x_5 = 0$. The grid search solution (black star) is (0.2, 0.2), and the PBnB approximated best strategy (red star) is (0.12, 0.27).

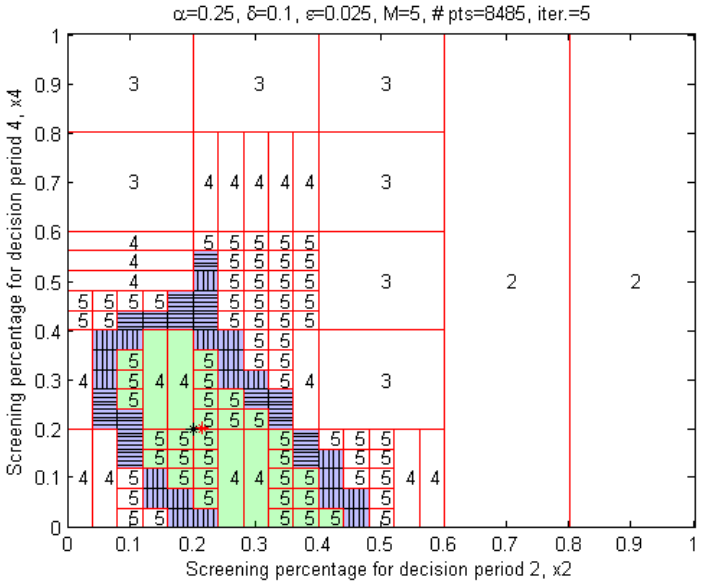

Figure 5.7: Approximated 0.1 level set for scenario 60-Hb with $x_1, x_3, x_5 = 0$. The grid search solution (black star) is (0.2, 0.2), and the PBnB approximated best strategy (red star) is (0.22, 0.20).

find a better solution than the grid search.

5.2 Portable Ultrasound Allocation for Orthopedic Diagnosis

5.2.1 Introduction

The hepatitis C budget allocation problem is a policy level decision with system details described by the disease progression in the system dynamics. In contrast, the portable ultrasound allocation problem focuses on the system design in an area hospital level, where the details of how patients flow through the system are modeled with a discrete-event simulation.

The US healthcare system is increasingly moving toward an emphasis on patient-centered care. Systems are testing alternative ways of delivering high quality and cost effective care by exploring new technologies. Previously, Adelman and Fishman [2] conducted a pilot study in which portable ultrasound was used to diagnose shoulder arthropathy within an outpatient surgical clinic. This new technology may reduce overall cost, and may affect patient health outcomes. We now extend the work in [2] to analyze the system and patient level impact of this technology to diagnose shoulder arthropathy as an alternative to the more common use of higher invasive image methods [64].

Previous research has demonstrated the effectiveness and efficacy of diagnostic ultrasound for shoulder arthropathy. This technique has been shown to have similar if not greater sensitivity and specificity for diagnosing shoulder arthropathy than more expensive and more invasive techniques like magnetic resonance imaging (MRI) or computed tomography (CT) at much lower cost to the healthcare system payer as well as reducing opportunity cost for patients [84, 87]. Based on the previously conducted pilot study [2], we now develop a discrete-event simulation model to demonstrate the potential system level effects of a population based implementation of portable ultrasound and outpatient setting for diagnosing shoulder arthropathy as a precursor to surgery. Our goal is to show the impact of an alternative, less expensive and less invasive procedure, intended to improve the quality of patient care within a patient-centered framework.

The following sections of this chapter are oriented as follows. The orthopedic care system

section describes the arthropathy diagnosis scheme and the modeling of the system as a discrete-event simulation model. The results section contains the analysis and optimized design of the problem.

5.2.2 Orthopedic Care System

Using a portable ultrasound for shoulder arthropathy provides a major change for a patient's pathway through the system. MRI is the most common method to diagnose shoulder arthropathy and is depicted in the upper part of Figure 5.8. It shows the patient pathway for using MRI as the imaging method, which includes the initial clinic visit, potential imaging center visit for MRI, second clinic visit and subsequent treatment. In this scenario, the patient may need to wait and travel to the imaging center before the second clinic visit. The lower part of Figure 5.8 illustrates the potential pathway for patients diagnosed by portable ultrasound imaging if needed. Since the imaging could be done immediately at the first clinic visit, the patient could save the time spent waiting for an appointment and travel time for both imaging and the second clinic visit as shown in Figure 5.8. The potential time savings and cost savings due to using portable ultrasound machines leads an organization to consider when to provide orthopedic care with portable ultrasound machines at current primary care clinics, which may result in more patient centered services and easier access to local orthopedic care.

The literature suggests that there is no significant difference in sensitivity and specificity of diagnosing rotator cuff tear between MRI and ultrasound [64]. However, the experiment was implemented with specialists who have ten years of experience in both MRI and ultrasound. The advantage of using a portable ultrasound is to let surgeons perform the imaging, however surgeons are typically not as experienced as an MRI specialist. We assume that the diagnosis using MRI provides excellent diagnosis quality, with 95% sensitivity and specificity, which means that 95% of patients that required surgery were diagnosed correctly and 95% of patients that did not require surgery were determined correctly after the MRI. The diagnosis quality of using portable ultrasound is defined by a parameter p, where the diagnosis quality is $95p\%$ as a relative value to MRI diagnosis. In other words, the portable

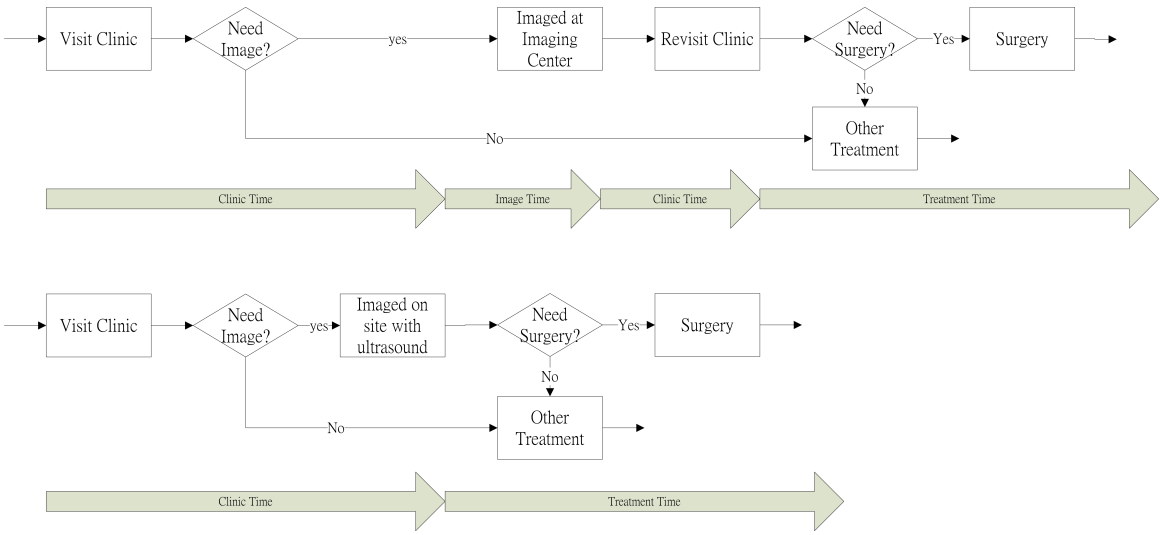

Figure 5.8: Patient pathways with MRI and portable ultrasound.

ultrasound provides $95p\%$ sensitivity and specificity.

In this analysis, two performance metrics are considered: cost and health utility loss. Cost includes cost of clinic service, MRI usage, ultrasound usage, and patient traveling for MRI. Health utility loss considers health loss of patients time before receiving the treatment and health utility loss for a not optimal treatment. The decisions of this portable ultrasound allocation problem are as follows. Each clinic must decide whether to purchase a portable ultrasound machine and arrange for surgeons to be trained in its use. In addition, the MRI capacity reserved for shoulder disorders at each specialty care center must also be determined.

5.2.3 Discrete-event Simulation and Optimization Modeling

The system is modeled as a discrete-event simulation model and includes three orthopedic clinics with imaging centers, and additional 14 potential orthopedic clinics, that currently are major primary care clinics with shoulder arthropathy patients. The 14 clinics are chosen because they have a high number of referrals to orthopedic care. Each primary care clinic is treated as a source of patient arrivals based on historical data that represents the rates of patient encounters in specialty care with diagnosis codes of shoulder arthropathy. In

the simulation model, the arrivals are modeled as Poisson distributions. The means for the patient arrivals are given in Table 5.2. Based on the historical data, we determine the probability that each patient clinical encounter may need imaging for diagnosing shoulder arthropathy. Based on the data, we estimate that 99% of patients that receive imaging need surgery. Depending on the imaging method, a probability based on the diagnosis quality determines whether the patient received an extra health utility loss for not receiving optimal treatments.

Table 5.3 contains the parameter settings of the orthopedic care discrete-event simulation model. Names of parameters are listed in the first column and the values are shown in the second column. Remarks indicate the sources of parameter values, including data, literature, and assumptions. Cost for each ultrasound purchase spreads the purchase price evenly into 10 years. Health utility loss for a patient receiving unnecessary surgery, i.e. only non-surgical treatment is required, is assumed to be equal to a full year loss before other treatment. Similarly, health utility loss for a patient receiving other treatment when surgery is required is assumed to be equal to a full year loss before surgery.

The decisions for the optimization model include the allocation of portable ultrasound at 17 clinics (14 primary care and 3 specialty centers) and MRI capacity at three imaging centers, assuming that the opportunity cost of reserving MRI capacity for orthopedic diagnosis is half of the total usage cost. There are 17 binary variables (indicating location of portable ultrasound machines) and 3 continuous variables (indicating amount of MRI capacity reserved per day). Evaluating potential combinations of all system designs is computationally impractical. Therefore, we apply optimization techniques to search for good enough solutions for our multiple objective model. One typical use for multiple objective analysis is to examine the trade-offs between different performance evaluations. Decision makers can directly investigate the trade-offs between objectives through an approximated Pareto optimal set. MOPBnB is used for Pareto optimal approximation and provides a certain statistical quality of the approximated Pareto set.

Table 5.2: The means for the daily patient arrivals at different major primary care clinics.

Location	Mean	Corresponding specialty care center
Burien	3.4500	
Central	18.7604	12.1821 Central and 6.5783 Bellevue
Northgate	11.504	8.5217 Central and 2.9823 Bellevue
Bellevue	13.3635	Bellevue
Factoria	5.0872	Bellevue
Everett	5.0872	Bellevue
Redmond	5.4324	Bellevue
Lynwood	3.2604	Bellevue
Northshore	3.2460	Bellevue
Renton	4.2912	Tacoma
FederalWay	4.7867	Tacoma
Olympia	13.9671	Tacoma
PortOrchard	7.0958	Tacoma
Puyallup	6.3589	Tacoma
Silverdale	7.8679	Tacoma
Tacoma	13.3635	Tacoma
Tacoma South	6.2676	Tacoma

Table 5.3: Parameter settings of portable ultrasound allocation simulation model.

Parameter	Value	Remark
Imaging probability	0.005	Data
Surgery probability	0.95	Assumed
MRI diagnosis quality	0.95	Assumed
Ultrasound relative diagnosis quality	0.7, 0.8, 0.9	
Cost for each clinic visit	300	Assumed
Cost for each MRI use	1500	[2]
Cost for each ultrasound use	114	[2]
Cost for each ultrasound purchase	4857/year	Assuming 10 year use
Health utility loss before a required surgery	0.000548 QALYs/day	0.2/per year Euro QALYs
Health utility loss before an other-treatment	0.00011 QALYs/day	Assumed 1/5 of surgery
Health utility loss for a patient receiving a surgery when only requires other treatment	0.04 QALYs	Assume one year lost
Health utility loss for a patient receiving an other-treatment while requires a surgery	0.16 QALYs	Assume one year lost
Clinic service time	1/200 day	Assumed
Travel cost for imaging	$.575/mile	UW travel

5.2.4 Results

This section reports the results of simulation and optimization model analyses of the impact of implementing the portable ultrasound machine. First, we run the simulation model under the current design (labeled MRI in Figure 5.9) with NO portable ultrasound and full MRI capacity at each imaging center (5 patients per day). An alternative option is to expand orthopedic care services to each primary care location (labeled Expand in Figure 5.9). Expanding orthopedic care has the potential to achieve better patient-centered services, with portable ultrasound machines having the potential for better patient satisfaction and better health outcomes. We consider different ultrasound diagnosis qualities; p is set to 0.7, 0.8, and 0.9. Depending on the demand, travel distance, and surgeon diagnosis quality using an ultrasound machine, it may not be worth expanding the orthopedic care to all current primary clinic locations. Therefore, we use a simulation optimization algorithm, MOPBnB, to analyze the trade-offs between the cost and heath utility by approximating the Pareto optimal set.

Figure 5.9 illustrates the relationship between the two objectives, cost and health utility loss, for solutions obtained under the MRI scenario, three Expand scenarios and the optimal scenarios using MOPBnB. The Expand scenarios compared to the MRI scenario indicate the cost saving while diagnosing patients with portable ultrasound. Also, for Expand scenarios, the decrease of 0.1 p increases the health utility loss. The MOPBnB algorithm provides several optimal solutions for each of the three diagnosis qualities p, showing that there exist certain combinations of allocation and expanding ultrasound for orthopedic care services to provide lower cost and health utility loss than the MRI or Expand scenarios.

Specifically, for high ultrasound diagnosis quality, p=0.9, there are four designs (blue stars) that approximate Pareto optimal solutions. Each of the designs includes whether to expand orthopedic care to each clinic and the suggested MRI capacity at each imaging center as shown in Table 5.4. The top left design, Design 1, has the highest health utility loss (3.32 QALYs) and lowest cost (0.964e7), which implements portable ultrasounds at 11 locations and only reserves 2.8 MRI patients per day through three imaging centers. At the other extreme, Design 4 with ultrasounds installed at five locations and 8.81 MRI capacity

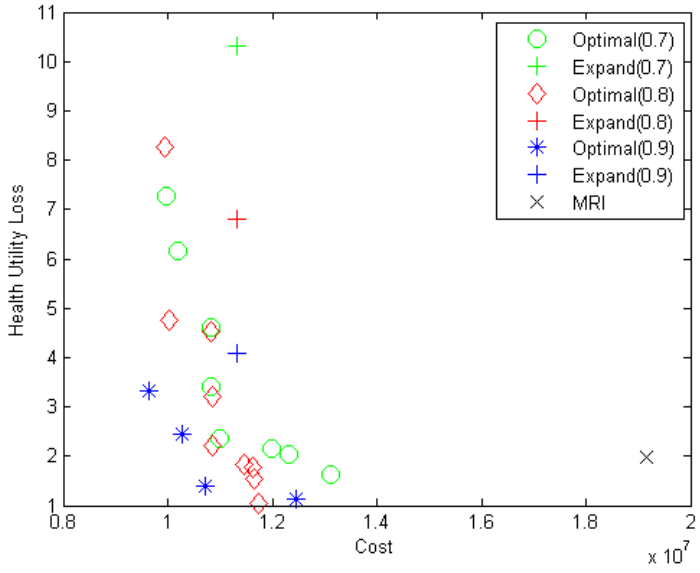

Figure 5.9: Approximated Pareto optimal set of different diagnosis quality p.

reserved, has the lowest health utility loss (1.14 QALYs) with a low cost of 1.246e7 cost. The cost saving of Design 4 cover the purchasing cost of five ultrasound machines. In addition, health utility loss associated with Design 4 is less than the MRI scenario. Therefore, we know that it is worthy to implement machines at five locations. Two locations, Burien and Northshore clinics, are suggested to implement ultrasound machine in all four designs.

Similarly, for p=0.8 and 0.7, there are also trade-offs between the approximated Pareto optimal designs, ranging from 7.27 QALYs loss with 0.994e7 cost to 1.04 loss QALYs with 1,174e7 cost for p=0.8 as shown in Table 5.5 and 7.27 QALYs loss with 0.997e7 cost to 1.64 loss QALYs with 1,3115e7 cost for p=0.7 as shown in Table 5.6.

Table 5.4: Pareto optimal designs of portable ultrasound allocation for $p = 0.9$.

decision variable	Design 1	Design 2	Design 3	Design 4
Burien	✓	✓	✓	✓
Central				
Northgate	✓		✓	
Bellevue		✓		✓
Factoria	✓	✓		
Everett	✓	✓	✓	
Redmond				
Lynwood	✓		✓	
Northshore	✓	✓	✓	✓
Renton	✓	✓		✓
FederalWay			✓	
Olympia				
PortOrchard	✓			✓
Puyallup				
Silverdale	✓		✓	
Tacoma	✓		✓	
Tacoma South	✓			
Central MRI	1.326	0.003	2.449	3.486
Bellevue MRI	0.644	3.966	3.767	0.446
Tacoma MRI	0.836	1.105	3.759	4.881

Table 5.5: Pareto optimal designs of portable ultrasound allocation for $p = 0.8$.

decision variable	Design 1	Design 2	Design 3	Design 4	Design 5	Design 6	Design 7	Design 8	Design 9
Burien									
Central									
Northgate	✓	✓	✓	✓			✓		✓
Bellevue									
Factoria	✓	✓	✓	✓	✓	✓		✓	
Everett	✓			✓		✓			✓
Redmond	✓	✓	✓		✓				
Lynwood	✓								
Northshore	✓	✓			✓		✓	✓	
Renton	✓	✓	✓			✓	✓		
FederalWay		✓	✓	✓	✓			✓	
Olympia	✓			✓	✓		✓		
PortOrchard						✓		✓	
Puyallup						✓		✓	✓
Silverdale		✓	✓		✓				
Tacoma						✓			
Tacoma South	✓	✓	✓		✓	✓	✓	✓	
Central MRI	0.097	0.529	1.668	2.05	2.564	2.956	2.049	0.795	1.522
Bellevue MRI	1.362	0.334	0.004	4.2	1.317	0.76	1.67	2.525	3.562
Tacoma MRI	0.26	2.113	0.769	1.211	3.676	2.026	2.52	1.583	4.912

Table 5.6: Pareto optimal designs of portable ultrasound allocation for $p = 0.7$.

decision variable	Design 1	Design 2	Design 3	Design 4	Design 5	Design 6	Design 7	Design 8
Burien	✓			✓				
Central								
Northgate	✓							
Bellevue		✓	✓					
Factoria		✓		✓		✓		✓
Everett		✓	✓		✓		✓	
Redmond	✓							
Lynwood				✓	✓			
Northshore			✓	✓			✓	✓
Renton		✓			✓	✓		
FederalWay	✓		✓			✓		✓
Olympia	✓	✓	✓		✓		✓	
PortOrchard				✓		✓		✓
Puyallup	✓	✓	✓					✓
Silverdale		✓	✓	✓		✓		
Tacoma	✓	✓		✓				
Tacoma South	✓			✓		✓		
Central MRI	0.451	3.433	3.438	2.136	0.55	4.193	1.936	4.103
Bellevue MRI	2.898	0.771	2.019	0.005	1.466	1.993	3.918	3.884
Tacoma MRI	0.622	1.283	1.255	1.061	3.967	0.943	3.406	1.114

Furthermore, the first three Pareto optimal designs of p=0.9, second through fifth designs of p=0.8, and first five designs of p=0.7 dominate Expand scenarios. Therefore, if decision makers are seeking low cost options, these designs are the candidates. On the other hand, when the decision makers weight health utility loss more than cost, other Pareto optimal designs should be considered.

5.3 Summary and Discussion

The hepatitis C screening and treatment allocation strategy problem illustrates PBnB for level set approximation for analyzing the sensitivity of decision variables by an approximated level set. The portable ultrasound machine is considered as a demonstration of trade-off analysis for multiple objectives by MOPBnB.

For the hepatitis C screening and treatment allocation strategy problem, we first identify two major decision periods for each scenario through a grid search on five decision periods. By focusing on two major decision periods, a visualization of the approximated level set for each scenario is provided. Using the level set, we analyze the relationship between the two major decision periods and provide information for decision makers on the sensitivity of the decisions. We also provide the impact of the implementation policy on population health outcomes and the trade-offs between screening and treatment efforts to eliminate a major disease in the U.S. healthcare systems.

Portable ultrasound machines are considered as a new technology that can change the system design to provide better patient-centered care for the orthopedic care. We use a discrete-event simulation model to evaluate different system designs with two performance metrics, cost and health utility loss. MOPBnB searched the Pareto optimal designs to understand the trade-offs between cost and health utility loss and construct design options for decision makers. The discrete-event simulation model has unique characteristics that may not be generalizable to other new technologies, however, we think this framework with MOPBnB is a good approach for the heath decision makers to analyze tradeoffs when considering some cost saving and patient-centered technologies.

Chapter 6

CONCLUSION

This dissertation addresses complex healthcare resource allocation problems requiring decision support by advanced analytic methods. A simulation optimization algorithm is introduced, Probabilistic Branch and Bound (PBnB), focusing on level set approximation for single objective problems and Pareto optimal set approximation for multiple objective problems. PBnB provides a set of solutions with probability bounds on the quality of the set, and it is capable of solving problems with mixed continuous and discrete variables. Three healthcare resource allocation problems are modeled with discrete-event simulation models to evaluate the performance of different system designs. First, an occupational health campaign determines a nurse staffing and laboratory test schedule, which has a small number of system designs. Hence, all designs are evaluated that are of interest to the decision makers. Second, a hepatitis C screening and treatment strategy requires the search for an optimal allocation of resources. PBnB for level set approximation is the approach used which allows the real-valued decision variables to be analyzed and provides further insights. Third, a portable ultrasound allocation problem for orthopedic diagnosis considers the replacement of MRI screening with portable ultrasound machines. Multiple objective PBnB is used to balance the cost and health utility loss, when substituting some amount of MRI capacity with lower cost ultrasound machines.

6.1 Probabilistic Branch and Bound

Probabilistic Branch and Bound (PBnB) is a partition-based random search simulation optimization algorithm for stochastic problems. It is beneficial for the decision makers as it is easier to analyze a system with noise, as given with a discrete-event simulation, and determine several good solutions. PBnB determines a set of solutions through an estimated bound on the performance. For a single objective problem, PBnB approximates a desirable

level set with quantile estimation. A level set benefits decision making through providing a set of well-performing solutions that allows decision makers to consider additional criteria that are hard to quantify and model. Also, the shape of an approximated level set represents the sensitivity of each decision variable. In a multiple objective problem, PBnB approximates the Pareto optimal set of solutions, by considering a bound on the closeness of a solution to the efficient frontier. The approximated efficient frontier and Pareto optimal set illustrate the trade-offs between objective functions and allow decision makers to select a design with their preference.

PBnB for level set approximation iteratively maintains, prunes or branches subregions of the bounded solution space (possibly with mixed continuous and discrete variables) based on an updated confidence interval of a target quantile. Finite-time probability bounds are derived on the maximum volume of incorrectly maintained or incorrectly pruned regions. Numerical results on noisy and non-noisy test functions demonstrate the performance of the algorithm. Tests on a sphere function, with spherical level sets, allow a comparison between the theoretical bounds and numerical results. Multiple objective probabilistic branch and bound (MOPBnB) focuses the goal on approximating the efficient frontier and the associated Pareto optimal set in the solution space. MOPBnB is also developed for both deterministic and noisy problems with mixed continuous and discrete variables. When the algorithm terminates, it provides a set of non-dominated solutions that approximates the Pareto optimal set and the associated objective function estimates that approximate the efficient frontier by pruning the subregions statistically far from the efficient frontier. Hence, the quality of the solutions is statistically analyzed using a measure of distance between solutions to the true efficient frontier. Numerical experiments with benchmark functions are presented to visualize the algorithm and its performance.

6.2 Healthcare Resource Allocation

The design and implementation of a concurrent campaign of influenza immunization and tuberculosis (TB) screening for health care workers (HCWs) aimed to reduce the number of clinic visits for each HCW. A discrete-event simulation model was developed to support resource allocation decisions in the planning and operation phases. Since the number

of possible designs is relatively small, all designs were evaluated for the decision makers. Furthermore, the simulation model was refined during the campaign and reused for the following years. The campaign was compressed to 100 days in 2010 and further compressed to 75 days in 2012 and 2013. With more than 5000 HCW arrivals in 2011, 2012 and 2013, the 14-day goal of providing TB screening results was achieved for each year and was reduced to approximately 4 days in 2012 and 2013.

The hepatitis C screening and treatment strategy is a problem regarding screening and treatment strategies under budget constraints for chronic hepatitis C birth-cohorts in the U.S. A Markov model of disease progression from [60] is able to evaluate health utility gain using quality-adjusted life years (QALYs) for each strategy. Using PBnB for a level set approximation provided an optimal strategy for budget allocation over ten years as well as a sensitivity analysis by an approximated set of good enough strategies. Specifically, first time periods with obvious dominant strategies were identified through grid search, and then PBnB was applied to identify the top 10 percent of strategies for two major decision periods. Approximating a set of the top 10 percent of strategies with PBnB provides decision makers the ability to explore good strategies. Also, a set of strategies indicates which decision time periods strongly impact the health utility gain.

A third healthcare problem considers using new, portable ultrasound machines to replace or supplement magnetic resonance imaging (MRI) for shoulder disorder diagnosis in orthopedic clinics. When implementing portable ultrasound machines, the health outcome of the patient needs to be considered in addition to costs. A discrete-event simulation model and MOPBnB were used to analyze the trade-off between the health outcomes of the patients and the cost of implementing the portable ultrasound machines and cost saving by reducing MRI usage. The decisions include purchasing and locating portable ultrasound machines, training the staff at appropriate clinics and the MRI capacity allocated for shoulder disorders. MOPBnB provides an approximated Pareto optimal set of system designs that allows decision makers to comprehensively understand the trade-offs.

6.3 Future Work

There is potential to further improve the PBnB algorithm since the current approach to sample size and replication is conservative and the number of subregions increases exponentially as the dimension increases. The current version makes no assumption on the black box function. Hence, a new version of PBnB with a lower sample size could be developed by assuming some function characteristics such as Lipschitz continuous. To avoid the increasing number of subregions, a sequential approach to approximate the target set of solutions could be considered. With the exception of the current promising subregion, other subregions can be merged, which limits the number of subregions for each iteration to a constant. This would improve the performance for large dimensional problems, and the target set will be captured partially and grows with further iterations.

When considering multiple objectives, the current version of MOPBnB is only driven by a distance metric, an indicator of closeness to the efficient frontier. The spread and the number of points could provide more comprehensive insights to the quality of MOPBnB.

There are also future research opportunities in the optimization of the hepatitis C screening and treatment strategy under budget constraints. Currently, the budget between the age-cohorts is determined by size of the cohort. However, health utility may be increased by allocating the budget unequally based on the characteristics of each cohort.

The portable ultrasound allocation problem currently uses the constant p to represent the diagnosis quality using the portable ultrasound. However, as surgeons gain experience, the diagnosis quality should improve over time. Hence, p could be described as a learning curve function to better represent the long-term performance of the orthopedic care center. There are other new medical technologies that could be analyzed under a similar framework involving a discrete-event simulation model and MOPBnB.

BIBLIOGRAPHY

[1] K. Aaby, R. L. Abbey, J. W. Herrmann, M. Treadwell, C. S. Jordan, and K. Wood. Embracing computer modeling to address pandemic influenza in the 21st century. *Journal of Public Health Management and Practice*, 12(4):365–372, 2006.

[2] S. Adelman and P. Fishman. Use of portable ultrasound machine for outpatient orthopedic diagnosis: an implementation study. *Perm J*, 17(3):18–22, 2013.

[3] M. A. Ahmed and T. M. Alkhamis. Simulation optimization for an emergency department healthcare unit in Kuwait. *European Journal of Operational Research*, 198(3):936–942, 2009.

[4] M. M. Ali, C. Khompatraporn, and Z. B. Zabinsky. A numerical evaluation of several stochastic algorithms on selected continuous global optimization test problems. *Journal of Global Optimization*, 31(4):635–672, 2005.

[5] M.M. Ali, C. Khompatraporn, and Z. B. Zabinsky. A numerical evaluation of several stochastic algorithms on selected continuous global optimization test problems. *Journal of Global Optimization*, 31:635–672, 2005.

[6] S. Andradottir. A global search method for discrete stochastic optimization. *SIAM Journal on Optimization*, 6:513–530, 1996.

[7] J. April, F. Glover, J. P. Kelly, and M. Laguna. Practical introduction to simulation optimization. In S. Chick, P. J. Saanchez, D. Ferrin, and D. J. Morrice, editors, *Proceedings of the 2003 Winter Simulation Conference*, pages 71–78, Piscataway, New Jersey, 2003.

[8] J. Banks. *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*. John Wiley & Sons, New York, 1998.

[9] R. E. Bechhofer, C. W. Dunnett, and M. Sobel. A two-sample multiple decision procedure for ranking means of normal populations with a common unknown variance. *Biometrika*, 41:170–176, 1954.

[10] G. Berchtold, H. Blaschke, F. Hanssmann, F. Liebl, S. L. Braun, W. Vogt, M. Eckert, G. Hoffmann, and S. Klose. Simulation modeling as a tool to evaluate alternative configurations of clinical laboratories. *SIMULATION*, 63(2):108–120, 1994.

[11] M. Berg. Accumulating and coordinating: Occasions for information technologies in medical work. *Computer Supported Cooperative Work (CSCW)*, 8(4):373–401, 1999.

[12] P. T. Van Berkel and J. T. Blake. A comprehensive simulation for wait time reduction and capacity planning applied in general surgery. *Health Care Management Science*, 10(4):373–385, 2007.

[13] L. Bianchi, M. Dorigo, L. M. Gambardella, and W. J. Gutjahr. A survey on metaheuristics for stochastic combinatorial optimization. *Natural Computing*, 8(2):239–287, 2008.

[14] K. Bodtker, L. Wilson, and W. Godolphin. Simulation modelling to assist operational management and planning in clinical laboratories. *SIMULATION*, 60(4):247–255, 1993.

[15] S.C. Brailsford, V.A. Lattimer, P. Tarnaras, and J.A. Turnbull. Emergency and on-demand healthcare: Modelling a large complex system. *Journal of the Operational Research Society*, 55:34–42, 2004.

[16] A. Brousselle and C. Lessard. Economic evaluation to inform health care decision-making: promise, pitfalls and a proposal for an alternative path. *Soc Sci Med*, 72(6):832–9, 2011.

[17] B. Cardoen, E. Demeulemeester, and J. Belien. Operating room planning and scheduling: A literature review. *European Journal of Operational Research*, 201(3):921–932, 2010.

[18] B. L. Carter, G. Ardery, J. D. Dawson, P. A. James, G. R. Bergus, W. R. Doucette, E. A. Chrischilles, C. L. Franciscus, and Y. Xu. Physician and pharmacist collaboration to improve blood pressure control. *Archives of Internal Medicine*, 169(21):1996–2002, 2009.

[19] CDC. Health, United States, 2012: With special feature on emergency care. Technical report, National Center for Health Statistics, CDC, Hyattsville, MD, 2013.

[20] CDC. Hepatitis C information for the public. Technical report, Centers for Disease Control and Prevention, Accessed at http://www.cdc.gov/hepatitis/hcv/cfaq.htm, Atlanta, GA, 2013.

[21] C. H. Chen and D. He. Intelligent simulation for alternatives comparison and application to air traffic management. *Systems Science and Systems Engineering*, 14(1):37–51, 2005.

[22] C. H. Chen, D. He, M. Fu, and L. H. Lee. Efficient simulation budget allocation for selecting an optimal subset. *INFORMS Journal on Computing*, 20(4):579–595, 2008.

[23] C. H. Chen, J. Lin, E. Yucesan, and S. Chick. Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10(3):251–270, 2000.

[24] H. C. Chen, L. Dai, C. H. Chen, and E. Yucesan. New development of optimal computing budget allocation for discrete event simulation. In S. Andradhottir, K. J . Healy, D. H. Withers, and B. L. Nelson, editors, *Proceedings of the 1997 Winter Simulation Conference*, pages 334–341, Atlanta, Georgia, 1997.

[25] J. Chhatwal, F. Kanwal, M. S. Roberts, and M. A. Dunn. Cost-effectiveness and budget impact of hepatitis C virus treatment with sofosbuvir and ledipasvir in the United States. *Annals of internal medicine*, 162:6:397–406, 2015.

[26] W. J. Conover. *Practical Nonparametric Statistics, 3rd Edition.* John Wiley & Sons, Inc., New York, 3rd edition, 1999.

[27] T. Csendes and J. Pintér. A new interval method for locating the boundary of level sets. *International Journal of Computer Mathematics*, 49(1-2):53–59, January 1993.

[28] K. Deb. *Multi-objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Chichester, New York, 2001.

[29] B. T. Denton, A. J. Miller, H. J. Balasubramanian, and T. R. Huschka. Optimal allocation of surgery blocks to operating rooms under uncertainty. *Operations Research*, 58(4-Part-1):802–816, 2010.

[30] S. Deo, K. Rajaram, S. Rath, U. S. Karmarkar, and M. B. Goetz. Planning for HIV screening, testing, and care at the Veterans Health Administration. *Operations Research*, 62:2:287–304, 2015.

[31] P. Dolan. *Chapter 32 The measurement of health-related quality of life for use in resource allocation decisions in health care, Handbook of Health Economics*, volume 1, Part B. Elsevier, 2000.

[32] M. F. Drummond. Resource allocation decisions in health care: A role for quality of life assessments? *Journal of Chronic Diseases*, 40(6):605–616, 1987.

[33] T. Eldabi, R. J. Paul, and T. Young. Simulation modelling in healthcare: reviewing legacies and investigating futures. *Journal of the Operational Research Society*, 58:262–270, 2007.

[34] D. Fone, S. Hollinghurst, M. Temple, A. Round, N. Lester, A. Weightman, K. Roberts, E. Coyle, G. Bevan, and S. Palmer. Systematic review of the use and value of computer simulation modelling in population health and health care delivery. *Journal of Public Health Medicine*, 24(4):325–335, 2003.

[35] M. C. Fu. Optimization for simulation: Theory vs. practice. *INFORMS Journal on Computing*, 14(3):192–215, 2002.

[36] M. C. Fu. *Handbook of Simulation Optimization*. Springer, New York, 2015.

[37] M. C Fu, C. H. Chen, and L. Shi. Some topics for simulation optimization. In S. J. Mason, R. R. Hill, L. Mnch, T. Jefferson O. Rose, and J. W. Fowler, editors, *Proceedings of the 2008 Conference on Winter Simulation*, pages 27–38, Miami, Florida, 2008.

[38] M. C. Fu, F. W. Glover, and J. April. Simulation optimization: a review, new developments, and applications. In M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, editors, *Proceedings of the 2005 Winter Simulation Conference*, pages 83–95, Orlando, Florida, 2005.

[39] M. L. Garcia, M. A. Centeno, C. Rivera, and N. DeCario. Reducing time in an emergency room via a fast-track. In C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman, editors, *Proceedings of the 1995 Winter Simulation Conference*, pages 1048–1053, Arlington, Virginia, 1995.

[40] J. D. Gibbons, I. Olkin, and M. Sobel. *Selecting and Ordering Populations: A New Statistical Methodology.* Wiley, New York, 1977.

[41] M. M. Guenal and M. Pidd. Discrete event simulation for performance modelling in health care: a review of the literature. *Journal of Simulation*, 4:42–51, 2010.

[42] A. Gupta, G. W. Evans, and S. S. Heragu. Simulation and optimization modeling for drive-through mass vaccination - a generalized approach. *Simulation Modelling Practice and Theory*, 37:99–106, 2013.

[43] J. A. Heim, H. Huang, Z. B. Zabinsky, J. Dickerson, M. Wellner, M. Astion, D. Cruz, J. Vincent, and R. Jack. Design and implementation of a combined influenza immunization and tuberculosis screening campaign with simulation modelling. *Journal of Evaluation in Clinical Practice*, 21(4):727–734, 2015.

[44] Y. C. Ho, C. G. Cassandras, C. H. Chen, and L. Dai. Ordinal optimisation and simulation. *Journal of the Operational Research Society*, 51:490–500, 2000.

[45] Y. C. Ho, Q. C. Zhao, and Q. S. Jia. *Ordinal optimization: Soft optimization for hard problems.* Springer, Berlin, Germany, 2007.

[46] H. Huang and Z. B. Zabinsky. Adaptive probabilistic branch and bound with confidence intervals for level set approximation. In R. Pasupathy, S. H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, editors, *Proceedings of the 2013 Winter Simulation Conference*, Washington D.C., 2013.

[47] H. Huang and Z. B. Zabinsky. Multiple objective probabilistic branch and bound for Pareto optimal approximation. In S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, editors, *Proceedings of the 2014 Winter Simulation Conference*, pages 3916–3927, Savannah, Georgia, 2014.

[48] H. Huang and Z. B. Zabinsky. Partition-based approach to level set approximation for simulation optimization. *Manuscript submitted*, 2016.

[49] H. Huang, Z. B. Zabinsky, J. A. Heim, and P. Fishman. Simulation optimization for medical imaging resource allocation. In *Extended Abstract of the 2015 INFORMS Healthcare Conference*, pages 3916–3927, Nashville, Tennessee, 2015.

[50] H. Huang, Z. B. Zabinsky, Y. Li, and S. Liu. Analyzing hepatitis C screening and treatment strategies using probabilistic branch and bound. In T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, editors, *Proceedings of the 2016 Winter Simulation Conference*, Washington D.C., 2016.

[51] D. F. Jones, S. K. Mirrazavi, and M. Tamiz. Multi-objective meta-heuristics: An overview of the current state-of-the-art. *European Journal of Operational Research*, 137(1):1–9, 2002.

[52] J. B. Jun, S. H. Jacobson, and J. R. Swisher. Applications of discrete-event simulation in health care clinics: a survey. *The Journal of the Operational Research Society*, 50(2):109–123, 1999.

[53] S. H. Kim and B. L. Nelson. A fully sequential procedure for indifference-zone selection in simulation. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 11(3):251–273, 2001.

[54] H. T. Kung, F. Luccio, and F. P. Preparata. On finding the maxima of a set of vectors. *Journal of the ACM*, 22:469–476, 1975.

[55] L. H. Lee, E. P. Chew, S. Teng, and Y. Chen. Multi-objective simulation-based evolutionary algorithm for an aircraft spare parts allocation problem. *European Journal of Operational Research*, 189(2):476–491, 2008.

[56] L. H. Lee, E. P. Chew, S. Teng, and D. Goldsman. Optimal computing budget allocation for multi-objective simulation models. In *Proceedings of the 2004 Winter Simulation Conference*, pages 586–594, Washington, D.C., 2004.

[57] L. H. Lee, E. P. Chew, S. Teng, and D. Goldsman. Finding the non-dominated Pareto set for multi-objective simulation models. *IIE Transactions*, 42(9):656–674, 2010.

[58] Y. Li. *Optimizing Hepatitis C Birth-Cohort Screening and Treatment Allocation Strategy.* Master thesis, Department of Industrial and Systems Engineering, University of Washington, Seattle, Washington, 2015.

[59] Y. Li, H. Huang, S. Liu, and Z. B. Zabinsky. Optimizing hepatitis C screening and treatment allocation strategy. In *2015 INFROMS Annual Meeting*, Philadelphia, 2015.

[60] Y. Li, H. Huang, Z. B. Zabinsky, and S. Liu. Optimizing implementation of hepatitis c birth-cohort screening and treatment strategies: Model-based projections. *Manuscript submitted*, 2016.

[61] R. C. Lin, M. Y. Sir, and K. S. Pasupathy. Multi-objective simulation optimization using data envelopment analysis and genetic algorithm: Specific application to determining optimal resource levels in surgical services. *Omega*, 41(5):881–892, 2013.

[62] S. Liu, L. E. Cipriano, M. Holodniy, and J. D. Goldhaber-Fiebert. Cost-effectiveness analysis of risk-factor guided and birth-cohort screening for chronic hepatitis C infection in the United States. *PLoS One*, 8:3:e58975, 2013.

[63] A. Marshall, C. Vasilakis, and E. El-Darzi. Length of stay-based patient flow models:

recent developments and future directions. *Health Care Management Science*, 8(3):213–220, 2005.

[64] C. L. Moore and J. A. Copel. Point-of-Care Ultrasonography. *New England Journal of Medicine*, 364(8):749–757, 2011.

[65] A. Murthy, G. De Angelis, D. Pittet, J. Schrenzel, I. Uckay, and S. Harbarth. Cost-effectiveness of universal MRSA screening on admission to surgery. *Clinical Microbiology and Infection*, 16(12):1747–1753, August 2010.

[66] M. Najafzadeh, K. Andersson, W. H. Shrank, A. A. Krumme, O. S. Matlin, T. Brennan, J. Avorn, and N. K. Choudhry. Planning for HIV screening, testing, and care at the Veterans Health Administration. *Operations research*, 62:2:287–304, 2015.

[67] B. L. Nelson, J. Swann, D. Goldsman, and W. Song. Simple procedures for selecting the best simulated system when the number of alternatives is large. *Operations Research*, 49(6):950–963, 2001.

[68] V. I. Norkin, G. C. Pflug, and A. Ruszczynski. A branch and bound method for stochastic global optimization. *Mathematical Programming*, 83:425–450, 1998.

[69] S. Ólafsson. Two-stage nested partitions method for stochastic optimization. *Methodology and Computing in Applied Probability*, 6:5–27, 2004.

[70] A. Oliver and C. Pritchard. Economic evaluations relating to diabetes: a descriptive review and their compliance with guidance. *Value Health*, 3 Suppl 1:7–14, 2000.

[71] G. Persad, A. Wertheimer, and E. J. Emanuel. Principles for allocation of scarce medical interventions. *The Lancet*, 373(9661):423–431, 2009.

[72] S. Petrou and J. Wolstenholme. A review of alternative approaches to healthcare resource allocation. *Pharmacoeconomics*, 18:33–43, 2000.

[73] J. Pintér. Globally optimized calibration of environmental models. *Annals of Operations Research*, 25(1):211–221, 1990.

[74] Y. Prasetio. Simulation-based optimization for complex stochastic systems. *Doctoral dissertation, University of Washington*, 2011.

[75] M. S. Rauner, W. J. Gutjahr, K. Heidenberger, J. Wagner, and J. Pasia. Dynamic policy modeling for chronic diseases: metaheuristic-based identification of pareto-optimal screening strategies. *Operations Research*, 58:5:1269–1286, 2010.

[76] P. P. Reid, W. D. Compton, J. H. Grossman, and G. Fanjiang. *Building a Better Delivery System: a New Engineering/Health Care Partnership*. The National Academies Press, Washington, D.C., 2005.

[77] Y. Rinott. On two-stage selection procedures and related probability-inequalities. *Communications in Statistics-Theory and Methods*, 7(8):799–811, 1978.

[78] W. H. Rogowski, S. D. Grosse, J. Schmidtke, and G. Marckmann. Criteria for fairly allocating scarce health-care resources to genetic tests: which matter most? *Eur J Hum Genet*, 2013.

[79] S. M. Shechter, C. L. Bryce, O. Alagoz, J. E. Kreke, J. E. Stahl, A. J. Schaefer, D. C. Angus, and M. S. Roberts. A clinically based discrete-event simulation of end-stage liver disease and the organ allocation process. *Med Decis Making*, 25(2):199–209, 2005.

[80] L. Shi and S. Ólafsson. Nested partitions method for stochastic optimization. *Methodology and Computing in Applied Probability*, 2 (3):271–291, 2000.

[81] L. Shi and S. Ólafsson. *Nested Partitions Method, Theory and Applications*. Springer, New York, 2009.

[82] D. Strech, G. Persad, G. Marckmann, and M. Danis. Are physicians willing to ration health care? conflicting findings in a systematic review of survey research. *Health Policy*, 90(2-3):113–24, 2009.

[83] G. Symon, K. Long, and J. Ellis. The coordination of work activities: Cooperation and conflict in a hospital context. *Computer Supported Cooperative Work (CSCW)*, 5(1):1–31, 1996.

[84] S. A. Teefey, D. A. Rubin, W. D. Middleton, C. F. Hildebolt, R. A. Leibold, and K. Yamaguchi. Detection and quantification of rotator cuff tears. *The Journal of Bone & Joint Surgery*, 86(4):708–716, 2004.

[85] E. Tekin and I. Sabuncuoglu. Simulation optimization: A comprehensive review on theory and applications. *IIE Transactions*, 36(11):1067–1081, 2004.

[86] USPSTF. CDC advise screening baby boomer birth cohort for hepatitis C. Technical report, CDC, 2013.

[87] M. Vlychou, Z. Dailiana, A. Fotiadou, M. Papanagiotou, I. V. Fezoulidis, and K. Malizos. Symptomatic partial rotator cuff tears: diagnostic performance of ultrasound and magnetic resonance imaging with surgical correlation. *Acta radiologica (Stockholm, Sweden : 1987)*, 50(1):101–5, January 2009.

[88] W. Wang. Adaptive random search for noisy and global optimization. *Doctoral dissertation, University of Washington*, 2011.

[89] M. L. Weng and A. A. Houshmand. Healthcare simulation: a case study at a local clinic. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, editors, *Proceedings of the 1999 Winter Simulation Conference*, pages 1577–1584, Phoenix, Arizona, 1999.

[90] WHO. World health statistics 2011. Technical report, World Health Organization, Geneva, 2011.

[91] J. Xu, S. Zhang, E. Huang, C. H. Chen, and L. H. Lee. Efficient multi-fidelity simulation optimization. In S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, editors, *Proceedings of the 2014 Winter Simulation Conference*, pages 46–57, Savannah, Georgia, 2014.

[92] W. L. Xu and B. L. Nelson. Empirical stochastic branch-and-bound for optimization via simulation. *IIE Transactions*, 45:7:685–698, 2013.

[93] Z. B. Zabinsky. Stochastic methods for practical global optimization. *Journal of Global Optimization*, 13:433–444, 1998.

[94] Z. B. Zabinsky, W. Wang, Y. Prasetio, A. Ghate, and J.W. Yen. Adaptive probabilistic branch and bound for level set approximation. In S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, editors, *Proceedings of the 2011 Winter Simulation Conference*, pages 46–57, Phoenix, Arizona, 2011.