5-2019

Approximate Dynamic Programming: Health Care Applications

Amir Ali Nasrollahzadeh
Clemson University, amiranasrollahzadeh@gmail.com

APPROXIMATE DYNAMIC PROGRAMMING: HEALTH CARE APPLICATIONS

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Industrial Engineering

---

by
Amir Ali Nasrollahzadeh
May 2019

Accepted by:
Dr. Amin Khademi, Committee Chair
Dr. B. Rae Cho
Dr. Scott J. Mason
Dr. Kevin M. Taaffe

Abstract

This dissertation considers different approximate solutions to Markov decision problems formulated within the dynamic programming framework in two health care applications. Dynamic formulations are appropriate for problems which require optimization over time and a variety of settings for different scenarios and policies. This is similar to the situation in a lot of health care applications for which because of the curses of dimensionality, exact solutions do not always exist. Thus, approximate analysis to find near optimal solutions are motivated. To check the quality of approximation, additional evidence such as boundaries, consistency analysis, or asymptotic behavior evaluation are required. Emergency vehicle management and dose-finding clinical trials are the two heath care applications considered here in order to investigate dynamic formulations, approximate solutions, and solution quality assessments. The dynamic programming formulation for real-time ambulance dispatching and relocation policies, response-adaptive dose-finding clinical trial, and optimal stopping of adaptive clinical trials is presented. Approximate solutions are derived by multiple methods such as basis function regression, one-step look-ahead policy, simulation-based gridding algorithm, and diffusion approximation. Finally, some boundaries to assess the optimality gap and a proof of consistency for approximate solutions are presented to ensure the quality of approximation.

Acknowledgments

First, I would like to thank my advisor, Dr. Amin Khademi, for his support, his mentorship, his patience, and the wonderful opportunities he created for me. He has been a wonderful advisor without whom completing a PhD degree would have not been this rewarding.

I would also like to thank my committee members, Dr. B. Rae Cho, Dr. Scott J. Mason, and Dr. Kevin M. Taaffe, for their continued support and encouragement, and for the constructive feedback they provided throughout my graduate studies. They set an example that I hope to emulate in future. I am also grateful to the Industrial Engineering department faculty, staff and my fellow graduate students for creating an open atmosphere full of motivation and opportunities.

I also like to particularly thank Dr. Maria E. Mayorga, and Dr. J. Cole Smith, for helping me write papers, prepare manuscripts, and for creating professional opportunities in academic life.

I am also grateful to my parents, Abolfazl Nasrollahzade and Razieh Adibi, my brother, Mohammad Nasrollahzadeh, and my friends Farhad Hasankhani, Maziar Fooladi, Mohammad Afkhami and Shervin Gholizade, for their love, friendship and support which makes studying abroad and away from home far easier than it is.

Table of Contents

List of Tables

List of Figures

Preface

It is safe to say that in almost every nation, health care systems are inadequate, inefficient and often costly to meet the demands. In the United States, a high-income country, 17.8% of GDP is spent on health care as of 2015 (National Center for Health Statistics 2016), nearly twice the amount of any other high-income or industrialized country. However, the outcomes do not justify such high costs and a plethora of studies and reports, which often include measures of performance such as access, equity, and responsiveness as well as measures of health, consistently rank the U.S. health care system as one of the least efficient systems among the most advanced industrialized countries (e.g., Schneider *et al.* 2017, World Health Organization 2000). Although, the challenges facing low and high-income nations are different, inefficiencies in health care systems such as inequality in access results in disturbing outcomes: the life expectancy gap between poor and rich people in some high-income countries (including the U.S.) is greater than the average life expectancy gap between low and high-income nations (Dwyer-Lindgren *et al.* 2017, World Health Organization 2015). The need for a more adequate, efficient and accessible health care system persists in both rich and poor nations. Design and implementation of a system able to deliver quality service given limited resources requires careful management and decision making. Brandeau *et al.* (2004) and Rais & Viana (2011) survey a wide range of applications in health care operations management and clinical practices where operation research methods support decision making procedures. In this dissertation, the focus is on the applications of dynamic programming methods in health care systems.

Dynamic programming formalization. Optimizing a system over time and a series of settings which usually requires a variety of actions arises in many situations in health care management. These situations often involve a sequential pattern between decision making and observing information with uncertainty, where a decision or an action at one period results in a probabilistic transition

from the current state to another state in the next period. Markov decision processes (MDPs) are a general method to model such dynamics under uncertainty. Dynamic programming formalizes a framework for optimization of a performance criterion with respect to current, past and uncertain future states of MDP models and binds the decisions made at each time into a policy. In fact, typical sources of uncertainty in health care systems such as different responses of patients to the same treatments requires the transitions to a new state to be different at each decision epoch. The uncertain transitions are accommodated within the dynamic programming framework since the transition probabilities governing the stochastic processes are permitted to be dependent on the decision at each epoch (Schaefer *et al.* 2005).

Because of this inherent flexibility in the dynamic programming framework, the literature on its use in health care is rich and covers a wide range of applications. For example, Green *et al.* (2006) and Patrick *et al.* (2008) applied dynamic programming frameworks to capacity allocation in diagnostic facilities and patient management in hospitals where facing diminishing government subsidies, the diagnostic facilities and hospitals are under immense pressure to reduce costs (Green 2005). Haijema *et al.* (2007) developed a dynamic programming approach for blood platelet production and inventory problem while considering complicating factors such as multiple types of demand and production lead time. The approach differs from classic supply chain and inventory management problems since blood products are perishable and their demands and supplies are random (Pierskalla 2005). Zaric & Brandeau (2001, 2002) presented dynamic programming formulations to allocate limited resources among competing prevention and treatment programs in controlling epidemics of infectious diseases. Zenios *et al.* (2000) and Alagoz *et al.* (2004) studied dynamic allocation schemes in organ transplants while considering the trade-offs between equity and efficiency. Maillart *et al.* (2008), Lee *et al.* (2008), and Khademi *et al.* (2015) investigated the screening and treatment practices and policies in breast cancer, dialysis therapy, and HIV treatments, respectively, by using dynamic programming methods. Wu *et al.* (2005) and Hall *et al.* (2008) provided dynamic programming solutions to vaccine formulary selection problem in immunization programs. Dynamic programming formulations to ambulance service planning (or emergency vehicle routing), one of the earliest applications of operation research in health care is discussed in Chapter 1. Response-adaptive design of dose-finding clinical trials and optimal stopping of adaptive dose-finding clinical trials are discussed in Chapters 2 and 3, respectively.

Exploiting the flexibility of dynamic programming framework in terms of the state space, the

action space, and the transition probabilities, all of which allow to capture the complexity of health care problems, comes with a cost: dynamic models tend to be extremely information intensive and thus become harder to solve exactly for real world problems (Schaefer *et al.* 2005). This problem, usually referred to as "the curses of dimensionality", arises because the general algorithm to solve a dynamic program involves evaluating a recursive value function for each decision epoch across the state and action spaces. As dimensions of a problem increase, the number of possible states, actions, or transitions grow exponentially and evaluating the value function becomes intractable (Powell 2007). Therefore, a fertile research area known as "approximate dynamic programming" (ADP) has been developed recently to address these issues and produce implementable high quality solutions.

Approximate dynamic programming. Since backward recursions to solve the value functions are not tractable if the state and action spaces are multidimensional, an alternative strategy is to iteratively step forward in time and estimate the value function. Using this estimation, a decision is made to optimize the estimated value function followed by a random transition to a new state generated in a Monte Carlo sample path simulation. Iterating over this procedure for a large enough set of sample paths allows the decision maker to approximate the value of a policy by taking a sample average (Powell 2007).

There are several different algorithms in the ADP literature which are able to estimate a value function in absence of proper information as a result of moving forward in time. Using such algorithms provides an upper (lower) bound for the true optimal value of the value function which is useful in comparing a range of policies with the current practices in health care applications. It can be shown that some of these algorithms converge to optimal solutions thus the quality of solutions are only dependent on computational power. However, the quality of solutions for some algorithms with unknown convergence properties remains in question. For such problems, assessing the quality of solution is analyzed by developing a bounding system and measuring the optimality gap.

In Chapter 1 of this dissertation, the ambulance dispatching and relocation problem is investigated in order to develop new policies for ambulance service planning in an emergency medical service (EMS) system. These policies determine, (i) which ambulance to dispatch when a call arrives to the system, (ii) is it beneficial to relocate an ambulance from one base to another such that a certain coverage level is maintained, and (iii) what to do with an ambulance which has finished its service. The problem is modeled within a dynamic programming framework and the solutions

are derived by an approximate dynamic programming method which uses a linear combination of functions to estimate the true value function. The linear combination of basis functions is tuned by a multiple linear regression model. Numerical evaluations of performance measures, e.g., the expected response time or the average fraction of lost calls, provide evidence that dynamic approximate policies such as considering multiple bases for redeploying an ambulance which has finished its service instead of sending the ambulance back to its original base significantly improves the performance of the system. The quality of solutions across multiple performance measures are assured by developing a lower bounding system which is used for quantifying the optimality gap. These solutions offer managerial insights for ambulance movements in an EMS system.

Chapter 2 studies the response-adaptive design of dose-finding clinical trials in which a number of volunteers are assigned to different dosage levels to identify a target dose. The policies should determine how to modify design elements such as allocation schemes at each decision epoch based on data collected so far to learn the target dose more efficiently. A state-of-the-art and a proposed approach for the design of dose-finding trials with unknown dose-response relationships are formulated using dynamic programming framework. The solutions to these dynamic models are computed via a "one-step look-ahead" policy which estimates the value function one decision epoch into the future by evaluating it for a large number of sample paths, i.e., a Monte Carlo simulation of value function. Several performance measures such as the variance of the target dose at the end of the trial and patient assignment patterns are derived to show a more efficient design of dose-finding clinical trial is possible without sacrificing the accuracy in learning the unknown dose-response relationship and thus the target dose while the trial is still in progress. However, the convergence properties of the state-of-the-art approach is not known and thus only upper bound solutions exist which do not ascertain if the true target dose will be eventually identified. This is not the case for the proposed approach. The consistency proof implies that the approximate dynamic programming policies are able to eventually learn the unknown dose-response relationship and thus the target dose with certainty.

Chapter 3 considers another aspect of adaptive designs in dose-finding clinical trials. In the previous chapter, it was shown that an adaptive design to identify a target dose of an unknown dose-response relationship is able to eventually learn the dose-response and thus the target dose. However, sampling more and more participants in the trial increases the costs and may not be necessary if enough evidence is already gathered. Therefore, an optimal stopping problem is motivated to

determine whether to (i) abandon the trial due to a lack of positive evidence about a significant improvement in the target dose versus placebo, (ii) continue the trial for a more promising significant improvement, (iii) or terminate the trial because enough evidence about a significant improvement is gathered and sampling more participants would only increase the costs. New information after sampling each participant is transformed to financial evidence in order to evaluate each decision in a dynamic programming framework. Multiple approximate solutions to the dynamic formulation, i.e., simulation-based gridding algorithm, one-step look-ahead policy, and diffusion approximation, are compared to show that there are cases for which classic approaches like the simulation-based gridding, and the widely used one-step look-ahead policies fall apart in finding high quality solutions. In particular, when the true dose-response curve is flat, both simulation-based algorithm and the one-step look-ahead policy prematurely terminate the trial half the times. In order to improve their performance, a statistical check measuring the accuracy of unknown dose-response curve estimation is added to both algorithms. However, numerical results show that the diffusion approximation method still outperforms both approximate procedures in terms of correctly deciding to abandon, continue, or terminate the trial.

Chapter 1

Real-Time Ambulance Dispatching and Relocation

Summary. In this chapter, we develop a flexible optimization framework for real-time ambulance dispatching and relocation. In addition to ambulance redeployment, we consider a general dispatching and relocation strategy by which the decision maker has the option to (i) select any available ambulance to dispatch to a call or to queue the call, and (ii) send an idle ambulance to cover the location of an ambulance just dispatched to a call. We formulate the problem as a stochastic dynamic program and because the state space is unbounded, an approximate dynamic programming (ADP) framework is developed to generate high-quality solutions. We assess the quality of our solutions by developing a lower bound on expected response time, and computing a lower bound on the expected fraction of late calls of any relocation policy. We test the performance of our policies and available benchmarks on an emergency medical services system in Mecklenburg County, NC. The results show that our policies are near-optimal and significantly outperform available benchmarks. In particular, our ADP policy reduces the expected response time and fraction of high priority late calls by 12% and 30.6% over the best available static benchmark in the case study. Moreover, they provide insights on the contribution of each dispatching, redeployment, and reallocation strategy.

Published: Nasrollahzadeh, A., Khademi, A., Mayorga, M., "Real-Time Ambulance Dispatching and Relocation", Manufacturing & Service Operations Management, 20(3): 467-480

1.1 Introduction

Motivation. Emergency medical services (EMS) provide out-of-hospital acute medical care and transport the sick or injured to hospitals for definitive care. Typically, EMS providers' performance is evaluated based on their response time (National Association of State EMS Officials n.d.), the amount of time that an ambulance takes to arrive to the scene of a call once the call is received, as reducing the response time is an essential factor in lowering patient mortality rates (Wilde 2013). In particular, a target for the proportion of urgent calls whose response time is less than a threshold is a common measure of performance. For example, the U.S. National Fire Protection Association suggests a target that 90% of emergency medical calls be reached by a first responder within four minutes followed by an Advanced Life Support response within eight minutes (NFPA 2010). Also, in North America a common target is reaching 90% of urgent urban calls within nine minutes (Fitch 2005).

Factors such as increased non-emergency calls, which (by law) require that an ambulance be dispatched, and insufficient funding have increased pressure on EMS providers to "do more with less" or, at best, to use the same level of resources to achieve response time targets set by municipalities or contracts (Ward 2014). This has spurred EMS providers to better manage their ambulances by using more complex dispatching and location policies. Studies of realistic settings show that the performance of static policies, those that send the closest ambulance and preassign a location to each ambulance, can be quite poor (Maxwell *et al.* 2010). Recently, the availability of real-time information to dispatchers via geographical information systems, and the affordability of computing power has facilitated using real-time ambulance management, which provides a platform that enables EMS providers to consider more sophisticated operational strategies to improve the performance of ambulance deployment policies. One potential strategy is ambulance relocation, which refers to repositioning idle ambulances in real-time to better respond to future calls. It is possible for some locations to be covered by more than one ambulance, therefore some ambulances might be idle at their locations while providing no additional coverage value. Note that an area is covered if an idle ambulance can reach it in a specific time threshold. Repositioning these ambulances to improve the coverage level is a strategy we call "ambulance reallocation." This strategy can improve the performance of EMS systems because idle ambulances at other locations can compensate for a "coverage hole" caused by dispatching the only ambulance covering a region. A second type of

strategy is to send an ambulance that just finished service to a new location rather than sending it to a preassigned base, which we call "ambulance redeployment." A third potential strategy is to decide which ambulance should serve a call (if immediately), which we call "ambulance dispatching," as sending the closest ambulance for every incident may be suboptimal (Swersey 1994). This strategy can significantly improve the performance of the system as supported by our numerical analysis. For example, suppose a high priority call arrives and the closest ambulance is 20 minutes away from the call location, however, another ambulance is currently in service just two minutes away from the call location and will be available in three minutes. Sending the closest ambulance immediately will result in a late call in this example and is shown to be suboptimal in realistic settings. We intend to develop a flexible mathematical framework to explore a variety of strategies for real-time ambulance operations management. Pursuant to this goal, we formulate the problem as a stochastic dynamic program and use an approximate dynamic programming (ADP) approach to produce efficient real-time dispatching and relocation policies.

Main Contributions and Results. In this chapter, we make the following contributions: (1) We develop a flexible optimization framework by simultaneously considering general dispatching, redeployment, and reallocation strategies for real-time stochastic dynamic ambulance operations management. We consider a general dispatching rule upon receiving a call in that the decision maker can send any available ambulance in addition to not serving the call immediately. Therefore, we let the model decide which ambulance should immediately be dispatched to a received call or the call has to wait for an ambulance in the near future. EMS providers are also motivated to spread out ambulances on the roads to meet the performance standards. To improve coverage level, we consider relocating an available ambulance to the location of an ambulance just dispatched to serve a call and we name it "ambulance reallocation."(2) In order to assess the quality of solutions produced by our ADP approach, we develop a novel lower bound on the expected response time of any relocation policy. To create this bound, we consider a lower bounding system as in Maxwell *et al.* (2014). However, instead of solving a maximum covering location problem, upon receiving a call in the original system, we reposition the available ambulances to minimize the expected response time by solving a different p-median integer program. (3) We develop new basis functions which estimate the expected response time of the calls in the system and modified some of the available basis functions in the literature to enhance the performance of the ADP policies. In particular, we introduce new basis

functions that estimate a future state of the system in which a busy ambulance becomes available, thus enabling the ADP algorithm to react to the future coverage level. These basis functions serve an important purpose in that, they allow the algorithm to delay or alter a dispatching or relocation decision in response to a situation by considering future costs of an appropriate response when a new ambulance configuration has emerged. (4) We measure the contribution of each strategy in terms of a variety of objective functions such as the expected discounted priority-adjusted response time and the expected discounted priority-adjusted fraction of late calls. We discover insights regarding the relative contribution of each strategy, as well as available benchmarks.

We test the performance of six static benchmarks in the literature on our data set to find the best static benchmark in terms of expected response time and fraction of late calls. Our analysis shows that Maximum Expected Covering Location Problem (MEXCLP) and Maximum Covering Location Problem (MCLP) outperform other static benchmarks when the objective is to minimize the expected fraction of late calls and the expected response time, respectively. Thus, the static policy, hereafter, refers to MEXCLP (MCLP) when the objective is to minimize the fraction of late calls (response time).

In addition, we consider five dynamic benchmarks, including a heuristic which has been reported to be efficient in the literature. In order to analyze the contribution of each dispatching, redeployment, and reallocation strategy on performance improvement, we design three dynamic benchmarks by adding each strategy to the static policy one at a time. That is, Benchmark 1 builds on the static policy by considering a general dispatching rule instead of sending the closest available ambulance; Benchmark 2 builds on the static policy by considering a redeployment strategy after an ambulance has finished serving a call; Benchmark 3 builds on the static policy by sending an available ambulance to the location of an ambulance just dispatched to serve a call. Benchmark 4, which consists of redeployment and reallocation strategies, is used to compute the optimality gap as both lower bounds assume the closest ambulance is dispatched; Benchmark 5 uses the dynamic heuristic relocation policy proposed by Jagtenberg *et al.* (2015) to evaluate its performance with respect to our relocation strategies.

Our results show that when the ADP objective function is to minimize the expected response time, ADP policies generated by Benchmarks 1, 2, and 3 outperform the MCLP static benchmark by 2.7%, 6.8%, and 1.3%, respectively. However, when the ADP objective function is to minimize the expected fraction of late calls, the ADP policies produced in Benchmarks 1, 2, and 3 outperform

the MEXCLP static benchmark by 13.5%, 21.3%, and 9.8%, respectively. This shows that each strategy can significantly improve the static benchmarks. Note that the contribution of a redeployment strategy is significantly greater than that of dispatching and reallocation strategies. Also, this observation is consistent in both ADP objective functions, i.e., minimizing response time and fraction of late calls. Furthermore, the expected frequency that Benchmark 2 deviates from the static benchmarks is significantly greater than other dynamic benchmarks. The ADP approach that simultaneously considers all three strategies outperforms both the static and dynamic benchmarks. In particular, when the ADP objective is to minimize the expected fraction of late calls, our ADP approach outperforms Benchmark 2, in expected response time and fraction of late calls by 4.3% and 14.5%, respectively. Benchmark 2 is similar to the setting studied in Maxwell *et al.* (2010), where only ambulance redeployment is considered. Our results suggest that expanding the action space beyond redeployment can significantly improve the performance of the system, e.g., 14.5% improvement in the fraction of late calls when the ADP objective is to minimize the fraction of late calls. Furthermore, Benchmark 1, which uses a general dispatching rule, provides novel insights to the discussion around optimality of policies that deviate from sending the closest available ambulance. Our results show that Benchmark 1 simultaneously reduces the fraction of late calls and response time as our ADP policy shifts the entire response time distribution toward shorter times (see Figure 2.3). This result is different from Jagtenberg *et al.* (2016) where deviating from sending the closest ambulance resulted in an improvement on fraction of late calls, but significant increase in response time.

1.2 Background

The literature on ambulance operations management is quite rich. Therefore, we briefly discuss previous works related to this chapter and refer the reader to the following survey papers and the references therein for a comprehensive review. Swersey (1994) and Brotcorne *et al.* (2003) reviewed deterministic and probabilistic ambulance location and relocation models. Also, Ingolfsson (2013) provided a survey on the analytical stochastic models focusing on ambulance station selection and ambulance allocation to stations with respect to performance measures such as response time.

Early models of ambulance location problem seek to minimize the number of ambulances required to respond to future calls for a determined time threshold or to maximize the demand

covered using a fixed fleet size; see Church & ReVelle (1974) and the references therein. These approaches did not consider the fact that when an ambulance is dispatched, the coverage level might fall below a minimum threshold. One possibility to address the unavailability of dispatched ambulances over time includes considering multiple coverage, i.e., demand points that are supposed to be covered by more than one vehicle. Gendreau *et al.* (1997) introduced the double standard model by including multiple coverage. Doerner *et al.* (2005) extended their work with respect to capacity constraints and different demand density in each location. Gendreau *et al.* (2001, 2006) developed dynamic models to formulate the ambulance repositioning problem, where the objective function is to maximize the total covered demand. Because these approaches require solving an integer program every time the dispatcher makes a decision, they are computationally very intensive. Also, these models are deterministic and do not capture the effect of randomness in the system.

Berman (1981a,b) used Markov decision theory to minimize the long-run cost of repositioning ambulances. They provided an exact dynamic programming approach to find available ambulances to compensate for coverage level drop induced by dispatched ambulances. However, these exact formulations are only tractable in oversimplified settings for a few number of ambulances over a small network of routes. Restrepo *et al.* (2009) used an Erlang loss function to compute the fraction of late calls, those not responded to within a time threshold, and embedded it into an optimization model to minimize the percentage of late calls by static deployment of ambulances. McLay & Mayorga (2013) formulated the ambulance dispatching problem as a Markov decision process to optimally dispatch ambulances to prioritized patients. Alanis *et al.* (2013) developed a two-dimensional Markov chain model to analyze ambulance repositioning according to a compliance table, which suggested where to reposition an ambulance based on the number of available ambulances. (In this setting, the closest ambulance is dispatched to serve the call. After the ambulance finishes its service, the decision maker seeks to reposition ambulances in such a way that maintains the configuration of the available ambulances similar to the one suggested by the compliance table.) Andersson & Varbrand (2007) measured the ability of an ambulance to cover a future call by introducing a "preparedness function," which approximates the value function in a dynamic program. However, in order to apply it to real-time applications even with small sets of available ambulances and relocation destinations, the dynamic relocation problem is solved heuristically. van Barneveld *et al.* (2016) designed a heuristic dynamic repositioning policy by minimizing the "unpreparedness function," which returns the expected penalty that the next request generates. In that setting, the

best relocation policy is found in terms of a "motion" from an origin base to a destination base. To prevent long transition times, a linear bottleneck assignment problem is solved to determine how available ambulances should move to reach the new configuration.

Mason (2013) developed a dynamic repositioning policy, which relocated ambulances in demand zones when coverage levels dropped below a threshold. However, repositioning idle ambulances every time coverage levels fall below a certain level may result in shortage of available ambulances at times of dispatch. Therefore, to limit the repositioning time, a neighborhood search strategy is developed to solve the base allocation problem, which leads to solutions that differ only slightly from the initial base locations. Jagtenberg *et al.* (2015) developed a heuristic approach to real-time ambulance relocation by maximizing the expected marginal contribution of each available ambulance to the coverage level. van Barneveld (2016) extended MEXCLP to incorporate a non-negative nondecreasing function of response time into the objective function to calculate performance measures related to response time instead of coverage level. By solving the extended formulation for different levels of available ambulances, compliance tables are obtained offline and when the number of available ambulances changes, an assignment problem is carried out to reconfigure the system, i.e., move the ambulances to new positions according to the compliance tables. Sudtachat *et al.* (2016) modified the steady state probabilities calculated in Alanis *et al.* (2013) and incorporated them into an integer program to maximize the coverage level in a single type ambulance and call priority system with zero-length queue. The resulting nested compliance policy, when out of compliance, requires at most one vehicle movement at a time to reconfigure accordingly. Bélanger *et al.* (2016) modified the double standard model to consider multi-period and dynamic settings with or without relocations. In the multi-period settings, ambulances were only relocated between the periods and returned to the same base throughout the period. In the dynamic settings, the double standard model is solved whenever an ambulance is dispatched if certain time has passed since the last relocation and a secondary coverage level falls below a threshold.

To make ambulance redeployment decisions in an uncertain dynamic setting, Maxwell *et al.* (2010) developed an ADP approach based on approximate policy iteration. They formulated the ambulance redeployment problem as a dynamic program and approximated the value function by an affine combination of basis functions. They used an iterative simulation-based procedure to estimate tunable parameters of the approximation. The objective is to minimize the fraction of late calls only through redeployment. Their model, however, does not consider ambulance reallocation and uses a

myopic dispatching rule, i.e., the closest ambulance is sent to a call, calls are served in decreasing order of priority, and a first-come first-served strategy is considered for each priority. Schmid (2012) also used ADP to real-time ambulance dispatching and relocation. Our approach is different both in problem scope and methodology used. In particular, Schmid (2012) did not consider the relocation of idle ambulances and assumed that an ambulance must be immediately dispatched to a call. In terms of ADP, Schmid (2012) used a general ADP framework based on aggregation and post-decision states. However, we develop an ADP approach specific to ambulance operations management by exploiting novel basis functions, as well as developing a lower bound on expected response time.

There is another stream of literature related to approximate dynamic programming. Many researchers have used ADP to come up with high-quality solutions for a variety of applications, e.g., allocating resources in service systems (Adelman 2007); resource allocation in healthcare (Bertsimas *et al.* 2013, Khademi *et al.* 2015); and supply chain management (Lai *et al.* 2010, Van Roy *et al.* 1997).

1.3 Problem Formulation

This section presents an infinite-horizon Markov decision process formulation of the problem. Let $\mathcal{L} := \{0, 1, 2, \ldots, L\}$ be the set of call locations and $\mathcal{B} := \{0, 1, 2, \ldots, B\}$ be the set of all ambulance bases. We assume a total of N ambulances are available and at most J calls are tracked. This assumption is not restrictive because one may consider a large J.

State space. An ambulance i is represented by $m_i = (f_i, o_i, d_i, t_i)$, where f_i is the status of the ambulance, o_i is the original location of the ambulance, d_i is the destination of the ambulance, and t_i is the start time of the latest movement of the ambulance. For the purposes of this chapter, it is sufficient to consider six possibilities for the status of an ambulance, i.e., $f_i \in \{0, 1, 2, 3, 4, 5\}$, where 0 shows that the ambulance is available at base, 1 shows that the ambulance is going to a call location, 2 shows that the ambulance is serving a call on scene, 3 shows that the ambulance is going to hospital, 4 shows that the ambulance has finished serving a call, and 5 shows that the ambulance is being reallocated and going to another base or the ambulance is going to a base after finishing service. Note that if ambulance i is idle in a location, the original location is set to the current location and the destination to null. Similarly, when an ambulance is serving a call on scene, we set the original

location to the call location and destination to null. We let vector $m = (m_1, m_2, \ldots, m_N) \in M$ represent the state of all ambulances. A call j is represented by $c_j = (g_j, l_j, p_j, q_j)$, where g_j is the status, l_j is the location, p_j is the priority, and q_j is the arrival time of the call. In particular, $g_j \in \{0, 1\}$, where 0 shows that the call is waiting for service and 1 shows that the call is assigned to an ambulance. When an ambulance reaches the call scene, the call is removed from the list. Aligned with literature, we consider two priority levels for a call, $p_j \in \{0, 1\}$, where 0 shows that the priority of a call is low, and 1 shows that the priority of the call is high (Maxwell *et al.* 2010). Extending the framework presented in this chapter to consider more priority levels is straightforward. We let vector $c = (c_1, c_2, \ldots, c_J) \in C$ represent the state of all calls.

Without loss of generality, we assume that decisions are made at transition times. In our model, transition times are associated with the following events: "call j arrives," "ambulance i is in transit to call j," "ambulance i arrives at the location of call j," "ambulance i is finished serving call j at scene," "ambulance i is finished serving call j at hospital," and "ambulance i arrives at a base." Let E be the set of all possible events. Therefore, the state space of the system is represented by $S := \{s = (\tau, e, m, c) : e \in E, m \in M, c \in C\}$, where τ corresponds to the current time.

Action space. The action space is described in four cases. We assume that dispatching, reallocating, and redeploying the ambulances are non-preemptive. One can relax this assumption by defining an event "consider preemption," which occurs with a certain frequency and upon occurrence, one may preempt any of the ambulance services and reconsider actions. However, Maxwell *et al.* (2010) showed that considering only the service preemption of ambulances that are returning to base significantly increases the computational effort while its benefit may be marginal.

Case 1: If call j arrives, the decision maker has two types of decision: (i) which ambulance should be immediately dispatched to serve the call (if any), and (ii) which ambulances should be reallocated to other bases (if any). Note that in this case, an ambulance is not necessarily dispatched upon receiving a call immediately. If this happens, the call will join the queue and will be served later. The rationale for considering reallocation decisions is that by spreading out the ambulances over the area it is likely that a location is only covered by one ambulance, thus, sending the ambulance to a call may cause a coverage hole. Since coverage level does not decrease unless an idle ambulance becomes unavailable, reallocation decisions are considered when an ambulance is dispatched. Because multiple reallocations in short intervals are expensive and could become a

burden on the ambulance crew (van Barneveld *et al.* 2016, Jagtenberg *et al.* 2015), we assume that reallocations are limited to at most one ambulance upon dispatching an ambulance. Let $\mathcal{M}(s)$ be the set of available ambulances, i.e., $\mathcal{M}(s) := \{i : f_i = 0\}$, $\mathcal{B}_1(s)$ represent the location of the ambulance just dispatched to a call, and $\mathcal{M}_1(s)$ represent the set of all available ambulances right after dispatching ambulance i when the state is s. If no ambulance is dispatched upon receiving a call, we set $\mathcal{B}_1(s) = \emptyset$ and do not consider ambulance reallocation. Note that when ambulances are in transit towards a base ($f_i = 5$), they are not considered available due to the non-preemption assumption. It is possible to use the event "consider preemption" to preempt an ambulance that is moving towards a base and dispatch it to a call. However, the benefit of such preemptions may be marginal (Maxwell *et al.* 2010).

Define $X_{i,j} = 1$ if ambulance i is assigned to call j, and $X_{i,j} = 0$, otherwise. Also, define $Y_{i,b} = 1$ if ambulance i is reallocated to location b, and $Y_{i,b} = 0$, otherwise. Therefore, if event e is of the type "a call arrives," and $\mathcal{J}_1(s)$ denotes a set that points to the index of the call, the action space is given by

$$A_1(s) := \left\{ (X_{i,j}, Y_{i,b}) : \sum_{i \in \mathcal{M}(s)} X_{i,j} \leq 1,\ j \in \mathcal{J}_1(s); \sum_{i \in \mathcal{M}_1(s)} Y_{i,b} \leq 1,\ b \in \mathcal{B}_1(s) \right\},$$

where the first constraint ensures that at most one ambulance is assigned to a received call j and the second constraint ensures that at most one ambulance is reallocated to the location of the dispatched ambulance.

Case 2: Let $\mathcal{Q}(s)$ denote the set of all calls waiting in the queue for ambulance assignment, i.e., $\mathcal{Q}(s) := \{j : g_j = 0\}$. If $\mathcal{Q}(s) = \emptyset$ and event e is of the type "ambulance i has finished serving a call at scene," or "ambulance i has finished serving a call at hospital," the decision is where to redeploy the ambulance. Let $\mathcal{M}_2(s)$ denote the set of available ambulance for redeployment in state s and $Z_{i,b} = 1$ if ambulance i is redeployed to location b, and $Z_{i,b} = 0$, otherwise. We set $\mathcal{M}_2(s) = \{i\}$ in this case. The action space is presented by

$$A_2(s) := \left\{ (Z_{i,b}) : \sum_{b \in \mathcal{B}} Z_{i,b} = 1,\ i \in \mathcal{M}_2(s) \right\},$$

where the constraint ensures that ambulance i is redeployed to only one location.

Case 3: If $\mathcal{Q}(s) \neq \emptyset$ and event e is of type "ambulance i has finished serving a call at scene,"

or "ambulance i has finished serving a call at hospital," or "ambulance i has arrived at a base," the decision is to dispatch an available ambulance to a call in the queue, or to redeploy it to a location. If there is more than one call in the queue, the calls are served in decreasing order of priority, and within a given priority level, they are served based on a first-come first-served rule. Let $\mathcal{J}_3(s)$ denote the set that points to the highest priority call with the longest waiting time in the queue and $\mathcal{M}_3(s)$ denote the set of available ambulances for redeployment when the state is s. Set $\mathcal{M}_3(s) = \{i\}$ in this case. The action space is given by

$$A_3(s) := \left\{ (X_{i,j}, Z_{i,b}) : \sum_{i \in \mathcal{M}(s)} X_{i,j} \leq 1,\ j \in \mathcal{J}_3(s);\ \sum_{b \in \mathcal{B}} Z_{i,b} = 1 - X_{i,j},\ i \in \mathcal{M}_3(s), j \in \mathcal{J}_3(s) \right\},$$

where the first constraint considers dispatching an ambulance to call j in the queue, and the second constraint ensures that the ambulance that has just become available will be redeployed to a location if it is not already assigned to a call.

Case 4: If an event is of type "ambulance i is in transit to call j," or "ambulance i arrives at the location of call j," we set $A(s) = \emptyset$.

Transitions. We assume that call arrivals in location l follow a non-homogeneous Poisson process with rate λ_l^τ at time τ. If an ambulance arrives at call j scene, it completes the service at the scene with probability ρ_j, and it will transfer the patient to a hospital with probability $1 - \rho_j$. We assume that travel times are deterministic, and the time required to serve a call at scene or taking a patient to hospital follows an arbitrary distribution with a finite mean, independent of call location. Note that if destination is a hospital, in addition to travel time, our historical data also considers both the service time on scene before going to hospital and the time that it takes to handover the patient to hospital personnel. We estimate all of the distributions by using historical data from Mecklenburg County, NC. Let s_κ be the state of the system when the κth event happens. The evolution of state s_κ can then be characterized by action a_κ, a random element $\omega(s_\kappa, a_\kappa)$, and a function F, i.e., $s_{\kappa+1} = F(s_\kappa, a_\kappa, \omega(s_\kappa, a_\kappa))$.

Objective function. We consider minimizing the expected discounted priority-adjusted total response time and the expected discounted priority-adjusted fraction of late calls as the primary ADP objective functions for the optimization framework. We also report other performance measures

such as response time of late calls and fraction of late high priority calls in our case study. Let $h(s_\kappa, a_\kappa, s_{\kappa+1})$ denote the cost of a transition from s_κ to $s_{\kappa+1}$, when action a_κ is taken. In order to minimize the expected discounted priority-adjusted response time, define

$$h(s_\kappa, a_\kappa, s_{\kappa+1}) = \begin{cases} w_1(\tau(s_{\kappa+1}) - q_j) & \text{A high priority call } j \text{ arrives and the event } e(s_{\kappa+1}) \\ & \text{is of the form "ambulance } i \text{ arrives at the scene of} \\ & \text{call } j\text{,"} \\ w_2(\tau(s_{\kappa+1}) - q_j) & \text{A low priority call } j \text{ arrives and the event } e(s_{\kappa+1}) \\ & \text{is of the form "ambulance } i \text{ arrives at the scene of} \\ & \text{call } j\text{,"} \\ 0 & \text{otherwise,} \end{cases}$$

where $(\tau(s_{\kappa+1}) - q_j)$ measures the response time of call j, and w_1 and w_2 are priority adjustment weights. This cost structure is flexible in that w_1 and w_2 can be tuned to capture the relative importance of high priority versus low priority calls.

Similarly, in order to minimize the long-run priority-adjusted fraction of late calls, define

$$h(s_\kappa, a_\kappa, s_{\kappa+1}) = \begin{cases} w_3\big(\mathbb{1}_{\{\tau(s_{\kappa+1}) - q_j \geq \triangle\}}\big) & \text{A high priority call } j \text{ arrives and the event } e(s_{\kappa+1}) \\ & \text{is of the form "ambulance } i \text{ arrives at the scene of} \\ & \text{call } j\text{,"} \\ w_4\big(\mathbb{1}_{\{\tau(s_{\kappa+1}) - q_j \geq \triangle\}}\big) & \text{A low priority call } j \text{ arrives and the event } e(s_{\kappa+1}) \\ & \text{is of the form "ambulance } i \text{ arrives at the scene of} \\ & \text{call } j\text{,"} \\ 0 & \text{otherwise,} \end{cases}$$

where $\triangle$ denotes the given time threshold and $\mathbb{1}_{\{\tau(s_{\kappa+1}) - q_j \geq \triangle\}}$ is an indicator function, which takes the value of one if the call is not responded within the time threshold. The cost structure can capture the relative importance of call priorities by tuning w_3 and w_4. Note that one might use different time thresholds for different priorities.

Optimality equation. Let $J_\pi(s)$ denote the expected total discounted cost when $s_0 = s$ under policy $\pi \in \mathcal{P}$, where $\mathcal{P}$ denotes the set of all stationary non-anticipative policies. That is,

$$J_\pi(s) = \mathbb{E}\left\{ \sum_{\kappa=1}^{\infty} \gamma^{\tau(s_\kappa)} h(s_\kappa, \pi(s_\kappa), s_{\kappa+1}) \middle| s_0 = s \right\}, s \in S, \pi \in \mathcal{P},$$

where $\pi(s_\kappa)$ denotes the action selected by policy π in state s_κ at time $\tau(s_\kappa)$, and $0 \leq \gamma < 1$ is a discount factor. The decision maker solves for $v(s) = \inf_{\pi \in \Pi} \left\{ J_\pi(s) \right\}$, where $\Pi \subseteq \mathcal{P}$ denotes the set of admissible policies under consideration and $v(s)$ satisfies the Bellman optimality equation

$$v(s) = \min_{a \in A(s)} \left\{ \mathbb{E}_a \left(h(s, a, s') + \gamma^{(\tau(s') - \tau(s))} v(s') \middle| s \right) \right\}, \qquad \forall s \in S, \tag{1.1}$$

where the expectation is taken with respect to action a and $s' = F(s, a, \omega(s, a))$ (Puterman 2005). Moreover, a stationary optimal policy exists, which is myopic relative to the optimal value function.

1.4 Approximate Solutions

Solving formulation (1.1) to optimality is impractical due to the curse of dimensionality. The state space of the system, S, is unbounded and the traditional methods do not apply. Section 1.4.1 adapts approximate policy iteration to produce high-quality solutions, which provide an upper bound on the optimal value function and Section 1.4.3 computes lower bounds on the long-run fraction of late calls and response time under any relocation strategy in order to assess the quality of the solutions.

1.4.1 Upper Bound

The standard policy iteration algorithm starts with an arbitrary policy π^0. At iteration n, it evaluates π^n by calculating $v^n(s)$ for all $s \in S$ by solving $v^n(s) = L_{\pi^n} v^n(s)$, where $L_{\pi^n} v^n(s) = \mathbb{E}\left\{ h(s, a, s') + \gamma^{\tau(s') - \tau(s)} v(s') \right\}$. Next, it improves the policy by choosing a myopic policy relative to v^n, i.e., $\pi^{n+1}(s) \in \arg\min_{d \in D^{MD}} \left\{ \mathbb{E}\big(h(s, a, s') + \gamma^{\tau(s') - \tau(s)} v^n(s') \big) \right\}$, where $d \in D^{MD}$ denotes a decision rule in the set of stationary Markovian deterministic policies (D^{MD}). This iterative procedure is continued until $\pi^{n+1} = \pi^n$ (Puterman 2005). Because the state space is unbounded, the policy evaluation and improvement steps are intractable in this problem. To overcome this issue, the value function is approximated by an affine combination of basis functions, i.e., $v(s) \approx \hat{v}(s) = \alpha_0 + \sum_{k=1}^{K} \alpha_k \phi_k(s)$, where each $\phi_k(s)$ is a basis function and α_k is its associated weight in the

approximation. The quality of the approximation depends on the choice of basis functions, which should be able to characterize the optimal value function (Powell 2007). Section 1.4.2 discusses our choice of basis functions in detail. Therefore, by replacing the value function with its approximation, the policy improvement step at iteration n of the approximate policy iteration involves solving

$$\pi^n(s) \in \underset{a \in A(s)}{\arg\min} \left\{ \mathbb{E}_a \left(h(s,a,s') + \gamma^{(\tau(s')-\tau(s))} \hat{v}(s') \middle| s \right) \right\}, \quad \forall s \in S, \tag{1.2}$$

where $\mathbb{E}_a(\cdot)$ denotes the expectation with respect to action a. We use Monte Carlo simulation to approximate $\mathbb{E}_a(\cdot)$ via a sample average. Starting from state s and taking action a, we simulate the system for one step, and use $\hat{v}(s)$ as the cost-to-go estimate. Because the simulation is only evaluated until the next event, enumerating all trajectories is manageable. The next event could be a call arrival or a busy ambulance completing one stage of its transition, which is either reaching a call scene, finishing service (at scene or hospital), or arriving at a location after finishing service.

Solving formulation (1.2) involves enumerating all actions for a given state. In our setting, this is manageable because if the event is "call j arrives," the decision maker has to determine which ambulance should be immediately dispatched to the call (if any) and which ambulance should be reallocated to the location of the ambulance just dispatched (if any). Let $|\mathcal{M}(s)|$ denote the number of available ambulances. The size of the action set will be $1 + |\mathcal{M}(s)|\big(|\mathcal{M}(s)| - 1\big)$. If the event is of the type "ambulance becomes available after serving a call" and no calls are in the queue, then the decision maker determines which location the ambulance should be redeployed to. This is equal to the number of locations, denoted by $|\mathcal{B}|$, which in our case study is 40. If there are calls in the queue and the event is of type "ambulance becomes available after serving a call," or "ambulance has just arrived at its location," then the decision maker determines which ambulance to dispatch to the call based on a first-come first-served rule, and if the decision is not to dispatch, which location the ambulance should be redeployed to, which is at most $|\mathcal{M}(s)| + |\mathcal{B}|$. Once the expectation is estimated for all actions, the decision that yields the smallest value is chosen by the policy. Formulation (1.2) provides $\hat{v}$-improving decision rules for a fixed state s. However, solving it for each state is not possible because the state space is unbounded. Therefore, in order to evaluate a policy, we use formulation (1.2) upon visiting a state on the fly in the Monte Carlo simulation. That is, it is solved only for states observed in simulation. In the settings of interest, our computational experiments demonstrate that solving formulation (1.2) for a state is instantaneous.

Next, we develop an algorithmic approach to estimate $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_k)$ and consequently derive high-quality solutions. Consider an appropriately large finite horizon, initialize $\alpha = \alpha^0$ and evaluate the policy associated with it for states in $\hat{S}$, where $\hat{S}$ is a subset of S. We construct $\hat{S}$ by sampling states that are more likely to be visited by the optimal policy and update $\hat{v}^n(s)$ at iteration n of the algorithm (de Farias & Van Roy 2004). For the policy evaluation step, we propose the following procedure. Start from an initial state s, use Monte Carlo to simulate the system, and upon observing a state, find actions by solving formulation (1.2), and calculate the total discounted cost for that realization of the system. Let $C^r(s)$ denote the total discounted cost of the realization of the system, starting from state s in replication r, i.e., the simulated value function for state s in replication r. Let R_s be the total number of replications of the Monte Carlo simulation for state s. To estimate α, solve the following optimization problem

$$\min_{\alpha} \sum_{s \in \hat{S}} \sum_{r=1}^{R_s} \left(C^r(s) - \alpha_0 - \sum_{k=1}^{K} \alpha_k \phi(s) \right)^2, \tag{1.3}$$

which minimizes the squared error between the approximate value function and the simulated value function. Note that formulation (1.3) is a regression model and computational experiments in our case study show that solving it is instantaneous. This procedure continues until convergence in some norm is achieved. Algorithm 1 formalizes this approach.

1.4.2 Basis Functions

This section describes the basis functions $\{\phi_k(\cdot) : k = 1, \ldots, K\}$ used for the value function approximation. Jagtenberg *et al.* (2015) noted that basis functions in the ambulance relocation literature may not produce high-quality solutions in general, but our computational results show

Algorithm 1 Approximate policy iteration

Set $n = 0$, $\epsilon > 0$ and $\alpha = \alpha^0$.
while $|\hat{v}^n(s) - \hat{v}^{n-1}(s)| > \epsilon$ or $n \neq 0$ **do**
  Policy improvement: Find a myopic policy induced by $\hat{v}^n(s)$ by solving formulation (1.2).
  Policy evaluation: Use Monte Carlo to simulate the system; find actions for each state visited by the simulation via solving formulation (1.2); calculate $C^r(s)$, the total discounted cost for each initial state s and replication r.
  Projection: Use $C^r(s)$ from the Monte Carlo simulation and solve formulation (1.3) to estimate α^{n+1} for the next iteration.
  Set $n \leftarrow n + 1$.

that our ADP approach based on the following basis functions produces near-optimal solutions.

Response time. This novel basis function estimates the expected response time of a call when the state of the system is s. To that end, let $r_l(s)$ denote the expected response time of a call in region l in state s, and set $\phi_1(s) = \sum_{l \in \mathcal{L}} \lambda_l^\tau r_l(s)$. Response time is comprised of travel time of an ambulance to reach a call plus potential waiting time of a call in queue for ambulance assignment. In order to estimate the expected waiting time of a call in a region, we develop an $M/G/c$ queueing system for each region, and estimate the expected waiting time in the queue. Let $N_l(s)$ denote the number of available ambulances (the number of servers in the $M/G/c$ queuing system) that cover location l when the state of the system is s. We consider a region covered by ambulance i, if the time that it takes for an available ambulance to reach to the center of the region, $\bar{l}$, is less than a threshold $\triangle$. Therefore, $N_l(s) = \sum_{i \in \mathcal{M}(s)} \mathbb{1}_{\{t(o_i,\bar{l}) \leq \triangle\}}$, where $t(x, y)$ denotes the travel time between location x and y. We estimate the service rate of an ambulance in region l, $\mu_l(s)$, by considering the average travel time in region l, plus the average time that an ambulance spends on scene, plus the average time of handing over a patient to hospital personnel. Because ambulances in a region may also serve other regions, we adjust the arrival rate of calls in region l by summing over call arrival rates of regions covered by an available ambulance in region l, using $\lambda'_{l,\tau}(s) = \sum_{u \in \mathcal{L}} \sum_{i \in \mathcal{M}_l(s)} \lambda_u^\tau \mathbb{1}_{\{t(o_i,\bar{u}) \leq \triangle\}}$, where $\mathcal{M}_l(s)$ denotes the set of available ambulances that cover region l, i.e., $\mathcal{M}_l(s) := \{i \in \mathcal{M}(s) : t(o_i, \bar{l}) \leq \triangle\}$ and λ_l^τ is the call arrival rate in region l at time τ. We use $\mu_l(s)$ and $\lambda'_{l,\tau}(s)$ to compute the expected waiting time in queue in an $M/M/c$ queuing system in region l, i.e., $W^{q,l}_{(M/M/c)}(s)$. The expected waiting time in the queue in an $M/G/c$ queuing system is then approximated by

$$W^{q,l}_{(M/G/c)} \approx \frac{1 + cv_l^2}{2} W^{q,l}_{(M/M/c)},$$

where cv_l denotes the coefficient of variation of the service time in region l (Allen 1980). Let $\bar{t}^l$ denote the average travel time within region l, then

$$r_l(s) = \mathbb{1}_{\{\mathcal{M}_l(s)=\emptyset\}} \left[W^{q,l}_{M/G/c}(s) + \bar{t}^l \right] + \mathbb{1}_{\{\mathcal{M}_l(s)\neq\emptyset\}} \left[\min_{i \in \mathcal{M}_l(s)} t(o_i, \bar{l}) \right].$$

Note that the queuing theory approach may not accurately estimate arrival rates, service rates, and

the number of the servers for a region. To overcome this issue, we calibrate the model by scaling the arrival rates and find the scaling factor through experimentation.

Future response time. Ambulances in transit can serve a (currently in queue or future) call after their service is finished. Therefore, the destination of the busy ambulances is as important as their current location. This is the underlying motivation for the second and the fourth basis functions. Let $\vec{s}$ denote the state that corresponds to the earliest time when one of the following events occur: "an ambulance finishes serving a call (at scene or hospital)," or "an ambulance arrives at a base." The future response time in state $\vec{s}$ is important because it evaluates the trade-off between immediate and future cost. Given that the current state of the system is $s = (\tau, e, m, c)$, we construct a new state $\vec{s}(s) = \left(\vec{\tau}(s), \vec{e}(s), \vec{m}(s), \vec{c}(s)\right)$, where $\vec{\tau}(s)$ denotes the time that the future state $\vec{s}$ will be visited and $\left(\vec{e}(s), \vec{m}(s), \vec{c}(s)\right)$ denotes predicted future event, ambulance status, and call status at time $\vec{\tau}(s)$. Also, $\vec{s}(\cdot)$ is determined by searching the earliest time that a busy ambulance becomes available. Predicting future events, ambulance statuses, call statuses and the earliest time that a busy ambulance becomes available is possible by searching the future event list in the simulation. We then set $\phi_2(s) = \phi_1(\vec{s})$. This basis function is novel in that $\vec{s}$ computes the state that corresponds to the earliest time that an ambulance becomes available compared to Maxwell *et al.* (2010) where the future state is computed by replacing the locations of all busy ambulances with their destinations.

Uncovered call rate. The third basis function computes the rate of uncovered calls. Recall that $N_l(s)$ is the number of available ambulances in region l in state s, and calls arrive with rate λ_l^τ from location l at time τ. If no ambulance covers region l, then the call may be late (Restrepo *et al.* 2009). We define the uncovered call rate by $\phi_3(s) = \sum_{l \in \mathcal{L}} \lambda_l^\tau \mathbb{1}_{\{N_l(s)=0\}}$.

Future uncovered call rate. The fourth basis function calculates the uncovered call rate for a future state $\vec{s}$, which is constructed in the same way discussed in the second basis function, i.e., $\phi_4(s) = \phi_3(\vec{s})$.

Unreachable calls. The fifth basis function computes the number of calls for which an ambulance is assigned but it cannot reach the scene within the time threshold $\triangle$, i.e.,

$$\phi_5(s) = \sum_{j=1}^{J} \mathbb{1}_{\{g_j=1\}} \sum_{i=1}^{N} \mathbb{1}_{\{f_i=\text{“ambulance } i \text{ is going to call scene } j\text{”}\}} \mathbb{1}_{\{t_i+t(o_i,\bar{l})-q_j\geq\triangle\}},$$

where t_i is the time that ambulance i started to move to the scene of call j. The above expression first checks whether the call is assigned to an ambulance and then checks whether the ambulance will fail to reach the call location within the time threshold (Maxwell *et al.* 2010).

Aggregated delay time. This novel basis function computes the aggregated delay time for calls in the queue for which an ambulance is assigned, but is not going to reach to the call scene within the time threshold $\triangle$, i.e.,

$$\phi_6(s) = \sum_{j=1}^{J} \mathbb{1}_{\{g_j=1\}} \sum_{i=1}^{N} \mathbb{1}_{\{f_i=\text{“ambulance } i \text{ is going to call scene } j\text{”}\}} \mathbb{1}_{\{t_i+t(o_i,\bar{l})-q_j\geq\triangle\}} (t_i + t(o_i, \bar{l}) - q_j \geq \triangle).$$

The indicator function $\mathbb{1}_{\{f_i=\text{“ambulance } i \text{ is going to call scene } j\text{”}\}} \mathbb{1}_{\{t_i+t(o_i,\bar{l})-q_j\geq\triangle\}}$ ensures that only late calls are counted.

1.4.3 Lower Bound

This section provides a lower bound on the expected total response time for a broad class of relocation policies over a finite time horizon. The bound is based on a lower bounding system in which the call arrival process is exactly the same as the original system, and the number of available ambulances just before the arrival of a call is greater than or equal to that in the original system under policy π. This is achieved by computing a stochastic lower bound on the service time distribution of ambulances in the original system, which is independent of the ambulance configuration in an EMS system, thus the relocation policies (Maxwell *et al.* 2014). Therefore, we simulate a multi-server queuing system (ambulances resemble servers and calls resemble customers) with the bounding service time distribution, where calls arrive according to the same process as the original system. However, just before the arrival of a call, the available ambulances are repositioned to minimize the expected response time by solving an integer program. Because we use the same bounding system as Maxwell *et al.* (2014), the same set of assumptions hold true.

Let D be the (random) number of calls over a horizon and T denote the (random) total

response time over the same horizon. The goal is to compute a lower bound on $\mathbb{E}(T)$ independent of relocation policy π, which is given by

$$\mathbb{E}(T) = \mathbb{E}\big(\sum_{j=1}^{\infty} T_j \mathbb{1}_{\{j\leq D\}}\big) = \sum_{j=1}^{\infty} \mathbb{E}\big(\mathbb{1}_{\{j\leq D\}}\mathbb{E}[T_j|\mathcal{A}_j, \tau_j, \mathcal{C}_j]\big) \geq \sum_{j=1}^{\infty} \mathbb{E}\big(\mathbb{1}_{\{j\leq D\}}\nu(\mathcal{A}_j)\big) = \mathbb{E}\big(\sum_{j=1}^{D} \nu(\mathcal{A}_j)\big),$$

where T_j is the response time of the jth call, τ_j is the arrival time of the jth call, $\mathcal{C}_j$ is the configuration of ambulances at time τ_j, $\mathcal{A}_j$ is the number of available ambulance at time τ_j, and $\nu : \{0, 1, \ldots, \mathcal{A}\} \to [0, \infty]$ is a decreasing function such that $\mathbb{E}(T_j|\mathcal{A}_j, \tau_j, \mathcal{C}_j) \geq \nu(\mathcal{A}_j)$. Maxwell *et al.* (2014) constructed a bounding system by a coupling of the ambulance dynamics such that the number of available ambulances in the bounding system, $\tilde{\mathcal{A}}_j$, at the arrival time of the jth call satisfies $\tilde{\mathcal{A}}_j \geq \mathcal{A}_j$ for all j almost surely. Therefore, $\mathbb{E}(T) \geq \mathbb{E}\big(\sum_{j=1}^{D} \nu(\mathcal{A}_j)\big) \geq \mathbb{E}\big(\sum_{j=1}^{D} \nu(\tilde{\mathcal{A}}_j)\big)$.

Having $\nu(\cdot)$ allows us to approximate the above expectation by simulating the bounding system. Let $\nu(\mathcal{A}_j)$ denote the response time when $\mathcal{A}_j$ ambulances are available at the arrival time of the jth call. For $1 \leq \mathcal{A}_j \leq \mathcal{A}$, $\nu(\mathcal{A}_j)$ is is the optimal objective function of the following integer program

$$\begin{aligned}
\nu(\mathcal{A}_j) = \min & \sum_{l=1}^{\mathcal{L}} d_l \sum_{k=1}^{|\mathcal{A}_j|} \sum_{b=1}^{\mathcal{L}} y_{kbl} t(b,l) && && (1.4)\\
\text{s.t.} & \sum_{k=1}^{|\mathcal{A}_j|} \sum_{b=1}^{\mathcal{L}} y_{kbl} = 1, && \forall l,\\
& y_{kbl} \leq x_{kb}, && \forall b, l, k,\\
& \sum_{b=1}^{\mathcal{L}} x_{kb} = 1, && \forall k,\\
& x_{kb} \in \{0,1\}, y_{kbl} \in \{0,1\}, && \forall b, l = 1, 2, \ldots, \mathcal{L} \text{ and } \forall k = 1, 2, \ldots, |\mathcal{A}_j|.
\end{aligned}$$

where y_{kbl} is an indicator taking a value of 1 if ambulance k is stationed at base b and is assigned to serve location l, x_{kb} taking a value of 1 if ambulance k is stationed at base b, $t(b, l)$ denotes the travel time between base b and location l, and d_l denotes the proportional call arrival rate in location l. The first constraint ensures that each location is served by exactly one ambulance, and the third constraint prevents an ambulance to be located at different bases at the same time. Thus, formulation (1.4) seeks to minimize the expected response time to the demand. We set $\nu(0) = \nu(1)$. We also use the "cover bound" developed in Maxwell *et al.* (2014) to assess the quality of our

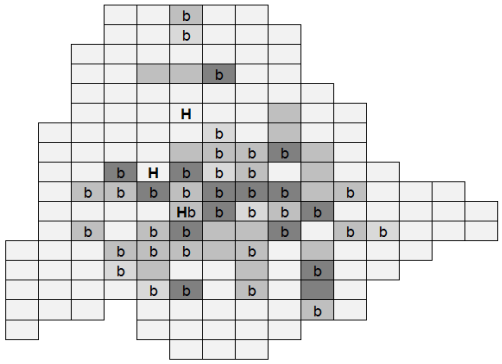


Figure 1.1: Mecklenburg County, NC.

solutions when the ADP objective function is minimizing the expected fraction of late calls.

1.5 Case Study: Mecklenburg County, NC

This section presents the result of implementing our ADP framework using data from the EMS provider in Mecklenburg County, which contains the city of Charlotte, and is the most populated and densely populated county in the state of North Carolina, with a population of over a million as of 2014 estimates (United States Census Bureau 2014). The EMS system in the county has, on average, 17 ambulances, three hospitals, and we consider 40 potential ambulance locations. We divide the county into 168 regions, where each region is a 2×2 mile square rectangle. As many EMS providers distribute ambulances along the road to meet the performance targets, we could consider all regions as a potential location for ambulances. However, in order to keep computations tractable, we limit the number of possible ambulance locations to 40 bases. Section 1.5.1 provides further details on the choice of base locations, which serve as the main contributing factor in designing static benchmarks. We assume that all ambulances are the same and turn-out time (the activation delay needed for the ambulance crew to get ready and depart the base) is 45 seconds. Travel times are calculated in rectilinear measurement (Manhattan distance) based on historical data from more than 40,000 incidents and travel speed is assumed to be constant. Figure 1.1 shows 168 zones in Mecklenberg County where darker regions correspond to higher call arrival rates, "b" denotes bases, and "H" points to hospitals in the county. We divide a day into four time intervals (12:00 am-06:00 am, 06:00 am-12:00 pm, 12:00 pm-06:00 pm, and 06:00 pm-12:00 am) and estimate

the rates λ_l^τ using historical data. A few regions on the borders of the county had too few data points to fit a distribution and are excluded from the study. The amount of time that an ambulance serves a call at scene has a normal distribution with mean and standard deviation of 54.18 and 15.18 minutes, respectively. Historical data show that 77% of calls are transferred to a hospital. The service time of calls that are transferred to a hospital has a normal distribution with a mean and standard deviation of 56.7 and 13.6 minutes. Note that this time includes the amount of time that an ambulance spends on the scene, travel time to hospital, and the time that it takes to handover the patient to the hospital. A call not reached within eight minutes is considered to be late. The simulation horizon is two weeks and we set $\gamma = 0.99$ per day. A sample of the 100 most visited states is used in formulation (1.3). Increasing the sample size to 200 and 500 states had minor effects on the results. We initialize the approximate policy iteration algorithm by setting $\alpha = (1, 1, 1, 1, 1, 1, 1)$ and $R_s = 5$ in each iteration of ADP (recall that R_s is the total number of replications of the Monte Carlo simulation for state s). After a warm-up period under the best static benchmark, we begin collecting the statistics at the extant state when the warm-up period ends. Priority adjustment weights (w_1, w_2) and (w_3, w_4) are such that high priority calls are 10 times more important than low priority calls. Each iteration of ADP takes about two days of CPU time on an Intel Core i7 3.4 GHz processor with 16 GB of RAM. However, this procedure is carried out off line and after estimating an appropriate α, solving formulation (1.2) is instantaneous, which is what an EMS system needs for real-time ambulance management.

1.5.1 Choice of Static Benchmark

This section investigates the performance of several static benchmarks and considers the best in terms of response time and the best in terms of fraction of late calls, as benchmarks to dynamic policies. A static policy sends the closest available ambulance to a call and returns an ambulance after finishing service to its predetermined base if no call is in the queue. If no ambulances are available, the call will join a queue and will be served according to a first-come first-served rule in a decreasing order of priority. In the absence of repositioning policies, identifying the base for each ambulance is the key to design efficient static benchmarks to ensure that a certain fraction of demand is reached within a specified response time target. Some models seek to maximize the fraction of demand covered by available ambulances, and others focus on minimizing the response time. We consider six frequently used models and refer the reader to van den Berg *et al.* (2016) for a

Table 1.1: Performance of static benchmarks

	Ambulance location models					
	MCLP	DSM	MEXCLP	MALP	ARTM	ERTM
Fraction of late calls (%)	25.3	25.6	24.4	27.1	48.3	24.7
Average response time (min.)	7.3	7.9	7.5	7.4	9.7	7.8

complete description and formulation of each model. Maximal Covering Location Problem (MCLP) maximizes the weighted number of demand locations covered by at least one ambulance. Double Standard Model (DSM) focuses on covering a demand location with two ambulances to prevent a coverage drop if an ambulance becomes busy. The DSM guarantees a certain level of coverage within the target response time for at least a fraction of demand and defines a second type of coverage with higher target response time that must be maintained for all demand locations. Maximum Expected Covering Location Problem (MEXCLP) maximizes the expected coverage of all demand locations by calculating the marginal contribution of each ambulance to coverage while considering that the ambulance might not be available with a certain probability called the "busy fraction," which is calculated by dividing the priority-adjusted total workload of the system in minutes by total ambulance capacity in minutes. Maximum Availability Location Problem (MALP) calculates the minimum number of available ambulances to guarantee a specific coverage level prior to formulating an instance of the model and uses it to maximize the covered demand. Average Response Time Model (ARTM) is equivalent to the p-median model applied to ambulance location problem and minimizes the average response time from the closest base. Expected Response Time Model (ERTM) is similar to MEXCLP in that it minimizes the expected response time by incorporating the probability of a demand location being served by the pth nearest ambulance.

The initial base of ambulances in the static benchmarks is determined by solving each model to optimality. The static benchmarks are simulated for two weeks and their performance is measured with respect to fraction of late calls and average response time. Table 1.1 shows that MEXCLP and MCLP outperform other models in the fraction of late calls and expected response time, respectively. Therefore, the performances of static benchmarks based on MEXCLP and MCLP are used to compare with that of dynamic benchmarks and ADP policy in terms of fraction of late calls and average response time, respectively.

1.5.2 Dynamic Benchmark Policies

We design three dynamic benchmarks to assess the contribution of each strategy, dispatching, redeployment, and reallocation. The fourth dynamic benchmark is used to test the quality of our solutions, and the fifth is a relocation heuristic designed by Jagtenberg *et al.* (2015). Benchmark 1 (dispatching only) allows the dispatcher to assign any available ambulance when a call is received. However, redeployment and reallocation decisions are not considered. That is, every ambulance in the EMS system is preassigned to a base and returns to that base after serving a call, and the repositioning of idle ambulances to the base of an ambulance, that was just dispatched to a call, is not considered. Note that if the dispatcher does not immediately send an available ambulance to a received call in this benchmark, the call will join a queue. After an ambulance finishes serving a call, the dispatcher decides whether the ambulance serves a call in the queue or returns to its preassigned base. Calls in the queue are served based on a first-come first-served rule in a decreasing order of priority. Benchmark 2 (redeployment only) sends immediately the closest available ambulance to a call and does not consider the possibility of ambulance reallocation after dispatching an ambulance. However, Benchmark 2 determines the redeployment policy, i.e., after an ambulance has finished serving a call, the dispatcher decides whether the ambulance serves a call in the queue or is redeployed to a base. Benchmark 2 is similar to the settings studied by Maxwell *et al.* (2010). Benchmark 3 (reallocation only) sends the closest available ambulance to serve a call and after an ambulance has finished serving a call, decides whether the ambulance serves a call in the queue or returns to its preassigned base. However, upon dispatching an ambulance to a call, Benchmark 3 considers the possibility of reallocating an available ambulance to the base of the ambulance that is just dispatched. The performance of Benchmark 4, which considers both redeployment and reallocation, is used to test the quality of the solutions by calculating the optimality gap with respect to the lower bounding system. The ADP approach, presented in Section 3.3, considers all of the dispatching, redeployment, and reallocation strategies simultaneously.

Relocation heuristic (Benchmark 5). Jagtenberg *et al.* (2015) developed a simple relocation heuristic, which is easy to implement and showed strong performance in some data sets. We use this heuristic as another benchmark. The dispatching policy in this benchmark is to send the closest ambulance to a received call; however, the relocation policy can reallocate an available ambulance

to a base, or redeploy an ambulance that just finished its service to a base that results in the largest marginal contribution to coverage according to the MEXCLP model. The marginal contribution of adding a kth ambulance to cover demand in region l is given by $E_k - E_{k-1} = \lambda_l^\tau(1-\alpha)\alpha^k$, where α denotes a "busy fraction" similar to MEXCLP and λ_l^τ is the demand (call arrival) rate in region l at time τ. The base that gives the largest marginal contribution over all demand is chosen as the destination for relocation.

1.5.3 Results and Managerial Insights

We compare the performance of ADP and benchmark policies with the static benchmarks with respect to four major measures, (i) average response time, (ii) fraction of late calls, (iii) average response time of late calls, and (iv) fraction of late high priority calls. Table 1.2 reports the performance of each policy when the ADP objective is to minimize the expected discounted priority-adjusted total response time. The average response time for the ADP policy is 6.5 ± 0.2 minutes (95% confidence interval) while the MCLP static benchmark is estimated to have an average response time of 7.3 ± 0.2 in 30 replications. Table 1.2 shows that the fraction of late calls for the ADP policy is significantly less than that of the MCLP static and other benchmarks. In particular, the fraction of late calls is $18.3 \pm 0.1\%$ for the ADP policy and $25.3 \pm 0.2\%$ for the MCLP static benchmark. The performance of the ADP policies for average response time of late calls and the fraction of high priority late calls is also significantly better than that of benchmarks. Similarly, Table 1.3 reports the performance of the ADP policy and benchmarks when the ADP objective is to minimize the expected discounted priority-adjusted fraction of late calls. Tables 1.2 and 1.3 show that the ADP policy improves various performance measures compared to other benchmarks.

Table 1.2: Performance of ADP policy and benchmarks for response time minimization

	Avg. response time (min.)	Fraction of late calls (%)	Avg. response time of late calls (min.)	Fraction of late high priority calls (%)
MCLP	7.3±0.2	25.3±0.2	11.7±0.1	4.9±0.2
Benchmark 1	7.1±0.1	22.5±0.1	10.5±0.1	4.2±0.1
Benchmark 2	6.8±0.1	20.2±0.1	9.1 ±0.1	3.8±0.1
Benchmark 3	7.2±0.1	22.9±0.1	10.4±0.1	4.5±0.1
Benchmark 4	6.7±0.1	19.2±0.2	9.0 ±0.2	3.5±0.2
Benchmark 5	6.8±0.1	19.7±0.1	10.1±0.1	3.7±0.2
ADP	6.5±0.2	18.3±0.1	8.9 ±0.2	3.4±0.2

Note: The ADP objective function in this table is to minimize the expected total discounted priority-adjusted response time, and the results are reported in 95% confidence intervals.

Table 1.3: Performance of ADP policy and benchmarks for late calls minimization

	Avg. response time (min.)	Fraction of late calls (%)	Avg. response time of late calls (min.)	Fraction of late high priority calls (%)
MEXCLP	7.5±0.2	24.4±0.2	12.7±0.2	4.6±0.2
Benchmark 1	7.1±0.1	21.1±0.1	11.2±0.1	3.9±0.1
Benchmark 2	6.9±0.1	19.2±0.1	10.1±0.1	3.5±0.1
Benchmark 3	7.3±0.1	22.0±0.1	10.7±0.1	4.2±0.1
Benchmark 4	6.8±0.1	18.3±0.2	9.9 ±0.2	3.4±0.2
Benchmark 5	6.8±0.1	19.7±0.1	10.1±0.1	3.7±0.2
ADP	6.6±0.2	16.4±0.1	9.2 ±0.1	3.1±0.1

Note: The ADP objective function in this table is to minimize the expected total discounted priority-adjusted fraction of late calls, and the results are reported in 85% confidence interval.

Our results indicate that the contribution of redeployment-only strategy is significantly greater than that of dispatching-only and reallocation-only strategies in improving the performance over static benchmarks in all measures. Our analysis shows that one reason for this observation may be that the expected proportion of time that the dispatching-only ADP strategy deviates from the best static benchmark is much less than the proportion of time that the redeployment-only ADP strategy deviates from it. Particularly, the redeployment-only ADP strategy sends the ambulance to its previous base after finishing service only 19% of times, while the dispatching-only ADP strategy immediately sends the closest ambulance to a received call nearly 70% of times. Further analysis of the dispatching-only ADP strategy shows that, conditioned on not immediately sending the closest ambulance, a non-closest ambulance is dispatched in nearly 87% of times, while calls are delayed in 13% of times. Moreover, the performance of the dispatching-only ADP strategy does not significantly change if the dispatcher is not allowed to queue a call when an ambulance is available. Although both high and low priority calls can be queued in our framework, our numerical analysis shows that only 1% of high priority calls are delayed. Our results also show that the reallocation-only ADP strategy relocates an idle ambulance to the base that just emptied in nearly 10% of times, and the reallocation flows are toward empty bases in high demand zones. Comparing the results for the relocation heuristic and Benchmark 4 shows that the relocation heuristic is an efficient policy when only relocation strategies are considered.

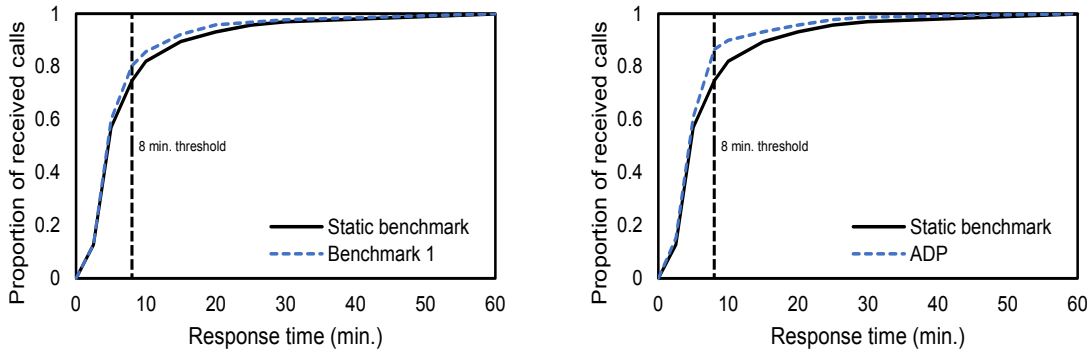
Figure 1.2 shows the empirical cumulative distribution function of the response times for the MCLP static benchmark to the dispatching-only strategy (a), and to the ADP policy (b). One could think that minimizing the expected discounted priority-adjusted total response time might involve the risk of losing some of the closer calls by trying to concentrate the optimization on calls with larger response times. Figure 1.2 suggests that the ADP policies do not abandon a few calls to

wait for a long time, instead, it shifts the entire distribution of response times to the left.

To illustrate the quality of solutions produced by the ADP framework, we compute a lower bound on the expected response time and fraction of late calls. Recall that we assumed a non-homogeneous Poisson process for call arrivals in Section 3.3. However, in reporting the results for comparing our lower bounds with Benchmark 4, an assumption of a constant call arrival rate for each location is forced in both the lower bounding system and Benchmark 4. Our results show that in the lower bounding system the expected response time and fraction of late calls are 5.1 minutes and 11.7%, respectively. We use Benchmark 4 to assess the quality of solutions with respect to the lower bounding system, because both Benchmark 4 and the lower bounding system use a myopic dispatching rule, i.e., immediately sending the closest ambulance and only rely on relocating available ambulances to improve performance, which in case of Benchmark 4 consists of redeployment and reallocation strategies. Our results show that the absolute difference between Benchmark 4 and lower bound on average response time (fraction of late calls) is 1.6 minutes (6.6%) when the objective function is to minimize the response time (fraction of late calls).

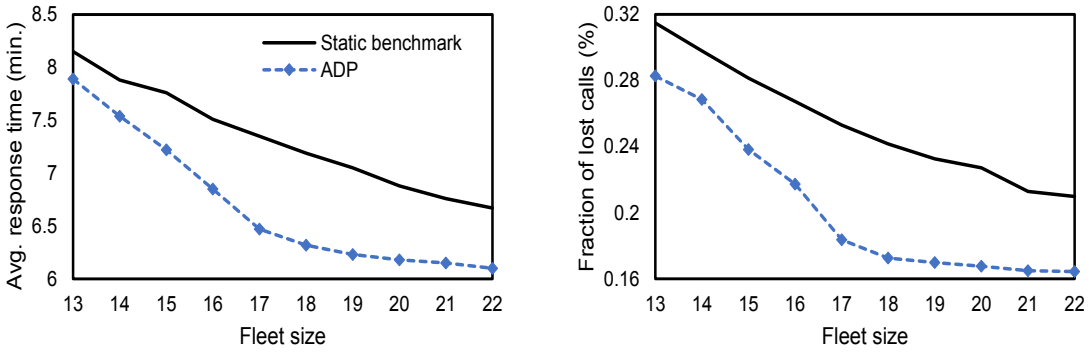
1.5.4 Sensitivity Analysis

Varying fleet size. We explore the effect of fleet size of the EMS provider on the performance of our ADP and the MCLP static benchmark. Figures 1.3(a) and 1.3(b) show the performance of ADP and MCLP static benchmark on response time and fraction of late calls for a variety of fleet sizes, respectively, and confirm that ADP policies consistently outperform the MCLP static benchmark in



(a) (b)

Figure 1.2: Empirical cumulative distributions of the response time



(a) (b)

Figure 1.3: Performance of ADP and MCLP static benchmark with different fleet sizes

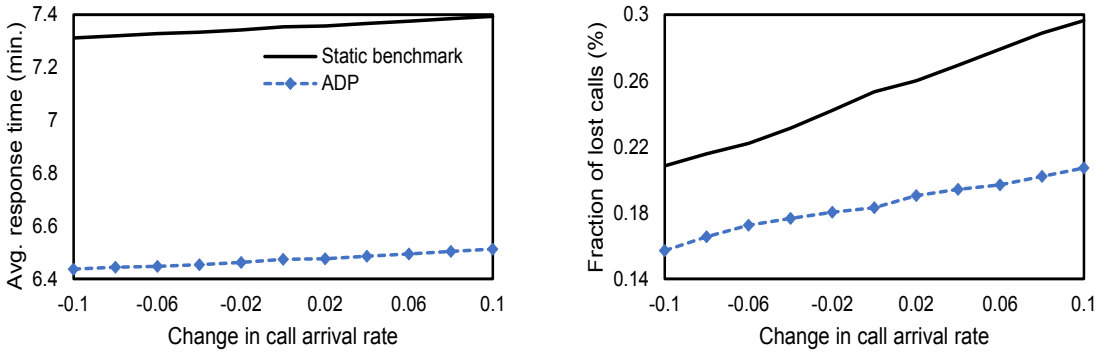
both measures. The difference in the gap between the performance of the ADP and the best static benchmark narrows when the fleet size becomes very large or small. If the fleet size is very small, the ambulances remain busy most of the time and less opportunity will be available for optimal repositioning of ambulances. On the other hand, if the fleet size is very large, there are always idle ambulances to respond to a call and intelligent repositioning of ambulances cannot significantly improve the performance. Table 1.4 shows the average utilization of ambulances for different fleet sizes, as given in Figure 1.3.

Time-dependent fleet size. We assume that the fleet size is constant in our simulation horizon whereas in practical situations, EMS managers may increase or decrease the number of emergency vehicles with respect to varying demand in different times of a day. In order to capture the effect of varying fleet sizes during different shifts, we compare the performance of time-dependent fleet size ADP with the best time-dependent fleet size static benchmark. We use the predicted fleet sizes provided in Rajagopalan *et al.* (2008) for each 6-hour time interval which guarantees at least 95% coverage while minimizing the cost of ambulance scheduling. Note that the fleet sizes are calculated based on a similar data set of Mecklenburg County, NC. Our results show that increasing the fleet

Table 1.4: Average utilization of ambulances for varying fleet sizes

Fleet size	13	14	15	16	17	18	19	20	21	22
Average utilization	46.9	45.1	43.3	40.8	38.6	36.2	34.7	32.7	30.9	28.1

Note: The ADP objective function in this table is to minimize the expected discounted priority-adjusted response time.



(a) (b)

Figure 1.4: Performance of ADP and MCLP static benchmark for different call arrival rates

size during rush hours may improve the performance in terms of fraction of late calls and expected response time. In particular, when the ADP objective is to minimize the expected response time, the fraction of late calls for a time-dependent fleet size ADP is 16.4% while the fraction of late calls for time-dependent fleet size MCLP static benchmark is 21.7%.

Varying travel times. Travel times are deterministic in our model and are computed from the actual travel times between the regions reported in our historical data set. Since travel times may increase especially during rush hours, we increase the ambulance travel times by 10% to test the performance of ADP policies in such settings. To that end, we use the ADP policies produced in the settings with no travel time increase. Results indicate that our ADP approach continues to consistently outperform the MCLP static benchmark and improvement percentage for response time and fraction of late calls, compared to the best static benchmark is 6.5% and 28.9%, respectively. It is worth mentioning that the improvement for both measurements is larger with longer travel times compared to the results in Section 5.3.

Varying call arrival rates. In order to explore the sensitivity of the ADP policy to changes in call arrival rates, we vary the call arrival rates in each region over the interval $[\lambda_l^\tau \pm 10\%\lambda_l^\tau]$. In order to test the performance of our ADP approach, decisions are continued to be made under the optimal policy with original call arrival rates. Figure 1.4 shows that the ADP policy continues to outperform the MCLP static benchmark. Ambulance bases for the static benchmark used in Figure

1.4 is also found by solving MCLP under the original call arrival rates.

Priority weights. Most of ambulance relocation studies only focus on optimizing the policy in response to high priority calls. This focus is usually justified either by considering all the calls to have high priorities or assuming that EMS providers are judged by their performance regarding to the highest priority calls (e.g., van Barneveld *et al.* 2016). In our numerical results, high priority calls weigh 10 times more than low priority calls. Table 1.5 reports the performance of each policy when the optimization is carried over just high priority calls, i.e., the weight of low priority calls in the ADP objective function is set to zero.

1.6 Conclusion

We formulated a real-time ambulance dispatching and relocation problem as a stochastic dynamic program and solved it via approximate dynamic programming. We extended the literature on real-time ambulance management via ADP, which considers only ambulance redeployment, in two dimensions. First, we considered a general dispatching strategy in which the decision maker can send any available ambulance to a received call in addition to having the option of not dispatching an ambulance immediately, but rather waiting for an ambulance that may become available soon. Second, we introduced an ambulance reallocation strategy in which the decision maker may send an available ambulance to the location of an ambulance just dispatched to a call. The ambulance reallocation strategy can improve performance by reducing the expected time that a region is uncovered, which is caused by dispatching the only ambulance that covers it. We tested the performance of policies generated by our ADP framework on an EMS system in Mecklenburg County, NC and

Table 1.5: Performance of ADP policy and benchmarks for high priority late calls minimization

	Avg. response time (min.)	Fraction of late calls (%)	Avg. response time of late calls (min.)	Fraction of late high priority calls (%)
MEXCLP	8.2	43.0	11.3	43.0
Benchmark 1	7.2	39.1	10.3	39.1
Benchmark 2	6.9	36.6	9.3	36.6
Benchmark 3	7.4	39.6	10.4	39.6
Benchmark 4	6.8	35.4	9.6	35.4
ADP	6.4	33.0	8.8	33.0

Note: The ADP objective function in this table is to minimize the expected discounted priority-adjusted fraction of late calls, and the results are reported in 95% confidence interval.

our results show that our policies significantly improve static benchmarks. In particular, our near-optimal policies reduce the response time and fraction of high priority late calls by 12% and 30.6% compared to the best static benchmarks.

We designed three benchmarks to analyze the contribution of each strategy, general dispatching, redeployment, and reallocation, by adding a strategy to the static policy one at a time. Our results show that each strategy significantly improves the static benchmarks and considering all three strategies simultaneously is significantly better than each strategy alone. We also showed that the redeployment strategy is the best when only one strategy could be added to the static policy. This observation, our analysis shows, is due to the fact that the expected frequency that the redeployment-only ADP policy deviates from the static benchmarks is significantly greater than the frequency that the dispatching-only ADP policy deviates from the static benchmarks. Allowing the dispatcher to queue a received call when an ambulance is available did not significantly improve the performance. Considering a general dispatching rule, redeployment and reallocation can significantly improve the performance of an EMS system.

Chapter 2

Response-Adaptive Design of Dose-Finding Clinical Trials

Summary. Identifying the right dose is one of the most important and challenging decisions that has to be made in drug development. Adaptive designs are promoted to conduct dose-finding clinical trials as they are more efficient and ethical compared to static designs. However, current techniques in response-adaptive designs of dose allocations are complex and need significant computational efforts, which is a major impediment for implementation in practice. This chapter provides a novel framework to produce high-quality dose allocation policies with significant reduction in complexity and computational effort, as well as novel properties to the learning problem. In addition, simulation results of a broad range of dose-finding studies reveal that the proposed policies perform competitively against the standard approach with significantly less computational effort. Moreover, consistency proof of the proposed policies ensure that the learning algorithm will eventually identify the correct target dose.

Manuscript: Nasrollahzadeh, A., Khademi, A., "Dynamic Programming to Response-Adaptive Dose-Finding Clinical Trials", Revision for submission to INFORMS Journal on Computing

2.1 Introduction

Motivation. Clinical trials are studies in which participants are assigned to one or more treatments to evaluate their effects on health-related outcomes. The objective is usually to determine whether new treatments are safe and effective by measuring certain responses in trial participants (National Institutes of Health 2014). The U.S. Food & Drug Administration (FDA) classifies the approval procedure of a medical product into four phases. Phase I studies a small group of volunteers with the disease/condition for several months to identify a safe dosage range and potential side effects. Phase II increases the number of participants up to several hundred, and extends the length of study up to two years. These studies are not large enough to determine if the drug will be beneficial, however, they provide safety evaluations and allow researchers to refine their methods for the next phases. In Phase III, 300-3000 volunteers are studied for a period of one to four years to confirm the drug's efficacy and to monitor its adverse reactions, particularly its long-term and rare side effects. Phase IV is carried out once the medical product has been approved by the FDA and involves several thousand volunteers and post-market safety monitoring (FDA 2017). The average cost of inventing, developing, and introducing a new drug to market has exponentially increased (see Figure 2.1) and it has surpassed $2.6 billion (Tufts 2014). The biggest drivers of this rise are expensive clinical trials (Roy 2012). Their costs depend on factors such as number of participants, locations of research facilities, complexity of the trial protocol, and the reimbursements provided to investigators. The total cost can reach $300-$600 million for large trials (Griffin *et al.* 2010). Phase II clinical trials constitute about 18% of pharmaceutical companies R&D expenditures while its probability of success remains almost half of that in Phase I (Roy 2012, Hay *et al.* 2014). Identifying the "right" dose, carried out in Phase II, is a critical step in drug development partly because of high attrition rates in Phase III clinical trials, the most costly phase, which may be due to inadequate dose selection, i.e., doses that are too low to achieve a desired benefit or doses that are too high and result in adverse reactions (Bornkamp *et al.* 2007).

In a standard (static) clinical trial, patients are randomly assigned to predetermined doses such that the number of patients allocated to each dose is roughly equal. Such a design may be inefficient. For example, if the slope of a dose-response curve is observed at a dose range not anticipated, equal assignment of patients to other dose ranges may lead to inefficient use of resources. These allocations may expose patients to toxic or ineffective doses which raises ethical concerns. In

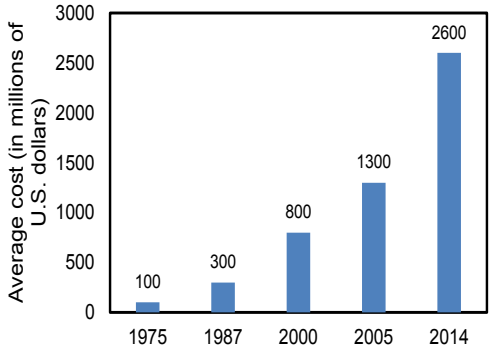


Figure 2.1: Exponential increase in cost of developing a new drug over time

addition, observations in the trial may indicate a larger variability in response to a particular dose, and thus fixed sample sizes cannot compensate for the unanticipated variability (Berry *et al.* 2002).

A better strategy is adaptive design in which modifications are made, while the study is in progress, as information accrues during the course of the trial. For example, the experimenters may increase the number of doses under consideration in the study, drop some doses from analysis, and change the patients' randomization procedure to avoid large sample sizes at doses where the shape of the curve is reasonably well estimated by the available data. Thus, adaptive designs tend to generally reduce length, total sample size, and costs of trials without compromising their integrity. Furthermore, such designs have an ethical motivation as they randomize patients to doses currently thought to be the best with greater probability (Rosenberger 1996).

Main contributions. In this chapter, we make the following contributions: Instead of using a Normal Dynamic Linear Model (NDLM) to approximate the dose-response curve as is done in the literature, we propose a new approximation to the curve such that the dynamic programming formulation of the problem enjoys conjugacy property, by which (i) we derive a couple of structural properties to the learning problem, (ii) reduce problem complexity and computational efforts, (iii) achieve faster convergence rates without sacrificing accuracy, and (iv) prove consistency of the design, i.e., learning the true underlying dose-response model when the number of patients becomes large almost surely, thus the right dose.

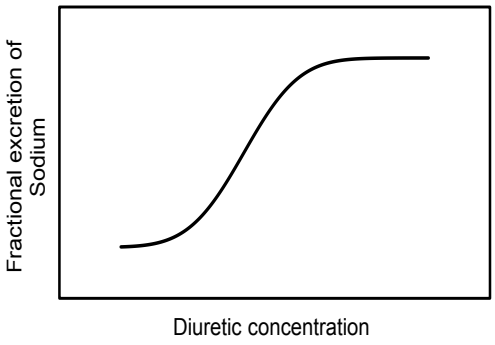


Figure 2.2: Dose-response curve of loop diuretic

2.2 Background

Before reviewing the vast literature on finding the right dose in a clinical trial, we describe the dose-response relationship, which is usually demonstrated by a curve and formulated as a random function of doses and unknown parameters. Then, we briefly review related works and highlight the distinctions of our approach.

2.2.1 Dose-Response Relationship

A dose-response curve identifies the relationship between the treatment dose of a drug and the patient's response usually measured by a numerical score. For example, consider the increase in fractional excretion of Sodium as the response to the amount of loop diuretic dose prescribed. Figure 2.2 shows a typical dose-response relationship for heart failure patients (Felker 2012). Let y denote the response score for a patient, and $z \in \mathcal{Z}$ denote the dose assigned to the patient, where $\mathcal{Z} := \{Z_j : j = 1, \ldots, J\}$ refers to the set of allowable doses, and Z_1 denotes the placebo. We let $f(z, \Theta)$ denote the dose-response curve as a function of dose z, parameterized by an unknown parameter vector $\Theta = (\theta_1, \ldots, \theta_J)'$. Hereafter, to ease the notation, we use index j in to refer to dose Z_j, e.g., $\theta_{Z_j} = \theta_j$. In particular, we assume that

$$y = f(z, \Theta) + \epsilon, \tag{2.1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ (see, e.g., Berry *et al.* 2002).

Existing literature focuses on three definitions for the right dose: 1) Minimum effective dose

defined as the smallest dose producing a particularly relevant response. 2) Maximum tolerable dose defined as the highest dose producing a desired response without unacceptable toxicity. 3) ED_{95} defined as the smallest dose at which 95% of the maximal response is achieved. In this chapter, we focus on estimating the ED_{95} formally represented as

$$\text{ED}_{95} = \min_{z} \big\{ z \in \mathcal{Z} : f(z, \Theta) \geq 0.95 f(z_{max}, \Theta) \big\}, \tag{2.2}$$

where z_{max} is a dose at which maximal response is observed. However, our proposed approach can be used for other definitions as well.

2.2.2 Related Works

There are three streams of literature related to this problem: literature on (i) optimal design of dose-finding trials in which finding a target dose is the objective, (ii) adaptive design of dose-finding trials in which sampling policies are adapted to observed responses while the trial is still in progress, and (iii) dynamic learning and knowledge gradient policies which demonstrate how dynamic learning techniques could be utilized in deriving adaptive policies.

Optimal design of dose-finding trials. In this line of literature, researchers have investigated efficient designs for estimating the target dose (right dose) by, e.g., minimizing its asymptotic variance under a particular dose-response model. For example, Biedermann *et al.* (2006), Wang (2006), Dette *et al.* (2008), Bretz *et al.* (2010), Dette *et al.* (2014) and Holland-Letz & Kopp-Schneider (2015) proposed optimal designs estimating the dose-response curve but did not consider response-adaptivity, and thus their designs are unable to modify dose range, sample size, or allocation scheme while the trial is still in progress. These designs are also dependent on prespecified dose-response models which are susceptible to misspecification of assumption and parameters. In contrast, the decisions in response-adaptive designs are subject to change as data is accrued and the dose-response models are not prespecified. In fact, the dose-response curves are assumed to be unknown and are usually approximated by a piecewise linear model which can estimate a wide range of practical dose-response relationships; see Berger & Wong (2009).

Response-adaptive clinical trials. In response-adaptive designs, patient allocation, dose range, and sample size are subject to modification when a new response is observed. Multi-armed bandit framework and Bayesian decision theory have been two of the most active lines of literature in

response-adaptive designs. In the multi-armed bandit approach, a decision maker selects a treatment based on observed information to maximize an expected (discounted) reward. For example, Cheng & Berry (2007), and Press (2009) developed response-adaptive two-armed bandits for sequential experiments such as clinical trials where information acquired during the trial was used to modify the allocation scheme and sampling size. Such designs were applicable only when two treatments are considered and their responses are binary (success or failure) in nature. Lai & Liao (2012) and Villar & Rosenberger (2018) extended the two-armed design to multi-armed bandits capable of comparing multiple treatments with continuous responses. However, the bandit structure is designed to identify the maximum reward when compared to a control treatment, and thus the policies derived are applicable for Phase III of clinical trials where a confirmatory study is necessary to test the benefits of new treatments versus a control treatment. For more details on benefits and challenges of applying multi-armed bandits in clinical trials, see Villar *et al.* (2015) and references therein.

Here, our focus is on the response-adaptive dose-finding clinical trials where Bayesian decision theory is utilized to design response-adaptive sequential sampling policies to identify a target dose. For example, Berry *et al.* (2002) and Müller *et al.* (2006) used NDLM, a piecewise linear model in which a point and the slope of each linear piece is updated in a Bayesian framework in order to approximate an unknown dose-response curve, and to formulate an adaptive dose allocation scheme (see details in Section 2.3). Weir *et al.* (2007) compared the standard Markov chain Monte Carlo (MCMC) simulation of NDLMs to estimate the dose-response curve with that of an importance sampling method. Furthermore, Krams *et al.* (2003) employed a similar approach in Acute Stroke Therapy by Inhibition of Neutrophils (ASTIN) clinical trial and used a fully Bayesian analysis for patient randomization and stopping criterion which was approved by the regulatory authorities. Smith *et al.* (2006), Warner *et al.* (2015), Lenz *et al.* (2015), Liu *et al.* (2017), and Holm Hansen *et al.* (2017) employed similar approaches in real dose-finding clinical trials or proof of concept studies. Similar to response-adaptive designs that employ NDLM to approximate an unknown dose-response curve, our approach also uses a piecewise linear model to estimate the dose-response curve. However, the induced conjugacy of our design eliminates the required time consuming MCMC simulation and has the benefit of consistency, in that the design learns the underlying true dose-response model perfectly and thus the right dose. Consistency of sequential allocation policies for dose-finding studies has not been addressed yet. In fact, our numerical analysis provides examples where the standard

approach fails to identify the target dose (Figures 2.15 and 2.16). Moreover, we exhibit several structural properties to the dose-response problem via dynamic programming techniques.

Dynamic learning and knowledge gradient. Next, we briefly review dynamic learning literature related to the problem in this chapter. For a comprehensive review of optimal learning, see Powell & Ryzhov (2012) and references therein.

Ranking and selection is a class of learning problems in which a risk-neutral decision maker seeks to find the best population, in terms of their expected value given a fixed budget to learn the unknown true distribution of the population (Gibbons *et al.* 1999). For this class of learning problems, Gupta & Miescke (1996) introduced the knowledge gradient (KG) algorithm for offline versions of ranking and selection problems where the algorithm chooses its future measurements by optimizing the one-step expected value function with respect to what is known so far. Frazier *et al.* (2008) extended the KG algorithm by assuming an independent multivariate normal prior on a piecewise linear approximation of an unknown function. Thus, the algorithm could learn the true function through Bayesian updating of prior moments. Frazier *et al.* (2009) further developed the method to accommodate correlated normal prior beliefs. Ryzhov *et al.* (2010) and Ryzhov & Powell (2011b) extended the algorithm to solve multi-armed bandits with exponential, Bernoulli, Poisson and uniform rewards. Ryzhov *et al.* (2012) adapted the method for a general class of online problems in multi-armed bandit literature. Furthermore, Negoescu *et al.* (2011) investigated KG policies in drug discovery problems; Ryzhov & Powell (2011a) applied it in information collection on graphs; Xie *et al.* (2016) developed a KG policy for pairwise sampling by common random numbers; and Wang *et al.* (2016) provided a KG policy for multi-armed bandits with binary responses. Edwards *et al.* (2017) reviews KG algorithms and identifies one important limitation, namely "dominated actions" which are chosen by KG in multi-armed bandit settings when in fact the chosen arms have inferior exploitation and exploration values for non Gaussian rewards. The objective in these studies is to maximize the expected total reward, where the reward function is equivalent to the posterior mean of the unknown parameter. In contrast, our approach minimizes the variance of the target dose, a non-linear function of the unknown parameter describing the dose-response curve.

There are several recent studies employing dynamic learning into clinical trials. Kotas & Ghate (2016, 2017) formulated a dynamic programming to optimal dose-finding and approximated the Bellman equation by suppressing uncertainty of the unknown parameters and random transitions between states. Ahuja & Birge (2016), Chick *et al.* (2017), and Negoescu *et al.* (2017) considered

adaptive two-armed bandits and implemented dynamic learning techniques to identify the most efficacious treatment in a variety of settings. However, as elaborated earlier, the structure of such designs are appropriate for Phase III clinical trials.

2.3 The State-of-the-Art Approach

In this section, we present a formal dynamic programming formulation for the existing response-adaptive dose-finding models (e.g., Berry *et al.* 2002) and formalize the one-step look-ahead policy available in the literature (e.g., Krams *et al.* 2003, Weir *et al.* 2007). The standard approach for a response-adaptive Phase II clinical trial studies the problem of allocating a number of homogeneous patients to a set of dose options. Patients are sequentially assigned to doses and their responses are observed before the assignment of the next patient. Assume that the investigator chooses a model to describe the dose-response relationship and uses data from patients to estimate model parameters and consequently identifies the target dose. Aligned with Berry *et al.* (2002), we use NDLM to describe the dose-response relationship as it provides the most flexibility to summarize the data. In particular, NDLMs are piecewise linear models that can approximate both monotonic and non-monotonic dose-response relationships. Furthermore, recursive methods to calculate moments of their posterior distribution already exist. Note that the following formulation is an approximation of the true dose-response relationship described in Section 2.2.1.

Let N denote the total number of patients to be sampled in the trial. Define y_j^k as the response of kth patient assigned to dose j, where $1 \leq k \leq N$. Following Berry *et al.* (2002), assume that the dose-response relationship is formalized by $f(z, \Theta) = \theta_z$. Therefore, according to formulation (2.1), the resulting dose-response model is

$$y_j^k = \theta_j + \epsilon_j^k, \qquad j = 1, \ldots, J,\ 1 \leq k \leq N. \tag{2.3}$$

Construct vector Y_j^n from responses of patients assigned to dose j when n observation has been made. Recall that Θ indicates the column vector $\Theta = (\theta_1, \ldots, \theta_J)'$ and z^n denotes the dose assigned to patient n. Therefore, the observed response after assigning dose z^n to patient n is $\hat{y}^{n+1} = \theta_{z^n} + \epsilon^{n+1}$. Note that conditional on Θ and z^n, the sampled observation $\hat{y}^{n+1}$ has a normal distribution $(\hat{y}^{n+1}|\Theta, z^n) \sim \mathcal{N}(\theta_{z^n}, \sigma^2)$. Define filtration $\mathscr{F}^n$ as the σ-algebra generated by sampling doses and

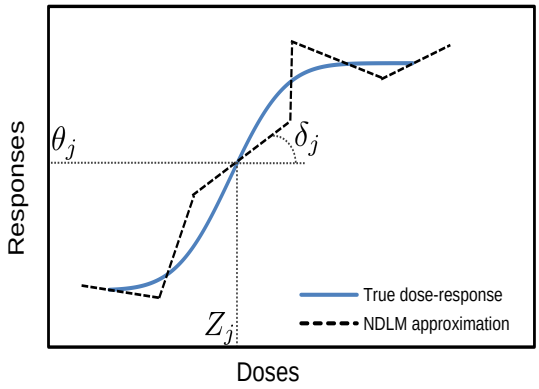


Figure 2.3: Standard piecewise linear approximation to dose-response curve

corresponding responses by time n, i.e., $\mathscr{F}^n$ is a σ-algebra generated by $z^0, \hat{y}^1, z^1, \hat{y}^2, z^2, \ldots, z^{n-1}, \hat{y}^n$. Notice that z^0 denotes the assignment dose before observing any response.

State space. The response-adaptive dose-finding problem is formulated within a dynamic programming framework. To that end, we set the decision epochs at times when a dose is allocated to patient n where $n \in \{0, \ldots, N\}$. No decision is made at decision epoch N. Using the piecewise linearity of the second order NDLM, we fit a locally linear curve to the true underlying dose-response relationship, i.e., for dose z close enough to dose j, the response is a straight line $\theta_j + (z - j)\delta_j$. Figure 2.3 shows such a piecewise linear approximation to the dose-response curve. Linear extrapolation will result in the relationship $\theta_j = \theta_{j-1} + \delta_{j-1}$ between one dose to the next (see, e.g., Berry *et al.* 2002).

Therefore, the evolution of parameters (θ_j, δ_j) from one dose to the next is adjusted by introducing normal residual errors in the following form

$$\begin{pmatrix} \theta_j \\ \delta_j \end{pmatrix} = \begin{pmatrix} \theta_{j-1} + \delta_{j-1} \\ \delta_{j-1} \end{pmatrix} + \begin{pmatrix} \nu_j \\ \omega_j \end{pmatrix}, \tag{2.4}$$

where ν_j and $\omega_j \sim \mathcal{N}(0, W_j)$. Assume that $W_j = W$ for all $j = 1, \ldots, J$, with known and fixed W. Let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_J)'$ indicate a column vector. In order to define the state space of the problem, let $q^n(\Theta, \boldsymbol{\delta})$ denote a probability distribution on parameters Θ and $\boldsymbol{\delta}$ given data $\mathscr{F}^n$, i.e., $q^n(\Theta, \boldsymbol{\delta}) = p(\Theta, \boldsymbol{\delta} | \mathscr{F}^n)$, which is the posterior distribution of Θ and $\boldsymbol{\delta}$ after observing responses of n patients. The state of the system is therefore $s^n = q^n(\Theta, \boldsymbol{\delta})$ and the state space $\mathcal{S}$ is the set of all

probability distributions on $(\Theta, \boldsymbol{\delta}|\mathscr{F}^n)$ such that

$$s^n \in \mathcal{S} := \left\{ p(\Theta, \boldsymbol{\delta}) : \theta_j = \theta_{j-1} + \delta_{j-1} + \nu_j,\ \delta_j = \delta_{j-1} + \omega_j,\ \nu_j \text{ and } \omega_j \sim \mathcal{N}(0, W_j),\ j = 1, \ldots, J \right\},$$

where $p(\cdot)$ denotes a probability density function (pdf) and conjugate hyperpriors for θ_1 and δ_1 are given. Note that $q^n(\Theta, \boldsymbol{\delta})$ is in proportion to a multivariate normal distribution where $(\theta_j, \delta_j|\mathscr{F}^n)$ follows a bivariate normal distribution, the moments of which are given by recursive equations (2.6). However, sampling from the joint distribution $q^n(\Theta, \boldsymbol{\delta})$ remains computationally quite challenging. Therefore, methods such as "forward filtering, backward sampling" have been proposed to sample from such distributions (West & Harrison 1997). Also, only $\mathbb{E}(\theta_1)$, $\mathbb{E}(\delta_1)$, $\text{Var}(\theta_1)$, and $\text{Var}(\delta_1)$ are fixed and the prior expectation and variance for δ_j and $\theta_j, j > 1$ are determined by the evolution equations (2.4).

Action space. This is described by whether dose $z^n = j$ is assigned to patient n. Let a_j^n denote the action prescribed for patient n when the state of the trial is $q^n(\Theta, \boldsymbol{\delta})$, i.e,

$$a_j^n = \begin{cases} 1 & \text{if dose } j \text{ is assigned to patient } n, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the action space is denoted by

$$A(s^n) := \left\{ a_j^n \in \{0, 1\}\, \forall j : \sum_{j=1}^{J} a_j^n = 1 \right\}.$$

Transitions. Assume that the response of a patient is observed before the assignment of a dose to the next patient. Given $a_j^n = 1$ and Θ, the response of a patient has a normal distribution $\hat{y}^{n+1} \sim \mathcal{N}(\theta_j, \sigma^2)$. We wish to find the posterior distribution of $(\Theta, \boldsymbol{\delta})$ given the response to action a_j^n is observed. Using Bayes' law,

$$q^{n+1}(\Theta, \boldsymbol{\delta}) = p(\Theta, \boldsymbol{\delta}|\mathscr{F}^{n+1}) \propto \mathcal{L}(Y_1^n, \ldots, Y_J^n|\Theta, \boldsymbol{\delta}) q^n(\Theta, \boldsymbol{\delta}),$$

where $\mathcal{L}(\cdot)$ is the likelihood function. Here, the action a_j^n determines z^n which along with its response $\hat{y}^{n+1}$ are included in $\mathscr{F}^{n+1}$.

West & Harrison (1997) provided a recursive algorithm to generate samples from the poste-

rior distribution $q^n(\Theta, \boldsymbol{\delta})$ under certain prior assumptions for a general multivariate NDLM. Notice that the size of Y_j^n is known at decision epoch n and denote it by L_j^n where $\sum_j L_j^n = n$. consider the following multivariate NDLM model

$$\begin{aligned} (Y_j^n)' &= F_j'\alpha_j + \mathcal{V}_j, \\ \alpha_j &= G_j\alpha_{j-1} + \Omega_j, \end{aligned} \tag{2.5}$$

where $\alpha_j = \binom{\theta_j}{\delta_j}$ is a 2-dimensional vector; F_j is a known $(2 \times L_j^n)$ dynamic regression matrix; G_j is a known (2×2) evolution matrix; $\mathcal{V}_j \sim \mathcal{N}(0, \mathcal{E}_j)$ and $\Omega_j \sim \mathcal{N}(0, \mathcal{W}_j)$ where $\mathcal{E}_j$ and $\mathcal{W}_j$ are known $(L_j^n \times L_j^n)$ and (2×2) variance matrices, respectively. Note that $\mathcal{V}_j$ and Ω_j are $(L_j^n \times 1)$ and (2×1) column vectors. In particular, $F_j = \left[\begin{smallmatrix} 1 \cdots 1 \\ 0 \cdots 0 \end{smallmatrix}\right]_{(2 \times L_j^n)}$, $G_j = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$, $\mathcal{E}_j$ is a diagonal matrix with ϵ_j^k as the elements of the main diagonal, and $\mathcal{W}_j$ is also a diagonal matrix with ν_j and ω_j on the main diagonal. Notice that formulation (2.5) is a general form of observation equations (2.3) and evolution equations (2.4). The posterior probability distribution on α_j given $\mathscr{F}^n$ is a bivariate normal $(\alpha_j|\mathscr{F}^n) \sim \mathcal{N}(m_j, C_j)$, where

$$\begin{aligned} m_j &= d_j + D_j\kappa_j, & Q_j &= F_j'R_jF_j + \mathcal{E}_j, \\ d_j &= G_jm_{j-1}, & C_j &= R_j - D_jQ_jD_j', \\ D_j &= R_jF_jQ_j^{-1}, & \kappa_j &= Y_j^n - f_j, \\ R_j &= G_jC_{j-1}G_j' + \mathcal{W}_j, & f_j &= F_j'd_j. \end{aligned} \tag{2.6}$$

Note that hyperparameters $m_1 = \left[\begin{smallmatrix} \mathbb{E}(\theta_1) \\ \mathbb{E}(\delta_1) \end{smallmatrix}\right]$ and $C_1 = \left[\begin{smallmatrix} \mathrm{Var}(\theta_1) & 0 \\ 0 & \mathrm{Var}(\delta_1) \end{smallmatrix}\right]$ are given. Carter & Kohn (1994) and Frühwirth-Schnatter (1994) developed a "forward filtering, backward sampling" (FFBS) algorithm to generate random samples from the full posterior distribution. In case of a general multivariate NDLM, the algorithm samples α_J from $(\alpha_J|\mathscr{F}^n) \sim \mathcal{N}(m_J, C_J)$ and then, for each $j = J-1, J-2, \ldots, 1$, samples α_j from $(\alpha_j|\alpha_{j+1}, \mathscr{F}^n) \sim \mathcal{N}(h_j, H_j)$ where

$$\begin{aligned} h_j &= m_j + B_j(\alpha_{j+1} - d_{j+1}), \\ H_j &= C_j - B_jR_{j+1}B_j', \\ B_j &= C_jG_{j+1}'R_{j+1}^{-1}. \end{aligned} \tag{2.7}$$

Thus, the algorithm moves forward from $j = 1$ to $j = J$, and computes m_j, C_j, R_j, B_j, and d_j. At $j = J$, α_J is sampled and the algorithm moves backwards from $j = J$ to $j = 1$ to compute h_j and

H_j at each step and samples α_j (see West & Harrison 1997, Ch. 15).

Objective function. In Phase II of the response-adaptive dose-finding trials, finding the target dose, e.g., ED_{95} is considered amongst the ultimate goals. The decision maker must choose a sequence of dose assignments such that learning the target dose is achieved quickly and accurately. Therefore, minimizing the variance of the target dose at the end of the trial which is equivalent to minimizing the uncertainty about the target dose is considered as the objective. Note that the target dose at the end of the trial is a random variable at times $n < N$. The expected cost at the end of the trial is

$$r_N(s^N) = \text{Var}\big(g(\Theta)|s^N\big), \tag{2.8}$$

where $s^N = q^N(\Theta, \boldsymbol{\delta})$, $g(\Theta) = \text{ED}_{95} = \min_z \big\{z \in \mathcal{Z} : f(z, \Theta) \geq 0.95 f(z_{max}, \Theta)\big\}$, and $r_n(s^n) = 0$ for $n = 0, \ldots, N-1$. Define policy π as a mapping from the state space to the action space, and let $l_\pi(s^0)$ denote the expected variance of ED_{95} with respect to $\mathscr{F}^N$ at the end of the trial under policy π when the initial prior on $(\Theta, \boldsymbol{\delta})$ is $s^0 = q^0(\Theta, \boldsymbol{\delta})$. The decision maker solves for $V^0(s^0) = \inf_{\pi \in \Pi} l_\pi(s^0)$, where Π is the set of non-anticipative admissible policies under consideration. The optimal value function is the unique solution to the Bellman equation

$$\begin{aligned} &V^n(s^n) = \min_{a_j^n \in A(s^n)} \mathbb{E}\Big\{V^{n+1}(s^{n+1})\Big|s^n, a_j^n\Big\},\ n = 0, \ldots, N-1, \\ &V^N(s^N) = \text{Var}\big(g(\Theta)\big|s^N\big). \end{aligned} \tag{2.9}$$

One-step look-ahead policy. Since the state space corresponds to a set of uncountable probability distributions, solving (2.9) is computationally infeasible. Therefore, a one-step look-ahead policy is proposed in the literature and carried out in dose-finding studies (e.g., Krams *et al.* 2003). This policy selects the dose with the minimum expected variance of ED_{95} while assuming the trial stops after assigning the next dose. The approach is formalized by Algorithm 2 which must run for each patient. In particular, after dose z^* is selected for assignment at time n, a true response $y_{z^*}^{n+1}$ is observed and is added to vector $Y_{z^*}^{n+1}$, which is used to update equations (2.6) and (2.7) and thus moments of prior distributions on $(\Theta, \boldsymbol{\delta})$, i.e., m_J, C_J, h_j, and H_j. Note that $\tilde{\Theta}, \tilde{\boldsymbol{\delta}}, \tilde{y}_{zm}$, and $\tilde{Y}_z$ as well as $\hat{\Theta}, \hat{\boldsymbol{\delta}}, \hat{m}, \hat{C}, \hat{h}$, and $\hat{H}$ are temporary and would be discarded at the end of each loop. Furthermore, in reporting the results, the "for" loop sampling $\Theta_{(m=1:M)}$ is parallelized using the "*foreach*" package in R. Notice that generating a posterior sample of $(\Theta, \boldsymbol{\delta}|\mathscr{F}^n)$ involves running the FFBS algorithm, which even with applying the one-step look-ahead policy is extremely time

Algorithm 2 One-step look-ahead policy for the state-of-the-art approach

for each dose $z \in \mathcal{Z}$ **do**
  Generate M samples of $(\tilde{\Theta}, \tilde{\boldsymbol{\delta}})$ from prior where $(\theta_J, \delta_J) \sim \mathcal{N}(m_J, C_J)$ and $(\theta_j, \delta_j) \sim \mathcal{N}(h_j, H_j)$

  for each sampled $\tilde{\Theta}_{(m=1:M)}$ **do**
    Simulate future observation $\tilde{y}_{zm} \sim \mathcal{N}(\tilde{\theta}_{zm}, \sigma^2)$.
    Add response $\tilde{y}_{zm}$ to vector $\tilde{Y}_z$ and update equations (2.6) and (2.7)
    Generate T posterior samples of $(\hat{\Theta}, \hat{\boldsymbol{\delta}})$ using the updated $\mathcal{N}(\hat{m}_J, \hat{C}_J)$ and $\mathcal{N}(\hat{h}_j, \hat{H}_j)$.
    for each sampled $\hat{\Theta}_{(t=1:T)}$ **do**
      Find $g(\hat{\Theta}_{(t)}) = \min_z\{z \in \mathcal{Z} : f(z, \hat{\Theta}_{(t)}) \geq 0.95 f(z_{max}, \hat{\Theta}_{(t)})\}$.
    Estimate the observed variance $U_{zm} = \mathrm{Var}\big[g(\hat{\Theta})\big|\mathscr{F}^n \cup (z, \tilde{y}_{zm})\big]$ using sample variance.
  Estimate the variance U_z for each dose by taking a Monte Carlo sample average $\sum_m \frac{U_{zm}}{M}$.
Select the dose z^* that minimizes U_z.

consuming. In the next section, we propose a new approach and show that it produces high-quality solutions more efficiently.

2.4 The Knowledge Gradient Approach

Recall that the standard approach in Section 2.3 approximates the dose-response relationship by a piecewise linear function, where at each dose j, a straight line with slope δ_j and point (j, θ_j) estimates the curve. We propose a novel approximation where the curve is approximated by connecting the points (j, θ_j) such that the slope between two consecutive doses is $\theta_{j+1} - \theta_j$. The proposed piecewise linear approximation enjoys conjugacy property over the states and the sampling distribution under the same assumptions as in Section 2.3, which significantly reduces the complexity of the problem and the computational effort needed to solve it. Figure 2.4 compares the proposed piecewise linear approximation with the NDLM approximation in a typical dose-response curve.

Similar to the standard approach, assume that the dose-response curve is of the form $f(z, \Theta) = \theta_z$ and samples from dose z are independent and normally distributed with mean Θ and known variance σ^2. One only needs to keep track of θ_j for each dose j in the proposed approach. Therefore, the state only includes the decision maker's belief regarding $\Theta = (\theta_1, \ldots, \theta_J)'$. Let $\mathbb{E}_n$ be the conditional expectation with respect to $\mathscr{F}^n$, i.e., $\mathbb{E}_n[\cdot] = \mathbb{E}[\cdot|\mathscr{F}^n]$. Defining $\mu^n := \mathbb{E}_n[\Theta]$ and $\Sigma_n := \mathrm{Cov}[\Theta|\mathscr{F}^n]$ while assuming a multivariate normal prior with mean μ^0 and positive semidefinite covariance matrix Σ^0 on the belief about Θ, i.e., $\Theta \sim \mathcal{N}(\mu^0, \Sigma^0)$, will result in a multivariate

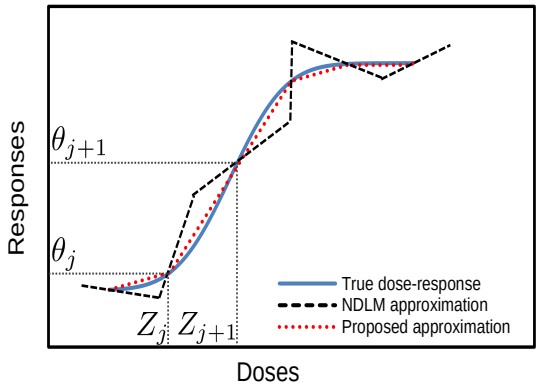


Figure 2.4: Standard vs. proposed piecewise linear approximation to dose-response curve

normal posterior belief about Θ with mean μ^n and covariance matrix Σ^n when conditioned on $\mathscr{F}^n$.

State and action spaces. Similar to Section 2.3, decision epochs are at the times when doses are assigned to patients. The state of the system at decision epoch n is also the probability distribution on parameter Θ given data $\mathscr{F}^n$, i.e., $q^n(\Theta) = p(\Theta|\mathscr{F}^n)$. However, in our proposed approach, the random variable Θ, our belief about the true response means of allowable doses, is normally distributed at time n. Therefore, the probability distribution on Θ is normal and can be completely described by its mean vector μ^n and covariance matrix Σ^n. Thus, define $s^n := (\mu^n, \Sigma^n)$ as the state of the trial at time n and set $\mathcal{S}$ to be the state space

$$s^n \in \mathcal{S} := \Big\{(\mu, \Sigma) : \mu \in \mathbb{R}^J,\, \Sigma \in \Psi\Big\},$$

where Ψ denotes the set of $J \times J$ positive semidefinite matrices. The action a_j^n has a value of 1 if doze j is assigned to patient n and 0 otherwise. The action space is denoted by $A(s^n) := \big\{a_j^n \in \{0,1\}\, \forall j : \sum_{j=1}^J a_j^n = 1\big\}$.

Transitions. Recall that our prior belief on Θ is a multivariate normal distribution. In addition, sample observations $\hat{y}^{n+1}$ are normally distributed. Therefore, the posterior distribution on Θ, which is specified by μ^{n+1} and Σ^{n+1} is also a multivariate normal distribution. The relationship between the prior and the posterior is characterized by state s^n, action a_j^n and the random response $\hat{y}^{n+1}$. Assuming that the covariance matrix Σ^n is nonsingular for now, μ^{n+1} and Σ^{n+1} can be written as (Gelman *et al.* 2004)

$$\begin{aligned} \mu^{n+1} &= \Sigma^{n+1}\big((\Sigma^n)^{-1}\mu^n + (\sigma^2)^{-1}\hat{y}^{n+1}e_j\big), \\ \Sigma^{n+1} &= \big((\Sigma^n)^{-1} + (\sigma^2)^{-1}e_j e_j'\big)^{-1}, \end{aligned} \tag{2.10}$$

where e_j is a J-vector of 0s and a single 1 at j^{th} index assuming $a_j^n = 1$. This formulation only holds when Σ^n is positive-definite and invertible, however, notice that $(\sigma^2)^{-1}e_j e_j'$ only changes one element of matrix $(\Sigma^n)^{-1}$. Using Sherman-Morrison formula to adjust the inverse of a matrix when only one element has changed, formulation (2.10) can be written in such a way that Σ^n is positive semidefinite and no longer needs to be invertible (Sherman & Morrison 1950),

$$\begin{aligned}\mu^{n+1} &= \mu^n + \frac{\hat{y}^{n+1} - \mu_j^n}{\sigma^2 + \Sigma_{jj}^n}\Sigma^n e_j', \\ \Sigma^{n+1} &= \Sigma^n - \frac{\Sigma^n e_j e_j' \Sigma^n}{\sigma^2 + \Sigma_{jj}^n}.\end{aligned} \tag{2.11}$$

Define $\tilde{\sigma}$ as a vector-valued function $\tilde{\sigma}\big(\Sigma, a_j^n\big) := \dfrac{\Sigma e_j}{\sqrt{\sigma^2 + \Sigma_{jj}}}$, and note that $\text{Var}[\hat{y}^{n+1} - \mu^n | \mathscr{F}^n] = \text{Var}[\theta_{z^n} + \epsilon^{n+1} | \mathscr{F}^n] = \sigma^2 + \Sigma_{jj}^n$. Define random variable $X^{n+1} := \dfrac{(\hat{y}^{n+1} - \mu^n)}{\sqrt{\text{Var}[\hat{y}^{n+1} - \mu^n | \mathscr{F}^n]}}$ by which formulation (2.11) is equivalent to

$$\begin{aligned}\mu^{n+1} &= \mu^n + \tilde{\sigma}\big(\Sigma^n, a_j^n\big)X^{n+1}, \\ \Sigma^{n+1} &= \Sigma^n - \tilde{\sigma}\big(\Sigma^n, a_j^n\big)\tilde{\sigma}'\big(\Sigma^n, a_j^n\big),\end{aligned} \tag{2.12}$$

where random variable X^{n+1} is standard normal when conditioned on $\mathscr{F}^n$.

Objective function. Similar to Section 2.3, consider minimizing the variance of the target dose at the end of the trial as our objective. Therefore, the expected cost at the end of the trial is $r_N(s^N) = \text{Var}\big(g(\Theta)\big|s^N\big)$ where $s^N = (\mu^N, \Sigma^N)$. At each decision epoch, the selected dose z^n is allowed to depend on samples by time n, that is $z^n \in \mathscr{F}^n$. Note that z^n is completely determined by a_j^n, i.e., if $a_j^n = 1$, then $z^n = j$. Thus, in order to ease the notation, hereafter, we use z^n to denote the selected dose by action a_j^n at time n. Define $\Pi := \big\{(z^0, \ldots, z^{N-1}) : z^n \in \mathscr{F}^n\big\}$ to be the set of measurement policies, where $\pi = (z^0, \ldots, z^{N-1})$ is an element in Π. Let $l_\pi(s^0)$ denote the expected variance of ED_{95} at the end of trial when the initial prior on θ is $\mathcal{N}(\mu^0, \Sigma^0)$. Choosing a policy that minimizes the expected cost is achieved by solving

$$V(s^0) = \inf_{\pi \in \Pi} l_\pi(s^0), \tag{2.13}$$

where $l_\pi(\cdot) = \mathbb{E}^\pi\Big\{\text{Var}_N\big[g(\Theta)\big]\Big|s^0 = (\mu^0, \Sigma^0)\Big\}$, $\mathbb{E}^\pi\big\{\,\cdot\,\big\}$ indicates expectation taken with respect to a fixed measurement policy π, and $\text{Var}_N(\cdot)$ is the variance with respect to $\mathscr{F}^N$. Defining a sequence

of value functions at each decision epoch $n \leq N-1$ as $V^n(s^n)$, the optimal value function is the unique solution to these Bellman equations

$$\begin{aligned} V^n(s^n) &= \min_{z^n} \mathbb{E}\Big\{ V^{n+1}(s^{n+1}) \Big| s^n, z^n \Big\}, \\ V^N(s^N) &= \mathrm{Var}\big(g(\Theta)\big|s^N\big). \end{aligned} \tag{2.14}$$

Define $\mathcal{Q}^n(s,z) := \mathbb{E}\Big[V^{n+1}\big(\eta(s^n, z^n, X^{n+1})\big) \Big| s^n = s, z^n = z \Big]$ for any $s \in \mathcal{S}$ as a function measuring the value of assigning dose z to patient n when the trial is in state s^n, where $\eta(\cdot)$ is a transition function which by using the updating equations (2.12) determines the next state, i.e., $s^{n+1} = \eta(s^n, z^n, X^{n+1})$. Denote $V^{n+1}(s^n)$ as the value of making no measurements while in state s^n. The following theorem states that the optimal policy always prefers to make a measurement.

Theorem 2.4.1 *The optimal policy always prefers to measure an alternative dose rather than to measure nothing at all, i.e, $Q^n(s,z) \leq V^{n+1}(s)$ for every $s \in \mathcal{S}$, $0 \leq n < N$ and $z \in \{1, \ldots, J\}$.*

Proof. The theorem is proven by induction on n. First, we show that the theorem holds for $n = N-1$.

$$\begin{aligned} Q^{N-1}(s,z) &= \mathbb{E}\Big\{ V^N\big(\eta(s^{N-1}, z^{N-1}, X^N)\big) \Big| s^{N-1} = s, z^{N-1} = z \Big\} \\ &= \mathbb{E}\Big\{ \mathrm{Var}\Big(g(\Theta) \Big| \eta(s^{N-1}, z^{N-1}, X^N) \Big) \Big| s^{N-1} = s, z^{N-1} = z \Big\} \\ &= \mathbb{E}\Big\{ \Big[\mathbb{E}\Big(g^2(\Theta) \Big| \eta(s^{N-1}, z^{N-1}, X^N) \Big) - \Big(\mathbb{E}\big(g(\Theta) \Big| \eta(s^{N-1}, z^{N-1}, X^N) \big) \Big)^2 \Big] \Big| \cdot \Big\} \\ &= \mathbb{E}\Big\{ \mathbb{E}\Big(g^2(\Theta) \Big| \eta(s^{N-1}, z^{N-1}, X^N) \Big) \Big| \cdot \Big\} - \mathbb{E}\Big\{ \Big(\mathbb{E}\big[g(\Theta) \Big| \eta(s^{N-1}, z^{N-1}, X^N) \big] \Big)^2 \Big| \cdot \Big\}, \end{aligned}$$

where the σ-field generated by s^{N-1} and z^{N-1} is a subset of the σ-field of the transition function at (s^{N-1}, z^{N-1}), i.e., $\sigma(s^{N-1}, z^{N-1}) \subseteq \sigma\big(\eta(s^{N-1}, z^{N-1}, X^N)\big)$, thus by using the tower property of conditional expectation

$$\begin{aligned} Q^{N-1}(s,z) &= \mathbb{E}\Big[g^2(\Theta) \Big| \cdot \Big] - \mathbb{E}\Big\{ \Big(\mathbb{E}\big[g(\Theta) \Big| \eta(s^{N-1}, z^{N-1}, X^N) \big] \Big)^2 \Big| \cdot \Big\} \\ &\leq \mathbb{E}\Big[g^2(\Theta) \Big| \cdot \Big] - \Big(\mathbb{E}\Big\{ \mathbb{E}\Big[g(\Theta) \Big| \eta(s^{N-1}, z^{N-1}, X^N) \Big] \Big| \cdot \Big\} \Big)^2 \\ &= \mathbb{E}\Big[g^2(\Theta) \Big| \cdot \Big] - \Big(\mathbb{E}\big\{ g(\Theta) \big| \cdot \big\} \Big)^2 \\ &= \mathrm{Var}\big[g(\Theta) \big| s^{N-1}, z^{N-1} \big] = V^N(s), \end{aligned}$$

where the inequality is the result of Jensen's inequality for convex functions. Therefore, $Q^{N-1}(s,z) \leq V^N(s)$. Now, suppose the induction hypothesis is true for $N-1,\ldots,n+1$. Then,

$$Q^{n+1}(s,z) = \mathbb{E}\Big\{ V^{n+2}\Big(\eta(s^{n+1}, z^{n+1}, X^{n+2})\Big)\Big| s^{n+1}=s, z^{n+1}=z\Big\} \leq V^{n+2}(s).$$

We must show that the induction hypothesis also holds for n. Recall that $V^n(s) = \min_z Q^n(s,z)$,

$$\begin{aligned}
Q^n(s,z) &= \mathbb{E}\Big\{ V^{n+1}\Big(\eta(s^n, z^n, X^{n+1})\Big)\Big| s^n=s, z^n=z\Big\} \\
&= \mathbb{E}\Big\{ \min_{z'} Q^{n+1}\Big(\eta(s^n, z^n, X^{n+1}), z'\Big)\Big| s^n=s, z^n=z\Big\} \\
&\leq \min_{z'} \mathbb{E}\Big\{ Q^{n+1}\Big(\eta(s^n, z^n, X^{n+1}), z'\Big)\Big| s^n=s, z^n=z\Big\} \\
&= \min_{z'} \mathbb{E}\Big\{ \mathbb{E}\Big[V^{n+2}\big(\eta(\eta(s^n, z^n, X^{n+1}), z', X^{n+2})\big)\Big|\eta(s^n, z^n, X^{n+1}), z'\Big]\Big|\cdot\Big\} \\
&= \min_{z'} \mathbb{E}\Big\{ V^{n+2}\Big[\eta\big(\eta(s^n, z^n, X^{n+1}), z', X^{n+2}\big)\Big]\Big|\cdot\Big\},
\end{aligned}$$

where the inequality is justified by the Jensen's inequality for concave functions, and the tower property of conditional expectation is applied. Note that $\eta\Big(\eta(s^n, z^n, X^{n+1}), z', X^{n+1}\Big)$ describes the state at which we arrive at time $n+2$ if we measure z^n first and then z'. However, since z^n and z' are fixed, the measurement order of z^n and z' does not affect the distribution of the state, thus we can change the order we measure z^n and z'. Therefore,

$$\begin{aligned}
Q^n(s,z) &\leq \min_{z'} \mathbb{E}\Big\{ V^{n+2}\Big[\eta\big(\eta(s^n, z^n, X^{n+1}), z', X^{n+2}\big)\Big]\Big|\cdot\Big\} \\
&= \min_{z'} \mathbb{E}\Big\{ V^{n+2}\Big[\eta\big(\eta(s^n, z', X^{n+2}), z^n, X^{n+1}\big)\Big]\Big|\cdot\Big\} \\
&= \min_{z'} \mathbb{E}\Big\{ \mathbb{E}\Big[V^{n+2}\Big(\eta\big(\eta(s^n, z', X^{n+2}), z^n, X^{n+1}\big)\Big)\Big|\eta(s^n, z', X^{n+2}), z^n\Big]\Big|\cdot\Big\} \\
&= \min_{z'} \mathbb{E}\Big\{ Q^{n+1}\Big(\eta(s^n, z', X^{n+2}), z^n\Big)\Big|\cdot\Big\}.
\end{aligned}$$

The induction hypothesis shows that $Q^{n+1}\Big(\eta(s^n, z', X^{n+2}), z^n\Big) \leq V^{n+2}\Big(\eta(s^n, z', X^{n+2})\Big)$ which results in

$$Q^n(s,z) \leq \min_{z'} \mathbb{E}\Big\{ Q^{n+1}\Big(\eta(s^n, z', X^{n+2}), z^n\Big)\Big|\cdot\Big\} \leq \min_{z'} \mathbb{E}\Big\{ V^{n+2}\Big(\eta(s^n, z', X^{n+2})\Big)\Big|\cdot\Big\} = V^{n+1}(s). \square$$

Theorem 2.4.1 shows that any extra measurement would be beneficial (not worse) to the value function at time n. In the following corollaries, the first suggests that there is no value in measuring a dose which is already known (its variance is zero), and the second corollary implies that the extra measurement should be made according to the optimal policy.

Corollary 2.4.1 *Let i, j denote any two doses where $i \neq j$, $n < N$, and $s = (\mu, \Sigma)$. If $\Sigma_{jj} = 0$ then $Q^n(s,i) \leq Q^n(s,j)$*

Proof. If $\Sigma^n_{jj} = 0$, then dose j is known almost surely and $\text{Cov}(\theta_j, \theta_t) = 0$ for all dose $t \in \{1, \ldots, J\}$. Therefore, the j^{th} row and column of the Σ^n matrix are equal to zero which results in $\tilde{\sigma}(s^n, j) = 0$. Recall that $\eta(s^n, j, X^{n+1})$ uses equations (2.12) to update the state. Since $\tilde{\sigma}(s^n, j) = 0$,

$$(\mu^{n+1}, \Sigma^{n+1}) = s^{n+1} = \eta(s^n, j, X^{n+1}) = s^n = (\mu^n, \Sigma^n).$$

Then, by applying Theorem 2.4.1,

$$Q^n(s^n, j) = \mathbb{E}\Big\{ V^{n+1}\big(\eta(s^n, j, X^{n+1})\big) \Big| s^n = s \Big\} = V^{n+1}(s^n) \geq Q^n(s^n, i). \square$$

Corollary 2.4.2 *$V^n(s) \leq V^{n+1}(s)$ for all states $s \in \mathcal{S}$.*

Proof. Theorem 2.4.1 shows that $\mathcal{Q}^n(s,z) \leq V^{n+1}(s)$. Therefore, $\min_z \mathcal{Q}^n(s,z) \leq V^{n+1}(s)$, where

$$\min_z \mathcal{Q}^n(s,z) = \min_{z \in \{1,\ldots,J\}} \mathbb{E}\Big\{ V^{n+1}\big(\eta(s^n, z^n, X^{n+1})\big) \Big| s^n = s, z^n = z \Big\} = V^n(s).$$

Thus, $V^n(s) \leq V^{n+1}(s)$.$\square$

Solving formulation (2.14) to optimality is impractical because of the state space continuity. Following the one-step look-ahead framework, we assume that the next patient in the trial will be the last, and allocate a dose to the next patient to minimize the expected value of a single period decision process. Note that $\text{Var}_n[g(\Theta)]$ is the value we would receive if we were to stop the trial at decision epoch n. Define the KG policy π^{KG} for every $s \in \mathcal{S}$ according to

$$\mathcal{X}^{\pi^{KG}}(s) \in \arg\min_z \mathbb{E}_n \Big\{ \text{Var}_{n+1}[g(\Theta)] - \text{Var}_n[g(\Theta)] \Big| s^n = s, z^n = z \Big\} \text{ for every } n < N, \qquad (2.15)$$

where $\mathcal{X}^{\pi^{KG}}(s^n)$ is a decision function which returns the dose selected in state s^n under the KG

policy π^{KG}, i.e., $\mathcal{X}^{\pi^{KG}}(s^n) := z^n$. In order to compute the KG policy, one needs to evaluate

$$\min_z \mathbb{E}_n\big\{\mathrm{Var}_{n+1}[g(\Theta)]\big|s^n = s, z^n = z\big\}$$

for every $s^n \in \mathcal{S}$ at each decision epoch n since $\mathrm{Var}_n[g(\Theta)]$ is constant with respect to $\mathscr{F}^n$. To that end, a similar approach to Algorithm 2 is applicable where random samples of Θ are generated by a multivariate normal distribution via equations (2.12) instead of computationally heavy FFBS approach. Algorithm 3 formalizes this approach. Since the recursive FFBS algorithm in the standard approach is replaced by a simple multivariate random generation process in our proposed approach, Algorithm 3 is significantly more efficient. Note that recursive equations (2.6) and (2.7) used to update the posterior after simulating a dummy future observation $\tilde{y}_{zm}$, are replaced with simple equations (2.12). Similar to Algorithm 2, Algorithm 3 must run for each patient. In particular, after dose z^* is assigned to a patient at time n, a true response $y_{z^*}^{n+1}$ is observed and is used to update μ^n and Σ^n for the next patient. Note that $\hat{m}$ and $\hat{\Sigma}$ are also temporary and should not be stored in memory.

2.5 Consistency of the Knowledge Gradient Policy

A measurement policy is "consistent" if it is able to learn the truth perfectly in the limit. In a response-adaptive dose-finding study, learning the true value of the target dose ED_{95} is achieved only if the true underlying dose-response relationship is known in the limit. In this section, we show that the knowledge gradient policy introduced in Section 2.4 learns the true proposed dose-response

Algorithm 3 Proposed knowledge gradient policy

for each dose $z \in \mathcal{Z}$ **do**
  Generate M samples of $\tilde{\Theta}$ from the prior $\mathcal{N}(\mu^{n-1}, \Sigma^{n-1})$.
  for each sampled $\tilde{\Theta}_{(m=1:M)}$ **do**
    Simulate future observation $\tilde{y}_{zm} \sim \mathcal{N}(\tilde{\theta}_{zm}, \sigma^2)$.
    Using $\tilde{y}_{zm}$, update $(\mu^{n-1}, \Sigma^{n-1})$ according to formulation (2.12) to obtain $(\hat{\mu}_{zm}^n, \hat{\Sigma}_{zm}^n)$.
    Generate T posterior samples of $\hat{\Theta}$ by sampling from $\mathcal{N}(\hat{\mu}_{zm}^n, \hat{\Sigma}_{zm}^n)$.
    for each sampled $\hat{\Theta}_{(t=1:T)}$ **do**
      Find $g(\hat{\Theta}_{(t)}) = \min_z\{z \in \mathcal{Z} : f(z, \hat{\Theta}_{(t)}) \geq 0.95 f(z_{max}, \hat{\Theta}_{(t)})\}$
    Estimate the observed variance $U_{zm} = \mathrm{Var}[g(\hat{\Theta})|\mathscr{F}^n \cup (z, \tilde{y}_{zm})]$ using sample variance.
  Evaluate expected variance U_z for each dose by taking a Monte Carlo sample average $\sum_m \frac{U_{zm}}{M}$.
Select the dose z^* that minimizes U_z.

model, thus the target dose, when the number of patients goes to infinity.

Consistency of some sampling policies has been studied in the literature. For example, reinforcement learning algorithms which force the measuring policy to explore all alternatives infinitely many times are consistent (Singh *et al.* 2000). However, in most response-adaptive designs, it is difficult to ensure that all alternatives are sampled frequently often to prove consistency. Frazier *et al.* (2008, 2009) derived the consistency conditions for a class of knowledge gradient policies, with independent and correlated normal prior beliefs in ranking and selection problems. Ryzhov *et al.* (2012) showed that the KG policy in a Gaussian multi-armed bandit problem finds the best alternative in the limit with probability one when the discount factor approaches one. Furthermore, Frazier & Powell (2011) provided a more general set of sufficient conditions for consistency of a broad class of sequential sampling policies. However, these methods do not directly apply because the objective function in our problem, which is minimizing the variance of the target dose, differs from typical ranking and selection objectives where one seeks to find the alternative with the highest mean.

In this section, we assume a multivariate normal prior on Θ with independent components, i.e., Σ^0 is diagonal with σ^2_{0j} elements. Therefore, the posterior is also normal with independent components. Sampling dose j may provide valuable information about the dose-response relationship through a reduction in uncertainty about the mean response in θ_j. In our case, this information (the reduction in uncertainty) is observable by measuring the reduction in variance of the target dose when dose j is sampled. Therefore, in order to know the dose-response curve perfectly, this uncertainty should approach zero in the limit for every dose, and thus a consistent policy should measure each dose infinitely often. However, it is possible that a measuring policy sticks to a set of doses for which the dose-response curve is already known perfectly thus providing no valuable information if sampled again. To avoid sticking to such doses, it is sufficient for a measurement policy to maintain an open neighborhood U, i.e., an open ball with arbitrary small radius $r_u > 0$, around states for which sampling particular doses has no value. The open neighborhood ensures that, when in such states, the measurement policy cannot encounter an infinite sequence of states for which measuring a dose would result in its variance to converge to zero and thus to stick.

Denote the mean and variance of θ_j at decision epoch n with $\hat{\mu}_{nj}$ and σ^2_{nj}, respectively. Using minimal-rank and mean-value parametrization of the exponential family distributions, the posterior distribution at time n can be described completely by $k^n = [\hat{\mu}_{nj}\lambda, -\lambda(\hat{\mu}^2_{nj} + \sigma^2_{nj})/2]$ where $\lambda = \frac{1}{\sigma^2}$ is the precision of $(\hat{y}^{n+1}|\Theta, z^n)$ distribution (Bickel & Doksum 2015). Define set $\mathcal{K}$ and its

closure as

$$\begin{aligned} \mathcal{K} &:= \Big\{ [u_j\lambda,\, -\lambda(u_j^2+v_j)/2]_{j\in\{1,\dots,J\}} : u\in\mathbb{R}^J, v\in\mathbb{R}^J_{++} \Big\}, \\ \mathrm{cl}(\mathcal{K}) &= \Big\{ [u_j\lambda,\, -\lambda(u_j^2+v_j)/2]_{j\in\{1,\dots,J\}} : u\in\mathbb{R}^J, v\in\mathbb{R}^J_{+} \Big\}, \end{aligned} \tag{2.16}$$

where $u=(u_1,\dots,u_J)$, $v=(v_1,\dots,v_J)$, and $\mathbb{R}_{++}$ denotes the set of strictly positive real numbers. Hereafter, the term "knowledge state" is used to refer to $k\in\mathcal{K}$, or $k\in\mathrm{cl}(\mathcal{K})$. Retrieving the mean and variance of θ_j from any knowledge state $k\in\mathrm{cl}(\mathcal{K})$ is done by functions $\hat{\mu}_j:\mathrm{cl}(\mathcal{K})\to\mathbb{R}$, and $\sigma_j^2:\mathrm{cl}(\mathcal{K})\to\mathbb{R}_+$ as follows

$$\hat{\mu}_j(k)=\frac{k_{j1}}{\lambda}, \qquad\qquad \sigma_j^2(k)=-\big(\frac{k_{j1}}{\lambda}\big)^2-2\frac{k_{j2}}{\lambda},$$

where $\hat{\mu}_{nj}=\hat{\mu}_j(k^n)$, and $\sigma_{nj}^2=\sigma_j^2(k^n)$ (see Frazier & Powell 2011).

To quantify the value of information for the KG policy, define $\mathrm{v}_z^{KG}(k)$ as a function that measures the incremental reduction in variance of the target dose after sampling dose z in knowledge state $k\in\mathrm{cl}(\mathcal{K})$, i.e.,

$$\mathrm{v}_z^{KG}(k)=\begin{cases} \mathbb{E}_n\Big[\mathrm{Var}_{n+1}\big(g(\Theta)\big)\Big|k^n=k,z^n=z\Big]-\mathbb{E}_n\Big[\mathrm{Var}_n\big(g(\Theta)\big)\Big|k^n=k,z^n=z\Big] & \text{if } \sigma_z^2(k)>0,\\ 0 & \text{if } \sigma_z^2(k)=0, \end{cases} \tag{2.17}$$

where $\mathbb{E}_n\big[\,\cdot\,\big]$ hereafter denotes the expectation with respect to probability measure ϕ^n on Θ determined by $k^n\in\mathrm{cl}(\mathcal{K})$ under which θ_j is distributed according to a normal distribution whose moments are given by functions $\hat{\mu}_j(k^n)$ and $\sigma_j^2(k^n)$ if $\sigma_j^2(k^n)>0$, or it is distributed according to $\hat{\mu}_j(k^n)$ almost surely if $\sigma_j^2(k^n)=0$. Note that $\mathbb{E}_n\Big[\mathrm{Var}_n\big(g(\Theta)\big)\Big|k^n=k,z^n=z\Big]$ is a constant at time n. Also, function $\mathrm{v}_z^{KG}(k)$ is well-defined for all $k\in\mathrm{cl}(\mathcal{K})$ because Jensen's inequality for convex functions and the tower property of conditional expectations imply that $\mathbb{E}_n\Big[\mathrm{Var}_{n+1}\big(g(\Theta)\big)\Big|k^n=k,z^n=z\Big]$ is bounded above by a constant term (see proof of Lemma 2.5.1), which implies $\mathrm{v}_z^{KG}(k)$ is non-positive. Therefore, $\mathrm{cl}\big(\mathrm{dom}(\mathrm{v}^{KG})\big)=\mathrm{cl}(\mathcal{K})$, where $\mathrm{dom}(\mathrm{v}^{KG})$ denotes the domain of function v^{KG}. Then, the KG policy defined in (2.15) can also be written in terms of the value of information, satisfying

$$\mathcal{X}^{\pi^{KG}}(k^n)\in\arg\min_z \mathrm{v}_z^{KG}(k^n). \tag{2.18}$$

Using the definition of value of information and its function v_z^{KG}, it is possible to partition the posterior states based on doses that have value or not. Define the partitioning sets M_z and M_*

as

$$
\begin{aligned}
M_z &:= \Big\{k \in \mathrm{cl}(\mathcal{K}) : \exists (k^n) \subseteq \mathrm{dom}(\mathrm{v}^{KG}) \text{ converging to } k \text{ with } \lim_{n\to\infty} \mathrm{v}_z^{KG}(k^n) = 0\Big\},\\
M_* &:= \Big\{k \in \mathrm{cl}(\mathcal{K}) : \forall (k^n) \subseteq \mathrm{dom}(\mathrm{v}^{KG}) \text{ converging to } k, \lim_{n\to\infty} \mathrm{v}_z^{KG}(k^n) = 0 \ \forall z \in \mathcal{Z}\Big\},
\end{aligned}
$$

where M_z is the set of knowledge states k^n for which sampling dose z does not have any informative value on the variance of the target dose at the end of the trial, i.e., sampling dose z does not reduce the variance of the target dose, and M_* is the set of knowledge states for which sampling any dose has no value. In addition to sets M_z and M_*, we partition the measurements $z \in \mathcal{Z}$ for each knowledge state $k \in \mathrm{cl}(\mathcal{K})$ according to measurements that have value or not. To that end, define $A_k := \{z \in \mathcal{Z} : k \in M_z\}$ for each $k \in \mathrm{cl}(\mathcal{K})$ to be the set of doses for which sampling provides no value in knowledge state k. As discussed earlier in this section, to guarantee consistency, the sampling policy should avoid measuring dose z when the trial is in state $k \in M_z$, and maintain an open neighborhood around the sets $M_z \setminus M_*$ in which measuring dose z has no value whereas other measurements do. The following lemma allows us to simplify the partitioning sets.

Lemma 2.5.1 *For each $z \in \mathcal{Z}$, $k \to \mathrm{v}_z^{KG}(k)$ is continuous on $\mathrm{dom}\big(\mathrm{v}^{KG}(k)\big)$ and can be extended continuously onto $\mathrm{cl}\big(\mathrm{dom}(\mathrm{v}^{KG})\big) = \mathrm{cl}(\mathcal{K})$.*

Proof. First, we show that for every $k \in \mathcal{K}$, $\mathbb{E}_n\Big[\mathrm{Var}_{n+1}\big(g(\Theta)\big)\Big|k^n = k, z^n = z\Big] \geq 0$ is finite and therefore, $\mathrm{dom}(\mathrm{v}^{KG}) = \mathrm{cl}(\mathcal{K})$. Using the tower property of conditional expectation and Jensen's inequality for convex functions,

$$
\begin{aligned}
\mathbb{E}_n\Big[\mathrm{Var}_{n+1}\big(g(\Theta)\big)\Big|k^n = k, z^n = z\Big] &= \mathbb{E}_n\Big\{\mathbb{E}_{n+1}\Big[g^2(\Theta)\Big] - \Big(\mathbb{E}_{n+1}\big[g(\Theta)\big]\Big)^2\Big|k^n = k, z^n = z\Big\}\\
&= \mathbb{E}_n\Big\{\mathbb{E}_{n+1}\Big[g^2(\Theta)\Big]\Big|\cdot\Big\} - \mathbb{E}_n\Big\{\Big(\mathbb{E}_{n+1}\big[g(\Theta)\big]\Big)^2\Big|\cdot\Big\}\\
&\leq \mathbb{E}_n\Big\{\Big[g^2(\Theta)\Big]\Big|\cdot\Big\} - \Big(\mathbb{E}_n\Big\{g(\Theta)\Big|\cdot\Big\}\Big)^2\\
&= \mathrm{Var}_n\Big[g(\Theta)\Big|k^n = k, z^n = z\Big],
\end{aligned}
$$

where $\mathrm{Var}_n\Big[g(\Theta)\Big|\cdot\Big]$ is finite for every $k \in \mathcal{K}$ because $g(\Theta)$ has finite support. Next, we show that $\mathrm{v}^{KG}(k)$ is continuous on its domain. Note that in formulation (2.17), $\mathbb{E}_n\Big[\mathrm{Var}_n\big(g(\Theta)\big)\Big|k^n = k, z^n = z\Big]$ is constant at time n. Therefore, it suffices to show continuity of $\mathbb{E}_n\Big[\mathrm{Var}_{n+1}\big(g(\Theta)\big)\Big|k^n = k, z^n = z\Big]$ for every $k \in \mathrm{cl}(\mathcal{K})$. Define $h(\Theta, k^n) = \mathrm{Var}_{n+1}\big[g(\Theta)\big|k^n\big]$ (for ease of notation we drop

$z^n = z$ in $\mathrm{Var}_{n+1}\big[g(\Theta)\big|k^n, z^n = z\big]$). The function $\mathrm{Var}_{n+1}[\cdot]$ is a random variable at time n, thus $\Theta \to h(\Theta, k^n)$ is measurable by definition for each n and $k^n \in \mathrm{cl}(\mathcal{K})$. Let $(k_*^n)_{n=1}^{\infty} \subseteq \mathrm{cl}(\mathcal{K})$ be a sequence converging to $k^* \in \mathrm{cl}(\mathcal{K})$ almost surely. Therefore,

$$\begin{aligned}
\lim_{n\to\infty} h(\Theta, k_*^n) &= \lim_{n\to\infty} \mathrm{Var}_{n+1}\big[g(\Theta)\big|k_*^n\big] \\
&= \lim_{n\to\infty} \Big\{ \mathbb{E}_{n+1}\big[g^2(\Theta)\big|k_*^n\big] - \Big(\mathbb{E}_{n+1}\big[g(\Theta)\big|k_*^n\big]\Big)^2 \Big\} \\
&= \lim_{n\to\infty} \mathbb{E}_{n+1}\big[g^2(\Theta)\big|k_*^n\big] - \lim_{n\to\infty} \Big(\mathbb{E}_{n+1}\big[g(\Theta)\big|k_*^n\big]\Big)^2 \\
&= \lim_{n\to\infty} \int g^2(\Theta) d\phi^n - \lim_{n\to\infty} \Big(\int g(\Theta) d\phi^n \Big)^2,
\end{aligned}$$

where ϕ^n denotes the probability measure on Θ given k_*^n. Since the probability measure ϕ^n is a continuous function of converging sequence k_n^*, continuous mapping theorem implies that ϕ^n converges weakly to ϕ^* when $\lim_{n\to\infty} k_*^n = k^*$. Note also that $g^2(\Theta)$ and $g(\Theta)$ are bounded, measurable and almost everywhere continuous functions. That is, the set where $g(\Theta)$ and $g^2(\Theta)$ are not continuous lies in $\mathbb{R}^{J-1}$ and thus has a measure of zero. Therefore, by applying continuous mapping theorem and Portmanteau's theorem (see, e.g., Billingsley 2013)

$$\begin{aligned}
\lim_{n\to\infty} \int g^2(\Theta) d\phi^n - \lim_{n\to\infty} \Big(\int g(\Theta) d\phi^n \Big)^2 &= \int g^2(\Theta) d\phi^* - \Big(\int g(\Theta) d\phi^* \Big)^2 \\
&= \mathbb{E}_{n+1}\big[g^2(\Theta)\big|k^*\big] - \Big(\mathbb{E}_{n+1}\big[g(\Theta)\big|k^*\big]\Big)^2 \\
&= \mathrm{Var}_{n+1}\big[g(\Theta)\big|k^*\big] = h(\Theta, k^*),
\end{aligned}$$

which shows that function h is almost everywhere continuous on its domain. Note that $g(\Theta) \in \mathcal{Z}$. Then by applying Popoviciu's inequality we have $h(\Theta, k^n) \leq \frac{1}{4}(J-1)^2 < \infty$. Therefore, $H(k^n) = \int_{\mathbb{R}^J} h(\Theta, k^n) d\Gamma(\Theta)$ exists by dominated convergence theorem, where $\Gamma(\Theta)$ is the desired probability measure, and thus

$$\begin{aligned}
H(k^n) = \int_{\mathbb{R}^J} h(\Theta, k^n) d\Gamma(\Theta) &= \mathbb{E}_n\Big[h(\Theta, k^n)\Big|k^n = k, z^n = z\Big] \\
&= \mathbb{E}_n\Big[\mathrm{Var}_{n+1}\big(g(\Theta)\big)\Big|k^n = k, z^n = z\Big].
\end{aligned}$$

Using continuity of $h(\Theta, k^n)$ through Lebesgue's dominated convergence theorem results in

$$\lim_{n\to\infty} \int_{\mathbb{R}^J} h(\Theta, k_*^n) d\Gamma(\Theta) = \int_{\mathbb{R}^J} \lim_{n\to\infty} h(\Theta, k_*^n) d\Gamma(\Theta) = \int_{\mathbb{R}^J} h(\Theta, k^*) d\Gamma(\Theta).$$

Therefore, $H(k^n) = \mathbb{E}_n\Big[\mathrm{Var}_{n+1}\big(g(\Theta)\big)\Big|k^n = k, z^n = z\Big]$ is continuous. Furthermore, the set $\mathrm{cl}(\mathcal{K}) = \Big\{\big[u_j\lambda,\, -\lambda(u_j^2 + v_j)/2\big]_{j\in\{1,\ldots,J\}}, u_j \in \mathbb{R}^J, v_j \in \mathbb{R}^J_+\Big\}$ has closure points such that $\sigma_j^2(k^n) = v_j = 0$. Note that if $\sigma^2_{z^n}(k^n) = 0$, then $k^{n+1} = k^n$. Thus,

$$\mathrm{v}^{KG}(k^n) = \mathbb{E}_n\Big[\mathrm{Var}_{n+1}\big(g(\Theta)\big)\Big|k^n = k, z^n = z\Big] - \mathbb{E}_n\Big[\mathrm{Var}_n\big(g(\Theta)\big)\Big|k^n = k, z^n = z\Big] = 0,$$

which is exactly equal to the value of information when $\sigma^2_{z^n} = 0$. Therefore, v^{KG} is continuous on $\mathrm{cl}\big(\mathrm{dom}(\mathrm{v}^{KG})\big) = \mathrm{cl}(\mathcal{K})$.□

Using Lemma 2.5.1, we can rewrite the sets M_z and M_* as

$$\begin{aligned} M_z &:= \Big\{k \in \mathrm{cl}\big(\mathrm{dom}(\mathrm{v}^{KG})\big) : \mathrm{v}_z^{KG}(k) = 0\Big\}, \\ M_* &:= \Big\{k \in \mathrm{cl}\big(\mathrm{dom}(\mathrm{v}^{KG})\big) : \mathrm{v}_z^{KG}(k) = 0\ \forall z \in \mathcal{Z}\Big\}, \end{aligned}$$

which by definition in formulation (2.17) can be further simplified to

$$\begin{aligned} M_z &:= \Big\{k \in \mathrm{cl}(\mathcal{K}) : \sigma_z^2(k) = 0\Big\}, \\ M_* &:= \Big\{k \in \mathrm{cl}(\mathcal{K}) : \sigma_z^2(k) = 0\ \forall z \in \mathcal{Z}\Big\}. \end{aligned}$$

The doses and posterior knowledge states are partitioned according to which doses have informative value about the variance of the target dose. However, we still need to show that the KG policy avoids sampling z when in states $M_z \setminus M_*$. To that end, define the set U for any $k \in \mathrm{cl}(\mathcal{K}) \setminus M_*$ as

$$U := \Big\{k' \in \mathrm{cl}(\mathcal{K}) : \min_{z\in A_k} \mathrm{v}_z^{KG}(k') > \max_{z\notin A_k} \mathrm{v}_z^{KG}(k')\Big\}. \tag{2.19}$$

If $A_k = \emptyset$, then $U = \mathrm{cl}(\mathcal{K})$. Note that minimum over empty set is $+\infty$ (and maximum is $-\infty$). Thus, there exist an open set in U such that the KG policy chooses an alternative which is not in A_k almost surely. Now, suppose that $A_k \neq \emptyset$.

Lemma 2.5.2 *If $A_k \neq \emptyset$, then U is open and $k \in U$ for any $k \in cl(\mathcal{K}) \setminus M_*$.*

Proof. Assuming $A_k \neq \emptyset$, we first show that $k \in U$. Consider any $k \in \mathrm{cl}(\mathcal{K}) \setminus M_*$ such that $\sigma_z^2(k) = 0$ for dose z. Therefore, sampling dose z does not change the posterior on Θ and $\mathbb{E}_n\Big[\mathrm{Var}_{n+1}\big(g(\Theta)\big)\Big|k^n =$

$k, z^n = z\Big] = \mathbb{E}_n\Big[\mathrm{Var}_n\big(g(\Theta)\big)\Big|k^n = k, z^n = z\Big]$, which implies that $\mathrm{v}_z^{KG}(k) = 0$. Thus,

$$\min_{z \in A_k} \mathrm{v}_z^{KG}(k) = 0. \tag{2.20}$$

Now, consider any $k \in \mathrm{cl}(\mathcal{K}) \setminus M_*$ such that $\sigma_z^2(k) > 0$. Using the tower property of conditional expectation and Jensen's inequality for convex function (see proof of Lemma 2.5.1),

$$\mathbb{E}_n\Big[\mathrm{Var}_{n+1}\big(g(\Theta)\big)\Big|k^n = k, z^n = z\Big] < \mathrm{Var}_n\Big[g(\Theta)\Big|k^n = k, z^n = z\Big] = \mathbb{E}_n\Big[\mathrm{Var}_n\big(g(\Theta)\big)\Big|k^n = k, z^n = z\Big],$$

where inequality is strict since $\Big[\mathbb{E}_{n+1}\big(g(\Theta)\big)\Big]^2$ is strictly convex almost surely if $\sigma_z^2(k) > 0$. Thus, $\mathrm{v}_z^{KG}(k) < 0$, and

$$\max_{z \notin A_k} \mathrm{v}_z^{KG}(k) < 0. \tag{2.21}$$

Therefore, equations (2.20) and (2.21) result in

$$\min_{z \in A_k} \mathrm{v}_z^{KG}(k) > \max_{z \notin A_k} \mathrm{v}_z^{KG}(k),$$

which implies that $k \in U$. Continuity of $k \to \mathrm{v}_z^{KG}(k)$, shown in Lemma 2.5.1, suffices to show that U is open. To that end, consider an open set $\tilde{K} \subset \mathrm{cl}(\mathcal{K})$ such that $\big(\mathrm{v}_z^{KG}\big)^{-1}(\tilde{K}) = U$. Since v_z^{KG} is continuous, the inverse image of any open set under v_z^{KG} is an open set, thus U is open.□

Therefore, for any $k \in U$, $\min_{z \in A_k} \mathrm{v}_z^{KG}(k') > \max_{z \notin A_k} \mathrm{v}_z^{KG}(k')$ where $k' \in \mathrm{cl}(\mathcal{K})$. Thus, the KG policy in knowledge state k, avoids sampling doses in set A_k for which there is no uncertainty about the belief because there are always doses such as $z \notin A_k$ where $\mathrm{v}_z^{KG}(k)$ shows better values. Note that any knowledge state in $k \in \mathrm{cl}(\mathcal{K}) \setminus M_*$ satisfies the inequality in (2.19). Therefore, KG policy does not stick to a set of doses and measures every dose infinitely many times and thus is consistent.

2.6 Numerical Analysis

This section presents the results of implementing our proposed algorithm and compares it to the standard approach used in, e.g., Berry *et al.* (2002), the ASTIN study by Krams *et al.* (2003), Smith *et al.* (2006), Müller *et al.* (2006), and Weir *et al.* (2007). We implement both

adaptive policies on three different types of dose-response curves, i.e., bell-shaped, sigmoid, and non-monotonic piecewise linear functions, and show the patient assignment patterns for each curve, as well as the fitted posterior dose-response curves. The true dose-response curves used in these experiments resemble a wide range of practical dose-response relationships. For example, bell-shaped and non-monotonic dose-response curves are studied in, e.g., Owen *et al.* (2014) and Bulayeva & Watson (2004), respectively, while sigmoid shaped curves are among the most occurring dose-response relationship studied (Gadagkar & Call 2015).

2.6.1 Simulation Initialization

In our experiments, we consider 11 doses which are placed equidistant. Typically this number in Phase II trials is between 4-12 doses (Berry *et al.* 2002). Aligned with the ASTIN trial, the first dose is considered as placebo with its response marking the baseline score for the treatment in the trial. Performance of the state-of-the-art and the proposed approaches are compared by developing three measurement policies within the one-step look-ahead framework. We use "NDLM" to refer to the measurement policy of the state-of-the-art approach (see Algorithm 2). "KG-I" and "KG-C" denote the policies developed for the proposed approach (see Algorithm 3) where KG-I denotes the policies in which no correlation is considered on prior beliefs about Θ whereas "KG-C" assumes an exponential covariance function on prior beliefs about Θ.

A simulation is carried out in order to compare the performance of these policies. At each decision epoch a patient arrives at the trial and is given a dose according to the latest posterior estimate of the dose-response curve which minimizes the variance of ED_{95}. The patient's response is then generated from the true distribution and is added to the data. To estimate the posterior dose-response curve, M and T sample sizes (parameters of the algorithms) are set to 500 and 1000 in both algorithms, and sequences of 60 and 200 patients are used in reporting the results. A thinning factor of 5 is considered for random variable generation in both Algorithms where every fifth randomly generated number was used in the simulation to avoid serial correlation in a sequence of random numbers. In reporting each performance measure, 30 simulations with different sequence of random numbers are considered. We assume that the variance of normal residual of the dose-response function is known in both approaches and its value is fixed at one unit. We conduct a sensitivity analysis on this variance, and show that our approach is robust with respect to variation in normal residual. The simulation is coded in R and is run on an Intel core i7 3.7 GHz processor

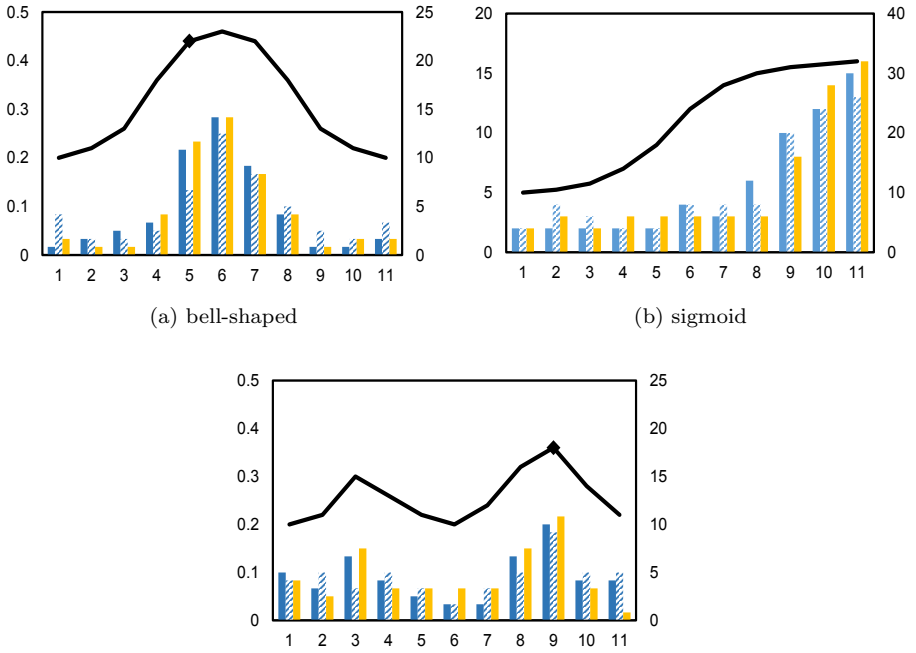
with 16 GB of RAM.

In case of "NDLM", m_1 and C_1 diagonals in $\mathcal{N}(m_1, C_1)$ are set to placebo and 100, respectively. This is to ensure that the prior carries little information about the beliefs on Θ. The observation variance matrix $\mathcal{E}_j$ is set to have diagonal values equal to the normal residual. No covariance structure is considered since observations are assumed to be independent of each other. Aligned with Weir *et al.* (2007) and West & Harrison (1997), the evolution variance matrix $\mathcal{W}_j$ is set to have diagonal values equal to a discounted structure $\frac{C_j(1-\gamma')}{\gamma'}$, where $\gamma' \in [0, 1]$. The discounted structure provides stability for the system and allows for information decay when moving from dose $j-1$ to dose j. We set $\gamma' = 0.6$ in our experiment, which produces the fastest convergence in the standard approach. Our sensitivity analysis show that the quality of solutions produced by the standard approach is highly sensitive to the choice of γ'.

In case of "KG-I", μ_0 and Σ_0 are set equal to m_1 and C_1 in the "NDLM" case. However, for "KG-C" policy, where correlation is considered about the beliefs on Θ, $\text{Cov}(\theta_i, \theta_j)$ is calculated by a Gaussian covariance function where $\text{Cov}(\theta_i, \theta_j) = \beta \exp\{-\gamma(i-j)^2\}$ where β is usually estimated by $\text{Var}(\theta_i)$ (Rasmussen & Williams 2006). The Gaussian structure of the covariance function allows for less correlation when doses are further apart. To keep symmetry of the covariance matrix, β is chosen to be equal to $\frac{\text{Var}(\theta_i) + \text{Var}(\theta_j)}{2} = 100$, and γ, the lengthscale factor is set to 0.01 for the sigmoid curve and one for the bell-shaped and piecewise linear curves where smaller values of the lenghtscale factor correspond to smoother changes between θs.

2.6.2 Results

Patient assignment. Figure 2.5 shows the patient assignment pattern to three dose-response curves for 60 patients. In particular, Figures 2.5(a), 2.5(b), and 2.5(c) show the results of patient assignments for bell-shaped, sigmoid and piecewise linear curves, respectively. Note that the horizontal axis shows the dose indices, the left-hand side vertical axis denotes the proportion of patients assigned to a dose while the solid line represents the dose-response curve, with score on the right-hand side vertical axis. Our results show that KG-I and KG-C outperform the NDLM approach in assigning more patients to the target dose; see Table 2.1. For example, KG-I and KG-C assign 21.1±0.3 (95% confidence interval) and 20.8±0.4 patients to the target dose on the bell-shaped curve, respectively, while NDLM assigns 20.1±0.3 patients to the target dose.



(a) bell-shaped

(b) sigmoid

(c) piecewise linear

Figure 2.5: Patient assignments to dose-response curves (sample size=60)
Note: The diamond on the dose-response identifies the ED_{95} dose.

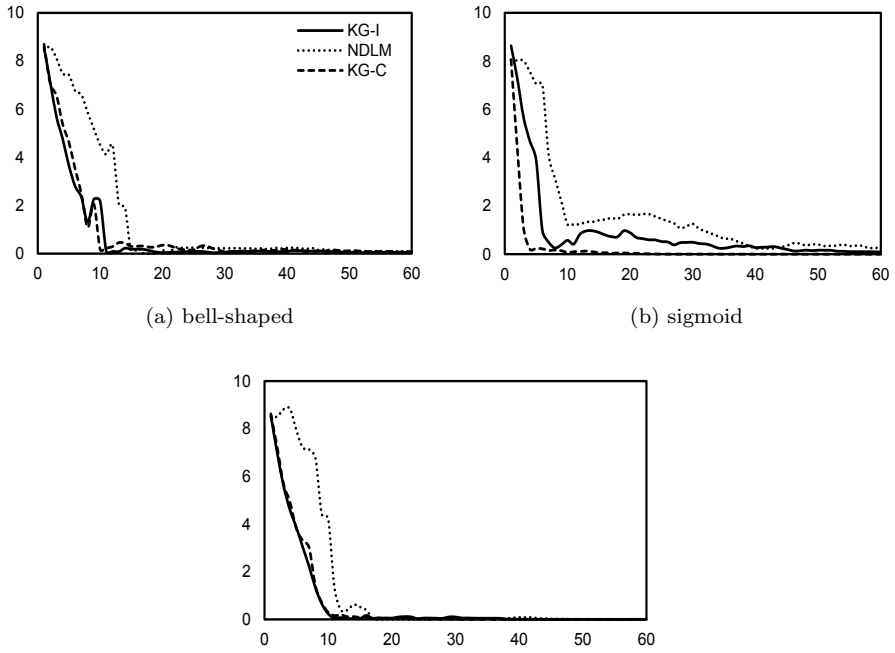
Table 2.1: Patient assignments to target dose (sample size=60)

	bell-shaped	sigmoid	piecewise linear
KG-I (%)	21.1±0.3	13.0±0.4	23.1±0.3
KG-C (%)	20.8±0.4	20.2±0.3	22.5±0.4
NDLM (%)	20.1±0.3	18.2±0.3	21.5±0.3

Note: The performance is reported in terms of proportion of patients assigned to the target dose out of 60 total patients in 95% confidence interval.

Posterior variance of the target dose. Figures 2.6(a), 2.6(b), and 2.6(c) show how the expected variance of ED_{95} for dose z^*, i.e, $\min_z \mathbb{E}_n\Big\{\text{Var}_{n+1}\big[g(\theta)\big|\mathscr{F}^n, z)\big]\Big\}$, changes under KG-I, KG-C, and NDLM policies during the trial for bell-shaped, sigmoid, and piecewise linear dose-response models, respectively. As can be seen, KG-I and KG-C policies achieve lower variances of the target dose for the same number of patients earlier in the trial, thus they learn the target dose more quickly than the NDLM policy. For example, in case of sigmoid curves, after simulating 15 patients on average, the expected variance under KG-C policy drops below 0.1 level, the KG-I policy achieves similar precision after 50 patients while NDLM policy never drops below 0.1 with 60 patients.

Posterior estimate of the dose-response curve. Figure 2.7 shows the estimated posterior dose-



(a) bell-shaped

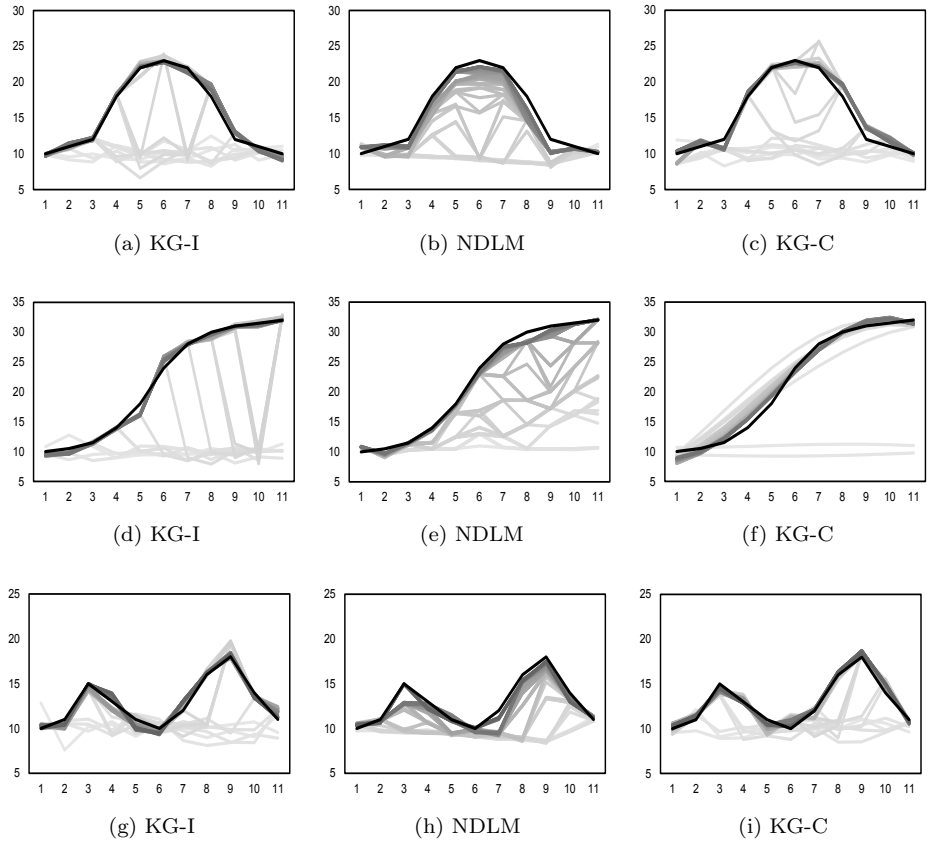
(b) sigmoid

(c) piecewise linear

Figure 2.6: Expected variance of ED_{95} for dose z^* (sample size=60)
Note: The horizontal axis shows the patients number and the vertical axis denotes the variance of the target dose.

response curves of KG-I, KG-C and NDLM policies with respect to the same three dose-response curves used in Figure 2.5 for 60 patients. In particular, Figures 2.7(a), 2.7(b), and 2.7(c) show the estimated posterior dose-responses of 60 patients for bell-shaped dose-response curve under KG-I, KG-C, and NDLM policies. Posterior dose-response estimates of sigmoid and piecewise linear models are shown in Figures 2.7(d)-2.7(i). Note that the vertical axis denotes the response score of a dose while the horizontal axis shows the dose indices. The solid black line represents the true dose-response curve whereas the gray lines show the estimated posterior fitted dose-responses with darker lines representing later fits in the simulation. As more patients are simulated in the trial, gray lines representing posterior estimates become darker and move closer to the true dose-response curve.

L^2 distance between estimated posteriors and the true dose-response curve. Figures 2.7(a)-2.7(i) show that both KG-I and KG-C are able to learn the dose-response curve at least as well as the NDLM policy. However, they do not demonstrate which algorithm achieves learning the true dose-response curve more efficiently. To differentiate the learning processes, Figure 2.8



(a) KG-I (b) NDLM (c) KG-C

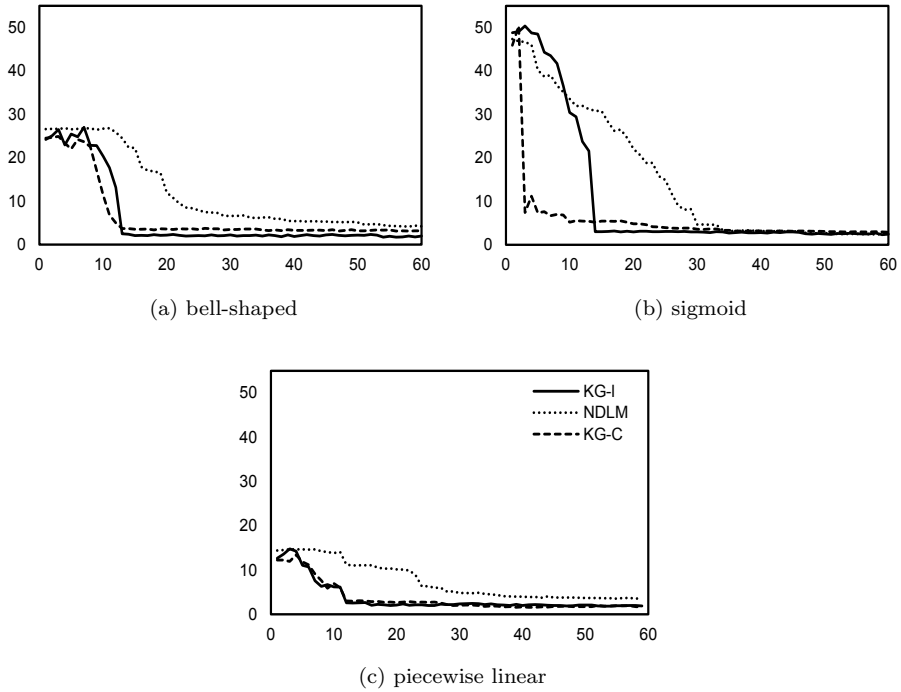
(d) KG-I (e) NDLM (f) KG-C

(g) KG-I (h) NDLM (i) KG-C

Figure 2.7: Posterior estimates to dose-response curves (sample size=60)

shows the L^2 distance between the posterior estimate means and the true dose-response model, i.e., $\left\|\mathbb{E}_n(\Theta) - \Theta^*\right\|$, where Θ^* represents the true value of Θ. The results suggest that both KG-I and KG-C polices outperform the NDLM policy in reducing the L^2 distance between the posterior estimate means to the dose-response curves and their true values, thus learning the true dose-response curve faster. Moreover, Figure 2.8(b) shows that KG-C reduces the L^2 distance considerably faster than KG-I in the sigmoid curve where correlation between doses and their responses is stronger. Notice that similar behavior is observable in Figure 2.6(b).

As mentioned earlier, one major impediment in applying the standard framework to design of dose-finding trials is its difficulty to implement and heavy computational effort requirement. Table 2.2 shows that proposed policies are significantly more efficient because they do not require the time consuming recursive FFBS algorithm to estimate the posterior dose-response curve. Note that the



(a) bell-shaped

(b) sigmoid

(c) piecewise linear

Figure 2.8: L^2 distance (sample size=60)
Note: The horizontal axis shows the patients number and the vertical axis denotes the L^2 distance

results in Table 2.2 are reported for simulating a single decision epoch along 30 sample paths, and the performance of NDLM policy is enhanced, compared to the original standard method, by parallelizing independent "for" loops in Algorithm 2.

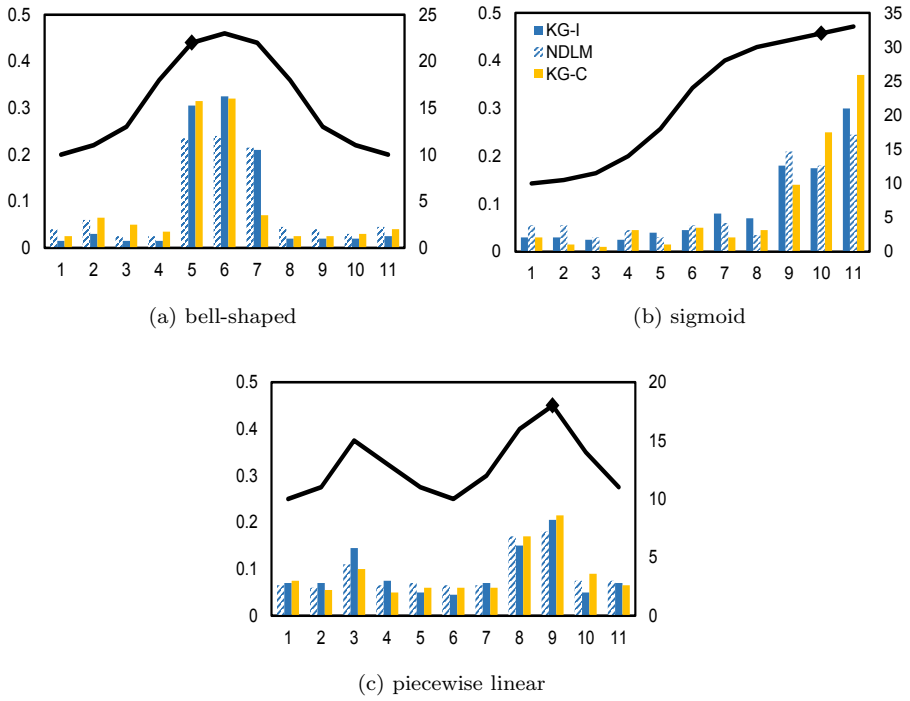
2.6.3 Sensitivity Analysis

Assignment pattern for 200 patients. Figure 2.9 shows the assignment pattern for bell-shaped, sigmoid, and piecewise linear models when 200 patients are simulated under KG-I, KG-C and NDLM policies. Our results show that KG-I, KG-C assign more patients to the target dose than the NDLM policy. Note that we omitted figures for posterior estimates and variance of the target dose because

Table 2.2: Computational time for a single decision (in hours)

	bell-shaped	sigmoid	piecewise linear
KG-I	0.16	0.17	0.18
KG-C	0.16	0.17	0.18
NDLM	5.63	5.73	6.18

simulating more than 60 patients does not affect the results in such a way to be shown clearly in figures similar to Figures 2.6 and 2.7.



(a) bell-shaped

(b) sigmoid

(c) piecewise linear

Figure 2.9: Patient assignments to dose-response curves (sample size=200)
Note: The horizontal axis shows the dose indices, the left-hand side vertical axis denotes the proportion of assignment, and the right-hand side vertical axis denotes the dose-response score. The diamond on the dose-response identifies the target dose.

Sensitivity to the Variance of the normal residual. In Section 2.2, the dose-response model is presented as

$$y = f(z, \theta) + \epsilon,$$

where ϵ is normally distributed by mean zero and variance σ^2, i.e., $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We assumed that σ^2 is known throughout this chapter and is fixed at one in reporting the results. Here, we conduct a sensitivity analysis on the variance of the normal residual and show that our proposed policies, similar to the "NDLM" policy, is robust with respect to variation of the normal error. The following results are reported for $\sigma^2 = 4$. Our results indicate that the proposed approaches, KG-I and KG-C are robust with respect to the variation in normal error and outperform the NDLM policy when the variance of the normal residual increases. However, it also assigns more patients to larger doses

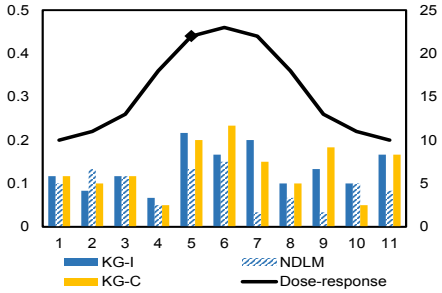


Figure 2.10: Sensitivity in terms of patient assignments to observation variance
Note: The horizontal axis shows the dose indices, the left-hand side vertical axis denotes the proportion of assignments, and the right-hand side vertical axis denotes the dose-response score. The diamond on the dose-response identifies the target dose.

on one side of the bell-shaped curve. For an instance of 60 patients, Figure 2.10 shows the patient assignment pattern for the bell-shaped curve where our proposed policies assign more patients to the target dose than the NDLM policy. Figure 2.11 demonstrates that the expected variance of ED_{95} for dose z^* reduces faster under KG-I and KG-C policies. Figure 2.12 shows that the L^2 distance between the posterior estimate means and the corresponding true values reduces faster under KG-I and KG-C policies, thus learning the true dose-response model is achieved more quickly. Figure 2.13 shows posterior estimates to the dose-response curve.

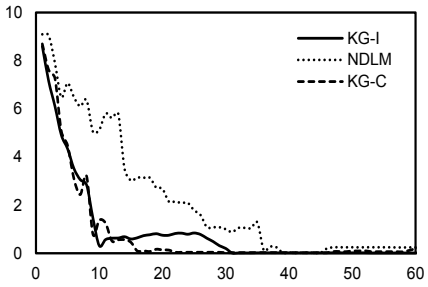


Figure 2.11: Sensitivity in terms of posterior variance of ED_{95} for dose z^* to observation variance
Note: The horizontal axis shows the patients number and the vertical axis denotes the variance of the target dose.

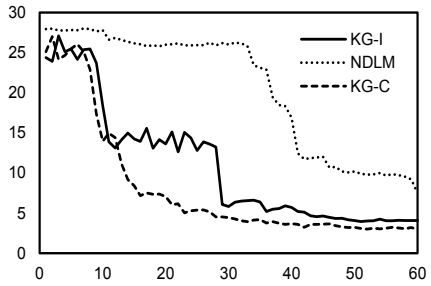
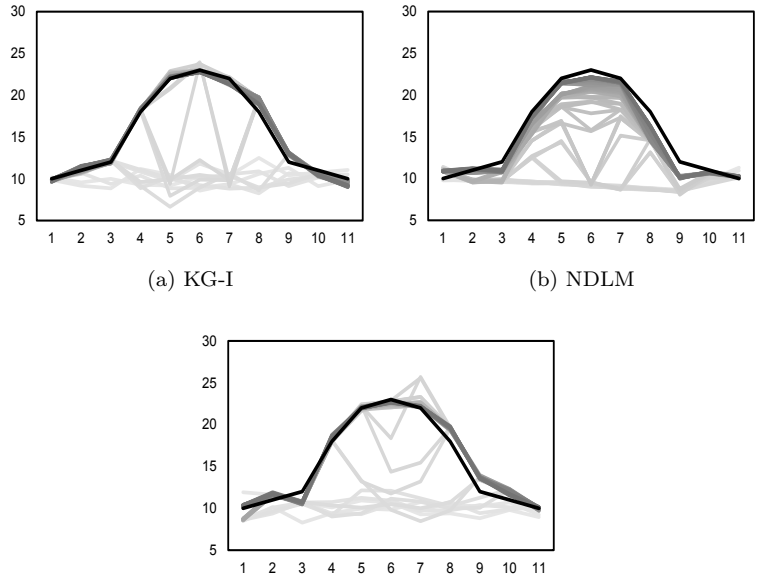


Figure 2.12: Sensitivity in terms of L^2 distance to observation variance
Note: The horizontal axis shows the patients number and the vertical axis denotes the L^2 distance between posterior estimates and the true model.

Sensitivity of the standard approach to the variance of evolution equations. The evolution equations presented in Section 2.3 in equation (2.5) show that

$$\alpha_j = \begin{pmatrix} \theta_j \\ \delta_j \end{pmatrix} = G_j \alpha_{j-1} + \Omega_j,$$



(a) KG-I

(b) NDLM

(c) KG-C

Figure 2.13: Sensitivity in terms of posterior estimates to observation variance
Notes: The horizontal axis shows the dose indices and the vertical axis denotes the dose-response score.

where $G_j = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ and Ω_j is distributed according to $\mathcal{N}(0, \mathcal{W}_j)$ where $\mathcal{W}_j$ is a known $(L_j^n \times L_j^n)$ matrix. In Section 3.5, we set the diagonal values of $\mathcal{W}_j$ equal to $\frac{C_j(1-\gamma')}{\gamma'}$ where γ' was set to 0.6. However, our results show that the NDLM policy is highly sensitive to the choice of γ'. Figure 2.14 compares the patients assignment to the target dose under NDLM policy when γ' changes by ± 0.1 for the bell-shaped curve. In particular, Figure 2.14 shows that when γ' changes by $+0.1$, the NDLM policy can not approximate the dose-response correctly (see Figure 2.15(c)), thus is unable to identify the target dose. Figure 2.15 shows the posterior estimates to the dose-response curve,

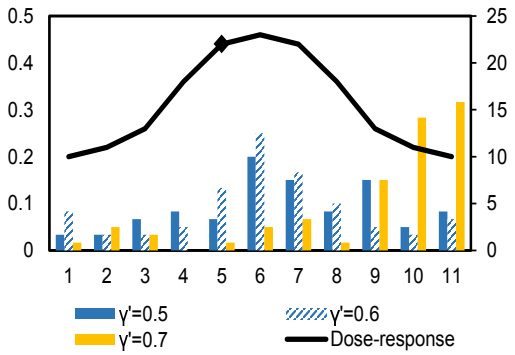
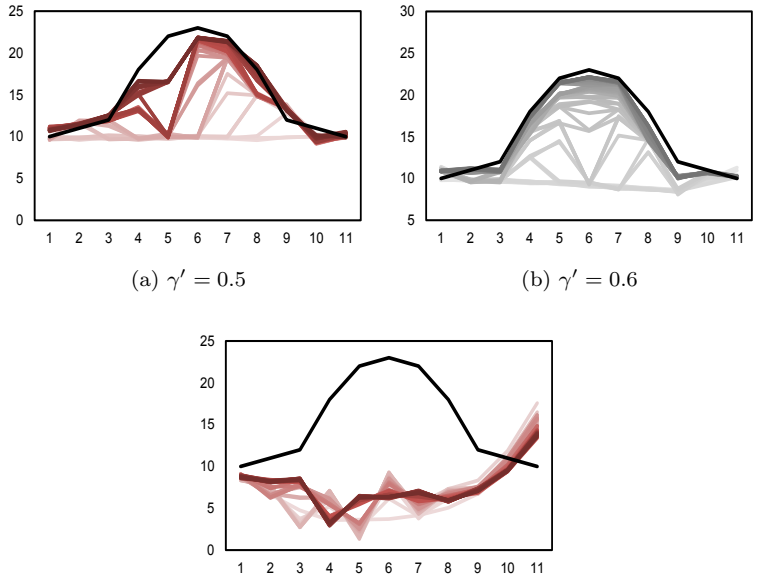


Figure 2.14: Sensitivity in terms of patient assignment to evolution variance

and Figure 2.16 demonstrates the L^2 distance between the estimated posterior means and the true value of bell-shaped dose-response curve. Notice that, slightest changes in γ' contribute to inefficient learning of the dose-response curve apparent in Figures 2.15 and 2.16, especially when γ' is changed by +0.1. These observations suggest that the one-step look-ahead policy in the standard approach may not be consistent. Heavy computational effort required by the NDLM policy reported in Table 2.2 increases this inefficiency in fine tuning the model parameters and implementing the design in practice.



(a) $\gamma' = 0.5$ (b) $\gamma' = 0.6$

(c) $\gamma' = 0.7$

Figure 2.15: Sensitivity in terms of posterior estimates to evolution variance
Note: The horizontal axis shows the dose indices and the vertical axis denotes the dose-response score.

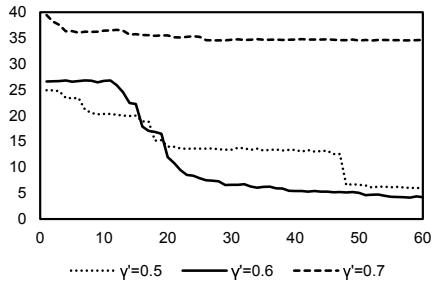


Figure 2.16: Sensitivity in terms of L^2 distance to evolution variance
Note: The horizontal axis shows the patients number and the vertical axis denotes the L^2 distance between posterior estimates and the true model.

2.7 Conclusion

In this chapter, we developed a novel framework to identify the target dose in response-adaptive Phase-II clinical trials, derived analytical insights regarding the learning process, and proposed a knowledge gradient policy to solve it. We also showed that our knowledge gradient policy is consistent in that it learns the dose-response model perfectly, and thus the target dose, if the number of patients tends to infinity. In contrast, we showed that the state-ot-the-art approach may fail to identify the target dose. To assess the quality of our solutions, we presented a dynamic programming formulation of the state-of-the-art approach (Section 2.3), which has been implemented in real clinical trials (e.g., Krams *et al.* 2003), and compared it with our proposed approach (Section 2.4). To that end, we created a simulation study and tested the performance of both approaches on practical instances in literature. Our results show that the proposed policies outperform the state-of-the-art approach in terms of solution quality and time efficiency. In particular, our results show that our policies assign the right dose to more patients than the standard approach. In addition, our policies learn the target dose faster as they reach smaller expected variances of ED_{95} with fewer number of patients. Finally, our knowledge gradient policies are far easier to implement because they reduce the complexity of the problem, and they require significantly less computational power and time. In particular, our policies are significantly more efficient than the standard approach in terms of runtime in an environment where computational burden of a design is a major challenge for implementation in practice. Therefore, our proposed approach may have a significant impact on how dose-finding studies will be conducted.

Chapter 3

Optimal Stopping of Adaptive Dose-Finding Clinical Trials

Summary. The ultimate goal in a dose-finding clinical trial is to identify the target dose such that its efficacy and adverse effects be tested afterwards in a confirmatory phase for a large population. Adaptive designs enable decision makers to terminate or abandon the trial early on because enough evidence is gathered for efficacy or futility, which can help reduce the costs among other significant benefits. The optimal stopping formulation for this problem has unique features because the target dose is not fixed at each decision epoch and its advantage over placebo is random. Therefore, consequences of such uncertainty in the next confirmatory trials should be considered in decision making. We implement a standard method developed for this problem and propose two methods: one based on a one-step look-ahead policy and the other based on a continuous version of the problem and approximating transitions by an Itô process. Our results reveal that if there is not a significant advantage over placebo in the true dose-response curve, the standard method has a low probability of selecting the right decision, which may have significant adverse consequences. In contrast, our method based on the continuous version produces high quality solutions. Motivated by these results, we propose a constraint on accuracy of dose-response curve estimation before deciding for stopping.

Manuscript: Nasrollahzadeh, A., Khademi, A., "Optimal Stopping of Adaptive Dose-Finding Clinical Trials", Under review at Manufacturing & Service Operations Management

3.1 Introduction

As mentioned in Section 3.1, dose-finding trials contribute significantly to the cost of a clinical trial. Therefore, authorities and pharmaceutical companies are motivated to reduce cost by designing a more efficient process where efficacy or futility decisions can be made as soon as evidence allows. Chapter 2 also identifies the main goal of dose-finding clinical trials to be searching for the "target dose," a critical step in the drug development process (Bornkamp *et al.* 2007). This is because a poor selection of the target dose may cause the Food and Drug Administration (FDA) to disapprove the next phase (Phase III), which is the most costly phase in drug development, due to futility (insignificant positive evidence) or adverse effects (exposure to unnecessary risk)(Snapinn *et al.* 2006). In particular, during 2000-2012, failure to select optimal drug doses was a leading factor for delay or denial of drug submissions in the first submission round by the FDA (Sacks *et al.* 2014). Moreover, the European Medicines Agency stresses the importance of rigorous/scientific dose finding by relying on model-based estimation, rather than hypothesis testing with pairwise comparisons (Mullard 2015).

Adaptive designs of dose-finding clinical trials generally can reduce the cost of conducting a clinical trial, in addition to having other benefits such as changing the patient randomization decisions to avoid allocating large samples to doses that are not beneficial, and thus decreasing the overall length of the trial. For example, in a fully sequential design, at each decision epoch, the decision maker can terminate the trial because there is already sufficient evidence that the target dose is efficacious or the decision maker can abandon the trial because there is not enough evidence that the drug is effective. Therefore, adaptive designs can significantly reduce the length and sample size of a trial which are key factors in increasing the overall costs (Berry *et al.* 2002), and thus optimal stopping of a clinical trial for efficacy or futility has a natural motivation in adaptive clinical trials.

This chapter studies the optimal stopping of an adaptive dose-finding clinical trial. There is a major difference between optimal stopping of a dose-finding clinical trial with that of a Phase III trial, which is the confirmatory phase. This is because if the decision is to terminate a dose-finding trial for efficacy, the next step is to run an extremely expensive Phase III trial for regularity approval and eventual marketing. Therefore, termination decisions in dose-finding trials should consider the probability that the confirmatory phase is successful and estimate the profit/loss upon success or failure.

We formulate the optimal stopping of an adaptive dose-finding trial as a finite-horizon stochastic dynamic program (SDP), where at each intermediate decision epoch the decision maker may abandon the trial for futility, continue the trial to collect more evidence about the dose-response curve, or terminate the trial for efficacy and move to the confirmatory phase. A key feature in this problem is that upon termination the decision maker has to consider the probability of success in the next clinical trial. Before further discussion on this key feature, note that the main goal of a dose-finding trial is to identify the target dose for which efficacy should be tested versus a standard treatment or placebo by a large patient population in Phase III. Therefore, in order to estimate the probability of success in Phase III, it is natural for the decision maker to consider the power of the hypothesis test $H_0 : df^* \leq 0$ versus $H_1 : df^* > 0$, where df^* denotes the expected response improvement over placebo or standard treatment (assuming higher response is more favorable) (Müller *et al.* 2006).

However, two main challenges are involved in the definition of df^* in a Bayesian setting: (i) The target dose is a random variable at the beginning of each decision epoch given the history of the states, actions, and observations up to said period; (ii) The expected response of any dose (including the target dose) is also a random variable at the beginning of each decision period given said history. Addressing such challenges requires a proper dose-response model and a SDP setup, which are discussed in Sections 3.2.1 and 3.3, respectively. The resulting SDP formulation, however, suffers heavily from the curse of dimensionality because the state space of the formulation is multidimensional and unbounded.

For this problem, Brockwell & Kadane (2003) proposed an approximation procedure, which was partially applied to optimal stopping of a fully Bayesian dose-finding trial (Berry *et al.* 2002). The approximation is based on discretizing the state space by a grid, using forward simulation until the last decision epoch to create sample paths, and using backward induction to estimate the value function in each cell of the grid at each time period. This method is computationally extremely time-consuming and Berry *et al.* (2002) stated that applying this method at each decision epoch in a fully adaptive design is "impractical." We implement this method in a fully adaptive setting by parallelization of different sample paths as a benchmark, in terms of solution quality and computational time, to our methods.

We propose two solution methods for this problem. The first one adapts the one-step look-ahead framework, in which the decision maker assumes that the next decision epoch is the terminal

time. This approach is computationally much less demanding than the benchmark method because it only requires one-step forward simulations. However, the induced stopping time by this method may happen earlier than the optimal stopping time (Proposition 3.4.1). In the second proposed method, we consider a two-armed bandit version of the problem with one unknown arm (the target dose) and one known arm (placebo), where the posterior of df^*, i.e, the advantage of target dose over placebo, happens to be normally distributed in our setup. Therefore, in a continuous sampling regime, a scaled mean of df^* follows an Itô process, which enables us to formulate a continuous-time Bellman equation for the continuous-time optimal stopping counterpart. By using Itô's lemma, we show that the optimal value function to the continuous-time Bellman equation satisfies a partial differential diffusion-advection equation with boundary conditions (Proposition 3.4.2). In addition, the solution to the partial differential equation depends only on the utility function (objective function of the decision maker) which can be found upfront, and identifies a continuation region over a mean response (vertical axis) and time (horizontal axis) coordinates, which is easy to understand and implement. This method is also computationally appealing because it bypasses forward simulations to find the optimal decision regions. Finally, we develop a heuristic to extend the results of one unknown arm setting to multiple unknown arms, which has to address the challenge that the target dose is a random variable and the observations may not belong to the true target dose.

We test the performance of the two proposed methods along the available benchmark via simulation. In addition to monetary value of a stopping decision, which is the primary objective function, we report the probability of correct decision at stopping time for each method. We test the results on two setting: one where there is a significant difference between the average response of the target dose and placebo (the ultimate decision should be termination), and one where said difference is negligible (the ultimate decision is abandonment). Our simulation results shed light on behavior and performance of each method.

3.2 Background

Optimal stopping is an important decision making problem and is studied in different communities because of its vast applications. For classical references on optimal stopping problems see Chow *et al.* (1971) and Peskir & Shiryaev (2006). The optimal stopping of a clinical trial has also received significant attention due to its importance. For an overview of advancements in optimal

stopping of clinical trials see a survey by Hee *et al.* (2016) from a Bayesian perspective and Jennison & Turnbull (1999), O'Brien & Fleming (1979), and Whitehead (1997) from a frequentist perspective, and references therein. Also, Stallard *et al.* (2001) reviewed different stopping rules for Phase II clinical trials. Here, we only focus on Bayesian decision theoretic methods developed for optimal stopping of dose-finding trials. One main method used in Bayesian decision theoretic designs is based on forward simulation of the trial to the end and using backward induction to estimate the value function over a grid to ultimately evaluate the stopping region. Details of this methodology is presented in Brockwell & Kadane (2003) and it is used in Berry *et al.* (2002), a fully adaptive trial, and is also adapted to binary outcomes in Jiang *et al.* (2013). Note that these authors use "constrained backward induction" terminology but we choose to use backward induction to ease exposition. However, this method is computationally demanding and we propose two more efficient algorithms to this problem which our simulation results confirm their competitive and superior performances depending on the settings.

Our first proposal is to adapt the one-step look-ahead approach to optimal stopping of dose-finding trials. This approach is used for sequential sampling in Bayesian settings by Gupta & Miescke (1996) and Chick & Inoue (2001). Frazier *et al.* (2008) applied this method for optimal stopping of a ranking and selection problem. Also, Branke *et al.* (2007) proposed two stopping rules that stop experimenting when: i) the probability of good selection exceeds a target, or ii) the expected opportunity cost exceeds a target in a ranking and selection setting. However, the optimal stopping of a dose-finding trial is different from a ranking and selection setup because the decision maker has to consider the effects of termination decision on the next confirmatory phase of the drug development process. This consideration is handled by a hypothesis test in which the significance of the advantage of the target dose over placebo, calculated by subtracting the placebo response from that of the (random) target dose, is tested. Therefore, application of this method to our problem deems its own analysis.

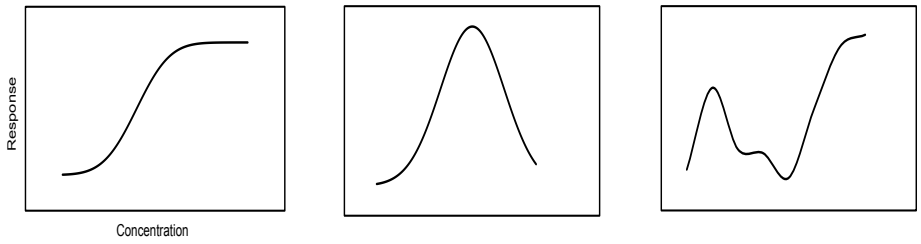
Our second proposed approach is inspired by a work presented in Chernoff (1961), where a diffusion approximation is used to test whether the mean of a normal distribution is positive or negative. Such a method is used for optimal learning of patient response types (Negoescu *et al.* 2017), local time method for targeting and selection (Ryzhov 2018), hiring and retention policies of workers (Arlotto *et al.* 2013), discounted economic analysis in sampling selection problem (Chick & Gans 2009), undiscounted economic analysis in sampling selection problem (Chick & Frazier 2012),

an application of the later in optimal stopping of a Phase III clinical trial (Chick *et al.* 2017), and in optimal stopping of a clinical trial with correlated treatments (Chick *et al.* 2018). However, the structure of our problem is different from previous studies because the decision maker has to consider the power of a hypothesis test and thus a need for a separate analysis. Moreover, the heuristic to extend the diffusion results to multiple doses setting is different from the heuristic that are proposed in the literature and is tailored to our setting; see Section 3.4.3 for details.

3.2.1 Dose-Response Model

The relationship between treatment dose of a drug and its induced response, e.g., change in a measurable medical outcome, is essential in dose-finding studies and is usually described by a curve or function referred to as a dose-response curve. To identify this relationship, we use the same model as in Section 3.2.1. For example, Figure 3.1 presents three typical dose-response curves where the sigmoid shape in Figure 3.1(a) is one of the most recurring dose-response relationships in theory and practice (Gadagkar & Call 2015). The target dose is also defined similar to Section 3.2.1 in equation (3.2), i.e., ED_{95}. Note that the dynamic formulation and the approximate solutions presented in this chapter are flexible enough to accommodate other definitions of target dose such as minimum effective dose (MED), or maximum tolerable dose (MTD) (see Chow 2003). The motivation for ED_{95} is that the highest response may correspond to high dosages (toxic doses) which may induce unwanted adverse side effects. However, there are several ways to incorporate toxicity. For example, one might assume that a safe dosage range is approved in Phase I of the trial, or model efficacy and toxicity jointly; see e.g., Zhang *et al.* (2006). In our setup, one could define a ratio of efficacy/toxicity as a function of $f(z, \Theta)$ instead of ED_{95} to consider limiting exposure to toxic levels of the drug (Berry *et al.* 2002).

There are two main classic approaches to estimate the dose-response curve and, in particular, $f(z, \Theta)$. The first approach considers a functional form for $f(z, \Theta)$ upfront and seeks to estimate the parameters Θ by using observations and prescribes decisions based on that (e.g., Kotas & Ghate 2018). For example, one may consider an $\text{E}_{\max}$ model as $f(z, \Theta) = \theta_0 + \frac{\theta_1 z}{\theta_2 + z}$, where $(\theta_0, \theta_1, \theta_2)$ are unknown parameters to be estimated. However, one major issue with this approach is susceptibility to model misspecification because information arising from observations might reveal that the true dose-response curve is bell-shaped and thus the pre-identified $\text{E}_{\max}$ model was



(a) Felker (2012) (b) Owen *et al.* (2014) (c) Bulayeva & Watson (2004)

Figure 3.1: Typical dose-response curves

misspecified. In addition, because we consider a normal distribution on error, thus the response, a Bayesian setup suffers from nonconjugacy, which poses significant computational challenges as one has to use time-consuming Markov chain Monte Carlo approaches to generate samples from the posterior distribution. Also, the state space of a SDP involves the set of all probability distributions on the unknown parameter, making the analysis for decision making extremely challenging since the parameters' marginal distributions might not be known. The second approach to estimating the dose-response curve is to use piecewise linear approximations to the curve, which addresses the challenges discussed above. In particular, because this method approximates the curve at each dose, it does not assume a functional form for the response upfront and it is therefore less likely to experience model misspecification error. Moreover, as we next discuss, a proper choice for curve approximation will result in conjugacy, which significantly reduces computational efforts and simplifies analysis.

In this chapter, similar to Section 3.4, we assume that at any given dose j, the response follows a model such that $y_j = \theta_j + \epsilon$, i.e., $f(Z_j, \Theta) = \theta_j$. By this first-order construction, the response of patients at dose j is normally distributed with unknown mean θ_j and known variance

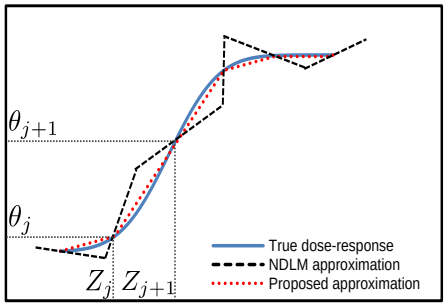


Figure 3.2: Two dose-response approximations

σ^2. We can interpret that such construction approximates the true dose-response curve by fitting a piecewise linear function connecting consecutive θ_js. Figure 3.2 shows a true response curve along with the piecewise linear approximation induced by the proposed model. More complicated models can also be considered to approximate the dose-response curve. For example, in Figure 3.2, a second order normal dynamic linear model (NDLM) is also used to approximate the true dose-response curve, by which a patient's response for doses close enough to dose $z \in \mathcal{Z}$ is estimated by fitting a straight line using the expected response at dose z and a slope at that dose as model parameters. This is the approach proposed by Berry *et al.* (2002) and applied in dose-finding trials by Krams *et al.* (2003), Warner *et al.* (2015), Lenz *et al.* (2015) and Liu *et al.* (2017); see Section 3.3. The NDLM model is flexible in approximating any dose-response curve and allows for a linear correlation structure with random deviations. Although our proposed one-step look-ahead policy can be adapted to more complicated dose-response models such as NDLM, the diffusion approximation method relies on the proposed first-order model and its extension to more complicated models is not clear since df^*, the advantage over placebo, will not follow an Itô process any longer. Moreover, it is shown in Chapter 2 that the first-order approximation model is competitive to the NDLM in representing the true dose-response curve when adapted in an optimal learning setting. In addition, the one-step look-ahead policy for the first order construction, with the objective of minimizing the variance of the target dose, is consistent as the number of patients grow to infinity. There is also another method known as Gaussian process regression which approximates the dose-response curve in a continuous fashion, where new measurements are not limited to a predetermined set of doses; see Powell & Ryzhov (2012). However, since the FDA only accepts the dosages for which patient responses are available, a discrete approach in approximating the dose-response curve is sufficient.

In this chapter, we only investigate the optimal stopping problem and fix the allocation policy and the dose-response approximation model to those proposed in Section 3.4. Therefore, we consider a Bayesian setup where the decision maker has a multivariate normal belief about the expected response Θ and updates her belief upon each observation of patient's response.

3.3 Problem Formulation

In this section, we present a stochastic dynamic programming formulation for the response-adaptive optimal stopping of dose-finding clinical trials. At each decision epoch, based on the

information accrued so far, an investigator decides whether to (i) abandon the trial due to lack of significant positive evidence about effectiveness of the treatment, (ii) continue the trial to collect more information if there is significant positive evidence about the effectiveness of the treatment with expectation of improvement, or (iii) terminate the trial for efficiency and move to a confirmatory study such as Phase III when effectiveness is verified by testing the treatment for a large population. The optimal stopping problem is coupled with an adaptive patient allocation decision, i.e., upon continuation, the decision maker has to allocate a dose to the next patient in the trial. The objective of the decision maker in patient allocation problem is to identify the target dose with higher accuracy, i.e., minimization of variance of the target dose, which is called D-optimal design in statistical literature (Berry *et al.* 2002). However, the objective of optimal stopping in this context is expressed in terms of monetary value as will be discussed later in this section. This objective, which is called net present value (NPV) in finance literature, is appropriate in stopping problems where sampling costs and rewards are financial measures themselves (Brealey *et al.* 2012). Therefore, in this chapter, we assume that the allocation decisions, upon continuation of the trial, are based on variance minimization objective and is independent of the optimal stopping problem. In particular, we follow a one-step look-ahead policy in allocating patients to treatments to minimize the variance of the target dose, proposed in Chapter 2.

Recall that Θ represents a vector of unknown expected responses corresponding to doses in set $\mathcal{Z}$ where the resulting dose-response function is formalized by $f(z, \Theta) = \theta_z$. Let n denote decision epochs, N be the total number of (potential) homogeneous patients in the trial, and $y^{n+1} = \theta_{z^n} + \epsilon^{n+1}$ be the observed response of patient $n+1$ after assignment to dose z^n where $(y^{n+1}|\Theta, z^n) \sim \mathcal{N}(\theta_{z^n}, \sigma^2)$. We assume that the response of a patient is observed before the next decision epoch. Define $\mathscr{F}^n$ as the sigma-algebra generated by $z^0, y^1, z^1, y^2, \ldots, z^{n-1}, y^n$. Note that z^0 is the assignment dose before observing any response and τ represents some stopping time at which $\mathscr{F}^\tau$ describes the accrued information gathered by sampling τ patients. We use y and $\hat{y}$ to denote true and simulated observations, respectively.

State space. Decision epochs are set at times when response of a patient is observed. We assume a possibly correlated multivariate normal prior on our belief about Θ, i.e., $\Theta \sim \mathcal{N}(\mu^0, \Sigma^0)$. Recall that said construction approximates the dose-response curve by a piecewise linear function using estimates θ_j at each dose Z_j. Observations y form a normal likelihood distribution resulting in a Bayesian conjugate setup where posterior distributions on Θ are also multivariate normal. Define

$\mu^n := \mathbb{E}[\Theta|\mathscr{F}^n]$, and $\Sigma^n := \text{Cov}[\Theta|\mathscr{F}^n]$ as posterior moments of the belief about Θ. At decision epoch n in the trial, an investigator decides about abandoning, continuing, or terminating only based on the current estimate of the dose-response curve, which is summarized by the posterior normal probability distribution on parameter Θ given historical information $\mathscr{F}^n$, i.e., $\mathbb{P}(\Theta|\mathscr{F}^n)$. This posterior can be completely described by state variable $s^n = (\mu^n, \Sigma^n)$. Thus, the state space $\mathcal{S}$ is defined as

$$s^n \in \mathcal{S} := \Big\{(\mu, \Sigma) : \mu \in \mathbb{R}^J, \Sigma \in \Psi\Big\} \cup \nabla,$$

where Ψ denotes the set of $J \times J$ positive semidefinite matrices, and ∇ denotes an absorbing state showing the end of the decision making process.

Action space. At each decision epoch, if enough evidence (in the form of current estimate of the dose-response) has emerged to suggest that an effective target dose is identified, and sampling more patients will not improve the estimate by a significant margin considering the cost of sampling, the investigator may decide to "terminate" the trial and switch to a confirmatory phase where the target dose is further tested to confirm its effectiveness. On the contrary, the decision maker might learn that the current estimate of the dose-response curve shows no signs of effectiveness, e.g., a flat dose-response curve, and sampling more patients will only increase trial costs, and thus the investigator may "abandon" the trial. However, if the current estimate of the dose-response curve suggests that an effective target dose may be identified and continuing the trial with more sampling potentially may lead to a significant improvement of the estimate and utility, then the investigator may "continue" the trial by allocating a dose to the next patient, observe the response, and update the current estimate of the dose-response curve. Recall that the allocation scheme is assumed to be given and independent of optimal stopping problem. Define $a^n(s) \in \{0, 1, 2\}$ as the decision variable when in state s, where "0" shows that the decision is to abandon the trial, "1" shows the continuation of the trial, and "2" shows that the trial is terminated. Thus, the action space is described by

$$A(s) := \Big\{a^n(s) \in \{0, 1, 2\}, \forall n \leq N\Big\},$$

where at stopping time $n = \tau$, or the last decision epoch $n = N$, $a^n \in \{0, 2\}$. For $s = \nabla$, set $A(s) := \emptyset$.

Transitions. Terminating or abandoning the trial at decision epoch n determines the stopping time as $\tau = n$, and the dynamic system transits to state ∇ where no more sampling is allowed and

the current estimate of the dose-response curve remains unchanged. However, if the decision is to continue the trial, a dose is selected according to an allocation policy and its observed response will be used in order to update the current estimate of the dose-response curve, i.e., transit to a new state. The new state $s^{n+1} = (\mu^{n+1}, \Sigma^{n+1})$ is described by

$$\begin{aligned} \mu^{n+1} &= \mu^n + \tilde{\sigma}\big(\Sigma^n, j\big)X^{n+1}, \\ \Sigma^{n+1} &= \Sigma^n - \tilde{\sigma}\big(\Sigma^n, j\big)\tilde{\sigma}'\big(\Sigma^n, j\big), \end{aligned} \tag{3.1}$$

where $a^n(s) = j$ denotes the allocated dose, $\tilde{\sigma}(\Sigma^n, j) := \frac{\Sigma^n e_j}{\sqrt{(\sigma^2 + \Sigma^n_{jj})}}$, e_j is a J-vector of 0s and a single 1 at the j^{th} index, and $X^{n+1} := \frac{y^{n+1} - \mu^n}{\sqrt{(\sigma^2 + \Sigma^n_{jj})}}$ is a standard normal random variable when conditioned on $\mathscr{F}^n$.

Objective function. We consider maximizing monetary equivalent of benefits acquired due to early termination or abandonment of the trial versus costs incurred by continuing the trial with more sampling. If the decision is to abandon the trial, i.e., $a^n = 0$, then no immediate reward or cost is incurred. If the decision is to continue the trial, i.e., $a^n = 1$, then only a sampling cost $c_1 > 0$ is paid. In case of termination, i.e., $a^n = 2$, immediate reward consists of the monetary value of the advantage over placebo, if such an advantage is significant, minus the setup/sampling cost in the confirmatory phase. Define utility function $u(a^n, s^n, \mathscr{F}^n)$ as the expected immediate benefit (reward$-$cost) incurred when deciding on action a^n in state s^n given information $\mathscr{F}^n$ by

$$u(a^n, s^n, \mathscr{F}^n) = \begin{cases} 0 & \text{if } a^n = 0, \\ -c_1 & \text{if } a^n = 1, \\ -c_1' n_p + c_2 m_n \mathbb{E}[\mathbb{1}_{\{B^n\}} | \mathscr{F}^n] & \text{if } a^n = 2, \end{cases} \tag{3.2}$$

where $c_1' n_p$ is the cost of sampling n_p patients in the confirmatory phase ($c_1' > 0$), and $c_2 > 0$ is the payoff per unit advantage of the current estimate of the target dose over placebo. Note that $m_n = \mathbb{E}[df^* | \mathscr{F}^n]$ denotes the expected advantage over placebo where $df^* = \theta_{z^*} - \theta_0$, z^* being the (random) target dose, and θ_0 is the known and fixed response of placebo. Since z^* is random with respect to $\mathscr{F}^n$, θ_{z^*} denotes the posterior expected response at dose z^*, and thus df^* identifies the posterior advantage over placebo. Furthermore, The indicator function $\mathbb{1}_{\{B^n\}}$ determines the significance of the advantage over placebo by considering the event B^n in which the null hypothesis

is rejected when comparing $H_0 : df^* \leq 0$ versus $H_1 : df^* > 0$. In particular,

$$B^n := \left\{ \frac{\sqrt{n_p}(\bar{y}_* - \bar{y}_0)}{\sqrt{2\sigma^2}} > q_\alpha \right\}, \tag{3.3}$$

where $\bar{y}_*$ and $\bar{y}_0$ denote n_p-sample average responses of the estimated target dose and placebo at decision epoch n, and q_α denotes $(1-\alpha)$ quantile of normal distribution with α being the significance level for the hypothesis. The expectation $\mathbb{E}[\mathbb{1}_{\{B^n\}}|\mathscr{F}^n]$ can be estimated with an arbitrary accuracy by Monte Carlo as follows: Create a sample from Θ and calculate the target dose, z^*, for said sample by equation (3.2); Create n_p samples from $\mathcal{N}(\theta_{z^*}, \sigma^2)$ and $\mathcal{N}(\theta_0, \sigma^2)$; Calculate $\bar{y}_*$ and $\bar{y}_0$ and identify whether the event B^n occurs; Continue this process for enough samples and take a sample average to estimate $\mathbb{E}[\mathbb{1}_{\{B^n\}}|\mathscr{F}^n]$. Note that the utility function defined in this section is tailored to our specific problem. However, both of our proposed methods can handle various utility functions as long as they are measurable with respect to the defined filtration.

Given that a decision to abandon or terminate the trial has been made at stopping time $n = \tau$, the optimal expected utility is given by

$$G(s^\tau) = \max_{a^\tau \in \{0,2\}} u(a^\tau, s^\tau, \mathscr{F}^\tau) = \max\Big\{0, -c_1' n_p + c_2 m_\tau \mathbb{E}[\mathbb{1}_{\{B^\tau\}}|\mathscr{F}^\tau]\Big\}. \tag{3.4}$$

Therefore, for every $n < \tau$, the decision has to be $a^n = 1$ and a sampling cost c_1 is paid as the expected immediate utility, i.e., $u(a^n = 1, s^n, \mathscr{F}^n) = -c_1$. Let $l_\pi(s^0)$ denote the expected utility at stopping time τ given historical information $\mathscr{F}^\tau$ under policy π when the initial prior on the belief about Θ is $s^0 = (\mu^0, \Sigma^0)$; that is,

$$l_\pi(s^0) = \mathbb{E}^\pi\Big\{ - c_1\tau + \max_{\pi(a^\tau)\in\{0,2\}} u\big(\pi(a^\tau), s^\tau, \mathscr{F}^\tau\big)\Big| s^0\Big\}, \qquad \forall \pi \in \Pi, \tag{3.5}$$

where Π is the set of all non-anticipative admissible policies, and the investigator selects a policy $\pi \in \Pi$ such that $V(s^0) = \sup_{\pi\in\Pi} l_\pi(s^0)$. Therefore, the optimal value function is the solution to the following optimality equations

$$\begin{aligned} &V(s^n) = \mathbb{E}^\tau \sup_{\tau \geq n+1} \Big\{ - c_1(\tau - n) + \mathbb{E}\big[V(s^\tau)\big|\mathscr{F}^n\big]\Big\}, \\ &V(s^\tau) = G(s^\tau), \end{aligned} \qquad \forall s \in \mathcal{S}. \tag{3.6}$$

The state space defined on our belief about the dose-response curve Θ is unbounded, and thus standard SDP techniques are computationally intractable. We describe and implement the status-quo algorithm to approximate this problem and propose two different alternative techniques to solve it.

3.4 Approximate Solutions

In this section, we explain three approximate methods. In particular, Berry *et al.* (2002) used a simulation-based gridding algorithm discussed in Section 3.4.1 to evaluate stopping times. This approach is computationally extremely expensive in implementation and gives rise to "static terminator" where for a few sample dose-response curves, a large number of trials are simulated forward in time to compute their average expected utility over a discretized grid by backward induction. These approximations are used statically to evaluate stopping times for "similar" dose-response curves. Section 3.4.2 proposes a one-step look-ahead policy to find stopping times. At each decision epoch, this policy assumes that the next decision epoch is the last, and thus selects the decision with the maximum expected utility, eliminating simulation of a large number of trials to the end. Finally, Section 3.4.3 proposes a diffusion approximation method within which the Bellman equation resulting from optimality equations (3.6) is approximated by a diffusion-advection partial differential equation from which stopping boundaries are derived. Because this approach depends only on prior information and utility function, it finds the stopping region upfront and does not require any forward simulation or backward induction while being far more efficient.

3.4.1 Simulation-Based Gridding Approximation

A full solution to the response-adaptive optimal stopping problem in Section 3.3 requires backward induction where final stage of observation is evaluated first, and then earlier stages are computable using optimal values of later stages. Note that the final stage can be enumerated because it does not involve future expectation of utility. However, since the state space is unbounded, this method is computationally intractable. Brockwell & Kadane (2003) and Müller *et al.* (2007) proposed a method in which the state space is discretized by a grid and a sufficient number of experiments are run to estimate the final stage value function. The key idea is that in each cell of this grid, say cell j, the value of termination and abandonment can be evaluated easily. The value of continuation is

the sample average (Monte Carlo) of all the cells which are visited in the next decision epoch by the experiments currently visiting cell j. We present the details of the approach for completeness and to clarify the differences between the assumptions used in this approach with those in our setup.

To construct the grid, Berry *et al.* (2002) assumed a normal prior on the advantage over placebo, i.e., $df^* \sim \mathcal{N}(m_0, \nu_0^2)$. Let (m_n, ν_n) denote the posterior mean and standard deviation of the advantage over placebo at ED_{95} at time n, i.e., $m_n = \mathbb{E}[df^*|\mathscr{F}^n]$ and $\nu_n^2 = \text{Var}[df^*|\mathscr{F}^n]$. Construct a bivariate grid over possible values of m and ν carefully considering their upper and lower bounds as follows. Given the allocation scheme and thus the allocation dose z^n, simulate trials $i = 1, \ldots, M$ by generating observations $\hat{y}_i^{(n+1:N)}$, and update the current estimate of the mean for dose-response curves Θ by calculating $\mu_i^{(n+1:N)}$ and $\Sigma_i^{(n+1:N)}$. In order to estimate m_n and ν_n^2, at each decision epoch, after the current Θ is evaluated, simulate samples from Θ, identify the target dose for each sample, and calculate the posterior mean and standard deviation of df^* through sample mean and variance estimation. Record the trajectory of each trial, i.e., the sequence of (m_n, ν_n), over the bivariate grid for (m, ν). For example, Figure 3.3(a) shows 30 trial trajectories of m_n on a simplified univariate grid (only m_n versus n) for $N = 10$ patients. It might be the case that some of the grid cells remain empty, i.e., no simulated trials resulted in m and ν values corresponding to that cell, which affects the quality of the approximation. To fix that, consider a particular (m_n, ν_n) corresponding to those cells as priors and simulate a number of trials starting from those cells. Thus, the entire grid is populated.

Remark 3.4.1 *Note that we assumed a correlated multivariate normal prior on our belief about Θ, that is $\Theta \sim \mathcal{N}(\mu_0, \Sigma_0)$. However, as noted before, z^*=ED_{95} is random with respect to $\mathscr{F}^n$, and thus θ_{z^*} is not distributed normally with respect to $\mathscr{F}^n$. In the simulation-based gridding algorithm, the actual unknown distribution of θ_{z^*} is approximated by a normal distribution in the literature. However, we do not make such an assumption in the proposed one-step look-ahead policy and the diffusion approximation method in Sections 3.4.2 and 3.4.3.*

To evaluate the optimal decision in each cell, start from the last decision epoch N when the continuation decision is not available and the optimal value function can be computed by equation (3.4). Denote A_j^n as the subset of indices $i \in \{1, \ldots, M\}$ whose trajectories terminate in the jth cell (which corresponds to a (m, ν) pair) in the grid (m_n, ν_n, n). For the last decision epoch N, this is demonstrated by darker trajectories which end up in a specific cell in Figure 3.3(a). The

termination utility function in the jth cell is evaluated by taking a sample average of the value functions corresponding to trial simulations whose trajectories terminated in that grid cell, i.e.,

$$\hat{U}_j^N(a^N = 2) \approx \frac{1}{|A_j^N|} \sum_{i \in A_j^N} u_j^N(a^N = 2, \hat{s}_i^N), \tag{3.7}$$

where $\hat{U}_j^N(a^N = 2)$ is the approximated utility function at decision epoch N in the grid cell j when the decision is to terminate the trial, $|\cdot|$ denotes set cardinality, and the utility function $u_j^N(a^N, \hat{s}_i^N)$ is known for $a^N \in \{0, 2\}$ and $\hat{s}_i^N = (m_j^N, \nu_j^N)$ for all $i \in A_j^N$ where m_j^N and ν_j^N correspond to the jth cell values for m and ν, respectively. Therefore, the expected utility of termination at the last decision epoch N is given by

$$u_j^N(a^N = 2, \hat{s}_i^N) = -c_1' n_p + c_2 m_j^N \mathbb{E}[\mathbb{1}_{\{B^N\}} | \mathscr{F}^N] \qquad \forall i \in A_j^N,$$

where $B^N := \left\{ \frac{\sqrt{n_p}(\bar{y}_* - \bar{y}_0)}{\sqrt{(2\sigma^2 + (\nu_j^N)^2)}} > q_\alpha \right\}$ with $\sqrt{2\sigma^2} + (\nu_j^N)^2$ denoting the posterior predictive variance of $\bar{y}_* - \bar{y}_0$, and thus the approximated value function in each cell of the grid at decision epoch N is

$$\hat{V}_j^{*,N} = \max\left\{0, \hat{U}_j^N(a^N = 2)\right\}, \tag{3.8}$$

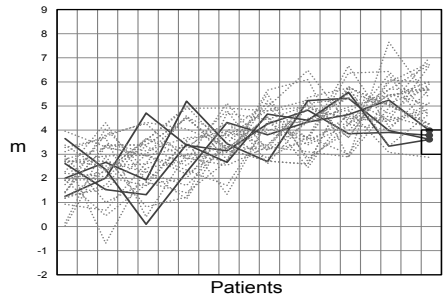
where if $\hat{V}_j^{*,N} = 0$, the optimal decision is to abandon the trial in the jth grid cell, i.e., $a_j^{*,N} = 0$. Otherwise, the optimal decision is to terminate the trial, $a_j^{*,N} = 2$. Working backwards, the utility function in the jth cell for $n < N$ when the decision is to continue the trial is given by the following sample average

$$\hat{U}_j^n(a^n = 1) \approx \frac{1}{|A_j^n|} \sum_{i \in A_j^n} \hat{V}_{j(i)}^{*,n+1}, \tag{3.9}$$

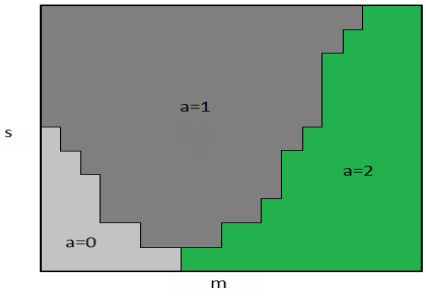
where $j(i)$ denotes a cell that trajectory i visits at decision epoch $n+1$. Therefore, the approximated value function in each cell of the grid at decision epoch $n < N$ is

$$\hat{V}_j^{*,n} = \max\left\{0, \hat{U}_j^n(a^n = 1), \hat{U}_j^n(a^n = 2)\right\}. \tag{3.10}$$

Enumerating the entire grid backwards until decision epoch n identifies the optimal decision and value function for each cell. Figure 3.3(b) shows a hypothetical example of optimal decisions on the grid (m, ν) at a particular decision epoch n. Algorithm 1 describes the gridding algorithm in



(a) An example of trajectories of simulated experiments on a grid over (m, n)



(b) Optimal decisions on a grid over (m, ν) at decision epoch n

Figure 3.3: An example of gridding approximation

an online fashion where after each observation the entire process is repeated to find the stopping region. At each period, dose allocation z^n is given and the state variable (μ^n, Σ^n) is known. Steps 3-14 describe the M forward simulations of the trial starting from decision epoch n to the end of trial N. In particular, for each $i \in \{1, \ldots, M\}$ and for any decision epoch $n \leq k \leq N$, dose z_i^k is determined by the fixed allocation policy. A future observation is simulated using the posterior predictive distribution. This observation is then used to update the state variable (μ_i^k, Σ_i^k). Using the updated state variable, a sample of T dose-response curves Θ are generated whereafter m_i^k and ν_i^k are evaluated using sample mean and sample variance, respectively. Steps 16-23 implement a similar procedure for the cells that are not populated already. Note that experiments $i \in \{1, \ldots, M\}$, or $i \in \{1, \ldots, M'\}$ are independent of each other and can run in parallel. We used the "*foreach*" package in the R programming language to parallelize the forward simulations and reduce run time. When the grid is fully populated, starting from step 25, we use backward induction to evaluate the optimal value function and thus the optimal decision for each cell in the grid. Starting from the last epoch N, the optimal value function is the maximum expected value of termination or abandonment for that cell since only these two decisions are available. The expected value of termination is easily evaluated because in each cell (m, ν) are known. Working backwards, in order to find the optimal value function and the optimal decision for each cell, we need to consider the expected value of continuation as well. This is achieved by tracking which experiments are visiting each grid cell in A_j^k at each decision epoch. Therefore, using equation (3.9), the utility of continuation in each cell j at decision epoch k is the average optimal value function of the cells at decision epoch $k+1$ which are visited by an experiment originating from cell j.

Algorithm 1 is run at each decision epoch n to evaluate the optimal value function and thus the optimal decision across the entire grid. Thereafter, at decision epoch n, true response y^n is observed, and using a similar procedure to steps $7-11$, (m_n, ν_n) are evaluated and the optimal decision is identified by finding the corresponding grid cell for (m_n, ν_n). Berry *et al.* (2002) applied this approach under a set of "typical dose-response curves" where the approximate value function for each grid cell was computed by taking the average of expected utilities under the same set of dose-response curves. Therefore, when a true observation from a dose-response curve investigated in the trial becomes available at decision epoch n, (m_n, ν_n) tuple is evaluated and depending on which grid cell it falls into, the optimal decision is identified. This approach may be problematic particularly when the unknown dose-response curve does not closely resemble those in the typical set. Furthermore, when the shape of the dose-response curve is unknown and a response-adaptive dynamic allocation scheme is trying to learn it, the resulting response-adaptive optimal stopping problem becomes computationally demanding since every time a true observation is received and the approximated dose-response curve is updated, a large number of forward simulations from decision epoch n to N are required to update the grid and evaluate the optimal decision for each cell. In fact, Berry *et al.* (2002) stated that this approach is "impractical" to be repeated when a new observation becomes available. In Sections 3.4.2 and 3.4.3, we propose alternative methods that are significantly more efficient and can be used in a fully sequential setting.

3.4.2 One-Step Look-Ahead Policy

Frazier *et al.* (2008) proposed a kind of one-step look-ahead policy (knowledge gradient) to optimal stopping of ranking and selection problems by assuming the experiment has to terminate at the next decision epoch. We adapt such a framework into the optimal stopping of a dose-finding trial with unique challenges. In particular, we consider three actions at each decision epoch, i.e, abandonment, continuation, and termination, whereas in most standard ranking and selection problems, only continuation and termination decisions are available. Furthermore, our utility function consists of $\mathbb{E}[\mathbb{1}_{\{B^n\}}|\mathscr{F}^n]$, which is emanated from evaluating the significance of the advantage over placebo via a hypothesis test, when the decision is to terminate the trial.

To quantify the value gained in continuing the trial, define $V^{\mathrm{KG}}_{a^n=1}(s)$ as a function that measures the difference between terminating or abandoning the trial at time n, and continuing the

Algorithm 4 Simulation-based Gridding Algorithm

Input. The allocation scheme z^n and state (μ^n, Σ^n).
######## Forward simulations: Populate the grid with M experiments ########
for $i := 1$ **to** M **do**
  for $k := n$ **to** N **do**
    Using the fixed and given allocation scheme, evaluate dose z_i^k.
    Simulate future observation $\hat{y}_i^{k+1} | \mu_i^k, \Sigma_i^k, z_i^k \sim \mathcal{N}(\mu_{i,z}^k, \sigma^2 + \Sigma_{i,zz}^k)$.
    Update the state using $\hat{y}_i^{k+1}$ to obtain $(\mu_i^{k+1}, \Sigma_i^{k+1})$ by equation (3.1).
    Generate T samples of dose-response $\Theta_{i,t}^k \sim \mathcal{N}(\mu_i^{k+1}, \Sigma_i^{k+1})$.
    Estimate the target dose $z_{i,t}^{*,k}$ for each $\Theta_{i,t}^k$ using equation (3.2).
    Let $df_{i,t}^{*,k} = f(z_{i,t}^{*,k}, \Theta_{i,t}^k) - f(0, \Theta_{i,t}^k)$.
    Estimate m_i^k and ν_i^k using T-sample mean and T-sample variance.
    Record the trajectory of (m_i^k, ν_i^k) in the grid (m, ν, k) for experiment i.
########## Forward simulations: Populate the empty cells ##########
for each empty cell j in the grid $(m, \nu, n : N)$ **do**
  Identify (m_j, ν_j, n_j).
  for $i := 1$ **to** M' **do**
    for $k := n_j$ **to** N **do**
      Repeat steps 5-12.
################## Backward induction ##################
for $k := N$ **to** n **do**
  for each cell j in the grid (m, ν, k) **do**
    Determine A_j^k defined in Section 3.4.1.
    if $k = N$ **then**
      Evaluate the optimal approximated value function by equation (3.8).
    else
      Evaluate the approximated utility of continuation by equation (3.9).
      Evaluate the optimal approximated utility by equation (3.10), thus the optimal decision.

trial incurring the cost of sampling and terminating or abandoning the trial at time $n+1$, i.e.,

$$V_{a^n=1}^{\mathrm{KG}}(s^n) = \mathbb{E}\Big\{ -c_1 + \max_{a^{n+1}\in\{0,2\}} u(a^{n+1}, s^{n+1}, \mathscr{F}^{n+1}) \Big| \mathscr{F}^n \Big\} - \max_{a^n\in\{0,2\}} u(a^n, s^n, \mathscr{F}^n), \qquad (3.11)$$

where the knowledge gradient policy π^{KG}, hereafter KG policy, decides to continue the trial, i.e., $a^{\pi^{KG}}(s^n) = 1$, when $V_{a^n=1}^{\mathrm{KG}}(s^n) > 0$. In the case that $V_{a^n=1}^{\mathrm{KG}}(s^n) \leq 0$, the optimal decision is identified by $a^{\pi^{\mathrm{KG}}}(s^n) \in \arg\max_{a^n\in\{0,2\}} u(a^n, s^n, \mathscr{F}^n)$. Note that $a^{\pi^{\mathrm{KG}}}(s)$ is a function returning the optimal decision selected when in state s^n under the KG policy π^{KG}. In order to evaluate $V_{a^n=1}^{\mathrm{KG}}(s^n)$, one needs to estimate both the current expected utility function $u(a^n, s^n, \mathscr{F}^n)$, and the one-step utility function $u(a^{n+1}, s^{n+1}, \mathscr{F}^{n+1})$ by taking a sample average (Monte Carlo). Algorithm 2 details this procedure. At each decision epoch n the state (μ^n, Σ^n) is given. To evaluate the utility of termination, a T-sized sample of dose-response curves Θ_t^n is generated. For each Θ_t^n, target dose

$z_t^{*,n}$ is evaluated via equation (3.2), and thus the advantage over placebo, i.e., df^* can be calculated where $df_{n,t}^* = f(z_t^*, \Theta_t^n) - f(0, \Theta_t^n)$. Estimate m_n, and ν_n using sample mean and sample variance. These values are later used in step 9 to evaluate the expected utility of termination. Then, n_p observations of y_* and y_0 are generated where $\hat{y}|\Theta, z \sim \mathcal{N}(\theta_z, \sigma^2)$ in order to estimate $\mathbb{E}[\mathbb{1}_{\{B^n\}}|\mathscr{F}^n]$ by Monte Carlo. The value of termination is then computed via equation (3.2).

Instead of simulating the entire trial to the last participant N to evaluate the value of continuation, a one-step look-ahead policy is implemented where at each decision epoch n, the next stage is assumed to be the last. Therefore, the value of continuation is computed by looking one step into the future. Starting from step 11 in Algorithm 5, the trial is simulated one-step into the future by generating future observations and updating the estimate of the dose-response curve Θ with respect to them. For each simulated observation, our belief about the dose-response curve is updated and the expected value of termination in the next stage is estimated. Taking a sample average over all these values results in an approximation of the expected utility of termination at decision epoch $n+1$. Since the expected value of abandonment is fixed to 0, one can approximate the value of continuation in equation (3.11) by taking the maximum over 0 and the approximated expected value of termination. The one-step look-ahead stopping rule is checked in step 17 of Algorithm 5. This approach replaces a large number of trial simulations from decision epoch n to N by one-step forward simulations of the trial, which significantly reduces the complexity and computational time of the algorithm.

The following result bounds the optimal decision from below, and shows that the KG policy may stop sooner than the optimal policy, i.e., whenever the KG policy decides to continue the trial, the optimal decision is also continuation of the trial. This proposition motivates a sensitivity analysis with respect to the history of the trial. In particular, we later show that stopping sooner than the optimal policy may result in low probability of correct decision in certain situations.

Proposition 3.4.1 *The optimal stopping time τ is bounded below by the KG stopping time τ^{KG}, i.e., $\tau^{KG} \leq \tau$.*

Proof. Consider the optimal stopping problem at time n when the system is in state s^n. Based on

Algorithm 5 One-Step Look-Ahead Policy

1: **Input.** State (μ^n, Σ^n) at the beginning of each decision epoch.
2: Generate T samples of $\Theta_t^n \sim \mathcal{N}(\mu^n, \Sigma^n)$.
3: Estimate the target dose $z_t^{*,n}$ using equation (3.2).
4: Let $df_{n,t}^* = f(z_t^*, \Theta_t^n) - f(0, \Theta_t^n)$.
5: Estimate m_n and ν_n using T-sample mean and T-sample variance.
6: Generate n_p observations of $\hat{y}_{*,t}^n - \hat{y}_{0,t}^n | \Theta_t^n$.
7: Check whether the event $B_t^n := \{\frac{\sqrt{n_p}(\bar{y}_{*,t} - \bar{y}_{0,t})}{\sqrt{2\sigma^2}} > q_\alpha\}$ holds true.
8: Estimate $\mathbb{E}[\mathbb{1}_{\{B^n\}} | \mathscr{F}^n]$ by taking a sample average over T samples.
9: Evaluate the value of termination, i.e., $a^n = 2$ using equation (3.2).
10: ##################One-step forward simulation####################
11: **for** $t := 1$ **to** T **do**
12: Simulate future observation $\hat{y}_t^{n+1} | \Theta_t^n, z^n \sim \mathcal{N}(\theta_{t,z^n}, \sigma^2)$.
13: Update the state using $\hat{y}_t^{n+1}$ to obtain $(\mu_t^{n+1}, \Sigma_t^{n+1})$ by equation (3.1).
14: Generate M posterior samples of $\Theta_{t,m}^{n+1} \sim \mathcal{N}(\mu_t^{n+1}, \Sigma_t^{n+1})$.
15: Repeat steps 3-9 to evaluate the value of termination by taking sample average of M values.
16: Evaluate $V_{a^n=1}^{\mathrm{KG}}$ via equation (3.11) by taking a sample average of T estimated termination values in the above "for" loop.
17: **if** $V_{a^n=1}^{\mathrm{KG}} > 0$ **then**
18: The optimal decision is to continue, go to step 2, $n \leftarrow n+1$.
19: **else**
20: Terminate or abandon the trial using the expected value of termination evaluated at step 9.

equation (3.11), the KG policy decides to continue the trial only if

$$\max\Big(0, -c_1' n_p + c_2 m_n \mathbb{E}[\mathbb{1}_{\{B^n\}} | \mathscr{F}^n]\Big) < \mathbb{E}^n\Big\{ - c_1 + \max_{a^{n+1} \in \{0,2\}} u(a^{n+1}, s^{n+1}, \mathscr{F}^{n+1}) \Big| \mathscr{F}^n \Big\},$$

where the left hand side of the inequality denotes the value of terminating or abandoning the trial at time n while the right hand side denotes the value of continuing the trial. To prove the proposition, it is enough to show that whenever the KG policy decides to continue the trial, the optimal policy also chooses continuation. The optimal policy decides to continue the trial if

$$\max\Big(0, -c_1' n_p + c_2 m_n \mathbb{E}[\mathbb{1}_{\{B^n\}} | \mathscr{F}^n]\Big) < \mathbb{E}^\tau \sup_{\tau \geq n+1} \Big\{ - c_1(\tau - n) + \mathbb{E}^n\big[\max_{a^\tau \in \{0,2\}} u(a^\tau, s^\tau, \mathscr{F}^\tau) \big| \mathscr{F}^n \big] \Big\},$$

where the value of termination or abandoning is equal to that of the KG policy, i.e., the left hand sides in both above inequalities are equal. Note that the supremum is taken over the set $\tau \geq n+1$

which contains $\tau = n+1$, and thus

$$\begin{aligned}
&\mathbb{E}^{\tau} \sup_{\tau \geq n+1} \Big\{ -c_1(\tau - n) + \mathbb{E}^n \big[\max_{a^\tau \in \{0,2\}} u(a^\tau, s^\tau, \mathscr{F}^\tau) \big| \mathscr{F}^n \big] \Big\} \\
&\qquad = \mathbb{E}^{\tau} \sup_{\tau \geq n+1} \Big\{ \mathbb{E}^n \big[-c_1(\tau - n) + \max_{a^\tau \in \{0,2\}} u(a^\tau, s^\tau, \mathscr{F}^\tau) \big| \mathscr{F}^n \big] \Big\} \\
&\qquad \geq \mathbb{E}^{n+1} \Big\{ \mathbb{E}^n \big[-c_1(n+1-n) + \max_{a^{n+1} \in \{0,2\}} u(a^{n+1}, s^{n+1}, \mathscr{F}^{n+1}) \big| \mathscr{F}^n \big] \Big\} \\
&\qquad = \mathbb{E}^n \Big\{ -c_1 + \max_{a^{n+1} \in \{0,2\}} u(a^{n+1}, s^{n+1}, \mathscr{F}^{n+1}) \Big| \mathscr{F}^n \Big\},
\end{aligned}$$

where the last equality is justified by the tower property of conditional expectation. Therefore, whenever the KG value of continuation is greater than abandoning or terminating the trial at time n, the optimal value of continuation is also greater and the optimal policy decides to continue the trial. □

3.4.3 Diffusion Approximation

Although the complexity and computational time of the knowledge gradient method is significantly better than the simulation-based gridding method, both require forward simulations to approximate the optimal solution to the value functions in (3.6). Instead, we propose a method that assumes a prior belief about the actual benefit of the target dose over placebo and approximates its increments over time by a continuous-time Wiener process, which enables us to analyze the optimal stopping boundaries offline. This framework is inspired by Chernoff (1961), where a diffusion approximation is used to sequentially test whether the drift of a Wiener process is positive. Brezzi & Lai (2002), Chick & Gans (2009), Chick & Frazier (2012) and others used this framework to approximate the solution of the Bellman equation in option pricing, ranking and selection, and multi-armed bandit settings. Our approach also approximates the stopping time of sequential normal means (i.e., advantage over placebo) by solving a continuous-time Bellman equation. To that end, we first consider a setting where there is a single unknown dose versus a known placebo and develop optimal stopping boundaries for it. Then, we design a heuristic that uses the said boundaries to create decisions in multiple doses settings. Details of this method are included to demonstrate the challenges when extended to multiple doses.

A single dose with unknown mean response versus a placebo with known mean response. For now, assume that the trial involves a placebo with known expected response and a single dose

with unknown expected response. In particular, without loss of generality assume that $y_0 \sim \mathcal{N}(0, \sigma^2)$ and $y_* \sim \mathcal{N}(\theta, \sigma^2)$, where θ is unknown and a prior $\theta \sim \mathcal{N}(m_0, \nu_0^2)$ is given, where we set $t_0 = \frac{\sigma^2}{\nu_0^2}$. For a single dose, the advantage over placebo is given by $df^* = \theta - 0 = \theta$; see Section 3.3. Therefore, at each time period, a sample from the dose with unknown mean is observed and the posterior on θ and, therefore, on df^* becomes $df^*|\mathscr{F}^n \sim \mathcal{N}(m_n, \frac{\sigma^2}{t_n})$ where

$$\begin{aligned} t_n &= t_0 + n, \\ m_n &= \frac{t_0}{t_n} m_0 + \frac{\sum_{i=1}^n \hat{y}_*^i}{t_n}. \end{aligned} \tag{3.12}$$

Note that in this setting df^* naturally follows a normal distribution. Recall that in the utility calculation there is an expectation to calculate, which by this construction has a closed form. In particular, we have

$$\begin{aligned} \mathbb{E}[\mathbb{1}_{\{B^n\}}|\mathscr{F}^n] &= \mathbb{P}\left\{ \frac{\sqrt{n_p}(\bar{y}_* - \bar{y}_0)}{\sqrt{(2\sigma^2 + \frac{\sigma^2}{t_n})}} > q_\alpha \middle| \mathscr{F}^n \right\} \\ &= 1 - \Phi\big(Q_\alpha(m_n, t_n)\big), \end{aligned}$$

where $Q_\alpha(m_n, t_n) = q_\alpha - \frac{m_n \sqrt{n_p}}{\sqrt{(2\sigma^2 + \frac{\sigma^2}{t_n})}}$, $2\sigma^2 + \frac{\sigma^2}{t_n}$ is the posterior predictive variance of $\bar{y}_* - \bar{y}_0$, and $\Phi(\cdot)$ denotes a normal cumulative distribution function.

Redefine the state variable $\hat{s} = (m_n, t_n)$, and using $\hat{s}^0 = (m_0, t_0)$, let $\tilde{l}_\pi(\hat{s}^0)$ denote the expected utility at stopping time τ under policy $\pi \in \Pi$ when the initial prior is parametrized by (m_0, t_0), i.e.,

$$\tilde{l}_\pi(\hat{s}^0) = \mathbb{E}^\pi \left[-c_1 \tau + \max\left\{0, -c_1' n_p + c_2 m_\tau \Big(1 - \Phi(Q_\alpha(m_\tau, t_\tau))\Big)\right\} \middle| \hat{s}^0 \right], \tag{3.13}$$

where the investigator selects a policy $\pi \in \Pi$ such that $V^*(\hat{s}^0) = \sup_{\pi \in \Pi} \tilde{l}_\pi(\hat{s}^0)$.

Define $x_0 = m_0 t_0$ and $x_n = x_0 + \sum_{i=1}^n \hat{y}_*^i$ where $m_n = \frac{x_n}{t_n}$. Using these definitions, the state variable can be rewritten as $\hat{s}^n = (x_n, t_n)$. Let $G(x_\tau, t_\tau)$ denote the optimal expected utility at the stopping time given by

$$G(x_\tau, t_\tau) = \max\left\{0, -c_1' n_p + c_2 \frac{x_\tau}{t_\tau} \Big(1 - \Phi\big(Q_\alpha(\frac{x_\tau}{t_\tau}, t_\tau)\big)\Big)\right\}. \tag{3.14}$$

Since the utility functions are uniformly bounded for any state and action, and the action space is

finite, there exits a Markovian and deterministic optimal policy (Bertsekas & Shreve 1996, Chapter 8). Therefore, the optimal policy to $V^*(m_0, t_0) = \sup_{\pi\in\Pi} \tilde{l}_\pi(m_0, t_0)$ is also the solution to the following Bellmen equation

$$\begin{aligned} B(x_n, t_n) &= \max\Big\{G(x_n, t_n), -c_1 + \mathbb{E}\big[B(x_{n+1}, t_{n+1})\big|x_n, t_n\big]\Big\}, \\ B(x_\tau, t_\tau) &= G(x_\tau, t_\tau), \end{aligned} \tag{3.15}$$

where $t_{n+1} = t_n + 1$, and $x_{n+1} = x_n + \hat{y}_*^{n+1}$.

Note that optimality equation (3.15) has a continuous state space and thus it is computationally intractable to solve. Therefore, in order to approximate the solution to the Bellman equation in (3.15), suppose that patients' responses are observed continuously rather than at discrete decision epochs t_n. Also, extend x_n to be a continuous real valued random variable for real valued t_n. Therefore, the cumulative sum $x_n = x_0 + \sum_{i=1}^n \hat{y}_*^i$ may be interpreted as accumulated diffusion of patients' responses where x_n is a Brownian motion with drift m_n and variance σ^2 per unit time, that is,

$$dx_n = m_n\, dt + \sigma\, dW_n, \tag{3.16}$$

where W_n is a standard Brownian motion. Note that the diffusion process is utilized to approximate the posterior mean of the dose level. Extend the definition of filtration $\mathscr{F}^n$ to be the natural sigma-algebra generated by the process $\{x_n\}_{n\in[t_0,t_n]}$, i.e., $\mathscr{F}_{ct}^{n\in[t_0,t_n]}$. Therefore, the continuous-time approximation of the Bellman equation in (3.15) is given by

$$\begin{aligned} B_{ct}(x_n, t_n) &= \max\Big\{G(x_n, t_n), -c_1\Delta t + \mathbb{E}\big[B(x_{n+\Delta t}, t_n + \Delta t)\big|\mathscr{F}_{ct}^n\big]\Big\}, \\ B_{ct}(x_\tau, t_\tau) &= G(x_\tau, t_\tau). \end{aligned} \tag{3.17}$$

The following proposition shows that $B_{ct}(x_n, t_n)$ is the solution to a free boundary problem with a partial differential diffusion-advection equation and two boundary conditions.

Proposition 3.4.2 *$B_{ct}(x_n, t_n)$ is the solution to the following partial differential equation in the continuation set $\mathscr{C} := \Big\{(x_n, t_n) : -c_1\Delta t + \mathbb{E}\big[B_{ct}(x_{n+\Delta t}, t_n + \Delta t)\big|\mathscr{F}_{ct}^n\big] > G(x_n, t_n)\Big\}$,*

$$0 = -c_1 + \frac{\partial B_{ct}(x_n, t_n)}{\partial t} + \frac{\partial B_{ct}(x_n, t_n)}{\partial x}\frac{x_n}{t_n} + \frac{1}{2}\frac{\partial^2 B_{ct}(x_n, t_n)}{\partial x^2}\sigma^2, \tag{3.18}$$

where $B_{ct}(x_n, t_n) = G(x_n, t_n)$ outside of the continuation set $\mathscr{C}$. The free boundary $\partial\mathscr{C}$ is given by

$$\begin{aligned} B_{ct}(x_n, t_n) &= G(x_n, t_n), \qquad \text{on } \partial\mathscr{C}, \\ \frac{\partial B_{ct}(x_n, t_n)}{\partial x} &= \frac{\partial G(x_n, t_n)}{\partial x}, \quad \text{on } \partial\mathscr{C}. \end{aligned} \tag{3.19}$$

Proof. Assuming that $B_{ct}(x_n, t_n)$ is twice differentiable in the continuation set $\mathscr{C}$, $B_{ct}(x_{n+\Delta t}, t_n + \Delta t)$ in equation (3.17) may be written according to the Taylor series expansion by

$$B_{ct}(x_n, t_n) = -c_1\Delta t + \mathbb{E}\Big[B_{ct}(x_n, t_n) + \frac{\partial B_{ct}(x_n, t_n)}{\partial t}\Delta t + \frac{\partial B_{ct}(x_n, t_n)}{\partial x}\Delta x + \frac{1}{2}\frac{\partial^2 B_{ct}(x_n, t_n)}{\partial x^2}(\Delta x)^2 + \mathcal{O}(\partial t)\Big],$$

where $\mathcal{O}(\partial t)$ denotes all the terms in the Taylor expansion with ∂t^2, $\partial t \partial x$, or higher degrees of differentiability in ∂t. Replacing Δx_n with $m_n\, dt + \sigma\, dW_n$, we have

$$\begin{aligned} B_{ct}(x_n, t_n) = -c_1\partial t + \mathbb{E}\Big[B_{ct}(x_n, t_n) &+ \frac{\partial B_{ct}(x_n, t_n)}{\partial t}\partial t + \frac{\partial B_{ct}(x_n, t_n)}{\partial x}(m_n\, \partial t + \sigma\, \partial W_n) \\ &+ \frac{1}{2}\frac{\partial^2 B_{ct}(x_n, t_n)}{\partial x^2}(m_n\, \partial t + \sigma\, \partial W_n)^2 + \mathcal{O}(\partial t)\Big]. \end{aligned}$$

Using Itô's lemma, and noting that ∂t^2 and $\partial t \partial W$ tend to zero faster than ∂W^2 when $\partial t \to 0$,

$$\begin{aligned} B_{ct}(x_n, t_n) = -c_1\partial t + \mathbb{E}\Big[B_{ct}(x_n, t_n) &+ \Big(\frac{\partial B_{ct}(x_n, t_n)}{\partial t} + \frac{\partial B_{ct}(x_n, t_n)}{\partial x}m_n + \frac{1}{2}\frac{\partial^2 B_{ct}(x_n, t_n)}{\partial x^2}\sigma^2\Big)\partial t \\ &+ \sigma\frac{\partial B_{ct}(x_n, t_n)}{\partial x}\partial W_n\Big], \end{aligned}$$

where ${\partial W_n}^2$ is substituted with ∂t. Noting that $\mathbb{E}[\partial W_n] = 0$ because W_n is a standard Brownian motion, we have

$$0 = -c_1 + \frac{\partial B_{ct}(x_n, t_n)}{\partial t} + \frac{\partial B_{ct}(x_n, t_n)}{\partial x}\frac{x_n}{t_n} + \frac{1}{2}\frac{\partial^2 B_{ct}(x_n, t_n)}{\partial x^2}\sigma^2,$$

where the terms $B_{ct}(x_n, t_n)$ cancel each other on both sides of the equality, the equation is divided by $\Delta t = \partial t$, and m_n is replaced with $\frac{x_n}{t_n}$. The first boundary is derived by the definition of the continuation set, and the second boundary is a so-called smooth pasting condition. □

Note that the boundaries to the continuation set $\mathscr{C}$ can be found without any trial simulation, which significantly reduces the complexity and computational effort required to obtain optimal stopping times. The solution to the free boundary problem in Proposition 3.4.2 is evaluated using a trinomial tree discretization method. To that end, create a grid over (x_n, t_n) for all $0 \le n \le N$ by

considering a rectangle $[t_0, t_0 + N] \times [\underline{x}, \overline{x}]$ where $\underline{x}$ and $\overline{x}$ are appropriately selected lower and upper bounds for x. These bounds are selected to be similar to the range considered in the simulation-based gridding algorithm. According to the trinomial tree method, the Brownian process defined in equation (3.16) in any given cell in the grid, i.e., (x_i, t_i), can move to one the following three different cells:

$$\begin{cases} (x_{i-1}, t_{i+1}) & \text{with probability } p_d, \\ (x_i, t_{i+1}) & \text{with probability } p_m, \\ (x_{i+1}, t_{i+1}) & \text{with probability } p_u, \end{cases}$$

where probabilities p_d, p_m, and p_u satisfy the following equations (see Ingber *et al.* 2001)

$$\begin{cases} p_u + p_m + p_d = 1, \\ \Delta x(p_u - p_d) = x\Delta t \\ (\Delta x)^2(p_u + p_d) = \tilde{\sigma}^2 \Delta t, \end{cases}$$

where Δx and Δt are carefully selected grid intervals, and $\tilde{\sigma} = \frac{2\sigma^2}{t} - \frac{2\sigma^2}{t+1}$ denotes the posterior variance at $t + 1$. To choose appropriate grid intervals, set Δt such that $\frac{1}{\Delta t}$ is equal to an integer value. In Section 3.5, we considered $\Delta t = 0.05$. Following Arlotto *et al.* (2010), we also assume that $p_u = p_d$, thus $p_u = p_d = \frac{\tilde{\sigma}^2 \Delta t}{2(\Delta x)^2}$, and $p_m = 1 - 2p_d$. Therefore, the probabilities p_d or p_u are maximized when $t = t_0$. Since $p_u + p_d \leq 1$, $p_u \leq p_{max} \leq 0.5$, we have $\Delta x = \frac{\sigma\sqrt{2\Delta t}}{\sqrt{2t_0(t_0+1)p_{max}}}$. We assume $p_{max} = 0.495$. The backward solution to the grid is given by

$$B(x_i, t_i) = p_u B(x_{i+1}, t_{i+1}) + p_m B(x_i, t_{i+1}) + p_d B(x_{i-1}, t_{i+1}),$$

where at the top or bottom row of cells in x axis, the cell values are extended in a linear fashion. Therefore, at $i = 0$, or $i = I$, we have

$$B(x_{I+1}, t) = 2B(x_I, t) - B(x_{I-1}, t),$$
$$B(x_{0-1}, t) = 2B(x_0, t) - B(x_1, t).$$

Note that in the last column of cells in t axis where $t = t_0 + N$, the cell values are calculated by the boundary condition $B(x_i, t_0 + N) = G(x_i, t_0 + N)$. After enumerating the entire grid by values of $B(x_i, t_i)$, grid cells for which $B(x_i, t_i) = G(x_i, t_i)$ are recorded and their (x_i, t_i) values are extracted.

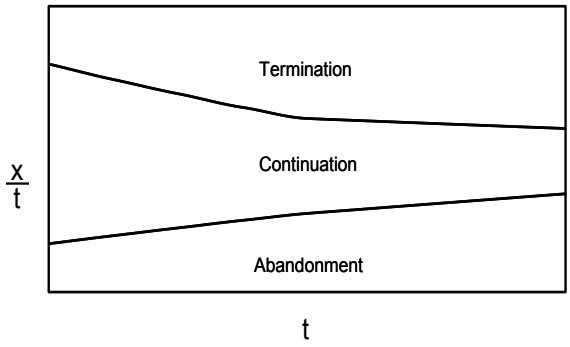


Figure 3.4: An example of boundaries of the continuation set

To apply the smooth pasting condition in the free boundary problem, we apply a smoothing spline package in R programming language to smooth the extracted (x_i, t_i) values; thus the lower and upper boundaries. To find the optimal decision at each decision epoch, observing true observations y^n, update the values of m_n, ν_n, and n, then calculate x_n and t_n by using the following formulas already given in Section 3.4.3.

$$t_0 = \frac{2\sigma^2}{\nu_0^2},$$

$$t_n = t_0 + n,$$

$$x_n = m_n t_n.$$

If (x_n, t_n) correspond to a cell inside the area marked by lower and upper boundaries, the trial is still in the continuation set and the optimal decision is to continue. However, if the corresponding cell falls outside of the boundaries, the decision is to stop the trial, i.e., terminate or abandon.

Figure 3.4 demonstrates an example of the solution to the free boundary problem. The area between the two lines denoted by $\mathcal{C}$ represents the continuation set, whereas $\mathcal{A}$ and $\mathcal{T}$ show abandonment and termination regions, respectively.

Multiple doses with unknown mean responses. In the previous part, we construct the continuation boundaries where there is only a single dose with unknown response. However, the original problem consists of multiple doses for which the mean response is unknown. Therefore, the target dose $z^* = \text{ED}_{95}$ is random and each continuation decision may yield a different target dose with respect to the sample path. This results in an unknown distribution for df^* when multiple doses are considered. In fact, if there was only a single dose with unknown response, the allocation dose

for continuation decisions and the target dose were similar, and posterior advantage over placebo, df^*, was distributed according to a normal distribution. However, in multiple doses setting, the allocation dose for continuation decisions may estimate a different target dose, which results in df^* to not enjoy conjugacy with respect to the patient's response. In the literature, a variety of heuristic approaches have been proposed to extend the results of a single alternative case to multiple doses settings. For example, Chick & Gans (2009) proposed a hierarchical approach in multi-armed bandit settings, where each armed is treated separately at first and an optimal stopping policy is identified assuming only one arm exists at a time. Then, Gittin's indices are evaluated for each arm to sequentially select which arm to sample from if any; see Glazebrook (1979). The authors also consider a fully sequential settings where each arm is simulated according to a one-step look-ahead policy and the the arm with the highest expected reward is selected to sample from. Chick & Frazier (2012) considered a system where at each time, one arm has an unknown mean and is compared in a single dose setting to another arm with known mean where the value of the known mean is equal to the maximum of the posterior expected reward of other arms. Then, a diffusion approximation problem is solved considering each arm at a time as the unknown arm, and the arm with the highest solution to the approximated Bellman equation is selected as the arm to continue to sample from. The algorithm stops if no arm produces a positive solution to the Bellman equation, and it is indifferent to solutions equal to zero. Chick *et al.* (2018) proposed an indexing policy by which the incremental value of sampling only from one arm beyond stopping point and selecting the current best arm is used to generate diffusion approximation subproblems for each arm. Then, the solutions to these problems are used to select the arm with the highest index to be sampled from. However, these approaches are dependent on assuming the reward in optimal stopping problem is equivalent to maximum expected reward of simulating the arm with the highest mean whereas in our formulation, the arm with the highest mean does not necessarily yields the maximum expected utility, and thus they are not directly applicable in our settings. Therefore, we propose the following heuristic to extend the single dose system to multiple doses case.

Recall that for each dose Z_j, there is a prior on the expected response θ_j. The diffusion approximation boundaries only depend on the prior and the shape of the utility function. Therefore, for each dose j, we construct the continuation boundaries upfront. The idea is that, at each decision epoch, we estimate $m_n^* = \frac{x_n^*}{t_n}$ of the target dose z^* and make decisions by considering the optimal region corresponding to dose z^* to check whether m_n^* falls into termination, abandonment,

or continuation regions. To that end, we create a sample from the posterior on Θ, and for each sample we use equation (3.2) to find the target dose. Then, we take a sample average to estimate the target dose and since said sample average may not be in $\mathcal{Z}$, we round it to the closest dose. Given the estimate of target dose z^*, we simply have $m_n^* = \mathbb{E}\{\theta_{z^*}|\mathscr{F}^n\}$. The decision is found by referring to the optimal decision region corresponding to dose z^* and checking whether m_n^* belongs to abandonment, continuation, or termination zone at t_n. Assuming a continuation decision at time n, a patient is assigned to a dose according to the allocation scheme. Its response y^{n+1} is observed and is used to update the estimate of the dose-response curve, i.e., Θ. Then, the above process continues until the stopping time or all patients are tested.

3.5 Numerical Analysis

In this section, we present implementation results of simulation-based gridding algorithm, one-step look-ahead policy, and diffusion approximation for a variety of settings. Since the performance of these solution methods may differ depending on the adaptive allocation scheme, we assume that the allocation algorithm is given and fixed to that in Section 3.4. To assess the quality of solution methods with respect to termination, abandonment, and continuation decisions, two different types of dose-response curves are tested: (i) A sigmoid curve with a significant advantage over placebo, one of the most recurring dose-responses in practice (e.g., Gadagkar & Call 2015). This curve is used to test the performance of different stopping rules with respect to continuation and termination decisions. Note that for this curve the optimal decision at stopping is to terminate the trial for efficacy. (ii) A flat dose-response curve, which is used to assess the quality of different algorithms when the correct decision at stopping is to abandon the trial for futility.

Note that the problem is modeled as a Bayesian Markov decision process and naturally it is optimal when assessed according to a fully Bayesian setup, i.e., problem instances, or in other words, true dose-response curves must be generated randomly from the same prior and the performance must be measured with respect to expectation under the particular assumed prior. However, because of computational difficulties in generating results for the simulation-based gridding approach, we assess the performance of these approximation methods with respect to two dose-response curves (a frequentist setting). Assessing different algorithms with respect to specific configuration is not without precedence particularly in clinical trials; see for example Berry *et al.* (2002), and Krams *et al.*

(2003). A sigmoid and a flat curve are considered to highlight the performance of these algorithms when facing favorable and unfavorable cases and to derive managerial insights with respect to the design of clinical trials.

3.5.1 Simulation Initialization

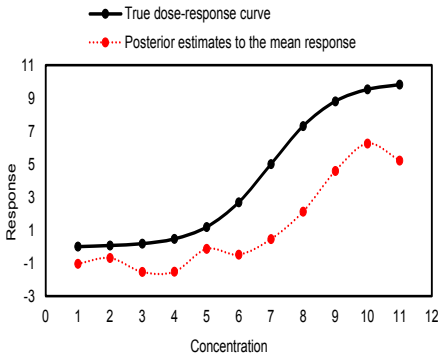
A typical number of doses under investigation in Phase II of clinical trials is between 4-12 (e.g., Berry *et al.* 2002). We considered 11 doses including placebo. The first dose is considered placebo and its known and fixed mean response marks the baseline score for any particular treatment. At each decision epoch if the decision is to continue the trial, a dose must be allocated to the next patient. We use a one-step look-ahead policy (knowledge gradient) to optimally select a dose which minimizes the one-step posterior variance of the target dose ED_{95}. Thereafter, the patient's response is generated from the true distribution and is used to update the posterior estimate of the dose-response curve. Aligned with literature (e.g., Berry *et al.* 2002), we assume that the stopping algorithm is applied only after observing the responses of a certain number of patients, e.g., 20, have already been through the trial. The total number of participants volunteered for the trial is assumed to be 400. We later show why one may be interested in applying the stopping rules only after a certain number of patients have been through the trial. We assume that the observation variance is known and is fixed at 100 units. A sensitivity analysis is also conducted on this assumption. The significance level is considered to be 1% across all experiments. We assume that the sampling cost $c_1 = 1000$, sampling cost in confirmatory phase $c_1' = 1000$, and reward per unit advantage over placebo $c_2 = 1,000,000$. The prior (μ_0, Σ_0) is set according to $\mu_0 = (0, \ldots, 0)$, and Σ_0 is initiated by a Gaussian covariance function where $\text{Cov}(\theta_i, \theta_j) = \beta \exp\{-\gamma(i-j)^2\}$ where β is usually estimated by $\text{Var}(\theta_i)$ (Rasmussen & Williams 2006). The Gaussian structure of the covariance function allows for less correlation when doses are further apart. To keep symmetry of the covariance matrix, β is chosen to be equal to $\frac{\text{Var}(\theta_i) + \text{Var}(\theta_j)}{2} = 100$, and γ, the lengthscale factor is set to 0.01 for both sigmoid and flat curves. A thinning factor of 5 is used in generating random variables where every fifth random variable created is used to avoid serial correlation in a computer generated sequence of random numbers. In reporting the results, 30 simulations with different sequence of random numbers are considered. The simulation is coded in R programming language and is run on an Intel core i7 3.7 GHz processor with 16 GB of RAM.

In case of the simulation-based gridding algorithm, recall that the advantage over placebo,

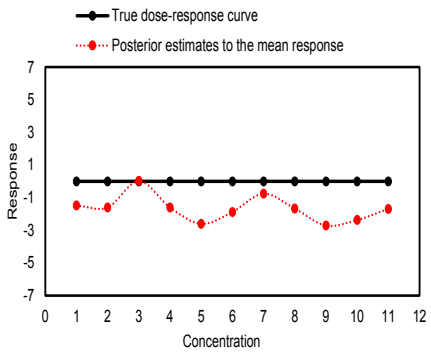
i.e., df^*, in the literature, is assumed to be normally distributed according to $\mathcal{N}(m, \nu^2)$ with respect to filtration $\mathscr{F}$. The prior values for m_0 and ν_0 are set equal to 0 and 10 to ensure that the prior carries little information about the belief on df^*. In constructing the grid over m and ν, we considered the range of m to be 20 units, i.e., $[0, 20]$, and the range of ν to be 10 unit, i.e., $[0, 10]$. The grid is divided into 40 and 20 intervals in the m and ν axes, respectively. We later conduct a sensitivity analysis on grid range and cell size. Initially, to populate the grid, $M = 1000$ experiments are run and their (m, ν) trajectories are recorded over the grid. Afterwards, from each empty cell in the grid, $M' = 10$ more simulations are initiated and their trajectories are recorded. Each experiment is a multi-step forward simulation from decision epoch n to N, which is equivalent to repeating a one-step forward simulation multiple times using the estimated dose-response curve at the end of each step as the prior dose-response curve estimate for the next step. To implement the algorithm in an online fashion, we parallelize forward simulations to speed up the computation. For more details regarding the implementation of the simulation-based gridding algorithm and the one-step look-ahead policy, we refer the readers to Algorithms 4 and 5.

For diffusion approximation, the prior values for m_0 and ν_0 are chosen to replicate those of the simulation-based gridding algorithm. We also assume a similar range for m as in the gridding algorithm, i.e., $m \in [0, 20]$. The discretization in diffusion approximation is different from the grid construction in the simulation-based gridding algorithm. Here, the grid is constructed over values of x and t. Since 20 patients have already been through the trial, t is considered to be in $[20+t_0, 400+t_0]$ where $t_0 = \frac{\sigma^2}{\nu_0^2}$. The details to calculate both axis intervals are given in Section 3.4.3. We later do a sensitivity analysis on the grid size for both the simulation-based gridding algorithm and the diffusion approximation method.

Because the simulation of the trial for all three methods is the same, we report the computational time required to find the stopping decision for each method. At each decision epoch, the gridding algorithm runs a forward simulation and uses backward induction which takes 1 hour on average. Note that in this method the computational time in early stages when there are many patients to consider is considerably longer than the later stages when fewer patients are left. At each time period, the one-step look-ahead policy takes about 30 seconds to find the decision. The diffusion approximation creates the stopping regions upfront and for a given dose allocation and its response, finding the stopping decision is instantaneous. These results confirm that the proposed



(a) Sigmoid curve



(b) Flat curve

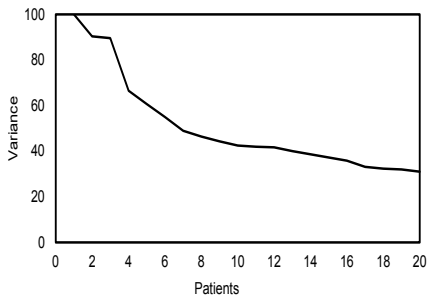
Figure 3.5: Posterior estimates to the dose-response curve after 20 patients

methods are much less demanding than the standard method.

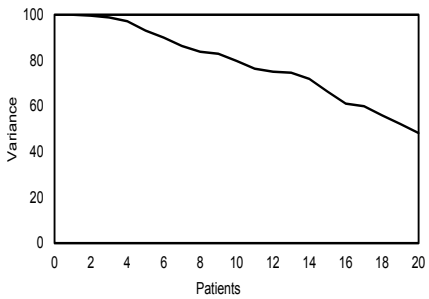
3.5.2 Results

State of dose-response estimation. Figures 3.5 and 3.6 show the state of the dose-response estimation after assigning 20 patients. In particular, Figures 3.5(a) and 3.5(b) show the posterior estimates to the dose-response curve where each point on the piecewise linear dotted line is the sample average of 30 posterior estimates of μ in $\Theta \sim \mathcal{N}(\mu, \Sigma)$ after observing 20 patients. Furthermore, Figures 3.6(a) and 3.6(b) show the maximum posterior variance where each point denotes the sample average of 30 posterior estimates of maximum Σ_{jj}, $j = 1, \ldots, J$ for each patient.

Expected utility and stopping time. Figure 3.7 shows the estimated expected utility $l_\pi(s^0)$ when stopping at patient τ. Recall that the objective is to select a policy, thus a stopping time, such that $l_\pi(s^0)$ is maximized, i.e., $\sup_{\pi \in \Pi} l_\pi(s^0)$. Figure 3.7(a) achieves its maximum expected utility

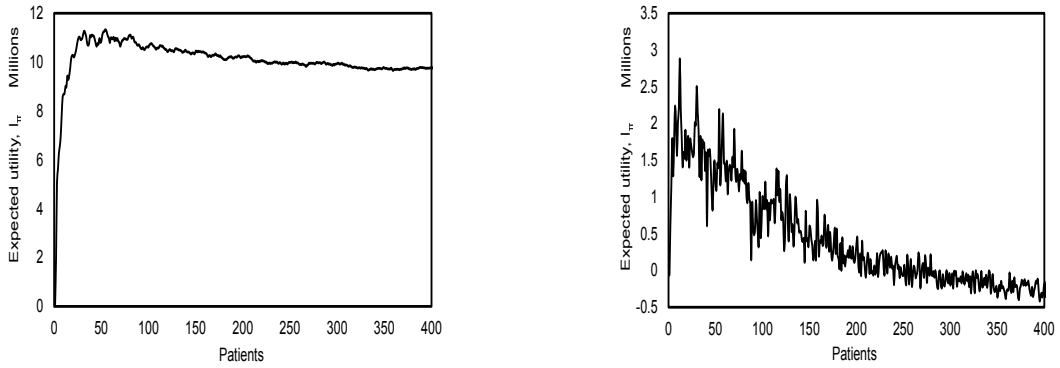


(a) Sigmoid curve



(b) Flat curve

Figure 3.6: Maximum posterior variance



(a) Sigmoid curve

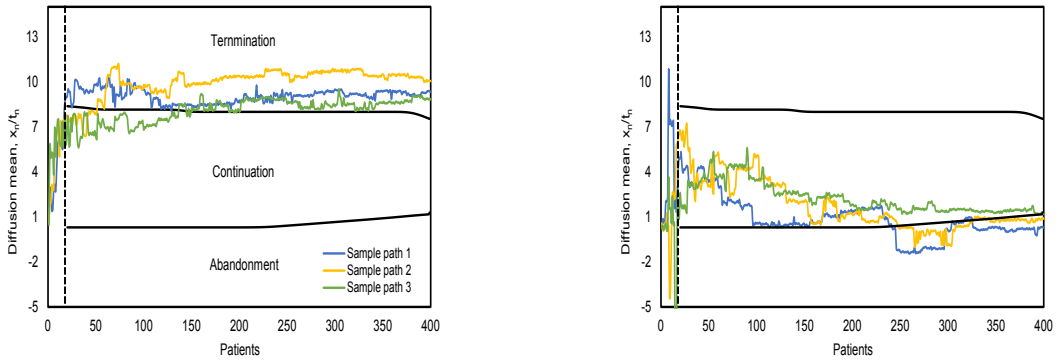
(b) Flat curve

Figure 3.7: Expected utility $l_\pi(s^0)$

at patient 54 where $l_\pi \cong 11,319,361$, whereas Figure 3.7(b) reaches its maximum expected utility at patient 30 where $l_\pi \cong 2,604,865$. Note that in the flat curve, Figure 3.7(b), the first maximum happens in the first 20 patient initialization and thus is not identified as the maximum expected utility. The best approximate solution method would choose a stopping time closest to 54 and 30 when the true dose-response curve is sigmoid and flat, respectively.

Tables 3.1 and 3.2 show the expected utility at stopping time, the average stopping time and the probability of correct decision (PCD) for the three stopping rules with respect to sigmoid and flat dose-response curves, respectively. In case a significant advantage over placebo exists, correct decision is to detect significance and terminate the trial. If the true dose-response curve is flat, abandoning the trial is considered as the correct decision. Notice that all three algorithms correctly terminate the sequential sampling process when the dose-response curve is sigmoid with significant advantage over placebo. The KG policy stops sooner but the expected utility at stopping time is higher for the simulation-based gridding algorithm. The diffusion approximation method achieves lower expected utility time and stops later. In particular, Figure 3.8(a) demonstrates a few diffusion paths crossing into the termination region from the continuation region. Note that the average stopping time and expected utilities reported in both Tables 3.1 and 3.2 are the average over 30 sample paths. Note that the reported probability of correct decision only considers the stopping decisions and is independent of the target dose selection in case of detecting a significance.

When sampling from a flat dose-response curve, the simulation-based gridding algorithm



(a) Sigmoid curve

(b) Flat curve

Figure 3.8: Diffusion paths

and the KG policy incorrectly terminate the trial most of the times. In case of KG policy, as soon as the next step expected utility is estimated to be less than the current one, the policy stops sampling. If the allocation policy overestimates the expected response of the target dose, the current expected utility may become positive and thus the incorrect decision to terminate the trial instead of abandoning. The simulation-based gridding algorithm performs better than the KG policy since it allows the forward simulation to continue until the end of the trial. However, the forward simulations depend on the predictive posterior distribution and if the prior does not provide accurate information, which is unlikely in early stages, then the multiple-steps look-ahead simulations may not accurately predict the future expected utilities and thus the incorrect decision. Table 3.2 shows that the diffusion approximation algorithm correctly abandons 96% of times when the dose-response curve is flat although the average abandonment time comes significantly further in the trial. For example, Figure 3.8(b) shows a few diffusion paths crossing into abandonment region. Furthermore, the expected utility at stopping time for the diffusion approximation algorithm, although lower than the simulation-based gridding and the KG policy, is closer to the true expected utility for the flat dose-response curve. Therefore, one might conclude that the simulation-based gridding and KG

Table 3.1: Performance of approximate solutions for sigmoid curve (20 patient initialization)

	Expected utility ($)	Stopping time	PCD
Simulation-based gridding	11,076,870	80	1
KG	10,791,346	38	1
Diffusion approximation	10,632,415	106	1

Note: Stopping times are reported in terms of number of patients going through the trial before an stopping decision is made.

algorithms do not accurately represent the true dose-response curve at the time of stopping. These results show that in this setting the standard method may produce significantly poor solutions, which may have severe consequences in terms of costs of the next phase and the health of future patients: see Rojas-Cordova & Hosseinichimeh (2018) for a discussion on consequences of misspecification errors in adaptive clinical trials.

One approach to address such a shortcoming of the gridding and KG policies is to start considering stopping decisions if enough evidence is gathered regarding the dose-response curve. This evidence may be interpreted as the accuracy of the dose-response estimation, i.e., the diagonal of the covariance matrix Σ in state variable s^n. In order to avoid tracking a J-sized vector, the maximum posterior variance is used to ensure the quality of estimation. Therefore, we consider stopping decisions if $\max_j \text{Var}[\theta_j|\mathscr{F}^n] \leq \bar{V}$, where $\bar{V}$ is a tuning threshold.

Table 3.2: Performance of approximate solutions for flat curve (20 patient initialization)

	Expected utility ($)	Stopping time	PCD
Simulation-based gridding	1,926,790	34	0.10
KG	1,840,765	28	0.10
Diffusion approximation	-3,910	277	0.96

Note: Stopping times are reported in terms of number of patients going through the trial before an stopping decision is made.

3.5.3 Sensitivity Analysis

Sensitivity to history $\mathscr{F}^n$. Motivated by our results, we propose applying the stopping rule only after a certain number of patients' responses have already been observed. As more patients' responses are added to the history, the accuracy of the estimation about Θ increases. This is because sampling dose j results in lowering Σ_{jj} which in turn is a measure of uncertainty about the dose-response estimation at dose j. Therefore, bounding $\max_j \text{Var}[\theta_j|\mathscr{F}^n]$ is ensuring a minimum level of accuracy about the state of dose-response curve estimation. Note that this modification does not contribute to the complexity of the stopping rule since $\text{Var}[\theta_j|\mathscr{F}^n] = \Sigma_{jj}$ is already available to the decision maker as part of the state space.

We propose the following heuristic: At each decision epoch, follow the standard method (or our proposed methods) to find stopping decisions; if the decision is to continue, continue; if the decision is to stop, check whether $\max_j \text{Var}[\theta_j|\mathscr{F}^n] \leq \bar{V}$ is satisfied; if it is satisfied, follow the

Table 3.3: Performance of approximate solutions for sigmoid curve (after $\max_j \text{Var}[\theta_j|\mathscr{F}^n] \leq 4$)

	Expected utility ($)	Stopping time	PCD
Simulation-based gridding	10,001,441	289	1
KG	9,957,969	282	1
Diffusion approximation	9,916,273	280	1

Note: Stopping times are reported in terms of number of patients going through the trial before an stopping decision is made.

decision; otherwise, continue.

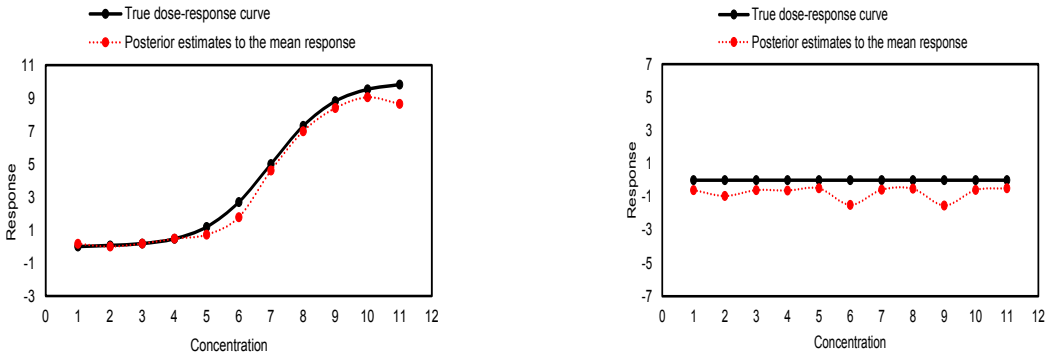
As mentioned before, one can tune $\bar{V}$ to change the amount of evidence gathered before the stopping decisions are applied. We consider $\bar{V} = 4$ units in presenting the results. Figure 3.9 shows the state of dose-response estimation in terms of posterior estimate to the dose-response curve when $\max_j \text{Var}[\theta_j|\mathscr{F}^n] \leq 4$ for the first time. Figure 3.10 shows the maximum posterior variance from the start of the trial until $\max_j \text{Var}[\theta_j|\mathscr{F}^n] \leq 4$ for the first time. In case of the sigmoid dose-response curve, $\max_j \text{Var}[\theta_j|\mathscr{F}^n] \leq 4$ when $n \geq 280$, and $\sup_{\pi\in\Pi} l_\pi(s^0) = 10,011,765$ which is achieved at patient 286. For a flat dose-response curve, $\max_j \text{Var}[\theta_j|\mathscr{F}^n] \leq 4$ when $n \geq 243$, and $\sup_{\pi\in\Pi} l_\pi(s^0) = -312$ is achieved at patient 296. Similar to Section 3.5, Tables 3.3 and 3.4 show the performance measures for the three stopping rule with respect to the sigmoid and flat dose-response curve, respectively.

Figure 3.11 shows the new continuation set boundaries and a few diffusion paths for sigmoid and flat dose-response curves. In case of the sigmoid dose-response curve, all diffusion paths would cross into the termination region before $\max_j \text{Var}[\theta_j|\mathscr{F}^n] \leq 4$. Thus, the average stopping time in Table 3.3 is reported to be equal to 280 because the diffusion algorithm decides on termination before further sampling. In case of the flat dose-response curve, a few diffusion paths cross into the abandonment region before $\max_j \text{Var}[\theta_j|\mathscr{F}^n] \leq 4$ is satisfied. However, a few diffusion paths remain in the continuation region a little longer and some do not cross into abandonment region at all which is why the average stopping time reported in Table 3.4 happens later and the probability of correct

Table 3.4: Performance of approximate solutions for flat curve (after $\max_j \text{Var}[\theta_j|\mathscr{F}^n] \leq 4$)

	Expected utility ($)	Stopping time	PCD
Simulation-based gridding	-11,199	284	0.69
KG	-12,889	244	0.64
Diffusion approximation	-12,571	304	0.96

Note: Stopping times are reported in terms of number of patients going through the trial before an stopping decision is made.



(a) Sigmoid curve

(b) Flat curve

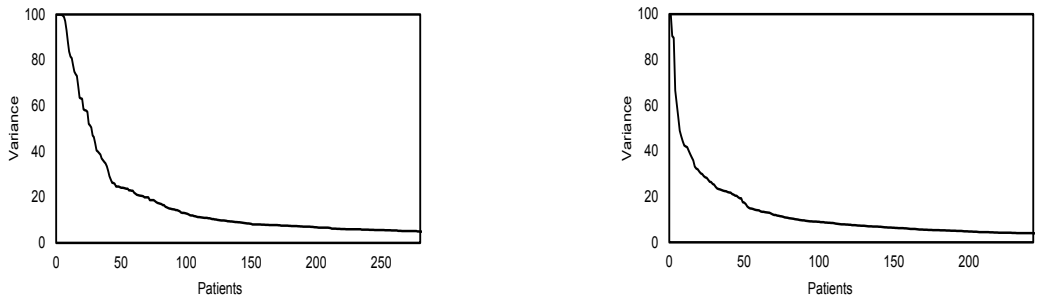
Figure 3.9: Posterior estimates to the dose-response curve when $\max_j \text{Var}[\theta_j|\mathscr{F}^n] \leq 4$

decision remains the same with respect to the results in Table 3.2.

Sensitivity to the variance of observation. The dose-response model in Section 3.2.1 is presented as

$$y = f(z, \Theta) + \epsilon,$$

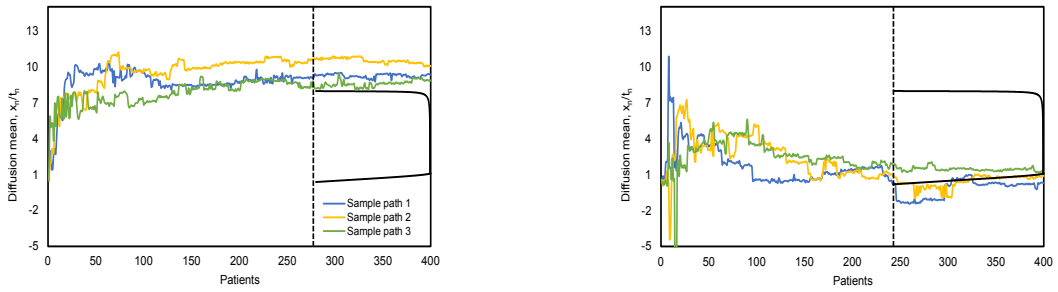
where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. In reporting the results in Section 3.5, we assumed that σ^2 is constant and equal to 100 units throughout the trial. However, this might not be the case in the real world, and thus we propose a sensitivity analysis with respect to the variance of observation, and show that our conclusions are robust. The following results are reported for $\sigma^2 = 1000$. In particular, Figure 3.12 shows the maximum posterior variance at each decision epoch for both sigmoid and flat dose-response curves until $\max_j \text{Var}[\theta_j|\mathscr{F}^n] \leq 4$ is satisfied. Note that to satisfy this condition in case of the sigmoid dose-response curve, $n \geq 815$, where the maximum expected utility is $8,590,869$ achieved at $n = 825$.



(a) Sigmoid curve

(b) Flat curve

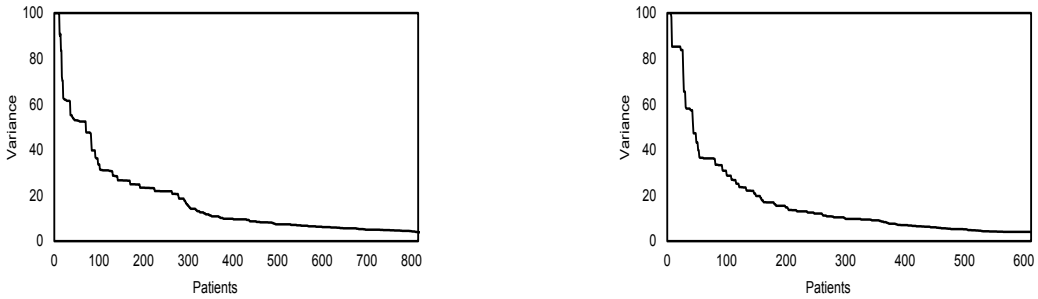
Figure 3.10: Maximum posterior variance until $\max_j \text{Var}[\theta_j|\mathscr{F}^n] \leq 4$



(a) Sigmoid curve (b) Flat curve

Figure 3.11: Diffusion paths after $\max_j \mathrm{Var}[\theta_j|\mathscr{F}^n] \leq 4$

For the flat dose-response curve, $n \geq 612$ satisfies the condition and results in a maximum expected utility of $-589,975$ at $n = 621$. Tables 3.5 and 3.6 denote the performance measurements of the three solution algorithm with respect to the sigmoid and flat dose-response curves, respectively. The results for the sigmoid dose-response curve are compatible with those of Tables 3.3 and 3.4 in Section 3.5. However, in case of the flat dose-response curve, the probability of correct decision has improved significantly for the simulation-based gridding algorithm and the one-step look-ahead policy. Notice that by increasing the variance of observation, the allocation algorithm requires a larger number of samples such that the cost of sampling cancels out any incorrect identification of improvement over placebo, and thus all algorithms abandon correctly.



(a) Sigmoid curve (b) Flat curve

Figure 3.12: Sensitivity of maximum posterior variance to observation variance

Sensitivity to discretization parameters. The simulation-based gridding algorithm and the diffusion approximation rely on discretization of the state space variables. In the simulation-based algorithm, the approximation is modified to depend on the true state variable, i.e., $s = (\mu, \Sigma)$, only

Table 3.5: Sensitivity of approximate solutions for sigmoid curve to observation variance

	Expected utility ($)	Stopping time	PCD
Simulation-based gridding	8,555,127	830	1
KG	8,550,258	815	1
Diffusion approximation	8,550,258	815	1

Note: Stopping times are reported in terms of number of patients going through the trial before an stopping decision is made.

Table 3.6: Sensitivity of approximate solutions for flat curve to observation variance

	Expected utility ($)	Stopping time	PCD
Simulation-based gridding	-627,207	616	1
KG	-627,207	613	1
Diffusion approximation	-792,184	612	1

Note: Stopping times are reported in terms of number of patients going through the trial before an stopping decision is made.

through $\tilde{s} = (m, \nu)$. In constructing the grid, the range over m is considered to be $[0, 20]$ divided into 40 intervals. The range over ν is considered to be $[0, 10]$ which is divided into 20 intervals. Our numerical analysis shows that expanding the range does not contribute to the solution in a significant way. However, doubling the number of intervals in each axis produced better results shown here in Table 3.7. Because of the very expensive computation efforts, we could never discretize the grid as finely as the diffusion approximation method.

In the case of diffusion approximation, the grid is constructed over another modified state variable $\hat{s} = (x, t)$. As described in details in Section 3.4.3, Δt is selected in such a way that $\frac{1}{\Delta t}$ is an integer. In reporting the results for the diffusion approximation method, we assumed $\Delta t = 0.05$. The interval in the x-axis is calculated by the procedure described in Section 3.4.3 and is dependent on Δt. Our numerical analysis showed that refining the grid cell sizes by changing $\Delta t = 0.01$, or expanding the range over x-axis does not contribute to the solution significantly. Note that the range for the t-axis cannot be changed.

Table 3.7: Sensitivity of simulation-based gridding to discretization

Sigmoid dose-response curve			
	Expected utility ($)	Stopping time	PCD
Grid interval=0.5	10,001,441	252	1
Grid interval=0.25	10,007,286	253	1
Flat dose-response curve			
Grid interval=0.5	-11,199	284	0.69
Grid interval=0.25	-8907	289	0.69

Note: Stopping times are reported in terms of number of patients going through the trial before an stopping decision is made.

3.6 Conclusion

In this chapter, we studied the optimal stopping problem of a fully adaptive dose-finding clinical trial with unique features, which separates it from standard optimal stopping problems. We implemented a standard algorithm (gridding) to approximately solve this problem and compared it with two methods that we proposed in terms of solution quality and computational effort. Our first proposed method assumes that the next decision epoch is the last one (KG) and produces stopping decisions accordingly. Our second proposal considers a two-doses continuous version of the sampling and stopping problem and creates an Itô process for the state transition by which solving the continuous Bellman equation coincides with solving a partial differential diffusion equation. We proposed a heuristic approach to extend the algorithm to multiple drug doses.

Our results show that if in the true dose-response curve the target dose has a significant advantage over placebo, all three methods make a right decision in terminating the trial for efficacy; the stopping time of KG is sooner, then the standard method, followed by the diffusion approximation. The estimate of the utility for the standard approach is higher than KG, followed by the diffusion approximation. However, if in the true dose-response curve the target dose does not have a significant advantage over placebo, the gridding and KG method perform extremely poorly in terms of probability of correct decision. In particular, these two methods decide on termination 90% of the times on average while the correct decision is abandonment, i.e., the error probability is 0.9 for these methods, which may have significant adverse consequences and is unacceptable for regularity approval. In fact, these two methods stop too early and significantly overestimate the benefits upon termination. In a stark contrast, the diffusion approximation method produced abandonment decision in 96% of time in this setting, producing only 4% error, and estimated the utility more accurately upon termination. The reason for such results is that the diffusion method stops the trial much later when it has enough evidence for making decisions.

Our results suggest that applying the standard method in a fully adaptive setting from early on, where a decision maker can stop or terminate the trial at each decision epoch, may have severe consequences when the true decision is to abandon. Motivated by such observations, we proposed a modified stopping rule, where the stopping decisions become activated only if the maximum posterior variance about the mean response Θ falls below a threshold. $\max_j \mathrm{Var}[\theta_j|\mathscr{F}]$ is a metric that measures the uncertainty about the whole dose-response curve and is available to decision

makers at each decision epoch because it is part of the state variable in both the allocation and the stopping problem. Our results show that using a constrained method significantly improves the performance of the simulation-based gridding algorithm and the KG policy.

Therefore, although considering financial evidence in designing a stopping rule for Phase II clinical trial is justified because of the usual high costs associated with sampling more participants particularly if measured with respect to the benefit they may provide, our results show that the standard method cannot guarantee a correct decision in certain situations, i.e., flat dose-response curves, and may lead to unnecessary and costly Phase III trials. However, we showed that utilizing the diffusion approximation method in optimal stopping of a dose-finding clinical trial consistently provides better probability of correct decision and may be a reliable method when considering financial evidence in stopping a trial.

Bibliography

Adelman, Daniel. 2007. Dynamic bid prices in revenue management. *Operations Research*, **55**(4), 647–661.

Ahuja, Vishal, & Birge, John R. 2016. Response-adaptive designs for clinical trials: Simultaneous learning from multiple patients. *European Journal of Operational Research*, **248**(2), 619–633.

Alagoz, Oguzhan, Maillart, Lisa M, Schaefer, Andrew J, & Roberts, Mark S. 2004. The optimal timing of living-donor liver transplantation. *Management Science*, **50**(10), 1420–1430.

Alanis, Ramon, Ingolfsson, Armann, & Kolfal, Bora. 2013. A Markov chain model for an EMS system with repositioning. *Production and Operations Management*, **22**(1), 216–231.

Allen, Arnold O. 1980. Queueing models of computer systems. *IEEE Computer Society*, 13–24.

Andersson, T., & Varbrand, Peter. 2007. Decision support tools for ambulance dispatch and relocation. *Journal of the Operational Research Society*, **58**(2), 195–201.

Arlotto, Alessandro, Gans, Noah, & Chick, Stephen. 2010. Optimal employee retention when inferring unknown learning curves. *Pages 1178–1188 of: Simulation Conference (WSC), Proceedings of the 2010 Winter*. IEEE.

Arlotto, Alessandro, Chick, Stephen E, & Gans, Noah. 2013. Optimal hiring and retention policies for heterogeneous workers who learn. *Management Science*, **60**(1), 110–129.

Bélanger, Valérie, Kergosien, Yannick, Ruiz, Angel, & Soriano, Patrick. 2016. An empirical comparison of relocation strategies in real-time ambulance fleet management. *Computers & Industrial Engineering*, **94**, 216–229.

Berger, Martijn PF, & Wong, Weng-Kee. 2009. *An Introduction to Optimal Designs for Social and Biomedical Research*. Vol. 83. John Wiley & Sons.

Berman, Oded. 1981a. Dynamic repositioning of indistinguishable service units on transportation networks. *Transportation Science*, **15**(2), 115–136.

Berman, Oded. 1981b. Repositioning of indistinguishable service units on transportation networks. *Computers and Operations Research*, **8**(2), 105–118.

Berry, Donald A, Mueller, Peter, Grieve, Andy P, Smith, Michael, Parke, Tom, Blazek, Richard, Mitchard, Neil, & Krams, Michael. 2002. Adaptive Bayesian designs for dose-ranging drug trials. *Pages 99–181 of: Case Studies in Bayesian Statistics*. Springer.

Bertsekas, Dimitir P, & Shreve, Steven. 1996. *Stochastic Optimal Control: The Discrete-Time Case*. New York: Academic Press.

Bertsimas, Dimitris, Farias, Vivek F, & Trichakis, Nikolaos. 2013. Fairness, efficiency, and flexibility in organ allocation for kidney transplantation. *Operations Research*, **61**(1), 73–87.

Bickel, Peter J, & Doksum, Kjell A. 2015. *Mathematical Statistics: Basic Ideas and Selected Topics*. Vol. 2. CRC Press.

Biedermann, Stefanie, Dette, Holger, & Pepelyshev, Andrey. 2006. Some robust design strategies for percentile estimation in binary response models. *Canadian Journal of Statistics*, **34**(4), 603–622.

Billingsley, Patrick. 2013. *Convergence of Probability Measures*. John Wiley & Sons.

Bornkamp, Björn, Bretz, Frank, Dmitrienko, Alex, Enas, Greg, Gaydos, Brenda, Hsu, Chyi-Hung, König, Franz, Krams, Michael, Liu, Qing, Neuenschwander, Beat, *et al.* 2007. Innovative approaches for designing and analyzing adaptive dose-ranging trials. *Journal of Biopharmaceutical Statistics*, **17**(6), 965–995.

Brandeau, Margaret L, Sainfort, François, & Pierskalla, William P. 2004. *Operations Research and Health Care: A Handbook of Methods and Applications*. Vol. 70. Springer Science & Business Media.

Branke, Jürgen, Chick, Stephen E, & Schmidt, Christian. 2007. Selecting a selection procedure. *Management Science*, **53**(12), 1916–1932.

Brealey, Richard A, Myers, Stewart C, Allen, Franklin, & Mohanty, Pitabas. 2012. *Principles of Corporate Finance*. Tata McGraw-Hill Education.

Bretz, Frank, Dette, Holger, & Pinheiro, Jose C. 2010. Practical considerations for optimal designs in clinical dose finding studies. *Statistics in Medicine*, **29**(7-8), 731–742.

Brezzi, Monica, & Lai, Tze Leung. 2002. Optimal learning and experimentation in bandit problems. *Journal of Economic Dynamics and Control*, **27**(1), 87–108.

Brockwell, Anthony E, & Kadane, Joseph B. 2003. A gridding method for Bayesian sequential decision problems. *Journal of Computational and Graphical Statistics*, **12**(3), 566–584.

Brotcorne, Luce, Laporte, Gilbert, & Semet, Frederic. 2003. Ambulance location and relocation models. *European Journal of Operational Research*, **147**(3), 451–463.

Bulayeva, Nataliya N, & Watson, Cheryl S. 2004. Xenoestrogen-induced ERK-1 and ERK-2 activation via multiple membrane-initiated signaling pathways. *Environmental Health Perspectives*, **112**(15), 1481.

Carter, Chris K, & Kohn, Robert. 1994. On Gibbs sampling for state space models. *Biometrika*, 541–553.

Cheng, Yi, & Berry, Donald A. 2007. Optimal adaptive randomized designs for clinical trials. *Biometrika*, **94**(3), 673–689.

Chernoff, Herman. 1961. Sequential tests for the mean of a normal distribution. *Pages 79–91 of: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. University of California Press.

Chick, Stephen E, & Frazier, Peter. 2012. Sequential sampling with economics of selection procedures. *Management Science*, **58**(3), 550–569.

Chick, Stephen E, & Gans, Noah. 2009. Economic analysis of simulation selection problems. *Management Science*, **55**(3), 421–437.

Chick, Stephen E, & Inoue, Koichiro. 2001. New two-stage and sequential procedures for selecting the best simulated system. *Operations Research*, **49**(5), 732–743.

Chick, Stephen E, Forster, Martin, & Pertile, Paolo. 2017. A Bayesian decision theoretic model of sequential experimentation with delayed response. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Chick, Stephen E, Gans, Noah, & Yapar, Ozge. 2018. *Bayesian Sequential Learning for Clinical Trials of Multiple Correlated Medical Interventions*. Accessed: January 2019, https://dx.doi.org/10.2139/ssrn.3184758.

Chow, Shein-Chung. 2003. *Encyclopedia of Biopharmaceutical Statistics*. Informa Health Care.

Chow, YS, Robbins, H, & Siegmund, D. 1971. *Great Expectations: The Theory of Optimal Stopping*. Houghton Mifflin, Boston.

Church, Richard, & ReVelle, Charles. 1974. The maximal covering location problem. *Papers of the Regional Science Association*, **32**(1), 101–118.

de Farias, Daniela, & Van Roy, Benjamin. 2004. On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research*, **29**(3), 462–478.

Dette, Holger, Bretz, Frank, Pepelyshev, Andrey, & Pinheiro, Jose. 2008. Optimal designs for dose-finding studies. *Journal of the American Statistical Association*, **103**(483), 1225–1237.

Dette, Holger, Kiss, Christine, Benda, Norbert, & Bretz, Frank. 2014. Optimal designs for dose finding studies with an active control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**(1), 265–295.

Doerner, Karl F., Gutjahr, Walter J., Hartl, Richard F., Karall, Michaela, & Reimann, Marc. 2005. Heuristic solution of an extended double-coverage ambulance location problem for Austria. *Central European Journal of Operations Research*, **13**, 325–340.

Dwyer-Lindgren, Laura, Bertozzi-Villa, Amelia, Stubbs, Rebecca W, Morozoff, Chloe, Mackenbach, Johan P, van Lenthe, Frank J, Mokdad, Ali H, & Murray, Christopher JL. 2017. Inequalities in life expectancy among U.S. counties, 1980 to 2014: Temporal trends and key drivers. *JAMA Internal Medicine*, **177**(7), 1003–1011.

Edwards, James, Fearnhead, Paul, & Glazebrook, Kevin. 2017. On the identification and mitigation of weaknesses in the knowledge gradient policy for multi-armed bandits. *Probability in the Engineering and Informational Sciences*, **31**(2), 239–263.

FDA. 2017. *The Drug Development Process: Clinical Research.* Accessed: August 2017, https://www.fda.gov/ForPatients/Approvals/Drugs/ucm405622.htm.

Felker, G Michael. 2012. Loop diuretics in heart failure. *Heart Failure Reviews*, **17**(2), 305–311.

Fitch, Jay. 2005. Response times: Myths, measurement, and management. *Journal of Emergency Medical Services*, **30**(9), 47–56.

Frazier, Peter, Powell, Warren, & Dayanik, Savas. 2009. The knowledge-gradient policy for correlated normal beliefs. *INFORMS Journal on Computing*, **21**(4), 599–613.

Frazier, Peter I, & Powell, Warren B. 2011. Consistency of sequential Bayesian sampling policies. *SIAM Journal on Control and Optimization*, **49**(2), 712–731.

Frazier, Peter I, Powell, Warren B, & Dayanik, Savas. 2008. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, **47**(5), 2410–2439.

Frühwirth-Schnatter, Sylvia. 1994. Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, **15**(2), 183–202.

Gadagkar, Sudhindra R, & Call, Gerald B. 2015. Computational tools for fitting the Hill equation to dose–response curves. *Journal of Pharmacological and Toxicological methods*, **71**, 68–76.

Gelman, Andrew, Carlin, John B, Stern, Hal S, & Rubin, Donald B. 2004. *Bayesian Data Analysis*. Chapman & Hall/CRC.

Gendreau, Michel, Laporte, Gilbert, & Semet, Frédéric. 1997. Solving an ambulance location model by tabu search. *Location Science*, **5**(2), 75–88.

Gendreau, Michel, Laporte, Gilbert, & Semet, Frederic. 2001. A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing*, **27**(12), 1641–1653.

Gendreau, Michel, Laporte, Gilbert, & Semet, Frederic. 2006. The maximal expected coverage relocation problem for emergency vehicles. *Journal of the Operational Research Society*, **57**(1), 22–28.

Gibbons, Jean Dickinson, Olkin, Ingram, & Sobel, Milton. 1999. *Selecting and Ordering Populations: A New Statistical Methodology*. SIAM.

Glazebrook, Kevin D. 1979. Stoppable families of alternative bandit processes. *Journal of Applied Probability*, **16**(4), 843–854.

Green, Linda V. 2005. Capacity planning and management in hospitals. *Pages 15–41 of: Operations Research and Health Care*. Springer.

Green, Linda V, Savin, Sergei, & Wang, Ben. 2006. Managing patient service in a diagnostic medical facility. *Operations Research*, **54**(1), 11–25.

Griffin, Robert, Lebovitz, Yeonwoo, English, Rebecca, *et al.* 2010. *Transforming Clinical Research in the United States: Challenges and Opportunities: Workshop Summary*. National Academies Press.

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

Haijema, René, van der Wal, Jan, & van Dijk, Nico M. 2007. Blood platelet production: Optimization by dynamic programming and simulation. *Computers & Operations Research*, **34**(3), 760–779.

Hall, Shane N, Jacobson, Sheldon H, & Sewell, Edward C. 2008. An analysis of pediatric vaccine formulary selection problems. *Operations Research*, **56**(6), 1348–1365.

Hay, Michael, Thomas, David W, Craighead, John L, Economides, Celia, & Rosenthal, Jesse. 2014. Clinical development success rates for investigational drugs. *Nature Biotechnology*, **32**(1), 40–51.

Hee, Siew Wan, Hamborg, Thomas, Day, Simon, Madan, Jason, Miller, Frank, Posch, Martin, Zohar, Sarah, & Stallard, Nigel. 2016. Decision-theoretic designs for small trials and pilot studies: A review. *Statistical Methods in Medical Research*, **25**(3), 1022–1038.

Holland-Letz, Tim, & Kopp-Schneider, Annette. 2015. Optimal experimental designs for dose–response studies with continuous endpoints. *Archives of Toxicology*, **89**(11), 2059–2068.

Holm Hansen, Christian, Warner, Pamela, Parker, Richard A, Walker, Brian R, Critchley, Hilary OD, & Weir, Christopher J. 2017. Development of a Bayesian response-adaptive trial design for the Dexamethasone for excessive menstruation study. *Statistical Methods in Medical Research*, **26**(6), 2681–2699.

Ingber, Lester, Chen, Colleen, Mondescu, Radu Paul, Muzzall, David, & Renedo, Marco. 2001. Probability tree algorithm for general diffusion processes. *Physical Review E*, **64**(5), 056702.

Ingolfsson, Armann. 2013. EMS planning and management. *Pages 105–128 of: Operations Research and Health Care Policy*. Springer.

Jagtenberg, CJ, Bhulai, Sandjai, & van der Mei, RD. 2015. An efficient heuristic for real-time ambulance redeployment. *Operations Research for Health Care*, **4**, 27–35.

Jagtenberg, CJ, Bhulai, Sandjai, & van der Mei, RD. 2016. Dynamic ambulance dispatching: Is the closest-idle policy always optimal? *Health Care Management Science*, 1–15.

Jennison, Christopher, & Turnbull, Bruce W. 1999. *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC.

Jiang, Fei, Jack Lee, J, & Müller, Peter. 2013. A Bayesian decision-theoretic sequential response-adaptive randomization design. *Statistics in Medicine*, **32**(12), 1975–1994.

Khademi, Amin, Saure, Denis R, Schaefer, Andrew J, Braithwaite, Ronald S, & Roberts, Mark S. 2015. The price of nonabandonment: HIV in resource-limited settings. *Manufacturing & Service Operations Management*, **17**(4), 554–570.

Kotas, Jakob, & Ghate, Archis. 2016. Response-guided dosing for rheumatoid arthritis. *IIE Transactions on Healthcare Systems Engineering*, **6**(1), 1–21.

Kotas, Jakob, & Ghate, Archis. 2017. *Optimal Bayesian learning of dose-response parameters from a cohort*. Accessed: March 2019, https://dx.doi.org/10.2139/ssrn.2630392.

Kotas, Jakob, & Ghate, Archis. 2018. Bayesian learning of dose-response parameters from a cohort under response-guided dosing. *European Journal of Operational Research*, **265**(1), 328–343.

Krams, Michael, Lees, Kennedy R, Hacke, Werner, Grieve, Andrew P, Orgogozo, Jean-Marc, Ford, Gary A, *et al.* 2003. Acute stroke therapy by inhibition of neutrophils (ASTIN) an adaptive dose-response study of UK-279, 276 in acute ischemic stroke. *Stroke*, **34**(11), 2543–2548.

Lai, Guoming, Margot, François, & Secomandi, Nicola. 2010. An approximate dynamic programming approach to benchmark practice-based heuristics for natural gas storage valuation. *Operations Research*, **58**(3), 564–582.

Lai, Tze Leung, & Liao, Olivia Yueh-Wen. 2012. Efficient adaptive randomization and stopping rules in multi-arm clinical trials for testing a new treatment. *Sequential Analysis*, **31**(4), 441–457.

Lee, Chris P, Chertow, Glenn M, & Zenios, Stefanos A. 2008. Optimal initiation and management of dialysis therapy. *Operations Research*, **56**(6), 1428–1449.

Lenz, Robert A, Pritchett, Yili L, Berry, Scott M, Llano, Daniel A, Han, Shu, Berry, Donald A, Sadowsky, Carl H, Abi-Saab, Walid M, & Saltarelli, Mario D. 2015. Adaptive dose-finding Phase 2 trial evaluating the safety and efficacy of ABT-089 in mild to moderate Alzheimer disease. *Alzheimer Disease & Associated Disorders*, **29**(3), 192–199.

Liu, Feng, Walters, Stephen J, & Julious, Steven A. 2017. Design considerations and analysis planning of a Phase 2a proof of concept study in rheumatoid arthritis in the presence of possible non-monotonicity. *BMC Medical Research Methodology*, **17**(1), 149.

Maillart, Lisa M, Ivy, Julie Simmons, Ransom, Scott, & Diehl, Kathleen. 2008. Assessing dynamic breast cancer screening policies. *Operations Research*, **56**(6), 1411–1427.

Mason, Andrew James. 2013. Simulation and real-time optimised relocation for improving ambulance operations. *Pages 289–317 of: Handbook of Healthcare Operations Management*. Springer.

Maxwell, Matthew S, Restrepo, Mateo, Henderson, Shane G, & Topaloglu, Huseyin. 2010. Approximate dynamic programming for ambulance redeployment. *INFORMS Journal on Computing*, **22**(2), 266–281.

Maxwell, Matthew S., Cao Ni, Eric, Tong, Chaoxu, Henderson, Shane G., Topaloglu, Huseyin, & Hunter, Susan R. 2014. A bound on the performance of an optimal ambulance redeployment policy. *Operation Reseach*, **62**(5), 1014–1027.

McLay, Laura A, & Mayorga, Maria E. 2013. A model for optimally dispatching ambulances to emergency calls with classification errors in patient priorities. *IIE Transactions*, **45**(1), 1–24.

Mullard, Asher. 2015. Regulators and industry tackle dose-finding issues. *Nature Reviews Drug Discovery*, **14**, 371–372.

Müller, Peter, Berry, Don A, Grieve, Andrew P, & Krams, Michael. 2006. A Bayesian decision-theoretic dose-finding trial. *Decision Analysis*, **3**(4), 197–207.

Müller, Peter, Berry, Don A, Grieve, Andy P, Smith, Michael, & Krams, Michael. 2007. Simulation-based sequential Bayesian design. *Journal of Statistical Planning and Inference*, **137**(10), 3140–3150.

National Association of State EMS Officials. *EMS performance measures: Recommended attributes and indicators for system and service performance.* Accessed: January 2016, https://www.nasemso.org/Projects/PerformanceMeasures/.

National Center for Health Statistics. 2016. *Health, United States, 2016: With Chartbook on Long-term Trends in Health.* Accessed: November 2017, https://www.cdc.gov/nchs/data/hus/hus16.pdf.

National Institutes of Health. 2014. *Notice of Revised NIH Definition of Clinical Trial.* Accessed: January 2016, http://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-015.html.

Negoescu, Diana M, Frazier, Peter I, & Powell, Warren B. 2011. The knowledge-gradient algorithm for sequencing experiments in drug discovery. *INFORMS Journal on Computing*, **23**(3), 346–363.

Negoescu, Diana M, Bimpikis, Kostas, Brandeau, Margaret L, & Iancu, Dan A. 2017. Dynamic learning of patient response types: An application to treating chronic diseases. *Management science*, **64**(8), 3469–3488.

NFPA. 2010. *NFPA 1710, standard for the organization and deployment of fire suppression operations, emergency medical operations, and special operations to the public by career fire departments.* National Fire Protection Association.

O'Brien, Peter C, & Fleming, Thomas R. 1979. A multiple testing procedure for clinical trials. *Biometrics*, **35**(3), 549–556.

Owen, Shawn C, Doak, Allison K, Ganesh, Ahil N, Nedyalkova, Lyudmila, McLaughlin, Christopher K, Shoichet, Brian K, & Shoichet, Molly S. 2014. Colloidal drug formulations can explain "bell-shaped" concentration–response curves. *ACS Chemical Biology*, **9**(3), 777–784.

Patrick, Jonathan, Puterman, Martin L, & Queyranne, Maurice. 2008. Dynamic multipriority patient scheduling for a diagnostic resource. *Operations research*, **56**(6), 1507–1525.

Peskir, Goran, & Shiryaev, Albert. 2006. *Optimal Stopping and Free-Boundary Problems.* Springer.

Pierskalla, William P. 2005. Supply chain management of blood banks. *Pages 103–145 of: Operations research and health care.* Springer.

Powell, Warren B, & Ryzhov, Ilya O. 2012. *Optimal Learning.* Vol. 841. John Wiley & Sons.

Powell, W.B. 2007. *Approximate Dynamic Programming: Solving the Curses of Dimensionality.* Hoboken, New Jersey: Wiley-Interscience.

Press, William H. 2009. Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research. *Proceedings of the National Academy of Sciences*, **106**(52), 22387–22392.

Puterman, M.L. 2005. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* New York, NY: Wiley-Intersceince.

Rais, Abdur, & Viana, Ana. 2011. Operations research in healthcare: A survey. *International Transactions in Operational Research*, **18**(1), 1–31.

Rajagopalan, Hari K, Saydam, Cem, & Xiao, Jing. 2008. A multiperiod set covering location model for dynamic redeployment of ambulances. *Computers & Operations Research*, **35**(3), 814–826.

Rasmussen, Carl Edward, & Williams, Christopher KI. 2006. *Gaussian Processes for Machine Learning.* Vol. 1. MIT press Cambridge.

Restrepo, Mateo, Henderson, Shane G, & Topaloglu, Huseyin. 2009. Erlang loss models for the static deployment of ambulances. *Health Care Management Science*, **12**(1), 67–79.

Rojas-Cordova, Alba C, & Hosseinichimeh, Niyousha. 2018. Trial termination and drug misclassification in sequential adaptive clinical trials. *Service Science*, **10**(3), 354–377.

Rosenberger, William F. 1996. New directions in adaptive designs. *Statistical Science*, 137–149.

Roy, Avik S. A. 2012. *Stifling New Cures: The True Cost of Lengthy Clinical Drug Trials*. Accessed: January 2016, http://www.manhattan-institute.org/html/fda_05.htm.

Ryzhov, Ilya O. 2018. The Local Time Method for Targeting and Selection. *Operations Research*, **66**(5), 1406–1422.

Ryzhov, Ilya O, & Powell, Warren B. 2011a. Information collection on a graph. *Operations Research*, **59**(1), 188–201.

Ryzhov, Ilya O, & Powell, Warren B. 2011b. The value of information in multi-armed bandits with exponentially distributed rewards. *Procedia Computer Science*, **4**, 1363–1372.

Ryzhov, Ilya O, Frazier, Peter I, & Powell, Warren B. 2010. On the robustness of a one-period look-ahead policy in multi-armed bandit problems. *Procedia Computer Science*, **1**(1), 1635–1644.

Ryzhov, Ilya O, Powell, Warren B, & Frazier, Peter I. 2012. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*, **60**(1), 180–195.

Sacks, Leonard V, Shamsuddin, Hala H, Yasinskaya, Yuliya I, Bouri, Khaled, Lanthier, Michael L, & Sherman, Rachel E. 2014. Scientific and regulatory reasons for delay and denial of FDA approval of initial applications for new drugs, 2000-2012. *JAMA*, **311**(4), 378–384.

Schaefer, Andrew J, Bailey, Matthew D, Shechter, Steven M, & Roberts, Mark S. 2005. Modeling medical treatment using Markov decision processes. *Pages 593–612 of: Operations research and health care*. Springer.

Schmid, Verena. 2012. Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *European Journal of Operational Research*, **219**(3), 611–621.

Schneider, Eric C., Sarnak, Dana O., Squires, David, Shah, Arnav, & Doty, Michelle M. 2017. *Mirror, Mirror 2017: International Comparison Reflects Flaws and Opportunities for Better U.S. Health Care*. Accessed: November 2017, http://www.commonwealthfund.org/interactives/2017/july/mirror-mirror/.

Sherman, Jack, & Morrison, Winifred J. 1950. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, **21**(1), 124–127.

Singh, Satinder, Jaakkola, Tommi, Littman, Michael L, & Szepesvári, Csaba. 2000. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, **38**(3), 287–308.

Smith, Michael K, Jones, Ieuan, Morris, Mark F, Grieve, Andrew P, & Tan, Keith. 2006. Implementation of a Bayesian adaptive design in a proof of concept study. *Pharmaceutical Statistics*, **5**(1), 39–50.

Snapinn, Steven, Chen, Mon-Gy, Jiang, Qi, & Koutsoukos, Tony. 2006. Assessment of futility in clinical trials. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, **5**(4), 273–281.

Stallard, Nigel, Whitehead, John, Todd, Susan, & Whitehead, Anne. 2001. Stopping rules for phase II studies. *British Journal of Clinical Pharmacology*, **51**(6), 523–529.

Sudtachat, Kanchala, Mayorga, Maria E, & Mclay, Laura A. 2016. A nested-compliance table policy for emergency medical service systems under relocation. *Omega*, **58**, 154–168.

Swersey, Arthur J. 1994. The deployment of police, fire, and emergency medical units. *Handbooks in Operations Research and Management Science*, **6**, 151–200.

Tufts. 2014. *Cost to Develop and Win Marketing Approval for a New Drug Is $2.6 Billion*. Accessed: August 2017, http://csdd.tufts.edu/news/complete_story/pr_tufts_csdd_2014_cost_study.

United States Census Bureau. 2014. *Mecklenburg County, North Carolina quick facts*. Accessed: January 2016, http://quickfacts.cencus.gov/qfd/states/37/37119.html.

van Barneveld, TC, Bhulai, S, & van der Mei, RD. 2016. The effect of ambulance relocations on the performance of ambulance service providers. *European Journal of Operational Research*, **252**(1), 257–269.

van Barneveld, Thije. 2016. The minimum expected penalty relocation problem for the computation of compliance tables for ambulance vehicles. *INFORMS Journal on Computing*, **28**(2), 370–384.

van den Berg, Pieter L, van Essen, J Theresia, & Harderwijk, Eline J. 2016. Comparison of static ambulance location models. *Pages 1–10 of: 2016 3rd International Conference on Logistics Operations Management (GOL)*. IEEE.

Van Roy, Benjamin, Bertsekas, Dimitri P, Lee, Yuchun, & Tsitsiklis, John N. 1997. A neuro-dynamic programming approach to retailer inventory management. *Pages 4052–4057 of: Decision and Control, 1997., Proceedings of the 36th IEEE Conference on*, vol. 4. IEEE.

Villar, Sofía S, & Rosenberger, William F. 2018. Covariate-adjusted response-adaptive randomization for multi-arm clinical trials using a modified forward looking Gittins index rule. *Biometrics*, **74**(1), 49–57.

Villar, Sofía S, Bowden, Jack, & Wason, James. 2015. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical Science*, **30**(2), 199.

Wang, Jixian. 2006. Optimal parametric design with applications to pharmacokinetic and pharmacodynamic trials. *Journal of Applied Statistics*, **33**(8), 837–852.

Wang, Yingfei, Wang, Chu, & Powell, Warren. 2016. The knowledge gradient for sequential decision making with stochastic binary feedbacks. *Pages 1138–1147 of: International Conference on Machine Learning*.

Ward, Michael J. 2014. Saving more lives? *Journal of Emergency Medical Services*, **39**(2), 46–53.

Warner, P, Weir, CJ, Hansen, CH, Douglas, A, Madhra, M, Hillier, SG, Saunders, PTK, Iredale, JP, Semple, S, Walker, BR, *et al.* 2015. Low-dose dexamethasone as a treatment for women with heavy menstrual bleeding: protocol for response-adaptive randomised placebo-controlled dose-finding parallel group trial (DexFEM). *BMJ Open*, **5**(1), e006837.

Weir, Christopher J, Spiegelhalter, David J, & Grieve, Andrew P. 2007. Flexible design and efficient implementation of adaptive dose-finding studies. *Journal of Biopharmaceutical Statistics*, **17**(6), 1033–1050.

West, M. J., & Harrison, P. J. 1997. *Byesian Forecasting and Dynamic Models*. Springer-Verlag.

Whitehead, John. 1997. *The Design and Analysis of Sequential Clinical Trials*. John Wiley & Sons.

Wilde, Elizabeth Ty. 2013. Do emergency medical system response times matter for health outcomes? *Health Economics*, **22**(7), 790–806.

World Health Organization. 2000. *The World Health Report 2000*. Accessed: November 2017, http://www.who.int/whr/2000/en/whr00_en.pdf.

World Health Organization. 2015. *Global Health Observatory Data: Life Expectency*. Accessed: November 2017, http://www.who.int/gho/mortality_burden_disease/life_tables/situation_trends_text/en/.

Wu, Joseph T, Wein, Lawrence M, & Perelson, Alan S. 2005. Optimization of influenza vaccine selection. *Operations Research*, **53**(3), 456–476.

Xie, Jing, Frazier, Peter I, & Chick, Stephen E. 2016. Bayesian optimization via simulation with pairwise sampling and correlated prior beliefs. *Operations Research*, **64**(2), 542–559.

Zaric, Gregory S, & Brandeau, Margaret L. 2001. Resource allocation for epidemic control over short time horizons. *Mathematical Biosciences*, **171**(1), 33–58.

Zaric, Gregory S, & Brandeau, Margaret L. 2002. Dynamic resource allocation for epidemic control in multiple populations. *Mathematical Medicine and Biology*, **19**(4), 235–255.

Zenios, Stefanos A, Chertow, Glenn M, & Wein, Lawrence M. 2000. Dynamic allocation of kidneys to candidates on the transplant waiting list. *Operations Research*, **48**(4), 549–569.

Zhang, Wei, Sargent, Daniel J, & Mandrekar, Sumithra. 2006. An adaptive dose-finding design incorporating both toxicity and efficacy. *Statistics in Medicine*, **25**(14), 2365–2383.