

 Open access • Journal Article • DOI:10.1002/SIM.3720

## Reinforcement learning design for cancer clinical trials — Source link

Michael R. Kosorok, Yufan Zhao

**Institutions:** University of North Carolina at Chapel Hill

**Published on:** 01 Jan 2009 - Cancer clinical trials (NIH Public Access)

**Topics:** Dynamic treatment regime and Reinforcement learning

Related papers:

- [Optimal dynamic treatment regimes](#)
- [Estimating Individualized Treatment Rules Using Outcome Weighted Learning](#)
- [Optimal Structural Nested Models for Optimal Sequential Decisions](#)
- [A Robust Method for Estimating Optimal Treatment Regimes](#)
- [An experimental design for the development of adaptive treatment strategies](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/reinforcement-learning-design-for-cancer-clinical-trials-3l8sy2spwc>

# REINFORCEMENT LEARNING DESIGN FOR CANCER CLINICAL TRIALS

by  
**Yufan Zhao**

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics, School of Public Health.

Chapel Hill  
2009

Approved by:

Michael R. Kosorok

Jason P. Fine

Yufeng Liu

Mark A. Socinski

Donglin Zeng

©2009  
Yufan Zhao  
ALL RIGHTS RESERVED

## ABSTRACT

YUFAN ZHAO: Reinforcement Learning Design for Cancer Clinical Trials

(Under the direction of Dr. Michael Kosorok)

There has been significant recent research activity in developing therapies that are tailored to each individual. Finding such therapies in treatment settings involving multiple decision times is a major challenge. In this dissertation, we develop reinforcement learning trials for discovering these optimal regimens for life-threatening diseases such as cancer. A temporal-difference learning method called  $Q$ -learning is utilized which involves learning an optimal policy from a single training set of finite longitudinal patient trajectories. Approximating the  $Q$ -function with time-indexed parameters can be achieved by using support vector regression or extremely randomized trees. Within this framework, we demonstrate that the procedure can extract optimal strategies directly from clinical data without relying on the identification of any accurate mathematical models, unlike approaches based on adaptive design. We show that reinforcement learning has tremendous potential in clinical research because it can select actions that improve outcomes by taking into account delayed effects even when the relationship between actions and outcomes is not fully known.

To support our claims, the methodology's practical utility is firstly illustrated in a virtual simulated clinical trial. We then apply this general strategy with significant refinements to studying and discovering optimal treatments for advanced metastatic stage IIIB/IV non-small cell lung cancer (NSCLC). In addition to the complexity of the NSCLC problem of selecting optimal compounds for first and second-line treatments based on prognostic factors, another primary scientific goal is to determine the optimal time to ini-

tiate second-line therapy, either immediately or delayed after induction therapy, yielding the longest overall survival time. We show that reinforcement learning not only successfully identifies optimal strategies for two lines of treatment from clinical data, but also reliably selects the best initial time for second-line therapy while taking into account heterogeneities of NSCLC across patients.

## ACKNOWLEDGMENTS

This dissertation could not have been written without my advisor, Dr. Michael Kosorok, who led me to this research field and patiently guided me through the dissertation process. I wish to express my deepest appreciation to him for his support, encouragement and mentoring throughout my doctoral studies.

I also would like to thank the rest of my committee members. I owe many thanks to Dr. Donglin Zeng for his encouraging inspirations, kind guidance, and enormously helpful discussions in completing this dissertation. I am deeply grateful to Dr. Mark Socinski for his in-depth knowledge of non-small cell lung cancer and insightful discussions on many occasions. I wish to express my sincere thanks to Dr. Jason Fine for his invaluable advice and kindness throughout this research. I also appreciate very much discussions with Dr. Yufeng Liu; working with him has been a wonderful learning experience for me.

I was very fortunate to have the opportunity to work at SAS Institute under the direction of Drs. Bob Rodriguez and Ying So. A special thanks goes to them for their financial support of my graduate studies and research.

I would also like to thank all of my friends in the Reinforcement Learning Group, particularly Kai Ding, Yiyun Tang, and Yingqi Zhao, whose friendship has made this journey more enjoyable and memorable.

Finally, it's impossible to have completed this journey without the love, support and encouragement from my dad, Dongtai Zhao, and my mom, Jun Xia. This dissertation is dedicated to them.

# CONTENTS

<b>LIST OF TABLES</b>	<b>viii</b>
<b>LIST OF FIGURES</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Literature Review</b>	<b>7</b>
2.1 Adaptive Design and Dynamic Treatment in Clinical Trials . . . . .	7
2.1.1 Adaptive Design . . . . .	7
2.1.2 Dynamic Treatment Regimes . . . . .	9
2.2 Optimal Controls for Drug Scheduling . . . . .	11
2.2.1 Mathematical Models for Cancer Treatment . . . . .	11
2.2.2 Finding Optimal Treatment Solutions . . . . .	17
<b>3 Reinforcement Learning, <math>Q</math>-Learning, and Their Approximations</b>	<b>20</b>
3.1 Introduction . . . . .	20
3.2 Reinforcement Learning and $Q$ -Learning . . . . .	22
3.2.1 Value Functions and the Bellman Equation . . . . .	24
3.2.2 Temporal-Difference Learning and $Q$ -Learning . . . . .	27
3.3 Support Vector Machines (SVM) . . . . .	30
3.4 Support Vector Regression (SVR) . . . . .	36
3.5 Extremely Randomized Trees (ERT) . . . . .	38
<b>4 Reinforcement Learning Treatment Strategies for A Virtual Cancer Trial</b>	<b>45</b>

4.1	Clinical Reinforcement Trials . . . . .	45
4.2	A Virtual Clinical Reinforcement Trial . . . . .	47
4.2.1	A Simple Chemotherapy Mathematical Model . . . . .	48
4.2.2	$Q$ -function Estimation and Optimal Regimen Discovery . . . . .	50
4.2.3	Simulation Results . . . . .	53
4.2.4	Summary of Virtual Cancer Trial Results . . . . .	55
<b>5</b>	<b>Reinforcement Learning Treatment Strategies Based on Support Vector Regression in a Non-small Cell Lung Cancer Trial</b>	<b>66</b>
5.1	Introduction . . . . .	66
5.2	Reinforcement Learning Model Refinement . . . . .	71
5.2.1	Patient Outcomes . . . . .	71
5.2.2	$Q$ -Learning Revisited . . . . .	73
5.3	Support Vector Regression for Censored Subjects . . . . .	75
5.4	Clinical Reinforcement Trial Conduct and Computational Strategy . . . . .	78
5.5	Simulation Study . . . . .	80
5.5.1	Data Generating Models . . . . .	81
5.5.2	Clinical Scenarios . . . . .	82
5.5.3	Simulation Methods and Results . . . . .	84
5.6	Summary of NSCLC trial results . . . . .	85
<b>6</b>	<b>Concluding Remarks</b>	<b>95</b>
6.1	Overview . . . . .	95
6.2	Future Research . . . . .	96
	<b>Bibliography</b>	<b>99</b>



# LIST OF TABLES

1	Summary of main simulation results for a general virtual cancer trial . . .	59
2	The scenarios studied in simulation study of an NSCLC virtual cancer trial	90
3	Comparisons between true optimal regimens and estimated optimal regimens for overall survival . . . . .	92

# LIST OF FIGURES

1	Helicopter in autonomous sustained hover . . . . .	40
2	Linear separating hyperplane, margin, and support vectors defined in SVM	41
3	Hinge loss function for SVM . . . . .	42
4	$\epsilon$ -insensitive loss function for SVR . . . . .	43
5	Procedure used by the ERT algorithm to build a tree . . . . .	44
6	One simple chemotherapy model . . . . .	57
7	Treatment plan and the procedure for estimating $Q$ -functions . . . . .	58
8	Plots of averaged value of “wellness” . . . . .	60
9	Plots of averaged value of “tumor size” . . . . .	61
10	Plots of averaged value of “wellness + tumor size” . . . . .	62
11	Example of the estimated optimal treatment for one patient . . . . .	63
12	Bar plots of averaged cumulative survival probability . . . . .	64
13	The averaged optimal sequential therapies . . . . .	65
14	Treatment plan and therapy options for an advanced NSCLC trial . . . . .	87
15	Four cases that determine the times $T_1$ , $C$ , $T_M$ , and $T_2$ . . . . .	88
16	Loss functions for $\epsilon$ -SVR-C . . . . .	89
17	Performance of optimal individualized regimens . . . . .	91
18	Sensitivity of the predicted survival to the sample size . . . . .	93
19	Boxplots of the predicted survival computed via $Q$ -learning with $\epsilon$ -SVR-C	94

# 1 Introduction

Discovering effective therapeutic regimens for life-threatening diseases is one of the central goals of medical research. Finding powerful and general methodologies for accomplishing this discovery is a major challenge. The prevailing approach is to develop candidate therapies in the laboratory using basic science and then to test those therapies in animals and then in human clinical trials. A major problem is that very few candidate treatments make it to human clinical trials and only about 10% of treatments making it to human clinical trials demonstrate enough efficacy to be approved for marketing (Hogberg, 2005; Food and Drug Administration, 2004). Typical regimens for some advanced cancer (such as breast cancer, lung cancer, and ovarian cancer) patients utilize a single agent in combination with some platinum-based compound, and consist of multiple stages of treatment (especially when relapse is common). For example, many studies demonstrate that three lines of treatment can improve survival for patients with advanced non-small cell lung cancer (NSCLC). For patients who present with a good performance status and stage IIIB/IV disease, platinum-based chemotherapy is the primary treatment which can offer a modest survival advantage over best supportive care alone. Approximately 50–60% of patients in recent first-line trials received second-line treatment (Sandler et al., 2006). Some patients who maintain a good performance status and tolerate therapy without significant toxicities will receive third-line therapy (Stinchcombe and Socinski, 2008).

A widely used approach is to give a maximum dosage of chemotherapy drug for some period of time, followed by a period of recuperation in which no drug is given. Although this therapeutic regimen can be easily clinically implemented, this may not be the best strategy for minimizing tumor burden. Such problems have motivated the vast literature

on drug-scheduling strategies. In the past few years, there has been extensive research on applications of adaptive design to clinical trials. Many investigators have developed various adaptive designs to efficiently identify clinical benefits of the treatment, and demonstrated that conducting adaptive designs can be very promising in clinical development. In general, adaptive designs for multiple courses of chemotherapy allow modification of randomization schedules based on varied probabilities of treatment assignment in order to increase the probability of success. In choosing treatments for successive courses, one of the popular adaptive designs to do this is the play-the-winner-and-drop-the-loser design, which is to repeat a treatment that is successful in a given course and otherwise switch to a different treatment. Thall, Millikan, and Sung (2000) provided a statistical framework for multi-course clinical trials involving some modifications of the play-the-winner-and-drop-the-loser strategy. In their proposed design, all treatments after the first course are assigned adaptively, thus increasing the amount of information available per patient. Thall et al. (2007) presented a Bayesian adaptive design for a trial comparing two-course strategies for treating metastatic renal cancer. Each patient is fairly randomized between two treatments at enrollment, and if a patient suffers a disease progression (s)he is then re-randomized among three treatments not given initially. One of the common features of these adaptive designs is the use of parametric models accounting for efficacy, toxicity, or time to some events (such as survival time). By defining a probability model, it's easy to study the design's operating characteristics under a range of parameterizations and clinical scenarios. However, as a result, it will lead to all individuals being assigned to the same level and type of treatment. Therefore, the limitation is not only to ignore the heterogeneity in treatment across individuals, but also to unsuccessfully incorporate the heterogeneity needed for optimal individualized treatment across time.

In addition to the challenge of taking into account accrued information in clinical trial designs, another major challenge is the examination of the long-term benefit of treatment due to delayed effects. If we consider the larger context of the overall therapeutic strategy,

in many clinical settings a regimen with a lower initial response rate still can be the best choice in the long run. This is quite plausible due to the potential for the regimen’s comparatively better delayed clinical benefit. For finding new treatment regimens with this motivation, one of the most promising approaches has been referred to variously as “dynamic treatment regimes” or “adaptive treatment strategies” (Murphy, 2005a). In contrast with classic adaptive designs, dynamic treatment regimes can allow dosage level and type to vary with time for subject-specific needs. As a consequence, the optimal strategy is able to provide information not only on the best treatment choice from the beginning but also treatment choices that maximize outcomes for a later time. Dynamic treatment regimes are recently emerging as a new paradigm for the treatment and long term management of chronic disease, and they have been utilized in some trials such as sequential multiple assignment randomized trials (SMART) (Murphy, 2005a) and drug and alcohol dependency studies (Murphy et al., 2007a). However, to date, there are no clinical trial methodologies for discovering new treatment regimens for life-threatening diseases. Thus, for diseases like cancer, the use of clinical trials for evaluation and not discovery remains the prevailing paradigm.

Over the last few decades, machine learning has become an active branch of artificial intelligence. Some of the fields studied in machine learning involve stochastic sequential decision processes, commonly referred to as reinforcement learning methods. The term “reinforcement” comes from studies of animal learning in experimental psychology, where it refers to the occurrence of an event, in the proper relation to a response, that tends to increase the probability that the response will occur again in the same situation. The standard reinforcement learning method considers a performance agent operating in discrete time, observing at time  $t$  the environmental state  $x_t$ , taking an action  $a_t$ , and receiving back information from the environment (the next state  $x_{t+1}$  and the instantaneous reward  $r_t$ ). The basic process of reinforcement learning involves trying a sequence  $a_t$  of actions, recording the consequences  $r_t$  of those actions, statistically estimating the

relationship between  $a_t$  and  $r_t$ , and then choosing the action that results in the most desirable consequence.

In this dissertation, we present a general reinforcement learning framework and related statistical and computational methods for use in the clinical research arena. Reinforcement learning has been applied to treating behavioral disorders, where each patient typically has multiple opportunities to try different treatments (Pineau et al., 2007). Murphy et al. (2007b) suggest  $Q$ -learning, which is one of the most important breakthroughs in reinforcement learning, for constructing decision rules for chronic psychiatric disorders, since these chronic conditions often require sequential decision making to achieve the best clinical outcomes. Moreover, reinforcement learning has been successfully applied to the segmentation of the prostate in transrectal ultrasound images. Due to its use of knowledge obtained from the previous input image, the reinforcement learning algorithm is potentially capable of finding the appropriate local value for sub-images and extracting the prostate image (Sahba, Tizhoosh, and Salama, 2008). However, reinforcement learning has not yet been applied to life-threatening diseases like cancer where individual patients do not have the luxury to try many different treatments. Our main aim is to illustrate the application of these methods to the discover of new treatment regimens for life-threatening diseases such as cancer. This is a paradigm shift from the standard clinical trial framework which is used for evaluating treatments but not for discovery. We consider trials in which each patient is randomized among a set of treatments at each stage and this treatment set consists of a continuous range of possibilities including, for example a continuous range of dose levels. Therefore, rather than being constrained to a finite list of pre-specified treatments, our method allows for more general multiplicities of treatments which may include a continuum of possibilities at each stage. Reinforcement learning design has two attractive features that make it a useful tool for extracting optimal strategies directly from clinical data. First, without relying on the identification of any accurate mathematical models, it carries out treatment selection sequentially

with time-dependent outcomes to determine which of several possible next treatments is best for which patients at each decision time. This feature not only helps us account for heterogeneity in treatment across individuals, but also possibly captures the best individualized therapies even when the relationship between treatments and outcomes is not fully known. Secondly, in contrast to focusing on short-term benefits, the proposed approach improves longer-term outcomes by considering delayed effect of treatments. Furthermore, we find that reinforcement learning design can extract the optimal treatment strategies while taking into account a drug’s efficacy and toxicity simultaneously, which is supported by our simulation studies.

The remainder of this dissertation is organized as follows. In Section 2.1, we provide a literature review of clinical trial design with particular attention given to adaptive design and dynamic treatment regimes. We review mathematical models which use optimal control theory to seek the solution in cancer treatments in Section 2.2.

In Section 3.2, we provide a detailed description of reinforcement learning and  $Q$ -learning. In Section 3.3, we first describe a support vector machine (SVM) method which makes fitting  $Q$ -functions feasible for clinical data sets; and then we discuss one of the extensions of SVM, support vector regression (SVR), associated with its application to reinforcement learning, in Section 3.4. Another modern technique for estimating  $Q$ -functions, extremely randomized trees (ERT), is presented in Section 3.5.

In Section 4.1, we first propose to develop a new design and analysis method that utilizes this special technology for a new kind of clinical trial for cancer, “clinical reinforcement trials”. To demonstrate the reinforcement learning’s potential in discovering optimal therapies, in Section 4.2, we apply our proposed method to a virtual randomized sequential trial, which is a simulation study consisting of 1000 patients. This study examines the performance of reinforcement learning via SVR and demonstrates that the therapy found using  $Q$ -learning is superior to any constant-dose regimen.

In Chapter 5, we specialize our overall approach to advanced metastatic stage IIIB/IV

NSCLC. By studying an extensive simulation, we refine our model to identify optimal two-line treatment strategies for an NSCLC trial that includes right censored patients. In addition, we demonstrate that our method can reliably select the best time to initiate second-line therapy for NSCLC.

Finally, we summarize our proposed methods in Chapter 6 and discuss some challenges for future research.



## **2 Literature Review**

### **2.1 Adaptive Design and Dynamic Treatment in Clinical Trials**

#### **2.1.1 Adaptive Design**

Due to steeply rising drug development costs and escalating patient safety concerns, there is increasing pressure on pharmaceutical companies and clinical researchers to reexamine traditional clinical trial techniques and increase the efficiency and safety of the clinical trial process. One potential way to address the challenges that are receiving significant attention from pharmaceutical companies, regulatory agencies, and clinical researchers involved is through adaptive designs. In recent years, the use of adaptive design methods in clinical research and development based on accrued data has become very popular due to its flexibility and efficiency. For instance, in 2006, the United States Food and Drug Administration (FDA) released a Critical Path Opportunities List that calls for advancing innovative trial designs, especially for the use of prior experience or accumulated information in trial design. This shows the encouragement for the use of innovative adaptive design methods in clinical trials and the potential use of other approaches in clinical research and development, such as Bayesian approaches in phase II/III studies.

Based on the review of interim data, it is not uncommon to modify a trial or statistical procedures in the middle of the conduct of clinical trials. The purpose is not only to efficiently identify clinical benefits of the test treatment under investigation, but also to increase the probability of success of clinical development. An adaptive design is defined

as a design that allows adaptations to design and statistical procedures of the trial after its initiation without undermining the validity and integrity of the trial (Chow, Chang, and Pong, 2005). Many recent publications refer to an adaptive design as a clinical trial design that uses accumulating data to decide on how to modify aspects of the study as it continues, without compromising the scientific method (Gallo et al., 2006).

Commonly considered adaptive design methods in clinical trials include, but are not limited to: adaptive randomization design, group sequential design, sample size re-estimation design, play-the-winner-and-drop-the-loser design, adaptive dose finding design, adaptive treatment-switching design, hypothesis-adaptive design, and adaptive seamless phase II/III trial design. As we mentioned earlier in Chapter 1, we concentrate on the patient's treatment which often involves multiple courses of chemotherapy. In choosing treatments for successive courses, the design common to this is the play-the-winner-and-drop-the-loser design, which is to repeat a treatment that is successful in a given course and otherwise switch to a different treatment. Thall et al. (2000) provided a statistical framework for multi-course clinical trials involving some modifications of the play-the-winner-and-drop-the-loser strategy. In their proposed design, all treatments after the first course are assigned adaptively, thus increasing the amount of information available per patient.

Most adaptive designs for multiple courses of chemotherapy allow modification of randomization schedules based on varied and/or unequal probabilities of treatment assignment in order to increase the probability of success. For instance, a randomized two-course, three-treatment acute leukemia trial with adaptive randomization has been developed in a Bayesian framework by Thall, Sung, and Estey (2002). A simulation study with the goal of selecting one best treatment, or selecting a best ordered pair of treatments has been investigated by Thall et al. (2000). In addition, in a lymphocyte infusion trial (Thall, Inoue, and Martin, 2002), an adaptive decision process was evaluated by determining the infusion time that has the highest probability of treatment success. One

of the common features of all of these adaptive designs is the use of parametric models accounting for efficacy, toxicity, or time to some events (such as survival time). By defining a probability model, it is easy to study the design’s operating characteristics under a range of parameterizations and clinical scenarios; however, all of these approaches will result in all individuals being assigned to the same level and type of treatment. Therefore, the limitation of these approaches is not only to ignore the heterogeneity in treatment across individuals, but also to unsuccessfully incorporate the heterogeneity needed for optimal individualized treatment across time. Another common feature of all of these adaptive designs is the use of accrued information. These designs choose all treatments after the first one adaptively based on the patient’s outcomes in earlier courses, and thus they don’t waste important information from previous patients. However, in some cases, treatments may show not only an immediate effect to patients but also a delayed effect to patients over time. To date, there are no adaptive designs for incorporating delayed effect in sequential decisions. The lack of these two characteristics in adaptive designs is the most important motivation for our proposed reinforcement learning design.

### **2.1.2 Dynamic Treatment Regimes**

Dynamic treatment regimes, which are also called adaptive treatment strategies (Murphy, 2005a), are recently emerging as a new paradigm for the treatment and long term management of chronic disease. In contrast with classic adaptive design, dynamic treatment regimes can allow dosage level and type to vary with time for subject-specific needs. Dynamic treatment regimes have been conducted in some trials such as sequential multiple assignment randomized trials (SMART) (Murphy, 2005a) and drug and alcohol dependency studies (Murphy et al., 2007a). Murphy (2003) provided a method for estimating optimal decision rules which will produce the optimal mean response at the end of the time period for each individual. Robins (2004) proposed models and developed methods

for making inference about the optimal regime in a multiple courses trial as well.

Using the notation of Murphy (2005a), let  $a_1, a_2, \dots, a_k$  be defined as a sequence of  $k$  treatment decisions for each individual patient at time  $t$  in  $\{1, 2, \dots, k\}$ .  $S_j$  denotes the patient's status at the beginning of the time interval  $j$ , in other words, it is the intermediate outcome available after decision  $a_{j-1}$  and prior to decision  $a_j$ . The response at the end of the time period is denoted by  $Y$ . Thus, the order of event occurrence is  $S_1, a_1, S_2, a_2, \dots, S_k, a_k, S_{k+1}, Y$ . Additionally, for convenience, we use a bar sign over a variable to denote that variable and all past values of the same variable, for example,  $\bar{S}_j = \{S_1, S_2, \dots, S_j\}$  and  $\bar{a}_j = \{a_1, a_2, \dots, a_j\}$ . Also, an adaptive treatment strategy is a sequence of decision rules, denoted as  $d_1, d_2, \dots, d_k$ . It is important to recognize that each rule  $d_j$  is based on the information available at time  $j$ , that is,  $\bar{S}_j, \bar{a}_{j-1}$ , and  $a_j$ . In many cases, the backward induction framework (from dynamic programming) is used to find the optimal decision rules by maximizing mean response for each time point. Formally, in the simplest case, when  $k = 2$ , the optimal adaptive treatment strategy is given by  $(d_1^*, d_2^*)$ , where

$$d_2^*(\bar{s}_2, a_1) = \arg \max_{a_2} E_{\bar{a}_2}[Y | \bar{S}_2 = \bar{s}_2].$$

If we define

$$V_2(\bar{s}_2; a_1) = \max_{a_2} E_{\bar{a}_2}[Y | \bar{S}_2 = \bar{s}_2],$$

then the optimal decision at time 1 could be expressed as follows:

$$d_1^*(s_1) = \arg \max_{a_1} E_{a_1}[V_2(\bar{S}_2; a_1) | S_1 = s_1].$$

Again, if denote

$$V_1(s_1) = \max_{a_1} E_{\bar{a}_1}[V_2(\bar{S}_2; a_1) | S_1 = s_1],$$

then the mean of  $Y$  when the optimal rules  $(d_1^*, d_2^*)$  are used to assign treatment is given by

$$E[V_1(S_1)] = E \left[ \max_{a_1} E_{a_1} \left[ \max_{a_2} E_{\bar{a}_2}[Y | \bar{S}_2] | S_1 \right] \right],$$

which is consistent with the expectation formula in Robins's paper. Murphy (2003) proposed semiparametric methods for estimating the optimal rules through the available experimental or observational longitudinal data, when the multivariate distribution of  $(\bar{S}_k, Y)$  is unknown. The parametric part was used to estimate those optimal rules by modeling the regret function, while the second part consisting of high or infinite dimensional parameters was modelled as a collection of nuisance parameters. Following Murphy's first approach, Robins (2004) investigated a number of estimating equations for finding optimal decision rules using structural nested mean models. Moodie, Richardson, and Stephens (2007) showed that Murphy's approach and Robins's are closely related, further more, Murphy's model is a special case of Robins's.

One of the most important advantages of these dynamic treatment regimes is the consideration of treatment delayed effects to patients. An optimal rule provides information not only on the best treatment choice from the beginning but also treatment choices that maximize outcomes for a later time. The ascertainment of the optimal adaptive treatment strategy is an optimization problem receiving attention by many researchers. Dynamic programming combined with computational methodology is one of the most promising approaches for finding optimal decisions. In particular, we will introduce in Chapter 3 reinforcement learning and its application to deal with this optimal individual-based regime finding problem.

## **2.2 Optimal Controls for Drug Scheduling**

### **2.2.1 Mathematical Models for Cancer Treatment**

Modern treatment methods for cancer include improved traditional surgery, chemotherapy and radiotherapy as well as immunotherapy. Modelling of the treatment process is

viewed as a potentially powerful tool in the development of improving treatment regimens. While biomedical research concentrates on the development of new drugs and experimental (in vitro) and clinical (in vivo) determinations of their treatment schedules, analysis of mathematical models can assist in testing various treatment strategies and searching for optimal ones.

Mathematical models for cancer chemotherapy treatments have a long history and have attracted extensive research over the past several decades. Several approaches to modelling chemotherapeutic induced cell-kill (killing of tumor cells) have been developed. One of the early approaches was by Schabel, Skipper, and Wilcox (1964) who proposed that cell-kill due to a chemotherapeutic drug was proportional to the tumor population. It states that for a fixed dose, the reduction of large tumors occurred more rapidly than for smaller tumors. Skipper's concept is referred to as the log-kill mechanism. Mathematically, the general form of the model under investigation is depicted by the differential equation:

$$\dot{N}(t) = rN(t)F(N) - G(N(t), t),$$

where  $N$  is the tumor size,  $r$  is the growth rate of the tumor,  $F(N)$  is the generalized growth function. For Skipper's model, Gompertzian growth is applied:

$$F(N) = \ln\left(\frac{\Theta}{N}\right).$$

And the function  $G(N(t), t)$  is the cell kill term, describing the pharmacokinetic (PK) and pharmacodynamic (PD) effects of the drug on the system. In Skipper's log-kill (i.e., percentage kill) hypothesis,

$$G(N, t) = \delta u(t)N,$$

where  $\delta$  is the magnitude of the dose and the control, and  $u(t)$  describes the time dependent pharmacokinetics of the drug. In some diseases, for example, Hodgkin's disease and acute lymphoblastic leukemia, Norton, and Simon (1977; 1986) found Skipper's model to be inconsistent with clinical observations. The reduction in large tumors was slower than

in histologically similar smaller tumors. Therefore, Norton and Simon hypothesize that the cell-kill is proportional to the growth rate (e.g., exponential, logistic, or Gompertz) of the tumor. In Norton-Simon's hypothesis,

$$G(N, t) = \delta u(t)F(N).$$

A third hypothesis notes that some chemotherapeutic drugs must be metabolized by an enzyme before being activated. This reaction is saturable due to the fixed amount of enzyme. Thus, Holford, and Sheiner (1981) developed the  $E_{max}$  model which describes cell-kill in terms of a saturable function of Michaelis-Menton form. In the  $E_{max}$  model,

$$G(N, t) = \frac{\delta u(t)N}{K + N}.$$

The model considered by Matveev and Savkin (2002) is a more complex one, wherein the negative effects of the tumor cells on the healthy cell population are also considered. This is a vital addition to the earlier three models which did not consider the interaction between the tumor and the healthy cells. It should be noted that the healthy cell population is also assumed to follow a Gompertz growth model with the cytotoxic drug killing both the cancerous as well as normal cells. The set of differential equations for this mathematical model is:

$$\begin{aligned}\dot{N}(t) &= \alpha N \ln \frac{\theta_N}{N} - \mathcal{L}_1(c)N \\ \dot{L}(t) &= \beta L \ln \frac{\theta_L}{L} - \mathcal{L}_2(c)L - \Xi(N)L \\ c &= c(t) \in [0, c_{max}],\end{aligned}$$

where  $N(t)$  is the population of the tumor cells,  $L(t)$  is the population of the normal (healthy) cells,  $\theta_N$  is the maximum allowable size of the tumor,  $\theta_L$  is the normal size of the healthy cell population and  $c(t)$  (the control) is the concentration of the cytotoxic drug at the tumor site. The inhibiting effect of the cancerous cells on the healthy cells is captured by the  $-\Xi(N)L$  term. The function  $\Xi(\cdot)$  is a strictly increasing function

of  $N$  and is continuously differentiable on the interval  $[0, \infty]$ . Some other theoretical studies and mathematical works have been conducted to investigate cancer treatment. For information on T cell sensitivity, see Chan, George, and Stark (2003). For more related models, see Panetta and Kirschner (1998), Swan (1986; 1990), de Pillis and Radunskaya (2001), and Murray (1990a; 1990b).

More recently, de Pillis et al. (2007a) investigated a mathematical model of tumor-immune interactions with chemotherapy, and strategies for optimally administering treatment. In their model, two immune components (effector-immune cells and circulating lymphocytes) were included. They used the count of circulating lymphocytes in a patient's bloodstream as a reflection of the strength of the patient's overall immune health. The system of differential equations describing the growth, death, and interactions of cell populations with a chemotherapy treatment is given by

$$\begin{aligned}\dot{T}(t) &= aT(1 - bT) - c_1NT - K_TMT \\ \dot{N}(t) &= \alpha_1 - fN + g\frac{T}{h+T}N - pNT - K_NMN \\ \dot{C}(t) &= \alpha_2 - \beta C - K_CMC \\ \dot{M}(t) &= -\gamma M + V_M(t),\end{aligned}$$

where  $T(t)$  is the tumor cell population,  $N(t)$  is the effector-immune cell population,  $C(t)$  is the circulating lymphocyte population, and  $M(t)$  is the chemotherapy drug concentration.

In addition to chemotherapy, recently, immunotherapies are quickly becoming an important component in the multi-pronged approaches being developed to treat certain forms of cancer. Immunotherapy refers to the use of natural and synthetic substances to stimulate the immune response. The goal of immunotherapy is to strengthen the body's own natural ability to combat cancer by enhancing the effectiveness of the immune system. See, for example, Farrar et al. (1999), Morecki et al. (1996), Muller et al. (1998), O'Byrne et al. (2000), and Stewart (1996). Immunological therapies include the use of



antigen and nonantigen specific agents such as cytokines. Cytokines have been used to treat melanoma, leukemia, lymphoma, neuroblastoma, Kaposi's sarcoma, mesothelioma, brain cancer, cancer of the kidney, and cancer of the cervix. Interleukin-2 (IL-2) is a cytokine that was approved by the US Food and Drug Administration (FDA) in 1992 for treatment of metastatic renal cell (kidney) cancer. IL-2 helps immune system cells reproduce more rapidly once they are in the patient, and it became the first cytokine approved for use alone in treating advanced cancer. Clinical trials of mixed chemo-immunotherapy are developed for metastatic melanoma treatment. For instance, a series of sequential Phase II trials were conducted at M.D. Anderson Cancer Center (Buzaid 2000; Buzaid and Atkins 2001). These trials were based on integrating of IL-2 and interferon-alpha (IFN- $\alpha$ ) with the CVD (cisplatin, vinblastine, and dacarbazine) regimen.

One of the first attempts to consider effects of immunotherapy within an appropriate mathematical model was made by Kirschner and Panetta (1998). de Pillis, Gu, and Radunskaya (2006) proposed and analyzed a mathematical model governing cancer growth on a cell population level with combination immunotherapy, chemotherapy and vaccine treatment. This model's characteristics are useful not only to gain a broad understanding of the specific system dynamics, but also to help guide the development of combination therapies. The model describes the kinetics of four populations (tumor cells and three types of immune cells), as well as two drug concentrations in the bloodstream, using a series of coupled ordinary differential equations (ODEs) based on the model developed by de Pillis and Radunskaya (2003) in their previous study. The populations at time  $t$  are denoted by:

- $T(t)$ , tumor cell population,
- $N(t)$ , total NK cell population,
- $L(t)$ , total CD8<sup>+</sup>T cell population,
- $C(t)$ , number of circulating lymphocytes (or white blood cells),

- $M(t)$ , chemotherapy drug concentration in the bloodstream,
- $I(t)$ , immunotherapy drug concentration in the bloodstream.

Bringing together the specific forms for each cell growth and interaction term leads to the full system of ODEs:

$$\dot{T}(t) = aT(1 - bT) - cNT - DT - K_T(1 - e^{-M})T \quad (2.1)$$

$$\dot{N}(t) = eC - fN + g\frac{T^2}{h + T^2}N - pNT - K_N(1 - e^{-M})N \quad (2.2)$$

$$\begin{aligned} \dot{L}(t) = & -mL + j\frac{D^2T^2}{k + D^2T^2}L - qLT + (r_1N + r_2C)T \\ & - uNL^2 - K_L(1 - e^{-M})L + \frac{p_I LI}{g_I + I} + v_L(t) \end{aligned} \quad (2.3)$$

$$\dot{C}(t) = \alpha - \beta C - K_C(1 - e^{-M})C \quad (2.4)$$

$$\dot{M}(t) = -\gamma M + v_M(t) \quad (2.5)$$

$$\dot{I}(t) = -\mu_I I + v_I(t) \quad (2.6)$$

$$D = d\frac{(L/T)^l}{s + (L/T)^l}, \quad (2.7)$$

where the time denoted functions  $v_L(t)$ ,  $v_M(t)$ , and  $v_I(t)$  are the drug intervention terms for tumor infiltrating lymphocyte (TIL), chemotherapy drug, and interleukin-2 (IL-2), respectively.

To obtain data which could mimic real-life clinical data, we will use time-domain simulations of a nonlinear ODE model. In Section 4.2, we will provide more detailed discussion for the main characteristics of our proposed model, before defining the data generation procedure itself.

## 2.2.2 Finding Optimal Treatment Solutions

Given a set of mathematical models (for example, ODEs), optimal control theory is one of the mathematical optimization methods for deriving control policies. It was originally introduced by Pontryagin et al. (1962) as a convenient method of finding a control law for a given system such that a certain optimality criterion is achieved. A control problem usually includes an objective functional that is a function of state and control variables. The objective functional takes one or more functions as an argument and returns a number.

There exists the inevitable trade-offs involved in cancer treatment because many analyses show that large amounts of chemotherapy will kill the tumor, but it may also kill the patient. In the context of mathematically modelling cancer growth with chemotherapy, because of this implicit understanding that chemotherapy has damaging side effects, it is common to frame an optimal control problem so that the total amount of drug is minimized (Matveev and Savkin 2002; Fister and Panetta 2003). It is appealing to use an optimal control strategy to accomplish this, since the solution of it may cure the patient as fast as possible with the minimized dose level.

A simple abstract framework goes as follows. Given a dynamical system with input of IL-2  $u_1(t)$ , input of chemotherapy  $u_2(t)$ , and the size of the tumor at the end of the treatment period  $C(t)$ , define an objective functional to be minimized. The objective functional is the sum of the path costs, which usually take the form of an integral over time, and the terminal costs, which is a function only of the terminal state,  $t_f$ . Thus, this objective functional typically takes the form

$$J(u_1, u_2) = \int_{t=0}^{t=t_f} \lambda_1 u_1(t) + \lambda_2 u_2(t) dt + C(t_f).$$

Here, finding the optimal treatment strategy is the equivalent of minimizing  $J$ .

In optimal control theory, establishing the existence of the solution is the first task. Mathematically, using the fact that the solution to each state equation is bounded, the

existence of an optimal control for many problems can be determined using the theories developed by Fleming and Rishel (1975), Seierstad and Sydsaeter (1987), and Hartl et al. (1995) (Filippov-Cesari's theorem). Then, characterizing the optimal control can be accomplished by using Pontryagin's maximum principle (a necessary condition) (Pontryagin et al., 1962), by solving the Hamilton-Jacobi-Bellman equation (a sufficient condition), or by using other conditions from Kamien and Schwarz (1991) or the generalized Legendre-Clebsch conditions (Krener, 1977).

Although optimal control theory is promising, in some situations finding the solution can be very challenging. The optimal solution depends on the complexity of the mathematical model, the objective functional and the state constraints. One disadvantage of optimal control is its sensitivity to the choice of objective functionals. For instance, Fisher and Donnelly (2005) demonstrated qualitatively different treatment strategies based on the use of different objective functionals. These differences show the importance of defining an objective functional that most accurately reflects the toxicities of a particular drug along with the objective of the treatment strategy. Some objective functionals can be theoretically analyzed more tractably than others. de Pillis et al. (2007a) provided an example to illustrate this. In their study, they analyzed two types (quadratic and linear controls) of objective functionals for the models of chemotherapy. In the quadratic case, the control quickly moves to a small value, then gradually decreases, however in the linear control it is essentially turned off (appears to be the so-called bang-bang control). Since the amount of drug being delivered to the patient is small, the quadratic control treatment is comparable to the linear bang-bang control case in that the tumor is reduced by the same magnitude over the same time frame. Treatments based on both functionals were successful in reducing the tumor. Although the quadratic and linear controls have similar behavior in the administration of the chemotherapy drug, applying the control in a linear fashion to their model is somewhat problematic. When dealing with linear controls, singular representations are difficult to determine, and the possibility of a sin-

gular control can not be ruled out. So in most situations, suitable objective functionals need to be defined to capture the tumor cell population and the amount of drug used for the therapy. It will be important to choose the correct functional and the appropriate constraints since it is difficult to say exactly what these are before analyzing the model.

Another limitation of optimal control is that usually only the simple models with a small number of variables can be analyzed theoretically. When the mathematical models are very complex with a large number of variables and parameters, it is difficult to seek optimal bang-bang solutions successfully. For example, see the model (2.1)–(2.7) of mixed immuno-chemotherapy of tumor in de Pillis et al. (2006; 2007b). Although there are conditions in which the controls exist singularly, and this may be the best strategy for minimizing a tumor burden, the characterizations of the singular control can not be explicitly determined in some settings.

# 3 Reinforcement Learning, $Q$ -Learning, and Their Approximations

## 3.1 Introduction

Our goal in this chapter is to introduce the reinforcement learning theory, which will be used to discover optimal therapies in clinical cancer trials. From a computer science perspective, reinforcement learning is the first field to address the computational issues that arise when learning from interaction with an environment in order to achieve long-term goals (Sutton and Barto, 1998). Moreover, in contrast with adaptive design and optimal control introduced in previous chapters, reinforcement learning ( $Q$ -learning) is a model-free method which can be used for finding individualized therapies. This approach we explore is much more focused on goal-directed learning from interaction with the environment than other approaches to machine learning.

Multiple scientific fields have made contributions to reinforcement learning — machine learning, operations research, control theory, psychology, and neuroscience, to name but a few. Reinforcement learning has been applied successfully in a number of areas, and has produced some successful practical applications. These applications range from robotics and control to industrial manufacturing and combinatorial search problems such as computer game playing (Kaelbling, Littman, and Moore, 1996). One example is that reinforcement learning has been used to teach an autonomous controller to fly a helicopter upside down (see Figure 1), demonstrating unequivocally the potential of reinforcement learning for solving problems that are complex and counter-intuitive (Ng et al., 2006).

Another most convincing application is TD-gammon, a system that learns to play the game of Backgammon by playing against itself and learning from the results, described by Gerald Tesauro in (Tesauro, 1994; 2002). TD-gammon reaches a level of play that is superior to even the best human players. Recently, there has been some interest in the application of reinforcement learning algorithms to problems from the fields of management science and operations research. In an interesting paper by Gosavi, Bandla, and Das (2002), reinforcement learning is applied to airline yield management, and the aim is to find an optimal policy for the denial/acceptance of booking requests for seats in various fare classes. A second example is Crites and Barto (1998), where reinforcement learning is used to find a (sub)optimal control policy for a group of elevators. In both the above papers, the authors report that reinforcement learning based methods outperform the best and most often used standard algorithms. A marketing application is described in Pednault, Abe, and Zadrozny (2002), where a target selection decision in direct marketing is seen as a sequential problem.

This chapter is organized as follows. In Section 3.2, we describe the definition of reinforcement learning in a simplified setting, one that does not involve multiple agents to act in more than one situation. In Section 3.2.1, we show why the value function (function of states or state-action pairs) is the unique solution to the Bellman equation. The temporal-difference learning is introduced and discussed in Section 3.2.2. At the end of Section 3.2.2, we take a step closer to the  $Q$ -learning problem, which is one of the most important breakthroughs in reinforcement learning. In addition, in Section 3.3–3.5, we discuss three recent flexible techniques from the machine learning literature, support vector machines (SVM), support vector regression (SVR), and extremely randomized trees (ERT), as our main methods to fit  $Q$ -functions and to learn an optimal policy using a training data set.

## 3.2 Reinforcement Learning and $Q$ -Learning

Inspired by related psychological theory in computer science, reinforcement learning is a sub-area of machine learning. A detailed account of the history of reinforcement learning is found in Sutton and Barto (1998). The basic process of reinforcement learning involves trying a sequence of actions, recording the consequences of those actions, statistically estimating the relationship between actions and consequences, and then choosing the action that results in the most desirable consequence. In our reinforcement learning design, the thing a patient interacts with is called the “environment”, which may indicate the complex system consisting of the human body and more sources of error and greater restrictions on what can be measured. While these interactions continually happen, we choose a sequence of actions applied to the patient and the environment responds to those actions and provides feedback. To be specific, we use  $S$  and  $A$  to denote random variables, where  $S$  represents the set of environmental “states” and  $A$  represents the set of possible “actions”. Here “states” may represent individual patient covariates and “actions” can be denoted by various treatments or dose levels. Both variables can be discrete or continuous. Define time-dependent variables  $\mathbf{S}_t = \{S_0, S_1, \dots, S_t\}$ , and similarly, define  $\mathbf{A}_t = \{A_0, A_1, \dots, A_t\}$ . We use lower case letters, such as  $s$  and  $a$ , to denote the realized values of the random variables  $S$  and  $A$ , respectively. Also, for convenience, define  $\mathbf{s}_t = \{s_0, s_1, \dots, s_t\}$ , and similarly,  $\mathbf{a}_t = \{a_0, a_1, \dots, a_t\}$ . We assume the finite longitudinal trajectories are sampled at random according to a distribution  $P$ . Such a distribution is composed of the unknown distribution of each  $S_t$  conditional on previous  $(\mathbf{S}_{t-1}, \mathbf{A}_{t-1})$ . We denote these unknown conditional densities as  $\{f_0, \dots, f_T\}$ , and denote expectations with respect to the distribution  $P$  as  $E$ .

As a consequence of a patient’s treatment, after each time step  $t$ , the patient receives a numerical reward  $r_t$ . This could be denoted as a function, which maps to a single number the key elements: previous state  $\mathbf{s}_t$ , action  $\mathbf{a}_t$ , and current state  $s_{t+1}$ . When



$t = 0, 1, \dots, T$ , this process can be described by

$$r_t = R(\mathbf{s}_t, \mathbf{a}_t, s_{t+1}).$$

Reinforcement learning is learning what to do, how to map situations from state space  $S$  to action space  $A$ , and depending on what our goal is, how to choose  $a_t$  to maximize or minimize the expected discounted return:

$$R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots + \gamma^T r_{t+T} = \sum_{k=0}^T \gamma^k r_{t+k}.$$

In this equation,  $\gamma$  is the discount rate ( $0 \leq \gamma \leq 1$ ), which means, rewards that are received in the future are geometrically discounted according to  $\gamma$ . Additionally, we can interpret  $\gamma$  in another way. It can be seen as a control to balance the agent's immediate rewards and future rewards. If  $\gamma = 0$ , we easily see that  $R_t = r_t$ , and we only need to learn how to choose  $a_t$  so as to maximize or minimize the immediate reward  $r_t$ . As  $\gamma$  approaches 1, we take future rewards into account more strongly. In the extreme case, when  $\gamma = 1$ , we fully maximize or minimize rewards over the long run.

Another key element of a reinforcement learning system is an exploration “policy”,  $p$ , which maps state  $\mathbf{s}_t$  and action  $\mathbf{a}_{t-1}$  to the probability  $p_t(a \mid \mathbf{s}_t, \mathbf{a}_{t-1})$  (the probability that action  $a$  is taken given history  $\{\mathbf{s}_t, \mathbf{a}_{t-1}\}$ ). If the policy is possibly non-stationary and non-Markovian but deterministic, we denote  $\pi_t(\mathbf{s}_t, \mathbf{a}_{t-1}) = a_t$ . In other words, policy  $\pi_t$  as a step in a sequence of decision rules  $\{\pi_1, \dots, \pi_T\}$  is an action. Let the distribution  $P_\pi$  denote the distribution of training data whereby the policy  $\pi$  is used to generate actions. Then we can denote expectations with respect to the distribution  $P_\pi$  by an  $E_\pi$ . Let  $\Pi$  be the collection of all policies, and expectation  $E_\pi$  ranges are over  $\pi \in \Pi$ . For simplicity and with no loss of generality, in our study, we mainly concentrate on the goal of discovering which treatment yields a maximized reward for a patient. So seeking the policy that maximizes the expectations with respect to the sum of the rewards over the time trajectories is the ultimate goal of the study.

### 3.2.1 Value Functions and the Bellman Equation

Efficiently estimating the value function is the most important component of almost all reinforcement learning algorithms. The value function is defined as a function of a state or state-action pair, and the function represents the total amount of reward an agent can expect to accumulate over the future, starting from a given state. Recalling that  $\Pi$  is the set of all policies, we define the value function  $V(s)$  to be the expected return when starting in  $s$  under a policy  $\pi \in \Pi$ . This is formally denoted as

$$V(s) = E_{\pi} [R_t \mid s_t = s] = E_{\pi} \left[ \sum_{k=0}^T \gamma^k r_{t+k} \mid s_t = s \right]. \quad (3.1)$$

We are more interested in defining the time-dependent value function for history  $(\mathbf{s}_t, \mathbf{a}_{t-1})$ , that is,

$$V_t(\mathbf{s}_t, \mathbf{a}_{t-1}) = E_{\pi} \left[ \sum_{k=0}^T \gamma^k r_{t+k} \mid \mathbf{S}_t = \mathbf{s}_t, \mathbf{A}_{t-1} = \mathbf{a}_{t-1} \right]. \quad (3.2)$$

Equation (3.1) and (3.2) are called the state-value functions for policy  $\pi$  and the action-value function for policy  $\pi$  in Sutton and Barto (1998, page 69), respectively.

A fundamental property of value functions used throughout reinforcement learning is that they satisfy particular recursive relationships. To see this, first let  $T = \infty$ , then we extend equation (3.2) as follows,

$$\begin{aligned} V_t(\mathbf{s}_t, \mathbf{a}_{t-1}) &= E_{\pi} \left[ \sum_{k=0}^{\infty} \gamma^k r_{t+k} \mid \mathbf{S}_t = \mathbf{s}_t, \mathbf{A}_{t-1} = \mathbf{a}_{t-1} \right] \\ &= E_{\pi} \left[ r_t + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid \mathbf{S}_t = \mathbf{s}_t, \mathbf{A}_{t-1} = \mathbf{a}_{t-1} \right] \\ &= E_{\pi} \left[ r_t + \gamma V_{t+1}(\mathbf{S}_{t+1}, \mathbf{A}_t) \mid \mathbf{S}_t = \mathbf{s}_t, \mathbf{A}_t = \mathbf{a}_t \right] \\ &= \sum_{a_t} \pi_t(\mathbf{s}_{t+1}, \mathbf{a}_t) \sum_{s'} \mathcal{P}_{ss'}^a \left[ \mathcal{R}_{ss'}^a + \gamma V_{t+1}(s') \right], \end{aligned}$$

where

$$\mathcal{P}_{ss'}^a = Pr\{s_{t+1} = s' \mid s_t = s, \mathbf{a}_t = a\}$$

and

$$\mathcal{R}_{ss'}^a = E \left[ r_t \mid s_t = s, \mathbf{a}_t = a, s_{t+1} = s' \right].$$

The last two equations are two forms of the *Bellman equations* for  $V_t(\mathbf{s}_t, \mathbf{a}_{t-1})$ . The Bellman equation was first introduced by Bellman (1957). The Bellman equation expresses the relationship between the value of a state and the values of its successor states: the value of the start state is equivalent to the value of the expected next state plus the expectation of the reward along the way. It is worth noting that the value function  $V_t(\mathbf{s}_t, \mathbf{a}_{t-1})$  is the unique solution to its Bellman equation.

Before we consider seek the best policy to maximize the reward, we describe the optimal value function and optimal policy here first. The optimal value function is simply defined as

$$V_t^*(\mathbf{s}_t, \mathbf{a}_{t-1}) = \max_{\pi \in \Pi} V_t(\mathbf{s}_t, \mathbf{a}_{t-1}) = \max_{\pi \in \Pi} E_{\pi} \left[ \sum_{k=0}^T \gamma^k r_{t+k} \mid \mathbf{S}_t = \mathbf{s}_t, \mathbf{A}_{t-1} = \mathbf{a}_{t-1} \right].$$

The optimal policy is defined as a policy which yields the value function  $V_t(\mathbf{s}_t, \mathbf{a}_{t-1})$  with the highest value. Although there may be more than one, we denote all the optimal policies by  $\pi^*$ . Based on the existence of an optimal policy, we can establish the *Bellman optimality equation*, which expresses the fact that the value of a state under an optimal policy must equal the expected return for the best action from the state. Thus, the Bellman optimality equation for  $V_t^*(\mathbf{s}_t, \mathbf{a}_{t-1})$  is derived as follows:

$$\begin{aligned} V_t^*(\mathbf{s}_t, \mathbf{a}_{t-1}) &= \max_{a_t} E_{\pi^*} \left[ \sum_{k=0}^{\infty} \gamma^k r_{t+k} \mid \mathbf{S}_t = \mathbf{s}_t, \mathbf{A}_{t-1} = \mathbf{a}_{t-1} \right] \\ &= \max_{a_t} E_{\pi^*} \left[ r_t + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid \mathbf{S}_t = \mathbf{s}_t, \mathbf{A}_{t-1} = \mathbf{a}_{t-1} \right] \\ &= \max_{a_t} E \left[ r_t + \gamma V_{t+1}^*(\mathbf{S}_{t+1}, \mathbf{A}_t) \mid \mathbf{S}_t = \mathbf{s}_t, \mathbf{A}_t = \mathbf{a}_t \right] \\ &= \max_{a_t} \sum_{s'} \mathcal{P}_{ss'}^a \left[ \mathcal{R}_{ss'}^a + \gamma V_{t+1}^*(s') \right]. \end{aligned}$$

It is clear that the optimal policy,  $\pi^*$ , must satisfy

$$\pi_t^*(\mathbf{s}_t, \mathbf{a}_{t-1}) \in \arg \max_{a_t} E \left[ r_t + \gamma V_{t+1}^*(\mathbf{S}_{t+1}, \mathbf{A}_t) \mid \mathbf{S}_t = \mathbf{s}_t, \mathbf{A}_t = \mathbf{a}_t \right].$$

Modern techniques in mathematical and computational areas have stimulated the developments of many methods for estimating the optimal value functions or optimal policies. Many of the existing methods can be categorized into one of the following two classes: dynamic programming and temporal-difference learning (Sutton and Barto, 1998). Bellman (1957) first provided the “dynamic programming” term to show how these methods are useful to a wide range of problems. Minsky (1961) first described the connection between dynamic programming and reinforcement learning. In classical dynamic programming methods, “policy evaluation” and “policy improvement” (Bellman 1957; Howard 1960) refer to the computation of the value function and the improved policy, respectively. The computation in both methods requires an iterative process. Combining these two methods together, we obtain two other methods called “policy iteration” and “value iteration” (Puterman and Shin 1978; Bertsekas 1987). Although dynamic programming can be applied to many types of problems, it is restricted to solving reinforcement learning problems under the Markov assumption. If this assumption is violated, it may not be possible to find an exact solution. Additionally, dynamic programming for solving reinforcement learning problems requires knowledge of a complete and accurate model of the environment. Specifically, for instance, it requires  $\mathcal{P}_{ss'}^a$  to be fully observed. This may be unrealistic in the clinical trial setting because of the heterogeneity in the model across individual patients.

In contrast, in reinforcement learning an agent does not necessarily know the reward function and the state-transition function. Both the reward and the new state that result from an action are determined by the environment, and the consequences of an action must be observed by interacting with the environment. In other words, reinforcement learning agents are not required to possess a model of their environment. This aspect distinguishes reinforcement learning from dynamic programming. In the next section we will discuss temporal-difference learning, which is a reinforcement learning algorithm that does not need such a model to find an optimal policy in an MDP.

### 3.2.2 Temporal-Difference Learning and Q-Learning

In the previous section we have defined optimal value functions and optimal policies, and we have reviewed the Bellman optimality equation and dynamic programming methods for obtaining an optimal policy based on the Markov property, assuming that we already have a model of the environment. Actually, even if we have a complete and accurate model of the environment's dynamics, it is usually not possible to directly compute an optimal policy by just solving the Bellman optimality equation. This section examines model-free learning, that is, temporal-difference (TD) learning, which was first introduced by Sutton (1988).

One fundamental expression of TD-learning is the incremental implementation. This implementation requires less memory for estimates and less computation. The general form is

$$\text{new estimate} \leftarrow \text{old estimate} + \text{stepsize} \left[ \text{target} - \text{old estimate} \right].$$

Specifically, if we replace *estimate* with value function, *target* with reward function, and denote *stepsize* as  $\alpha$ , then in this case TD learning becomes

$$V_t(\mathbf{S}_t, \mathbf{A}_{t-1}) \leftarrow V_t(\mathbf{S}_t, \mathbf{A}_{t-1}) + \alpha \left[ r_t + \gamma V_{t+1}(\mathbf{S}_{t+1}, \mathbf{A}_t) - V_t(\mathbf{S}_t, \mathbf{A}_{t-1}) \right]. \quad (3.3)$$

Roughly speaking, the TD method bases its incremental implementation in part on an existing estimate. Recalling the Bellman equation in the previous section, we know that

$$\begin{aligned} V_t(\mathbf{s}_t, \mathbf{a}_{t-1}) &= E_\pi \left[ \sum_{k=0}^{\infty} \gamma^k r_{t+k} \mid \mathbf{S}_t = \mathbf{s}_t, \mathbf{A}_{t-1} = \mathbf{a}_{t-1} \right] \\ &= E_\pi \left[ r_t + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid \mathbf{S}_t = \mathbf{s}_t, \mathbf{A}_{t-1} = \mathbf{a}_{t-1} \right] \\ &= E_\pi \left[ r_t + \gamma V_{t+1}(\mathbf{S}_{t+1}, \mathbf{A}_t) \mid \mathbf{S}_t = \mathbf{s}_t, \mathbf{A}_t = \mathbf{a}_t \right]. \end{aligned}$$

In these equations, under a policy  $\pi$ , each  $V$  represents the true value of a state-action pair, but this is not known. Thus, in (3.3), the TD target uses the current estimate  $V$

instead of the true  $V$ . TD learning as discussed above is also known as TD(0) learning, which is a special case of TD( $\lambda$ ) learning. Almost any TD( $\lambda$ ) learning belongs to the “eligibility traces” problem. For more details on these issues, see Sutton and Barto (1998) and Kaelbling, Littman, and Moore (1996).

One of the most important off-policy TD-learning methods is Watkins’  $Q$ -learning (Watkins, 1989; Watkins and Dayan, 1992).  $Q$ -learning handles discounted infinite-horizon Markov decision process (MDP). It requires no prior knowledge, is exploration insensitive and easy to implement, and is so far one of the most popular and seems to be the most effective model-free algorithm for learning from delayed reinforcement. In the situation where we don’t have any information about the transition function or the probability distribution of the random variables, such a model-free method can be used to find optimal strategies from the unknown system.

$Q$ -learning no longer requires estimating the value function, it estimates a  $Q$ -function instead. The algorithm therefore utilizes such a  $Q$ -function which calculates the quality of a state-action combination as follows:

$$Q : S \times A \rightarrow \mathbb{R}.$$

The motivation of  $Q$ -learning is that once the  $Q$  functions have been estimated, we only need to know the state to determine an action, without the knowledge of a transition model that tells us what state we might go to next. Before learning has started,  $Q$  returns a fixed value which is chosen by the designer. Then, at each time point  $t$ , the learner is given a reward value which is calculated for each combination of a state  $s_t \in S_t$ , and action  $a_t \in A_t$ . The core of the algorithm is a simple value iteration update. It assumes the old value and makes a correction based on the new information as follows (Sutton and Barto, 1998):

$$Q_t(s_t, a_t) \leftarrow Q_t(s_t, a_t) + \alpha_t(s_t, a_t) \times \left[ r_t + \gamma \max_{a_{t+1}} Q_{t+1}(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t) \right],$$

where  $r_t$  is the current reward given at time  $t$ ,  $\alpha_t(s_t, a_t) \in (0, 1]$  the learning rate (or

learning step-size).  $\alpha_t(s_t, a_t)$  is a constant which determines to what extent the newly acquired information will override the old information, that is, how fast learning takes place. A factor of 0 will make the learner not learn anything, while a factor of 1 would make the learner fully update based on the most recent information. We can interpret  $\gamma$  as a control to balance a learners' immediate rewards and future rewards. As  $\gamma$  approaches 1, we take future rewards into account more strongly. In the following context, we let  $\gamma = 1$ , which means we fully maximize rewards over the long run. For simplicity of computation, we ignore the step-size (let  $\alpha_t(s_t, a_t) = 1$ ) for the rest of the article. All results hold with minor modifications when the step-size effects are considered.

From a statistical perspective, the optimal time-dependent  $Q$ -function is

$$Q_t^*(\mathbf{s}_t, \mathbf{a}_t) = E \left[ r_t + \gamma V_{t+1}^*(\mathbf{S}_{t+1}) \mid \mathbf{S}_t = \mathbf{s}_t, \mathbf{A}_t = \mathbf{a}_t \right].$$

Note that since

$$V_t^*(\mathbf{s}_t) = \max_{a_t} Q_t^*(\mathbf{s}_t, \mathbf{a}_{t-1}, a_t),$$

it is relatively easy to determine an optimal policy, which satisfies

$$\pi_t^*(\mathbf{s}_t, \mathbf{a}_{t-1}) = \mathit{arg} \max_{a_t} Q_t^*(\mathbf{s}_t, \mathbf{a}_{t-1}, a_t).$$

One-step  $Q$ -learning has the simple recursive form

$$Q_t(\mathbf{S}_t, \mathbf{A}_t) \leftarrow r_t + \gamma \max_{a_{t+1}} Q_{t+1}(\mathbf{S}_{t+1}, \mathbf{A}_t, a_{t+1}). \quad (3.4)$$

Under some appropriate and rigorous assumptions,  $Q_t$  has been shown to converge to  $Q^*$  with probability 1 (Watkins and Dayan, 1992). More general convergence results were proved by Jaakkola, Jordan, and Singh (1994) and Tsitsiklis (1994).

In learning a non-stationary non-Markovian policy with one set of finite horizon trajectories (training data set)

$$\{S_0, A_0, r_0, S_1, A_1, r_1, \dots, A_T, r_T, S_{T+1}\},$$

we denote the estimator of the optimal  $Q$ -functions based on this training data by  $\widehat{Q}_t$ , where  $t = 0, 1, \dots, T$ . According to the recursive form of  $Q$ -learning in (3.4), we must

estimate  $Q_t$  backwards through time  $t = T, T - 1, \dots, 1, 0$ , that is, estimate  $Q_T$  from the last time point back to  $Q_0$  at the beginning of the trajectories. And we set  $Q_{T+1}$  equal to 0 in the first equation, yielding

$$Q_T(\mathbf{S}_T, \mathbf{A}_T) \leftarrow r_T + \gamma \max_{a_{T+1}} Q_{T+1}(\mathbf{S}_{T+1}, \mathbf{A}_T, a_{T+1}).$$

In order to estimate each  $Q_t$ , we denote  $Q_t(\mathbf{s}_t, \mathbf{a}_t; \theta)$  as a function of a set of parameters  $\theta$ , and we allow the estimator to have different parameter sets for different time points  $t$ . Once this backwards estimation process is done, we save the sequence  $\{\widehat{Q}_0, \widehat{Q}_1, \dots, \widehat{Q}_T\}$  for estimating optimal policies,

$$\widehat{\pi}_t = \arg \max_{a_t} \widehat{Q}_t(\mathbf{s}_t, \mathbf{a}_t; \theta_t),$$

where  $t = 0, 1, \dots, T$ , and thereafter use these optimal policies to test or predict for a new data set.

There are many other promising learning methods based on modification or extension of  $Q$ -learning, for example, Blatt, Murphy, and Zhu (2004) proposed  $A$ -learning. However, some properties of these methods have not yet been carefully investigated. Due to the simple equation expressions and the minimal amount of computation, we restrict our attention to  $Q$ -learning for discovering effective therapeutic regimens in our clinical settings.

### 3.3 Support Vector Machines (SVM)

As we mentioned earlier in Chapter 2, either adaptive design or optimal control must proceed by using explicit mathematical models. This requirement yields a limitation for discovering optimal dynamic treatment regimens that are tailored to individual patient needs. Thus we introduced a powerful technique from computer science and statistics — reinforcement learning, specifically  $Q$ -learning — to our clinical trial design setting.



$Q$ -learning could circumvent this situation by its emphasis on learning through the individual’s interaction with its environment, without relying on any complete models of the environment. We call this application to clinical trial design as “Reinforcement Learning Design” or “ $Q$ -learning Design”.

In Section 3.3–3.5, our main aim is to estimate the  $Q$ -function for finding the corresponding optimal policy. However, challenges may arise due to the complexity of the structure of true  $Q$ -function, including the high-dimension of the states variable  $S$ , the high-dimension of the action variable  $A$ , or when the action space is continuous. In order to obtain the estimator of interest, many authors consider different approaches in recent years. Murphy (2005b), Blatt, Murphy, and Zhu (2004) and Tsitsiklis and van Roy (1996) showed that  $Q$ -learning estimating can be viewed as approximating least squares value iteration. The parameters  $\theta_t$  for the  $t$ -th  $Q$ -function satisfy

$$\theta_t \in \arg \min_{\theta} \mathbb{E}_n \left[ r_t + \max_{a_{t+1}} Q_{t+1}(\mathbf{S}_{t+1}, \mathbf{A}_t, a_{t+1}; \theta_{t+1}) - Q_t(\mathbf{S}_t, \mathbf{A}_t; \theta) \right]^2.$$

This is consistent with the one-step update of Sutton and Barto (1998) with  $\gamma = 1$ , and furthermore, it is generalized to permit function approximation and non-stationary  $Q$ -functions. Another simple and standard estimating form is in Murphy et al.’s (2007) method. They claimed that  $Q$ -learning is a generalization of the familiar regression model. In their sequential multiple assignment randomized trial (SMART) design, there are only two treatment decisions. Thus construction of the decision rules should be addressed from the second decision to the first decision (backwards). For instance, in the second decision, two treatment options are available. If we denote  $A_2$  as the second decision, it is coded as 1 if the switch is assigned and is coded as 0 if augmentation is assigned. Based on the SMART data, the regression model for  $Q_2$  is

$$Q_2(S, A_2; \theta) = \beta_0 + \beta_1 S + (\beta_2 + \beta_3 S) A_2,$$

where  $\theta = (\beta_0, \beta_1, \beta_2, \beta_3)$  and  $S$  indicates the state value (a summary of side effects) up to the end of the first decision point. When the dimension of actions is low, linear regression

methods should be adequate, but in more extreme cases these methods can be questionable. Considering the one possible set  $\{a_0, a_1, \dots, a_n\}$  with  $n \geq 3$ , the linear regression method may only yield the optimal decision as  $a_0$  or  $a_n$  due to the  $\max_a Q(\mathbf{S}, \mathbf{A}, a; \theta)$  term in the  $Q$ -learning implementation, therefore, quadratic regression or higher order polynomial regression may be desired for estimating the  $Q$ -function. The complex and unclear structure of the  $Q$ -function has motivated the vast literature on nonparametric machine learning and statistical methods.

In this section, we introduce support vector machines (SVMs) as our main technique for fitting  $Q$ -functions. The foundation of SVMs was developed by Vapnik (1995). SVMs have received increasing attention from the statistical community as well as from computer science and engineering, and they keep gaining popularity due to many attractive features and promising empirical performance. The SVM paradigm is originally designed for the classification problem, and it provides a compromise between the parametric and the nonparametric approaches. SVMs are often involved in the solution of learning the relationship between the  $\mathbf{x}$  and  $\mathbf{y}$  variables in a training data set  $\{(\mathbf{x}_i, \mathbf{y}_i) \in X \times Y\}_{i=1}^n$ . In  $Q$ -learning, the variable  $X$  may be replaced by  $\{S, A\}$  that represents states and actions information, and  $Y$  may be replaced by  $r$  that represents numerical rewards. In this section, we first illustrate the basic ideas of SVM for the typical two-group classification problem. Then we briefly discuss support vector regression (SVR) as an extension of SVM in Section 3.4.

The classification problems solved by SVMs can be restricted to consideration of the two-class problem without loss of generality. A classification task usually involves training and testing data which consist of some data instances. Each instance in the training set contains one “class label” ( $y$ ) and several “attributes” ( $\mathbf{x}$ ). SVM is used as a statistical technique for classifying samples  $\{(\Phi(\mathbf{x}_i), y_i) \in X \times Y\}_{i=1}^n$ , where  $y_i = +1$  or  $y_i = -1$  indicates the two possible classes, and  $\Phi$  is a function mapping the attributes  $\mathbf{x}_i$  into a “feature space”. This nonlinear transformation  $\Phi$  guarantees that any data set

becomes arbitrarily separable as the data dimension grows (Cover 1965).

Denoting  $\mathbf{w}^T \Phi(\mathbf{x}) + b = 0$  as any separating hyperplane in the feature space, we can rescale  $\mathbf{w}$  and  $b$  so that the following equations hold for  $i = 1, \dots, n$ :

$$\mathbf{w}^T \Phi(\mathbf{x}_i) + b \begin{cases} \geq 1, & \text{if } y_i = +1 \\ \leq -1, & \text{if } y_i = -1. \end{cases}$$

The distance between two classes  $2/\|\mathbf{w}\|$  is called the margin. SVM works, roughly, by finding the hyperplane in the feature space which separates the  $y_i = +1$  class from the  $y_i = -1$  class with the largest margins. If the mapped data have become linearly separable, the following equation has to be solved by maximizing the margin:

$$\begin{aligned} & \min_{\mathbf{w}, b} \|\mathbf{w}\|^2 \\ & \text{subject to } y_i(\mathbf{w}^T \Phi(\mathbf{x}_i) + b) \geq 1, \quad i = 1, \dots, n. \end{aligned} \quad (3.5)$$

Consider the solution of equation (3.5), and denote it by  $\mathbf{w}^*$  and  $b^*$ . Points  $\Phi(\mathbf{x}_i)$  that satisfy

$$y_i \left[ (\mathbf{w}^*)^T \Phi(\mathbf{x}_i) + b^* \right] = 1$$

are called support vectors (a sparse solution). As seen in Figure 2, SVMs calculate a linear hyperplane by looking for margin maximation, so the solution only depends on the support vectors (Cortes and Vapnik 1995). Usually support vectors just represent a small fraction of the sample, therefore, this fact implies that, the evaluation of the decision function  $D^*(\mathbf{x}) = (\mathbf{w}^*)^T \Phi(\mathbf{x}_i) + b^*$  is computationally efficient. This attractive property is especially useful when dealing with data sets with a low ratio of sample size to dimension (for example, microarray data analysis). SVMs take advantage of this sparsity in the data and are effective even for problems where the data is of dimensionality as large as the number of samples.

A positive definite function,  $K(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{\infty} \lambda_i \Phi_i(\mathbf{x}_1)^T \Phi_i(\mathbf{x}_2)$  called the kernel (Mercer 1909), plays an important role in SVMs. If we restrict  $f(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x})$  to

belong to a reproducing kernel Hilbert space (RKHS), then these functions can be expressed in an alternative form  $f(\mathbf{x}) = \sum_j \alpha_j K(\mathbf{x}_j, \mathbf{x})$ , with  $\mathbf{w}$  replaced by  $\sum_j \alpha_j \Phi(\mathbf{x}_j)$ . As a result, the knowledge of the explicit mapping  $\Phi$  and the vector  $\mathbf{w}$  is not needed, we need only know the kernel  $K$  in its close form. Some basic examples of kernels (with some kernel parameters) are:

1. Linear kernel  $K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$ : the mapping is the identity.
2. Polynomial kernel  $K(\mathbf{x}, \mathbf{y}) = (\gamma \mathbf{x}^T \mathbf{y} + c)^d, \gamma > 0$ : which maps the data into a finite dimensional vector space.
3. Gaussian kernel or Radial Basis Function (RBF)  $K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2)$ : which maps the data into an infinite dimensional space.
4. Sigmoid kernel  $K(\mathbf{x}, \mathbf{y}) = \tanh(\gamma \mathbf{x}^T \mathbf{y} + c)$ : which is a multi-layer perceptron.

We now consider the more general case where the mapped data remain nonseparable. Let  $L$  indicate a loss function, then SVMs address this nonseparability problem by finding a function  $f$  that minimizes an empirical error of the form  $\sum_{i=1}^n L(f(\mathbf{x}_i), y_i)$ . The specific loss function is called hinge loss (Figure 3), defined as

$$L(f(\mathbf{x}_i), y_i) = (1 - y_i f(\mathbf{x}_i))_+, \quad (3.6)$$

with  $(x)_+ = \max(x, 0)$ . Many authors (Ivanov 1976; Phillips 1962; Tikhonov and Arsenin 1977) express the search for a max-margin classifier as a convex optimization problem that maximizes the margin between the data points with a hinge loss penalization for miss-classified or almost miss-classified data points. The most widely used setting minimizes Tikhonov's regularization functional, which consists of solving the optimization problem:

$$\min_{f \in H_K} \frac{1}{n} \sum_{i=1}^n (1 - y_i f(\mathbf{x}_i))_+ + \mu \|f\|_K^2, \quad (3.7)$$

where  $\mu > 0$  controls the trade-off between the fit of the solution  $f$  to the data (measured by  $L$ ) and the approximation capacity of the function space that  $f$  belongs to (measured

by  $\|f\|_K$ ). Based on the work of Kimeldorf and Wahba (1970), it is easy to show that (3.7) can be restated as

$$\min_{\mathbf{w}, b} \frac{1}{n} \sum_{i=1}^n (1 - y_i(\mathbf{w}^T \Phi(\mathbf{x}_i) + b))_+ + \mu \|\mathbf{w}\|^2. \quad (3.8)$$

In order to avoid the nondifferentiable problem due to the hinge loss function in (3.6), Lin et al. (2002) demonstrated that solving problem (3.8) is equivalent to solving

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T \Phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, \dots, n, \end{aligned} \quad (3.9)$$

where  $\xi_i$  are called slack variables and  $C = 1/2\mu n$ . Although (3.9) is the most widely used SVM formulation, in practice, it is usually changed to a standard optimization problem (convex and quadratic) equipped with Lagrange multipliers ( $\lambda_i$ ):

$$\begin{aligned} \min_{\lambda} \quad & \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^n \lambda_i \\ \text{subject to} \quad & \sum_{i=1}^n y_i \lambda_i = 0, \\ & 0 \leq \lambda_i \leq C, i = 1, \dots, n. \end{aligned}$$

In conclusion, SVMs operate within the framework of regularization theory by minimizing an empirical risk. SVMs are consistent (with good asymptotic properties), and their empirical error converges to the expected error, and under some conditions, converges to the Bayes optimal rule. A significant advantage of the SVM is that sparse solutions to classification problems are usually obtained. This fact facilitates the application of SVMs to problems that involve data with high dimensional attributes.

In  $Q$ -learning, we define attributes  $\mathbf{x}_{it} \in \mathbf{S}_t \times \mathbf{A}_t$ ,  $i = 1, \dots, n$ ,  $t = 0, \dots, T$ , where  $\mathbf{S}_t = \{S_0, S_1, \dots, S_t\}$  and  $\mathbf{A}_t = \{A_0, A_1, \dots, A_t\}$ , and we assign the label index  $y_{it}$  to each reward value  $r_{it}$ . In many cases the reward function maps  $(\mathbf{s}_t, \mathbf{a}_t, s_{t+1})$  to a set which

consists of some discrete integer number, and if the size of the set is larger than 2, it is a multicategory classification problem. These kind of problems are often treated as a sequence of binary classifications. For example, the “one versus rest” approach solves  $k$  binary problems through sequential training. But this method may be suboptimal and may yield poor performance due to the absence of a dominating class (Lee, Lin and Wahba 2004). Liu and Shen (2006) proposed a novel multicategory  $\psi$ -learning to treat all classes simultaneously.  $\psi$ -learning can deliver accurate multi-class prediction and outperform its SVM counterpart. Other multi-class classification methods can be found in Crammer and Singer (2001; 2003) and Lee, Lin and Wahba (2004). However, when the number of the classes is large (more than 4) or in the extreme case where  $r_t$  is continuous, and the numerical value is not only the label index but it has meaning, then the multicategory learning methods mentioned above may not be adequate. Therefore, support vector regression (SVR), one of the most popular extensions of SVM, is motivated and discussed in the next section.

### 3.4 Support Vector Regression (SVR)

SVMs were developed to solve the classification problem, but recently they have been extended to the domain of regression problems (Vapnik, Golowich, and Smola 1997). From a mathematical perspective, the support vector regression function is also derived within the RKHS context. In contrast with SVM, one of the popular loss functions involved in SVR is known as the  $\epsilon$ -insensitive loss function (Figure 4), which is defined as

$$L(f(\mathbf{x}_i), y_i) = (|f(\mathbf{x}_i) - y_i| - \epsilon)_+,$$

where  $\epsilon > 0$  (Vapnik, 1995). That is, as long as the absolute difference between the actual and the predicted values is less than  $\epsilon$ , the empirical loss is zero, otherwise there is a cost which grows linearly. SVR is more general and flexible than least-square regression, since

it allows a predicted function that has at most  $\epsilon$  deviation from the actually obtained targets  $y_i$  for all the training data. Other possible loss functions include quadratic loss, Laplace loss, and Huber loss. Similar to equation (3.7), by using the  $\epsilon$ -insensitive loss function, the following optimization problem arises:

$$\min_{f \in H_K} \frac{1}{n} \sum_{i=1}^n (|f(\mathbf{x}_i) - y_i| - \epsilon)_+ + \mu \|f\|_K^2. \quad (3.10)$$

Once more, similar to (3.8), problem (3.10) can be restated as

$$\min_{\mathbf{w}, b} \frac{1}{n} \sum_{i=1}^n (|\mathbf{w}^T \Phi(\mathbf{x}_i) + b - y_i| - \epsilon)_+ + \mu \|\mathbf{w}\|^2. \quad (3.11)$$

Since the  $\epsilon$ -insensitive loss function is also nondifferentiable, (3.11) can be solved by appropriate optimization methods, that is,

$$\min_{\mathbf{w}, b, \xi, \xi'} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n (\xi_i + \xi'_i),$$

$$\begin{aligned} \text{subject to} \quad & (\mathbf{w}^T \Phi(\mathbf{x}_i) + b) - y_i \leq \epsilon + \xi_i, \\ & y_i - (\mathbf{w}^T \Phi(\mathbf{x}_i) + b) \leq \epsilon + \xi'_i, \\ & \xi_i, \xi'_i \geq 0, \quad i = 1, \dots, n. \end{aligned} \quad (3.12)$$

In practice, the following dual convex quadratic formulation with Lagrange multipliers ( $\lambda_i$ ) is used:

$$\begin{aligned} \min_{\boldsymbol{\lambda}, \boldsymbol{\lambda}'} \quad & \frac{1}{2} (\boldsymbol{\lambda} - \boldsymbol{\lambda}')^T K(\mathbf{x}_i, \mathbf{x}_j) (\boldsymbol{\lambda} - \boldsymbol{\lambda}') - \sum_{i=1}^n (y_i - \epsilon) \lambda'_i + \sum_{i=1}^n (y_i + \epsilon) \lambda_i, \\ \text{subject to} \quad & \sum_{i=1}^n (\lambda_i - \lambda'_i) = 0, \\ & 0 \leq \lambda_i, \lambda'_i \leq C, \quad i = 1, \dots, n. \end{aligned} \quad (3.13)$$

The ideas underlying SVR are similar but slightly differ from those within the margin-based classification scheme. In  $\epsilon$ -insensitive SVR (3.12), the slack variables  $\xi_i$  and  $\xi'_i$  allow for some data points in the feature space to stay outside the confidence band determined

by  $\epsilon$ . In other words, the goal is to find a function that has at most  $\epsilon$  deviation from the actually obtained targets  $y_i$  for all the training data. Errors with deviation larger than  $\epsilon$  are not accepted. Once the above formulation is solved to get the optimal  $\lambda_i$  and  $\lambda'_i$ , the approximation function at  $\mathbf{x}$  is given by:

$$f(\mathbf{x}) = \sum_{i=1}^n (\lambda'_i - \lambda_i) K(\mathbf{x}_i, \mathbf{x}) + b. \quad (3.14)$$

There are several examples where SVR are successfully used in practice, and they perform better than other classification methods. See Chen, Chang, and Lin (2001) and Smola and Scholkopf (2004). To achieve good performance by using SVM/SVR, some procedures such as data scaling, kernel and related parameter selection need to be examined very carefully. We discuss those procedures in detail in our simulation example in Section 4.2.

### 3.5 Extremely Randomized Trees (ERT)

The complex and unclear structure of the  $Q$ -function has also partly motivated the vast literature on nonparametric statistical methods and machine learning. Ernst, Geurts, and Wehenkel (2005) and Geurts, Ernst, and Wehenkel (2006) proposed an extremely randomized trees (ERT) method, which is called the Extra-Trees algorithm, for batch mode reinforcement learning. Unlike classical classification and regression trees such as Kd-tree or pruned CART tree, this nonparametric method builds a model in the form of the average prediction of an ensemble of regression trees (called a random forest). Moreover, each tree built by this algorithm consists of strongly randomizing both attribute and cut-point choice while splitting a tree node. In addition to the number of trees  $G$ , this method depends on one parameter, called  $K$ , the maximum number of cut-direction tests at each node, and  $n_{\min}$ , the minimum number of elements at each leaf required to split a node. The choice of an appropriate value of  $G$  depends on the resulting compro-



mise between computational requirements and prediction accuracy.  $K$  determines the strength of the randomization, for  $K = 1$ , the splits are chosen totally independent of the output variable. A larger  $n_{\min}$  yields smaller trees but higher bias. The ERT algorithm builds  $G$  trees using the training data set. To determine a test at a node for each tree, this algorithm randomly selects  $K$  attributes with  $K$  randomized cut-points. A score is calculated for each test and then the one which has the highest value is kept. The algorithm stops splitting a node when the number of elements in the node is less than  $n_{\min}$ . The complete ERT algorithm is given in Figure 5.

Compared to standard tree-based regression methods, ERT successfully leads to significant improvements in precision. Additionally, it can dramatically decrease variance while at the same time decreasing bias, and it is very robust to outliers. ERT has been recently demonstrated in a simulation of HIV infection (Ernst et al., 2006) and adaptive treatment of Epilepsy (Guez et al., 2008). While this algorithm reveals itself to be very effective to extract a well-fitted  $Q$  from the data set, it has one drawback: the computational efficiency is relatively low especially with increasing sample size of patients in the training data set.



Figure 1: Helicopter in autonomous sustained hover (Ng et al., 2006).

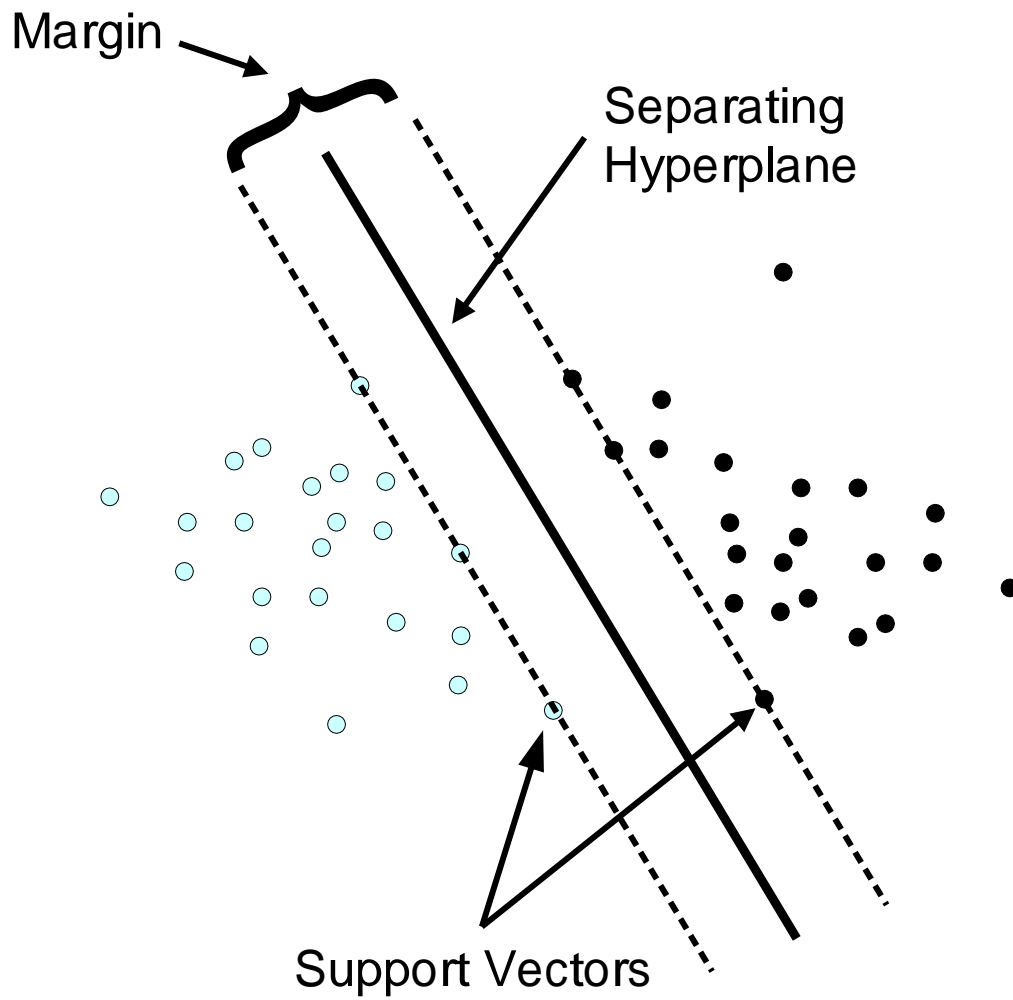


Figure 2: Linear separating hyperplane, margin, and support vectors defined in Support Vector Machines (SVM).

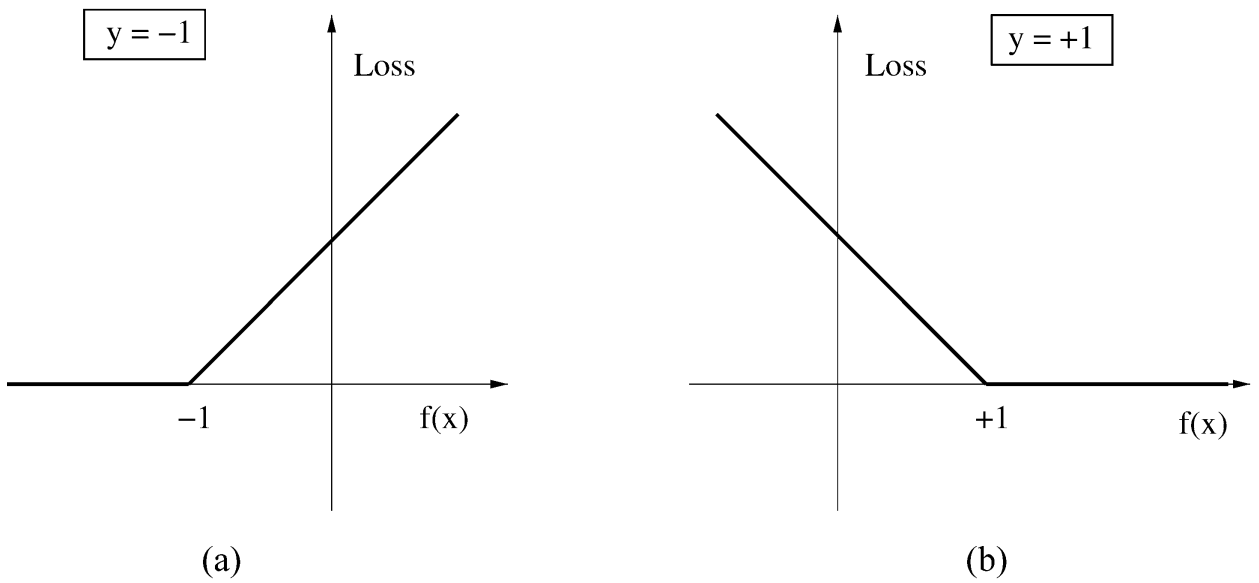


Figure 3: Hinge loss function  $L(f(\mathbf{x}_i), y_i) = (1 - y_i f(\mathbf{x}_i))_+$ . In (a),  $L(f(\mathbf{x}_i), -1)$ ; in (b),  $L(g(\mathbf{x}_i), +1)$ .

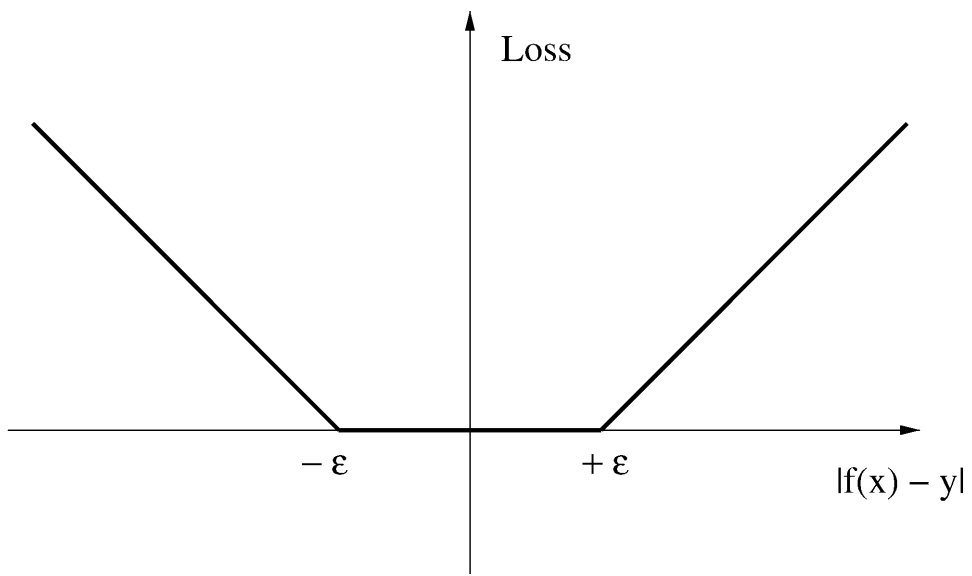


Figure 4:  $\epsilon$ -insensitive loss function  $L(f(\mathbf{x}_i), y_i) = (|f(\mathbf{x}_i) - y_i| - \epsilon)_+$ .

---

**Build\_a\_tree**( $\mathcal{TS}$ )Input: a training set  $\mathcal{TS}$ Output: a tree  $T$ ;

- If
  - (i)  $\#\mathcal{TS} < n_{min}$ , or
  - (ii) all input variables are constant in  $\mathcal{TS}$ , or
  - (iii) the output variable is constant over the  $\mathcal{TS}$ ,return a leaf labeled by the average value  $\frac{1}{\#\mathcal{TS}} \sum_l o^l$ .
- Otherwise:
  1. Let  $[i_j < t_j] = \text{Find\_a\_test}(\mathcal{TS})$ .
  2. Split  $\mathcal{TS}$  into  $\mathcal{TS}_l$  and  $\mathcal{TS}_r$  according to the test  $[i_j < t_j]$ .
  3. Build  $T_l = \text{Build\_a\_tree}(\mathcal{TS}_l)$  and  $T_r = \text{Build\_a\_tree}(\mathcal{TS}_r)$  from these subsets;
  4. Create a node with the test  $[i_j < t_j]$ , attach  $T_l$  and  $T_r$  as left and right subtrees of this node and return the resulting tree.

**Find\_a\_test**( $\mathcal{TS}$ )Input: a training set  $\mathcal{TS}$ Output: a test  $[i_j < t_j]$ :

1. Select  $K$  inputs,  $\{i_1, \dots, i_K\}$ , at random, without replacement, among all (non constant) input variables.
  2. For  $k$  going from 1 to  $K$ :
    - (a) Compute the maximal and minimal value of  $i_k$  in  $\mathcal{TS}$ , denoted respectively  $i_{k,min}^{\mathcal{TS}}$  and  $i_{k,max}^{\mathcal{TS}}$ .
    - (b) Draw a discretization threshold  $t_k$  uniformly in  $]i_{k,min}^{\mathcal{TS}}, i_{k,max}^{\mathcal{TS}}[$
    - (c) Compute the score  $S_k = \text{Score}([i_k < t_k], \mathcal{TS})$
  3. Return a test  $[i_j < t_j]$  such that  $S_j = \max_{k=1, \dots, K} S_k$ .
- 

Figure 5: Complete algorithm used by extremely randomized trees (ERT) to build a random forest (Geurts et al., 2006).  $\mathcal{TS}$  denotes training set,  $(i^l, o^l)$  denotes input-output pair.

## 4 Reinforcement Learning Treatment Strategies for A Virtual Cancer Trial

### 4.1 Clinical Reinforcement Trials

In the previous chapter, we introduced reinforcement learning to cancer clinical trials. The main advantage offered by using reinforcement learning in clinical trials is that discovery is included in the trial itself, not just evaluation as is the case for standard Phase III clinical trials. There are a number of ways this discovery occurs, but one of the key ways is by the manner in which patient differences are leveraged to enable discovery of effective treatments missed by standard clinical trial designs. Suppose, for example, we have two drugs,  $A$  and  $B$ , and a continuous biomarker  $X$  which varies from patient to patient. Suppose also that when  $X$  is less than or equal to its median value  $M$ , treatment  $A$  is twice as effective as treatment  $B$ ; but that when  $X$  is greater than  $M$ , treatment  $A$  is half as effective as treatment  $B$  (i.e., it is harmful). In a standard randomize trial, the benefit of treatment  $A$  would be washed out and completely undetected. However, if we leveraged individual differences, we would find out that treatment  $A$  is very effective when given to people for whom  $X \leq M$  but not given when  $X > M$ . Thus the proposed approach is not just a nuanced improvement over standard clinical trials but a paradigm shift in methods for discovering effective treatments.

In this section, we propose a new design and analysis method for a new kind of clinical trial for life threatening diseases, “clinical reinforcement trials”. The design for these trials consists of three aspects:

First, a finite, reasonably small set of decision times is identified. These times could be either specific time points measured from trial onset or decision points in the treatment process such as the starting times of a each new line of cancer treatment. For example, in the simulation study below, we create a synthetic cancer treatment setting where patients are monitored monthly for six months and treatment for each month is determined based on patient biomarker values available at the beginning of the month. As a second example, in NSCLC, it may be more appropriate to have one decision time at the beginning of the first line of treatment, a second decision time at the beginning of the second line of treatment, and possibly a third decision time at the beginning of the third line of treatment. The third line is currently only available for certain patients and there is only one FDA approved third line treatment, and so decision possibilities are severely limited at the third decision time. Note that the decision time in this instance is really a stage of treatment and not a calendar time. Other decision time sets, including hybrid variants of the previous two examples, are also possible.

Second, for each decision time, a set of possible treatments to be randomized is identified. The choice of treatments can be a continuum as mentioned earlier or a finite set and can include restrictions which may be functions of observed variables such as biomarkers. For example, in our simulations we restrict the dose of chemotherapy at the first decision time to be above a threshold so that all patients are guaranteed some initial treatment. When the set of treatments is finite, the proposed design reduces to a SMART design.

Third, a utility function is identified which can be assessed at each time point and which contains an appropriately weighted combination of outcomes available at each interval between decision times and at the end of the final treatment interval. In our simulation study below, we use a combination of tumor size and overall patient health as our utility function.

Once the design has been determined, patients are then recruited into the study



and randomized to the treatment set under the protocol restrictions at each decision point, outcome measures used to compute patient state and utility are obtained, and each patient is followed through to completion of the protocol or until the end of the trial. The patient data is collected and  $Q$ -learning is applied, in combination with either SVR or ERT applied at each time point as described above, to estimate the optimal treatment rule as a function of patient variables and biomarkers, at each decision time. We allow the  $Q$ -functions to differ from decision time to decision time. We will show in the simulation study below that our proposed approach is able to generate treatment rules that lead to improved patient outcomes. One open question which we will pursue in a later paper is sample size guidelines. Fortunately, it appears from our simulation studies that the sample sizes required are similar to and not larger than the sizes required for typical phase III trials.

## 4.2 A Virtual Clinical Reinforcement Trial

In this section we simulate a sequentially randomized clinical reinforcement trial as a numerical example to examine the performance of the proposed design and methodology. To demonstrate that the optimal therapy found using  $Q$ -learning is superior to any other regimens, the treatments at each course are specified in terms of a continuum of dose levels of a single drug, and the comparisons we consider are between the optimal regimen identified from our proposed clinical reinforcement trial procedure and various constant-dose regimens. We first present a simple mathematical model for disease and chemotherapy which we will be using for our study. We then present the specific implementation of  $Q$ -learning which we will use for the simulation. This section concludes with a presentation of the results of the simulation study.

### 4.2.1 A Simple Chemotherapy Mathematical Model

As discussed in Section 2.2.1, there exists a large volume of literature concerning mathematical models of chemotherapy. To construct a set of training data reflecting a hypothetical cancer trial, we need a simple chemotherapy mathematical model capable of describing the fundamental principles governing tumor progression and responses to therapy. The goal for a chemotherapy mathematical model is to allow for sufficient complexity so that the model will qualitatively generate clinically observed in vivo tumor growth patterns, while simultaneously maintaining sufficient simplicity to admit analysis. Thus, inspired by discussions in Section 2.2.1, a sophisticated model we present must exhibit: (1) tumor growth in the absence of chemotherapy; (2) patients' negative wellness outcomes in response to chemotherapy; (3) the drug's capability for killing tumor cells while also increasing toxicity; (4) an interaction between tumor cells and patient wellness. To obtain data which satisfy these requirements, we propose a system of ordinary difference equations (ODE) modeled as follows:

$$\begin{aligned}\dot{W}_t &= a_1(M_t \vee M_0) + b_1(D_t - d_1), \\ \dot{M}_t &= \left[ a_2(W_t \vee W_0) - b_2(D_t - d_2) \right] \times \mathbf{1}\{M_t > 0\},\end{aligned}\tag{4.1}$$

where time (with month as unit)  $t = 0, 1, \dots, T - 1$ . Note that these changing rates yield a piecewise linear model over time. Without loss of trade-off between toxicity and efficacy, the piecewise linear model can be implemented very easily. For simplicity, we here consider tumor size instead of number of tumor cells.  $M_t$  denotes the tumor size at the specified time,  $M_0$  indicates the value of tumor size when the patient is at the beginning of the study.  $W_t$  measures the negative part of wellness (toxicity), similarly,  $W_0$  indicates the initial value of patient's wellness.  $D_t$  denotes the chemotherapy drug level. The value of other different parameters for the model are fixed as:  $a_1 = 0.1$ ,  $a_2 = 0.15$ ,  $b_1 = 1.2$ ,  $b_2 = 1.2$ ,  $d_1 = 0.5$  and  $d_2 = 0.5$ . The indicator function term  $\mathbf{1}\{M_t > 0\}$  in (4.1) represents the feature that when tumor size is absorbed to 0, the patient has been

cured, and there is no future recurrence of the tumor. Note that this model is not meant to reflect a specific cancer but to reflect a generic plausible cancer created for illustration.

Before generating simulated clinical data, it is easy to notice that the dynamic model has two state variables  $(W_t, M_t)$  and one action (treatment) variable  $(D_t)$ . The state variables can be obtained via:

$$W_{t+1} = W_t + \dot{W}_t,$$

$$M_{t+1} = M_t + \dot{M}_t,$$

where  $t = 0, 1, \dots, T - 1$  are the  $T$  decision times we will utilize in our simulated trial design. We generate a simulated clinical reinforcement trial with  $N = 1000$  patients (replicates) with each simulated patient experiencing 6 months ( $T = 6$ ) of treatment based on this ODE model. The initial values  $W_0$  and  $M_0$  for each patient are generated from independent uniform  $(0, 2)$  deviates. The treatment set consists of doses of a chemotherapy agent with acceptable dose range of  $[0, 1]$ , where the value 1 corresponds to the maximum acceptable dose. The values chosen for chemotherapy drug level  $D_0$  are simulated from the uniform  $(0.5, 1)$  distribution, moreover,  $D_1, \dots, D_5$  are drawn according to a uniform distribution in the interval  $(0, 1)$ . Thus our treatment set is restricted differently at decision time  $t = 0$  than at other decision times to reflect a requirement that patients receive at least some drug at onset of treatment. Various other distribution settings for the action space are possible, and clinical researchers have tremendous flexibility when designing clinical reinforcement trials.

Figure 6 provides a disease progression example of one patient to show dynamic treatment results with influence of different levels of chemotherapy drug. The system is clearly sensitive to the chemotherapy dosing regimen. Note that when the dose level switches to low, the tumor size grows to a dangerous level. Moreover, the toxicity increases (decreases) once the dosage is changed to a higher (lower) level. By applying reinforcement learning to this crude computer model, we aim at uncovering the ideal regimen which has the best trade-off between efficacy and toxicity.

### 4.2.2 $Q$ -function Estimation and Optimal Regimen Discovery

We now return to  $Q$ -learning. Based on the proposed ODE model, we can generate a simulated clinical trial that provides a set of simulated finite horizon trajectories (the training data),

$$\{S_{0i}, A_{i0}, r_{i0}, S_{i1}, A_{i1}, r_{i1}, \dots, A_{i5}, r_{i5}, S_{i6}\}_{i=1}^{1000},$$

where each two-dimensional state variable  $S_t$  consists of  $(W_t, M_t)$ , and each continuous action variable  $A_t$  is a dose level  $D_t$ . In terms of optimality criterion, we seek effective regimens that maximize a sum of numerical rewards over six months. We assume each reward only depends on the states observed right before and after each action, that is, when  $t = 0, 1, \dots, 5$ ,

$$r_t = R(s_t, a_t, s_{t+1}).$$

We decompose this reward function  $r_t$  into three parts:  $R_1(D_t, W_{t+1}, M_{t+1})$  due to survival status,  $R_2(W_t, D_t, W_{t+1})$  due to wellness effects, and  $R_3(M_t, D_t, M_{t+1})$  due to tumor size effects. It can be described by:

$$R_1(D_t, W_{t+1}, M_{t+1}) = -60, \quad \text{if patient died,}$$

otherwise,

$$R_2(W_t, D_t, W_{t+1}) = \begin{cases} 5 & \text{if } W_{t+1} - W_t \leq -0.5, \\ -5 & \text{if } W_{t+1} - W_t \geq 0.5, \\ 0 & \text{otherwise,} \end{cases}$$

$$R_3(M_t, D_t, M_{t+1}) = \begin{cases} 15 & \text{if } M_{t+1} = 0, \\ 5 & \text{if } M_{t+1} - M_t \leq -0.5, \text{ but } M_{t+1} \neq 0, \\ -5 & \text{if } M_{t+1} - M_t \geq 0.5, \\ 0 & \text{otherwise.} \end{cases}$$

In most phase III clinical trials, the primary endpoint of clinical interest is the overall survival (OS), that is why we put  $-60$  as a high penalty for patient's death. Additionally, we assigned the relative high value  $15$  as a bonus when a patient is cured.

We assume that survival status depends on both toxicity and tumor size. For each time interval  $(t-1, t], t = 1, \dots, 6$ , we define the hazard function as  $\lambda(t)$ , where

$$\log\lambda(t) = \mu_0 + \mu_1 W_t + \mu_2 M_t,$$

and  $\mu_0, \mu_1$ , and  $\mu_2$  are constant pre-specified parameters. In particular, assigning  $\mu_1 = \mu_2 = 1$  indicates that we consider wellness and tumor size to have an equally weighted influence on the survival rate. The survival function is

$$\Delta F(t) = \exp[-\Delta\Lambda(t)],$$

where  $\Delta\Lambda(t) = \int_{t-1}^t \lambda(s)d(s)$  is the cumulative hazard function. The reason the term  $R_1(D_t, W_{t+1}, M_{t+1})$  is expressed as a function of  $W_{t+1}$  and  $M_{t+1}$  is that the hazard function is only determined by the states at the end of each time interval. The conditional probability of death for each time interval is  $p = 1 - \Delta F(t)$ . The survival status (with death coded as 1) is drawn according to a Bernoulli distribution  $B(p)$ . Overall, by letting  $\gamma = 1$  (we would like to fully consider maximizing rewards in the long run), the one-step  $Q$ -learning with recursive form is utilized:

$$Q_t(S_t, A_t) \leftarrow r_t + \max_{a_{t+1}} Q_{t+1}(S_{t+1}, A_t, a_{t+1}),$$

where  $r_t = R_1(D_t, W_{t+1}, M_{t+1}) + R_2(W_t, D_t, W_{t+1}) + R_3(M_t, D_t, M_{t+1}), t = 0, \dots, 5$ .

To obtain the estimator  $\widehat{Q}_t$ , we apply SVR and ERT respectively for fitting  $Q_t$  backward, and save the results as  $\{\widehat{Q}_5, \widehat{Q}_4, \dots, \widehat{Q}_0\}$ . Figure 7 illustrates the treatment plan and relevant  $Q$ -function estimation procedures. Because of the inner product property of the kernel in SVM/SVR, scaling the data before applying SVR is very important. Another advantage for scaling is to avoid states with greater numeric ranges dominating those with smaller numeric ranges. In our simulation studies, every variable is scaled to

zero mean and unit variance, and the center and scale values are saved and used for later predictions. To do fitting of  $\widehat{Q}_t$  via SVR, we select the Gaussian kernel (or Radial Basis Function),  $K(\mathbf{x}, \mathbf{y}) = \exp(-\zeta\|\mathbf{x} - \mathbf{y}\|^2)$ , because the Gaussian kernel can nonlinearly map samples into a higher dimensional space. Consequently, it can handle the case when the relation between rewards (labels) and states and actions (attributes) is nonlinear. In the SVR approach there are two hyperparameters involved with the Gaussian kernel:  $\zeta$  and the tuning parameter  $C$ . To maximize the performance of the proposed method, we apply a grid search to choose  $C$  and  $\zeta$  by using cross-validation. Trying exponentially growing sequences of  $C$  and  $\zeta$  is recommended as a practical method to identify good hyperparameters. Specifically, for each  $t$  in our simulated example, given a straightforward coarse grid search with  $C = 2^{-5}, 2^{-3}, \dots, 2^{15}$  and  $\zeta = 2^{-15}, 2^{-13}, \dots, 2^3$ , we apply cross-validation to each candidate pair  $(C, \zeta)$ , and then select the pair that yields the highest cross-validation rate. To fit  $\widehat{Q}_t$  via ERT, we need to be careful with the choice of parameters  $G$ ,  $K$  and  $n_{\min}$ . Based on empirical studies, Geurts *et al.* (2006) suggest that the default value of  $K$  should be equal to the number of attributes in the regression problem. Thus we fix  $K$  as the dimension of state variables plus the dimension of action variables, which is equal to 3 in our case. To maintain good precision and small bias,  $G$  and  $n_{\min}$  have been chosen equal to 50 and 2, respectively.

In order to evaluate how the above estimated treatment policies perform, we generate an additional 200 patients having initial values  $W_0$  and  $M_0$  randomly chosen from the same uniform distribution adopted in the training data. Based on the sequential estimator  $\{\widehat{Q}_5, \widehat{Q}_4, \dots, \widehat{Q}_0\}$ , when  $t = 0, 1, \dots, 5$ , the individualized optimal policy calculations are carried out using:

$$\widehat{\pi}_t = \arg \max_{a_t} \widehat{Q}_t(s_t, a_t; \widehat{\theta}_t).$$

The entire algorithm for  $Q$ -function estimation and optimal regimen discovery is summarized as follows:

1. Inputs: a set of training data consisting of attributes  $\mathbf{x}$  (states  $s_t$ , actions  $a_t$ ) and

- index  $\mathbf{y}$  (rewards  $r_t$ ), i.e.,  $\{(s_t, a_t, r_t)_i, t = 0, \dots, T, i = 1, \dots, N\}$ .
2. Initialization: Let  $t = T + 1$  and  $\widehat{Q}_{T+1}$  be a function equal to zero on  $\mathbf{S}_t \times \mathbf{A}_t$ .
  3. Iterations: repeat computations until stopping conditions are reached ( $t = 0$ ):
    - (a)  $t \leftarrow t - 1$ .
    - (b)  $Q_t$  is fitted with the support vector regression (SVR) or extremely randomized trees (ERT) through the following recursive equation:
 
$$Q_t(s_t, a_t) = r_t + \max_{a_{t+1}} Q_{t+1}(s_{t+1}, a_{t+1}).$$
    - (c) Use cross-validation to choose tuning parameters  $C$  and  $\zeta$  to fit  $Q_t$  via SVR with Gaussian kernel; choose plausible values of parameters  $K, G, n_{min}$ , and fit  $Q_t$  via ERT ( $K = 3, G = 50, n_{min} = 2$  in our simulation).
  4. Given the sequential estimates of  $\{\widehat{Q}_0, \widehat{Q}_1, \dots, \widehat{Q}_5\}$ , the sequential individualized optimal polices  $\{\widehat{A}_0, \dots, \widehat{A}_5\}$  for new patients in the testing dataset can be predicted one by one.

### 4.2.3 Simulation Results

In our analysis, we first evaluate the operating characteristics of 10 different constant doses (0.1, 0.2, ..., 0.9, 1.0). For comparison, we also evaluate patients' subsequent outcomes ( $W_t$  and  $M_t$ ) conducted by our estimated optimal regimens. In addition, we examine the properties of cumulative survival probability and the computed optimal strategies. All of these numbers are averaged over 200 repeated simulations. When the simulated testing trial ends at  $t = 6$ , all the results of our comparison are summarized in Table 1.

We used a sample size of 1000 for our simulated clinical reinforcement trial and estimated the optimal treatment policy using both SVR and ERT. For the sake of simplicity, unless stated explicitly otherwise, we only show figure results for the SVR method, since we obtain very similar results when we estimate optimal therapy using ERT. On Figure 8 and 9, trajectories (wellness and tumor size, respectively) that would have been observed by putting the patients on constant-dose regimens have been plotted. Note that the wellness measure has been inverted so that larger values represent worse health. This is to make comparisons with tumor size more direct. We test the behavior of estimated optimal regimens on 200 new simulated patients by comparing the outcomes using  $\hat{\pi}_t$  from the  $\hat{Q}_t$  ( $t = 0, \dots, 5$ ) against the results obtained using 10 different fixed  $D_t$  ( $t = 0, \dots, 5$ ) in the ODE model. As shown in both Figure 8 and Figure 9, the optimal regimens derived from  $Q$ -learning do not have better performance compared to some constant dosing regimens. This is not beyond our expectation. Because when higher dose level decreases tumor size, it can bring a higher toxicity simultaneously, and vice versa. However, due to our reward functions structure, the estimated optimal policies have an appealing feature to seek a good balance between toxicity and efficacy. Figure 10 illustrates that the estimated optimal regimen is absolutely superior to any constant-dose regimen when we combine toxicity and efficacy ( $W_t + M_t$ ) as one comparison criterion. Table 1 agrees with this conclusion by respectively presenting  $W_6 + M_6 = 3.269$  (SVR) or  $W_6 + M_6 = 3.194$  (ERT) as the lowest number compared to the others. Most notably, although the regimen derived from simulated data shows suboptimal results in the first three months, it achieves the best performance eventually. These findings agree well with reinforcement learning’s substantially powerful long-run capabilities.

Figure 11 provides the dynamic optimal regimen for an individual patient as well as the effect values (toxicity and efficacy) in the whole trial. This simulated patient comes into the trial with initial condition  $W_0 = 0.30$  and  $M_0 = 1.05$ . Optimal therapy begins with a very high dose  $D_0 = 1.00$  aimed at reducing the patient’s tumor burden.



The patient is then monitored for the following month and then treated with another two consecutive high doses ( $D_1 = 0.74$ ,  $D_2 = 1.00$ ). In the third month, the tumor size suddenly reaches 0, i.e., the patient has been cured. As expected, we find that the dosage to be administrated rapidly reduces to 0 in the following months. Patients who recover after three months will not receive high dosing anymore because the high dose will likely result in unnecessarily high toxicity. As we can see, rather than the constant dose level for each  $t$ , optimal therapy usually has an up-and-down structure due to its adaptive properties. This is an important result to demonstrate that the optimal policy can be approximated very well by reinforcement learning.

At last, compared to all fixed-level doses, Table 1 and Figure 12 clearly show that the therapy found using the  $Q$ -learning approach with either SVR and ERT has better performance in terms of cumulative survival probability (CSP) over 6 months. Table 1 also shows that both SVR and ERT appear to perform equally well with comparable computational burden. Additionally, we plot the average optimal strategies (regimens) in Figure 13. As we can see, rather than the constant dose level for each  $t$ , optimal therapy usually has an up-and-down structure due to its adaptive feature.

#### 4.2.4 Summary of Virtual Cancer Trial Results

We have developed a reinforcement learning method for discovering effective therapeutic regimens in clinical trial design. To investigate the validity of such a purely data (model-free) driven approach, we have generated clinical data by relying on a set of hypothetical (and simplistic) but plausible ODE models. Based on these simulated data, we have found that reinforcement learning is indeed able to identify individualized optimal regimens in clinical trials which consists of multiple courses of treatment. Such regimens can reduce tumor burden while taking into account a drug's toxicity. Treatment delay effects, which is an important issue that must be considered for longer term outcomes, are fully assessed

by this method. Another appealing feature of our approach is the incorporation of  $Q$ -learning methodology with SVR and ERT. Hence even in a data set comprised of high-dimensional attributes, our method is capable of obtaining promising results without much computational burden.

Since a choice of reward function plays a crucial role in reinforcement learning, therefore, it is very important to consider alternative rewards directly reflecting primary endpoints (such as overall survival, progression-free survival, side effects, etc.) in clinical trial designs. One of many feasible approaches is to perform retrospective analysis to identify clinical factors that influence the outcome of patients treated with chemotherapy drugs, and to build a model that can be used in practice to predict long-term survival in this patient population. Such a model may assist us in building a more plausible reward function, and thereafter determining a regimen which is as close as possible to an optimal policy. To conduct such clinically relevant reward functions, we believe that close collaboration with clinical researchers is required. An interesting illustrative example of a related strategy is shown by Ernst et al. (2006). They consider discounted instantaneous costs (which is a continuous function directly associated with actions) as their reward function: the rationale behind this comes from a validated and identified HIV model (Adams et al., 2004).

Since the work of this study is motivated by the clinical question of proper treatment for Stage IIIB/IV NSCLC, as examined by several clinical trials conducted at the UNC Lineberger Comprehensive Cancer Center (LCCC), an important application is to refine our model to more accurately reflect NSCLC and the associated treatment issues. The goal of the study is to compare strategies for multiple lines of treatment for patients with advanced NSCLC who have not been treated previously with systemic therapy. In Chapter 5 we will apply reinforcement learning to discover individualized optimal regimens while restricting attention to first-line and second-line only, since there is only one approved agent (Erlotinib) indicated for third-line treatment (Shepherd et al., 2005).

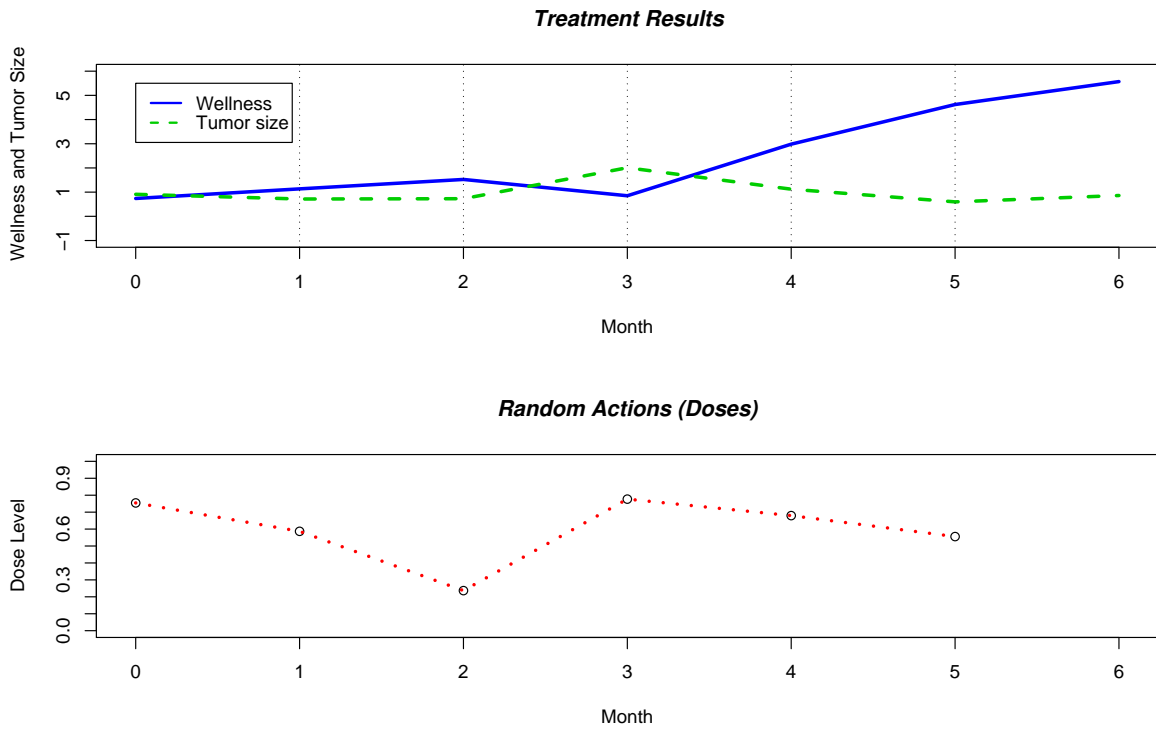


Figure 6: Representation of the disease progression for a patient treated with a randomized chemotherapy drug. The solid curve represents the negative part of patient's wellness, the dashed curve represents the tumor size, and the dotted curve represents the randomized treatment.

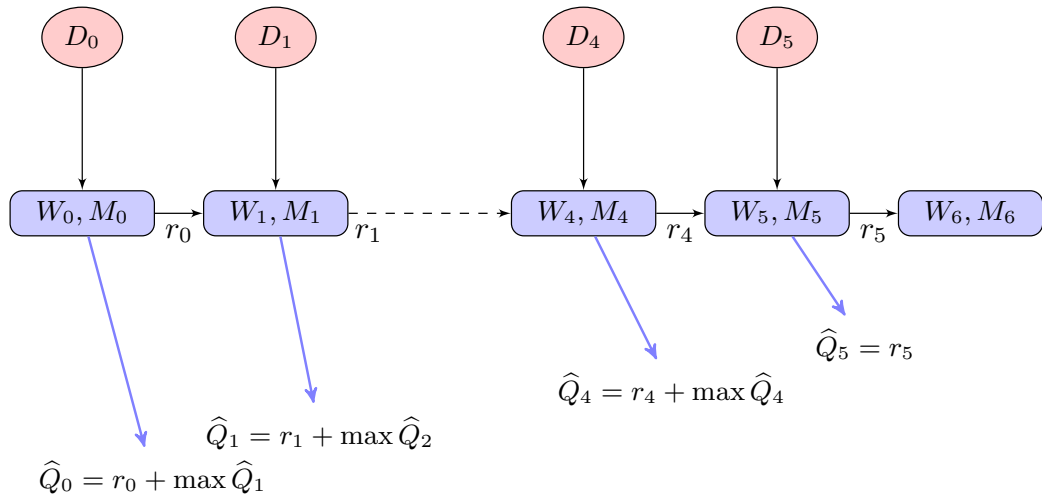


Figure 7: Treatment plan and the procedure for obtaining the sequential estimator  $\{\hat{Q}_5, \hat{Q}_4, \dots, \hat{Q}_0\}$ .

Table 1: Summary of main simulation results

	Optimal Regimens		Constant-Dose Regimens									
	RL <sup>1</sup>	RL <sup>2</sup>	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$W_6$	<b>1.276</b>	<b>1.328</b>	-0.411	0.129	0.669	1.217	1.783	2.375	3.016	3.705	4.421	5.141
$M_6$	<b>1.993</b>	<b>1.866</b>	4.737	4.017	0.300	2.654	2.133	1.658	1.203	0.812	0.496	0.257
$W_6 + M_6$	<b>3.269</b>	<b>3.194</b>	4.326	4.146	3.970	3.870	3.916	4.033	4.219	4.517	4.917	5.397
CSP	<b>0.442</b>	<b>0.441</b>	0.240	0.292	0.345	0.377	0.363	0.331	0.275	0.189	0.061	0.003

53

Note: Results such as the wellness ( $W_6$ ), tumor size ( $M_6$ ), wellness plus tumor size ( $W_6 + M_6$ ), and cumulative survival probability (CSP) over 6 months are given for the optimal regimen using reinforcement learning methods via SVR (RL<sup>1</sup>) and ERT (RL<sup>2</sup>), and for the constant-dose regimen which ranges from 0.1 to 1.0. All numbers are averaged over 200 patients. Entries for superior strategies are given in boldface type.

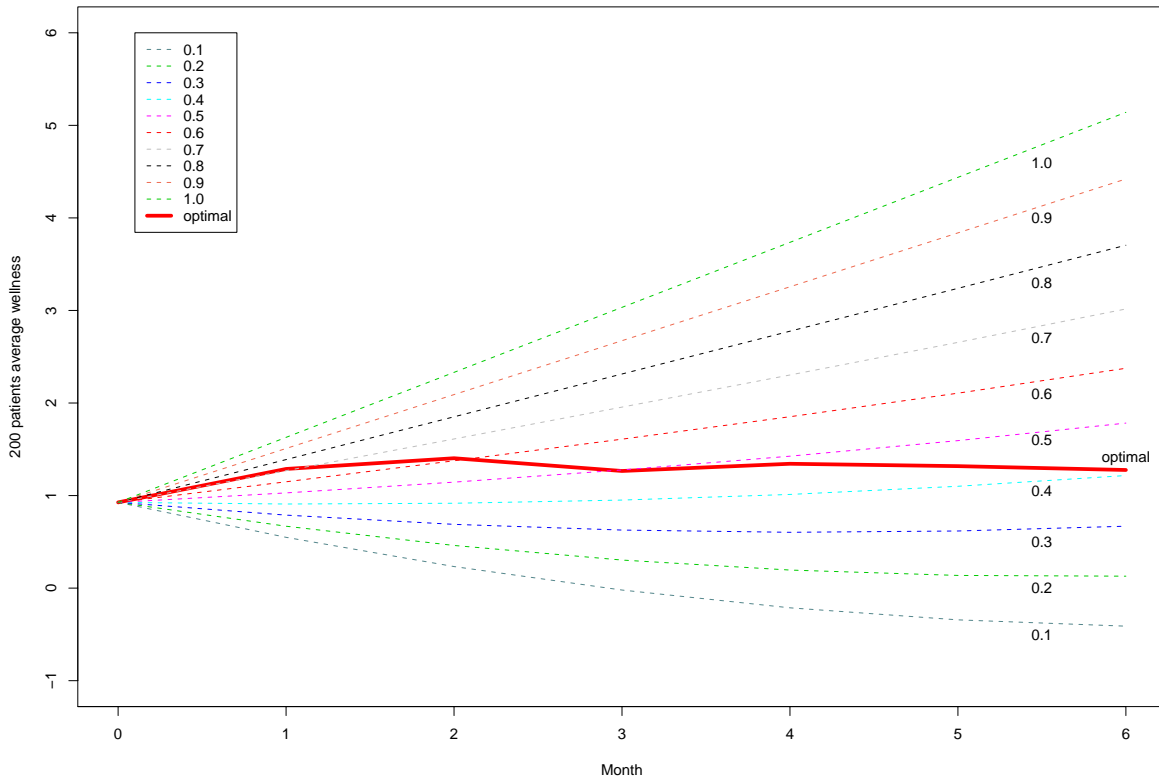


Figure 8: Plots of averaged value of “wellness” for 10 different constant-dose regimens compared to optimal regimen. The results are based on 200 patients. Dashed curves represent the constant-dose regimens, and a solid curve represents the optimal regimen.

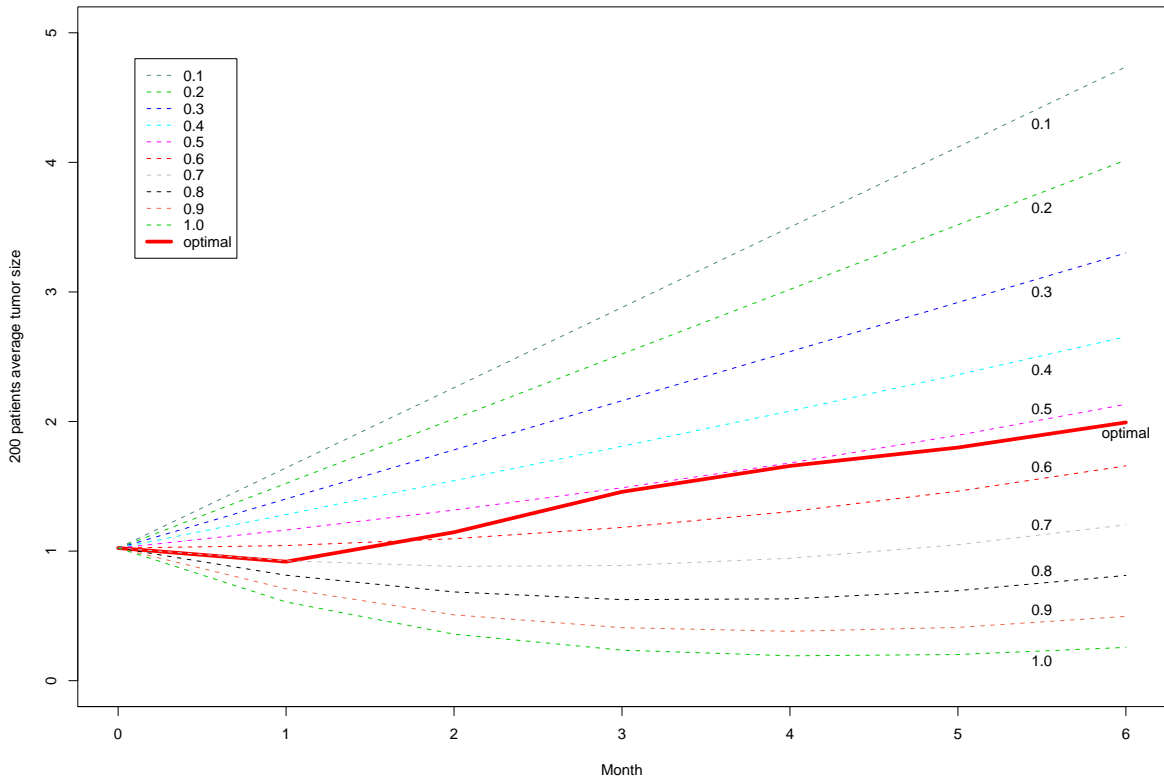


Figure 9: Plots of averaged value of “tumor size” for 10 different constant-dose regimens compared to optimal regimen. The results are based on 200 patients. Dashed curves represent the constant-dose regimens, and a solid curve represents the optimal regimen.

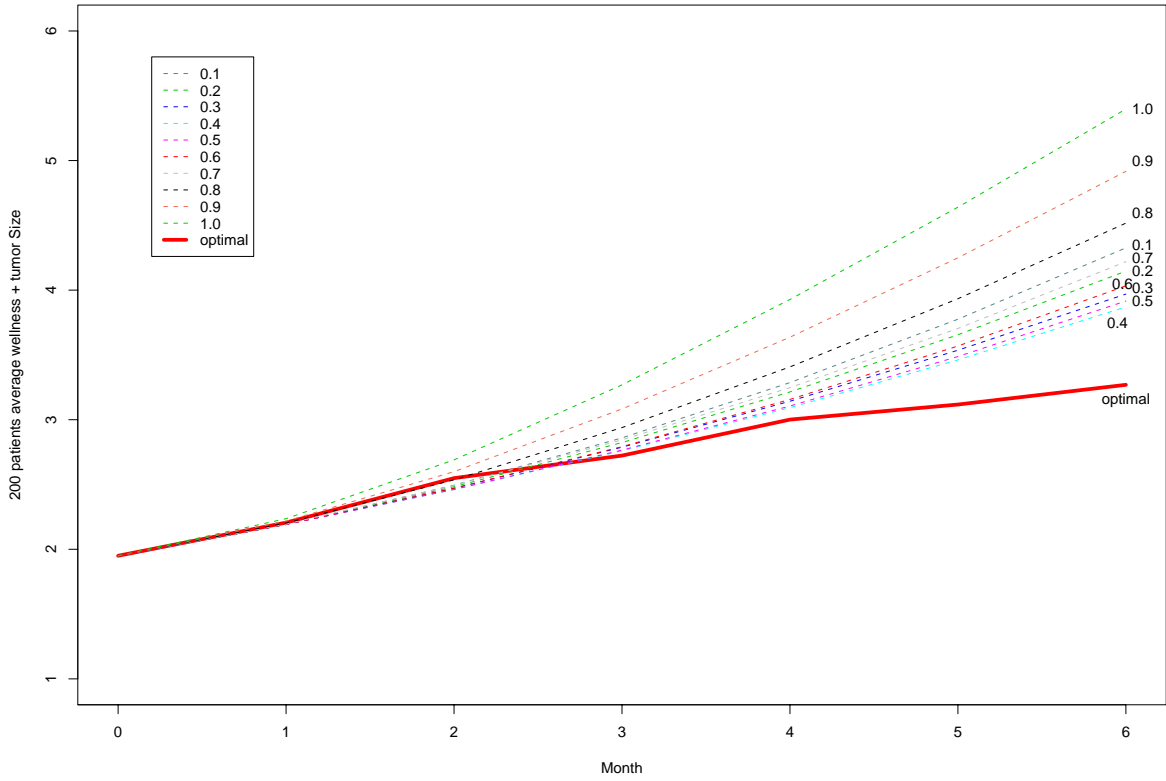


Figure 10: Plots of averaged value of “wellness + tumor size” for 10 different constant-dose regimens compared to optimal regimen. The results are based on 200 patients. Dashed curves represent the constant-dose regimens, and a solid curve represents the optimal regimen.



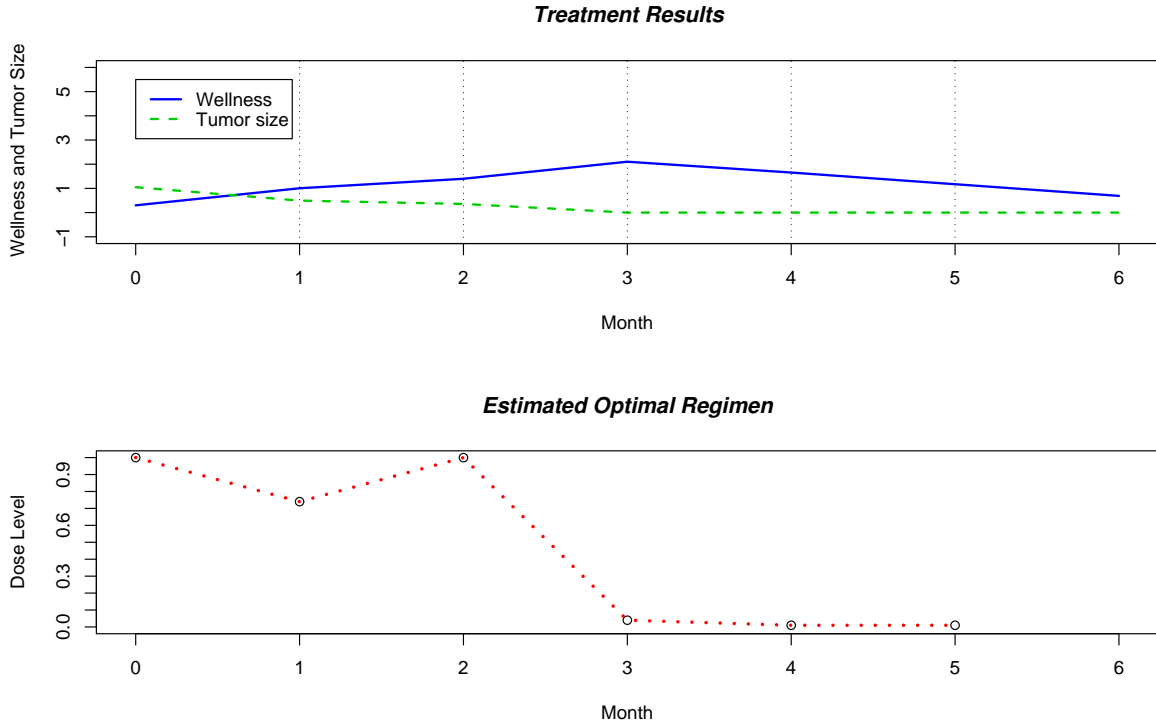


Figure 11: Representation of the optimal treatment for a patient with  $W_0 = 0.30$  and  $M_0 = 1.05$ . The optimal treatment sequence ( $D_t \in \{1.00, 0.74, 1.00, 0.04, 0.01, 0.01\}$ ) is computed by the reinforcement learning methods on clinical data generated by 1000 patients. The solid curve represents the negative part of patient's wellness, the dashed curve represents the tumor size, and the dotted curve represents the estimated optimal regimen.

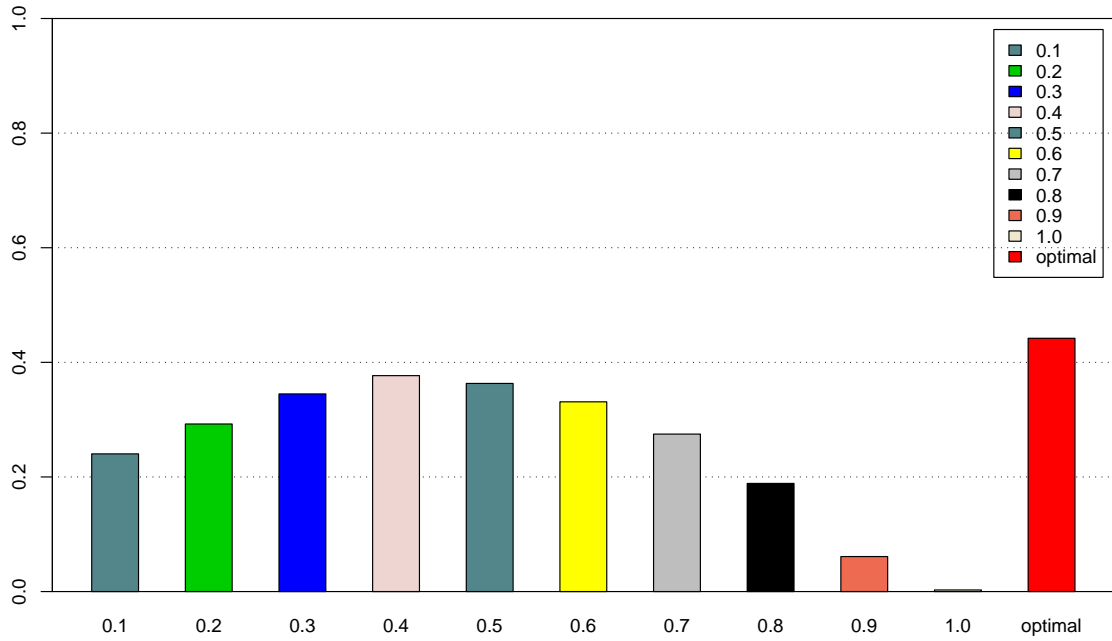


Figure 12: Bar plots of averaged cumulative survival probability at 6 months for 10 different constant-dose regimens compared to optimal regimen. The results are based on 200 patients.

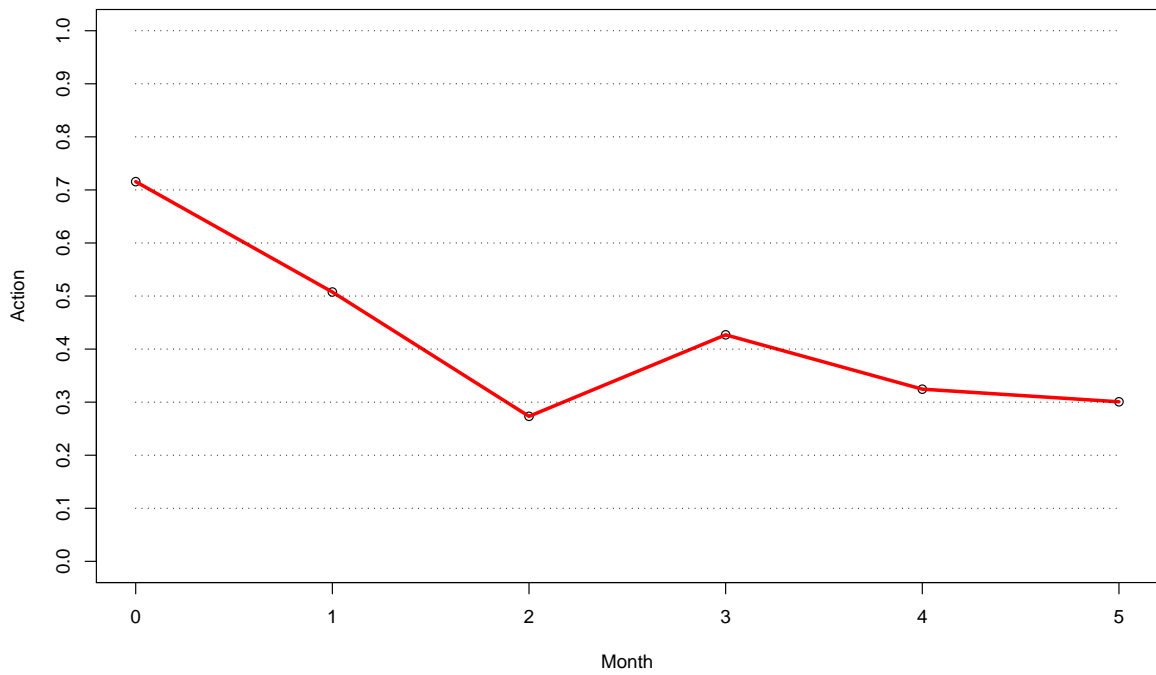


Figure 13: The averaged optimal sequential therapies ( $D_t \in \{0.72, 0.51, 0.27, 0.43, 0.32, 0.30\}$ ) for 200 patients. Dashed curves represent constant-dose regimens, and a solid curve represents the optimal regimen.

# 5 Reinforcement Learning Treatment Strategies Based on Support Vector Regression in a Non-small Cell Lung Cancer Trial

## 5.1 Introduction

In Chapter 4 we utilized reinforcement learning to discovery optimal regimen for a virtual cancer trial. In this chapter, we will further extend our methodology to directly address the assessment of first and second lines of treatment in advanced non-small cell lung cancer (NSCLC).

There has been significant recent research activity in developing therapies that are tailored to each individual. Finding such therapies in treatment settings involving multiple decision times is a major challenge. For example, in treating advanced NSCLC, patients typically experience two or more lines of treatment, and many studies demonstrate that three lines of treatment can improve survival for patients. Discovering tailored therapies for these patients is a very complex issue since effects of covariates (such as established prognostic factors or biomarkers) must be modelled within the multi-stage structure. In this dissertation, we present a new kind of NSCLC clinical trial, based on reinforcement learning methods from computer science, that statistically finds an optimal individualized treatment plan at each decision time which is a function of available patient prognostic information.

For NSCLC patients who present with a good performance status and stage IIIB/IV disease, platinum-based chemotherapy is the primary treatment which offers a modest

survival advantage over best supportive care (BSC) alone. First-line treatment primarily consists of doublet combinations of platinum compounds (cisplatin or carboplatin) with gemcitabine, pemetrexed, paclitaxel, or vinorelbine (Scagliotti et al., 2008; Sandler et al., 2006; Pirker et al., 2008). These drugs modestly improved the therapeutic index of therapy, but no combination seemed superior. More recently, the addition of bevacizumab, a monoclonal antibody against vascular endothelial growth factor (VEGF), to carboplatin and paclitaxel has been shown to produce a higher response rate and longer progression-free survival and overall survival times (Sandler et al., 2006). However, this phase III study was only designed to investigate patients with histologic evidence of non-squamous cell lung cancer. Therefore, in first-line treatment of NSCLC trial, a very important clinical question is what tailored treatment to administer based on each individual's prognostic factors (including the patient's histology type, toxicity profile, smoking history, and VEGF level, etc.), among many approved first-line treatments.

All patients with advanced NSCLC who initially received a platinum-based first-line chemotherapy inevitably experience disease progression. Approximately 50-60% of patients on recent phase III first-line trials received second-line treatment (Sandler et al., 2006). Similar to the first-line regimen, three FDA approved second-line agents (docetaxel, pemetrexed, and erlotinib) appear to have similar response and overall survival efficacy but very different toxicity profiles (Shepherd et al., 2000; Ciuleanu et al., 2008; Shepherd et al., 2005). The choice of agent should also mainly depend on a number of factors, including the patient's comorbidities, toxicity from previous treatments, and the risk for neutropenia. A better understanding of prognostic factors in the second-line setting may allow clinicians to better select patients for second-line therapy and lead to better designed second-line trials.

The current standard treatment paradigm is to initiate second-line therapy at the time of disease progression. Recently there have been two phase III trials that have investigated other possible timings of initiating second-line therapy (Fidias et al., 2007;

Ciuleanu et al., 2008). Both of these trials have revealed a statistically significant improvement in the progression-free survival, and a trend towards improved survival for the earlier use of second-line therapy. However, in terms of considering overall survival as the primary endpoint, a nonsignificant difference has been also revealed by these two trials. Stinchcombe and Socinski (2009) claimed that even under the best of circumstances not all patients will benefit from the early initiation of second-line therapy. Hence the proper selection of patients is also critical to determining the proper time for initiation. Hence, despite the difficulty of discovering the individualized superior therapies in second-line treatment, another primary challenge is to determine the optimal time to initiate second-line therapy, either to receive treatment immediately after completion of platinum-based therapy, or to delay to another time prior to disease progression, whichever results in the largest overall survival probability. The goal is to provide patients with non-cross-resistant therapies capable of obtaining better response rates and longer survival time.

Some patients who maintain a good performance status and tolerate therapy without significant toxicities will receive third-line therapy (Stinchcombe and Socinski, 2008). Since there is only one FDA approved agent (Erlotinib) available for third-line treatment, we restrict our attention to finding optimal therapies for first-line and second-line only.

Figure 14 illustrates the treatment plan and clinically relevant patient outcomes. Therapy begins with first-line platinum-based doublets aimed at improving survival and palliating disease-related symptoms without undue toxicity. The patient is then delivered to no more than 8 cycles of treatment as recommended by the American Society of Clinical Oncology (ASCO). Socinski and Stinchcombe (2007) suggest the standard initial duration of platinum-based therapy should be 3 to 4 cycles since four of the five trials investigating the duration of platinum therapy in the first-line setting have revealed equivalent survival with the shorter duration of therapy. Due to the effects of the initial treatment, generally patients experience disease progression within a median of 3-6

months, and the median survival time observed is 8 to 10 months (Schiller et al., 2002; Sandler et al., 2006). Approximately 30–40% of patients survive 1 year, and less than 15% survive 2 years (Bunn and Kelly, 1998). If the first line of treatment is successfully completed without progression or death, then a second line of therapy is administered sometime between the completion of first-line treatment and the time of first evidence of disease progression. Patients with a good performance status in second-line trials have a median survival duration of approximately 9 months (Stinchcombe and Socinski, 2008). Given the noncurative nature of chemotherapy in advanced NSCLC, the overall survival time is defined as the primary endpoint.

The primary scientific goal of the trial is to select optimal compounds for first and second-line treatments as well as the optimal time to initiate second-line therapy based on prognostic factors yielding the longest averaged survival time. We create such new trial based on a reinforcement learning method, called  $Q$ -learning, for maximizing the averaged survival time of subgroup patients as a function of prognostic factors, treatment decisions, and optimal timing. We take the reinforcement learning approach because decisions must be made adaptively to various individuals during the trial, and this problem is especially acute in multi-stage treatment. In Chapter 4 we introduced  $Q$ -learning to cancer clinical communities and created a clinical reinforcement trial for discovering effective therapeutic regimens. By taking into account a drug’s efficacy and toxicity simultaneously, we demonstrated that reinforcement learning methodology not only captures the optimal individualized therapies successfully, but also is able to improving longer-term outcomes by considering delayed effects of treatment. While the trial proposed in Chapter 4 used to identify the optimal treatment shares similarity to some cancers, the structure (referred to optimal timing) is very different from NSCLC, so significant refinement for different optimal strategies identification is needed. Another challenge may also arise due to the right censoring phenomena in realistic trials, and we will address this issue in our new clinical reinforcement trial for NSCLC. In addition, in Chapter 4

we just utilized simplistic integer numbers as the reward function in  $Q$ -learning to trade off efficacy against toxicity. Thus it is important to consider some more plausible utility functions such as progression-free survival, overall survival, or quality of life to reflect the primary endpoint directly. In general cases, based on different reward functions chosen by clinicians, optimal treatment strategies found by a clinical reinforcement trial could be possibly more than one. In our NSCLC trial, we focus our attention on overall survival and treat it as the reward function, since this is arguably the most crucial clinical outcome, although quality of life is also important.

The design has two main components: a clinical reinforcement trial for fair randomization of patients among the different therapies in first and second-line treatments, as well as time of initiating second-line therapy, and a confirmatory phase III trial for finding and validating the optimal individualized therapies. Each new patient in the confirmatory trial is more likely to be assigned at appropriate treatments and timing having longest overall survival time, based on the performance of estimated optimal policies which are learned from the clinical reinforcement trial. In order to successfully handle the complex fact of heterogeneity in treatment across individuals and the possibility of right censored individuals in an NSCLC trial, we incorporate a modified SVR called  $\epsilon$ -SVR-C within a  $Q$ -learning framework to fit  $Q$ -functions for each decision point.

The remainder of this chapter is organized as follows. In Section 5.2 – 5.3, we provide a detailed description of the patient outcomes and refined  $Q$ -learning framework, followed by an introduction to  $\epsilon$ -SVR-C for estimating  $Q$ -functions with censored observations. The NSCLC trial conduct and related computational issues are presented in Section 5.4. In Section 5.5, we present a simulation study of the design to discover individualized optimal treatment strategies. We close with a summary in Section 5.6.



## 5.2 Reinforcement Learning Model Refinement

### 5.2.1 Patient Outcomes

Let  $t_1$  and  $t_2$  denote the decision time of first and second line treatment, respectively. Given first-line chemotherapy, the indicator of the time to disease progression is denoted by  $T_P$ .  $t_2$  is also the time at the completion of first-line treatment, and is a fixed value usually less than  $T_P$  and determined by the number of cycles delivered in the first line of chemotherapy. Denote the time of initiating second-line therapy by  $T_M$ . Thus, according to the description of treatment plan in Section 5.1,  $T_M \in [t_2, T_P]$ . At the end of first-line therapy,  $t_2$ , clinicians make a decision when to start  $T_M$ . We let  $T_D$  denote the time of death from the start of therapy, i.e., the overall survival time. For patients who have died before  $t_2$ , let  $T_1$  denote the time from  $t_1$  to patient's death. In this case,  $T_D = T_1$ . Similarly, for patients who live beyond  $t_2$ , let  $T_2$  denote the time from  $t_2$  to patient's death. Thus, for this kind of patient,  $T_D = t_2 + T_2$ . Note that  $T_D$  may less than  $T_M$ .

Because of the possibility of right censoring, we define the patient's censored time by  $C$  and the indicator of censoring by  $\delta$ . Right censoring may be due to several reasons, including an adverse event so severe that therapy cannot be continued or the patient chooses not to receive further therapy. We assume for simplicity that censoring is independent of death in this thesis. For convenience, we let  $Y_D = I(T_D \wedge C > t_2)$  and  $\nu = Pr(Y_D = 1)$ , so  $T_2$  is defined only if  $Y_D = 1$  and  $\delta = 0$ . Denoting the last follow-up time by  $T^0$ , we then can define  $T^0 = T_D \wedge C \wedge t_2 + I(T_D \wedge C > t_2)(T_2 \wedge (C - t_2))$ . The settings for determining  $T_1$ ,  $C$ ,  $T_M$ , and  $T_2$  are summarized in Figure 15, including the possibilities of death or right censoring either before or after second-line therapy.

Denote patient covariate values at the  $i$ th line by  $\mathbf{O}_i = (O_{i1}, \dots, O_{iq})$  for  $i = 1, 2$ . Such covariates can include prognostic variables or biomarkers thought to be related to outcome. In first-line therapy, we assume that the death time  $T_1$  depends on the

covariates  $\mathbf{O}_1$  and possible treatment  $D_1$  according to a possible function

$$[T_1 \mid \mathbf{O}_1, D_1] \sim f_1(\mathbf{O}_1, D_1; \alpha_1),$$

where decision  $D_1$  only consists of a finite set of agents  $d_1$ . If the patient survives long enough to be treated by second-line therapy, we assume that the disease progression time  $T_P$  follows another distribution

$$[T_P \mid \mathbf{O}_1, D_1] \sim f_2(\mathbf{O}_1, D_1; \alpha_2).$$

In addition, to account for the effects of initial timing of second-line therapy on survival,  $T_2$  is given by

$$[T_2 \mid \mathbf{O}_2, D_1, D_2, T_M] \sim f_3(\mathbf{O}_2, D_1, D_2, T_M; \alpha_3),$$

where  $D_2$  consists of a finite set of agents  $d_2$  and  $T_M$  is a continuum of initiation times for second-line therapy as described above. Therefore, this study is designed to identify the the initiation time,  $T_M$ , that is associated with the best combination of treatments  $d_1$  and  $d_2$ , while maintaining longest survival  $T_D$ . Due to heterogeneities among patients, biomarker-treatment interactions, and the large number of possible shapes of  $T_2$  as functions of  $T_M$ , functions  $f_1$ ,  $f_2$ , and  $f_3$  can be linear or non-linear and may vary between different groups of patient. Thus, incorporating  $\mathbf{O}_i$  into model  $f_i$  ( $i = 1, 2, 3$ ) is quite challenging, and such model-based approaches can easily become intractable (Thall et al., 2007). Another important issue is accounting for delayed effects of first-line therapy. It is possible that the treatment having a short disease progression time  $T_P$ , by administering first-line therapy, is a good strategy for two-stage treatment protocols in terms of overall survival time. Thall et al. (2007) claimed that the conventional model-based approaches are not capable of handling this situation very well. Based on clinical data, reinforcement learning is not only a model-free method which carries out treatment selection sequentially with time-dependent outcomes to determine optimal individualized therapy, but it can also improve longer-term outcomes by taking into account delayed effects of treatments.

### 5.2.2 Q-Learning Revisited

As mentioned in Section 3.2.2, the  $Q$ -learning (Watkins, 1989; Watkins and Dayan, 1992) is one of the most widely used reinforcement learning methods. The core of the algorithm is a simple value iteration update. It assumes the old value and makes a correction based on the new information as follows (Sutton and Barto, 1998):

$$Q_t(s_t, a_t) \leftarrow Q_t(s_t, a_t) + \alpha_t(s_t, a_t) \times \left[ r_t + \gamma \max_{a_{t+1}} Q_{t+1}(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t) \right], \quad (5.1)$$

where  $r_t$  is the current reward given at time  $t$ ,  $\alpha_t(s_t, a_t) \in (0, 1]$  the learning rate (or learning step-size). We let  $\gamma = 1$  to fully maximize rewards over the long run. For simplicity of computation, we ignore the step-size (let  $\alpha_t(s_t, a_t) = 1$ ) for the rest of the article. All results hold with minor modifications when the step-size effects are considered. Then model (5.1) can be simplified to one-step simple recursive form

$$Q_t(s_t, a_t) \leftarrow r_t + \max_{a_{t+1}} Q_{t+1}(s_{t+1}, a_{t+1}). \quad (5.2)$$

The  $Q$ -learning algorithm attempts to find a policy  $\pi$  that maps states to actions the learner ought to take in those states.  $\pi$  is possibly deterministic, non-stationary, and non-Markovian. We denote the optimal policy by  $\pi_t^*$ , which satisfies

$$\pi_t^* = \arg \max_{a_t} Q_t(s_t, a_t).$$

In Chapter 4 we performed a simulation study of a simple  $Q$ -learning approach with 6 decision time points for discovering optimal dosing for treatment of a generic cancer. While the results were encouraging, there remains much work to do before these methods can be applied to realistic cancer scenarios. For example, in Chapter 4, the choice of treatments at each decision time point is taken from a continuum of dosing levels. However, in NSCLC treatment with two decision time points, the action variables in the second stage become two-dimensional ( $d_2$  and  $T_M$ ). Due to this significantly different structure, a new methodology and model are needed. Moreover, the presence of cen-

soring in the reward outcome means that a fundamentally new approach is required for estimating the  $Q$  functions.

In our clinical setting we respectively denote state and action random variables by  $O_i$  and  $D_i$  for  $i = 1, 2$ . This is consistent with notations of prognostic factors or biomarkers and treatment options used in Section 5.2.1. As mentioned in Section 5.1, we consider survival time as the primary reward function. Specifically, by performing a treatment  $d_1$ , where  $d_1 \in D_1$ , the patient can transit from first line to second line. Such treatment associated with prognostic factors provides the patient a progression time  $T_P$  and  $T_1$  ( $T_1$  is defined only if  $Y_D = 0$ ).  $T_P$  is only used for determining  $T_M$ . Moreover,  $D_2$ , which consists of two dimensional action variables including a possible discrete action (agent)  $d_2$  mixed with a continuous action (time)  $T_M$ , provides the patient a survival time  $T_2$ . While taking into account possible right censoring in the first stage, such a reward function can be formally defined as  $T_1 \wedge C$ , plus the corresponding censoring indicator, if  $Y_D = 0$  or  $t_2$  if  $Y_D = 1$ , where  $T_1$  satisfies

$$T_1 \sim R_1(o_1, d_1).$$

In the second stage, the reward function is defined by  $T_2 \wedge (C - t_2)$ , where  $T_2$  satisfies

$$T_2 \sim R_2(o_2, d_1, d_2, T_M).$$

Functions  $R_1$  and  $R_2$  coincide with  $f_1$  and  $f_3$  and are not observable in realistic trials. In  $Q$ -learning, because for every state there are a number of possible treatments that could be taken, each treatment within each state has a value according to how long the patient will survive due to completion of that treatment. The scientific goal of our study is to find an optimal policy to maximize patients' overall survival time  $T_D$ . This is accomplished by learning which treatment (including starting time for second-line therapy) is optimal for each state.

While learning a non-stationary non-Markovian optimal policy with one set of finite

horizon trajectories (also called a training data set)

$$\{O_1, D_1, T_D \wedge C \wedge t_2, O_2, D_2, T_2 \wedge (C - t_2), \delta\},$$

we denote the estimation of the optimal  $Q$ -functions based on this training data by  $\widehat{Q}_t$ , where  $t = 1, 2, 3$ . According to the recursive form of  $Q$ -learning in (5.2), we must estimate  $Q_t$  backwards through time, that is, use the estimate  $Q_3$  from the last time point back to  $Q_1$  at the beginning of the trial. For convenience we set  $Q_3$  equal to 0. In order to estimate each  $Q_t$ , we denote  $Q_t(O_t, D_t; \boldsymbol{\theta}_t)$  as a function of a set of parameters  $\boldsymbol{\theta}_t$ , and we allow the estimator to have different parameter sets for different time points  $t$ . Once this backwards estimation process is done, we save the sequence of  $\widehat{Q}_1$  and  $\widehat{Q}_2$ , and we thereafter use them to respectively estimate optimal treatment policies

$$\widehat{\pi}_1 = \arg \max_{d_1} \widehat{Q}_1(o_1, d_1; \boldsymbol{\theta}_1)$$

and

$$\widehat{\pi}_2 = \arg \max_{d_2, T_M} \widehat{Q}_2(o_2, d_2, T_M; \boldsymbol{\theta}_2),$$

for new patients in a testing dataset. These estimated optimal policies should also be evaluated in a follow-up confirmatory phase III trial comparing the optimal policy or policies with the standard of care.

### 5.3 Support Vector Regression for Censored Subjects

A strength with  $Q$ -learning is that it is able to compare the expected survival of the available treatments without requiring a model of the relationship. To achieve this, the main task is to estimate the  $Q$  functions for finding the corresponding optimal policy. However, challenges may arise due to the complexity of the structure of the true  $Q$  function, specifically, the non-smooth maximization operator in recursive equations (5.2).

In the previous chapter we applied SVR as our main method to fit  $Q$  functions and learn optimal policies using a training data set. Instead of the hinge loss function used

in SVM, one of the popular loss functions involved in SVR is known as the  $\epsilon$ -insensitive loss function (Vapnik, 1995), which is defined as

$$L(f(\mathbf{x}_i), y_i) = (|f(\mathbf{x}_i) - y_i| - \epsilon)_+, \quad (5.3)$$

where  $\epsilon > 0$ .

Note that we have in the prior chapter assumed that all patients are followed up until they die. In conducting an NSCLC trial, a common problem is the right censoring caused by patients who do not complete the study and drop out of the study without further measurements. Possible reasons for patients dropping out of the study include, adverse reactions, lack of improvement, unpleasant study procedures, and other factors related or unrelated to the trial procedure and treatments. For simplicity, we assume that right censoring is independent of death.

In general, we denote interval censored data by  $(\mathbf{x}_i, l_i, u_i)_{i=1}^n$ . If the patient experiences the death event and  $T_D$  is observed rather than being interval censored then we include  $T_D$  and denote such observation as  $(\mathbf{x}_i, y_i)$ . In other words, when we observe  $T_D$  exactly ( $\delta = 0$ ), we let  $l_i = u_i = y_i$ . Note that by letting  $u_i = +\infty$  we can easily obtain a right censored observation  $(\mathbf{x}_i, l_i, +\infty)$ .

One naive way to handle censored data within  $Q$ -learning by using SVR is to consider only those samples for which the survival time  $T_D$  are known exactly. Such an approach which totally ignores censored data will reduce the sample size for statistical analysis and inference. Thus the more patients that are censored, or the earlier they are censored, the more unreliable the results will be. An SVR procedure that targets interval censored subjects was introduced by Shivaswamy, Chu, and Jansche (2007). The key component of their procedure is a loss function, defined as

$$L(f(\mathbf{x}_i), l_i, u_i) = \max(l_i - f(\mathbf{x}_i), f(\mathbf{x}_i) - u_i)_+.$$

However, this loss function dose not have  $\epsilon$ -insensitive properties, that is, it does not allow  $\epsilon$  or other deviations from the predicted  $f(\mathbf{x}_i)$ , especially when  $l_i = u_i = y_i$ . In

this article, we propose a modified SVR algorithm with  $\epsilon$ -insensitive loss function (called  $\epsilon$ -SVR-C) to make use of both survival time  $T_D$  and censoring time  $C$  in the data set and to reduce the potential bias which may be caused by performing a classical SVR with censored data.

Given the interval censored data set  $(\mathbf{x}_i, l_i, u_i)_{i=1}^n$ , our modified loss function is defined as

$$L(f(\mathbf{x}_i), l_i, u_i) = \max(l_i - \epsilon - f(\mathbf{x}_i), f(\mathbf{x}_i) - u_i - \epsilon)_+. \quad (5.4)$$

The main difference between (5.3) and (5.4) is that  $y_i$  is separated into two parts which are replaced by  $l_i$  and  $u_i$ , respectively. We remark that this loss function does not penalize values of  $f(\mathbf{x}_i)$  if it is between  $l_i - \epsilon$  and  $u_i + \epsilon$ . On the other hand, the cost grows linearly if this output is more than  $u_i + \epsilon$  or less than  $l_i - \epsilon$ . Figure 16 shows the loss function of the modified SVR. Note that when  $u_i = +\infty$ , this loss function becomes one sided, which means there is no empirical error if  $f(\mathbf{x}_i) \geq l_i - \epsilon$ . In addition, when the data is not observed as censored, our modified SVR algorithm reduces to the classical SVR.

The parameter  $\epsilon$  can be useful if the desired accuracy of the approximation can be specified beforehand. Note that when  $\epsilon = 0$ , our approach reduces to the method proposed by Shivaswamy et al. (2007). Based on some small simulation studies, the performance of our method is not very sensitive to the choice of  $\epsilon$  (say, from 0 to 0.1). This means, in our study,  $\epsilon$ -SVR-C's performance is very close to Shivaswamy et al. (2007)'s method. However, since  $\epsilon$ -insensitive tube is useful to control the proportion of support vectors involved in approximation, we prefer  $\epsilon$ -SVR-C throughout this thesis.

Denoting index sets  $L = \{i : l_i > -\infty\}$  and  $U = \{i : u_i < +\infty\}$ , the corresponding modified SVR optimization formulation is:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi, \xi'} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \left( \sum_{i \in L} \xi_i + \sum_{i \in U} \xi'_i \right), \\ \text{subject to} \quad & (\mathbf{w}^T \Phi(\mathbf{x}_i) + b) - u_i \leq \epsilon + \xi_i, \quad i \in U, \\ & l_i - (\mathbf{w}^T \Phi(\mathbf{x}_i) + b) \leq \epsilon + \xi'_i, \quad i \in L, \end{aligned}$$

$$\xi_i \geq 0, i \in L; \quad \xi'_i \geq 0, i \in U.$$

Similarly to classical SVR, the dual can be presented as follows by introducing Lagrange multiplier  $\lambda_i$ :

$$\begin{aligned} \min_{\boldsymbol{\lambda}, \boldsymbol{\lambda}'} \quad & \frac{1}{2}(\boldsymbol{\lambda} - \boldsymbol{\lambda}')^T K(\mathbf{x}_i, \mathbf{x}_j)(\boldsymbol{\lambda} - \boldsymbol{\lambda}') - \sum_{i \in L} (l_i - \epsilon)\lambda'_i + \sum_{i \in U} (u_i + \epsilon)\lambda_i, \\ \text{subject to} \quad & \sum_{i \in L} \lambda'_i - \sum_{i \in U} \lambda_i = 0, \\ & 0 \leq \lambda_i, \lambda'_i \leq C, \quad i = 1, \dots, n. \end{aligned}$$

Once the above formulation is solved to get the optimal  $\lambda_i$  and  $\lambda'_i$ , the approximate function at  $\mathbf{x}$  is given by:

$$f(\mathbf{x}) = \sum_{i=1}^n (\lambda'_i - \lambda_i) K(\mathbf{x}_i, \mathbf{x}) + b.$$

Based on results for non-censored  $Q$ -learning with classical SVR, it is expected that the  $\epsilon$ -SVR-C behaves similarly, with the estimated policies  $\hat{\pi}$  being more robust to censored data and being more optimal than results where the censored patients are simply ignored. To verify this comparison, a small simulation study will be reported in Section 5.5.

## 5.4 Clinical Reinforcement Trial Conduct and Computational Strategy

Different populations of patients with NSCLC appear to have different clinical and molecular characteristics, so clinical trials that investigate the activity of different agents, and incorporate patient selection based on clinical factors, are required. The goal of this clinical reinforcement trial is to compare two-line treatment strategies for patients with NSCLC who have not been treated previously with systemic therapy. As mentioned in Section 5.1, while many new single agents with potential clinical efficacy currently are being produced at an increasing rate, the number of doublet combinations in the first



line that can be evaluated clinically is limited. Considering the number of possible agents that may be of interest in the second line, the limitations are far greater.

Without loss of generality, suppose for simplicity that strategies are based on four FDA approved therapies (either single agents or doublets), which we denote by  $A_i, i = 1, \dots, 4$ . In our study we assume that the second line treatment must be different from the first. When designing the trial, two of the four agents  $A_1$  and  $A_2$  are selected for first-line treatment, while  $A_3$  and  $A_4$  are selected for second line. A total of  $n$  patients are recruited into the trial and fairly randomized at enrollment between  $A_1$  and  $A_2$ , and each patient is followed through to completion of first-line treatment, given the patient is not dead or lost to follow-up from the study. We fix this duration  $t_2 - t_1$  as 2.8 months, although other lengths are possible, depending on the number of cycles of treatment. At the end of first-line treatment, patients are randomized again between agents  $A_3$  and  $A_4$ . Moreover, another important clinical decision that needs to make at this point is when to initiate the second-line treatment. Thus, the initiation for second-line treatment could be randomized to as early as  $t_2$  or as late as  $T_P$  (recall that  $T_P$  denotes the time of patient’s disease progression). At the end of the trial, the patient data is collected and  $Q$ -learning is applied, in combination with SVR applied at each time point, to estimate the optimal treatment rule as a function of patient variables and biomarkers, at  $t_1$  and  $t_2$ .

The trial described above was motivated by the desire to compare several agents as well as timing in a randomized fashion, the belief that different agents combined with different timing given consecutively may have interactive effects for separate population of patients, and the desire to determine a sound basis for selecting individualized optimal strategies for evaluation in a future clinical trial. Computationally, the entire algorithm for  $Q$ -function estimation and optimal treatment discovery is summarized as follows:

1. Inputs: If  $t = 1$ , a set of training data consists of attributes  $\mathbf{x}_i$  (states  $o_1$ , actions  $d_1$ ) and index  $y_i$  (rewards  $T_1 \wedge C$ ), i.e.,  $\{(o_1, d_1, T_1 \wedge C, \delta)_i, i = 1, \dots, n\}$ ; if  $t = 2$ ,

a set of training data  $\{(o_2, d_2, T_M, T_2 \wedge (C - t_2), \delta)_j, j = 1, \dots, n'\}$ , where  $n' \leq n$  since patients may die or be censored before second-line therapy.

2. Initialization: Let  $\widehat{Q}_3$  be a function equal to zero.
3.  $Q_2$  is fitted with  $\epsilon$ -SVR-C through the following equation:

$$Q_2(o_2, d_2, T_M) = T_2 \wedge (C - t_2).$$

4.  $Q_1$  is fitted with  $\epsilon$ -SVR-C through the following equation:

$$Q_1(o_1, d_1) = T_1 \wedge C \wedge t_2 + Y_D \times \max_{d_2, T_M} Q_2(o_2, d_2, T_M).$$

5. For the SVR computations in steps 3 and 4, if a Gaussian kernel is applied, we use a straightforward coarse grid search with  $C = 2^{-5}, 2^{-3}, \dots, 2^{15}$  and  $\zeta = 2^{-15}, 2^{-13}, \dots, 2^3$ , evaluated at each candidate pair  $(C, \zeta)$ , and then select the one that yields the highest cross-validation rate.
6. Given  $\widehat{Q}_1$  and  $\widehat{Q}_2$ , the individualized optimal policies  $\widehat{\pi}_1$  and  $\widehat{\pi}_2$  for application to future patients are computed.
7. Evaluate  $\widehat{\pi}_1$  and  $\widehat{\pi}_2$  in a confirmatory phase III trial to compare the optimal policies with the standard of care.

## 5.5 Simulation Study

To demonstrate that the tailoring therapy for NSCLC found by using the proposed clinical reinforcement trial is superior, we employ an extensive simulation study to assess the proposed approach on virtual clinical trials of patients, and then evaluate using Phase III trial-like comparisons between the estimated optimal regimen and the various possible fixed treatments.

### 5.5.1 Data Generating Models

Based on historical research, it is well known that the rate of disease progression or death for patients with advanced NSCLC increases over time. Consequently, in order to generate simulation data, we simply consider that  $T_P$ ,  $T_1$ , and  $T_2$  follow different exponential distributions. Many alternative models are also possible.

Let  $\exp(x)$  denote an exponential distribution with mean  $e^x$ . For a patient given first-line treatment  $d_1$ , we assume death time distribution

$$[T_1 | D_1] \sim \exp(\alpha_{D_1} + \beta_{D_1}W_1 + \kappa_{D_1}M_1 + \tau_{D_1}W_1M_1). \quad (5.5)$$

If  $T_1 > t_2$ , we assume disease progression time distribution

$$[T_P | D_1] \sim \exp(\alpha_{D_1}^P + \beta_{D_1}^P W_1 + \kappa_{D_1}^P M_1 + \tau_{D_1}^P W_1 M_1). \quad (5.6)$$

Given  $t_2$ ,  $T_M$  is uniformly generated from  $[t_2, t_2 + 4]$  (4 months interval). If  $T_P \leq T_M$ , then let  $T_M = T_P$ . In addition, for a patient given second-line treatment  $d_2$  and initiation time  $T_M$ , we assume the death time

$$[T_2 | D_1, D_2] \sim \exp(\alpha_{D_{12}} + \beta_{D_{12}}W_2 + \kappa_{D_{12}}M_2 + h(T_M; \boldsymbol{\varphi})). \quad (5.7)$$

Given  $T_D = T_1$  or  $t_2 + T_2$  and patient censored probability  $p^c$ , we generate right censored time  $C$  uniformly from interval  $[t_1, t_1 + 24]$  (2 years interval), where the censoring indicator is drawn according to a Bernoulli distribution  $B(p^c)$ . Note that in our simulation study we straightforwardly use exponential pdfs (5.5)–(5.7) to replace  $(f_1, f_2, f_3)$ , which are mentioned in the notation of Section 5.2.1. For the sake of simplicity, in these density functions only two state variables such as quality of life (QOL) and tumor size are considered as patient prognostic factors or biomarkers to be related to outcome, and they are denoted by  $W_t$  and  $M_t$  ( $t = 1, 2$ ), respectively. We consider these two factors because they are patient based, easy to be measured, can predict therapeutic benefit after treatment of chemotherapy, and more importantly, they are significant prognostic factors

for survival (Socinski et al., 2007). In addition, state variables for the next decision are generated by simple dynamic models  $W_2 = W_1 + T_M \dot{W}_1$  and  $M_2 = M_1 + T_M \dot{M}_1$ .

Recall that  $\nu$  is the probability for the event that the patient can live beyond  $t_2$ . The parameter vector for patients who only experience first-line treatment is

$$\boldsymbol{\theta}_1 = (\alpha_{D_1}, \beta_{D_1}, \kappa_{D_1}, \tau_{D_1}),$$

otherwise, it is

$$\boldsymbol{\theta}_2 = (\alpha_{D_1}^P, \beta_{D_1}^P, \kappa_{D_1}^P, \tau_{D_1}^P, \nu, \alpha_{D_{12}}, \beta_{D_{12}}, \kappa_{D_{12}}, \boldsymbol{\varphi}).$$

Parameter vectors  $\boldsymbol{\theta}_1$  and  $\boldsymbol{\theta}_2$  as well as the shape of time-related function  $h(T_M; \boldsymbol{\varphi})$  vary among different patients. Note that two patients who receive different decisions with the same first-line treatment, say  $(A_1, A_3)$  and  $(A_1, A_4)$ , both contribute data for learning in stage 1.

### 5.5.2 Clinical Scenarios

To construct a set of scenarios reflecting the interaction between two lines of treatment, we temporarily assume that a large portion of patients survive long enough to be treated by second-line therapy, that is, we specified  $\nu = 0.8$  for all patients. Except for  $\nu$ , each clinical scenario under which we will evaluate the design in the simulation study is built by a unique set of fixed values of  $(\alpha_{D_1}^P, \beta_{D_1}^P, \kappa_{D_1}^P, \tau_{D_1}^P, \alpha_{D_{12}}, \beta_{D_{12}}, \kappa_{D_{12}})$ . The remaining fixed parameter values needed for the simulations are those that determine how  $T_2$  varies as a function of  $T_M$ . To implement this, we specified four corresponding model-based cases of each function  $h(T_M; \boldsymbol{\varphi})$  in terms of their numerical values at each  $T_M$ . All of these underscore the importance of specifying the optimal regimen to target a subpopulation of patients with distinct characteristics.

Hence, to facilitate interpretation of reinforcement learning strategies for capturing individualized therapies, four scenarios are specified and summarized in Table 2. In group

1 and 4, initial timing of second-line therapy for survival time ( $T_2$ ) are functions that form an inverse-U (quadratic) shape with  $T_M$ , while initial timing in group 2 and 3 for  $T_2$  are functions that linearly decrease and increase with  $T_M$ , respectively. Each group thus consists of a combination  $(A_i, A_j)$  as well as  $T_M$  timing from Table 2 (where  $i = 1, 2$  and  $j = 3, 4$ ), with the fixed values of  $\alpha_{D_1}^P, \beta_{D_1}^P, \kappa_{D_1}^P, \tau_{D_1}^P, \alpha_{D_{12}}, \beta_{D_{12}}, \kappa_{D_{12}}$ , and  $\varphi$  as described above.

Note that whatever combination of two-line treatment  $(A_i, A_j)$  is evaluated, all patients within one group share the same trend of  $T_2$  versus  $T_M$ . However, we assume there is only one strategy that will yield the longest survival time in each group. For convenience, we denote “1, 2, 3” as the location of optimal initiation of second-line therapy, defined as “immediate, intermediate, delayed”, respectively. For example, as claimed in the last column in Table 2,  $A_1A_32$  indicates that the two-line treatments  $(A_1, A_3)$  along with an intermediate initiation time point is the optimal regimen for group 1. The inverse-U-shaped function  $T_2$  for  $T_M$  corresponds to the case where patients have relatively low QOL at enrollment but relatively large tumor size, hence, this optimal intermediate initiation of second-line therapy is recommended to delay treatment in a short time for patients who may have severe symptoms and low tolerance of chemotherapy, but not to be fully delayed due to the possibility of death. In scenario 2, due to the good QOL and large tumor size at enrollment, it is optimal for the second-line therapy to begin immediately after first-line therapy, hence,  $A_1A_41$  is the optimal regimen for these patients. Similarly, in scenario 3, treatment  $A_2A_33$  is considered the superior treatment since we believe fully that delaying the initiation of second-line therapy at the time of disease progression will improve survival and palliate symptoms. Although scenario 4 has optimal regimen  $A_2A_42$ , due to the flat shape of  $T_2$  versus  $T_M$ , there is no significant improvement between delaying and not delaying the initiation of second-line therapy.

### 5.5.3 Simulation Methods and Results

First, according to various  $(W_1, M_1)$  as described in Table 2, a clinical reinforcement trial of size  $N = 100$  is generated for each group (total  $n = 400$ ), and we repeat this simulation for 10 times.  $\widehat{Q}_1$  and  $\widehat{Q}_2$  are computed via the algorithm given in Section 5.3. Then predicted optimal strategies are computed via an independent phase III confirmatory trial of size 100 per group, generated from the same mechanism as its corresponding reinforcement trial. For comparison, we assign all test patients to  $(A_i, A_j) \times \{\text{immediate, intermediate, delayed}\}$ , which consists of 12 combinations in total. Patients' outcomes (overall survival) conducted by our estimated optimal regimens and 12 different fixed regimens are all evaluated. All of these results are averaged over 400 patients in each regimen in the confirmatory trial. As shown in Figure 17, among regular regimens, by assigning all testing patients to  $A_2A_33$  will yield the averaged longest survival as 14.71 months. It thus appears that, in terms of adaptively selecting best strategies for each group, the optimal regimen obtained by  $Q$ -learning with  $\epsilon$ -SVR-C is superior due to the averaged survival of 16.45 months (with standard deviation 0.063) over 10 trials. Because of this encouraging result, it is worthwhile to deeply investigate whether our approximations are close to the exact solution. To carefully examine this comparison, we assign patients of each group to the corresponding true optimal regimen described in Table 2. Since the reinforcement trial was simulated 10 times with a size of 400, the minimum, maximum, and mean values of averaged predicted survival for each group are computed based on these 10 trials, respectively. The results are summarized in Table 3. The averaged predicted survival over all groups is shown as 16.446, this number along with minimum 16.065 and maximum 16.624 are all pretty close to true optimal survival 16.554. In terms of estimation, under each of the scenarios 1–3 our methods perform very similarly and slightly underestimates the true optimal survival. In contrast, our method slightly overestimates the true optimal survival in scenario 4.

Second, although our  $Q$ -learning method with  $N = 100$  per group using  $\epsilon$ -SVR-C leads to an apparently small bias for estimating individualized optimal regimens, an examination of performance influenced by the sample size is worthwhile. We repeated the simulations 10 times for each specified sample size while varying  $N$  from 2 to 600 per group. The results are illustrated in Figure 18, which shows that the method’s reliability is very sensitive to  $N$  when  $N \leq 80$ , with the averaged survival for the estimated optimal strategy increasing from 14.017 when  $N = 2$  to 16.320 when  $N = 80$ . The boxplots also show that both the deviation and estimation bias of predicted survival are getting smaller when the sample size becomes larger. When  $N \geq 100$ , our methods appear to do a very reliable job of selecting the best strategy. Hence, in the setting we study here, the sample sizes required to reach excellent approximation are similar to and not larger than the sizes required for typical phase III trials.

Third, in order to compare performance of the  $\epsilon$ -SVR-C for censored subjects to non-censored and ignoring the censored cases, from 400 training samples over 10 simulations run, we randomly select a fraction of the samples (based on censored data simulation described in Section 5.5.1) so that they become right censored patients. We repeat the comparisons with reinforcement trial which has 25%, 50%, and 75% censored proportion, respectively. The boxplots are presented in Figure 19. Evidently, in terms of averaged predicted survival in all cases, the  $\epsilon$ -SVR-C algorithm outperforms the method which totally ignores censored data, particularly when the censored proportion is large.

## 5.6 Summary of NSCLC trial results

We have proposed an adaptive reinforcement learning design for conducting a clinical trial of multiple lines of treatment in a group of patients with advanced NSCLC. The incorporation of  $Q$ -learning with the proposed  $\epsilon$ -SVR-C appears to successfully identify optimal treatment strategies tailored to a proper subpopulation of patients. While our

method has been utilized for the two decision points at hand, the general principals and algorithms of this approach could be applied, with suitable modification, to design future trials having similar goals but for possibly different diseases.

We provided an explicit simulation to evaluate the performance of  $\epsilon$ -SVR-C in reinforcement trials. Our analysis and simulation conclude that the  $Q$ -learning procedure with  $\epsilon$ -SVR-C can handle censored data and simultaneously maintain good estimation, and is a practical choice for reinforcement learning designs with higher levels of censoring. When there are no censored subjects, the  $\epsilon$ -SVR-C reduces to classical SVR. When there are large portion of censored subjects, the  $\epsilon$ -SVR-C is much more robust and effective than the naive method which just simply ignores the censored data. More simulations and theoretical studies of  $\epsilon$ -SVR-C are needed in the future.



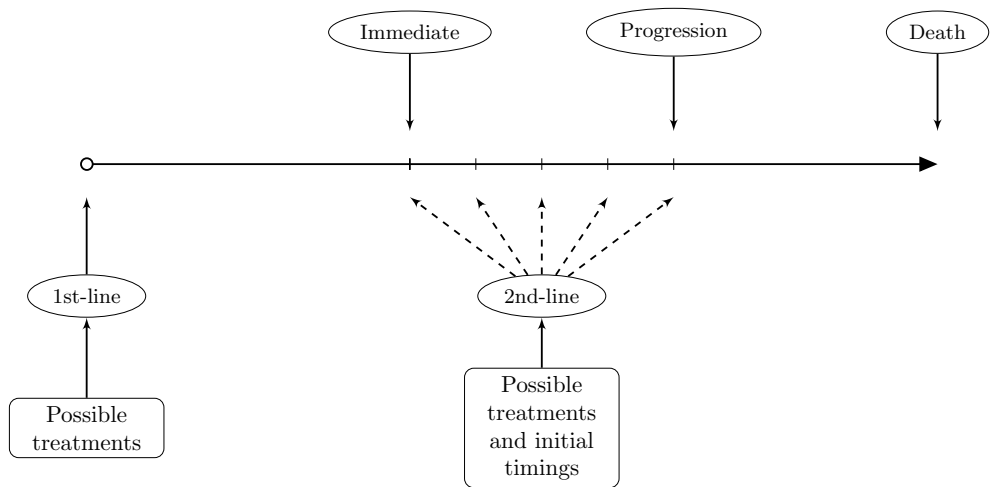


Figure 14: Treatment plan and therapy options for an advanced NSCLC trial.

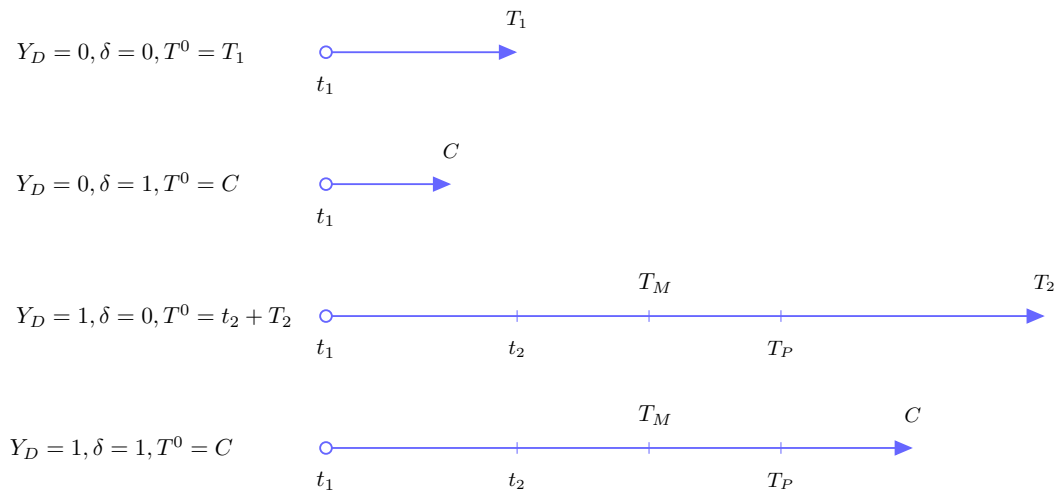
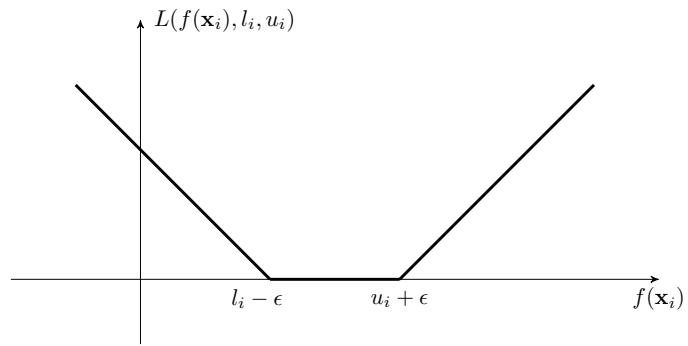
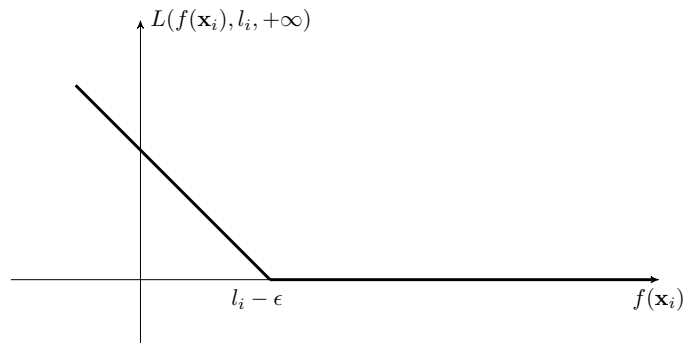


Figure 15: The four cases that determine the times  $T_1$ ,  $C$ ,  $T_M$ , and  $T_2$ . In each case, the time of last follow-up is indicated by a right triangle. Note that all times originate at  $t_1$  except  $T_2$  which originates at  $t_2$ .




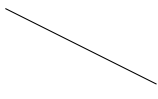
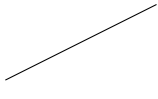
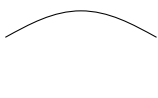
(a)



(b)

Figure 16: Loss functions of  $\epsilon$ -SVR-C for interval censored data (a) and right censored data (b).

Table 2: The scenarios studied in the simulation. Sample size = 100/group.

Group	State Variables Status		Timing	Optimal Regimen
1	$W_1 \sim N(0.25, \sigma^2)$ $M_1 \sim N(0.75, \sigma^2)$	$W_1 \downarrow M_1 \uparrow$		$A_1A_32$
2	$W_1 \sim N(0.75, \sigma^2)$ $M_1 \sim N(0.75, \sigma^2)$	$W_1 \uparrow M_1 \uparrow$		$A_1A_41$
3	$W_1 \sim N(0.25, \sigma^2)$ $M_1 \sim N(0.25, \sigma^2)$	$W_1 \downarrow M_1 \downarrow$		$A_2A_33$
4	$W_1 \sim N(0.75, \sigma^2)$ $M_1 \sim N(0.25, \sigma^2)$	$W_1 \uparrow M_1 \downarrow$		$A_2A_42$

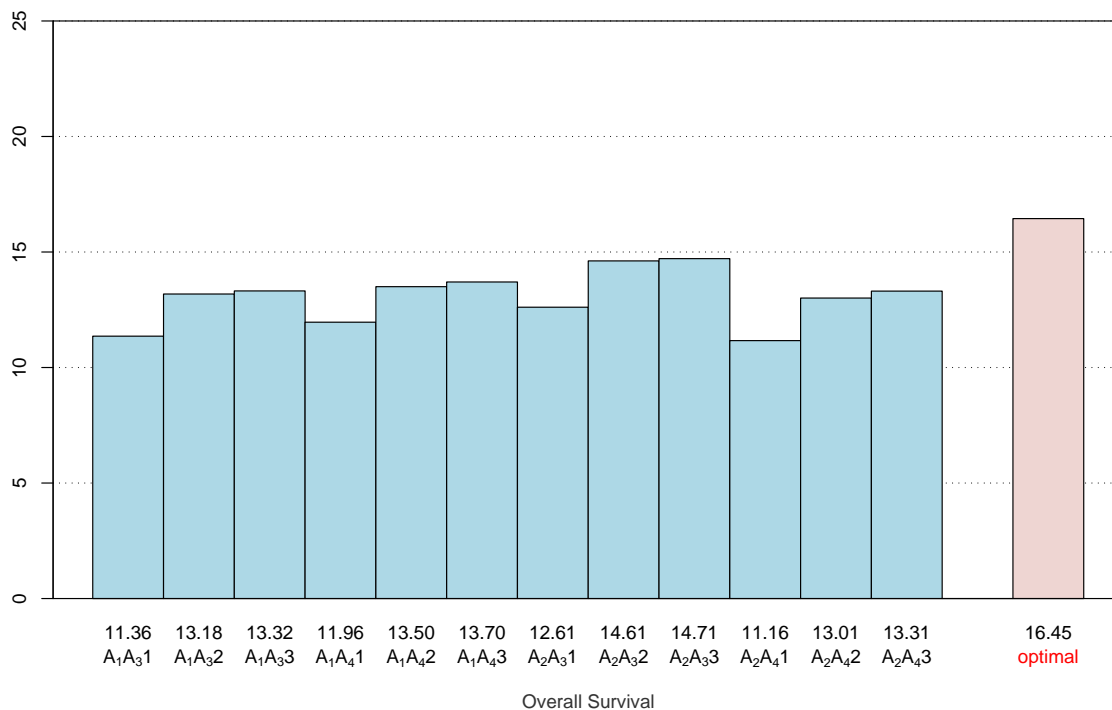


Figure 17: Performance of optimal individualized regimens versus other 12 combinations. The same confirmatory phase III trial was used in this comparison. The 12 bars in the left indicate results of 12 different fixed regimens, while the bar in the right indicate results of optimal regimens. The optimal regimens will yield averaged survival time of 16.45 months with standard deviation 0.063 (over 10 trials), which is longer than all fix regimens.

Table 3: Comparisons between true optimal regimens and estimated optimal regimens for overall survival (month). Each training dataset is of size 100/group with 10 simulation runs. Testing dataset is of size 100/group.

Group	Optimal regimen	True survival	Predicted survival		
			Min	Mean	Max
1	$A_1A_32$	14.773	14.072	14.593	14.769
2	$A_1A_41$	15.343	14.941	15.197	15.341
3	$A_2A_33$	17.614	17.060	17.417	17.576
4	$A_2A_42$	18.487	18.188	18.578	18.810
Average		16.554	16.065	16.446	16.624

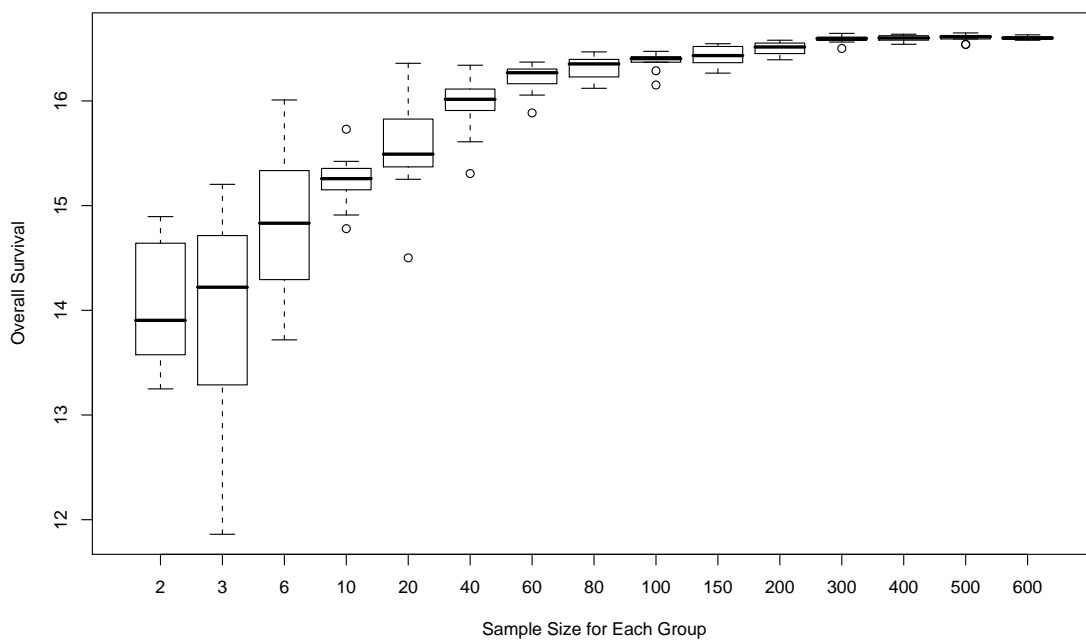


Figure 18: Sensitivity of the predicted survival to the sample size.

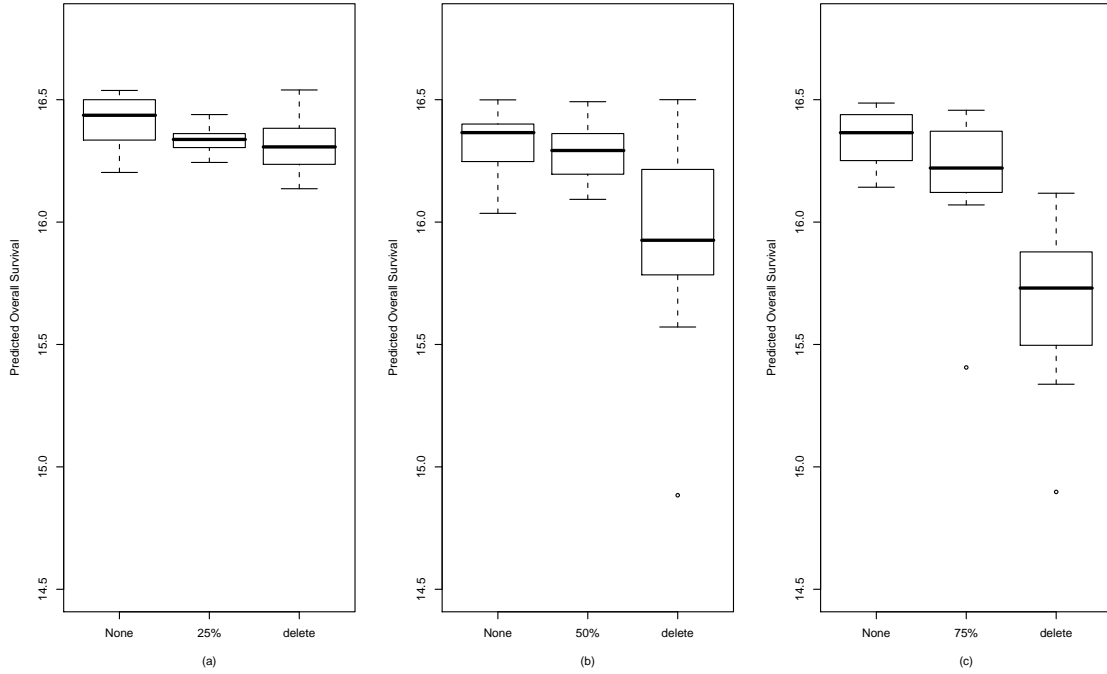


Figure 19: Boxplots of the predicted survival computed via  $Q$ -learning with  $\epsilon$ -SVR-C based on a reinforcement trial with 25% (a), 50% (b), and 75% (c) fraction of right censored subjects. In each case ((a),(b), or (c)), based on the same trial, 3 boxplots indicate performance of non-censored, right-censored, and for ignoring censoring, respectively.



## 6 Concluding Remarks

### 6.1 Overview

In this dissertation, we have developed a groundbreaking new approach to cancer clinical trials which uses reinforcement learning ( $Q$ -learning) techniques from computer science to discover optimal tailoring treatment regimens for cancer. The idea of using reinforcement learning in clinical trials is a paradigm shift from the standard approach — of selecting the best treatment from a small set of pre-defined treatment options assigned to an assumed-to-be-homogeneous group of patients — to evaluating a continuum of treatment options and optimizing over a varied range of patients with different clinical histories and symptoms. Our work was motivated by real NSCLC trial examples from University of North Carolina Lineberger Comprehensive Cancer Center protocol 9719, “Phase III randomized trial comparing a defined duration versus continuous administration of combination chemotherapy in advanced non-small cell carcinoma of the lung”, and protocol 2003, “Phase II randomized trial comparing weekly administration of taxol with IIIB/IV non-small cell lung cancer”.

In Chapter 4, we performed a simulation study of a simple reinforcement learning approach for discovering optimal dosing for treatment of a generic cancer. The disease model is based on a simple differential equation that balances a simulated chemotherapy agent’s efficacy and toxicity. By utilizing  $Q$ -learning, the clinical reinforcement trial was able to find the best treatment rule for dosing of the agent based on biomarkers available from the patient. Our approach uses SVR and ERT to fit  $Q$  functions at each decision point. Clearly, the optimal treatment is superior by 6 months after initiation

of treatment, although it is not optimal at 2 months. This demonstrate the ability of reinforcement learning to not only adapt to individual patient needs but to discover the proper tradeoff between short and long term effects of treatment.

In Chapter 5, we significantly refined our reinforcement learning model for an advanced NSCLC trial to account for changes of the best two lines of treatment along with optimal times of initiating second-line therapy across patients. Our  $\epsilon$ -SVR-C method incorporated with  $Q$ -learning models overall survival in the timing, covariate effects, and appropriate patient heterogeneity, while taking into account right censored patients in the loss function of SVR using all information. Our approach is powerful since covariate effects of patient are embedded in the design within the multi-stage structure. Our simulations show that the method does a good job of assigning patients in favor of a superior treatment, and that it does this reliably within subgroups when there is treatment-covariate interaction. Simulation studies also demonstrate that our method requires relatively small sample sizes to approximate true optimal therapies, and sensitivity studies indicate that the correct selection probabilities increase with sample size.

## 6.2 Future Research

There are a number of challenges we expect to face in future research. First of all, in Chapter 4 we have defined the reward as a straightforward function to map states and actions into some integer numbers (15, 5, 0,  $-5$  and  $-60$ ). This simplistic reward function construction along with the  $Q$ -learning represents an attractive way for trading off efficacy against toxicity and death. However, it is unclear how changing these numbers affects the resulting optimal regimens identified during discovery of effective therapeutic strategies. Understanding the robustness of  $Q$ -learning to numerical reward choices is an interesting problem and clearly deserves further investigation.

Secondly, in Chapter 4 we observed that with sample size  $N = 1000$  for a clinical

reinforcement trial, using SVR or ERT leads to an apparently small bias for estimating optimal regimens. The evidence for this is the confirmed success of the discovered treatment regimen on an independent sample of 200 simulated patients. Similarly, in Chapter 5 we studied the prediction accuracy of our method with varying sample sizes in an NSCLC trial. The posterior analysis shows that with sample size  $N \geq 100$  per group our method can yield a small estimation bias. Although both results indicate that good estimation can be achieved when sample size is relatively small, this assumption may be violated in many settings due to the complexity associated with the performance of the approximation on the  $Q$  function, the high-dimensional state or action space, the horizon time  $T$ , the connection with SVR or ERT, and more importantly, estimation accuracy. Therefore, an interesting but potentially difficult question would be: how to determine an appropriate sample size for a clinical reinforcement trial, which allows utilizing the SVR or ERT to fit  $Q$  and can be guaranteed to reliably obtain a treatment policy that is very close to the true optimal one? This sample size calculation is related to the statistical learning error problem. Recently, there has been considerable interest in studying the generalization error for  $Q$ -learning. Murphy (2005) derived finite sample upper bounds in a closely related setting which depend on the number of observations in the training set, the number of decision points, the performance of the approximation on the training set, and the complexity of the approximation space. We believe further development of this theory is needed to better understand how the performance of  $Q$ -learning with SVR is related to the sample size of the training data in clinical reinforcement trials. We hope that this dissertation will serve to stimulate interest in these issues.

In Chapter 5, we discussed some possible extensions of our method regarding the right censored observations. We would like to briefly revisit this issue that arose and discuss how it relates to more realistic problems in NSCLC clinical trials. We mentioned that right censored cases include patients who drop out of the study without further measurements, and we mentioned that the classical SVR method may need to be modified

to  $\epsilon$ -SVR-C to account for such right censored observations. By doing this,  $Q_2$  is fitted with  $\epsilon$ -SVR-C through:

$$Q_2(o_2, d_2, T_M) = T_2 \wedge (C - t_2), \quad (6.1)$$

and furthermore,  $Q_1$  is fitted with  $\epsilon$ -SVR-C through:

$$Q_1(o_1, d_1) = T_1 \wedge C \wedge t_2 + Y_D \times \max_{d_2, T_M} Q_2(o_2, d_2, T_M). \quad (6.2)$$

This assumption of independent censoring may be too simplistic to handle more complex situations in NSCLC trials. For example, given first-line therapy, some patients determine to drop out of the study due to adverse reactions or lack of improvement and are not willing to participate in the second-line therapy. However, these dropout patients can still be followed-up until the patients' death or administrative censoring. That is,  $O_2$  can not be measured during this process but  $T_D$  or  $C$  can be. Such issues have motivated the development of a possible alternative to the  $Q$ -learning procedure described in Chapter 5.

To achieve this alternative, we propose a modified  $Q$ -learning design incorporated with  $\epsilon$ -SVR-C. Let  $D$  denote the indicator of a dropout event.  $D = 1$  indicates patients dropout at or before  $t_2$  without any measurements as well as without evaluation of second-line therapy. In place of  $Q_2$  in (6.1), we then define  $Q'_2$  at  $t_2$  for patients who have status  $D = 1$ , which is fitted with  $\epsilon$ -SVR-C through:

$$Q'_2(o_1, d_1) = T_2 \wedge (C - t_2). \quad (6.3)$$

Compared to (6.1),  $o_1$  and  $d_1$  are embedded in  $Q'_2$ . In addition, the corresponding  $Q_1$  is modified as:

$$\begin{aligned} Q_1(o_1, d_1) &= T_1 \wedge C \wedge t_2 \\ &+ Y_D \times \left[ D \times Q'_2(o_1, d_1) + (1 - D) \times \max_{d_2, T_M} Q_2(o_2, d_2, T_M) \right]. \end{aligned} \quad (6.4)$$

Note that equation (6.4) is reduced to (6.2) when  $D = 0$ . This extension for accounting for patients who dropout but who are followed would be conceptually straightforward if

the indicator  $D$  is observed. Thus, when  $D$  is unknown, a continuing challenge will be to develop methods which model  $D$  efficiently to ensure that  $Q$ -learning is viable in these settings.

In future research, we also plan to address the following issues:

- (1) NSCLC clinical trial design. We will develop a protocol for a Stage IIIB/IV NSCLC clinical reinforcement trial. This will include identifying and refining all of the needed aspects which have been described in Chapter 5. Part of this process will involve identifying what and how to randomize at various decision points in a manner that is consistent with standards of clinical practice and avoids randomizing to inferior treatments. This approach is quite new and may involve several iterations before a suitable and efficient design is achieved.
- (2) Adaptation to other cancers. The general principals and methods of this approach are very adaptive to other cancers in addition to NSCLC, such as breast and colon cancers, and we plan to develop general guidelines to pursue this. As part of this, we expect to be able to start identifying specific other cancer treatment questions, and then to use the differences between these cancers to formulate a general process for developing reinforcement trials in a broad range of cancer settings.
- (3) Creation of software tools. Clearly, we will develop user-friendly software to implement our reinforcement learning method freely for public use. We will also develop software for the clinical trial design and analysis of clinical reinforcement trials. We believe both of these phases will be valuable to other clinical researchers.

## Bibliography

- Adams, B., Banks, H., Kwon, HD., and Tran, H. (2004). Dynamic multidrug therapies for HIV: Optimal and STI control approaches. *Mathematical Biosciences and Engineering* 1, 223-241.
- Bellman, RE. (1957). *Dynamic programming*. Princeton University Press, Princeton.
- Bertsekas, DP. (1987). *Dynamic programming: deterministic and stochastic models*. Prentice-Hall, Englewood Cliffs, NJ.
- Blatt, D., Murphy, SA., and Zhu, J. (2004). *A-learning for approximate planning*. Unpublished manuscript.
- Bulzebruck, H., Bopp, R., Drings, P., Bauer, E., Krysa, S., Probst, G., van Kaick, G., Muller, KM., and Vogt-Moykopf, I. (1992). New aspects in the staging of lung cancer. Prospective validation of the International Union Against Cancer TNM classification. *Cancer* 70, 1102-1110.
- Bunn, P. and Kelly, K. (1998). New chemotherapeutic agents prolong survival and improve quality of life in non-small cell lung cancer: A review of the literature and future directions. *Clinical Cancer Research* 4, 1087-1100.
- Buzaid, AC. (2000). Strategies for combining chemotherapy and biotherapy in melanoma. *Cancer Control* 7, 185-197.
- Buzaid, AC. and Atkins, M. (2001). Practical guidelines for the management of biochemotherapy related toxicity in melanoma. *Clinical Cancer Research* 7, 2611-2619.
- Chan, C., George, A., and Stark, J. (2003). T cell sensitivity and specificity – kinetic proofreading revisited. *Discrete and Continuous Dynamical Systems Series B* 3, 343-360.
- Chen, BJ., Chang, MW., and Lin, CJ. (2004). Load forecasting using support vector machines: A study on EUNITE competition 2001. *IEEE Transactions on Power Systems* 19, 1821-1830.
- Chow, SC., Chang, M., and Pong, A. (2005). Statistical consideration of adaptive methods in clinical development. *Journal of Biopharmaceutical Statistics* 15, 575-

- Ciuleanu, TE., Brodowicz, T., Belani, CP., Kim, J., Krzakowski, M., Laack, E., Wu, Y., Peterson, P., Adachi, S., and Zielinski, CC. (2008). Maintenance pemetrexed plus best supportive care (BSC) versus placebo plus BSC: A phase III study. *Journal of Clinical Oncology* 26, May 20 supplement, abstract 8011.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning* 20, 273-297.
- Cover, TM. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers* 14, 326-334.
- Crammer, K. and Singer, Y. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research* 2, 265-292.
- Crammer, K. and Singer, Y. (2003). Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research* 3, 951-991.
- Crites, RH. and Barto, AG. (1998). Elevator group control using multiple reinforcement learning agents. *Machine Learning* 33, 235-262.
- de Pillis, L. and Radunskaya, A. (2001). A mathematical tumor model with immune resistance and drug therapy: an optimal control approach. *Journal of Theoretical Medicine* 3, 79-100.
- de Pillis, LG. and Radunskaya, A.E. (2003). Immune response to tumor invasion. In: *Bathe, K. (Ed.), Computational Fluid and Solid Mechanics* vol. 2, MIT Press, Cambridge, MA, 1661-1668.
- de Pillis, LG., Gu, W., and Radunskaya, AE. (2006). Mixed immunotherapy and chemotherapy of tumors: modeling, applications and biological interpretations. *Journal of Theoretical Biology* 238, 841-862.
- de Pillis, LG., Gu, W., Fister, KR., Head, T., Maples, K., Murugan, A., Neal, T., and Yoshida, K. (2007a). Chemotherapy for Tumors: an analysis of the dynamics and a study of quadratic and linear optimal controls. *Mathematical Biosciences* 209, 292-315.

- de Pillis, LG., Fister, KR., Gu, W., Collins, C., Daub, M., Gross, D. Moore, J., and Preskill, B. (2007b). Seeking bang-bang solutions of mixed immuno-chemotherapy of tumors. *Electronic Journal of Differential Equations* 171, 1-24.
- Ernst, D., Geurts, P., and Wehenkel, L. (2005). Tree-based batch model reinforcement learning. *Journal of Machine Learning Research* 6, 503-556.
- Ernst, D., Stan, G-B., Goncalves, J., and Wehenkel, L. (2006). Clinical data based optimal STI strategies for HIV: a reinforcement learning approach. *Proceedings of the 45<sup>th</sup> IEEE Conference on Decision and Control*.
- Ettinger, DS., Akerley, W., Bepler, G. et al. (2007). NCCN clinical practice guidelines in oncology: Non-small cell lung cancer. V1.
- Farrar, JD., Katz, KH., Windsor, J., Thrush, G., Scheuermann, RH., Uhr, JW., and Street, NE. (1999). Cancer dormancy. VII. A regulatory role for CD8+ T cells and IFN-gamma in establishing and maintaining the tumor-dormant state. *Journal of Immunology* 162, 2842-2849.
- Fidias, P., Dakhil, S., Lyss, A., Loesch, D., Waterhouse, D., Cunneen, J., Chen, R., Treat, J., Obasaju, C., and Schiller, J. (2007). Phase III study of immediate versus delayed docetaxel after induction therapy with gemcitabine plus carboplatin in advanced non-small-cell lung cancer: Updated report with survival. *Journal of Clinical Oncology* 25, June 20 supplement, LBA7516.
- Fister, KR. and Donnelly, JH. (2005). Immunotherapy: an optimal control theory approach. *Mathematical Biosciences* 3, 499-510.
- Fister, KR. and Panetta, JC. (2003). Optimal control applied to competing chemotherapeutic cell-kill strategies. *SIAM Journal on Applied Mathematics* 63, 1954-1971.
- Fleming, WH. and Rishel, RW. (1975). *Deterministic and Stochastic Optimal Control*. Springer-Verlag, New York.
- Food and Drug Administration (2004). Innovation or stagnation: Challenge and opportunity on the critical path to new medical products. *White paper*.
- Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M., and Pinheiro, J. (2006). Adaptive design in clinical drug development - an executive summary of the



- PhRMA Working Group (with discussions). *Journal of Biopharmaceutical Statistics* 16, 275-283.
- Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning* 11, 3-42.
- Gosavi, A., Bandla, N., and Das, TK. (2002). A reinforcement learning approach to a single leg airline revenue management problem with multiple fare classes and over-booking. *IIE Transactions* 34, 729-742.
- Guez, A., Vincent, R., Avoli, M., and Pineau, J. (2008). Adaptive treatment of Epilepsy via batch-mode reinforcement learning. *Innovative Applications of Artificial Intelligence* March 25, 1671-1678.
- Hanna, N., Shepherd, FA., Fossella, FV., Pereira, JR., De Marinis, F., von Pawel, J., Gatzemeier, U., Tsao, TC., Pless, M., Muller, T., Lim, HL., Desch, C., Szondy, K., Gervais, R., Shaharyar, Manegold, C., Paul, S., Paoletti, P., Einhorn, L., Bunn, PA. Jr. (2004). Randomized phase III trial of pemetrexed versus docetaxel in patients with non-small-cell lung cancer previously treated with chemotherapy. *Journal of Clinical Oncology* 22, 1589-1597.
- Hartl, RF., Sethi, SP., and Vickson, RG. (1995). A survey of the maximum principles for optimal control problems with state constraints. *SIAM Review* 2, 181-218.
- Hogberg, T. (2005). Widening bottlenecks in drug discovery Glimpses from Drug Discovery Technology Europe. *Drug Discovery Today* 10, 820-822.
- Holford, N. and Sheiner, L. (1981). Pharmacokinetic and pharmacodynamic modeling in vivo. *CRC Critical Reviews in Bioengineering* 5, 273-322.
- Howard, R. (1960). *Dynamic programming and Markov processes*. MIT Press, Cambridge, MA.
- Ivanov, VV. (1976). *The theory of approximate methods and their application to the numerical solution of singular integral equations*. Noordhoff International, Leyden.
- Jaakkola, T., Jordan, MI., and Singh, SP. (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation* 6, 1185-1201.

- Kaelbling, LP., Littman, ML., and Moore, AW. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research* 4, 237-285.
- Kamien, MI. and Schwartz, NL. (1991). *Dynamic optimization: the calculus of variations and optimal control in economics and management. Advanced Textbooks in Economics*. Second Ed., vol. 31, North-Holland, Amsterdam.
- Kimeldorf, GS. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics* 41, 495-502.
- Kirschner, D. and Panetta, JC. (1998). Modeling immunotherapy of the tumor-immune interaction. *Journal of Mathematical Biology* 37, 235-252.
- Krener, AJ. (1977). The high order maximal principle and its application to singular extremals. *SIAM Journal on Control and Optimization* 15 (2), 256.
- Lavori, PW., Rush, AJ., Wisniewski, SR., Alpert, J., Fava, M., Kupfer, DJ., Nierenberg, A., Quitkin, FM., Sackeim, HA., Thase, ME. and Trivedi, M. (2001). Strengthening clinical effectiveness trials: equipoise-stratified randomization. *Biological Psychiatry* 50, 792-801.
- Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association* 99, 67-81.
- Lin, Y. (2002). Support vector machines and the Bayes rule in classification. *Data Mining and Knowledge Discovery* 6, 259-275.
- Liu, Y. and Shen, X. (2006). Multicategory  $\psi$ -learning. *Journal of the American Statistical Association* 101, 500-509.
- Matveev, AS. and Savkin, AV. (2002). Application of optimal control theory to analysis of cancer chemotherapy regimens. *Systems & Control Letters* 46, 311-321.
- Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A* 209, 415-446.

- Moodie, EEM., Richardson, TS., and Stephens, DA. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics* 63, 447-455.
- Morecki, S., Pugatsch, T., Levi, S., Moshel, Y., and Slavin, S. (1996). Tumor-cell vaccination induces tumor dormancy in a murine model of B-cell leukemia/lymphoma (BCL1). *International Journal of Cancer* 65, 204-208.
- Muller, M., Gounari, F., Prifti, S., Hacker, HJ., Schirmacher, V., and Khazaie, K. (1998). EblacZ tumor dormancy in bone marrow and lymph nodes: active control of proliferating tumor cells by CD8+ immune T cells. *Cancer Research* 58, 5439-5446.
- Murphy, SA. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B* 65 Part 2, 331-366.
- Murphy, SA. (2005a). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* 24, 1455-1481.
- Murphy, SA. (2005b). A generalization error for  $Q$ -learning. *Journal of Machine Learning Research* 6, 1073-1097.
- Murphy, SA., Lynch, KG., Oslin, D., McKay, JR., and TenHave, T. (2007a). Developing adaptive treatment strategies in substance abuse research. *Drug and Alcohol Dependence* 88S, S24-S30.
- Murphy, SA., Oslin, DW., Rush, AJ., and Zhu, J. (2007b). Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychopharmacology* 32, 257-262.
- Murray, JM. (1990a). Optimal control for a cancer chemotherapy problem with general growth and loss functions. *Mathematical Biosciences* 98, 273-287.
- Murray, JM. (1990b). Some optimal control problems in cancer chemotherapy with a toxicity limit. *Mathematical Biosciences* 100, 49-67.
- Norton, L. and Simon, R. (1977). Tumor size, sensitivity to therapy, and design of treatment schedules. *Cancer Treatment Reports* 61, 1307-1317.
- Norton, L. and Simon, R. (1986). The Norton-Simon hypothesis revisited. *Cancer Treatment Reports* 70, 163-169.

- Ng et al. (2006). Inverted autonomous helicopter flight via reinforcement learning. In *Experimental Robotics IX: The 9th International Symposium on Experimental Robotics*. New York: Springer, 363-371.
- O'Byrne, K.J., Dalgleish, A.G., Browning, M.J., Steward, W.P., and Harris, A.L. (2000). The relationship between angiogenesis and the immune response in carcinogenesis and the progression of malignant disease. *European Journal of Cancer* 36, 151-169.
- Ormoneit, D. and Sen, S. (2002). Kernel-based reinforcement learning. *Machine Learning* 49, 161-178.
- Pednault, E., Abe, N., and Zadrozny, B. (2002). Sequential cost-sensitive decisionmaking with reinforcement learning. *Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining* 259-268.
- Phillips, D.L. (1962). A technique for the numerical solution of certain integral equations of the first kind. *Journal of the Association for Computing Machinery* 9, 84-97.
- Pineau, J., Bellemare, M.G., Rush A.J., Ghizaru, A., and Murphy, S.A. (2007). Constructing evidence-based treatment strategies using methods from computer science. *Drug and Alcohol Dependence* 88S, S52-S60.
- Pirker, R., Szczesna, A., von Pawel, J., Krzakowski, M., Ramlau, R., Park, K., Gatzemeier, U., Bajeta, E., Emig, M., and Pereira, J.R. (2008). FLEX: A randomized, multicenter, phase III study of cetuximab in combination with cisplatin/vinorelbine (CV) versus CV alone in the first-line treatment of patients with advanced non-small cell lung cancer (NSCLC). *Journal of Clinical Oncology* May 20 suppl, abstract 3.
- Pontryagin, L.S., Boltyanskii, V.G., Gamkrelidze, R.V., and Mishchenko, E.F. (1962). *The Mathematical Theory of Optimal Processes*. Gordon and Breach, New York.
- Puterman, M.L. and Shin, M.C. (1978). Modified policy iteration algorithms for discounted Markov decision problems. *Management Science* 24, 1127-1137.
- Robins, J.M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium on Biostatistics* D. Y. Lin and P. Heagerty (eds), Springer, New York, 189-326.
- Rummery, G.A. and Niranjan, M. (1994). *On-line Q-learning using connectionist sys-*

*tems*. Technical Report CUED/F-INFENG/TR166, Cambridge University.

- Rush, A.J., Crismon, M.L., Kashner, T.M., Toprac, M.G., Carmody, T.J., Trivedi, M.H., Suppes, T., Miller, A.L., Biggs, M.M., Shores-Wilson, K., Witte, B.P., Shon, S.P., Rago, W.V., and Altshuler, K.Z. (2003). Texas Medication Algorithm Project, Phase 3 (TMAP-3): Rationale and Study Design. *Journal of Clinical Psychiatry* 64, 357-369.
- Sandler, A., Gray, R., Perry, M.C., Brahmer, J., Schiller, J.H., Dowlati, A., Lilenbaum, R., and Johnson, D.H. (2006). Paclitaxel-Carboplatin alone or with bevacizumab for non-small-cell lung cancer. *The New England Journal of Medicine* 355, 2542-2550.
- Scagliotti, G.V., Parikh, P., von Pawel, J., Biesma, B., Vansteenkiste, J., Manegold, C., Serwatowski, P., Gatzemeier, U., Digumarti, R., Zukin, M., Lee, J.S., Mellemegaard, A., Park, K., Patil, S., Rolski, J., Goksel, T., de Marinis, F., Simms, L., Sugarman, K.P., and Gandara, D. (2008). Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non-small-cell lung cancer. *Journal of Clinical Oncology* 26.
- Schabel, F., Skipper, H., and Wilcox, W. (1964). Experimental evaluation of potential anticancer agents. XIII. On the criteria and kinetics associated with curability of experimental leukemia. *Cancer Chemother Report* 25, 1-111.
- Schiller, J.H., Harrington, D., Belani, C.P., Langer, C., Sandler, A., Krook, J., Zhu, J., and Johnson, D.H. (2002). Comparison of four chemotherapy regimens for advanced non-small cell lung cancer. *New England Journal of Medicine* 346, 92-98.
- Schneider, L.S., Tariot, P.N., Lyketsos, C.G., Dagerman, K.S., Davis, K.L., Davis, S., Hsiao, J.K., Jeste, D.V., Katz, I.R., Olin, J.T., Pollock, B.G., Rabins, P.V., Rosenheck, R.A., Small, G.W., Lebowitz, B., and Lieberman, J.A. (2001). National Institute of Mental Health clinical antipsychotic trials of intervention effectiveness (CATIE): Alzheimer disease trial methodology. *American Journal of Geriatric Psychiatry* 9, 346-360.
- Seierstad, A. and Sydsaeter, K. (1987). *Optimal control theory with economic applications*. North Holland, Amsterdam, 3rd edition.
- Shepherd, F.A., Dancey, J., Ramlau, R., Mattson, K., Gralla, R., O'Rourke, M., Levitan, N., Gressot, L., Vincent, M., Burkes, R., Coughlin, S., Kim, Y., and Berille, J. (2000). Prospective randomized trial of docetaxel versus best supportive care

- in patients with non-small-cell lung cancer previously treated with platinum-based chemotherapy. *Journal of Clinical Oncology* 18, 2095-2103.
- Shepherd, FA., Pereira, JR., Ciuleanu, T., Tan, EH., Hirsh, V., Thongprasert, S., Campos, D., Maoleekoonpiroj, S., Smylie, M., Martins, R., van Kooten, M., Dediu, M., Findlay, B., Tu, D., Johnston, D., Bezjak, A., Clark, G., Santabarbara, P., and Seymour, L. (2005). Erlotinib in previously treated non-small-cell lung cancer. *The New England Journal of Medicine* 353, 123-132.
- Shivaswamy, P., Chu, W., and Jansche, M. (2007). A Support Vector Approach to Censored Targets. *Proceedings of the International Conference on Data Mining*. Omaha, NE.
- Smola, A. and Scholkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing* 14, 199-222.
- Socinski, MA., Crowell, R., Hensing, TE., Langer, CJ., Lilenbaum, R., Sandler, AB., and Morris D. (2007). Treatment of non-small cell lung cancer, stage IV. ACCP evidence-based clinical practice guidelines. *Chest* 132, 3, supplement.
- Socinski, MA. and Stinchcombe, TE. (2007). Duration of first-line chemotherapy in advanced non small-cell lung cancer: less is more in the era of effective subsequent therapies. *Journal of Clinical Oncology* 25, 5155-5157.
- Stewart, TH. (1996). Immune mechanisms and tumor dormancy. *Medicina (Buenos Aires)* 56, 74-82.
- Stinchcombe, TE. and Socinski, MA. (2008). Considerations for second-line therapy of non-small cell lung cancer. *The Oncologist* 13 (suppl 1), 28-36.
- Sutton, RS. (1988). Learning to predict by the method of temporal differences. *Machine Learning* 3(1), 9-44.
- Sutton, RS. and Barto, AG. (1998). *Reinforcement learning: an introduction*. MIT Press, Cambridge, MA.
- Swan, GW. (1986). Optimal control using the Verhulst-Pearl equation. *Bulletin of Mathematical Biology* 48, 381-404.
- Swan, GW. (1990). Role of optimal control theory in cancer chemotherapy. *Mathemat-*

*ical Biosciences* 101, 237-284.

Tesauro, G. (1994). TD-gammon, a self-teaching Backgammon program, achieves master-level play. *Neural Computation* 6, 215-219.

Tesauro, G. (2002). Programming Backgammon using self-teaching neural nets. *Artificial Intelligence* 134, 181-199.

Thall, PF., Millikan, RE., and Sung, HG. (2000). Evaluating multiple treatment courses in clinical trials. *Statistics in Medicine* 19, 1011-1028.

Thall, PF., Inoue, LYT., and Martin, TG. (2002). Adaptive decision making in a lymphocyte infusion Trial. *Biometrics* 58, 560-568.

Thall, PF., Sung, HG., and Estey, EH. (2002). Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *Journal of the American Statistical Association* 457, 29-39.

Thall, PF., Wooten, LH., Logothetis, CJ., Millikan, RE., and Tannir, NM. (2007). Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Statistics in Medicine* 26, 4687-4702.

Tikhonov, AN. and Arsenin, VY. (1977). *Solutions of ill-posed problems*. Wiley, New York.

Tsitsiklis, JN. (1994). Asynchronous stochastic approximation and Q-learning. *Machine Learning* 16, 185-202.

Tsitsiklis, JN. and Van Roy, B. (1996). Feature-based methods for large scale dynamic programming. *Machine Learning* 22, 59-94.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer, New York.

Vapnik, V., Golowich, S., and Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems* 9, 281-287.

Watkins, CJCH. (1989). *Learning from Delayed Rewards*. Ph.D. Thesis, King's College, Cambridge, UK.

Watkins, CJCH. and Dayan, P. (1992). *Q*-learning. *Machine Learning* 8, 279-292.