Dynamic, convex, and robust optimization with Bayesian learning for response-guided dosing

Jakob Kotas

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

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Archis Ghate, Chair

Minsun Kim

Emanuil Todorov

Zelda Zabinsky

Program Authorized to Offer Degree:
Applied Mathematics

University of Washington

Abstract

Dynamic, convex, and robust optimization with Bayesian learning for response-guided dosing

Jakob Kotas

Chair of the Supervisory Committee:
Associate Professor Archis Ghate
Industrial & Systems Engineering and Applied Mathematics

Medical treatment commonly involves the administration of drug doses at multiple time-points. Intuitively, the higher the doses, the higher the likelihood of disease control as well as the risk of adverse effects and of logistical inconvenience. Since an individual patient's response to treatment is uncertain, the need to effectively balance this trade-off pervades all of medicine. In response-guided dosing (RGD), the goal is to adaptively tailor doses to each individual patient's stochastic evolution of disease condition over multiple treatment sessions. Several clinical experts, in editorial and review papers, have commented that despite a strong surge of interest in RGD, a quantitative, dynamic decision-making framework has been missing. The **research objective** of this dissertation is to apply stochastic dynamic programming (DP), convex optimization, and Bayesian learning methods to develop such a mathematically rigorous framework to facilitate dosing decisions in RGD. The ultimate goal of this framework is to administer the right dose to the right patient at the right time.

RGD for rheumatoid arthritis. The first chapter begins with a stochastic DP framework to facilitate RGD in rheumatoid arthritis, which adapts biologic doses over the treatment course based on each patient's observed evolution of the 28-joint disease activity score (DAS28). The goal is to balance the DAS28 attained at the end of the course with the weighted total dose administered. Numerical experiments and sensitivity analyses using

data from the OPTION trial are performed, which are found to be monotone.

A general stochastic DP formulation for RGD. The specific rheumatoid arthritis formulation is then generalized for other diseases. The DP allows for an arbitrary dose-response function, and balances the disutility of doses with the disutility of the disease condition reached. We prove that under assumptions on the underlying functions, there exists an optimal dosing policy which is monotone with respect to patient state, and provide several examples where these conditions are met.

Robust RGD. We then study a robust counterpart of the stochastic DP model of the previous chapter, where the pmf of the distribution of the stochastic dose-response parameter is unknown but is assumed to belong to an interval uncertainty set. We show that the inner maximization problem of the robust Bellman's equations is a linear program with a closed-form solution. We prove monotonicity of optimal dose with respect to both disease state and an ambiguity parameter, and illustrate monotonicity via simulation.

Optimal Bayesian learning of dose-response parameters from a cohort. In this chapter, we study the problem of finding optimal RGD policies while learning the unknown distribution on a stochastic dose-response parameter from a cohort of patients. We provide a Bayesian stochastic DP formulation, though exact solution of Bellman's equations is computationally intractable. We therefore present two approximate control schemes and analyze the monotonicity, stationarity, and separability structures of the resulting dosing strategies, which are exploited in efficient, approximate solution of our problem. Numerical experiments are completed, and results are compared to non-Bayesian methods.

Optimal stopping for RGD. In the final chapter, we consider an optimal stopping variant of RGD, where the decision-maker is allowed to end treatment prematurely. This could occur, for example, when the patient responds well quickly so that further treatment is unnecessary. We numerically demonstrate that for some problems, it is optimal to stop treatment in states better than a certain threshold.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost I would like to express my utmost appreciation to Dr. Archis Ghate, my PhD advisor. He taught me most of what I know and I am endlessly grateful for the countless hours he spent with me discussing research, and for his invaluable guidance and advocacy which have pushed my career forward; not to mention his generous funding of my research appointments and conference travel through his grants. Besides his vast intellect, he is humble, open-minded, friendly, and an excellent communicator. I have been honored to have him as an advisor.

I am deeply indebted to Dr. Minsun Kim, whose work in applying optimization methods to radiation oncology problems spun off many interesting research directions that eventually resulted in several PhD theses, including my own. I also appreciate the unique insight she brought as both an applied mathematician and practicing clinician to our work on the reaction-diffusion model for GBM treatment.

To Dr. Zelda Zabinsky, thank you so much for your letter of recommendation and for agreeing to attend my defense remotely while on sabbatical.

I would also like to express my sincere thanks to Dr. Emo Todorov and Dr. Anne Goodchild, who have sacrificed their scarce and valuable time to serve on my doctoral committee.

I cannot thank Dr. Bernard Deconinck, Dr. Mark Kot, and Dr. John Sylvester enough for supporting my interest in teaching. They were generous with teaching assignments, visited my classroom and gave valuable feedback, and wrote letters for my job applications. Besides giving me wonderful opportunities to teach that I have greatly enjoyed, I would not have succeeded in finding a job without them. I would like to also thank the anonymous student who nominated me, and Dr. Nathan Kutz and my students who wrote letters in support

of that nomination for the Excellence in Teaching Award. Thank you to the Applied Math Department for honoring me twice with the Boeing Teaching Award.

Thank you also to Lauren Lederer, a stellar administrator who was extremely helpful in helping me navigate beauracracy, meet deadlines, send letters, and countless other tasks.

To my peers in the Applied Mathematics Department, present and past, thank you for the camaraderie and the many illuminating discussions. A special thanks goes to Scott Moe, whose presence in Lewis 214 the last three years was always appreciated. Thank you to Saumya Sinha for her collaboration on our work on robust response-guided dosing. Thank you also to Bethany Lusch for introducing me to my advisor.

I am deeply appreciative of Dr. Richard Rand's patient and lucid explanations of topics in dynamical systems, and Dr. Christoffer Heckman's collaboration and guidance. They stoked my interest in applied mathematics research and kickstarted my career. I also thank them for supporting me as I transitioned from my master's to PhD.

To the many professors and teachers of my past who sparked my intellectual curiosity, thank you. In particular, thanks to Dr. Steven Strogatz, whose Nonlinear Dynamics & Chaos class persuaded me to study applied mathematics in graduate school.

Thank you to my dear friends who celebrated my successes and helped me up when I stumbled during my time in graduate school. I could not have survived without your care and compassion.

And to my parents, thank you for your unconditional love, for the sacrifices you made for me, and for teaching me to make use of my talents.

Lastly, I would like to thank the UW Applied Mathematics Department, the Washington Research Foundation, the Achievement Rewards for College Scientists (ARCS) Foundation, and the National Science Foundation for their generous financial support.

DEDICATION

This dissertation is dedicated to Scott Smedinghoff (1987-2016),
an irreplaceable friend
and the most brilliant mathematician I have known.

"Why else am I going to grad school in math?
if not to play piano concertos"

Chapter 1

RESPONSE-GUIDED DOSING FOR RHEUMATOID ARTHRITIS

1.1 Background and motivation

Rheumatoid arthritis (RA) is an auto-immune disease that usually strikes between the ages of 35 and 50 [33, 77]. Its specific trigger is unknown. About 1.3 million adults in the US and 1% of the world's population suffer from RA [79]. It is a debilitating condition that can affect the whole body, but mainly occurs in joints especially in the hands and feet. RA leads to deformity, disability, pain, loss of productivity, and loss of quality of life. It thus has a considerable impact on our society [38].

RA patients were traditionally treated with non-biologic disease modifying antirheumatic drugs (DMARDs) [187]. Methotrexate is the most commonly used non-biologic DMARD [196]. However, many patients on methotrexate continue to exhibit inflammation and progressive joint destruction [3, 30, 39, 141, 159, 187]. These patients are increasingly being treated with a combination of methotrexate and biologic agents such as adalimumab, certolizumab, etanercept, golimumab, infliximab, abatacept, rituximab, and tocilizumab [3, 125]. The ATTRACT, ASPIRE, BeST, PREMIER and other clinical studies have shown the benefit of such combination therapy [31, 39, 42, 54, 82, 106, 114, 127, 142, 177, 180, 187, 194]. Combination therapy with methotrexate and a biologic agent is becoming the gold standard of RA treatment [43, 192].

Standard biologic treatment for RA follows a one-size-fits-all approach. The aforementioned eight biologic agents have now been included in the 2012 guidelines for RA treatment by the American College of Rheumatology (ACR) [170]. Table 1.1 lists their typical dosage.

It is estimated that millions of patients worldwide had been treated with three of the

Biologic agent	Dosage
infliximab	3-10 mg per kg every 4-8 weeks
etanercept	50 mg weekly
adalimumab	40 mg monthly
certolizumab	200 mg every other week or 400 mg monthly
golimumab	50 mg per month
abatacept	500-1000 mg every 4 weeks according to weight
rituximab	2 separate 1000 mg doses 2 weeks apart every 6 months
tocilizumab	4-8 mg per kg every 4 weeks

Table 1.1: Typical dosing regimen for eight biologic agents listed in the 2012 ACR guidelines [170]. This table is adapted from [3].

above eight biologic agents by 2011 [177]. Annual sales for a biologic with the brand-name Humira equaled $9.3 billion, thus making it the largest-selling drug (of any kind for any disease) in 2012; two other brand-name biologic products, Enbrel and Remicade, were also in the top five on this list [35, 99]. The overall world-market for RA medicines is expected to reach $38.5 billion in 2017 with a majority of the share going to biologic agents [150, 152].

The challenges in treating RA with biologic agents include:

- **uncertain response** — biologic agents are manufactured inside living organisms and are structurally complex, large molecules about 200-1000 times the size of other small-molecule drugs; thus they are highly sensitive and difficult to characterize [84];

- **financial cost** — annual payer cost of biologic treatment ranges between $15,000 to $20,000, which is about three times that of standalone methotrexate treatment [38, 133];

- **side effects** — biologics carry a risk of infections and hence patients are screened

annually for tuberculosis, receive annual influenza vaccination, and hepatitis B vaccination [3, 93, 145];

- **logistic inconvenience** — biologics are administered either intravenously or subcutaneously, whereas methotrexate is often administered orally [3, 193].

Owing to the aforementioned challenges, RA treatment with biologic agents must be planned judiciously [3, 188]. Recent clinical trials have therefore considered response-guided dosing (RGD) with biologic agents. RGD is also called "tight control" [69, 160]. In contrast to one-size-fits-all therapy, the idea in RGD is to adjust dose levels (this is often called titration) based on the observed evolution of the 28-joint disease activity score (DAS28). Potential benefits of RGD include a reduction in over- and under-dosing, improved cost-effectiveness, safer treatment regimens, lesser inconvenience to patients, and better disease-control. As such, the ultimate goal in RGD is to administer the right dose to the right patient at the right time.

DAS28 is a nonnegative score where higher numbers indicate higher RA activity [62, 63]. It is a composite score based on the number of tender joints, the number of swollen joints, a numerical value of the patient's global health and the patient's Erythrocyte Sedimentation Rate in mm/hr. It can be easily calculated using a simple formula and has been validated as a measure of disease activity in several clinical trials [63, 189, 190, 49, 194, 198]. The European League Against Rheumatism (EULAR) criteria shown in Table 1.2 are often employed to categorize patient-response to RA treatment at two time-points during RGD [63, 190]. DAS28 values of 3.2 and 2.6 are typically considered as the thresholds for low disease activity and remission, respectively.

In Flendrie et al. [59], RA treatment with infliximab was initiated at a dose of 3mg/kg. Treatment was administered intravenously at weeks 0, 2, 6, and every 8 weeks thereafter until week 38. Patients were classified as good responders, moderate responders, and non-responders as per the above EULAR criteria depending on their DAS28 at week 14. Treat-

initial DAS28	DAS28 improvement at endpoint		
	> 1.2	> 0.6 but ≤ 1.2	≤ 0.6
DAS28 ≤ 3.2	good response	moderate response	no response
$3.2 <$ DAS28 ≤ 5.1	moderate response	moderate response	no response
$5.1 <$ DAS28	moderate response	no response	no response

Table 1.2: The EULAR response criteria based on DAS28 measurements [63, 190]. For example, the table shows that if the initial DAS28 score is more than 5.1 (last row) and the improvement in DAS28 at an endpoint is between 0.6 and 1.2 (second column) then the patient is categorized as a non-responder.

ment at 3mg/kg was continued for good responders. For moderate responders and non-responders, dose was tailored to 6 mg/kg or 10 mg/kg depending on the subsequent evolution of their DAS28. This dose escalation showed a statistically significant reduction in disease activity for moderate responders. However, non-responders continued to experience high disease activity despite dose escalation. Rahman et al. [153] showed that increasing infliximab dose to 4.5 mg/kg in patients with inadequate response to an initial infliximab dose of 3 mg/kg might be effective. Durez et al. [50] also reached a similar conclusion.

Van den Broeder [47] conducted a trial with adalimumab, wherein the dosing intervals were fixed for each patient either at 2 weeks or at 4 weeks. Patients were examined every 8 weeks. The starting dose was 3 mg/kg and then the dose was gradually decreased to 1 mg/kg, 0.5 mg/kg and 0.25 mg/kg in weeks 8, 16, and 24, respectively. In the event of a disease flare (a DAS28 increase of 1.2 or a DAS28 increase of 0.6-1.2 if it resulted in a DAS28 of more than 5.1), dose was increased by one step to the earlier level. This dose titration reduced the total amount of adalimumab given to the patients by 67 percent without any statistically significant increase in DAS28. Second, the weekly dose administered to different patients varied from 4.1 mg to 130 mg. The authors remarked that their study "demonstrated the

principle of dose titration and the advantages of this approach compared with the common one-size-fits-all standard dosing schemes." They further commented, "this approach will save costs and may prevent long term side effects." Van der Mass et al. [188] monitored DAS28 to reduce infliximab doses. They did not find any statistically significant difference in the patients' quality of life after such dose reduction. Average cost reduction was €3474 per patient.

Based on the results of the aforementioned and other clinical trials, consensus seems to be emerging about the practical value of RGD and the need for tools that guide its implementation. Indeed, in their meta-analysis of six clinical trials, Schipper et al. [160] concluded "efforts should be made to implement tight control"; "systematic disease activity monitoring should be combined with treatment adjustments"; and "consensus about optimal treatment needs to be achieved." Mease [132] also reached a similar conclusion: "there is currently no gold standard for assessing disease activity and outcomes, so patients may not receive optimal treatment over time"; "implementation of tight control into routine care will require quick and simple validated tools for defining treatment targets and monitoring disease activity"; "ultimately, tight control should lead to treatment optimization and will provide the best chances of improvement and remission for patients." In Smolen et al. [176], a committee of more than sixty experts from around the world noted that "validated composite measures of disease activity should be used in routine clinical practice to guide treatment decisions"; "measures of disease activity must be obtained and documented regularly"; "treatment should be adjusted at least every three months." Van Vollenhoven [192] stated: "there is a feeling of sadness when it turns out that it is most likely that a very large number of patients have been treated for many years with dosages that were unnecessarily high"; "we, as rheumatologists in practice and in academia, must take on the responsibility for determining the optimal use of antirheumatic drugs." Palmer and El Mledany [145] concluded that "treatment of RA should be mapped out dynamically." Finally, referring to four extensive literature reviews on tight control [91, 101, 159, 161], Smolen and Aletaha [175] stated that "they all revealed that a tight control strategy with set rules for treatment

adaptations was associated with a superior outcome when compared with unsystematic monitoring and change of therapy."

Unfortunately, there is currently no consensus or guidelines on *how* to dynamically adapt doses based on the measured evolution of DAS28 scores for individual patients. Specifically, there is no systematic, quantitative decision-making framework to achieve this and the dose changes in the aforementioned trials seem somewhat ad-hoc. In this chapter, we make some initial progress in establishing such a framework, using tocilizumab as an example, via stochastic Dynamic Programming (DP).

1.2 A concrete stochastic DP formulation

We consider a treatment course with T sessions wherein DAS28 measurements are made and a biologic agent is administered. These sessions are indexed by $t = 1, 2, \ldots, T$ and the time-interval between two consecutive sessions is assumed to be constant, say four weeks, as is common in RA treatment. The DAS28 score measured in the tth session is denoted by DAS28_t and the biologic dose chosen for this session after measuring DAS28_t is denoted by d_t. Doses d_t belong to the interval $D \triangleq [0, \bar{d}]$, where $\bar{d} < \infty$ is the maximum permissible biologic dose in one session.

We assume that DAS28 scores evolve according to the dose-response model

$$\ln(\text{DAS28}_{t+1}) = \ln(\text{DAS28}_t) + \ln \kappa_2 - \ln(\kappa_1 + \kappa_2 + d_t) + \Theta, \tag{1.1}$$

where $\kappa_1, \kappa_2 > 0$ are parameters. Here, Θ are independent and identically distributed (iid) random variables that represent uncertainty in response. The use of such stochastic noise in dose-response functions is standard in the pharmacology literature (see, for instance, Chapter 4 of [32]). Our dose-response model (1.1) was derived from the well-known Michaelis-Menten formula (see page 52 in Chapter 6 of [143], Chapter 9 of [121], pages 144-145 of [134], and [100]) as described in the next paragraph.

We first assumed that the nominal relative change in DAS28 scores is given by a variation

of the standard Michaelis-Menten formula as

$$\frac{\text{DAS28}_t - \text{DAS28}_{t+1}}{\text{DAS28}_t} = \frac{\kappa_1 + d_t}{\kappa_1 + \kappa_2 + d_t}. \tag{1.2}$$

We believe that this formula is appropriate for modeling DAS28 evolution owing to its following desirable properties. The parameter κ_1 accounts for the so-called placebo effect (see, for instance, Chapter 9 of [121]), which models the change in DAS28_t when the biologic dose is zero. This change could, for example, be induced by a standard dose of methotrexate during combination therapy as in many clinical studies [174]. Setting $\kappa_1 = 0$ yields the standard Michaelis-Menten formula. Parameter κ_2 is interpreted as the biologic dose at which the relative change is $1/2$ when there is no placebo effect ($\kappa_1 = 0$). Note that the relative change in DAS28 is guaranteed to be a nonnegative fraction for any nonnegative biologic dose. The relative change asymptotically approaches one as the biologic dose approaches infinity. The relative change is positive, that is, DAS28_{t+1} is strictly smaller than DAS28_t for any biologic dose. Moreover, higher biologic doses induce higher relative change. This Michaelis-Menten model is a variation of the Emax model and of the Hill's equation, and these types of functions are commonly used to model dose-response in a variety of diseases and conditions (RA [112, 120, 124, 125], hepatitis C and AIDS [166, 182], hyperlipidemia [56], and hypertension [147]). After algebraic simplification, formula (1.2) yields $\text{DAS28}_{t+1}/\text{DAS28}_t = \kappa_2/(\kappa_1 + \kappa_2 + d_t)$. DAS28 dynamics (1.1) were then obtained after taking natural logarithms of both sides and then adding the stochastic noise term Θ. The reason for taking logarithms here was that it converts the original multiplicative model into an equivalent additive model that, as we shall see in Chapter 2, is analytically more convenient.

Our stochastic DP includes two "costs" that capture the fundamental trade-off in RGD. Higher doses potentially lead to better disease-control but at the same time may induce adverse effects and may be logistically and financially inconvenient for patients. Lower doses have the opposite effect. It is mainly for these reasons that the aforementioned clinical literature tracks the total dose administered over the treatment course in relation to the DAS28 score reached at the end of the treatment course. Thus, in our model, the first type

of cost is incurred in each treatment session and is given by cd_t in session t, where $c > 0$ is a constant seen as the coefficient of dose-aversion. This cost models the idea that the lower the total dose the better. The second cost is incurred at the end of the treatment course and equals DAS28_{T+1}. This is based on the maximalist school of thought, whereby the decision-maker attempts to reach as small a disease score as possible [10, 163, 175].

The decision maker's goal is to find an optimal dosing policy. That is, to derive a dose level in every possible DAS28 score in every session so as to minimize the total expected cost accumulated over the treatment course, given that the initial DAS28 score is DAS28_1. Let $J_t(\text{DAS28}_t)$ denote the minimum total expected cost accumulated by the end of the treatment course, given that the DAS28 score at the beginning of the tth session is DAS28_t. These optimal cost-to-go functions $J_t(\cdot)$ are unique solutions of Bellman's equations

$$J_t(\text{DAS28}_t) = \min_{d_t \in D} \Big\{ cd_t + E(J_{t+1}(\text{DAS28}_{t+1})) \Big\}, \ \forall \text{DAS28}_t \geq 0, \text{ and } t = 1, 2, \ldots, T, \quad (1.3)$$

with the boundary condition $J_{T+1}(\text{DAS28}_{T+1}) = \text{DAS28}_{T+1}$ for all $\text{DAS28}_{T+1} \geq 0$. Here, E denotes the expected value with respect to the random variable Θ. Doses that attain the above minima define an optimal policy. These equations can be solved approximately with backward recursion using discretization and state truncation if needed [25].

In the next section, we calibrate our model parameters using information available in the clinical literature, and perform numerical experiments and sensitivity analyses to gain insights into the behavior of the resulting dosing policy.

1.3 Calibration, numerical experiments, and sensitivity analyses

Our base-case parameter estimates below are based on information available in the OPTION study [174], which tracked the DAS28 scores of 622 patients on combination therapy with methotrexate and the biologic tocilizumab administered at four-week intervals over twenty four weeks. These patients were divided into three cohorts. The first cohort (size 204) did not receive tocilizumab and hence it is the placebo cohort. The second cohort (size 213) received a tocilizumab dose of 4mg/kg every four weeks and the third cohort (size 205)

received 8mg/kg every four weeks. The average initial DAS28 score in each cohort was 6.8.

1.3.1 Estimating base-case dose-response parameters κ_1, κ_2

We first estimated base-case values of parameters κ_1, κ_2 in our nominal dose-response model (1.2). Although the entire data set from the OPTION study is not available to us, we were able to surmise (by a visual inspection of Figure 2C in [174]) the numerical values of the mean DAS28 scores for the three cohorts over the twenty four week treatment course. These values are listed in Table 1.3 below. We used the resulting fifteen data points as

week	DAS28 (dose 0)	DAS28 (dose 4mg/kg)	DAS28 (dose 8mg/kg)
4	6.49	5.42	5.26
8	6.34	5.57	4.98
12	6.02	5.09	4.27
16	5.94	4.69	3.98
20	5.91	4.61	3.79
24	5.34	4.06	3.45

Table 1.3: Visually surmised DAS28 data from Figure 2C in [174].

input to the nonlinear regression subroutine nlinfit in MATLAB. This subroutine uses the Levenberg-Marquardt algorithm (see Chapter 10 of [140]) for least squares regression and the parameter estimates are sensitive to the initial guess provided by the user. We tried about fifty thousand combinations of initial guesses for the pair κ_1, κ_2 chosen from the grid $\{0, 0.01, 0.02, \ldots, 10\} \times \{0, 1, 2, \ldots, 500\}$ and selected the one with the smallest sum of squared errors. This led to the estimates $\kappa_1 = 4.5295, \kappa_2 = 124.1593$ as summarized in Table 1.4. Since our estimation method is crude, we will perform sensitivity analyses later in this section using different values of the biologic dose parameter κ_2 spread around this

base-case value while holding the placebo parameter κ_1 at its base-case value.

1.3.2 Estimating a base-case value of c

A base-case value of the cost coefficient c was derived as follows. We formulated a deterministic problem with $T = 7$ sessions (corresponding to twenty four weeks of therapy at four week intervals) and derived a value of c such that the dose $d = 6$ mg/kg would be optimal (since the OPTION clinical study used doses 4mg/kg and 8mg/kg, we chose the dose at the midpoint of these two values for our base-case calculations). That is, we found a value of c such that $d = 6$ mg/kg was optimal to the convex problem of minimizing $7cd + \text{DAS28}_1(\kappa_2/(\kappa_1 + \kappa_2 + d))^7$ with $\text{DAS28}_1 = 6.8$, $\kappa_1 = 4.5295$, and $\kappa_2 = 124.1593$. This objective function was obtained by calculating the total cost of 7 doses at level d each and of the terminal DAS28 reached after a 7-fold recursive application of the Michaelis-Menten formula starting with DAS28_1. Equating the derivative to zero and then substituting $d = 6$, we obtained $c = 0.028557$ as summarized in Table 1.4. We interpret this as inferring c from the one-size-fits-all approach. This derivation of c can be viewed as "inverse optimization" [4, 36, 55, 87]. Again, we will perform sensitivity analyses later in this section using different values of c spread around this base-case value.

1.3.3 Structure of optimal policy and optimal value function

Base-case values of our model parameters are listed in Table 1.4 below. Figure 1.1 illustrates our optimal dosing policy. Note that it is monotone — higher doses are administered in higher DAS28 scores. Moreover, doses are increasing over time; that is, for the same DAS28 score, a higher dose is administered in latter treatment sessions. Figure 1.2 illustrates the shape of our optimal value function. Note that it is convex and increasing in the natural logarithm of DAS28. It is also increasing over time; that is, for the same DAS28 score, the optimal value function takes a larger value in latter sessions. Figure 1.3 shows seven histograms of doses administered to a cohort of 1000 simulated patients in the seven treatment sessions (since

parameter	base-case value
κ_1	4.5295 (mg/kg)
κ_2	124.1593 (mg/kg)
c	0.028557 $(\text{mg/kg})^{-1}$
Θ	iid Normal$(0, 0.05^2)$
$\bar{d}$	10 (mg/kg)
DAS28_1	6.8

Table 1.4: Base-case values of our model parameters. In all our numerical experiments that used backward induction to solve Bellman's equations, Θ was truncated to the range $(-3\sigma, 3\sigma)$, where σ denotes the standard deviation; it was also discretized at intervals of 0.01σ. We note here that the Normal distribution is often used in the RA literature on disease activity scores [63, 90, 110, 171]. Doses were discretized using intervals of 0.01 mg/kg and states $\ln(\text{DAS28}_t)$ were discretized using intervals of 0.05. A linear interpolation of the corresponding discretized value function was used in our backward recursion procedure.

we discretized our state-space, linear interpolation was used to obtain dose levels at states that were off-grid). The first histogram shows that the initial doses to all simulated patients are identical because their DAS28 scores were assumed to be identical and equal to 6.8. This initial dose is in fact can be read-off from the y-axis in Figure 1.1 as the dose corresponding to a DAS28 of 6.8 on the x-axis in week 0. As the treatment progresses, the DAS28 scores of different simulated patients evolve stochastically as per the Michaelis-Menten formula, thus creating a "spread" of DAS28 scores. Patients with higher DAS28 scores receive higher doses and patients with lower DAS28 scores receive lower doses, as prescribed by our optimal policy in Figure 1.1. This leads to a spread of doses as depicted in the histograms in Figure 1.3. These histograms quantitatively illustrate the potential benefit of RGD — to avoid over- or under-dosing patients by administering the right dose to the right patient at the right time.

Similarly, Figure 1.4 shows histograms of the total doses administered over the seven sessions and the terminal DAS28 scores reached for this cohort of 1000 simulated patients.

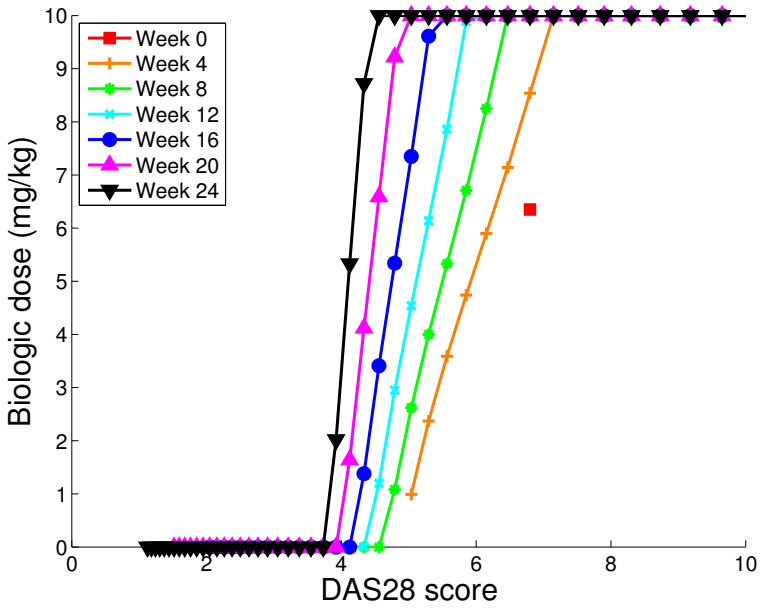


Figure 1.1: Optimal policy for the seven sessions in the treatment course; increasing in the natural logarithm of DAS28 scores (the reason for using natural logarithm on the X-axis as the state will become clear in Chapter 2). All parameters were fixed at their base-case values as in Table 1.4.

1.3.4 Sensitivity to problem parameters

In our model, biologic efficacy (or, in other words, response profile of a patient) is characterized by the value of κ_2. Higher values of κ_2 imply that a larger dose is needed to produce the same effect; lower values indicate that a smaller dose is needed to induce the same effect. Thus, intuitively, when other model parameters are fixed, the dose prescribed by the

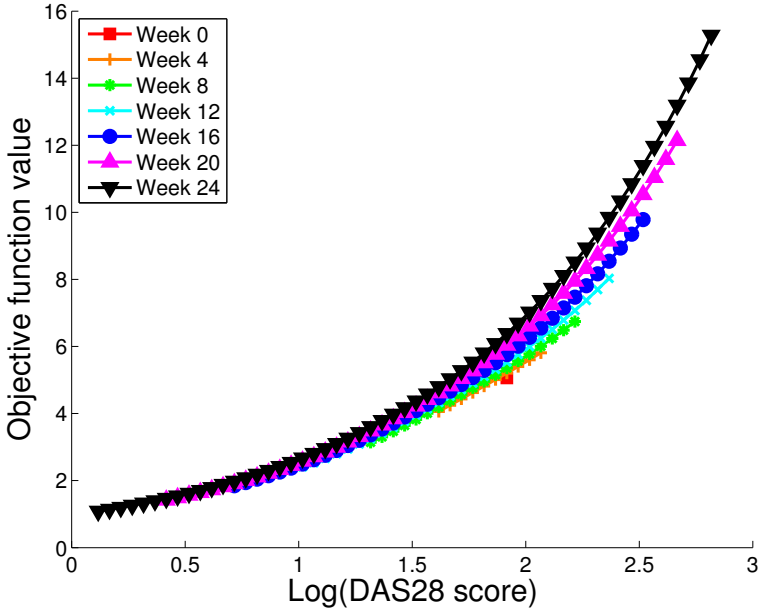


Figure 1.2: Optimal value function for the seven sessions in the treatment course; increasing and convex in the natural logarithm of DAS28 scores. All parameters were fixed at their base-case values as in Table 1.4.

optimal policy could vary non-monotonically with increasing κ_2. Figure 1.5 shows that for our specific combination of base-case parameter values, the optimal doses are decreasing in increasing κ_2 values.

We also study the sensitivity of optimal doses to κ_1. Larger values of κ_1 mean that the placebo (methotrexate in our case) is more effective. Thus, intuitively, a smaller biologic dose should be sufficient to achieve the same outcome. This intuition is quantified in Figure 1.6, which illustrates that in each treatment session, the optimal biologic dose in a fixed state is smaller for larger values of κ_1.

In our model, c characterizes the decision-maker's aversion to dose. The larger the value of c, the larger the aversion. Thus, intuitively, when other model parameters are fixed, the dose prescribed by the optimal policy should be decreasing in c. Figure 1.7 quantifies this

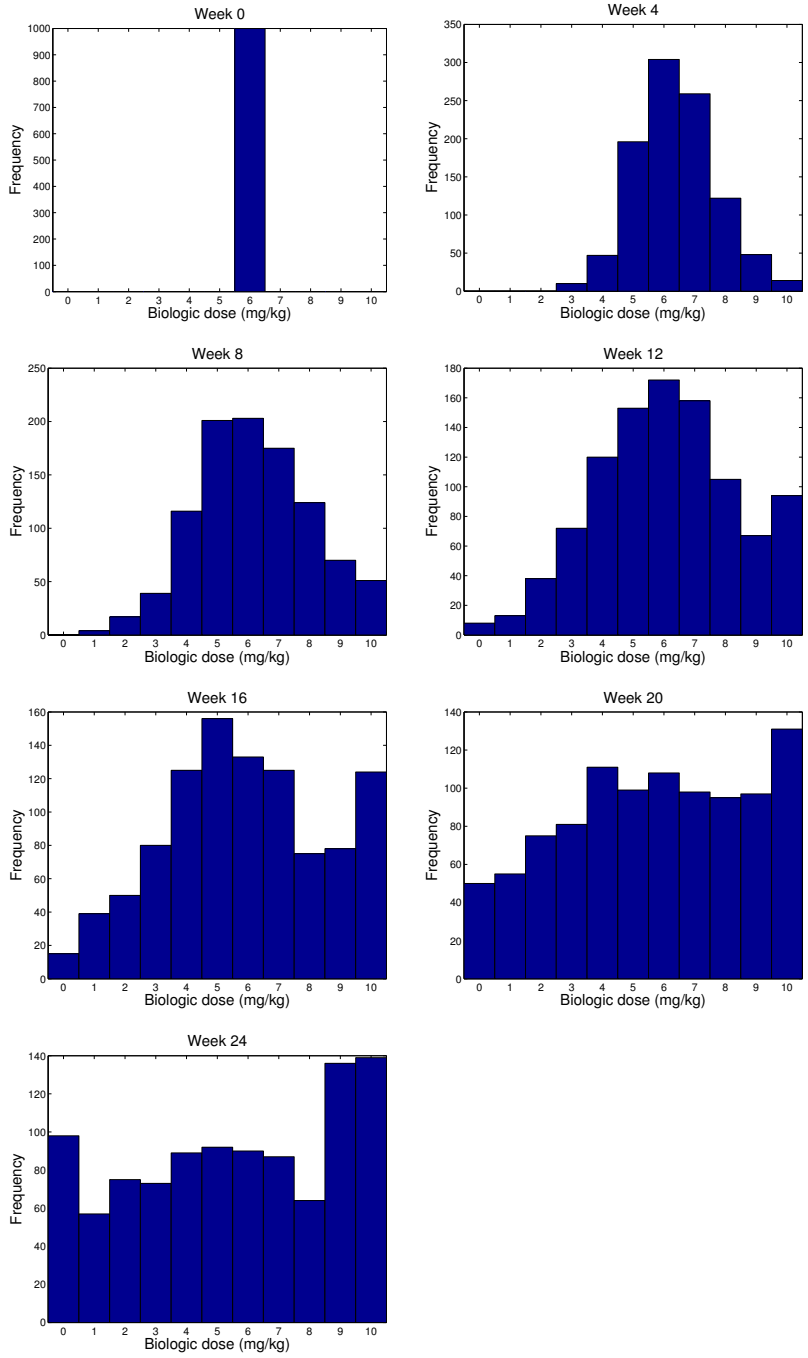
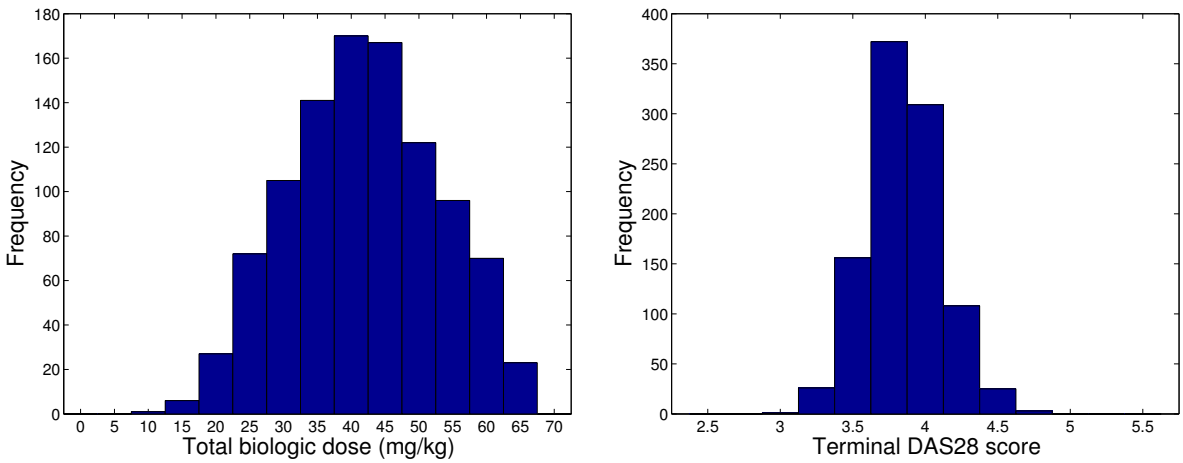


Figure 1.3: Histograms of doses administered to a cohort of 1000 (simulated) patients. All parameters were fixed at their base-case values as in Table 1.4.



(a) (b)

Figure 1.4: All parameters were fixed at their base-case values as in Table 1.4. (a) Histogram of total dose administered, over the seven-session treatment course, to a cohort of 1000 (simulated) patients. (b) Histogram of DAS28 score reached at the end of the seven-session treatment course for a cohort of 1000 (simulated) patients.

intuition.

The value of the dose-aversion coefficient c is somewhat subjective. This is similar to virtually all other decision-making problems that attempt to balance two competing objectives. For example, in the celebrated Markowitz portfolio optimization problem in finance, the value of the risk-aversion coefficient is subjective [128]. Similarly, in cancer radiotherapy optimization, the effect of radiation dose on the tumor is balanced against the toxic effect of dose on nearby organs-at-risk using weighting coefficients [53, 167]. In Figure 1.8, we therefore present a so-called efficient frontier. It makes explicit the trade-off between total dose administered and the DAS28 score reached at the end of the treatment courses. An efficient frontier could serve as a decision-tool in RGD for RA. A doctor could, for instance, choose a (total dose, DAS28 score) point on this frontier that he/she is comfortable with and

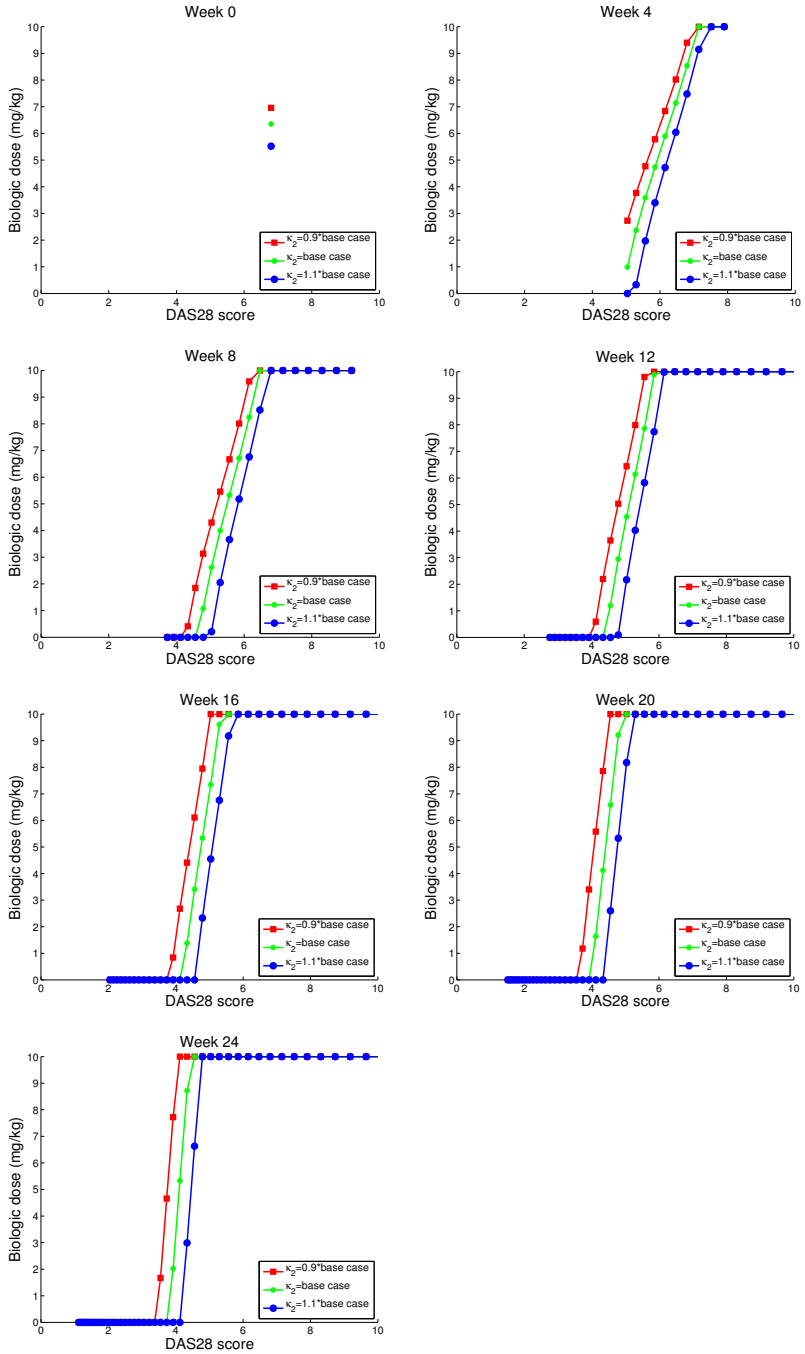


Figure 1.5: Sensitivity of optimal policy to κ_2. All other parameters were fixed at their base-case values as in Table 1.4.

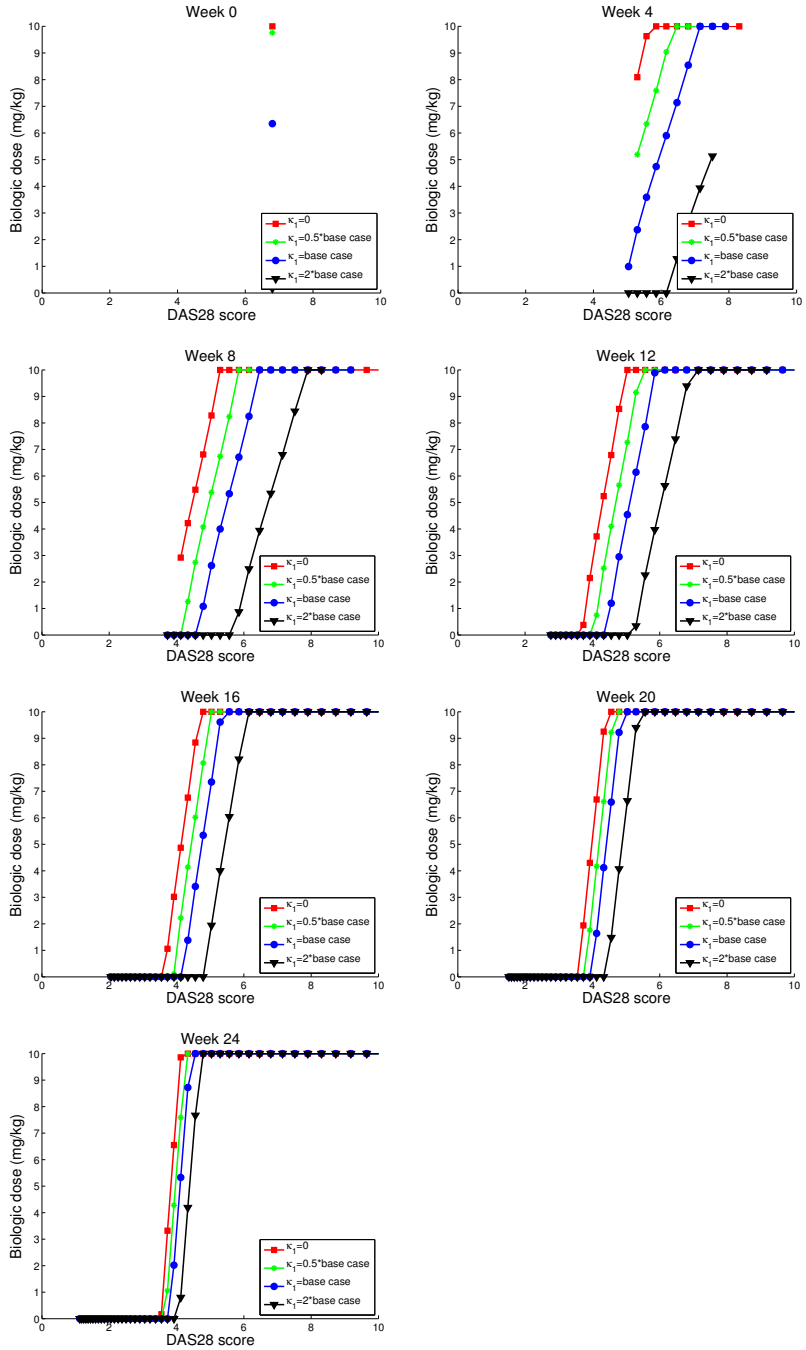


Figure 1.6: Sensitivity of optimal policy to κ_1. All other parameters were fixed at their base-case values as in Table 1.4.

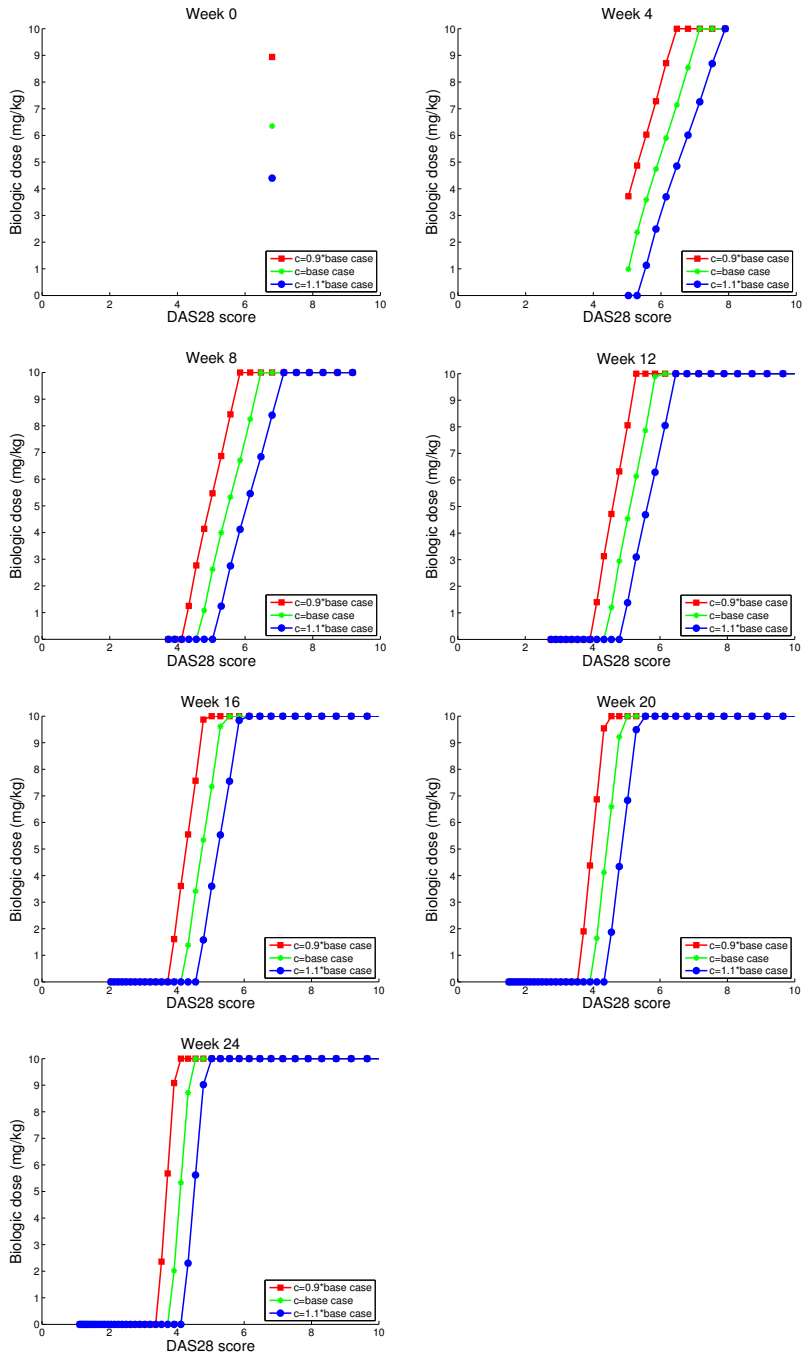


Figure 1.7: Sensitivity of optimal policy to c. All other parameters were fixed at their base-case values as in Table 1.4.

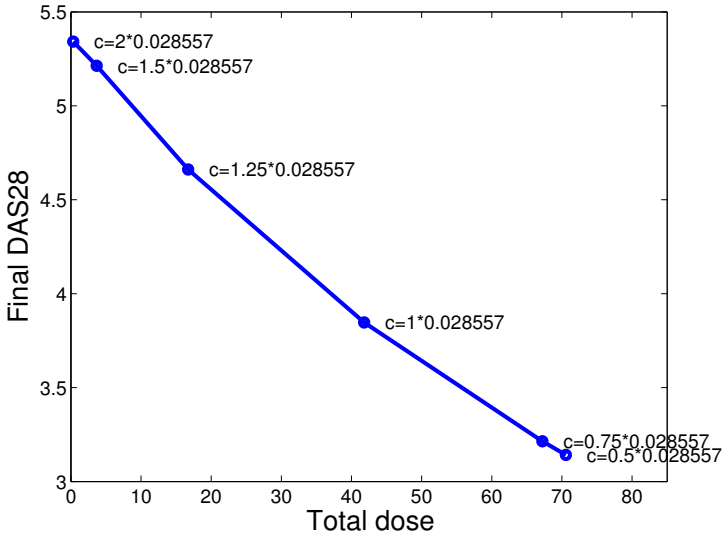


Figure 1.8: Efficient frontier defined by several values of c. All other parameters were fixed at their base-case values as in Table 1.4. The frontier was obtained by fitting a spline in MATLAB through the six data points shown in the figure. The six data points were obtained by averaging the total dose administered and the terminal DAS28 score reached on implementing our dosing policy over 1000 independent simulations.

use the implied value of c to derive an optimal RGD policy for his/her patients.

We also consider sensitivity to σ, the standard deviation of the normally distributed random variable appearing in the state dynamics. Because we do not have data from the existing literature to estimate σ, our base-case value of 0.05 is arbitrary; we thus vary it to understand its effect on optimal dose. This is illustrated in Figure 1.9. We note that for intermediate values of optimal dose, where the optimal dose is neither zero nor $\bar{d}$, higher σ curves have higher slope. This suggests a more aggressive treatment strategy that administers a larger incremental change in optimal dose between two given disease states when the disease progression is more uncertain.

Finally, we investigate the sensitivity to κ_1 of the improvement in objective value offered

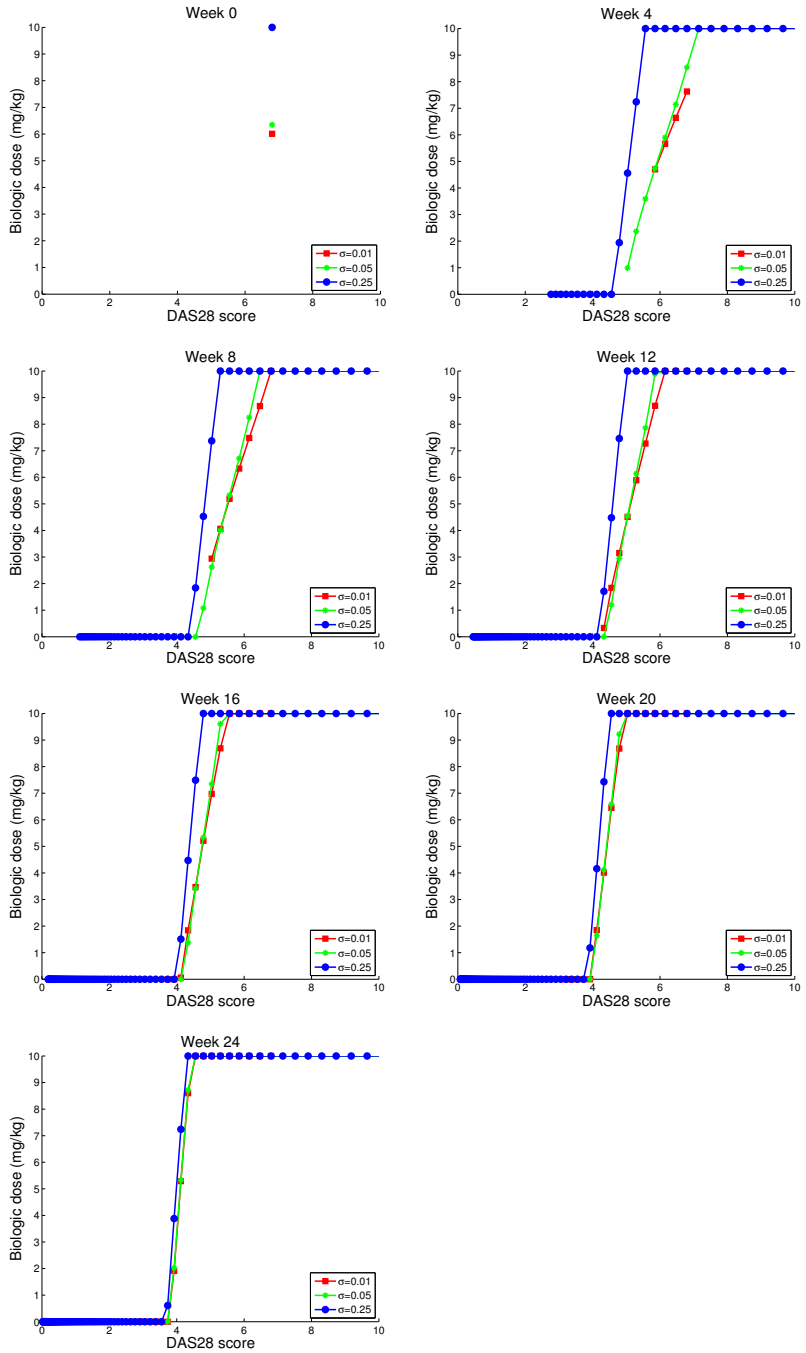


Figure 1.9: Sensitivity analysis using several values of σ. All other parameters were fixed at their base-case values as in Table 1.4.

by our optimal policy as compared to three other dosing policies in Figure 1.10. The first policy does not use a biologic agent for treatment and thus only administers methotrexate (this is the traditional therapy for RA) — we call this the 0 dose policy. The other two treatment strategies are from the OPTION trial — one administers a constant biologic dose of 4mg/kg and the other administers a constant biologic dose of 8mg/kg. The figure shows that our optimal policy outperforms these three policies for all values of κ_1. For small values of κ_1, where methotrexate is least effective, the 0 dose policy performs the worst. This is consistent with the statement in Section 1.1 that biologic treatment, and more strongly, optimal biological treatment, is potentially most beneficial for patients who respond poorly to methotrexate. The 4mg/kg policy performs better than the 0 dose policy. The objective value of the 8mg/kg policy is closest to that of the optimal policy. Recall, however, that even though the objective values of these two policies are similar, the doses they administer to individual patients would be different. The constant dose policy administers a dose of 8mg/kg to all patients (thus possibly under- and over-dosing patients), whereas our optimal policy administers smaller doses to patients who respond well and higher doses to patients who respond poorly. For large values of κ_1, where methotrexate is most effective, the objective value of the 0 dose policy is close to our optimal policy. This is to be expected because when methotrexate is very effective, the optimal policy should also administer small or almost no dose of the biologic. The 4mg/kg policy performs worse than the 0 dose policy. The 8mg/kg policy performs even worse. The reasoning behind these observation also applies to the entire continuum of intermediate κ_1 values.

Since the value of our objective function (Y-axis in Figure 1.10) may be difficult to interpret clinically, we break it into its two clinically meaningful components: the total dose delivered and the terminal DAS28 score reached, in Figure 1.11. In other words, roughly speaking, the Y-axis in Figure 1.10 equals the Y-axis in Figure 1.11(b) plus the coefficient of risk aversion times the Y-axis in Figure 1.11(a). We do wish to remind the readers, however, that any two policies cannot really be compared by looking solely at the Y-axis values either

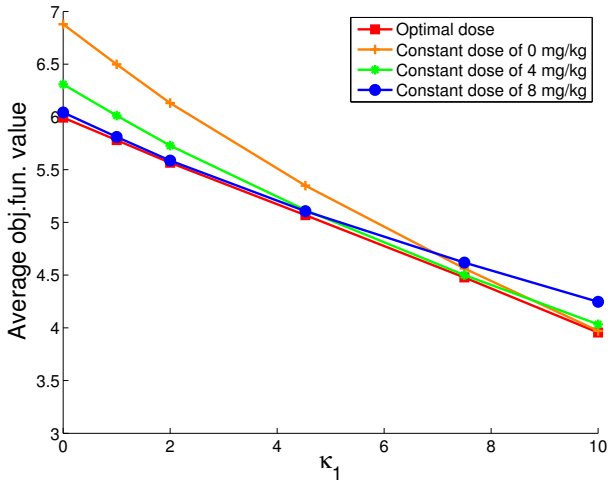
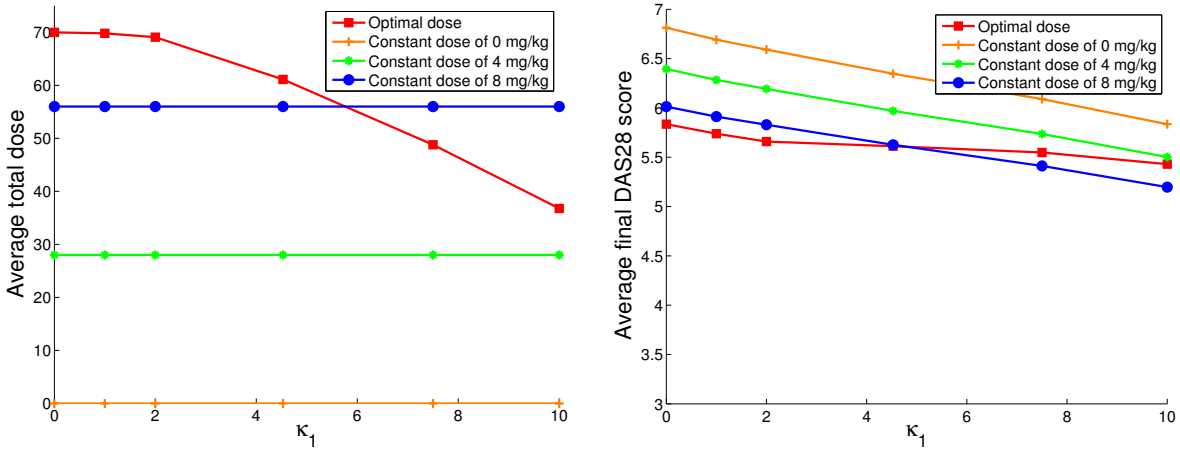


Figure 1.10: Comparison of three constant-dose policies with our optimal policy, as a function of κ_1.



(a) (b)

Figure 1.11: Comparison of the two components of our objective function for the three constant-dose policies from Figure 1.10 and our optimal policy, as a function of κ_1.

in Figure 1.11(a) or in Figure 1.11(b), because our stochastic DP does not optimize these individual components. Nevertheless, these two figures do make explicit the fundamental

trade-off in dosing. Figure 1.11(a) shows (obviously) that the total dose delivered by the three constant dose policies does not change with κ_1 (methotrexate effectiveness), whereas the total dose for the optimal policy does depend on κ_1. The optimal policy administers a higher total dose than these three policies for lower values of κ_1 (where methotrexate is not very effective) and attains a lower terminal DAS28 score. For higher values of κ_1, where methotrexate is effective, the optimal policy administers a lower dose than the largest constant dose policy, sacrificing the terminal DAS28 score a little bit. The optimal policy administers a higher dose than the two smaller constant dose policies and attains a lower terminal DAS28 score, at all values of κ_1.

In Chapter 2, we generalize the model of this chapter and prove the monotonicity result we observed numerically in Figure 1.1.

Chapter 2

A GENERAL STOCHASTIC DP FORMULATION FOR RESPONSE-GUIDED DOSING

The key ideas in our stochastic DP approach for biologic RGD in RA were demonstrated through a detailed, concrete example in Sections 1.2 and 1.3. In this chapter, we extend this example by allowing for general disease scores, dose-response dynamics, and cost functions. Our motivation for this generalization is two-fold as described next.

2.1 Motivation for a general formulation

Firstly, our generalization below broadens the applicability of the DP approach to other RA disease activity scores that may become available in the future (see [66] for a discussion of disease activity scores in RA), to other dose-response functions that a decision-maker may choose to fit to his/her data, and to other cost functions such as those consistent with the treat-to-target school of thought (where the goal is to bring the disease score below a remission threshold [145]) as opposed to the maximalist one. Perhaps more importantly, this generalization further broadens the applicability of our mathematical framework to other diseases and conditions where RGD is potentially helpful. As outlined next, examples include hepatitis C, LDL cholesterol lowering statin therapy, and AIDS.

A common treatment for chronic hepatitis C virus infections is a once-weekly dose of 180 μg pegylated interferon in combination with a daily dose of 1000 - 1200 mg ribavirin for 48 weeks [137, 200]. Recent clinical studies have used viral RNA measurements from an initial phase of therapy to adjust subsequent dose levels [45, 130, 200].

Guidelines for coronary heart disease (CHD) recommend low density lipoprotein (LDL) cholesterol levels of $<$100 mg/dL for patients with CHD, $<$130 mg/dL for patients with two

or more risk factors, and <160 mg/dL for patients with fewer than two risk factors [57]. One approach to CHD risk management is to monitor LDL levels and adjust LDL-reducing statin therapy to meet these targets [57] and there is some debate about whether or not this strategy is optimal [75, 76, 103].

The SMART trial for AIDS [70] studied a CD4 threshold based on-and-off dosing method where treatment was started (or re-started) when the CD4 count fell below 250 cells/mm^3 and it was continued only as long as this count stayed below 350. Other trials have also studied similar CD4-guided strategies [11, 12, 34, 44, 116, 80, 122].

We believe that our general stochastic DP approach here could guide RGD for such diseases and conditions using disease scores such as the viral load, LDL cholesterol levels, and CD4 counts.

2.2 Bellman's equations for the general formulation

Let T denote the number of treatment sessions, indexed by $t = 1, 2, \ldots, T$, in a treatment course. The time-interval between two consecutive treatment sessions may be in hours, days, weeks, or months depending on the disease. For simplicity of notation, we assume that these intervals are equal. At the beginning of each treatment session, the physician observes a numerical score of the patient's disease condition, and chooses a dose for that session. These numerical scores belong to a convex set $X \subseteq \mathbb{R}$. Smaller real numbers in this set represent less severe disease. The disease condition at the beginning of treatment session t is denoted by $x_t \in X$. The dose level chosen by the physician for this session after observing x_t is denoted by d_t. Possible dose levels d_t belong to the interval $D \triangleq [0, \bar{d}] \subset \mathbb{R}$, where $\bar{d}$ is a finite upper bound on permissible dose levels.

For $t = 1, 2, \ldots, T$, disease conditions evolve according to dynamics

$$x_{t+1} = f(x_t, d_t; \Theta), \text{ for } x_t, x_{t+1} \in X, \text{ and } d_t \in D, \tag{2.1}$$

where Θ are iid random variables that take values from a set $\Omega \subseteq \mathbb{R}$. These random variables are assumed to possess a probability density function $p(\cdot)$. Our results also hold when random

variables Θ are discrete, and in that case, $p(\cdot)$ denotes the probability mass function and all integrals below are replaced by sums. We assume that the response function $f(\cdot, \cdot; \theta)$ is continuous over $X \times D$ for each $\theta \in \Omega$. As in our RA example in Section 1.2, dose-response dynamics of the form (2.1) can be derived by starting with a nominal dose-response model such as exponential, exponential linear-quadratic, logistic, Michaelis-Menten, Hill's, Emax, power law, Gompertz, and beta-Poisson (see [172]). These nominal dose-response functions have been used in the medical literature for a variety of diseases and conditions including RA, hepatitis C, hyperlipidemia, AIDS, diabetes, and hypertension [46, 94, 100, 121, 124, 134, 147, 166, 182].

Aversion to dose is modeled using a continuous cost function $c : D \to \mathbb{R}_+$. Since D is compact, continuity of $c(\cdot)$ implies that it is bounded. Examples include linear, quadratic, and exponential functions. Aversion to disease conditions x_{T+1} at the end of the treatment course is modeled using a continuous and bounded cost function $h : X \to \mathbb{R}_+$. Examples include linear, quadratic, exponential, and ramp. The cost in the ramp function is zero up to a threshold and then increases with disease score; this can be used to model the treat-to-target approach.

Our concrete model in Section 1.2 is a special case of this general model, where $x_t = \ln(\text{DAS28}_t)$, $f(x_t, d_t; \Theta) = x_t + \ln \kappa_2 - \ln(\kappa_1 + \kappa_2 + d_t) + \Theta$, $c(d_t) = cd_t$, and $h(x_t) = \exp(x_t)$. Let $J_t(x_t)$ denote the minimum total expected disutility accumulated by the end of the treatment course, given that the disease condition at the beginning of the tth session is x_t. These optimal cost-to-go functions $J_t(\cdot)$ are unique solutions of Bellman's equations

$$J_t(x_t) = \min_{d_t \in D} \Big\{ c(d_t) + \int_{\Omega} J_{t+1}(f(x_t, d_t; \theta_t)) p(\theta) d\theta \Big\}, \ \forall x_t \in X, \text{ and } t = 1, 2, \ldots, T, \qquad (2.2)$$

with the boundary condition $J_{T+1}(x) = h(x)$ for all $x \in X$. As we shall see, problem (2.2) involves optimizing a continuous function over the nonempty compact set D and hence it has an optimal solution. Doses that attain the above minima define an optimal response-guided dosing policy. The set of doses that attain the minimum in (2.2) for state $x_t \in X$ in session t is denoted by $A_t^*(x_t) \subseteq D$. Bellman's equations (2.2) can be approximately solved easily

using discretization of X (along with truncation if needed) and of D.

2.2.1 Relation to other dynamic optimization models in treatment planning

Recent surveys of dynamic optimization models in medical treatment planning are available in [6, 158]. The idea of adapting treatment or diagnostic decisions to the observed stochastic evolution of a "health state" was recently employed in [7, 8, 40, 48, 107, 156, 165]. A common feature of these papers is that their decisions are of the wait/do not wait-type, and do not involve a sequential choice of dose levels. The stochastic evolution of the health state occurs according to a natural history progression model without intervention, and the decision process ends when a choice to not wait is made. A control-limit policy, where one waits if and only if the health state is better than a certain threshold, is typically found to be optimal. In particular, Alagoz et al. [7, 8] and Sandikci et al. [156] model reject/accept decisions of end-stage liver disease patients on a waiting list, who are offered livers from living or cadaveric donors. Shechter et al. [165] investigate the optimal time to initiate antiretroviral therapy for AIDS and hence adapt wait/initiate decisions to observed CD4 counts. Denton et al. [48] and Kurt et al. [107] study the optimal time to begin statin treatment for diabetes patients and hence also make wait/initiate decisions. Chhatwal et al. [40] consider a radiologist's choice of whether to perform a biopsy or to wait until the next annual mammogram, based on a risk score derived from the latest annual mammogram.

The work in [97] is an exception to this common theme of wait/do not wait decisions. There, the authors made dose level (high/low) decisions for response-guided radiotherapy and numerically found that the optimal policy had a control-limit structure in some cases. This idea was then extended to intensity modulated radiation therapy, where radiation intensity profiles were adjusted based on the spatiotemporal evolution of tumor cell density [98]. Another exception to the wait/do not wait theme are papers where a sequence of treatment modalities is chosen for metastatic tumors [18, 19] and AIDS [164]. Continuous time deterministic control methods, which utilize differential equations to model viral dynamics, have been used to derive dosing strategies for viral infections such as AIDS [109, 197].

In a different line of work, researchers have proposed stochastic compartment models to make dosing decisions for diseases such as diabetes. These models discretize the human body into a finite number of chambers and use stochastic differential and algebraic equations to model the transport of a drug through these chambers in continuous time. The state in these models often includes drug concentration in each compartment and controls relate to dose levels. Bayesian statistics is employed to learn patient-specific parameter distributions whereas drug concentration measurements are subject to stochastic errors. Continuous time stochastic control techniques are then applied to maintain drug concentrations in blood plasma inside these compartments within a preset range. The resulting problems are computationally intractable even in the single-compartment case, and hence approximate control schemes need to be devised and numerically compared. To the best of our knowledge, structural results about optimal policies are not available. Examples of this approach include Hu et al. [81], Bayard et al. [17], Jelliffe et al. [89], Acikgoz and Diwekar [2], Schumitzky [162], and references therein. However, these models do not consider the stochastic progression of disease condition, and hence do not implement RGD as envisioned in this chapter. Finally, papers mentioned in this and the previous paragraph only include numerical experiments and do not seek to prove the structure of their treatment strategies.

Our dosing policy in Section 1.3 had an intuitive, monotone structure that can in general be exploited to speed up backward induction calculations [151]. Thus, the question arises as to whether or not this structure holds in our general model. We investigate this question in the next section.

2.3 Monotonicity of optimal dosing policy

We claim, under two assumptions on the disease condition dynamics and on the cost functions, that in each treatment session there exist optimal doses that increase as the disease condition worsens.

Assumption 2.3.1 (Monotone, supermodular, and convex disease-response)**.** *For each re-*

alization $\theta \in \Omega$ of the random variable Θ,

- *the function $f(x, \cdot; \theta)$ is decreasing in dose for each fixed $x \in X$, and the function $f(\cdot, d; \theta)$ is increasing in disease condition for each fixed $d \in D$;*

- *the function $f(\cdot, \cdot; \theta)$ is supermodular over $X \times D$; that is, $f(u, a; \theta) - f(u, b; \theta) \geq f(x, a; \theta) - f(x, b; \theta)$ for all $x, u \in X$ such that $u \geq x$ and for all $a, b \in D$ such that $a \geq b$;*

- *the function $f(\cdot, \cdot; \theta)$ is convex over $X \times D$; that is, for any $x, y \in X$ and any $a, b \in D$, it satisfies the inequality $f(\lambda x + (1-\lambda)y, \lambda a + (1-\lambda)b; \theta) \geq \lambda f(x, a; \theta) + (1-\lambda) f(y, b; \theta)$ for every $\lambda \in [0, 1]$.*

Assumption 2.3.2 (Increasing, and convex costs)**.** *The cost function $c(\cdot)$ is increasing and convex over D; cost function $h(\cdot)$ is increasing and convex over X.*

The first item in Assumption 2.3.1 is natural and expresses that (i) for a fixed pre-session disease condition, a higher dose results in a better post-session disease condition, and (ii) given a fixed dose, the post-session disease condition is increasing in the pre-session disease condition.

Supermodularity as in the second item holds when the marginal difference in disease response for high and low doses is higher in worse disease conditions. In fact, in most special cases of (2.1), the response function will be additively separable in dose and disease condition (see examples below), and hence it will be both supermodular and submodular [186]. As a result, we expect that supermodularity of dose response will hold trivially in essentially all special cases of (2.1).

Again, in most special cases of (2.1), the response function will be additively separable in dose and disease condition (see examples below), and hence (joint) convexity as in the third item of Assumption 2.3.1 will be equivalent to componentwise convexity. This means that (i) the magnitude of marginal improvement in disease condition decreases with higher

doses, and (ii) for a fixed dose, the marginal improvement in disease condition increases with worsening disease conditions.

Linear dynamics is perhaps the simplest function that satisfies Assumption 2.3.1. Here, we provide several other examples of dose-response functions that are available in the medical literature and that also satisfy Assumption 2.3.1.

1. **Exponential** (see [172]): Consider a disease with nonnegative scores y_t that are nominally assumed to evolve according to $y_{t+1} = y_t \exp(-\alpha d_t)$ for some parameter $\alpha > 0$. As in the Michaelis-Menten formula discussed in Section 1.2, this function has the desirable properties that $y_{t+1} = y_t$ if $d_t = 0$ and y_{t+1} asymptotically approaches zero as $d_t \to \infty$. Taking natural logarithms on both sides, defining $x_t = \ln(y_t)$ and then adding the stochastic noise term, this nominal formula yields the linear dynamics $x_{t+1} = x_t - \alpha d_t + \Theta$. It is easy to see that $f(x_t, d_t; \Theta) = x_t - \alpha d_t + \Theta$ satisfies Assumption 2.3.1. The exponential dose-response function has been used to model evolution of LDL cholesterol levels under phytosterol treatment [46].

2. **Power law** (see [129]): Again consider a disease with nonnegative scores y_t that are nominally assumed to evolve according to $y_{t+1} = y_t(1 + d_t)^{-\alpha}$ where $\alpha > 0$ is a parameter. This function also has the aforementioned desirable properties and after taking logarithms and adding stochastic noise as above yields $x_{t+1} = x_t - \alpha \ln(1 + d_t) + \Theta$ where Assumption 2.3.1 holds. The power law function has been used to model the effect of chemotherapy on tumor cells [73].

3. **Beta-Poisson** (see [184]): This is a generalization of the above power law function such that $y_{t+1} = y_t(1 + d_t/\beta)^{-\alpha}$ where $\beta > 0$ is an additional parameter. Again, taking logarithms and adding stochastic noise yields dynamics where Assumption 2.3.1 holds. This function has been used in quantitative risk assessment of exposure to pathogenic microorganisms.

4. **Michaelis-Menten**: This function was discussed in detail earlier in this chapter.

Assumption 2.3.2 holds when (i) there is an increasing aversion to higher doses and the marginal aversion increases with dose, and (ii) there is an increasing aversion to worsening disease conditions and the marginal aversion increases with worsening disease conditions. The decision-maker's preference to lower doses and better disease conditions is a natural assumption; increasing marginal aversion is a characteristic of risk-averse decision-makers in medicine [52].

We are now ready to state our monotonicity result.

Theorem 2.3.3. *Under the above assumptions, optimal dose levels increase with worsening disease conditions in each treatment session. More precisely, in any session t, if dose level $a(x) \in A_t^*(x)$ for some $x \in X$, then, corresponding to every $u \in X$ with $u \geq x$, there exists a dose level $b(u) \in A_t^*(u)$ such that $b(u) \geq a(x)$.*

2.3.1 Proof of Theorem 2.3.3

The proof employs backward induction on $t = T+1, T, \ldots, 1$. In particular, we first show that if $J_{t+1}(\cdot)$ is increasing, convex, continuous, and bounded, then in treatment session t, there exist optimal decisions with the monotone structure described in Theorem 2.3.3. We then show that $J_t(\cdot)$ inherits these four properties from $J_{t+1}(\cdot)$. Theorem 2.3.3 then follows because the utility function $h(\cdot)$, that is, $J_{T+1}(\cdot)$, is assumed to possess these properties.

To begin the proof, we define the Q_t-function of DP for $t = 1, 2, \ldots, T$ as

$$Q_t(x_t; d_t) \triangleq c(d_t) + \int_{\Omega} J_{t+1}(f(x_t, d_t; \theta))p(\theta)d\theta, \ \forall x_t \in X, \ d_t \in D, \tag{2.3}$$

and note from (2.2) that

$$J_t(x_t) = \min_{d_t \in D} \{Q_t(x_t; d_t)\}. \tag{2.4}$$

We first show that $Q_t(\cdot;\cdot)$ has decreasing differences over $X \times D$. That is,

$$\Big[Q_t(u; a) - Q_t(u; b)\Big] - \Big[Q_t(x; a) - Q_t(x; b)\Big] \leq 0, \tag{2.5}$$

for all $x, u \in X$ such that $u \geq x$, and all $a, b \in D$ such that $a \geq b$. To see that (2.5) holds, observe from the definition of the Q_t-function in (2.3) that the left hand side in (2.5) equals

$$\int_{\Omega} \Big[J_{t+1}(f(u,a;\theta)) - J_{t+1}(f(u,b;\theta)) - J_{t+1}(f(x,a;\theta)) + J_{t+1}(f(x,b;\theta)) \Big] p(\theta) d\theta. \tag{2.6}$$

To prove that this expectation is nonpositive, it suffices to show that the expression inside [ ] in the integrand is nonpositive for every realization $\theta \in \Omega$ of the random variable Θ. By Assumption 2.3.1, we have,

$$f(u,a;\theta) \leq f(u,b;\theta), \tag{2.7}$$

$$f(x,a;\theta) \leq f(x,b;\theta), \tag{2.8}$$

$$f(x,a;\theta) \leq f(u,a;\theta), \text{ and} \tag{2.9}$$

$$f(x,b;\theta) \leq f(u,b;\theta). \tag{2.10}$$

As shown in Figure 2.1, there are exactly two possibilities for the relative positions of $f(u,a;\theta)$, $f(u,b;\theta)$, $f(x,a;\theta)$, and $f(x,b;\theta)$ that are consistent with inequalities (2.7)-(2.10). In both cases, by supermodularity of disease response in Assumption 2.3.1, we have,

$$f(u,a;\theta) - f(u,b;\theta) \geq f(x,a;\theta) - f(x,b;\theta). \tag{2.11}$$

Hence,

$$J_{t+1}(f(u,a;\theta)) - J_{t+1}(f(u,b;\theta)) \leq J_{t+1}(f(x,a;\theta)) - J_{t+1}(f(x,b;\theta)) \tag{2.12}$$

because $J_{t+1}(\cdot)$ is increasing and convex (see Figure 2.1 again). This implies that (2.6) is nonpositive and (2.5) holds.

Inequality (2.5) implies that in session t, there exists an optimal policy that is increasing in disease conditions as claimed in Theorem 2.3.3. We show this by contradiction. So suppose not. That is, there are two states, which we denote x and u where $u \geq x$, with the following property in session t: dose $a \in D$ is optimal in x and every dose $b \in D$ that is optimal in u satisfies $b < a$. Then $Q_t(x;a) \leq Q_t(x;b)$ by optimality of a in x, and $Q_t(u;b) < Q_t(u;a)$

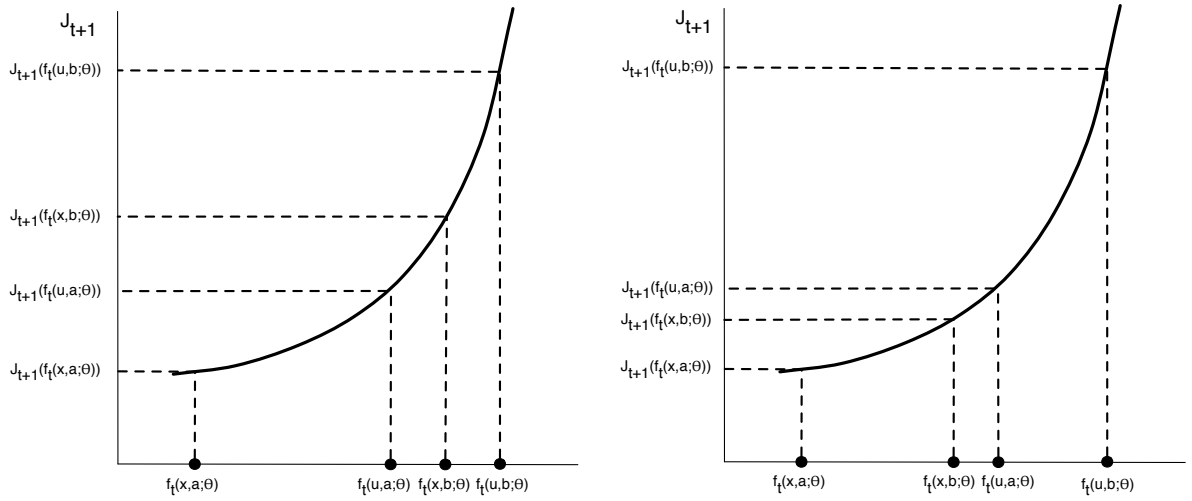


Figure 2.1: A visual aid to the proof that $Q_t(\cdot,\cdot)$ has decreasing differences. This figure shows the two possibilities for the relative positions of $f(u,a;\theta)$, $f(u,b;\theta)$, $f(x,a;\theta)$, and $f(x,b;\theta)$ that are consistent with inequalities (2.7)-(2.10). Corresponding values of the convex, increasing function $J_{t+1}(\cdot)$ satisfy inequality (2.12).

because b is optimal and a is not optimal in u. Adding these inequalities yields a contradiction to the decreasing differences property defined in (2.5).

We now show that $J_t(\cdot)$ is increasing in disease condition; that is, $J_t(u) \geq J_t(x)$ for all $x, u \in X$ such that $u \geq x$. Suppose dose $a \in D$ is optimal in x and dose $b \in D$ is optimal in u in session t. Then,

$$\begin{aligned} J_t(u) - J_t(x) &= Q_t(u;b) - Q_t(x;a) \geq Q_t(u;b) - Q_t(x;b) \\ &= \int_{\Omega} \Big[J_{t+1}(f(u,b;\theta)) - J_{t+1}(f(x,b;\theta)) \Big] p(\theta) d\theta \geq 0. \end{aligned}$$

Here, the first inequality follows because $a \in D$ is optimal in x and hence $Q_t(x;a) \leq Q_t(x;b)$. The second inequality follows because for every realization $\theta \in \Omega$ of the random variable Θ, $f(u,b;\theta) \geq f(x,b;\theta)$ by Assumption 2.3.1 and hence $J_{t+1}(f(u,b;\theta)) \geq J_{t+1}(f(x,b;\theta))$ as $J_{t+1}(\cdot)$ is increasing.

We show in addition that $Q_t(\cdot;\cdot)$ is convex over $X \times D$. Let $\lambda \in [0,1]$, $x, y \in X$, and $a, b \in D$. For brevity, we define the shorthand $z = \lambda x + (1-\lambda)y$, and $d = \lambda a + (1-\lambda)b$. We have,

$$\begin{aligned}
Q_t(z;d) &= c(d) + \int_\Omega \Big[J_{t+1}(f(z,d;\theta))\Big] p(\theta) d\theta \\
&\leq c(d) + \int_\Omega \Big[J_{t+1}(\lambda f(x,a;\theta) + (1-\lambda) f(y,b;\theta))\Big] p(\theta) d\theta \\
&\leq c(d) + \int_\Omega \Big[\lambda J_{t+1}(f(x,a;\theta)) + (1-\lambda) J_{t+1}(f(y,b;\theta))\Big] p(\theta) d\theta \\
&\leq \lambda c(a) + (1-\lambda) c(b) + \int_\Omega \Big[\lambda J_{t+1}(f(x,a;\theta)) + (1-\lambda) J_{t+1}(f(y,b;\theta))\Big] p(\theta) d\theta \\
&= \lambda Q_t(x;a) + (1-\lambda) Q_t(y;b).
\end{aligned}$$

Here, the first inequality holds because $f(\cdot,\cdot;\theta)$ is convex by Assumption 2.3.1 and $J_{t+1}(\cdot)$ is increasing. The second inequality holds because $J_{t+1}(\cdot)$ is convex. The third inequality results from convexity of $c(\cdot)$ by Assumption 2.3.2. This shows that $Q_t(\cdot;\cdot)$ is convex over $X \times D$.

Using a standard argument in convex optimization (see Section 3.2.5 of [29]), convexity of $Q_t(\cdot;\cdot)$ implies from (2.4) that $J_t(\cdot)$ is convex over X. To see this, let $\lambda \in [0,1]$ and $x, y \in X$. Suppose that $J_t(x) = Q_t(x;a)$ for some $a \in D$, that is, $a \in D$ is optimal in x, and that $J_t(y) = Q_t(y;b)$ for some $b \in D$, that is, $b \in D$ is optimal in y. We have,

$$\begin{aligned}
J_t(\lambda x + (1-\lambda)y) &= \min_{d \in D} \{Q_t(\lambda x + (1-\lambda)y; d)\} \\
&\leq Q_t(\lambda x + (1-\lambda)y; \lambda a + (1-\lambda)b) \\
&\leq \lambda Q_t(x;a) + (1-\lambda) Q_t(y;b) \\
&= \lambda J_t(x) + (1-\lambda) J_t(y),
\end{aligned}$$

where the first inequality holds because $\lambda a + (1-\lambda)b \in D$, and the second inequality follows from convexity of $Q_t(\cdot;\cdot)$ over $X \times D$.

We now prove that $Q_t(\cdot;\cdot)$ is continuous and bounded over $X \times D$. Fix any $(x;d) \in X \times D$, and consider a convergent sequence $(x^n;d^n) \in X \times D$ such that $\lim_{n\to\infty}(x^n;d^n) = (x;d)$. We have,

$$Q_t(x^n;d^n) = c(d^n) + \int_\Omega J_{t+1}(f(x^n,d^n;\theta))p(\theta)d\theta. \tag{2.13}$$

By continuity of $f(\cdot,\cdot;\theta)$ over $X \times D$, $\lim_{n\to\infty} f(x^n,d^n;\theta) = f(x,d;\theta)$ for each $\theta \in \Omega$. This implies, by continuity of $J_{t+1}(\cdot)$ over X, that

$$\lim_{n\to\infty} J_{t+1}(f(x^n,d^n;\theta)) = J_{t+1}(\lim_{n\to\infty} f(x^n,d^n;\theta)) = J_{t+1}(f(x,d;\theta)),$$

for each $\theta \in \Omega$. Similarly, by continuity of $c(\cdot)$ over D, we have, $\lim_{n\to\infty} c(d^n) = c(\lim_{n\to\infty} d^n) = c(d)$. We define a sequence of functions $r_n : \Omega \to \mathbb{R}$ by

$$r_n(\theta) = J_{t+1}(f(x^n,d^n;\theta)).$$

Since $J_{t+1}(\cdot)$ is bounded over X, there exists a constant M such that $|r_n(\theta)| \leq M$ for all n and all θ. Moreover, $\int_\Omega Mp(\theta)d\theta = M$, and hence, the dominated convergence theorem (see Theorem 10.27 in [13]), when applied to (2.13), implies that

$$\begin{aligned}\lim_{n\to\infty} Q_t(x^n;d^n) &= c(d) + \int_\Omega \lim_{n\to\infty} \Big\{J_{t+1}(f(x^n,d^n;\theta))\Big\}p(\theta)d\theta \\ &= c(d) + \int_\Omega J_{t+1}(f(x,d;\theta))p(\theta)d\theta = Q_t(x;d).\end{aligned}$$

Thus $Q_t(\cdot;\cdot)$ is continuous over $X \times D$. Since $c(\cdot)$ and $J_{t+1}(\cdot)$ are bounded over D and X, respectively, it follows that $Q_t(\cdot;\cdot)$ is also bounded over $X \times D$.

Now recall from (2.4) that for every $x \in X$, $J_t(x)$ is the minimum of $Q_t(x;d)$ over $d \in D$. Continuity of $J_t(\cdot)$ over X then follows from Berge's Maximum Theorem (see Theorem 17.31 in [9]). We nevertheless provide a detailed proof here for the sake of completeness (also see [181]). Consider a convergent sequence of states $x^n \in X$ such that $\lim_{n\to\infty} x^n = x$, and a corresponding sequence of optimal doses $d^n \in D$ in session t. Because D is compact, d^n has a convergent subsequence d^{n_k}. Let $\lim_{k\to\infty} d^{n_k} \triangleq d^* \in D$. We show that d^* is optimal in x in

session t. Suppose, by way of contradiction, that d^* is not optimal in x in session t. Then there exists some $d \in D$ such that

$$Q_t(x; d) < Q_t(x; d^*). \tag{2.14}$$

Now define the subsequence $a^{n_k} = d(1 - 1/n_k)$; note that $a^{n_k} \in D$ and $\lim_{k\to\infty} a^{n_k} = d$. Using continuity of $Q_t(\cdot;\cdot)$ in (2.14) we get

$$\lim_{k\to\infty} Q_t(x^{n_k}; a^{n_k}) = Q_t(x; d) < Q_t(x; d^*) = \lim_{k\to\infty} Q_t(x^{n_k}; d^{n_k}).$$

This implies that for sufficiently large k,

$$Q_t(x^{n_k}; a^{n_k}) < Q_t(x^{n_k}; d^{n_k}),$$

thus contradicting the optimality of d^{n_k} in x^{n_k}. Thus d^* must be optimal in x in session t, and in particular,

$$\lim_{k\to\infty} J_t(x^{n_k}) = \lim_{k\to\infty} Q_t(x^{n_k}; d^{n_k}) = Q_t(x; d^*) = J_t(x).$$

This shows that $J_t(\cdot)$ is continuous over X. The fact that $J_t(\cdot)$ is also bounded over X follows from (2.4) since $Q_t(\cdot;\cdot)$ is bounded over $X \times D$. This completes our proof of Theorem 2.3.3 by induction.

Monotonicity of optimal doses appears to be an intuitive property. We show by counterexamples in the next section, however, that it may not always hold.

2.3.2 Counterexamples to optimality of monotone dosing

The counterexamples here are constructed to illustrate that Assumptions 2.3.1 and 2.3.2 are not vacuous, and in particular, that a violation of an assumption can lead to optimal doses that are not increasing in disease conditions. As discussed in Section 2.3, the monotonicity properties of our dose-response and cost functions do not seem restrictive. Thus, below we construct (hypothetical) counterexamples where these properties hold but one of our (more restrictive) convexity assumptions is violated.

Counterexample 1: $h(\cdot)$ concave (risk-seeking decision-maker)

Consider a single-session, deterministic special case of our model wherein $X = \mathbb{R}_+$, $D = [0, 1]$ (that is, $\bar{d} = 1$), $f(x, d; \theta) = x - d$, $c(d) = 0.005d^2$, and $h(x) = 1 - e^{-x}$. This problem satisfies Assumptions 2.3.1 and 2.3.2 except notice that $h(\cdot)$ is now a concave function; that is, the decision-maker is risk-seeking with respect to the terminal disease condition. Since this is a single-session problem, we drop the subscript t and write

$$Q(x; d) = 0.005d^2 + 1 - e^{-x+d}.$$

That is, to find an optimal dose in disease condition x, we simply need to minimize the function $Q(x; d)$ over $d \in [0, 1]$. We consider two states: $x = 1$ and $x = 6$. When $x = 1$, $Q(1; d) = 0.005d^2 + 1 - e^{d-1}$ is a decreasing function of d over the interval $[0, 1]$ and hence dose $d = 1$ is uniquely optimal in state $x = 1$. When $x = 6$, $Q(6; d) = 0.005d^2 + 1 - e^{d-6}$ is a strictly convex function of d over the interval $D = [0, 1]$ and dose $d \approx 0.3527$ is its unique minimizer. In particular, no monotone policy is optimal.

Counterexample 2: $f(\cdot, \cdot; \theta)$ not convex

We again consider a single-session, deterministic special case of our model wherein $X = \mathbb{R}_+$, $D = [0, 1]$ (that is, $\bar{d} = 1$), $f(x, d; \theta) = 1 - e^{-x+d}$, $c(d) = 0.005d^2$, and $h(x) = x$. This problem satisfies Assumptions 2.3.1 and 2.3.2 except notice that $f(\cdot, \cdot; \cdot)$ is not a convex function. This results in

$$Q(x; d) = 0.005d^2 + 1 - e^{-x+d},$$

which is the same Q that was found in Counterexample 1 above. Thus, the rest of the discussion follows exactly as in Counterexample 1 to conclude that no monotone policy is optimal.

A robust counterpart of the stochastic DP in this chapter is presented in the next chapter.

Chapter 3

ROBUST RESPONSE-GUIDED DOSING

3.1 Background and motivation

One limitation of the stochastic DP of Chapter 2 is that the decision-maker is assumed to know at the outset the probability mass function (pmf) of the dose-response parameter. Any *a priori* estimate of this pmf, however, is subject to estimation errors. To tackle the resulting ambiguity, we present here a robust counterpart of the model of Chapter 2 (henceforth called the "nominal" model).

In our robust formulation, the pmf of the dose-response parameter will be assumed to belong to an uncertainty set. Uncertainty sets are often composed of pmfs that are in some sense "close to" a nominal pmf, which may have been estimated *a priori* from a clinical trial [21, 88, 139]. Roughly speaking, the decision-maker then follows a conservative approach whereby he/she attempts to find a dosing policy that minimizes the worst-case expected disutility over all pmfs from this uncertainty set. Examples of uncertainty sets include the interval set, the maximum likelihood set, the relative entropy set, and the ellipsoidal set. Although our general robust RGD model accommodates any of these sets, we illustrate our results in detail using the interval model. We show that the so-called inner maximization problem in the Bellman's equations for robust RGD with the interval uncertainty set is a linear program (LP) that can be solved analytically. Moreover, an optimal solution to this inner problem, that is, the worst-case pmf, does not depend on the observed disease condition and the dose chosen. This in turn implies that there exists a monotone dosing policy that is optimal to the robust stochastic DP, thus extending the main theoretical result from the nominal model. This extension is not only of theoretical interest but also significantly simplifies the computation of our robust optimal dosing policy. In particular, we show that

the state-action invariant structure of the worst-case pmf makes the robust problem only as hard to solve as the nominal problem. We further analyze in Section 3.3 a specific and common single-parameter formulation of the interval uncertainty set and provide a simple condition on the dose-response formula under which optimal doses vary monotonically with this parameter. We conclude by presenting numerical results on a hypothetical disease with an inverse-power dose-response function, and giving a counterexample to monotonicity when one of our assumptions is violated.

3.2 The robust stochastic DP

As in Chapter 2, let T denote the number of sessions, indexed by $t = 1, 2, \ldots, T$, in a treatment course. At the beginning of each session, the decision-maker observes a numerical score of the patient's disease condition, and chooses a dose for that session. These numerical scores belong to a compact interval $X \subseteq \mathbb{R}$. Smaller numbers in this set represent less severe disease. The disease condition at the beginning of session t is denoted by $x_t \in X$. The dose level chosen by the decision-maker for this session after observing x_t is denoted by d_t. Dose levels d_t belong to the interval $D \triangleq [0, \bar{d}] \subset \mathbb{R}$, where $\bar{d}$ is a finite upper bound on permissible dose levels.

For $t = 1, 2, \ldots, T$, disease conditions evolve according to dynamics $x_{t+1} = f(x_t, d_t; \Theta) = x_t + f_0(d_t; \Theta)$, for $x_t, x_{t+1} \in X$ and $d_t \in D$. All standard dose-response functions such as linear, Michaelis-Menten, inverse-power, Emax, Hill's, exponential, exponential linear-quadratic, power law, Gompertz, and Beta-Poisson can be expressed in this additively separable form (after taking logarithms in some cases). A detailed descriptions of these functions was given in section 2.3. Here, Θ are independent and identically distributed dose-response parameters in sessions t. Independence across sessions is somewhat restrictive although common in the literature (see, for example, Chapter 4 of [32], and also [162]). This assumption was employed in the model of Chapter 2 as well, and it holds when consecutive sessions are "sufficiently separated" from a biochemical viewpoint. Random variables Θ take values from a finite set $\Omega \triangleq \{\theta_1, \theta_2, \ldots, \theta_n\}$. We assume that the function $f_0(\cdot; \theta)$ is continuous over D

for each $\theta \in \Omega$.

Aversion to dose is modeled using a continuous disutility function $c : D \to \mathbb{R}_+$. Since D is compact, continuity of $c(\cdot)$ implies that it is bounded. Examples include linear, quadratic, and exponential functions. Aversion to disease conditions x_{T+1} at the end of the treatment course is modeled using a continuous and bounded disutility function $h : X \to \mathbb{R}_+$. Examples include linear, quadratic, exponential, and ramp (where the disutility is zero up to a disease-condition threshold and grows linearly thereafter).

Assumption 3.2.1 (Monotone and convex dose-response)**.** *The function $f_0(\cdot;\theta)$ is decreasing and convex in dose over D for every $\theta \in \Omega$.*

Assumption 3.2.2 (Increasing and convex disutilities)**.** *The disutility function $c(\cdot)$ is increasing and convex over D; the disutility function $h(\cdot)$ is increasing and convex over X.*

A detailed justification for these assumptions was provided in Chapter 2. In particular, Assumption 3.2.2 encodes a risk-averse decision-maker. Several examples of clinically relevant functions that satisfy these assumptions were also listed in Chapter 2; as such, we do not believe these assumptions to be particularly restrictive.

Pursuing standard practice in robust stochastic DP [21, 88, 139], we assume that the pmf of Θ is only known to lie in some set $\mathcal{P}$. In the robust optimization parlance, set $\mathcal{P}$ is called the uncertainty set and is composed of all "plausible" pmfs. This set is often chosen so that it includes all pmfs that are "close to" some nominal pmf. More precisely, let $\Delta \triangleq \{p(\cdot) : p(\theta) \geq 0, \sum_{\theta\in\Omega} p(\theta) = 1\}$ be the probability simplex in $\mathbb{R}^n$, and let $\mathcal{P} \subseteq \Delta$. Then, the worst-case total expected disutility is minimized by solving the robust Bellman's equations

$$\tilde{J}_t(x_t) = \min_{d_t\in D} \left\{ \overbrace{\max_{p_t(\cdot)\in\mathcal{P}} \left(c(d_t) + \sum_{\theta\in\Omega} \tilde{J}_{t+1}(x_t + f_0(d_t;\theta))p_t(\theta) \right)}^{\text{"inner problem"}} \right\}, \ \forall x_t \in X, \text{ and } t = 1, 2, \ldots, T, \tag{3.1}$$

with the boundary condition $\tilde{J}_{T+1}(x) = h(x)$ for all $x \in X$. Here, $\tilde{J}_t(\cdot)$, for $t = 1, 2, \ldots, T+1$, are called the robust optimal cost-to-go functions, and an optimal solution to the inner

problem is called a worst-case distribution.

The robust stochastic DP is computationally tractable if the inner maximization problem is easy to solve. This occurs, for example, when $\mathcal{P}$ is chosen to be a convex set. This, combined with the linearity (in $p_t(\cdot)$) of the objective function, implies that the inner problem is convex. Some examples of convex uncertainty sets are interval, maximum likelihood, ellipsoidal and entropy [21, 88, 139]. We focus on the interval uncertainty model in the subsequent discussion.

The interval model is motivated by statistical estimates of confidence intervals on the pmf components. It can also be obtained by projecting the ellipsoidal or maximum likelihood uncertainty sets onto the coordinate axes [21, 139]. In this model [21], the uncertainty set $\mathcal{P}$ is defined such that the probabilities $p(\theta)$, for $\theta \in \Omega$, belong to an interval. More precisely, $\mathcal{P} \triangleq \left\{p(\theta) \in \Delta : p_L(\theta) \leq p(\theta) \leq p_H(\theta)\right\}$, for some constants $p_L(\theta)$ and $p_H(\theta)$, for $\theta \in \Omega$. We assume, to avoid trivialities, that $\mathcal{P}$ is non-empty. A necessary condition for this is that $\sum_{\theta\in\Omega} p_L(\theta) \leq 1$ and $\sum_{\theta\in\Omega} p_H(\theta) \geq 1$.

In the interval uncertainty model, the inner problem is an LP, and in fact, we are able to find a closed-form solution for this LP. Moreover, it turns out that the resulting worst-case pmf does not depend on t, x_t, or d_t. This implies that the robust problem is as easy to solve as the nominal problem and it also enables us to prove a monotonicity result in Theorem 3.2.4 below, analogous to Theorem 3.2.4 from the nominal case, under the following additional assumption.

Assumption 3.2.3 (Monotonicity in dose-response parameter)**.** *The dose-response function $f_0(d;\cdot)$ is decreasing over Ω for each $d \in D$.*

This assumption implies that larger dose-response parameters induce a larger improvement in the disease condition. Our proofs below can be rewritten for a variation of this assumption where the improvement is increasing in the dose-response parameter. All dose-response models known to us satisfy one of these two types of monotonicity; Assumption 3.2.3 thus is not restrictive.

Theorem 3.2.4. *Suppose Assumptions 3.2.1, 3.2.2, and 3.2.3 hold. Then, optimal dose levels increase with worsening disease conditions in each session in the robust stochastic DP with interval uncertainty.*

Proof of Theorem 3.2.4: The proof employs backward induction on $t = T+1, T, \ldots, 1$. We first show that if $\tilde{J}_{t+1}(\cdot)$ is increasing, convex, continuous, and bounded, then in treatment session t, there exist optimal doses that are monotone. We then show that $\tilde{J}_t(\cdot)$ inherits these four properties from $\tilde{J}_{t+1}(\cdot)$. The theorem then follows because the utility function $h(\cdot)$, that is, $\tilde{J}_{T+1}(\cdot)$, is assumed to possess these properties.

So suppose, as the inductive hypothesis, that $\tilde{J}_{t+1}(\cdot)$ is increasing, convex, continuous, and bounded. For each fixed state-action pair (x_t, d_t), the inner problem

$$\tilde{Q}_t(x_t, d_t) \triangleq \max_{p_t(\cdot)\in\mathcal{P}} \left(c(d_t) + \sum_{\theta\in\Omega} \tilde{J}_{t+1}(x_t + f_0(d_t;\theta))p_t(\theta) \right) \tag{3.2}$$

is an n-variable LP in variables $p_t(\theta)$, for $\theta \in \Omega$. This LP has an optimal solution because its feasible region $\mathcal{P}$ is compact. We denote this optimal solution by $p_t^*(x_t, d_t; \theta)$, for $\theta \in \Omega$. With this notation, we have that the robust optimal cost-to-go functions are given by

$$\tilde{J}_t(x_t) = \min_{d_t\in D} \tilde{Q}_t(x_t, d_t) = \min_{d_t\in D} \left(c(d_t) + \sum_{\theta\in\Omega} \tilde{J}_{t+1}(x_t + f_0(d_t;\theta))p_t^*(x_t, d_t;\theta) \right). \tag{3.3}$$

Claim 1: Problem (3.2) has an optimal solution that does not depend on $t, (x_t, d_t)$ and hence $p_t^*(x_t, d_t; \theta)$ can in fact be written as $p^*(\theta)$ for $\theta \in \Omega$.

Proof of Claim 1: For any fixed session t and fixed state-action pair (x_t, d_t), the inner maximization problem is equivalent to the LP

$$\max\ g(p) \triangleq \sum_{\theta\in\Omega} \tilde{J}_{t+1}(x_t + f(d_t;\theta))p(\theta) \tag{3.4}$$

$$p_L(\theta) \le p(\theta) \le p_H(\theta) \text{ and } 0 \le p(\theta) \le 1, \text{ for } \theta \in \Omega;\ \sum_{\theta\in\Omega} p(\theta) = 1. \tag{3.5}$$

We henceforth simplify constraints (3.5) by assuming without loss of generality that $0 \le p_L(\theta) \le p_H(\theta) \le 1$ for all $\theta \in \Omega$. Suppose, without loss of generality, that the elements of Ω

are organized in ascending order as $\theta_1 \leq \ldots \leq \theta_n$. Then, Assumption 3.2.3, combined with the fact that $\tilde{J}_{t+1}(\cdot)$ is increasing, implies that the coefficients $a(\theta_i) \triangleq \tilde{J}_{t+1}(x_t + f_0(d_t, \theta_i))$ of the variables $p(\theta_i)$ in the LP objective function $g(\cdot)$, are decreasing in $i = 1, 2, \ldots, n$. Since $a(\theta_1)$ is the largest coefficient, we would like to attach the largest possible probability mass to θ_1. Thus, if there is a feasible pmf $p(\cdot)$ for which $p(\theta_1) = p_H(\theta_1)$, we choose $p^*(\theta_1) = p_H(\theta_1)$. Otherwise, we still choose the largest feasible value that $p(\theta_1)$ may take and continue. This idea helps us claim that the solution $p^*(\cdot)$ defined below in Equation (3.7) is optimal. But first, we need some additional notation. Let k be the smallest index (called the "switching index") in $\{1, 2, \ldots, n\}$ such that

$$\sum_{i=1}^{k} p_H(\theta_i) + \sum_{i=k+1}^{n} p_L(\theta_i) \geq 1. \tag{3.6}$$

Also let $r_k = 1 - \sum_{i=1}^{k-1} p_H(\theta_i) - \sum_{i=k+1}^{n} p_L(\theta_i)$. We will prove below that the pmf

$$p^*(\theta_i) = \begin{cases} p_H(\theta_i), & \text{if } i \leq k-1 \\ r_k, & \text{if } i = k \\ p_L(\theta_i), & \text{if } i \geq k+1, \end{cases} \tag{3.7}$$

schematically illustrated in Figure 3.1, is optimal for the above LP.

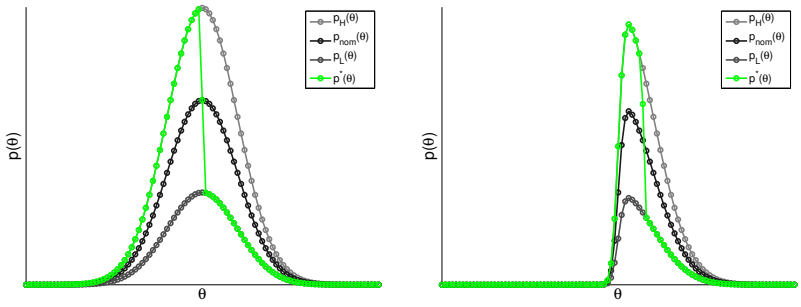


Figure 3.1: Illustration of the worst-case pmf defined in Equation (3.7) for two different nominal pmfs.

To show that (3.7) is optimal to (3.4), observe first that $p^*(\cdot)$ is well-defined since $k = n$ satisfies condition (3.6). Also, by construction, the components $p^*(\theta_i)$ sum up to 1. Moreover, components $p^*(\theta_i)$ for $i = 1, 2, \ldots, k-1$ and $i = k+1, k+2, \ldots, n$ are, by construction, between $p_L(\theta_i)$ and $p_H(\theta_i)$. So, for $p^*(\cdot)$ to be feasible, we need to check whether $p_L(\theta_k) \leq r_k \leq p_H(\theta_k)$. We do this as follows. Recall that k is the smallest index for which (3.6) holds. Therefore, by the definition of r_k, we have, $r_k = 1 - \sum_{i=1}^{k-1} p_H(\theta_i) - \sum_{i=k+1}^{n} p_L(\theta_i) \leq p_H(\theta_k)$. Now, to show that $r_k \geq p_L(\theta_k)$, we consider two cases. Firstly, if $k = 1$, then $r_1 = 1 - \sum_{i=2}^{n} p_L(\theta_i) \geq p_L(\theta_1)$ by the necessary condition for $\mathcal{P}$ to be non-empty. Secondly, if $k \neq 1$, we know that $k-1$ does not satisfy (3.6). Therefore,

$$\sum_{i=1}^{k-1} p_H(\theta_i) + \sum_{i=k}^{n} p_L(\theta_i) < 1 \;\Rightarrow\; p_L(\theta_k) < 1 - \sum_{i=1}^{k-1} p_H(\theta_i) - \sum_{i=k+1}^{n} p_L(\theta_i) = r_k.$$

Thus, we have shown that $p_L(\theta_k) \leq r_k \leq p_H(\theta_k)$ as required for feasibility of $p^*(\cdot)$.

Now, to prove optimality, let $p(\cdot)$ be any feasible solution of (3.4). Then,

$$\begin{aligned}
g(p^*) - g(p) &= \sum_{i=1}^{n} a(\theta_i)(p^*(\theta_i) - p(\theta_i)) \\
&= \sum_{i=1}^{k-1} a(\theta_i)(p_H(\theta_i) - p(\theta_i)) + a(\theta_k)\left(1 - \sum_{i=1}^{k-1} p_H(\theta_i) - \sum_{i=k+1}^{n} p_L(\theta_i) - p(\theta_k)\right) \\
&\quad + \sum_{i=k+1}^{n} a(\theta_i)(p_L(\theta_i) - p(\theta_i)) \\
&= \sum_{i=1}^{k-1} a(\theta_i)(p_H(\theta_i) - p(\theta_i)) + \sum_{i=k+1}^{n} a(\theta_i)(p_L(\theta_i) - p(\theta_i)) \\
&\quad + a(\theta_k)\left(-\sum_{i=1}^{k-1} p_H(\theta_i) - \sum_{i=k+1}^{n} p_L(\theta_i) + \sum_{i=1}^{k-1} p(\theta_i) + \sum_{i=k+1}^{n} p(\theta_i)\right) \\
&= \sum_{i=1}^{k-1} (a(\theta_i) - a(\theta_k))(p_H(\theta_i) - p(\theta_i)) + \sum_{i=k+1}^{n} (a(\theta_i) - a(\theta_k))(p_L(\theta_i) - p(\theta_i)) \geq 0.
\end{aligned}$$

Here, the last inequality holds because each summand is nonnegative. Hence, $p^*(\cdot)$ as described in (3.7) is optimal. Furthermore, we emphasize that this optimal solution does not depend on t, x_t, or d_t. This is because the above proof of optimality only uses the fact

that $a(\theta_j) - a(\theta_i) \geq 0$ for $j > i$, and in particular, does not utilize the specific values of these coefficients. In other words, this optimal solution depends only on the ordering of the coefficients $a(\theta_i)$, which is invariant with respect to t, x_t, d_t. We thus rewrite (3.2) as $\tilde{Q}_t(x_t, d_t) = c(d_t) + \sum_{\theta \in \Omega} \tilde{J}_{t+1}(x_t + f_0(d_t; \theta))p^*(\theta)$.

End of Proof of Claim 1.

Proofs of the next three claims are identical to similar results in Chapter 2.

Claim 2: $\tilde{Q}_t(\cdot;\cdot)$ has decreasing differences over $X \times D$, and hence optimal doses in session t are increasing in disease conditions.

Proof of Claim 2: Consider any $x^+ \geq x^-$ and $d^+ \geq d^-$. Then $\tilde{Q}_t(x^+; d^+) - \tilde{Q}_t(x^+; d^-) = c(d^+) - c(d^-) + \sum_{\theta \in \Omega} \big(\tilde{J}_{t+1}(x^+ + f_0(d^+; \theta)) - \tilde{J}_{t+1}(x^+ + f_0(d^-; \theta))\big)p^*(\theta) \leq c(d^+) - c(d^-) + \sum_{\theta \in \Omega} \big(\tilde{J}_{t+1}(x^- + f_0(d^+; \theta)) - \tilde{J}_{t+1}(x^- + f_0(d^-; \theta))\big)p^*(\theta) = \tilde{Q}_t(x^-; d^+) - \tilde{Q}_t(x^-; d^-)$. Here, the inequality follows because $f_0(d^+; \theta) \leq f_0(d^-; \theta)$ by the first half of Assumption 3.2.1 and because $\tilde{J}_{t+1}(\cdot)$ is increasing and convex. This proves decreasing differences, which imply monotonicity of optimal doses via a standard argument by contradiction.

End of Proof of Claim 2.

Claim 3: $\tilde{J}_t(\cdot)$ is increasing in disease condition.

Proof of Claim 3: Suppose dose d^+ is optimal in state x^+ and dose d^- is optimal in state $x^- \leq x^+$. Then, $\tilde{J}_t(x^+) = \tilde{Q}_t(x^+; d^+) = c(d^+) + \sum_{\theta \in \Omega} \big(\tilde{J}_{t+1}(x^+ + f_0(d^+; \theta))\big)p^*(\theta) \geq c(d^+) + \sum_{\theta \in \Omega} \big(\tilde{J}_{t+1}(x^- + f_0(d^+; \theta))\big)p^*(\theta) = \tilde{Q}_t(x^-; d^+) \geq \tilde{Q}_t(x^-; d^-) = \tilde{J}_t(x^-)$. Here, the first inequality holds because $\tilde{J}_{t+1}(\cdot)$ is increasing and the second inequality holds because d^- is optimal in x^-.

End of Proof of Claim 3.

Claim 4: $\tilde{Q}_t(\cdot;\cdot)$ is convex over $X \times D$, and $\tilde{J}_t(\cdot)$ is convex over X.

Proof of Claim 4: First note that $\tilde{J}_{t+1}(x_t + f_0(d_t;\theta))$ is convex over $X \times D$ for every $\theta \in \Omega$. This holds because $\tilde{J}_{t+1}(\cdot)$ is an increasing convex function and $x_t + f_0(d_t;\theta)$ is convex by the second half of Assumption 3.2.1 (see Section 3.2.4 of [29]). Thus, $\sum_{\theta\in\Omega} \tilde{J}_{t+1}(x_t + f_0(d_t;\theta))p_t^*(\theta)$ is also convex because expectation preserves convexity. Convexity of $\tilde{Q}_t(\cdot;\cdot)$ then follows because $c(\cdot)$ is convex by Assumption 3.2.2. Finally, $\tilde{J}_t(\cdot)$ is convex because convexity is preserved under minimization (see Section 3.2.5 of [29]).

End of Proof of Claim 4.

Claim 5: $\tilde{Q}_t(\cdot;\cdot)$ and $\tilde{J}_t(\cdot)$ are continuous and bounded.

Proof of Claim 5: By a technical argument identical to that in Chapter 2; omitted.

End of Proof of Claim 5.

Claims 2, 3, 4, and 5 restore the inductive hypothesis.

End of Proof of Theorem 3.2.4.

In addition to facilitating the above proof of monotonicity, the worst-case pmf in (3.7) offers a computational advantage. Since this worst-case pmf does not depend on t, x_t, or d_t, we only need to find it once at the outset of the Bellman's recursion (3.1). This reduces the robust Bellman's equations to those in a nominal stochastic DP where the pmf equals the worst-case pmf (3.7).

3.3 Monotonicity with respect to ambiguity level

We now further analyze one common interval uncertainty set that is characterized by a single parameter $0 \leq \delta \leq 1$ (see [21]). Specifically, the lower and upper bounds are chosen as

$$p_L(\theta) = (1-\delta)p_{\text{nom}}(\theta), \quad \text{and} \quad p_H(\theta) = (1+\delta)p_{\text{nom}}(\theta), \quad \text{for all } \theta \in \Omega,$$

where $p_{\text{nom}}(\cdot)$ is some nominal pmf over Ω. In Theorem 3.3.2 below, we prove that optimal doses are monotonically increasing in δ under the following assumption.

Assumption 3.3.1. *The dose-response function $f_0(\cdot;\cdot)$ has increasing differences over $D\times\Omega$; that is, for $d^+ \geq d^-$ and for $\theta^+ \geq \theta^-$, $f_0(d^+;\theta^+) - f_0(d^-;\theta^+) \geq f_0(d^+;\theta^-) - f_0(d^-;\theta^-)$.*

Theorem 3.3.2. *Suppose the dose response function satisfies all assumptions of Theorem 3.2.4, as well as Assumption 3.3.1. Then robust optimal doses are increasing in δ.*

Let us first examine the worst-case pmf $p_t^*(\cdot)$ for this interval uncertainty model. Recall that the switching index k is the smallest integer in $\{1, 2, \ldots, n\}$ which satisfies

$$\sum_{i=1}^{k}(1+\delta)p_{\text{nom}}(\theta_i) + \sum_{i=k+1}^{n}(1-\delta)p_{\text{nom}}(\theta_i) \geq 1. \tag{3.8}$$

For $\delta = 0$, this is true for any integer m and the worst-case distribution is the same as the nominal distribution. For $0 < \delta \leq 1$, we show that the switching index k is independent of δ and depends only on the nominal pmf. This holds because of the following algebraic argument. Consider any $0 < \delta \leq 1$ and any index m between 1 and n. Then,

$$(1-\delta)\sum_{i=1}^{m}p_{\text{nom}}(\theta_i) + (1+\delta)\sum_{i=m+1}^{n}p_{\text{nom}}(\theta_i) \geq 1 \text{ iff } \delta\left[-\sum_{i=1}^{m}p_{\text{nom}}(\theta_i) + \sum_{i=m+1}^{n}p_{\text{nom}}(\theta_i)\right] \geq 0$$
$$\text{iff } \sum_{i=m+1}^{n}p_{\text{nom}}(\theta_i) \geq \sum_{i=1}^{m}p_{\text{nom}}(\theta_i). \tag{3.9}$$

Thus, condition (3.8) is equivalent to inequality (3.9), which depends only on the nominal pmf. Thus, the switching index k does not depend on the parameter δ.

Now, to examine the dependence of the optimal dose on δ, we define the function $W_t(d,\delta) \triangleq c(d) + \sum_{\theta\in\Omega} J_{t+1}(x + f_0(d;\theta))p^*(\theta)$; this equals the worst-case expected value for a fixed dose d when the uncertainty is parametrized by δ. For this discussion, the state x_t is assumed to be fixed and hence we have suppressed the dependence of W_t on x_t in our notation. The robust optimal cost-to-go function is obtained by minimizing $W_t(\cdot,\delta)$ over all doses d. First, we derive an expression for W_t that is convenient for the rest of our proof below (we suppress the subscript nom for brevity).

$$W_t(d,\delta) = c(d) + (1+\delta)\sum_{i=1}^{k-1}\tilde{J}_{t+1}(x + f_0(d;\theta_i))p(\theta_i) + (1-\delta)\sum_{i=k+1}^{n}\tilde{J}_{t+1}(x + f_0(d;\theta_i))p(\theta_i)$$

$$+\left(1-\sum_{i=1}^{k-1}(1+\delta)p(\theta_i)-\sum_{i=k+1}^{n}(1-\delta)p(\theta_i)\right)\tilde{J}_{t+1}(x+f_0(d;\theta_k))$$

$$=c(d)+(1+\delta)\sum_{i=1}^{k-1}\tilde{J}_{t+1}(x+f_0(d;\theta_i))p(\theta_i)+(1-\delta)\sum_{i=k+1}^{n}\tilde{J}_{t+1}(x+f_0(d;\theta_i))p(\theta_i)$$

$$+\left(p(\theta_k)-\delta\left(\sum_{i=1}^{k-1}p(\theta_i)-\sum_{i=k+1}^{n}p(\theta_i)\right)\right)\tilde{J}_{t+1}(x+f_0(d;\theta_k))$$

$$=c(d)+\sum_{i=1}^{n}\tilde{J}_{t+1}(x+f_0(d;\theta_i))p(\theta_i)+\delta\sum_{i=1}^{k-1}\tilde{J}_{t+1}(x+f_0(d;\theta_i))p(\theta_i)$$

$$-\delta\sum_{i=k+1}^{n}\tilde{J}_{t+1}(x+f_0(d;\theta_i))p(\theta_i)-\delta\left(\sum_{i=1}^{k-1}p(\theta_i)-\sum_{i=k+1}^{n}p(\theta_i)\right)\tilde{J}_{t+1}(x+f_0(d;\theta_k))$$

$$=c(d)+\sum_{i=1}^{n}\tilde{J}_{t+1}(x+f_0(d;\theta_i))p(\theta_i)+\delta\sum_{i=1}^{k-1}p(\theta_i)\left[\tilde{J}_{t+1}(x+f_0(d;\theta_i))-\tilde{J}_{t+1}(x+f_0(d;\theta_k))\right]$$

$$+\delta\sum_{i=k+1}^{n}p(\theta_i)\left[\tilde{J}_{t+1}(x+f_0(d;\theta_k))-\tilde{J}_{t+1}(x+f_0(d;\theta_i))\right].$$

Now, to prove that the optimal dose is increasing in δ, we show that the function $W_t(\cdot,\cdot)$ has decreasing differences. Let $0\le\delta^-\le\delta^+\le 1$ and $d^+,d^-\in D,\ d^+\ge d^-$. Let $S=[W_t(d^+;\delta^+)-W_t(d^-;\delta^+)]-[W_t(d^+;\delta^-)-W_t(d^-;\delta^-)]$. For convenience, denote $\tilde{J}_{t+1}(x+f_0(d^+,\theta_i))$ by $\tilde{J}^{+i}$, $\tilde{J}_{t+1}(x+f_0(d^-,\theta_i))$ by $\tilde{J}^{-i}$, and so on. Then, noting that terms that are functions of d alone (and not of δ) cancel out, S can be written after some algebraic simplification, as

$$S=(\delta^+-\delta^-)\sum_{i=1}^{k-1}p(\theta_i)\left[\tilde{J}^{+i}-\tilde{J}^{+k}-\tilde{J}^{-i}+\tilde{J}^{-k}\right]$$

$$+(\delta^+-\delta^-)\sum_{i=k+1}^{n}p(\theta_i)\left[\tilde{J}^{+k}-\tilde{J}^{+i}-\tilde{J}^{-k}+\tilde{J}^{-i}\right].$$

We show that the term inside the summation is always non-positive. Consider any two doses $d^+\ge d^-$ and dose-response parameters $\theta^+\ge\theta^-$. Since the dose-response function is decreasing in dose and the uncertainty parameter θ, one of the following two orderings must hold:

$$\text{either } f_0^{++}\le f_0^{+-}\le f_0^{-+}\le f_0^{--},\quad \text{or}\quad f_0^{++}\le f_0^{-+}\le f_0^{+-}\le f_0^{--}.$$

Here, once again, f_0^{++} denotes $f_0(d^+;\theta^+)$, and so on. Further, since $f_0(\cdot;\cdot)$ has increasing differences in (d,θ), it satisfies $f_0^{--} - f_0^{-+} \geq f_0^{+-} - f_0^{++}$. Then, as $\tilde{J}_{t+1}$ is a convex increasing function, we have, $\tilde{J}^{+-} - \tilde{J}^{++} - \tilde{J}^{--} + \tilde{J}^{-+} \leq 0$. Here, $\tilde{J}^{++} \triangleq \tilde{J}_{t+1}(x + f_0(d^+;\theta^+))$, and so on. Thus, the term inside the summation is always non-positive and $W_t(\cdot,\cdot)$ has decreasing differences over $D \times [0,1]$.

Decreasing differences imply Theorem 3.3.2 via a standard argument by contradiction (omitted).

Theorem 3.3.2 also holds if the dose-response function is monotonically increasing in θ, and has decreasing differences over $X \times D$. This can be shown via a straightforward algebraic modification (omitted) of the above proof. Assumption 3.3.1 requires that the marginal improvement offered by a larger dose be increasing in the dose-response parameter and hence could be restrictive. Nevertheless, it is a mathematical necessity as we shall see in the next section.

3.4 Numerical results

We conducted experiments for a hypothetical disease with inverse-power response, where state y_t evolves as $y_{t+1} = \frac{1}{(\Theta+d_t)^k} y_t$ for some $k > 0$. Taking logarithms, and setting $x_t = \ln y_t$, this reduces to $x_{t+1} = x_t - k\ln(\Theta + d_t)$. The function $f_0(d_t;\Theta) = -k\ln(d_t + \Theta)$ satisfies Assumptions 3.2.1, 3.2.3, and 3.3.1. We set $c(d) = cd$; here, as in Chapter 2, $c > 0$ is called the coefficient of dose aversion. The terminal disutility function $h(x)$ was chosen to equal $\exp(x)$. Since x is the logarithm of the disease condition y, this exponential function amounts to trading-off the total dose given against the final disease state reached as is common in the clinical literature on RGD. These disutility functions were also employed in Chapter 2 and they satisfy Assumption 3.2.2. We report numerical results for $k = 1$, using $T = 10$ sessions with initial disease condition $x_1 = 1$ and the dose levels normalized to $D = [0,1]$. The uncertain dose-response parameters Θ are supported on a discretized subset with 61 equally-spaced grid points of the interval $[1,2]$. The nominal pmf is chosen to be a discretized Normal distribution centered at 1.5 on this grid. As in Chapter 1, the value of c is found

by solving an inverse optimization problem, so that a dose of 0.5 is optimal for a 10-session deterministic DP with dose-response parameter 1.5. Here, 0.5 and 1.5 are midpoints of the respective intervals. The parameter δ is varied between 0 and 1 in increments of 0.2, where $\delta = 0$ recovers the nominal DP. The results are in Figures 3.2, 3.3, and 3.4.

Figure 3.2 shows, as expected from Theorems 3.2.4 and 3.3.2, that robust optimal doses are increasing in worsening disease conditions and in the level of uncertainty δ. We also found, as expected, that the cost-to-go was increasing in δ because the inner maximization is performed over a larger uncertainty set. This is consistent with Figure 3.4, which shows that the price of robustness is increasing in δ.

3.5 Counterexample to Theorem 3.3.2

Finally, we were able to construct a counterexample where the conclusion of Theorem 3.3.2 was numerically observed to fail because this function was increasing in the dose-response parameter but had increasing differences.

Consider the Michaelis-Menten function, with

$$f_0(d_t; \Theta) = \ln\Big(\frac{\Theta}{\Theta + d_t}\Big).$$

Assumption 3.2.3 stated that the dose-response function $f_0(d;\cdot)$ is decreasing over Ω for each $d \in D$. This Michaelis-Menten function violates that assumption. We previously mentioned that the choice of taking $f_0(d;\cdot)$ to be decreasing over Ω for each $d \in D$ was arbitrary and that the function $f_0(d;\cdot)$ need only be monotone. However, if $f_0(d;\cdot)$ is in fact increasing over Ω for each $d \in D$, then assumption 3.3.1 would need to also be flipped: namely, $f_0(\cdot;\cdot)$ needs to have *decreasing* differences over $D \times \Omega$. This Michaelis-Menten function in fact has increasing differences. Thus, in any event, it does not satisfy one of assumptions 3.2.3 or 3.3.1, while still satisfying assumptions 3.2.1 and 3.2.2.

Numerically, we see that this Michaelis-Menten counterexample exhibits the opposite behavior to theorem 3.3.2. That is, the optimal dose is increasing in *decreasing* δ. This is shown in Figure 3.5.

One question that arises from this chapter is how such a nominal pmf of the dose-response parameter could be determined in the first place. One method would by using a group-averaged pmf, as determined by a clinical trial. In the following chapter, we present a Bayesian learning framework to find such a pmf by dosing a cohort of patients.

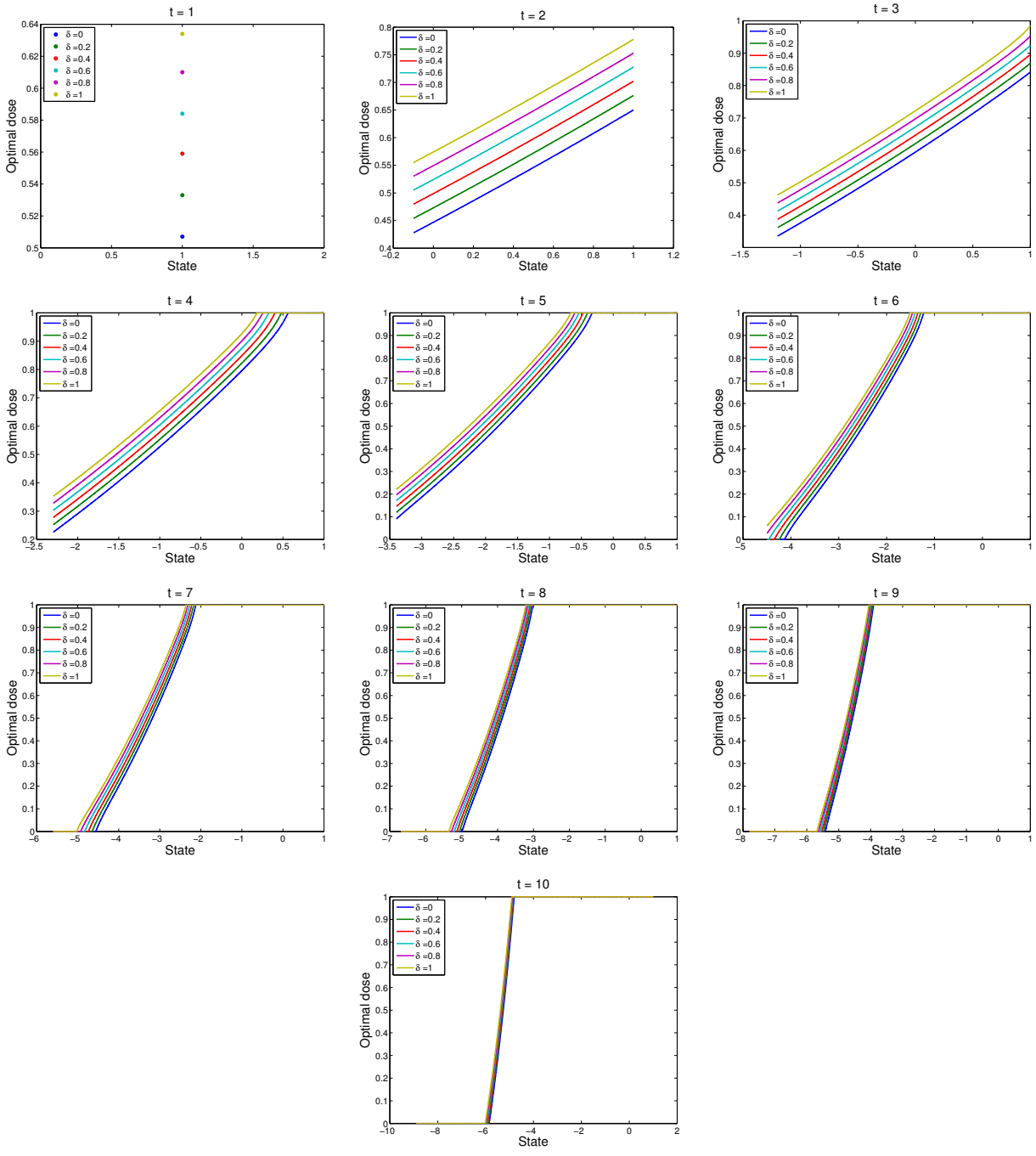


Figure 3.2: Robust optimal dose, session-by-session. As in Theorems 3.2.4 and 3.3.2, robust optimal doses are increasing in worsening disease conditions and in increasing dose-response ambiguity.

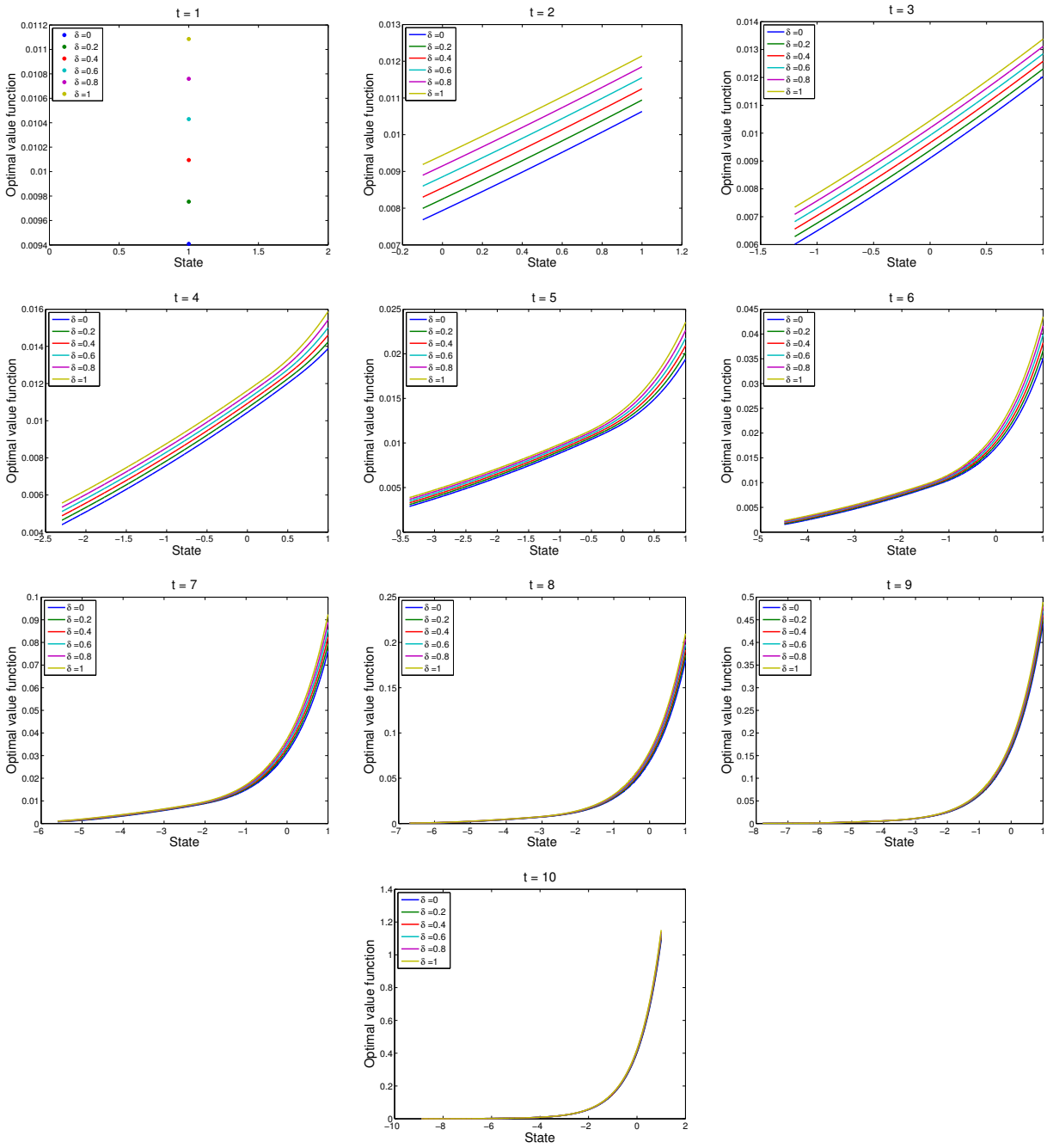


Figure 3.3: Robust objective function, session-by-session. Objective function increases with increasing δ since the inner maximization is performed over a larger uncertainty set.

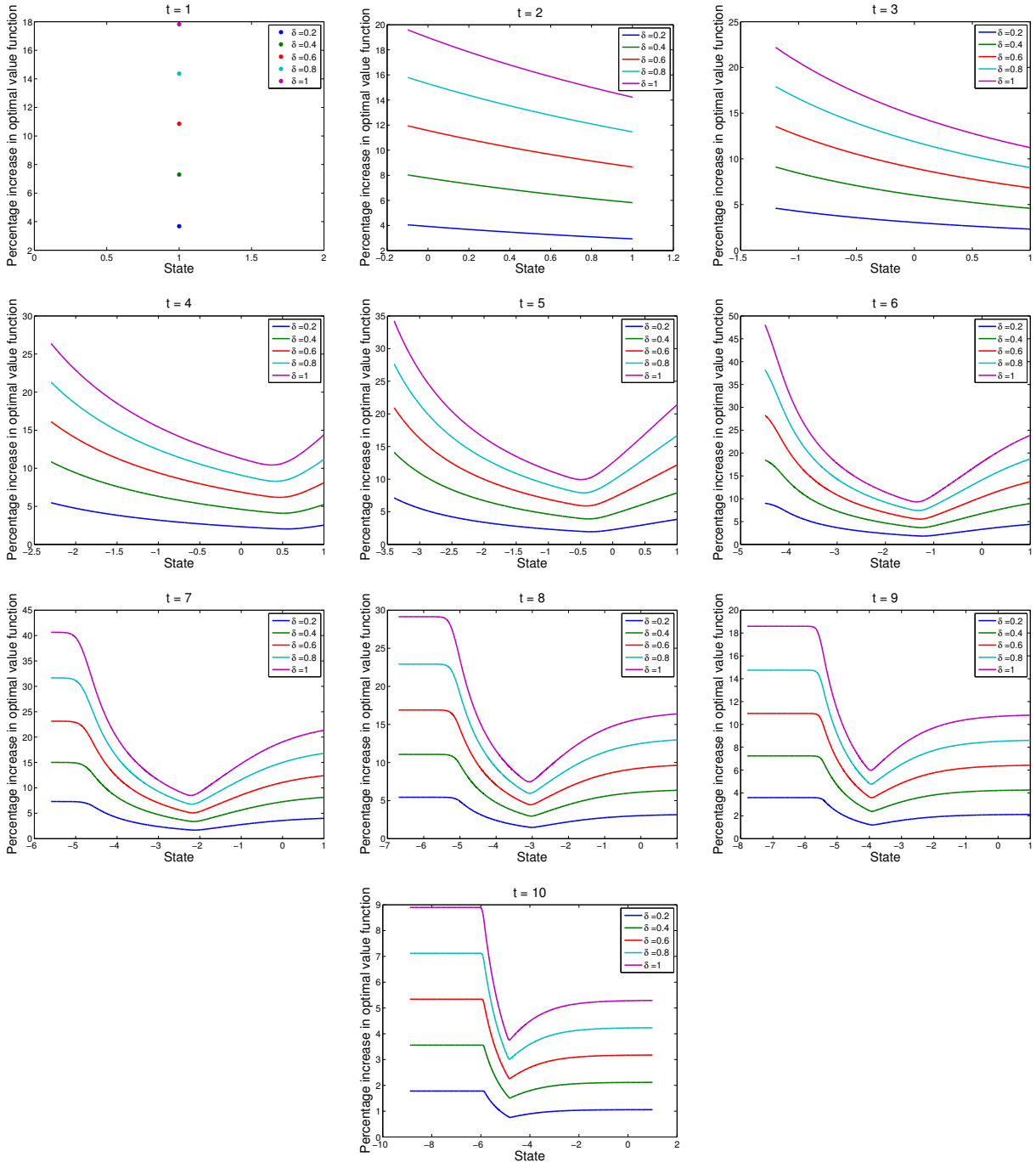


Figure 3.4: Percentage increase in the robust optimal cost-to- go function over the nominal optimal cost-to-go function, session-by-session. A higher ambiguity in dose-response results in a larger price of robustness.

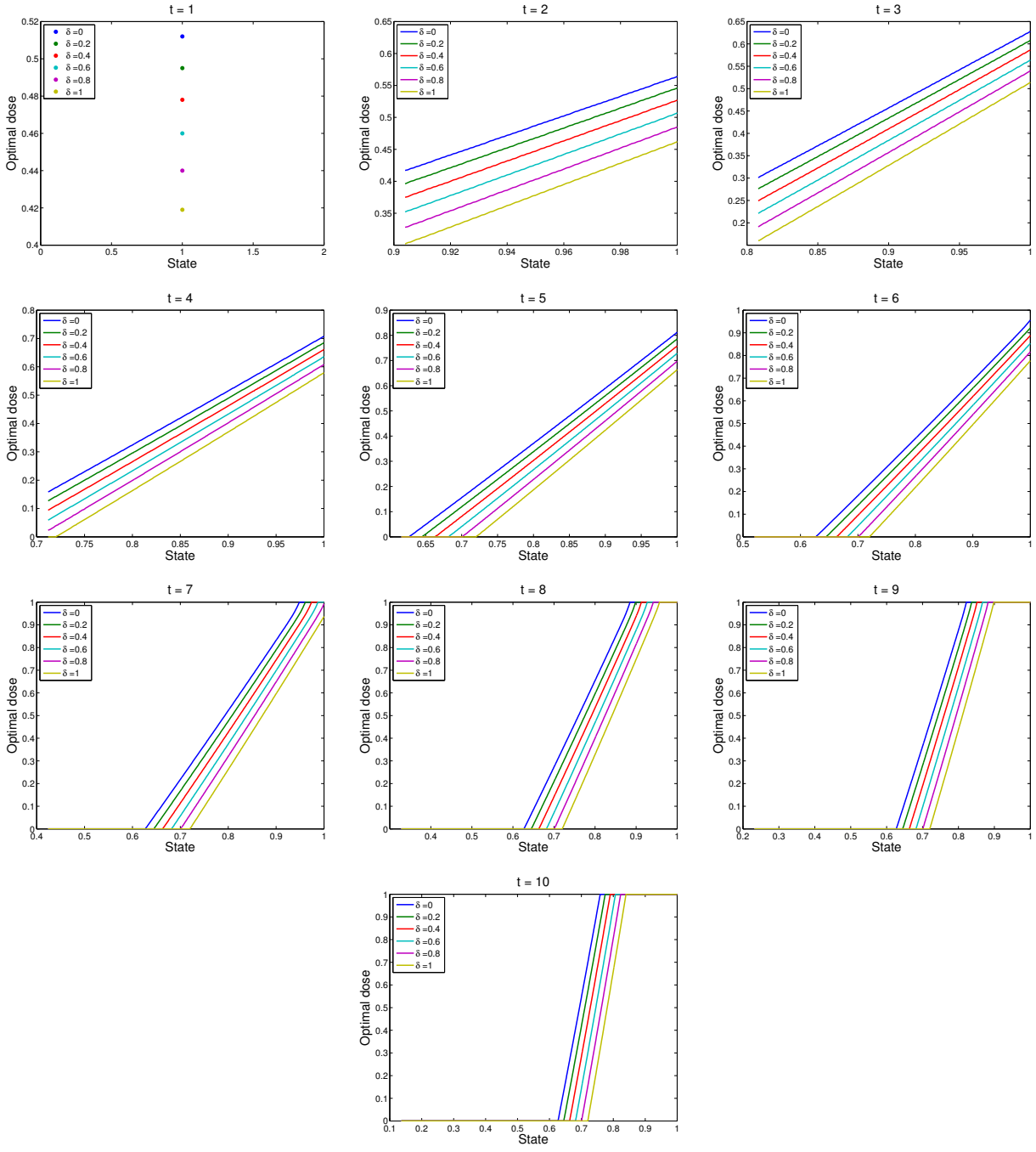


Figure 3.5: Counterexample to Theorem 3.3.2, using a Michaelis-Menten state transition function. Optimal dose is not increasing with increasing δ due to a violation of the decreasing-differences assumption on $f_0(\cdot;\cdot)$.

Chapter 4

OPTIMAL BAYESIAN LEARNING OF DOSE-RESPONSE PARAMETERS FROM A COHORT

4.1 Background and motivation

As we have discussed in previous chapters, dose-response formulas can be employed to model the evolution of patients' numerical score of disease condition over multiple treatment sessions. For instance, the fraction of surviving tumor-cells as a function of radiation dose is modeled using an exponential linear-quadratic in radiobiology [72]. The evolution of LDL cholesterol levels under phytosterol treatment was described using the exponential function in [46] and via the Michaelis-Menten formula in [56]. The reduction in blood pressure as a function of diuretic dose was modeled using the exponential function in [147]. The Michaelis-Menten formula was used to model the evolution of 28-joint disease activity scores (DAS28) in rheumatoid arthritis in Chapter 1. Other examples of dose-response functions include logistic, Hill's, Emax, Gompertz, power law, and beta-Poisson (see [73, 94, 100, 121, 125, 134, 147, 166, 172, 182]).

One main utility of such dose-response functions is in determining optimal doses that balance disease control against the risk of adverse effects or against the "cost" of administering high doses. In the preface to an edited book on dose-response studies, Ting [185] emphasized that:

"*establishing the dose-response relationship is one of the most important activities in developing a new drug.*"

In the United States, the Food and Drug Administration (FDA) has long-espoused the use

of dose-response curves for this purpose. Indeed, the FDA guidelines [60] stated:

"*[dose-response] information can help identify an appropriate starting dose, the best way to adjust dosage to the needs of a particular patient, and a dose beyond which increases would be unlikely to provide added benefit or would produce unacceptable side effects.*"

Dose-response functions often include one crucial parameter. Thus, if one intends to make dosing decisions based on such a function, one must rely on an estimate of this parameter.

4.2 Challenges in estimating dose-response

The common approach to the above estimation problem is to run a clinical study where a cohort of patients is perhaps divided into multiple groups. Distinct groups could receive different fixed doses. By observing how disease conditions of patients in various groups evolve, one could estimate the parameter of an assumed dose-response model using regression. For instance, a meta-analysis of fifty such trials on nine different biologic agents for treating rheumatoid arthritis is provided in [125]. Demonty et al. [46] presented a meta-analysis of eighty four such clinical studies for characterizing the dose-response of LDL cholesterol levels to phytosterols. A meta-analysis of thirty similar clinical studies for characterizing the dose-response of blood pressure to diuretics is reported in [147]. One controversial aspect of such designs is that patients could be treated "sub-optimally" for the sake of gathering data: for example, some patients receiving a placebo could benefit from the drug; on the other hand, some patients in the high dose group might benefit from lowering their dose in order to decrease side effects. A cross-over design, where a patient receives a sequence of distinct doses also suffers from a similar limitation. An illuminating discussion of this and related issues, from philosophical and ethical viewpoints, is available in [113]. Ting [185] discussed the challenges in selecting doses in clinical trials from a statistical viewpoint.

In clinical studies for RGD, a cohort of patients could be divided into say two groups — one group receives some form of RGD whereas the other group receives conventional, one-

size-fits-all treatment. To the best of our knowledge, such trials have not used dose-response functions or explicit objective functions to guide their dosing strategies perhaps because the dose-response parameter is not known when the trial begins. Consequently, it is again difficult to rigorously claim that these dosing strategies are "optimal" for the cohort involved. Along similar lines, Murphy et al. [136] commented:

"*despite the activity in evaluating adaptive treatment strategies, the development of data collection and analytic methods that directly inform the construction of adaptive treatment strategies lags behind.*"

4.2.1 Emerging consensus about the need for learning dose-response

The aforementioned and related challenges/limitations have been documented, in a broad range of contexts, over at least the last two decades. A consensus seems to be emerging regarding the need for learning dose-response better. For instance, the 1994 FDA guidelines [60] stated:

"*historically, drugs have often been initially marketed at what were later recognized as excessive doses ... . This situation has been improved by attempts to find the smallest dose with a discernible useful effect or a maximum dose beyond which no further beneficial effect is seen, but practical study designs do not exist to allow for precise determination of these doses. Further, expanding knowledge indicates that the concepts of minimum effective dose and maximum useful dose do not adequately account for individual differences and do not allow a comparison, at various doses, of both beneficial and undesirable effects. Any given dose provides a mixture of desirable and undesirable effects, with no single dose necessarily optimal for all patients.*"

These guidelines further added:

"*in adjusting the dose in an individual patient after observing the response to an initial dose, what would be most helpful is knowledge of the shape of individual dose-response curves, which is usually not the same as the population (group) average dose-response curve. Study designs that allow estimation of individual dose-response curves could therefore be useful in guiding titration, although experience with such designs and their analysis is very limited.*"

In 2010, a working group on dose-ranging studies [148] observed:

"*improper dose regimen selection, due to lack of sufficient dose-response (DR) knowledge for both efficacy and safety, remains one of the key drivers of the high failure rate currently observed in clinical drug development. Several initiatives aimed at improving the efficiency and success rate of drug development programs, such as the US Food and Drug Administration (FDA) Critical Path Initiative and PhRMA's Pharmaceutical Innovation Steering Committee (PISC) Working Groups, have identified the need to improve dose selection and, more broadly, DR knowledge as one of their key issues.*"

The working group recommended:

"*it is clear from the working group's results that high-quality dose selection can only be achieved with a combination of increased investment in dose-response learning and an optimized selection of design and analysis strategies.*"

In 2014, Pinheiro et al. [149] noted:

"*finding the right dose is a critical step in pharmaceutical drug development. The problem of selecting the right dose or dose range occurs at almost every stage throughout the process of developing a new drug ... . In recent years, considerable effort has been spent on improving the efficiency of dose finding throughout drug development ... . Despite these*

efforts, however, improper dose selection for confirmatory studies, due to lack of sufficient dose-response knowledge for both efficacy and safety at the end of Phase II, remains a key driver of the ongoing pipeline problem experienced by the pharmaceutical industry ... ."

Most recently, in 2014, Laber et al. [108] concluded:

"*the increasing ability to collect, manipulate, and access patient-level data combined with a rapidly growing interest in personalized medicine among clinical scientists has created an unprecedented opportunity to improve the quality of healthcare using data. The potential gains of a data-driven treatment are especially high in the context of chronic illness in which the treatments must be adaptive to the uniquely evolving health status of each patient.*"

Motivated by these observations, we seek to answer the following specific research question: how can we find optimal response-guided doses for a cohort of patients *while* learning the parameter of an assumed dose-response function? We offer a Bayesian framework to this end.

4.2.2 Overview of our contributions

Modeling contributions

We propose a Bayesian stochastic dynamic programming (DP) formulation (see [105] for a classic survey of this area), where the system state corresponds to the numerical scores of the disease conditions of the cohort at the beginning of each treatment session. The decisions correspond to the doses administered to the cohort in each session. The immediate cost is given by a disutility function that models patients' aversion to doses. The disease conditions evolve according to a single-parameter dose-response function. The decision-maker assumes that, in each session, the population distribution of this parameter is Categorical, independently of everything else; recall that this is the most general discrete distribution with finite

support. The probability mass function (pmf) of this Categorical distribution is not known to the decision-maker. The decision-maker begins with a Dirichlet prior on this pmf. The Dirichlet hyperparameters form the information state in our Bayesian stochastic DP. Based on the observed evolution of the cohort's disease conditions, the decision-maker updates his prior belief over the treatment course. Recall that the Dirichlet distribution is conjugate to the Categorical pmf. This implies that the belief evolution is characterized by a simple update of the information state. The quality of the disease condition reached at the end of the treatment course is modeled via a disutility function for the cohort. The decision-maker's goal is to minimize the total expected disutility of the doses given to the cohort over the treatment course plus that of the disease conditions reached by the cohort at the end of the course.

Mathematical analyses, computational methods, and simulations

This Bayesian stochastic DP is high-dimensional and non-linear. The dimension of the state vector is equal to the size of the cohort plus the number of hyperparameters (which in this case equals the number of values the Categorical random variable can take). The dimension of the decision vector is equal to the size of the cohort. Since we expect the cohort-size to be of the order of a hundred, exact solution of this stochastic DP is computationally difficult. We therefore propose a semi-stochastic certainty equivalent control (CEC) approach to solve this Bayesian stochastic DP approximately. In this method, at the beginning of each session, the decision-maker attempts to choose a dose vector assuming that the Categorical pmf is proportional to the Dirichlet hyperparameters. This results in a clairvoyant stochastic DP that is still high-dimensional. We are able to show, under natural convexity and monotonicity assumptions on various functions in our model, that a monotone dosing policy is optimal in this clairvoyant stochastic DP (Proposition 4.5.3). According to this strategy, worse disease conditions for the cohort call for higher doses to the cohort. Unfortunately, however, although this monotonicity result is of independent theoretical interest and provides insights into our problem, it may not lead to sufficient computational savings in solving the

clairvoyant stochastic DP when the cohort is large. Fortunately, when the cohort's disutility functions are given by sums of its members' disutility functions, that is, additively separable, the high-dimensional clairvoyant stochastic DP decomposes across individual patients and hence can be solved efficiently (Proposition 4.5.4 and Corollary 4.5.5). When this additive separability does not hold, we resort to a deterministic version of CEC, where in each session, the parameter value is fixed at the Categorical mean implied by the current Dirichlet hyperparameters. The resulting decision problem is convex in dose vectors over the remaining treatment sessions (Proposition 4.5.6). We prove that it has a stationary optimal solution (Lemma 4.5.7 and Corollary 4.5.8). That is, it is optimal to implement an identical dose vector in all remaining sessions. This reduces problem-dimension and efficiently yields an optimal dose vector. We test our ideas via computer simulations on an example that uses the Michaelis-Menten dose-response function. Through these simulations, we investigate the effect of cohort size and prior misspecification, and also compare the performance of various dosing policies.

4.3 Literature review

To our knowledge, ours is the first attempt to provide a mathematical decision-making framework to adaptively learn the distribution of a dose-response parameter while optimally dosing a cohort in a response-guided trial. We include below a review of the existing relevant literature on adaptive clinical trials and on dynamic optimization in treatment planning. Papers are categorized into different subsections according to either the topic of their investigation or the methodology they employ.

4.3.1 Literature on adaptive trials for learning of dose-response

Essentially all existing literature on learning dose-response in an adaptive manner focuses on estimating the minimum effective dose (MED) or the maximum tolerable dose (MTD). MED and MTD estimation is done in early phase clinical trials so that the resulting dose-range could be employed in subsequent studies. The MED can be roughly defined as the dose that

produces a response that is more than a certain threshold above placebo [27]. The MTD is based on dose-response of toxicity and can be defined roughly as the dose that produces a pre-specified probability of toxicity [155]. Adaptive approaches to MTD estimation include the continual reassessment method [144], the escalation with overdose control method [15], the Bayesian decision theoretic approach [199], and the random walk rules method [51]. Variations and extensions of these methods are described in [16, 28, 65, 71, 86]. A recent comprehensive review of such adaptive methods is available in [83]. As we outlined in Section 4.2.1, the FDA has expressed some concerns about excessive reliance on the concepts of MED and MTD.

4.3.2 Broader literature on adaptive clinical trials

Our work can be viewed as belonging to the broad area of adaptive clinical trials. In contrast to traditional or "fixed" trials, adaptive clinical trials allow for pre-specified modifications to the trial's implementation based upon interim data. Adaptive trials have several advantages over fixed trials: they can provide the same information more efficiently through the use of a smaller patient cohort or shorter treatment schedule or they can reveal more detailed information on the treatment's effect by giving better estimates of the dose-response relationship [61]. Berry [24] commented that the Bayesian approach is "*an ideal tool for building adaptive designs.*" He also stated that disadvantages of adaptive trials include the complexities in designing, running and interpreting them; the resulting higher logistical demands they impose; difficulty in getting regulatory approval; limited funding availability from pharmaceutical companies; and risk to information security because "*modifications that occur during the trial may convey information outside the sphere of confidentiality ...* ." Some of these limitations were also noted by the FDA [61].

Adaptive clinical trials are not a new idea; in 1978, Berry [22] considered a scenario in which patients are treated one-by-one in one of two treatment schemes, with the chance of success or failure for each scheme gradually being learned using Bayesian techniques as more patient outcomes are observed. However, it was not until more recently that adaptive trials

have been implemented; the first drug to be approved by the FDA based on a Bayesian approach was Pravigard, a cholesterol-lowering drug, in 2003 [23].

Several aspects of an adaptive clinical trial can be altered based on accumulated data. For a more detailed description of the types of adaptive designs, see the review articles by Chow and Chang [41] and Berry [23, 24]. Below, we include a non-exhaustive illustrative list of various types of adaptive designs. Many adaptive trials amount to allocation problems, with patients being assigned to one of several treatment strategies (for example, high dose, low dose, and placebo). In adaptive randomization, patients begin treatment at staggered intervals, and the sequential allocation decision of a particular patient is based on the observed response of earlier patients [37, 58, 111]. In sample size re-estimation and drop-the-losers designs, the number of patients allocated to each treatment is allowed to vary during the course of treatment; either increasing by incorporating new patients to the trial, or by dropping existing patients [85, 168, 202]. In addition to adaptive allocation designs, other aspects of the trial can be adaptive. In group sequential design, the entire trial is allowed to end prematurely due to safety concerns, or a determination of futility or efficacy [68, 102]. In biomarker-adaptive design, adaptations are allowed based on patients' biomarkers such as genomic traits [64, 169]. In adaptive hypothesis design, changes to the trial hypothesis are allowed, for example, changing from a superiority hypothesis to a non-inferiority hypothesis.

4.3.3 Literature on multi-armed bandit problems in clinical trials

While most of the above literature appeared in medical statistics journals, the Operations Research community has recently shown interest, from an optimization viewpoint, in designing adaptive clinical trials. This includes the recent works on adaptive allocation by Ahuja and Birge [5] and Negoescu et al. [138]. These papers applied the classic theory of multi-armed bandit problems in discrete- and continuous-time, respectively, to response-adaptive trial designs (see [20, 26, 67, 92, 95, 96, 154]). Consequently, the problem statement, formulation, states, decisions, stochastic state transitions, reward structure, mathematical analyses, and computational approaches in these two papers are different from our work here.

In Ahuja and Birge, a discrete-time Bayesian adaptive Markov decision process model was presented for allocating patients to a finite number of treatments. A constant number n of patients is allocated in each time-period. The state corresponds to the cumulative number of observed patients to date receiving a specific treatment and reaching a particular health outcome. The set of health outcomes is finite. Patient allocations are probabilistic; so, for each one of the n patients, the controls correspond to the probabilities of assigning that patient to various treatments. They derived various insightful properties of optimal allocations and compared their approach with allocation rules proposed earlier in Berry [22]. Results were also applied to simulating a stent trial.

Negoescu et al. considered a continuous-time framework, where the decision-maker allocates the current time-interval $[t, t + dt)$ between two arms. Any fractional split between arms is allowed. One of the arms is considered safe and the other one risky. The safe arm brings instantaneous Brownian rewards and induces life events from a Poisson process. The risky arm can belong to one of two types — good or bad. The risky arm also brings Brownian rewards and induces life events according to a Poisson process; the rewards and the rate of this Poisson process depend on the type of the arm, which is unknown to the decision-maker. Life events also fetch lump-sum rewards. At time zero, the decision-maker starts with an initial belief on the type of the risky arm, which is updated as time wears on. The goal is to find an allocation policy that maximizes the total expected discounted reward over a finite planning-horizon. A particularly appealing feature of this work was that it provided a complete analytical characterization of an optimal allocation policy. Simulation results for multiple sclerosis were presented.

4.3.4 Literature on dynamic optimization for adaptive treatment strategies

Finally, there is a large and rapidly expanding body of work on fully/partially observable stochastic DP models for adaptive treatment strategies. Surveys of such models are available in [6, 158]. Murphy and co-authors [135, 136] outlined an abstract stochastic DP/reinforcement learning framework for designing adaptive treatment strategies; they ap-

plied this approach to a simulation study of adaptive interventions in conduct disorders and drug abuse in children. The idea of adapting treatment or diagnostic decisions to the observed stochastic evolution of a "health state" according to a natural history progression model without intervention was recently employed in [7, 8, 40, 48, 107, 123, 156, 165, 201]. The decisions in these papers are of the wait/do not wait-type, and do not involve a sequential choice of dose levels. The decision process in such models ends when a choice to not wait is made. A control-limit policy is typically found to be optimal.

The work in [97] is an exception to this theme of wait/do not wait decisions. There, the authors made dose level (high/low) decisions for response-guided radiotherapy; in some cases, they numerically found on small-scale examples that the optimal policy had a control-limit structure. This model was extended to intensity modulated radiation therapy in [98]. A sequence of treatment modalities was optimized for metastatic tumors in [18, 19] and for AIDS in [164]. Mason et al. [131] presented a bi-criteria Markov decision process formulation for coordinated management of coexisting risk factors due to blood pressure and cholesterol in type-2 diabetes patients. Helm et al. [78] presented a dynamic linear Gaussian systems approach to optimize the timing of periodic examination of glaucoma patients. Continuous-time deterministic control methods, which utilize differential equations to model viral dynamics, have been used to derive dosing strategies for AIDS [109, 197]. None of these papers optimize doses to a cohort of patients while learning parameters of dose-response functions in a clinical trial framework.

Some researchers have proposed stochastic compartment models to make dosing decisions. These models discretize the human body into a finite number of chambers and use stochastic differential-algebraic equations to model the transport of a drug through these chambers. Continuous-time stochastic control techniques are then applied to maintain drug concentrations in blood plasma inside these compartments. Examples include Hu et al. [81], Bayard et al. [17], Jelliffe et al. [89], Acikgoz and Diwekar [2], Schumitzky [162], and references therein. However, these models do not consider the stochastic progression of disease condition, and hence do not implement RGD as we envision here. Most importantly, again,

they do not optimize doses to a cohort of patients while learning dose-response parameters in a clinical trial setting.

4.4 Model

Our model here is an extension of our previous stochastic DP model for RGD of Chapter 2. The earlier model did not consider a cohort of patients, instead focusing on treatment for a single patient, and assumed that the pmf of the dose-response parameter was known to the decision-maker at the beginning of the treatment course. That is, response-learning was not incorporated. Here, we propose a stochastic DP formulation that attempts to optimally learn this pmf during a clinical trial on a cohort of patients.

We consider a cohort with n patients indexed by $i = 1, 2, ..., n$. The dose-response of the cohort is assumed to be homogeneous and patients respond to treatment independently of one another. Treatment occurs in T sessions indexed by $t = 1, 2, ..., T$. The state variable $x_{t,i} \in X \subseteq \mathbb{R}$ represents the disease condition of patient i at the beginning of session t; larger states correspond to worse disease conditions. We represent the disease state of the whole cohort as the vector $\vec{x}_t \triangleq (x_{t,1}, x_{t,2}, ..., x_{t,n}) \in \mathbb{R}^n$. The decision variable $d_{t,i} \in D \triangleq [0, \bar{d}] \subset \mathbb{R}$ represents the dose to be administered to patient i in session t; we represent the dose administered to the whole cohort $\vec{d_t} \triangleq (d_{t,1}, d_{t,2}, ..., d_{t,n}) \in D^n$. Here, $\bar{d}$ is the largest permissible dose and we use the shorthand $D^n \triangleq \underbrace{[0, \bar{d}] \times [0, \bar{d}] \times ... \times [0, \bar{d}]}_{n \text{ times}}$.

The dose-response function for each patient is taken to be $x_{t+1,i} = f(x_{t,i}, d_{t,i}; \Theta)$; here Θ is a nonnegative dose-response parameter. In this chapter, as in Chapter 3, we again assume that this dose-response function can be expressed in the form

$$x_{t+1,i} = x_{t,i} + f_0(d_{t,i}; \Theta). \tag{4.1}$$

We remind the reader that essentially all standard dose-response functions such as linear, linear-quadratic, exponential, exponential linear-quadratic, Michaelis-Menten, Hill's, Emax, power law, beta-Poisson, and logistic can be expressed in this form (after taking logarithms

in some cases) as shown in Chapter 2. This assumption about the structure of the dose-response function is merely for notational convenience in some of our subsequent proofs. We will often write the dose-response of the entire cohort using the vector versions $\vec{f}$ and $\vec{f_0}$ of these functions.

We view the parameter Θ as a Categorical random variable that takes k possible values listed in the finite set $\Omega \triangleq \{v_1, v_2, \ldots, v_k\}$. The pmf $\vec{p} \triangleq (p_1, p_2, \ldots, p_k) \in \mathbb{R}^k_{++}$ of this Categorical random variable is not known to the decision-maker at the beginning of the trial. Here, we have used p_j, for $j = 1, 2, \ldots, k$, to denote the probability that Θ takes the value v_j. The decision-maker assumes a Dirichlet prior with hyperparameters $\vec{a}_1 \triangleq (a_{1,1}, a_{1,2}, \ldots, a_{1,k}) \in \mathbb{R}^k_{++}$ for this Categorical pmf at the beginning of the trial. That is, at the beginning of the trial, the prior probability density function of the Categorical pmf $\vec{p}$ is given by

$$g_1(\vec{p}, \vec{a}_1) \triangleq \frac{1}{B(\vec{a}_1)} \prod_{j=1}^{k} (p_j)^{a_{1,j}-1}, \tag{4.2}$$

where $B(\vec{a}_1)$ is the multinomial Beta function that is expressed in terms of Gamma functions as

$$B(\vec{a}_1) \triangleq \frac{\prod_{j=1}^{k} \Gamma(a_{1,j})}{\Gamma(\sum_{j=1}^{k} a_{1,j})}. \tag{4.3}$$

Recall here that the Gamma function is defined by $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt$ [1].

We use $\theta_{t,i}$, for $i = 1, 2, ..., n$, to denote the independent and identically distributed realization of the random variable Θ for patient i in session t. Note that the realizations $\theta_{t,i}$ are not observed directly, but rather can be calculated from our disease dynamics given $\vec{x}_t$, $\vec{d}_t$, and after observing $\vec{x}_{t+1}$. We think of this calculation procedure, perhaps with some abuse of notation, as $\theta_{t,i} = f^{-1}(x_{t,i}, d_{t,i}, x_{t+1,i})$. Recall that the Dirichlet prior is conjugate to the Categorical pmf. This implies that after calculating the realizations $\theta_{t,i}$, for $i = 1, 2, ..., n$,

the posterior distribution remains Dirichlet with its hyperparameters updated according to

$$a_{t+1,j} \leftarrow a_{t,j} + \sum_{i=1}^{n} I_j(\theta_{t,i}, v_j),\ j = 1, 2, \ldots, n. \tag{4.4}$$

Here, $I_j(\theta_{t,i}, v_j)$ is the indicator function that equals one when $\theta_{t,i} = v_j$ and equals zero otherwise. We compactly denote this "information update" step via a function ϕ as $\vec{a}_{t+1} = \phi(n; \vec{a}_t; \vec{d}_t; \vec{x}_t, \vec{x}_{t+1})$. This property of the Categorical-Dirichlet pair enables us to use $\vec{a}_t$ as the information state in our Bayesian stochastic DP. We assume that for any fixed value $\theta \in \Omega$ of the dose-response parameter, the function $f_0(\cdot; \theta)$ is continuous over D. We will often use $\vec{\Theta}$ to denote the random vector of dose-response parameters for the cohort in one session, and similarly, $\vec{\theta}$ to denote a realization of this random vector in that session.

The cohort's aversion to the adverse effects of dose in each session is modeled using a continuous, nonnegative disutility function $c : D^n \to \mathbb{R}_+$. Since D^n is compact, $c(\cdot)$ is bounded. Several examples of dose-disutility functions $c_i(d_{t,i})$ such as linear, linear-quadratic, exponential for individual patients were discussed in Chapter 2. Here, the cohort's disutility is an appropriate composition of these individual functions: for example, the cohort's disutility could be the average of individual disutilities; or the cohort's disutility could be the worst, that is, the largest, among all individual disutilities. The cohort's aversion to disease conditions $\vec{x}_{T+1}$ at the end of the trial is modeled using a continuous, nonnegative, and bounded disutility function $h : X^n \to \mathbb{R}_+$, where $X^n \triangleq \underbrace{X \times X \times ... \times X}_{n \text{ times}}$. Several examples of disutility functions $h_i(x_{t,i})$ such as linear, ramp, linear-quadratic, exponential, and logarithmic for individual patients were discussed in Chapter 2. The cohort's disutility $h(\cdot)$ is again an appropriate composition of these individual functions. The decision-maker's goal is to minimize the total expected disutility of the cohort over the clinical trial while learning, in a Bayesian manner, the pmf of the dose-response parameter.

Bellman's equations for this Bayesian stochastic DP can be written as follows. Here, for convenience and clarity of exposition, we use two optimal cost-to-go functions: $J_t(\cdot)$ are the "post-information-update" optimal cost-to-go functions for $t = 1, 2, ..., T+1$; and $V_t(\cdot)$ are the "pre-information-update" cost-to-go functions for $t = 2, 3, ..., T+1$. Then, for

$t = 1, 2, \ldots, T$, we have,

$$J_t(\vec{a}_t; \vec{x}_t) = \min_{\vec{d}_t \in D^n} \Big[c(\vec{d}_t) + E_{\vec{a}_t}[V_{t+1}(\vec{a}_t; \vec{d}_t; \vec{x}_t, \vec{x}_{t+1})] \Big], \tag{4.5}$$

$$x_{t+1,i} = f(x_{t,i}, d_{t,i}; \Theta), \ i = 1, 2, \ldots, n, \tag{4.6}$$

$$V_{t+1}(\vec{a}_t; \vec{d}_t; \vec{x}_t, \vec{x}_{t+1}) = J_{t+1}(\vec{a}_{t+1}; \vec{x}_{t+1}), \tag{4.7}$$

$$\vec{a}_{t+1} = \phi(n; \vec{a}_t; \vec{d}_t; \vec{x}_t, \vec{x}_{t+1}), \tag{4.8}$$

with boundary condition

$$J_{T+1}(\vec{a}_{T+1}; \vec{x}_{T+1}) = h(\vec{x}_{T+1}). \tag{4.9}$$

Here, we have included the subscript $\vec{a}_t$ on the expectation operator E to emphasize that the expectation of the cost-to-go function on the right hand side of (4.5) is taken with respect to the hyperparameters of the decision-maker's current Dirichlet belief.

Exact solution of these Bellman's equations is computationally intractable. We shed more light on this issue in the next section and then provide computational methods for approximate solution. We use the words increasing and decreasing in the weak sense to mean nondecreasing and nonincreasing, respectively, in this chapter. All comparisons between vectors are made componentwise.

4.5 Computational methods for approximate solution

The state and decision vectors in the above Bayesian stochastic DP are high-dimensional and continuous. Thus, a standard implementation of Bellman's backward recursion algorithm is not implementable as it would require that the minimization in (4.5) be performed for an *uncountable* number of states. Thus a state discretization would be necessary, which is again computationally impractical because of the high-dimension of the state. Moreover, even if such a state discretization were possible, each minimization in a discrete version of the problem would require the solution of a non-linear stochastic programming problem which itself is impractical due to the curse of action-space dimensionality. For these reasons, we pursue two approximate solution methods: semi-stochastic certainty equivalent control

and certainty equivalent control (see [25]). We analyze various properties of the resulting stochastic and deterministic optimization problems in the next few sections.

4.5.1 Semi-stochastic certainty equivalent control

Our Bayesian stochastic DP includes two levels of uncertainty. The first is rooted in the fact that the decision-maker does not know the pmf of the dose-response parameter. The second, and this uncertainty would be present even if the decision-maker knew this pmf, is in the evolution of the cohort's disease conditions. The former is thought of as the "higher level" uncertainty and the latter as the "lower level" uncertainty. The idea in semi-stochastic CEC is to suppress the higher level uncertainty, focusing only on the lower level one. This is usually done by replacing the values of the higher level random variables by their expected values.

Specifically, we implement semi-stochastic CEC as follows. When the state is $(\vec{x}_t, \vec{a}_t)$ at the beginning of the tth period of the trial, the decision-maker chooses doses for that period assuming that the components of the Categorical pmf $\vec{p}$ equal the expected values of the corresponding Dirichlet components and that there will be no further learning in the remaining periods. That is, doses are chosen in period t assuming that

$$p_j = \frac{a_{t,j}}{\sum_{j=1}^{k} a_{t,j}}, \ j = 1, 2, \ldots, n. \tag{4.10}$$

In other words, decisions are made in period t by solving the "clairvoyant" Bellman's equations

$$J_t(\vec{x}_s) = \min_{\vec{d}_s \in D^n} \Big[c(\vec{d}_s) + E J_{s+1}(\vec{f}(\vec{x}_s, \vec{d}_s; \vec{\Theta})) \Big], \ s = t, t+1, \ldots, T, \tag{4.11}$$

with boundary condition

$$J_{T+1}(\vec{x}_{T+1}) = h(\vec{x}_{T+1}). \tag{4.12}$$

Here, with some abuse of notation, we continue to use $J_t(\cdot)$ to denote the optimal cost-to-go function for the clairvoyant problem where the pmf of Θ is fixed as given by (4.10).

In particular, the expected value on the right hand side of (4.11) is taken with respect to this fixed pmf. After administering doses as prescribed by an optimal policy to this clairvoyant stochastic DP in the current period, the state evolves stochastically and the process is repeated until the trial is completed. Specifically, a sequence of T clairvoyant stochastic DPs are solved: the first one includes T periods, the second one includes $T-1$ periods, and ultimately, the last one includes a single period. This algorithm is summarized below.

Semi-stochastic certainty equivalent control

INITIALIZE: Set $t = 1$ and begin with a given initial state $(\vec{x}_1, \vec{a}_1)$.
DO WHILE $t \leq T$,

- **Step 1.** let the state at the beginning of session t be $(\vec{x}_t, \vec{a}_t)$;
- **Step 2.** fix the pmf $\vec{p}$ of the dose-response parameters for all patients and all remaining sessions as given in Equation (4.10);
- **Step 3.** obtain an optimal dosing policy for the resulting clairvoyant stochastic DP over periods $t, t+1, \ldots, T$ by solving Bellman's equations (4.11)-(4.12);
- **Step 4.** administer the dose vector $\vec{d}_t^*$ prescribed by this policy to the cohort in session t;
- **Step 5.** observe the cohort's disease state $\vec{x}_{t+1}$ at the beginning of the next session, and calculate the implied realizations $\theta_{t,i}$ of the dose-response parameters of all patients using $\theta_{t,i} = f^{-1}(x_{t,i}, d_{t,i}^*, x_{t+1,i})$; then update the information state using $\vec{a}_{t+1} = \phi(n; \vec{a}_t; \vec{d}_t^*; \vec{x}_t, \vec{x}_{t+1})$;
- **Step 6.** update $t \leftarrow t+1$ and go to Step 1 above.

END DO

Note, however, that the clairvoyant stochastic DP is still computationally expensive because $\vec{x}_t$ and $\vec{d}_t$ are high-dimensional. In the next two sections we establish, under two different sets of assumptions, two key properties of optimal policies for this clairvoyant stochastic DP. These properties provide insight into the corresponding dosing decisions and simplify our solution procedure.

Monotonicity of optimal dosing policy under convexity assumptions

We claim, under two assumptions on the dose-response function and the disutility functions, that in each session, there exist optimal dose vectors that increase as the disease condition worsens.

Assumption 4.5.1 (Monotone and convex disease-response)**.** *For each fixed value $\theta \in \Omega$ of the dose-response parameter:*

A. *the function $f_0(\cdot;\theta)$ is decreasing in dose; and*

B. *the function $f_0(\cdot;\theta)$ is convex over D.*

Assumption 4.5.2 (Increasing and convex disutilities)**.** *The disutility functions satisfy:*

A. *$c(\cdot)$ is increasing and convex over D^n; and*

B. *$h(\cdot)$ is increasing and convex over X^n.*

Assumption 4.5.1A is natural as it embodies the intuition that for a fixed pre-session disease condition, a higher dose results in a lower (hence, better) post-session disease condition. This monotonicity of dose-response is a standard assumption in the literature [155, 157], and holds for all aforementioned dose-response functions. The convexity property in Assumption

4.5.1B means that the magnitude of marginal improvement in disease condition decreases with higher doses. Assumption 4.5.2A holds when there is an increasing aversion to higher doses and the marginal aversion increases with dose; Assumption 4.5.2B holds when there is an increasing aversion to worsening disease conditions and the marginal aversion increases with worsening disease conditions. The decision-maker's preference to lower doses and better disease conditions is a natural and standard assumption; increasing marginal aversion is a characteristic of risk averse decision-makers in medicine and this seems especially appropriate in the clinical trials context [52]. We believe that Assumption 4.5.1B is the only somewhat restrictive assumption because we are aware of at least one dose-response function (the aforementioned exponential linear-qudratic function from radiobiology) where convexity fails to hold. Unfortunately, however, it does not appear possible to drop this assumption and still hope for monotonicity to hold. In fact, in the clairvoyant single-patient model of Chapter 2, we were able to construct a counterexample where this convexity of dose-response is violated and all our other assumptions hold but there is no monotone dosing policy that is optimal.

Proposition 4.5.3. *Under the above assumptions, optimal doses increase with worsening disease conditions in each session. More precisely, in any session t, if dose vector $\vec{a}(\vec{x}_t)$ is optimal for some $\vec{x}_t \in X^n$, then, corresponding to every $\vec{u}_t \in X^n$ with $\vec{u}_t \geq \vec{x}_t$, there exists an optimal dose level $\vec{b}(\vec{u}_t)$ such that $\vec{b}(\vec{u}_t) \geq \vec{a}(\vec{x}_t)$.*

Proof. The proof employs backward induction on $t = T+1, T, \ldots, 1$. We first show that if $J_{t+1}(\cdot)$ is increasing, convex, continuous, and bounded, then in treatment session t, there exist optimal decisions with the monotone structure described in Proposition 4.5.3. We then show that $J_t(\cdot)$ inherits these four properties from $J_{t+1}(\cdot)$. Proposition 4.5.3 then follows because the disutility function $h(\cdot)$, that is, $J_{T+1}(\cdot)$, is assumed to possess these properties.

Define $Q_t(\vec{x}_t; \vec{d}_t) \triangleq c(\vec{d}_t) + EJ_{t+1}(\vec{f}(\vec{x}_t, \vec{d}_t; \vec{\Theta}))$ so that $J_t(\vec{x}_t) = \min_{\vec{d}_t \in D^n} Q_t(\vec{x}_t; \vec{d}_t)$. Monotonicity would follow if $Q_t(\cdot;\cdot)$ had decreasing differences as characterized by the inequality

$$Q_t(\vec{x}_t^+; \vec{d}_t^+) - Q_t(\vec{x}_t^+; \vec{d}_t^-) \leq Q_t(\vec{x}_t^-; \vec{d}_t^+) - Q_t(\vec{x}_t^-; \vec{d}_t^-) \tag{4.13}$$

for all $\vec{x}_t^+, \vec{x}_t^- \in X^n$ such that $\vec{x}_t^+ \geq \vec{x}_t^-$ and for all $\vec{d}_t^+, \vec{d}_t^- \in D^n$ such that $\vec{d}_t^+ \geq \vec{d}_t^-$. To see this, suppose, in order to obtain a contradiction, not. That is, there are two states, which we denote $\vec{x}_t^+$ and $\vec{x}_t^-$ where $\vec{x}_t^+ \geq \vec{x}_t^-$ with the following property in session t: dose $\vec{d}_t^+ \in D^n$ is optimal in $\vec{x}_t^-$ and every dose $\vec{d}_t^- \in D^n$ that is optimal in $\vec{x}_t^+$ satisfies $\vec{d}_t^- < \vec{d}_t^+$. Then, $Q_t(\vec{x}_t^-; \vec{d}_t^+) \leq Q_t(\vec{x}_t^-; \vec{d}_t^-)$ by optimality of $\vec{d}_t^+$ in $\vec{x}_t^-$; and $Q_t(\vec{x}_t^+; \vec{d}_t^-) < Q_t(\vec{x}_t^+; \vec{d}_t^+)$ because $\vec{d}_t^-$ is optimal and $\vec{d}_t^+$ is not optimal in $\vec{x}_t^+$. Adding these inequalities yields a contradiction to the decreasing differences property of $Q_t(\cdot;\cdot)$ as in (4.13).

Now we show that $Q_t(\cdot;\cdot)$ has decreasing differences as in (4.13). Observe from the definition of $Q_t(\cdot;\cdot)$ that

$$\begin{aligned}
&Q_t(\vec{x}_t^+; \vec{d}_t^+) - Q_t(\vec{x}_t^+; \vec{d}_t^-) - Q_t(\vec{x}_t^-; \vec{d}_t^+) + Q_t(\vec{x}_t^-; \vec{d}_t^-) \\
&= E\Big[J_{t+1}(\vec{f}(\vec{x}_t^+, \vec{d}_t^+; \vec{\Theta})) - J_{t+1}(\vec{f}(\vec{x}_t^+, \vec{d}_t^-; \vec{\Theta})) - J_{t+1}(\vec{f}(\vec{x}_t^-, \vec{d}_t^+; \vec{\Theta})) + J_{t+1}(\vec{f}(\vec{x}_t^-, \vec{d}_t^-; \vec{\Theta}))\Big].
\end{aligned}$$

Therefore, it suffices to show that the quantity in square brackets on the right hand side above is nonpositive for every realization $\vec{\theta}$ of the random vector $\vec{\Theta}$. Toward this end, we fix some realization $\vec{\theta}$ and Equation (4.1) gives:

$$\vec{f}(\vec{x}_t^+, \vec{d}_t^+; \vec{\theta}) - \vec{f}(\vec{x}_t^+, \vec{d}_t^-; \vec{\theta}) = \vec{f}(\vec{x}_t^-, \vec{d}_t^+; \vec{\theta}) - \vec{f}(\vec{x}_t^-, \vec{d}_t^-; \vec{\theta}). \tag{4.14}$$

Moreover, Equation (4.1) and Assumption 4.5.1A imply that

$$\begin{aligned}
\vec{f}(\vec{x}_t^-, \vec{d}_t^+; \vec{\theta}) \leq \vec{f}(\vec{x}_t^+, \vec{d}_t^+; \vec{\theta}) \leq \vec{f}(\vec{x}_t^+, \vec{d}_t^-; \vec{\theta}) \\
\vec{f}(\vec{x}_t^-, \vec{d}_t^+; \vec{\theta}) \leq \vec{f}(\vec{x}_t^-, \vec{d}_t^-; \vec{\theta}) \leq \vec{f}(\vec{x}_t^+, \vec{d}_t^-; \vec{\theta}).
\end{aligned}$$

This set of inequalities, when combined with (4.14) and with the inductive assumption that $J_{t+1}(\cdot)$ is increasing and convex, leads to

$$J_{t+1}(\vec{f}(\vec{x}_t^+, \vec{d}_t^+; \vec{\theta})) - J_{t+1}(\vec{f}(\vec{x}_t^+, \vec{d}_t^-; \vec{\theta})) - J_{t+1}(\vec{f}(\vec{x}_t^-, \vec{d}_t^+; \vec{\theta})) + J_{t+1}(\vec{f}(\vec{x}_t^-, \vec{d}_t^-; \vec{\theta})) \leq 0$$

as required.

Suppose dose $\vec{d}_t^+$ is optimal in $\vec{x}_t^+$ and $\vec{d}_t^-$ is optimal in $\vec{x}_t^-$ in session t. Then, we have,

$$J_t(\vec{x}_t^+) - J_t(\vec{x}_t^-) = Q_t(\vec{x}_t^+; \vec{d}_t^+) - Q_t(\vec{x}_t^-; \vec{d}_t^-) \geq Q_t(\vec{x}_t^+; \vec{d}_t^+) - Q_t(\vec{x}_t^-; \vec{d}_t^+)$$

$$= E\Big[J_{t+1}(\vec{f}(\vec{x}_t^+, \vec{d}_t^+; \vec{\Theta})) - J_{t+1}(\vec{f}(\vec{x}_t^-, \vec{d}_t^+; \vec{\Theta}))\Big] \geq 0.$$

Thus, $J_t(\cdot)$ is increasing. Here, the first inequality follows because $\vec{d}_t^-$ is optimal in $\vec{x}_t^-$, and the second inequality holds because the quantity inside the square brackets of the final expression is nonnegative for every realization of the random vector $\vec{\Theta}$.

In order to show that $J_t(\cdot)$ is convex, we first show that $Q_t(\cdot;\cdot)$ is convex. Let $\vec{x}_t^\circ = \lambda\vec{x}_t^- + (1-\lambda)\vec{x}_t^+$ and $\vec{d}_t^\circ = \lambda\vec{d}_t^- + (1-\lambda)\vec{d}_t^+$ with $\lambda \in [0,1]$, $\vec{x}_t^-, \vec{x}_t^+ \in X^n$, and $\vec{d}_t^-, \vec{d}_t^+ \in D^n$. We have,

$$\begin{aligned}
Q_t(\vec{x}_t^\circ, \vec{d}_t^\circ) &= c(\vec{d}_t^\circ) + EJ_{t+1}(\vec{f}(\vec{x}_t^\circ, \vec{d}_t^\circ; \vec{\Theta})) \leq c(\vec{d}_t^\circ) + EJ_{t+1}(\lambda\vec{f}(\vec{x}_t^-, \vec{d}_t^-; \vec{\Theta}) + (1-\lambda)\vec{f}(\vec{x}_t^+, \vec{d}_t^+; \vec{\Theta})) \\
&\leq c(\vec{d}_t^\circ) + E\Big[\lambda J_{t+1}(\vec{f}(\vec{x}_t^-, \vec{d}_t^-; \vec{\Theta})) + (1-\lambda)J_{t+1}(\vec{f}(\vec{x}_t^+, \vec{d}_t^+; \vec{\Theta}))\Big] \\
&\leq \lambda c(\vec{d}_t^-) + (1-\lambda)c(\vec{d}_t^+) + E\Big[\lambda J_{t+1}(\vec{f}(\vec{x}_t^-, \vec{d}_t^-; \vec{\Theta})) + (1-\lambda)J_{t+1}(\vec{f}(\vec{x}_t^+, \vec{d}_t^+; \vec{\Theta}))\Big] \\
&= \lambda Q_t(\vec{x}_t^-; \vec{d}_t^-) + (1-\lambda)Q_t(\vec{x}_t^+; \vec{d}_t^+).
\end{aligned}$$

Thus, $Q_t(\cdot;\cdot)$ is convex. Here, the first inequality holds because $\vec{f}(\cdot,\cdot;\vec{\theta})$ is convex for each realization $\vec{\theta}$ of the random vector $\vec{\Theta}$ (this can be easily shown to hold by Assumption 4.5.1B) and $J_{t+1}(\cdot)$ is increasing. The second inequality follows because $J_{t+1}(\cdot)$ is convex. The third inequality holds because $c(\cdot)$ is convex.

We now show that $J_t(\cdot)$ is convex.

$$\begin{aligned}
J_t(\vec{x}_t^\circ) &= \min_{\vec{d}\in D} Q_t(\vec{x}_t^\circ; \vec{d}) \leq Q_t(\vec{x}_t^\circ; \vec{d}_t^\circ) \leq \lambda Q_t(\vec{x}_t^-; \vec{d}_t^-) + (1-\lambda)Q_t(\vec{x}_t^+; \vec{d}_t^+) \\
&= \lambda J_t(\vec{x}_t^-) + (1-\lambda)J_t(\vec{x}_t^+).
\end{aligned}$$

Here, the first inequality holds because $\vec{d}_t^\circ$ is feasible and the second inequality follows because $Q_t(\cdot,\cdot)$ is convex.

Finally, continuity and boundedness of $J_t(\cdot)$ can be established by an argument similar to Chapter 2. That argument is omitted here for brevity. This completes the proof by backward induction. □

This monotonicity property provides structural insights into our clairvoyant dosing policy for the cohort. It is also of independent theoretical interest. In the next section, under a

different set of assumptions, we prove another property of our clairvoyant dosing policy that achieves significant computational savings.

Problem decomposition under additively separable disutilities

One possible way to compose the cohort's disutility functions $c(\cdot)$ and $h(\cdot)$ is by adding the individual patient's disutilities $c_i(\cdot)$ and $h_i(\cdot)$, respectively. That is,

$$c(\vec{d}_t) = \sum_{i=1}^{n} c_i(d_{t,i}), \tag{4.15}$$

$$h(\vec{x}_{T+1}) = \sum_{i=1}^{n} h_i(x_{T+1,i}). \tag{4.16}$$

We then have

Proposition 4.5.4. *Suppose the cohort's disutility functions are additively separable as in (4.15)-(4.16). Then, for each $t = 1, 2, \ldots, T+1$, the optimal cost-to-go functions $J_t(\cdot)$ in the clairvoyant stochastic DP (4.11)-(4.12) are given by $J_t(\vec{x}_t) = \sum_{i=1}^{n} J_{t,i}(x_{t,i})$, where the individual optimal cost-to-go functions $J_{t,i}(\cdot)$ are obtained by solving the clairvoyant individual Bellman's equations*

$$J_{t,i}(x_{t,i}) = \min_{d_{t,i} \in D} \Big[c_i(d_{t,i}) + E J_{t+1,i}(f(x_{t,i}, d_{t,i}; \Theta)) \Big], \tag{4.17}$$

with boundary condition

$$J_{T+1,i}(x_{T+1,i}) = h_i(x_{T+1,i}). \tag{4.18}$$

Specifically, an optimal dosing policy for the cohort in Bellman's equations (4.11)-(4.12) obtained by collating the individual optimal dosing policies in Bellman's equations (4.17)-(4.18).

Proof. We provide a proof by backward induction. We start at the boundary $t = T$. We know from (4.12) and (4.16) that the claim is trivially true at this boundary. Now suppose the claim is true for some $t + 1$. We show that the claim is true for t. We have, from

the clairvoyant Bellman's equations, the inductive hypothesis, linearity of expectation, and (4.15) that

$$J_t(\vec{x}_t) = \min_{\vec{d}_t \in D^n} \Big[\sum_{i=1}^{n} c(d_{t,i}) + \sum_{i=1}^{n} EJ_{t+1,i}(x_{t+1,i}) \Big] = \sum_{i=1}^{n} \min_{d_{t,i} \in D} \Big(c(d_{t,i}) + EJ_{t+1,i}(x_{t+1,i}) \Big) \tag{4.19}$$

$$= \sum_{i=1}^{n} J_{t,i}(x_{t,i}). \tag{4.20}$$

This proves the claim. □

Corollary 4.5.5. *If the individual disutilies are identical across the cohort, that is, the functions $c_i(\cdot)$ and $h_i(\cdot)$ do not depend on i, then an optimal policy for the cohort's clairvoyant stochastic DP can be obtained simply by solving a clairvoyant stochastic DP for (any) one patient.*

Proof. Follows immediately from Proposition 4.5.4. □

The assumption that disutilities are invariant across patients is natural because the cohort is assumed to be homogeneous. We emphasize that Proposition 4.5.4 and Corollary 4.5.5 do not need any convexity or monotonicity assumptions on the various functions in our model. This corollary significantly simplifies the implementation of semi-stochastic CEC because, in period t, optimal doses for the cohort can be found simply by solving a one-dimensional clairvoyant stochastic DP instead of solving an n-dimensional clairvoyant stochastic DP. The Bellman's equations for the one-dimensional clairvoyant stochastic DP can be approximately solved efficiently by discretization as described, with or without convexity assumptions, in detail in Chapter 2.

4.5.2 Certainty equivalent control

CEC works by suppressing, in addition to the higher level uncertainty, the lower level uncertainty as well. Specifically, in period t, when the state is $(\vec{x}_t, \vec{a}_t)$, decisions are made by assuming that the pmf $\vec{p}$ of the unknown dose-response parameter Θ is given as in (4.10)

and that the dose-response parameter (for all patients and for all remaining sessions) equals the expected value implied by this pmf. That is, the dose-response parameter is given by

$$\omega \triangleq \sum_{j=1}^{k} v_j \left(a_{t,j} / \sum_{j=1}^{k} a_{t,j} \right). \tag{4.21}$$

This leads to a deterministic control problem, which is then solved to obtain an optimal sequence of doses for the cohort for all remaining periods. However, only the doses for the first period from this sequence are administered to the cohort while the others are discarded. The state then evolves stochastically and the process is repeated until the trial is completed. Thus, in one complete run of CEC, a total of T deterministic control problems are solved; the first one includes T periods, the second includes $T-1$ periods, and ultimately, the last one involves only one period. This algorithm is summarized below.

Certainty equivalent control

INITIALIZE: Set $t = 1$ and begin with a given initial state $(\vec{x}_1, \vec{a}_1)$.
DO WHILE $t \leq T$,

- **Step 1.** let the state at the beginning of session t be $(\vec{x}_t, \vec{a}_t)$;

- **Step 2.** fix the dose-response parameters for all patients and all remaining sessions at w as given in Equation (4.21); use $\vec{w} \triangleq \underbrace{(\omega, \omega, \ldots, \omega)}_{n \text{ times}}$ as the vector of parameters for the cohort in each of the remaining sessions;

- **Step 3.** solve the deterministic optimization problem

$$\min \sum_{s=t}^{T} c(\vec{d}_s) + h\big(\vec{x}_t + \sum_{s=t}^{T} \vec{f}_0(\vec{d}_s, \vec{\omega})\big) \tag{4.22}$$

subject to

$$\vec{d}_s \in D^n, \ s = t, t+1, \ldots, T, \tag{4.23}$$

to obtain an optimal sequence of dose vectors $\vec{d}_t^*, \vec{d}_{t+1}^*, \ldots, \vec{d}_T^*$;

- **Step 4.** discard $\vec{d}^*_{t+1}, \ldots, \vec{d}^*_T$ and administer the dose vector $\vec{d}^*_t$ to the cohort in session t;

- **Step 5.** observe the cohort's disease state $\vec{x}_{t+1}$ at the beginning of the next session, and calculate the implied realizations $\theta_{t,i}$ of the dose-response parameters of all patients using $\theta_{t,i} = f^{-1}(x_{t,i}, d^*_{t,i}, x_{t+1,i})$; then update the information state using $\vec{a}_{t+1} = \phi(n; \vec{a}_t; \vec{d}^*_t; \vec{x}_t, \vec{x}_{t+1})$;

- **Step 6.** update $t \leftarrow t + 1$ and go to Step 1 above.

END DO

The key step in CEC thus involves solving the deterministic optimization problem (4.22)-(4.23). The objective function in this problem is continuous and the feasible region is compact; so this problem has an optimal solution. It is a non-linear problem in decision vectors $\vec{d}_t, \ldots, \vec{d}_T$; thus its dimension is $n(T - t + 1)$. In the next section, we establish a sufficient condition under which this problem is convex. In the following section, we prove that when this problem is convex, it has a stationary optimal solution; that is, there is an optimal solution where the dose vectors administered to the cohort are identical in all remaining sessions. This stationarity property reduces the dimension of the convex deterministic optimization problem to n, thus enabling its efficient solution.

Convexity of the deterministic problem in CEC

Proposition 4.5.6. *Suppose Assumptions 4.5.1B and 4.5.2A,B hold. Then problem (4.22)-(4.23) is convex.*

Proof. It suffices to show that the objective function to be minimized is convex in $(\vec{d}_t, \vec{d}_{t+1}, \ldots, \vec{d}_T)$ over $D^{n(T-t+1)}$ for any fixed $\vec{x}_t \in X^n$ and any fixed $\vec{w}$. By Assumption 4.5.2A, $c(\vec{d}_s)$ is convex. The nonnegative weighted sum of convex functions is convex; therefore,

$\sum_{s=t}^{T} c(\vec{d}_s)$ is convex. It remains to show that $h(\vec{x}_t + \sum_{s=t}^{T} \vec{f}_0(\vec{d}_s, \vec{w}))$ is convex. Once this is established, we again use the property that the nonnegative weighted sum of two convex functions is convex to complete the proof. We define a vector function $\vec{g} : \mathbb{R}^{n(T-t+1)} \to \mathbb{R}^n$ as $\vec{g}(\vec{d}_t, \vec{d}_{t+1}, \ldots, \vec{d}_T) \triangleq \vec{x}_t + \sum_{s=t}^{T} \vec{f}_0(\vec{d}_s, \vec{w})$. Here, we have suppressed the dependence on $\vec{x}_t$ and $\vec{w}$ as these two are fixed. Thus, we need to prove that $h(\vec{g}(\vec{d}_t, \vec{d}_{t+1}, \ldots, \vec{d}_T))$ is convex. Since $h(\cdot)$ is increasing and convex by Assumption 4.5.2B, it suffices to prove that each component of the vector function $\vec{g}(\cdot)$ is convex (see page 100 of [29]). For $i = 1, 2, \ldots, n$, the ith component of this vector function equals $x_{t,i} + \sum_{s=t}^{T} f_0(d_{s,i}, w)$, which is convex by Assumption 4.5.1B. This completes the proof. □

Optimality of stationary dosing decisions in convex problems

In this section, we prove that when the deterministic optimization problem (4.22)-(4.23) is convex as shown above, it has a stationary optimal solution.

We derive this stationarity property from a more general result about optimality of symmetric solutions to convex problems. So, digressing in terms of notation and topic a bit, we let $u^1, u^2, \ldots, u^N$ be a sequence of decision variables, where $u^t \in \mathbb{R}^M$ for some integer $M \geq 1$, and where $N \geq 2$ is some integer. Consider functions $g_0(u^1; u^2; \ldots; u^N) : \mathbb{R}^{MN} \to \mathbb{R}$ and $g_l(u^1; u^2; \ldots; u^N) : \mathbb{R}^{MN} \to \mathbb{R}$ for $l = 1, 2, \ldots, L$, where $L \geq 0$ is some nonnegative integer. Let Π be the set of all permutations of the set $\{1, 2, \ldots, N\}$ and let π be any generic element of Π. Also, let π_t, for $t = 1, 2, \ldots, N$, denote the number in the tth position in π. We say that function g is *symmetric* when, for any $\pi \in \Pi$, we have $g(u^1; u^2; \ldots; u^N) = g(u^{\pi_1}; u^{\pi_2}; \ldots; u^{\pi_N})$. The lemma below proves optimality of symmetric solutions to convex, symmetric problems. Such results have been previously established in the theoretical optimization literature; see [195] for a historical account on sufficient conditions for symmetric optimal solutions. We nevertheless provide a proof here for the sake of completeness in a format that is convenient for our application at hand.

Lemma 4.5.7. *Consider the optimization problem*

$$\min\ g_0(u^1; u^2; \ldots; u^N), \tag{4.24}$$

$$g_l(u^1; u^2; \ldots; u^N) \leq 0,\ l = 1, 2, \ldots, L. \tag{4.25}$$

Suppose function g_l, for $l = 0, 1, 2, \ldots, L$ are convex over $\mathbb{R}^{MN}$ and symmetric. If the above problem has an optimal solution, then it has a symmetric optimal solution, that is, an optimal solution where $u^1 = u^2 = \ldots = u^N$.

Proof. Suppose $(\mu^1, \mu^2, \ldots, \mu^N)$ is optimal to the above problem. Consider another solution $\nu \in \mathbb{R}^{MN}$, given by,

$$\nu = \frac{1}{N!} \sum_{\pi \in \Pi} (\mu^{\pi_1}; \mu^{\pi_2}; \ldots; \mu^{\pi_N}). \tag{4.26}$$

Consider a random variable $V \in \mathbb{R}^{MN}$ that takes $N!$ different values $(\mu^{\pi_1}; \mu^{\pi_2}; \ldots; \mu^{\pi_N})$ with equal probabilities $1/N!$ for all $\pi \in \Pi$. Then notice that ν as defined in (4.26) is the expected value of this random variable and also that ν is symmetric. Because g_l is convex, Jensen's inequality [29] and the law of the unconscious statistician imply that

$$g_l(\nu) = g_l(E(V)) \leq E(g_l(V)) = \frac{1}{N!} \sum_{\pi \in \Pi} g_l(\mu^{\pi_1}; \mu^{\pi_2}; \ldots; \mu^{\pi_N}) = \frac{1}{N!} \sum_{\pi \in \Pi} g_i(\mu^1; \mu^2; \ldots; \mu^N). \tag{4.27}$$

Here, the last equality holds because g_l is symmetric. Since $(\mu^1; \mu^2; \ldots; \mu^N)$ is feasible, we know that $g_l(\mu^1; \mu^2; \ldots; \mu^N) \leq 0$. Thus, the above equation implies that $g_i(v) \leq 0$, and in particular, this holds for $l = 1, 2, \ldots, L$. That is, ν is feasible to the above optimization problem. By an identical thought process, we see that

$$g_0(\nu) = g_0(E(V)) \leq E(g_0(V)) = \frac{1}{N!} \sum_{\pi \in \Pi} f(\mu^{\pi_1}; \mu^{\pi_2}; \ldots; \mu^{\pi_N}) = \frac{1}{N!} \sum_{\pi \in \Pi} g_0(\mu^1; \mu^2; \ldots; \mu^N) \tag{4.28}$$

$$= g_0(\mu^1; \mu^2; \ldots; \mu^N). \tag{4.29}$$

But since $(\mu^1; \mu^2; \ldots; \mu^N)$ is optimal, the above inequality implies that $g_0(\nu) = g_0(\mu^1; \mu^2; \ldots; \mu^N)$. That is, ν is also optimal. □

Corollary 4.5.8. *Suppose Assumptions 4.5.1B and 4.5.2A,B hold. Then problem (4.22)-(4.23) has an optimal solution where $\vec{d}_t^* = \vec{d}_{t+1}^* = \ldots = \vec{d}_T^*$.*

Proof. The objective function is convex as shown in the proof of Proposition 4.5.6. It is easy to see that the objective function is also symmetric. Finally, the constraints are linear and trivially symmetric. The result then follows by Lemma 4.5.7. □

4.6 Simulation results

In our simulation study, disease condition evolution was modeled using the logarithmic form of the Michaelis-Menten function

$$x_{t+1,i} = x_{t,i} + \ln \Theta - \ln(\Theta + d_{t,i}), \tag{4.30}$$

where $x_{t,i}$ is the natural logarithm of patient i's disease condition before session t; $d_{t,i}$ is the dose administered to patient i in session t; and Θ is the unknown dose-response parameter to be learned. This function satisfies all assumptions made in Section 4.5. We also comment that higher values of Θ here mean that the drug is less effective.

We employed the Michaelis-Menten function in our work on single-patient RGD without learning for rheumatoid arthritis in Chapter 1. That chapter includes a detailed discussion of the properties of this function and also its derivation. There, we used clinical data from [174] to estimate various problem parameters and performed extensive sensitivity analyses. Below, we simply use arbitrary, representative values of these parameters for a hypothetical disease as all our qualitative conclusions are insensitive to specific numbers used in our simulations.

The dose-response parameter Θ was allowed to take on values from the interval $[5, 50]$. The maximum allowable dose $\bar{d}$ was fixed at 1. The number of sessions, T, was fixed at 10. In each session, we used the linear average disutility function

$$c(\vec{d_t}) = \frac{c}{n} \sum_{i=1}^{n} d_{t,i}, \tag{4.31}$$

where c is called the coefficient of dose aversion. For the terminal disutility function, we used the exponential average

$$h(\vec{x}_{T+1}) = \frac{1}{n} \sum_{i=1}^{n} \exp(x_{T+1,i}). \tag{4.32}$$

Note that, since $x_{T+1,i}$ represents the logarithm of the patient i's final disease state, $\exp(x_{T+1,i})$ is simply patient i's final disease state. Thus, equivalently, $h(\vec{x}_{T+1})$ is the average of final disease states in the cohort. Consequently, the combination of disutility functions (4.31)-(4.32) is, in a sense, the simplest possible way to capture the trade-off between dose levels and disease conditions. In fact, clinical trials on RGD often report the total dose administered versus the final disease condition reached in their results; as such, our choice of functions $c(\cdot)$ and $h(\cdot)$ here is consistent with this thought process. As in Chapter 1, the value of c in (4.31) was obtained by inverse optimization in a deterministic dosing problem for a single arbitrary patient i. That is, the value of c was taken to be the one that would result in an optimal constant dose of 0.5 in every session for a patient whose initial logarithmic disease state is $x_{1,i} = 1$, with $\theta = 27.5$. Here, the dose of 0.5 was chosen as the midpoint of the allowable dosing interval $[0, 1]$; similarly, $\theta = 27.5$ was the midpoint of the allowable range of Θ values $[5, 50]$. This inverse optimization resulted in $c = 0.08107427$.

We provide a comparison of our two approximate control schemes, semi-stochastic CEC and CEC, against other solution methods without learning: optimal dosing based simply on the uninformed prior with no updating ("optimal uninformed dosing") and constant dose across all patients and sessions equal to the midpoint of the range of allowed dose values ("constant midpoint dosing"). We also present a clairvoyant solution in which the decision-maker knows the true distribution on Θ a priori; this hypothetical solution method is not realistic in practice, but presents a perfect-case benchmark to compare the other methods against.

For each of these solution methods, we look at two scenarios of the true distribution: one in which the median of the true distribution is lower than the median of the range of allowed values of Θ, and one in which the opposite is true. The first scenario thus represents a situation in which the drug is more effective on average than was anticipated before treatment, and learning reveals that to be the case over the course of T=10 treatment sessions. We call this scenario the "optimistic" scenario. The second, which we call the "pessimistic" scenario, is one in which the drug is less effective on average than was anticipated before treatment.

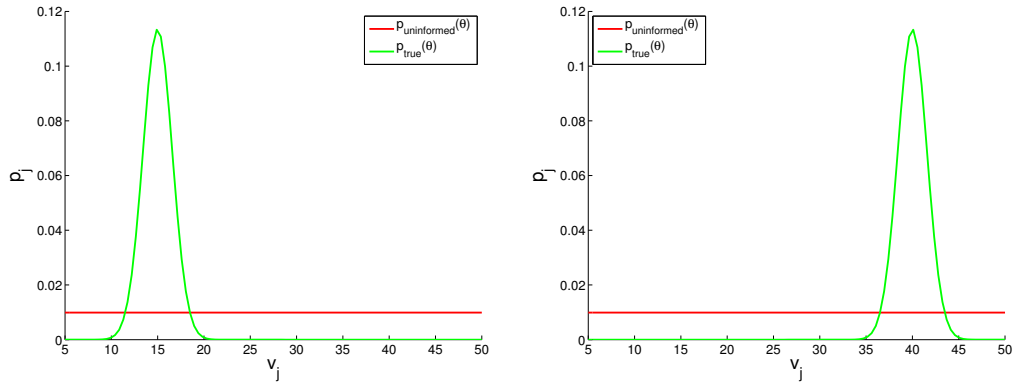


Figure 4.1: Illustration of scenarios considered in our simulations. In the optimistic scenario, shown on the left, the true distribution is a Normal distribution with mean 15 and standard deviation 2.5, truncated and discretized to 101 bins between 5 and 50, while the uninformed distribution is uniform, discretized into 101 bins between 5 and 50. In the pessimistic scenario, shown on the right, the same is true except that the true distribution's mean is 40.

In both cases, we begin with a uniform Dirichlet prior: $(a_{1,1}, a_{1,2}, ..., a_{1,101}) = (1, 1, ..., 1)$, which corresponds to an initial guess $\vec{p} \triangleq (p_1, p_2, ..., p_{101}) = (1/101, 1/101, ..., 1/101)$. Figure 4.1 illustrates these two scenarios.

For each of these scenarios, and all solution methods, we first report results for a simulated cohort of 100 patients, whose initial states are drawn from a Normal distribution with mean 1 and standard deviation 0.1. In Figure 4.2, we show the dose prescribed in each session to each patient in the cohort for each solution method for the optimistic scenario. In Figure 4.3, the same is shown for the pessimistic scenario.

First we focus on the optimistic scenario of Figure 4.2. This is the scenario in which the drug is more effective on average than was anticipated before treatment. We see the doses prescribed by the hypothetical perfect-case clairvoyant solution across all sessions are

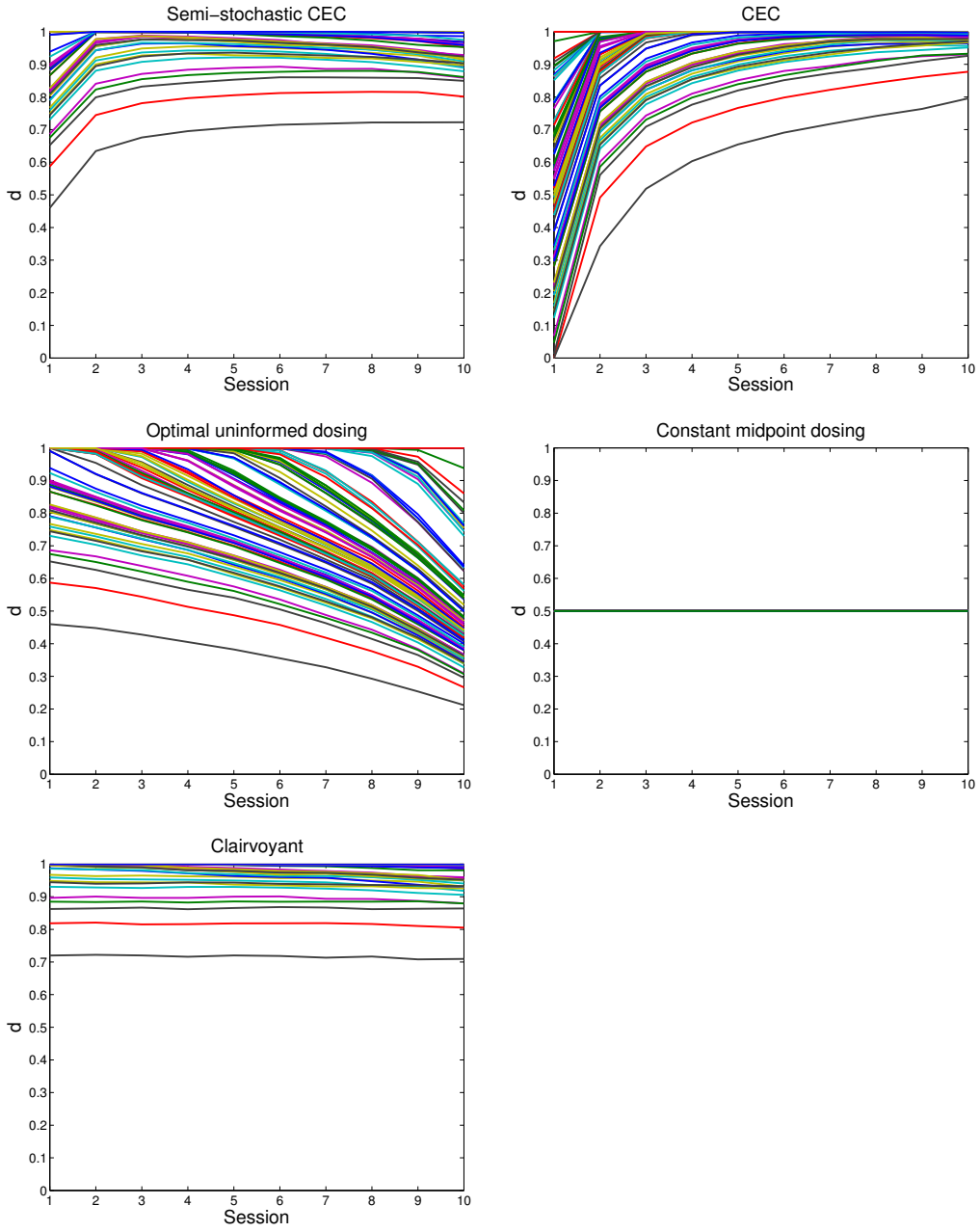


Figure 4.2: Comparison of doses prescribed by different solution methods for a cohort of 100 patients over 10 treatment sessions under the optimistic scenario, where the drug is more effective on average than was anticipated before treatment.

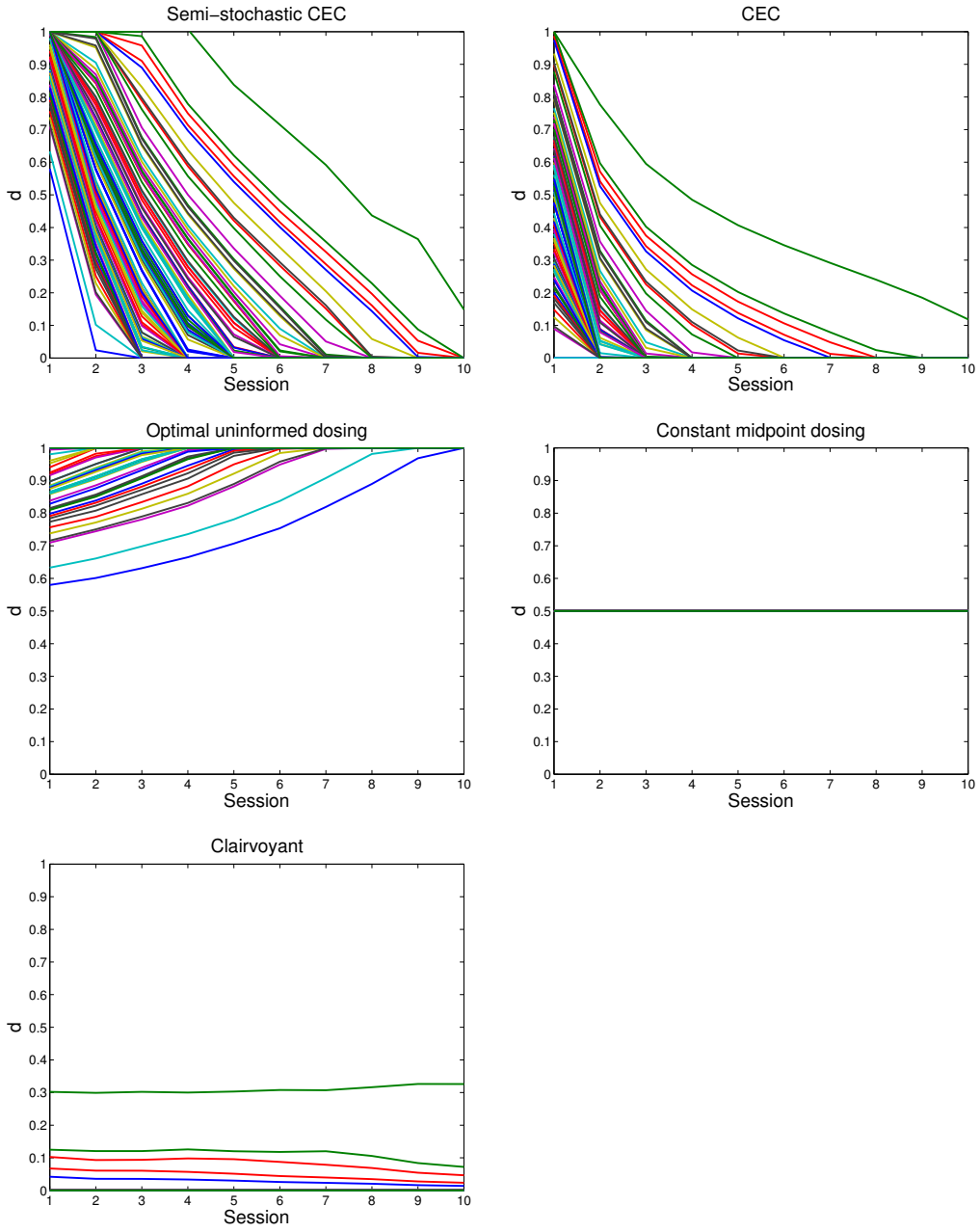


Figure 4.3: Comparison of doses prescribed by different solution methods for a cohort of 100 patients over 10 treatment sessions under the pessimistic scenario, where the drug is less effective on average than was anticipated before treatment.

near the upper range of allowable doses between 0 and 1. This is intuitive since the drug is highly effective and thus the benefit of giving a high dose which will greatly improve patients' health states outweighs the associated costs. In semi-stochastic CEC and CEC, we see that the doses prescribed to the cohort begin over a wide range but generally trend upward as the decision-maker becomes increasingly aware of the high effectiveness of the drug. On the other hand, in optimal uninformed dosing, where the dosing scheme is optimized for the uninformed uniform prior which is never updated, we see the opposite trend: doses tend to be moving downward, which would seem to be suboptimal; we will verify this intuition quantitatively later in Figure 4.4. Finally, in constant midpoint dosing, the same dose of 0.5 is given to all patients in all sessions and information about our belief on the distribution is never updated nor in fact even used in decision-making.

Next, we discuss the pessimistic scenario of Figure 4.3. This is the scenario in which the drug is less effective on average than was anticipated before treatment. Many of the same observations that were made about trends in dosing prescriptions for the optimistic scenario hold here as well, but in the reverse sense. The hypothetical perfect-case clairvoyant solution prescribes doses at the low end of the decision space of [0, 1]– in fact, most of the 100 patients are clustered at a dose of 0 in all sessions. Both our semi-stochastic CEC and CEC methods begin with a range of doses, not knowing anything about the underlying distribution, and gradually lower those doses as it becomes clear the drug is less effective and thus the associated cost outweighs the benefit in patient health state. Yet, the optimal uninformed dosing strategy seems to take the reverse approach, which again is suboptimal, as we will show in Figure 4.4; and the constant midpoint dosing once again gives a flat 0.5 dose to all patients in all sessions.

In Figure 4.4, we compare these different solution methods in each of the two scenarios for a variety of cohort sizes. Cohort sizes were taken to be 10, 30, 100, and 300, and are indicated on the x-axis. For each cohort size, the initial patient states were once again drawn from a Normal distribution with mean 1 and standard deviation 0.1. On the y-axis we plot the average value of the objective function reached over 100 independent simulations. Note

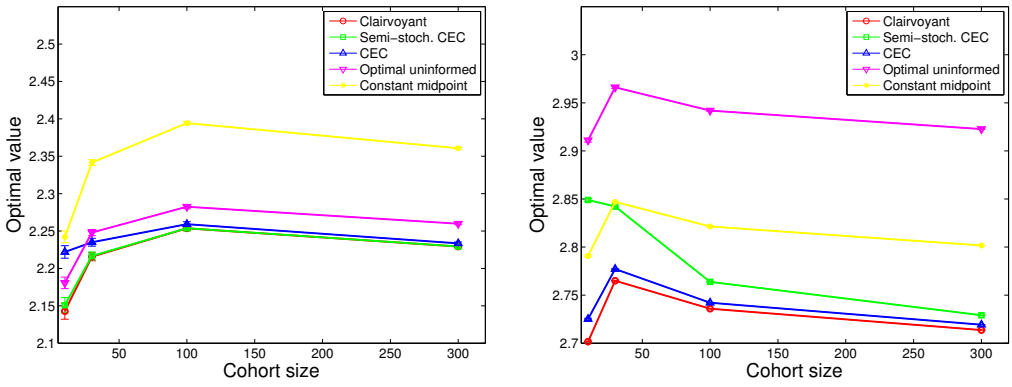


Figure 4.4: Comparison of objective function values obtained by different solution methods for different cohort sizes, averaged over 100 independent simulations. The left figure shows the optimistic scenario while the right figure shows the pessimistic scenario. Clairvoyant is a hypothetical perfect-case in which the decision-maker knows the true distribution a priori. Semi-stochastic CEC and CEC are our approximate control schemes. Optimal uninformed and constant midpoint are dosing methods that do not employ learning.

that this objective function represents a "per-patient" cost, so that the values reached can be compared across cohort size.

We make several observations. First, on the whole, objective function values reached in the optimistic scenario are better than the objective function values reached in the pessimistic scenario. This is expected due to the increased effectiveness of the drug in the optimistic scenario as compared to the pessimistic scenario.

Second, for both scenarios, in cohorts of 30 and larger, we note the ordering of methods relative to each other: the most optimal is the hypothetical perfect-case clairvoyant solution. Next are our approximate control schemes of semi-stochastic CEC and CEC. Finally are the dosing schemes that do not employ learning: optimal uninformed dosing and constant midpoint dosing. We emphasize that although we call the former method "optimal" uninformed dosing, that this dosing is optimized against an uninformed distribution which is never up-

dated, and not the true distribution; thus in a sense it represents the "right answer to the wrong problem." The latter method of constant midpoint dosing does not employ learning or optimization techniques whatsoever. We do not have a theoretical basis for determining which of semi-stochastic CEC or CEC should perform better, nor do we for the two non-learning methods, but we do see that these two pairs are ordered as expected for cohorts of 30 and larger in both the optimistic and pessimistic scenarios.

Third, we note that objective function values of the semi-stochastic CEC and CEC methods approach those of the clairvoyant method as the cohort size increases. This is due to the fact that a larger cohort size represents more observations, and thus faster convergence of our belief on the distribution to the true distribution.

Fourth, we mention that the shape of each curve (that is, the upward trend of all curves in the optimistic scenario and the upward bump followed by lowering trend of all curves in the pessimistic scenario) is not informative, for the reason that different cohort sizes begin with a different sampling of initial patient conditions. The only intuition we have on these shapes is that they should asymptote to horizontal as the cohort size goes to infinity due to the law of large numbers. Admitting that this reasoning is perhaps not entirely rigorous, we do see such a flattening trend in our plots.

Finally, we focus specifically on the cohort of 100 for both optimistic and pessimistic scenarios. The average optimal value reached among 100 independent simulations, for each dosing method in both scenarios, is given in Table 4.1. In these tables we also denote the percentage loss in optimality for each dosing method as compared to the hypothetical, perfect-case clairvoyant solution method. From these tables, we see that in both the optimistic and pessimistic scenarios, our approximate control schemes of semi-stochastic CEC and CEC fare significantly better than the non-learning methods. In the optimistic scenario, even comparing the worse of our two approximate control schemes (CEC) with the better of the two non-learning methods (optimal uninformed dosing), we are able to reach within 0.02457% of perfect as compared to 1.2803%, an improvement by a factor of over 5. Similarly, in the pessimistic scenario, comparing even the worse of our two approximate control scemes

Method	% loss in optimality
Clairvoyant	0
Semi-stochastic CEC	0.0056%
CEC	0.2457%
Optimal uninformed	1.2803%
Constant midpoint	6.2429%

Method	% loss in optimality
Clairvoyant	0
Semi-stochastic CEC	1.0207%
CEC	0.2261%
Optimal uninformed	7.5288%
Constant midpoint	3.1230%

Table 4.1: Loss in optimality incurred by different dosing methods under the optimistic scenario (upper table) and pessimistic scenario (lower table) for a cohort of 100 patients, averaged over 100 independent simulations.

(semi-stochastic CEC) with the better of the two non-learning methods (constant midpoint dosing), we reach within 1.0207% of perfect as compared to 3.1230%, an improvement of a factor of over 3. Obviously these factors increase if we compare the better of our methods with either of the non-learning methods.

One limitation of all chapters so far is that the total number of treatment sessions T was assumed to be known *a priori*. In the final chapter, we present an optimal stopping extension where treatment is allowed to end early.

Chapter 5

OPTIMAL STOPPING FOR RESPONSE-GUIDED DOSING

5.1 Background and motivation

Treatment paradigms for various diseases allow for stopping due to adverse events, and in some cases guidelines have been constructed for when to stop treatment. For some diseases, a recommendation to stop treatment is made typically at the end of a gradual tapering-down of dose for patients who respond well to treatment and are considered to be in remission. For others, patients are given a standard dose and the treatment decision at each time step is of the stop/do-not-stop type. In addition, stopping treatment may occur for patients in poor disease states due to a finding of futility or a desire to switch to a different drug or type of treatment.

Discontinuation of pharmalogical therapy has been studied in a number of diseases. For rheumatoid arthritis (RA), a protocol for discontinuing the biologic agent infliximab has been developed by Maas et al.: patients whose 28-joint disease activity score (DAS28) is below 3.2, and have received stable dose for at least 6 months, have their doses tapered down by 25% of the original dose every 8-12 weeks until discontinuation of treatment is acheived or the patient experiences a flare-up [188]. Another study on adapting dose of the biologic agent infliximab based on patient response ended up stopping treatment for 7 of 76 patients due to adverse events [59]. One meta-analysis compared gradual lowering of dose (also called "down-titration") and discontinuation versus continuation of the drugs adalimumab and etanercept in RA patients with low DAS28 scores with mixed results, finding that stopping treatment produces benefits in some, but not all patients [191].

Infliximab is also used to treat other inflammatory bowel diseases (IBD) including Crohn's disease and ulcerative colitis (UC). Other studies have focused mainly on patient outcomes

after the decision to stop infliximab treatment. Several studies have been conducted on the risk of IBD disease relapse after a decision to interrupt treatment of infliximab [117, 126, 178]. A prevalence study found that an "important proportion" of RA patients in remission were directed to down-titrate or discontinue treatment the drug, indicating that the stopping decision is not uncommon in practice, though a patient-specific numerical framework does not exist [115]. One study found that 62% of patients who stopped a second-line drug in combination therapy for RA did not experience a flare within one year; yet patients who continued the second-line drug had a lower chance of flare [183]. A meta-analysis of flare rates for RA patients with low DAS28 scores or in remission found that "more than one-third of patients" may down-titrate or stop disease-modifying anti-rheumetic drugs (DMARD) without risk of a flare for one year [104].

While there is interest in the decision to stop treatment and the repercussions thereof, we are unaware of any literature which combines the decision to stop or not-to-stop with a patient-specific, response-guided dosing framework. In this chapter, we extend the stochastic DP model of Chapter 2 to allow for stopping treatment as an alternative to administering dose in any session. In essence, this adds an additional option to the decision-space, so that in any session t a dose $d \in D$ may still be administered, or a decision to stop may be made. If the decision-maker stops treatment, then no dose may be administered in future sessions, and as a result no future per-session costs are incurred. The final objective function value is simply the boundary value function evaluated at the current disease state. Through numerical simulation we find that when stopping is optimal, it is optimal below a threshold disease state.

5.2 Model

The model is an extension of the stochastic DP model for RGD on a single patient of Chapter 2. That model took the number of equally-spaced treatment sessions T to be known a priori, with sessions indexed by $t = 1, 2, ..., T$. In this chapter, we also take T to be known, but allow for the possibility of ending treatment early. Thus, T can be thought of as a loose

upper bound on the number of treatment sessions. In session t, we can as before choose to give a dose, which is denoted d_t, where $d_t \in D \triangleq [0, \bar{d}]\ \forall\ t$ where $\bar{d} < \infty$ is the maximum permissible dose in one session; alternatively a decision to stop treatment can be made. For the purposes of graphical illustration, we denote the decision to stop treatment by $d_t = -1$. This value is not physically meaningful, but as it lies outside the permitted dose range $D \triangleq [0, \bar{d}]$, it allows for simple visualization.

If treatment is terminated when the state is x_t, the patient derives a terminal disutility $h(x_t)$. If treatment is continued, then a dose level d_t for that session must be chosen. For all $x_t \in X$, and $t = 1, 2, \ldots, T$, Bellman's equations thus change to

$$J_t(x_t) = \min\left\{ \min_{d_t \in D} \left(c(d_t) + \int J_{t+1}(f(x_t, d_t; \Theta)) d\Theta \right), h(x_t) \right\}, \text{ with } J_{T+1}(x) = h(x), \quad (5.1)$$

where, as before, $J_t(\cdot)$ is the optimal cost-to-go function in session t, $c(\cdot)$ is the per-session cost due to adverse effects, $f(\cdot, \cdot; \cdot)$ is the state transition function, and $p(\cdot)$ is the probability density of the stochastic variable Θ.

Since we will seek solutions to this problem numerically, Θ is again taken to be a Categorical random variable that takes k possible values listed in the finite set $\Omega \triangleq \{v_1, v_2, \ldots, v_k\}$. p_j, for $j = 1, 2, \ldots, k$ denotes the probability that Θ takes the value v_j. The distribution is assumed to be known, perhaps as a result of completing a trial on a cohort of patients, as described in Chapter 4. Again, $f(\cdot, \cdot; \cdot)$ is taken to be additively separable in state: $f(x_t, d_t; \theta) = x_t + f_0(d_t; \Theta)$.

Bellman's equations can then be written as:

$$J_t(x_t) = \min\left\{ \min_{d_t \in D} \left(c(d_t) + \sum_{j=1}^{k} p_j J_{t+1}(x_t + f_0(d_t; v_j)) \right), h(x_t) \right\}, \text{ with } J_{T+1}(x) = h(x), \quad (5.2)$$

5.3 Motivating examples

Naively, one may suspect that an optimal stopping formulation offers no additional benefit to the original model, as it could be considered equivalent to giving a dose 0 in all future sessions. That is, stopping in session t would mean that $d_t = d_{t+1} = d_{t+2} = ... = d_T = 0$.

However, this reasoning is incomplete. Looking back at the original model where we have monotonicity of optimal dose, even if the optimal dose in session t for some state x_t is $d_t^* = 0$, stochasticity in disease-response could bring the state x_{t+1} to a higher level where the optimal dose $d_{t+1}^* > 0$. Therefore, by choosing to stop treatment, we forego any possible improvements later from giving a positive dose. If the cost function is such that $c(0) > 0$, we also see that the decision to stop avoids incurring those future fixed costs. Thus, allowing the decision to stop is indeed a non-trivial extension of our original model of Chapter 2.

5.3.1 Analytical solution of a 1-period problem with Bernoulli distributed noise

First let us consider a hypothetical disease that uses the log Michaelis-Menten transition function with additive noise:

$$x_{t+1} = x_t + f_0(d_t; \theta) = x_t + \ln \kappa - \ln(\kappa + d_t) + r\Theta \tag{5.3}$$

The final state cost function is exponential: $h(x) = \exp(x)$ and the per-session cost function is linear: $c(d) = ad + b$. For algebraic simplicity we take the random variable Θ to be Bernoulli distributed, which is a special case of the Categorial distribution with only two possible outcomes:

$$\Theta \sim \begin{cases} \theta_-/r, & \text{prob.} \quad p \\ \theta_+/r, & \text{prob.} \quad 1-p \end{cases} \tag{5.4}$$

We define:

$$S \triangleq E[e^{r\Theta}] = pe^{\theta_-} + (1-p)e^{\theta_+} \tag{5.5}$$

And we assume $T = 1$, indicating a one-period problem. Since $T = 1$, stopping is optimal for x which satisfy

$$e^x < \min_d \left\{ ad + b + E[e^{x - \ln(\kappa+d) + \ln\kappa + r\Theta}] \right\} = \min_d \left\{ ad + b + \frac{\kappa e^x S}{\kappa + d} \right\} \tag{5.6}$$

Equating the partial derivative with respect to d of the quantity in the curly brackets to zero reveals the optimal dose d^*:

$$a - \frac{\kappa e^x S}{(\kappa + d)^2} = 0 \Rightarrow d^* = \max\left\{ \sqrt{\frac{\kappa e^x S}{a}} - \kappa, 0 \right\} \tag{5.7}$$

where we have assumed $\bar{d} > \sqrt{\frac{\kappa e^{\bar{x}} S}{a}} - \kappa$ and $\bar{x}$ is the largest state considered. This corresponds to a situation with no upper bound on dose. Monotonicity of the quantity $\sqrt{\frac{\kappa e^{x} S}{a}} - \kappa$ with respect to x implies that d^* is a piecewise-defined function of x:

$$d^* = \begin{cases} 0, & x \leq \ln\left(\frac{\kappa a}{s}\right) \\ \sqrt{\frac{\kappa S}{a}} e^{x/2} - \kappa, & x > \ln\left(\frac{\kappa a}{s}\right) \end{cases} \tag{5.8}$$

Note that d^* represents the optimal dose for the problem without allowing for stopping. Equation 5.6 becomes:

$$e^x < a d^* + b + \frac{\kappa e^x S}{\kappa + d^*} \tag{5.9}$$

Case 1a. $x \leq \ln\left(\frac{\kappa a}{S}\right)$ *and* $S < 1$

In this case, by equation 5.8, $d^* = 0$. Then equation 5.9 becomes:

$$e^x < b + e^x S \Rightarrow e^x(1 - S) < b \tag{5.10}$$

Since $S < 1$, stopping is optimal for $x < \ln\left(\frac{b}{1-S}\right)$. Thus there are two sub-cases depending on the ordering of $\ln\left(\frac{\kappa a}{S}\right)$ and $\ln\left(\frac{b}{1-S}\right)$.

Subcase 1ai. $\ln\left(\frac{b}{1-S}\right) < \ln\left(\frac{\kappa a}{S}\right)$

Stopping is optimal over $x < \ln\left(\frac{b}{1-S}\right)$.

Subcase 1aii. $\ln\left(\frac{b}{1-S}\right) \geq \ln\left(\frac{\kappa a}{S}\right)$

Stopping is optimal over $x \leq \ln\left(\frac{\kappa a}{S}\right)$.

Case 1b. $x \leq \ln\left(\frac{\kappa a}{S}\right)$ *and* $S \geq 1$

In this case, $e^x(1 - S) < b$ holds for all x since $e^x(1 - S) \leq 0$ and $b > 0$. Therefore stopping is optimal over $x \leq \ln\left(\frac{\kappa a}{S}\right)$.

Case 2a. $x > \ln\left(\frac{\kappa a}{S}\right)$ *and* $\kappa a - b \geq \kappa a S$

In this case we have $d^* = \sqrt{\frac{\kappa S}{a}} e^{x/2} - \kappa$ and equation 5.9 becomes:

$$e^x - 2\sqrt{\kappa a S} e^{\frac{x}{2}} + \kappa a - b < 0. \tag{5.11}$$

Equation 5.11 is quadratic in $e^{x/2}$. The discriminant is $4(\kappa a S - (\kappa a - b))$ which is nonpositive by assumption in this case. Thus, there is no solution and stopping is never optimal for any $x > \ln\left(\frac{\kappa a}{S}\right)$.

Case 2b. $x > \ln\left(\frac{\kappa a}{s}\right)$ *and* $0 < \kappa a - b < \kappa a S$

In contrast to case 2a, $\kappa a - b < \kappa a S$ implies the discriminant is positive. Thus there exists a real solution for $e^{x/2}$. Equation 5.11 becomes

$$\sqrt{\kappa a S} - \sqrt{\kappa a S - (\kappa a - b)} < e^{x/2} < \sqrt{\kappa a S} + \sqrt{\kappa a S - (\kappa a - b)}. \tag{5.12}$$

$\kappa a - b > 0$ implies that $-(\kappa a - b) < 0 \Rightarrow \kappa a S - (\kappa a - b) < \kappa a S \Rightarrow \sqrt{\kappa a S - (\kappa a - b)} < \sqrt{\kappa a S} \Rightarrow \sqrt{\kappa a S} - \sqrt{\kappa a S - (\kappa a - b)} > 0$. Thus both the left and right expressions in equation 5.12 are positive. Solving for x gives the stopping region:

$$2\ln\left(\sqrt{\kappa a S} - \sqrt{\kappa a S - (\kappa a - b)}\right) < x < 2\ln\left(\sqrt{\kappa a S} + \sqrt{\kappa a S - (\kappa a - b)}\right). \tag{5.13}$$

There are three subcases depending on the ordering of $2\ln\left(\sqrt{\kappa a S} - \sqrt{\kappa a S - (\kappa a - b)}\right)$, $2\ln\left(\sqrt{\kappa a S} + \sqrt{\kappa a S - (\kappa a - b)}\right)$, and $\ln\left(\frac{\kappa a}{S}\right)$.

Subcase 2bi. $2\ln\left(\sqrt{\kappa a S} - \sqrt{\kappa a S - (\kappa a - b)}\right) < 2\ln\left(\sqrt{\kappa a S} + \sqrt{\kappa a S - (\kappa a - b)}\right) \leq \ln\left(\frac{\kappa a}{S}\right)$

In this subcase, no $x > \ln\left(\frac{\kappa a}{S}\right)$ is optimal.

Subcase 2bii. $2\ln\left(\sqrt{\kappa a S} - \sqrt{\kappa a S - (\kappa a - b)}\right) \leq \ln\left(\frac{\kappa a}{S}\right) < 2\ln\left(\sqrt{\kappa a S} + \sqrt{\kappa a S - (\kappa a - b)}\right)$

In this subcase, stopping is optimal for $\ln\left(\frac{\kappa a}{S}\right) < x < 2\ln\left(\sqrt{\kappa a S} + \sqrt{\kappa a S - (\kappa a - b)}\right)$.

Subcase 2biii. $\ln\left(\frac{\kappa a}{S}\right) < 2\ln\left(\sqrt{\kappa aS} - \sqrt{\kappa aS - (\kappa a - b)}\right) < 2\ln\left(\sqrt{\kappa aS} + \sqrt{\kappa aS - (\kappa a - b)}\right)$

In this subcase, stopping is optimal for x such that $2\ln\left(\sqrt{\kappa aS} - \sqrt{\kappa aS - (\kappa a - b)}\right) < x < 2\ln\left(\sqrt{\kappa aS} + \sqrt{\kappa aS - (\kappa a - b)}\right)$.

Case 2c. $x > \ln\left(\frac{\kappa a}{s}\right)$ *and* $\kappa a - b \leq 0$

Here we again have a positive discriminant, but now $\kappa a - b \leq 0$ implies $\sqrt{\kappa aS} - \sqrt{\kappa aS - (\kappa a - b)} \leq 0$. Thus the left expression in equation 5.12 is nonpositive, and the left inequality is always satisfied since $e^{x/2} > 0$. Solving for x gives:

$$x < 2\ln\left(\sqrt{\kappa aS} + \sqrt{\kappa aS - (\kappa a - b)}\right). \tag{5.14}$$

There are two subcases depending on the ordering of $2\ln\left(\sqrt{\kappa aS} + \sqrt{\kappa aS - (\kappa a - b)}\right)$ and $\ln\left(\frac{\kappa a}{S}\right)$.

Subcase 2ci. $2\ln\left(\sqrt{\kappa aS} + \sqrt{\kappa aS - (\kappa a - b)}\right) \leq \ln\left(\frac{\kappa a}{S}\right)$

Here stopping is non-optimal for all $x > \ln\left(\frac{\kappa a}{S}\right)$.

Subcase 2cii. $2\ln\left(\sqrt{\kappa aS} + \sqrt{\kappa aS - (\kappa a - b)}\right) > \ln\left(\frac{\kappa a}{S}\right)$

Here stopping is optimal for $\ln\left(\frac{\kappa a}{S}\right) < x < 2\ln\left(\sqrt{\kappa aS} + \sqrt{\kappa aS - (\kappa a - b)}\right)$.

At this point, to reduce the cases under consideration, we focus on the system when $S = 1$. This corresponds to a situation in which the expected value of the multiplicative noise term is 1; therefore, the objective function value is the same on average when no dose is given. One such combination: $\theta_- = -0.25$ and $\theta_+ = 0.25$, with $p = 0.5622$ gives $S = 1$. To compare with a numerical simulation, we choose values $\kappa = a = 1$, and $b = 0.1$. The problem reduces to two cases: case 1b and 2b. Case 1b shows that stopping is optimal for $x \leq 0$. Case 2b shows that stopping is optimal when $x > 0$ and $2\ln(1 - \sqrt{0.1}) < x < 2\ln(1 + \sqrt{0.1})$, but since $2\ln(1 - \sqrt{0.1}) < 0$, this reduces to $0 < x < 2\ln(1 + \sqrt{0.1}) \approx 0.5495$. Combining both cases, the analytical solution to this problem shows x optimal for $x < 0.5495$....

A numerical simulation was performed on this example, and with a grid spacing of $x = 0.01$, stopping was found to be optimal for $x \leq 0.54$ and non-optimal for $x \geq 0.55$, agreeing with our analytical result. This is illustrated in Figure 5.1.

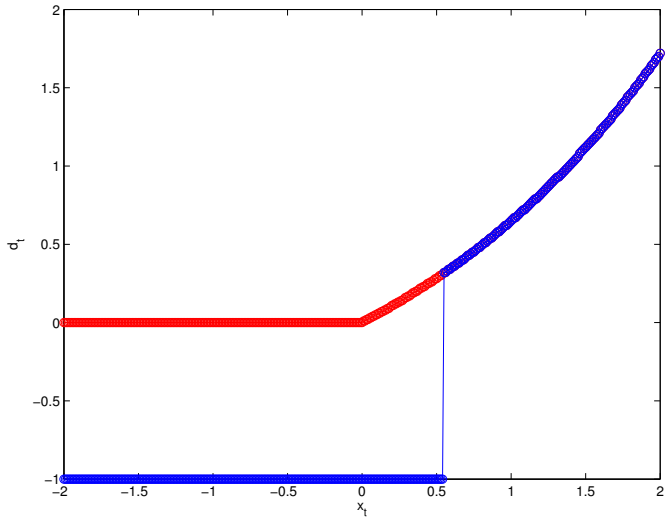


Figure 5.1: Optimal policy for the 1-period problem with Bernoulli distributed noise. $d_t = -1$ indicates a decision to stop treatment. Blue represents the optimal policy for the problem allowing stopping; red represents the optimal policy if stopping is not allowed.

5.3.2 Numerical solution of a 3-period problem with Normally distributed noise

As another motivating example, consider another hypothetical disease that also uses the log Michaelis-Menten transition function with additive noise of equation 5.3.

The parameter values were taken to be $\kappa = 100$ and $r = 0.25$. The final state cost function is exponential: $h(x) = \exp(x)$, and per-session cost function is linear: $c(d) = 0.015d + 0.1$. The patient is treated over $T = 3$ sessions with a maximum dose $\bar{d} = 1$, and the distribution

on Θ is taken to be:

$$\Theta \sim \begin{cases} v_1 = -1, & \text{prob.} & p_1 \\ v_2 = -0.98, & \text{prob.} & p_2 \\ v_3 = -0.96, & \text{prob.} & p_3 \\ v_4 = -0.94, & \text{prob.} & p_4 \\ \quad \dots & & \dots \\ v_{100} = 0.98, & \text{prob.} & p_{100} \\ v_{101} = 1, & \text{prob.} & p_{101} \end{cases} \tag{5.15}$$

with a truncated, discretized Normal distribution over the nonpositive values of v_j:

$$p_i = \begin{cases} \frac{\exp((-3+0.12(i-1))^2/2)}{\sum_{j=1}^{51} \exp((-3+0.12(j-1))^2/2)} & i = 1, 2, ..., 51 \\ 0 & i = 52, 53, ..., 101 \end{cases} \tag{5.16}$$

The fact that this pmf is skewed entirely toward nonpositive values indicates that the patient's state will be better in the next session even without treatment, since $f_0(0; \Theta) = 0.25\Theta$.

The problem was solved numerically through backward induction of the Bellman's equations (5.2), discretizing state and dose on a grid of width 0.01.

Again, the Bellman's equations with stopping (5.2) can be solved via backward induction by discretizing the state-action space. For this particular example, the numerical results of Figure 5.2 show that the optimal decision in the first session is a nonzero dose for $x_1 \geq 0.79$, a zero dose for $-0.15 \leq x_1 \leq 0.78$, and stopping treatment for $x_1 \leq -0.16$.

Another example of the benefit of stopping would be one in which the decision to give zero dose occurs for no state; rather, the optimal decision is either to give a positive dose or stop. In such a situation, the decision to stop early has a lower associated cost than giving zero dose now for the chance of reaching a higher state later. Such a picture is shown in Figure 5.3 which used the same values as Figure 5.2 except that the per-session cost function $c(d)$ was changed to $c(d) = 0.005d + 0.1$. In this figure, the optimal policy is to give a dose of $d_1 = 1$ in the first session when $x_1 \geq -0.17$ and to stop treatment in the first session when $x_1 \leq -0.18$. This example thus represents an extreme case where either stopping

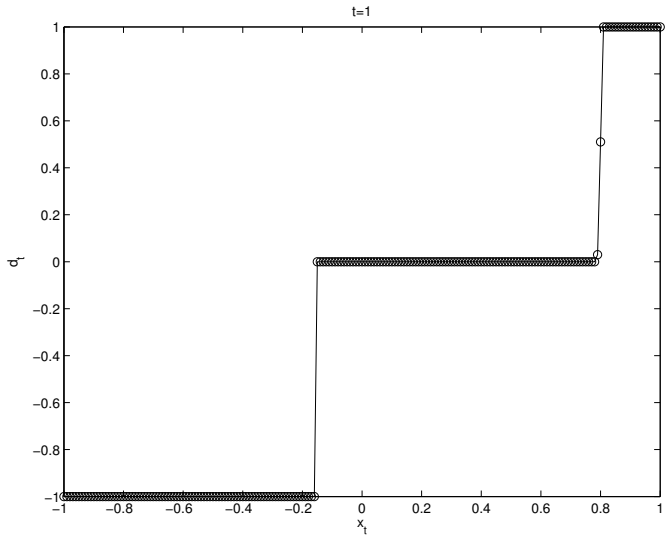


Figure 5.2: Optimal dose for the first period in a 3-period problem. $d_t = -1$ indicates a decision to stop treatment. This picture illustrates a scenario in which there are intervals of x_t where it is optimal to stop treatment, to continue treatment but give zero dose in session 1, or to give positive dose in session 1.

or maximum dose is optimal, with the transition happening as sharply as possible. This is probably in part due to the large value of κ relative to $\bar{d}$. Recall that in the original (non-log) Michaelis-Menten formulation, κ represented the dose required to bring the state down by a factor of two in one session; thus, a monotone increasing policy where $\bar{d} \ll \kappa$ will quickly saturate to $\bar{d}$.

5.4 Sensitivity analysis of a 3-period problem with tent function-distributed noise

For the purposes of performing sensitivity analyses, we consider the problem of a hypothetical disease treated over $T = 3$ sessions with the possibility of stopping in any session $t \leq T$. The state-transition function is again taken to be the log Michaelis-Menten with additive noise

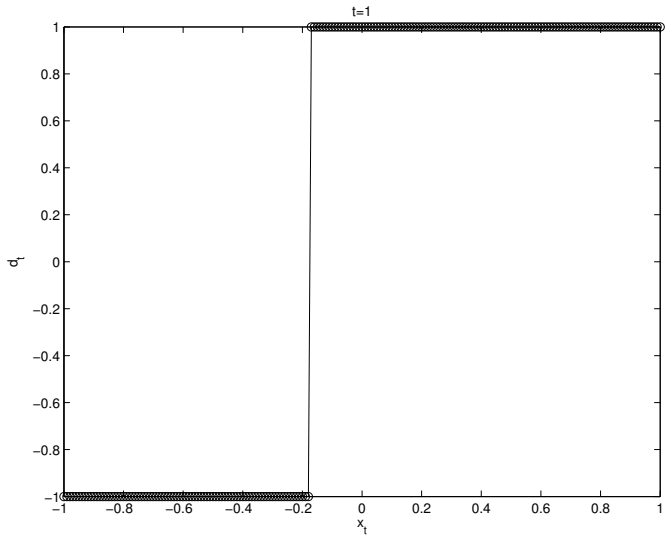


Figure 5.3: Optimal dose for the first period in a 3-period problem. $d_t = -1$ indicates a decision to stop treatment. This picture illustrates a scenario in which the cost to give zero dose in session 1 is always outweighed by the benefit of stopping early, but a positive dose can be given above a threshold state.

of equation (5.3) of the motivating examples.

In the absense of stochasticity ($\theta = 0$), the parameter κ represents the dose necessary to reduce the state by $\ln 2$. The multiplicative factor r parameterizes the level of uncertainty in the stochastic state evolution. Initial values were taken to be $\kappa = 10$ and $r = 0.25$; sensitivity analyses will be performed on these values later. The maximum permissible dose was taken to be $\bar{d} = 1$. The state and dose were both discretized on a grid spacing of 0.01.

The terminal state cost function was taken to be exponential: $h(x) = \exp(x)$. The dose-aversion cost function was taken to be linear: $c(d) = ad + b$. The parameters $a > 0$ and $b > 0$ were varied. Here, a represents the marginal cost of increasing the dose by one due to adverse effects, while b represents a fixed per-session cost perhaps due to administrative overhead. a was varied from 0 to 0.5 in increments of 0.01, and b was varied from 0 to 0.2 in

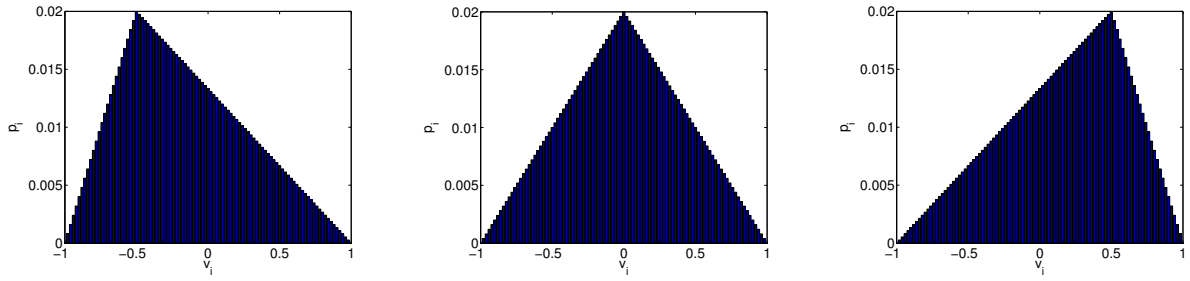


Figure 5.4: Three pmf's considered: left-skewed, symmetric, and right-skewed tent functions.

increments of 0.05.

For ease of investigating different probability mass functions on θ, a single-parameter tent function was chosen. The unnormalized pmf is:

$$q_i \triangleq \begin{cases} \frac{i-1}{i_{peak}-1}, & i \leq i_{peak} \\ \frac{i_{max}-i}{i_{max}-i_{peak}}, & i > i_{peak} \end{cases} \tag{5.17}$$

where $i = 1, 2, ..., i_{max}$, $i_{max} = 101$, and i_{peak} parameterizes the peak of the tent function. To turn this into a pmf, it is normalized:

$$p_i = \frac{q_i}{\sum_{j=1}^{i_{max}} q_j} \tag{5.18}$$

Three values of i_{peak} were considered: 26, 51, and 76, resulting in left-skewed, symmetric, and right-skewed tent function pmfs, respectively. These pmfs are illustrated in Figure 5.4.

5.4.1 Left-skewed tent function

Due to the form of the additive noise term $+r\Theta$ on our state transition function, a left-skewed tent function physically represents a situation in which the patient's health state is more likely to improve than deteriorate in the absence of therapy.

Results of numerical simulation are shown in Figure 5.5. We see in several of the subplots a distinction between a dose of -1, which indicates a decision to stop treatment, and dose 0,

which gives no dose at the moment but keeps open the possibility of giving dose later by not stopping.

For a fixed value of b and for every session t, we observe monotonicity of dose (nondecreasing dose from left to right along any horizontal slice of any subplot). We also observe that the lowest state for which nonnegative dose is prescribed is increasing with increasing a for all values of b and all sessions t. This is due to the increased marginal cost of dose in the per-session cost function $c(d) = ad + b$: as cost increases, the disease state at which treatment "turns on" also should increase.

In comparing different values of b, we first observe that the optimal policy never stops treatment when $b = 0$. This is intuitive, as b represents the fixed per-session cost (recall that $c(d) = ad + b$. With $b = 0$, the patient state is more likely to improve than deteriorate with any dose in [0,1]. Thus, there is never any reason to stop treatment early, as continuing treatment until the end is always more likely to achieve a lower disease state. Increasing b from 0 has several effects: most notably, the appearance of a region in a-state space where stopping is optimal. For $b > 0$, the fact that the patient's health state is more likely to improve than deteriorate is counterbalanced by the fact that the fixed cost b will be incurred in future sessions if treatment is continued. Thus, it is conceivable that this trade-off could lead to different results for optimality depending on the location in $x_t - a$ space. In particular, we see that the region in $x_t - a$ space where stopping is optimal moves in from the left because the lower the patient's disease state, the less likely it is that the patient's state will jump high enough to warrant a positive dose in a future session. Thus, it is not worth the fixed costs b that will be incurred in future sessions, and stopping now becomes optimal. The larger b is, the greater this effect– hence the movement of the stopping/zero-dose threshhold to the right as b increases. In between we see the existence of a "wait-and-see" region, where zero dose is given but treatment is not stopped. This region represents states which are relatively high so that there is a significant chance the state will jump up to a higher state where dose should be given; but relatively low so that no dose need be given at the present state. Thus, in this region, the decision maker does not stop treatment but waits to see how the state

changes in future sessions.

5.4.2 Symmetric tent function

The symmetric tent function physically represents a situation in which the patient's health state is as likely to improve as deteriorate in the absense of therapy. Results of numerical simulation are shown in Figure 5.6. For these simulations, we see the lack of a zero-dose region for all values of b and all sessions t. This indicates that treatment is only continued when a positive dose is given. With a decreased likelihood of patient improvement, and especially when $b > 0$ but even when $b = 0$, the costs of giving zero dose now for the possibility of being able to give a positive dose later are outweighed by the benefits of stopping now.

5.4.3 Right-skewed tent function

The right-skewed tent function physically represents a situation in which the patient's health state is more likely to deteriorate than improve in the absense of therapy. Results of numerical simulation are shown in Figure 5.7. As for the symmetric tent function, we see the lack of a zero-dose region for all values of b and all sessions t.

5.4.4 Effect of changing r

The parameter r quantifies the size of the additive noise term in the state transition function. Thus, a larger r represents more uncertainty in the patient's disease state regardless of dose given. The 3-period problem with left-skewed tent function of section 5.4.1 was considered, setting $b = 0.1$ to be constant, but varying r over the values $0.05, 0.1, 0.25, 0.5, 1$. Numerical results are shown in Figure 5.8.

The main change in optimal policy is the expansion of the "wait-and-see" region with respect to x_t-space with increasing r. For intermediate values of x_t, there is less sense in adpoting a wait-and-see policy if the uncertainty is very small– if the patient is in a good

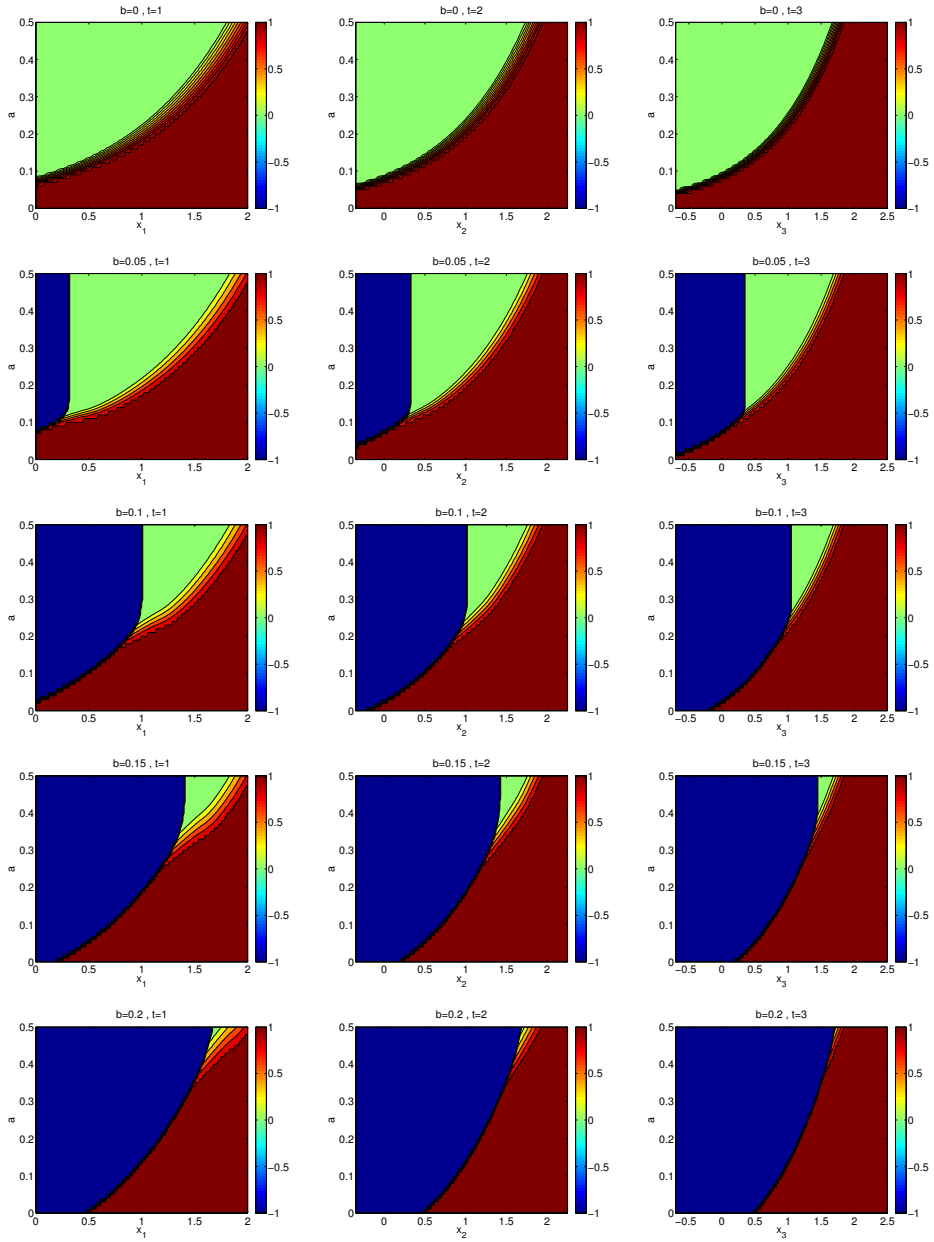


Figure 5.5: Contour plots of optimal dose for a three-period problem using the linear per-session cost function $c(d) = ad + b$, and a left-skewed tent distribution for the additive noise θ. Columns correspond to session number t and rows to varying values of b. Color bar corresponds to optimal dose on [0,1]. A dose of -1 indicates a decision to stop treatment.

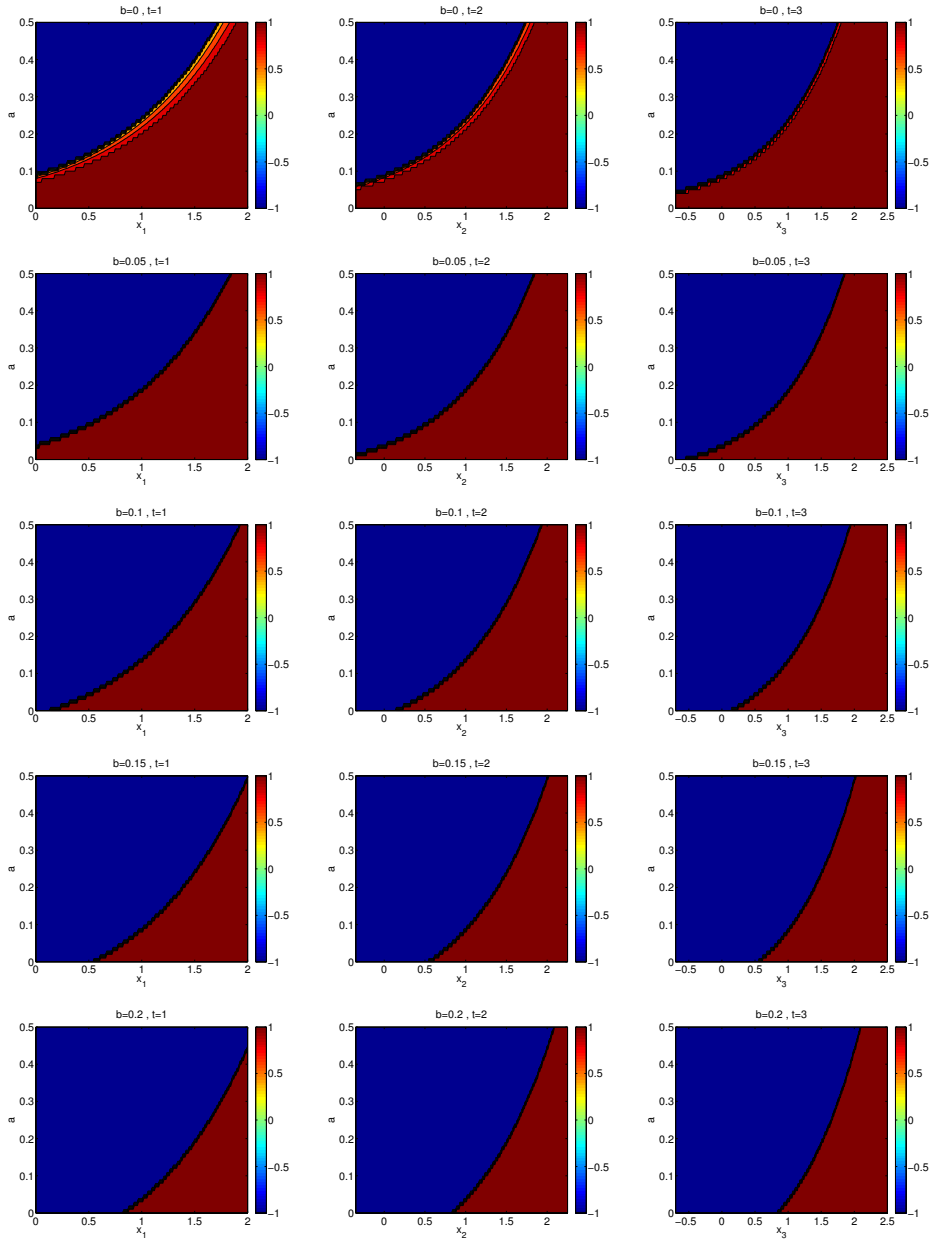


Figure 5.6: Contour plots of optimal dose for a three-period problem using the linear per-session cost function $c(d) = ad + b$, and symmetric tent distribution for the additive noise θ. Columns correspond to session number t and rows to varying values of b. Color bar corresponds to optimal dose on [0,1]. A dose of -1 indicates a decision to stop treatment.

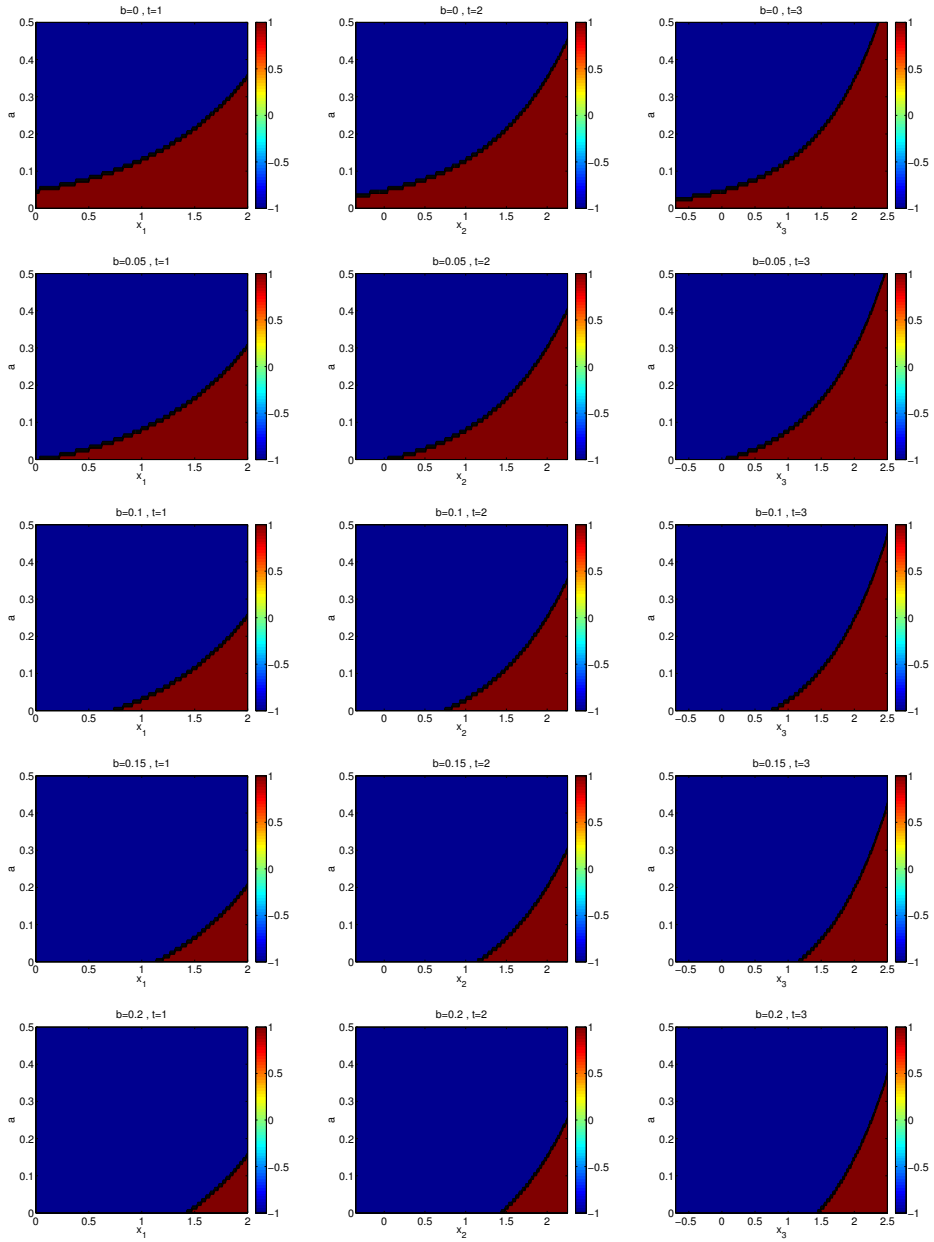


Figure 5.7: Contour plots of optimal dose for a three-period problem using the linear per-session cost function $c(d) = ad + b$, and a right-skewed tent distribution for the additive noise θ. Columns correspond to session number t and rows to varying values of b. Color bar corresponds to optimal dose on [0,1]. A dose of -1 indicates a decision to stop treatment.

enough health state to receive zero dose now, it's unlikely they will jump upward to need a positive dose later. Thus, we see the collapse of this region for small values of r. For larger values of r, it is more likely that such a jump will occur.

5.4.5 Effect of changing κ

If we consider the deterministic state-transition function by eliminating the additive noise term: $f(x_t, d_t) = x_t + \ln\kappa - \ln(\kappa + d)$, then κ represents the dose required to reduce x_t by $\ln 2$. If x_t represents the log of state, as we took for example in the rheumatoid arthritis example of Chapter 1, then κ is the dose required to reduce the state by a factor of 2. Thus, a larger value of κ relative to $\bar{d}$ represents a situation where the disease is more resistant to dose, and thus the amount of dose required to achieve the same final disease state is higher.

Numerical results are shown in Figure 5.9, which was generated assuming $b = 0.1$, $r = 0.25$, $\bar{d} = 1$, and a and κ were varied. Note that the threshold between the "wait-and-see" region and the stopping region does not appear to change with respect to κ. This is because κ affects neither the terminal disease cost function $h(x)$ nor the cost of giving zero dose $c(0) = 0a + b = b$ nor the state-transition function with zero dose $f(x_t, 0; \Theta) = x_t + \ln\kappa - \ln(\kappa + 0) + r\Theta = x_t + r\Theta$. Thus, the only effect of κ is in determining the threshold where positive dose is given. We notice that optimal doses tend to decrease with increasing κ for a particular fixed a, b, r, and $\bar{d}$. When κ is on the order of $\bar{d}$, we see interesting non-monotone results. This dose is no longer necessarily monotonically increasing with increasing disease state, as we saw in Chapter 2, as we have no proof for the existence of a monotone optimal policy in the stopping case, but it is interesting to observe an actual counterexample of this.

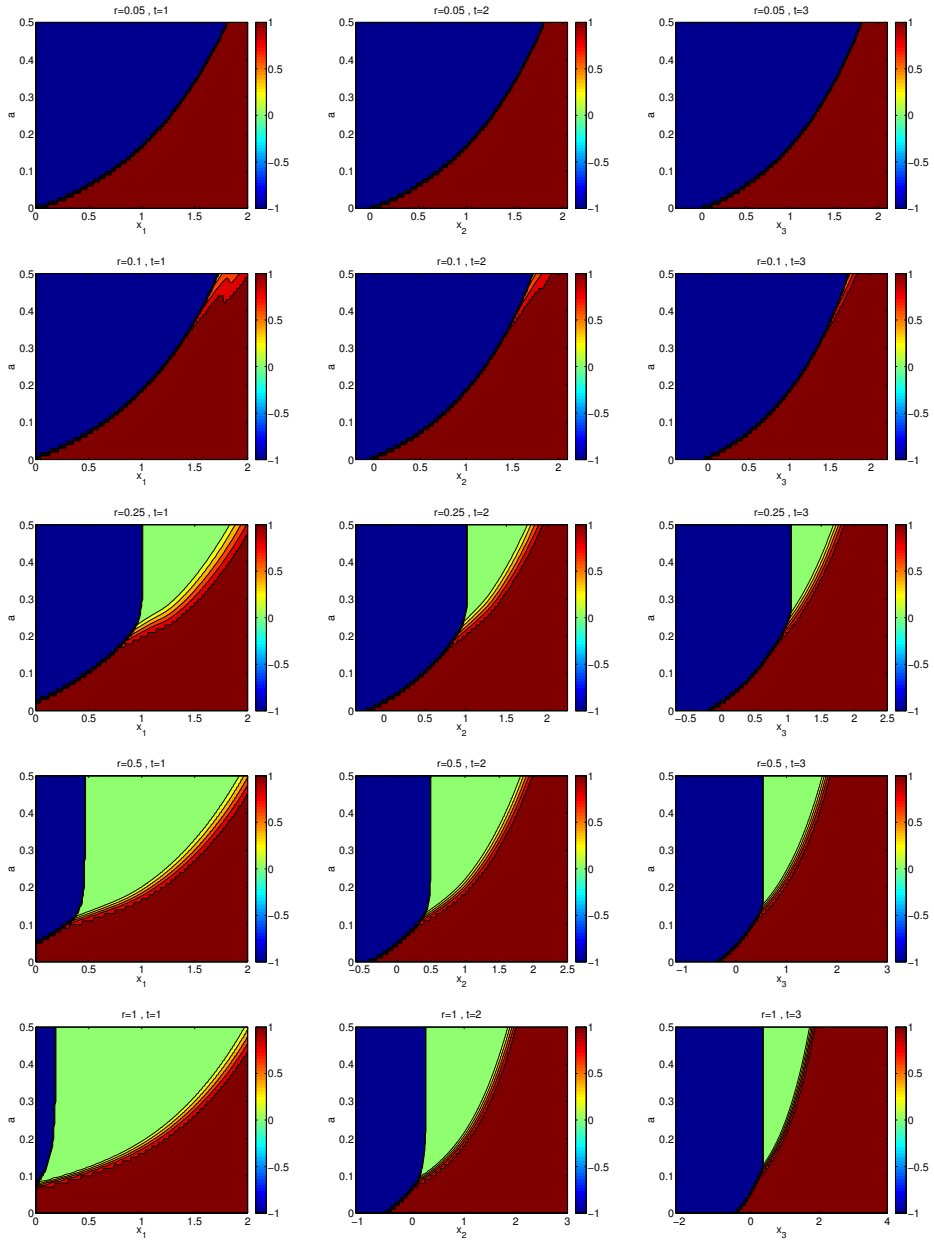


Figure 5.8: Contour plots of optimal dose for the left-skewed 3-period problem, varying the parameter r, which quantifies the size of the noise term in the state transition function. Columns correspond to session number t and rows to varying values of r. Color bar corresponds to optimal dose on [0,1]. A dose of -1 indicates a decision to stop treatment.

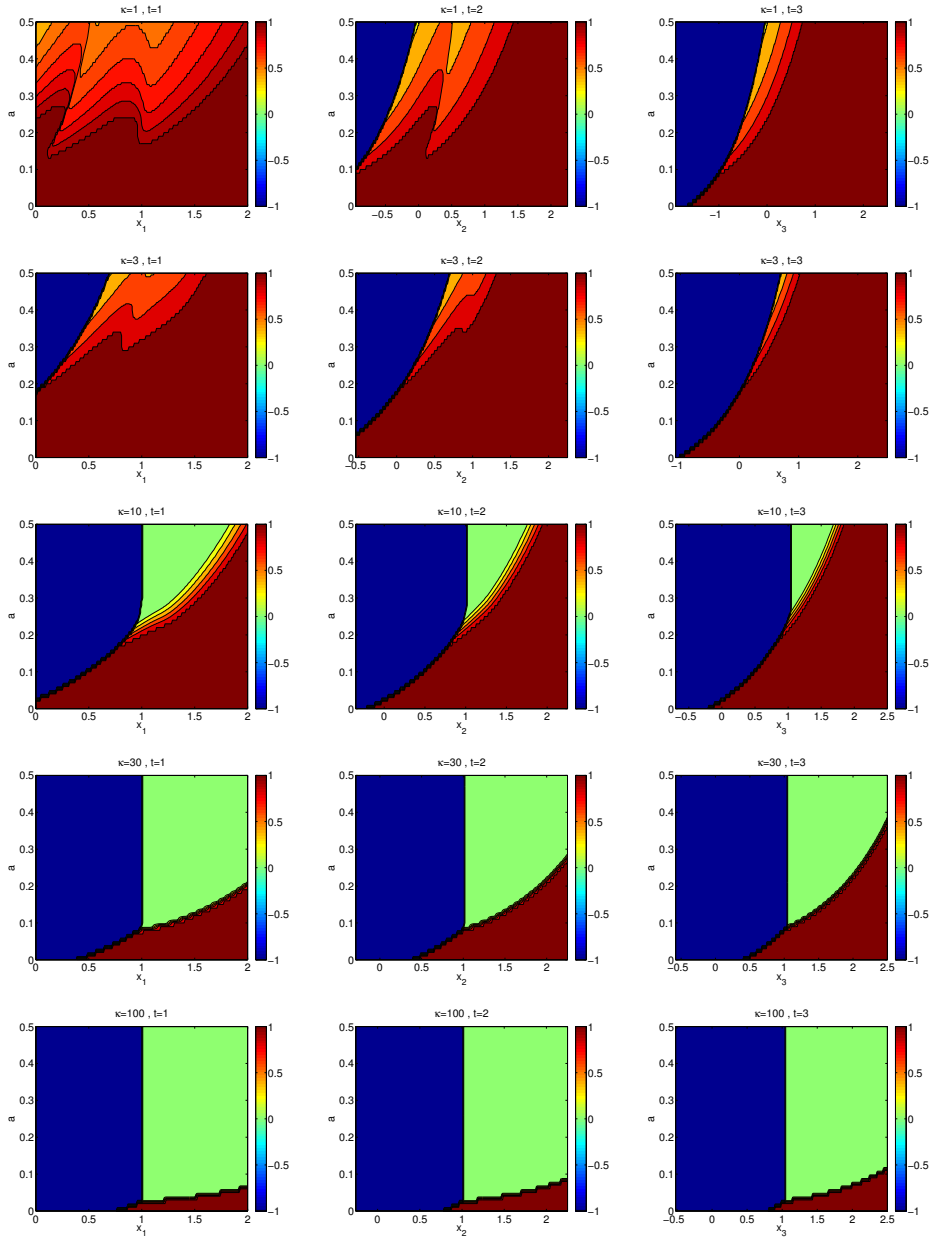


Figure 5.9: Contour plots of optimal dose for the left-skewed 3-period problem, varying the parameter κ, which quantifies the dose-response. Columns correspond to session number t and rows to varying values of κ. Color bar corresponds to optimal dose on [0,1]. A dose of -1 indicates a decision to stop treatment.

5.5 Stopping for rheumatoid arthritis

We now reconsider the rheumatoid arthritis problem based on OPTION trial data of Chapter 1. All parameters were set to the same values as Chapter 1, except we consider a slightly different cost function. In Chapter 1 we took the cost function to be $c(d) = 0.028557d$, where the coefficient came from solving the inverse optimization problem for the deterministic case. In the first frame of Figure 5.10, and with this cost function, we see that stopping is never optimal for the range of states we consider. However, by changing the cost function to an affine type, allowing for a fixed per-session cost: $c(d) = 0.028557d + b$, where $b \geq 0$ is a constant, we see that stopping indeed becomes optimal for some of the lowest disease states. This is intuitive, as for a very low disease state, the certain fixed per-session cost in upcoming sessions outweighs the possibility that the disease state will rise to a point where dose would be given, so a decision to stop is optimal. For intermediate state values, again we see a "wait-and-see" policy, where no dose is given but neither is treatment stopped, as the possibility of a flare-up to the point that positive dose is optimal is higher. For still higher disease states, we observe that a positive dose is given. Finally, we note that as b increases, the threshold state below which stopping is optimal increases, until the point where the "wait-and-see" region collapses. Again this is reasonable, as the decision to stop becomes more appealing the higher the per-session cost is. Eventually, we reach the point where the per-session fixed cost is so high that we only *do not* stop treatment if the patient's disease state is very high– otherwise, stopping is optimal.

A non-monotone policy is observed numerically for the case where $b = 0.15$ in Figure 5.10; a zoomed-in picture of the non-monotone region is shown in Figure 5.11. Again, this is possible since we have no monotonicity proof for the stopping problem.

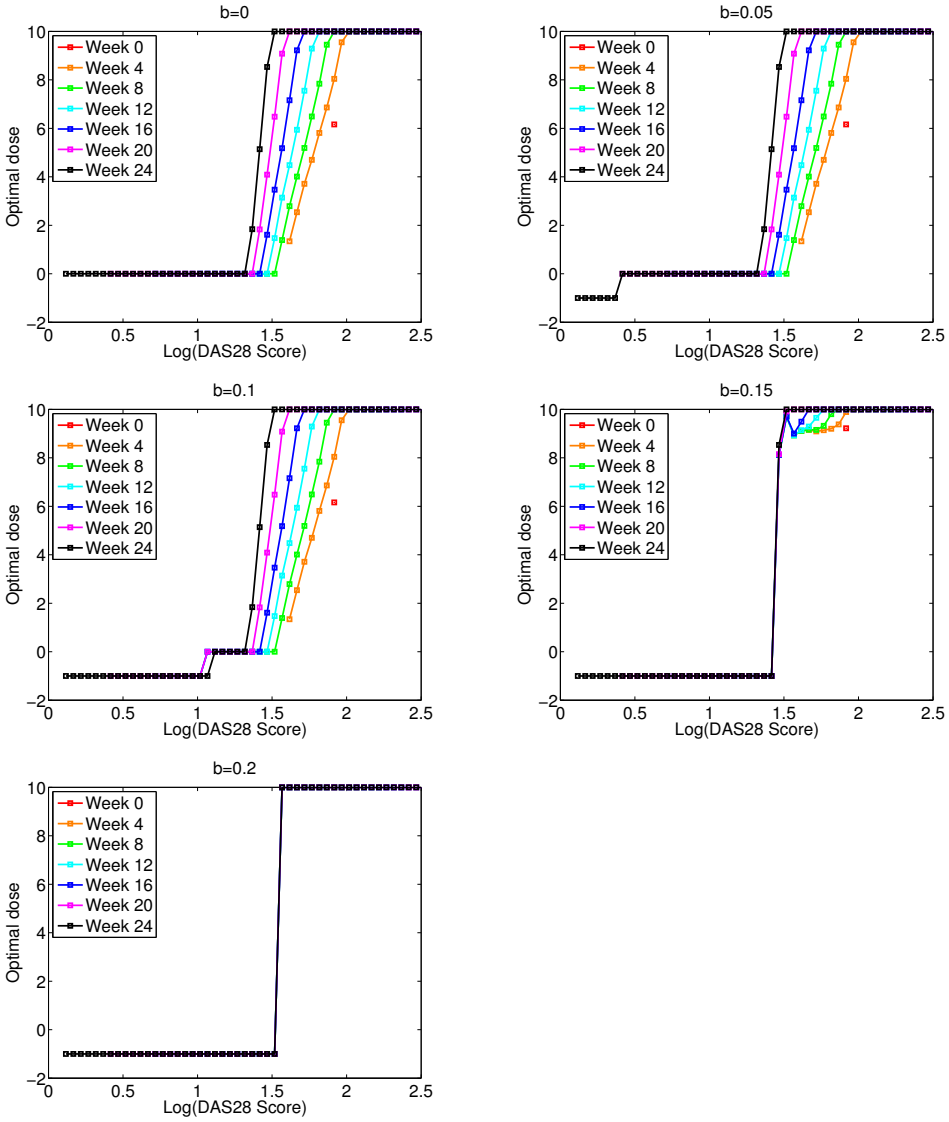


Figure 5.10: Optimal policy for the rheumatoid arthritis example of Chapter 1. All parameters and functions are the same as Chapter 1, except the cost function includes a fixed per-session cost b: $c(d) = 0.028557d + b$. $b = 0$ corresponds to precisely the example of Chapter 1 but allowing for the possibility of stopping. A dose of -1 indicates a decision to stop treatment.

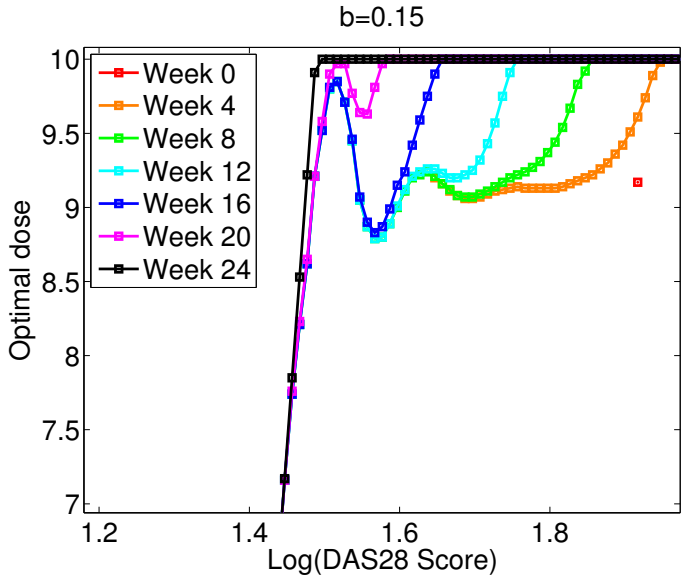


Figure 5.11: Optimal policy for the rheumatoid arthritis example of Chapter 1, except with the cost function $c(d) = 0.028557d + 0.15$. This is a zoom-in of Figure 5.10 with $b = 0.15$. Note the non-monotone dosing policy.

5.6 Monotonicity of stopping threshold state with respect to time

In our numerical experiments, we have observed that if stopping is ever optimal, it is optimal below a threshold state. Let us define this threshold state in session t as x_t^*. One question that naturally arises in the multi-period problem is whether x_t^* is a function of t, and if so, if x_t^* is monotone in one direction or the other.

We notice that in the $b = 0.1$ panel of Figure 5.10, x_t^* appears to increase between the week 20 and week 24 sessions. However, the picture is incomplete for two reasons. First, the initial state of $\log 6.8$ is significantly higher than x_t^* in any session such that it takes several sessions for the states to spread out enough due to the noise term until the states can be low enough that stopping ever occurs. Second, the grid-spacing in the x dimension is too coarse. We remedy both these issues with in Figure 5.12, which begins with a range of initial states over $1 \leq x_1 \leq 2.5$, and reduces the x-dimension spacing from 0.05 to 0.001. A zoom-in of Figure 5.12 around x_t^* is shown in Figure 5.13.

Figure 5.13 suggests that x_t^* is anything but constant, and in fact appears to monotonically increase with respect to session number. That is, the less treatment time remaining, the higher the threshold state between stopping and not stopping.

This result is not only numerically observed, but provable. A proof for a general DP problem with stopping is found in section 4.4, volume 1 of [25]. For convenience we provide a counterpart of this proof using our notation.

By Bellman's equations 5.2, it is optimal to stop at time t for all states in the set

$$T_t = \left\{ x \middle| h(x) \leq \min_{d \in D} c(d) + E\big[J_{t+1}(x + f_0(d; \Theta))\big] \right\} \tag{5.19}$$

Equation 5.2 along with the boundary condition of the DP, $J_T(x_T) = h(x_T)$, implies that

$$J_{T-1}(x) \leq J_T(x) \quad \forall x \tag{5.20}$$

Using equation 5.2 along with the stationarity property of our problem and the mono-

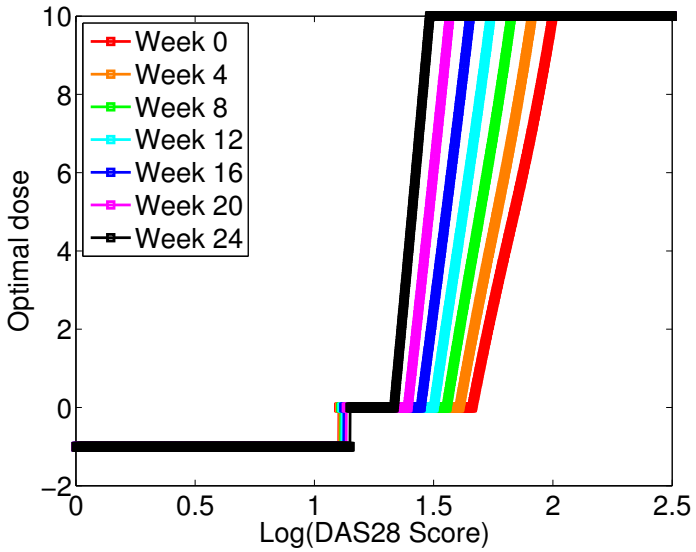


Figure 5.12: Optimal policy for the rheumatoid arthritis example of Chapter 1, except with the cost function $c(d) = 0.028557d + 0.1$. A wide initial range of states and reduced x-grid spacing is shown. A dose of -1 indicates a decision to stop.

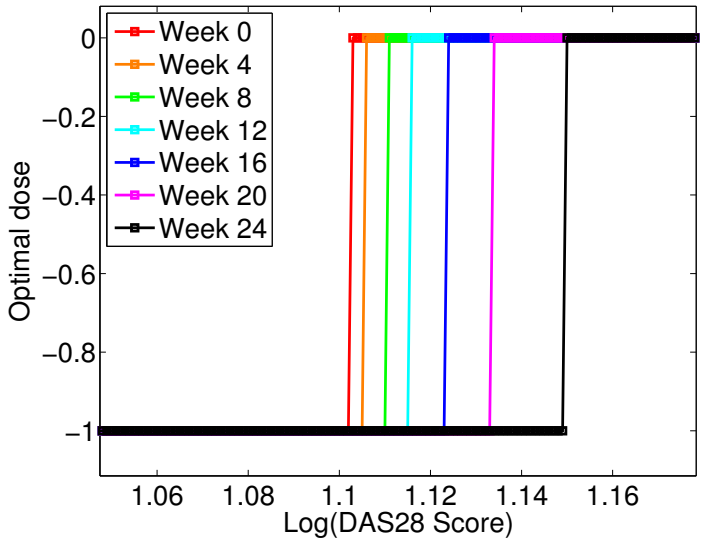


Figure 5.13: Optimal policy for the rheumatoid arthritis example of Chapter 1, except with the cost function $c(d) = 0.028557d + 0.1$. The plot is a zoomed-in version of Figure 5.12 around the threshold area between stopping and not stopping. A dose of -1 indicates a decision to stop.

tonicity property of DP, we obtain via induction

$$J_t(x) \leq J_{t+1}(x) \quad \forall x, t. \tag{5.21}$$

Using this fact, we see

$$T_1 \subset T_2 \subset \ldots \subset T_{N-1}. \tag{5.22}$$

In our numerical simulations we have observed that $T_t = (-\infty, x_t^*)$ for all t. This combined with 5.22 gives that the upper limit of the stopping set, x_t^*, increases monotonically with t.

We now discuss the conclusions of this dissertation, chapter-by-chapter.

Chapter 6

CONCLUSIONS AND FUTURE WORK

6.1 Chapter 1: RGD for rheumatoid arthritis

In Chapter 1, we established a rigorous mathematical framework for response-guided dosing. The resulting dosing policy was illustrated using data on RA. This policy administers higher doses in worse disease conditions. Our numerical simulations suggest that such monotone dosing performs better than three treatment strategies reported in the RA literature. Our sensitivity analyses reveal that optimal doses decrease with increasing aversion to dose and also with increasing efficacy of the placebo (methotrexate). The optimal dosing policy becomes more aggressive (the marginal increase in dose over two given states is higher) as response-uncertainty increases. Such sensitivity analyses provide quantitative guidelines for choosing doses in practice. For example, while it is known in the clinical literature that combination treatment with a biologic agent could be beneficial for patients who do not respond well to methotrexate alone, our methodology quantifies the benefit of various combination dosing strategies as a function of sensitivity to methotrexate. Dose-histograms from simulation experiments could provide a physician with a distribution of doses that would be administered to his/her patients if our methodology were implemented. Similarly, a physician could choose a point on the efficient frontier and then compute a corresponding dosing policy via the implied coefficient of risk aversion.

6.2 Chapter 2: A general stochastic DP formulation for RGD

In Chapter 2, we generalized the model of Chapter 1 to apply to other diseases, dose-response dynamics, and cost functions. We were able to prove that dose-monotonicity holds essentially when dose-response is convex and the decision-maker is risk-averse. To test the benefits of

such a dosing policy in practice, one would need to run a clinical trial where a group of patients is treated according to this policy and other groups are treated with constant doses in each session. The resulting treatment outcomes would need to be compared to quantify potential benefits if any. Our methodology, via computer simulations, could help physicians better-design such trials. Indeed, Murphy et al. [136] have espoused the use of such a priori simulations in the design of clinical trials.

One limitation of our model is the assumption that the model-parameters have been estimated before treatment begins. This limitation is addressed and alleviated in Chapters 3 and 4.

6.3 Chapter 3: Robust RGD

In Chapter 3 we presented a robust counterpart to our RGD framework. In our original framework, we assumed that the decision-maker knew the pmf of the dose-response parameter *a priori*. The robust counterpart softens this assumption by allowing the decision-maker to optimize against the worst case among a range of pmf's; in the language of robust optimization, this is the uncertainty set. Our analysis focused on the interval uncertainty set. However, the robust Bellman's equations (3.1) allow for other uncertainty sets such as entropy, maximum likelihood, and ellipsoidal. The inner problems for these sets are not LPs but they are convex and, at least for entropy and maximum likelihood, can be solved efficiently via a bisection algorithm [88, 139].

For the case of the interval uncertainty set, we have shown that the inner maximization problem in the Bellman's equations is an LP with a closed-form solution. We showed that this solution– that is, the worst-case pmf– does not depend on the state-action pair. This further implied the existence of a monotone optimal policy, which was the key result of our nominal (non-robust) model of Chapter 2. In addition, the fact that the worst-case pmf is independent of state-action pair makes the robust problem as difficult to solve as the nominal problem. We then focused on a specific case where the size of the uncertainty set was tuned with a single parameter δ, which we called the ambiguity level. For this case, we showed

that there exists a monotone optimal policy with respsect to δ under further assumptions on the dose-response function $f(\cdot;\cdot)$.

Numerical results agreed with our theoretical findings. For the inverse-power dose-response function, Figures 3.2, 3.3, and 3.4 showed monotonicity of dose with respect to state and δ. We also found a counterexample, the Michaelis-Menten dose-response function, which violated an assumption and numerically resulted in an optimal policy which was not monotone increasing in δ.

6.4 Chapter 4: Optimal Bayesian learning of dose-response parameters from a cohort

Adaptive clinical trials, in which pre-specified trial modifications are allowed based upon observed data during the trial, offer some advantages over the so-called fixed trials. Adaptive techniques are sometimes used during early stages of drug development to determine the minimum effective dose and the maximum tolerable dose. In Chapter 4, we introduced an adaptive scheme for learning the pmf of an unknown dose-response parameter while optimally dosing a cohort of patients. Our framework allows for essentially all single-parameter dose-response models. These include functions such as Michaelis-Menten, exponential, and power law. Furthermore, our approach can in principle be applied to a variety of diseases that are treated pharmaceutically over multiple treatment sessions, such as rheumatoid arthritis, LDL cholesterol, and high blood pressure.

Our framework was based upon a Bayesian stochastic DP. Because of the high dimensionality of the problem and corresponding intractability of an exact solution, we presented two approximate methods of solution: semi-stochastic CEC and CEC. Under reasonable assumptions that the dose-response function be monotone and convex, and the cost functions be increasing and convex, we proved desirable properties of these solution methods. For semi-stochastic CEC, we proved monotonicity of optimal dose with worsening disease conditions. We have also shown that the semi-stochastic CEC problem decomposes across patients when the disutility functions are additively separable, which yields significant sav-

ings in computation time. For CEC, we proved convexity of the deterministic problem, and the optimality of stationary dosing decisions.

Finally, we presented a simulation example to illustrate the empirical behavior of our approach. We worked with the following model primitives: a Michaelis-Menten dose-response function with a dose-response parameter Θ; a linear average per-session disutility function and exponential average terminal disutility function; a discretized Normal distribution of Θ to be learned, and a uniform initial prior on Θ. We emphasize that our framework is general and that none of these particular choices were necessary in order to run our model but were only chosen as an illustrative example. Our numerical results brought forth several intuitive properties of the doses delivered by our approximation methods relative to those administered by the clairvoyant approach and by two other simplistic strategies that do not adaptively learn the dose-response parameter. Our simulation runs compared the two approximation methods and found that they performed close to the clairvoyant approach. We hope that our work offers at least a small step toward better learning dose-response while treating a cohort of patients as envisioned by several expert panels.

Our Bayesian stochastic DP assumes that a dose-response function is given as an input to the learning problem and attempts to learn the parameter of this function while optimally dosing a cohort. A natural extension of this involves a more difficult problem where a set $\{f^1, f^2, \ldots, f^M\}$ of M different dose-response functions, each with a single parameter, is given. The pmf over these functions is $\vec{Q} \triangleq (Q^1, Q^2, \ldots, Q^M)$ such that $\sum_{m=1}^{M} Q^m = 1$. Given f^m, the pmf of its parameter is denoted by $\vec{p}^m \triangleq (p_1^m, \ldots, p_{k_m}^m)$. Here, k_m is the number of possible values this parameter can take, the corresponding set of values being $\{v_1^m, \ldots, v_{k_m}^m\}$. The decision-maker now wishes to learn $\vec{Q}$ as well as $\vec{p}^m$, for $m = 1, 2, \ldots, M$, while optimally dosing a cohort of patients. This problem could provide an interesting direction for future research.

6.5 Chapter 5: Optimal stopping for RGD

Finally, in Chapter 5, we presented an extension of the stochastic DP model of Chapter 2 where the decision-maker can decide to stop treatment at any treatment session. If a decision to stop is made in the current period, all future per-session costs are avoided, and the patient's final disease state is taken to be the current state. Intuitively, we expect the decision to stop will be optimal, if ever, at low disease states. At these states, the future per-session cost outweighs the benefit of lowering the disease state through treatment. In some cases, we may also find the existence of a "wait-and-see" region, where zero dose is given in a particular session, but the decision to stop is not made– this can incur a per-session cost, but the possibility of giving a dose later to lower the disease state outweighs that per-session cost, so we continue. At the highest disease states, positive doses are given as the benefit of reducing disease state wins out over the per-session costs.

Again, the problem is solved using backwards induction of the Bellman's equations with stopping (5.2) by discretizing the state-action space. First, we presented a motivating example before moving in to a more detailed example of a 3-period problem. We considered a linear affine cost function $c(d) = ad + b$ and investigated the effects of changing the fixed per-session cost b as well as the coefficient of cost per unit dose a. We took the probability mass function of the stochastic noise term to be a single-parameter tent function, and observed the effects of skewing the tent function left and right. We also considered a case where the fixed per-session cost b is 0 in the final session $t = T$, but nonzero in earlier sessions and found similar results.

Finally, we reconsidered the rheumatoid arthritis example of Chapter 2 again, but this time allowing for stopping. For the original problem, stopping was never optimal over the states considered. However, by adding a fixed per-session cost to the cost function $c(d) = 0.28557d + b$, we found that stopping is optimal for some of the lowest disease states. This indicates that stopping is optimal when the future fixed per-session costs outweigh the potential benefit of giving dose later. We also observed an optimal policy that was not

monotone, despite fulfilling the assumptions for monotonicity of Chapter 2. This is certainly possible as we have no monotonicity proof for the stopping framework, but it is interesting to note that such a counterexample did occur.

In the literature, stopping is mentioned not only for patients in very low disease states (remission,) but also sometimes for very high disease states, as this indicates a failure of the drug to have an effect on the patient. In practice, this would often indicate the need to switch to a different drug or treatment scheme. As we were only considering the dose of a single drug in our framework, this situation did not arise for us, but could be another interesting direction for future work.

6.6 Further future work

In addition to the extensions of each chapter mentioned above, there are several more directions that the work in this dissertation can be extended. Here we give a brief outline of two.

6.6.1 Hard dosing constraints

In our DP model, the aversion to dose in each session was modeled with a cost function $c : d \to \mathbb{R}$. With the assumption that $c(\cdot)$ is increasing in d, this amounts to a soft constraint on dose given: a higher dose is penalized more than a lower dose by virtue of the fact that there is a higher associated cost. An alternative framework considers dose to be a hard constraint. The simplest example would be the case where the total dose over multiple sessions must not exceed a threshold. Here we present an alternative formulation where information about the total adverse effect of all previous dosing sessions is included in the DP state.

For $t = 1, 2, \ldots, T + 1$, let E_t denote the total adverse effect of doses used in sessions $1, 2, \ldots, t-1$. Thus, we have $E_1 = 0$. For $t = 1, \ldots, T$, let B_t be the maximum total adverse effect that can be tolerated in the first t sessions. Intuitively, these parameters should satisfy $B_1 \leq B_2 \leq \ldots \leq B_T$. Let $\phi : D \to \mathbb{R}_+$ be an increasing, convex function that quantifies

the adverse effect of dose administered in one session. For instance, in radiobiology, upper limits are given on the biologically equivalent dose (BED) given to certain organs-at-risk in the vicinity of the tumor. BED is given by the formula $\phi(d_t) = d_t + d_t^2/(a/b)$, where $a, b \in \mathbb{R}_+$ are parameters. Another, simpler example would be where $\phi(d_t) = d_t$. Then we have $E_{t+1} = E_t + \phi(d_t)$ and dose $d_t \geq 0$ must be chosen such that $E_{t+1} \leq B_{t+1}$. We model this scenario by defining a state variable $s_t = B_t - E_t$. This is the slack in adverse effect of doses delivered in sessions $1, 2, \ldots, t-1$, as compared to the tolerance limit B_t. Then notice that $E_{t+1} = B_t - s_t + \phi(d_t)$ and hence $s_{t+1} = B_{t+1} - E_{t+1} = B_{t+1} - B_t + s_t - \phi(d_t)$. Define parameters $\Delta B_t \triangleq B_{t+1} - B_t \geq 0$ for $t = 1, 2, \ldots, T$. Thus, given the state variable s_t, dose $d_t \geq 0$ must be chosen such that $s_{t+1} \geq 0$, i. e., such that $\phi(d_t) \leq \Delta B_t + s_t$. Consequently, starting with the boundary condition $J_{T+1}(x, s) = \psi(x)$ for all $x \in X$ and $0 \leq s \leq B_T$, Bellman's equations (2.2) now change to

$$J_t(x_t, s_t) = \min_{\{d_t \geq 0 : \phi(d_t) \leq \Delta B_t + s_t\}} \int_{\Omega} J_{t+1}(f_t(x_t, d_t; \theta_t), \Delta B_t + s_t - \phi(d_t)) p(\theta) d\theta, \tag{6.1}$$

for all $x_t \in X$, all $s_t \in [0, B_t]$, and $t = 1, 2, \ldots, T$.

Owing to its two-dimensional state, this DP is computationally more challenging to solve as compared to its counterpart (2.2). It will be interesting in the future to develop efficient computational procedures for solving this DP. We believe that this will be possible because a monotone policy where doses increases with worsening disease conditions x_t and increasing slacks s_t is likely to be optimal for (6.1).

6.6.2 Partial observations

Suppose that when the disease state is x_t, its measurement is given by $z_t = \eta_t x_t + \Phi_t$, where η_t is a known coefficient and Φ_t are iid random variables representing measurement noise that does not depend on disease response uncertainties Θ, disease states, doses, and measurements. Suppose random variables Φ_t take values from a set $\Lambda \subseteq \Re$, and have a probability density function $\pi(\cdot)$. Then this problem with imperfect state information can be converted into a stochastic DP whose state at the beginning of session t is given by

a probability distribution $\mathcal{P}_t$ over X. This is called the belief distribution and is known to be a sufficient statistic for the information history of the controlled stochastic process. Unfortunately, the belief distributions and hence the states of this stochastic DP are infinite dimensional, thus making exact solution intractable.

First we can use discretization and truncation to approximate the problem as a partially-observable stochastic DP with finite states, finite actions, and finite observations. Then the belief state is finite-dimensional although it is still continuous. We can then exploit the fact that the cost-to-go functions in the corresponding DP are known to be piecewise linear. Although this DP is known to be computationally hard in theory, the hope is that we will be able to efficiently find good policies in some problems ([118, 119, 146, 173, 179]). This was, for instance, the case in mammography screening problems in [14] and treatment for ischemic heart disease in [74].

BIBLIOGRAPHY

[1] M Abramowitz and I A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover, New York, New York, USA, 1972.

[2] S U Acikgoz and U M Diwekar. Blood glucose regulation with stochastic optimal control for insulin-dependent diabetic patients. *Chemical Engineering Science*, 48(3):1227, 2010.

[3] S K Agarwal. Biologic agents in rheumatoid arthritis: an update for managed care professionals. *Journal of Managed Car Pharmacy*, 17(9 Supplement B):S14–S18, 2011.

[4] R K Ahuja and J B Orlin. Inverse optimization part I: Linear programming and general problem. *Operations Research*, 49(5):771–783, 2001.

[5] V Ahuja and J Birge. Response-adaptive designs for clinical trials: simultaneous learning from multiple patients. `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2126906`, 2014.

[6] O Alagoz, H Hsu, A J Schaefer, and M S Roberts. Markov decision processes: A tool for sequential decision making under uncertainty. *Medical Decision Making*, 30(4):474–483, 2010.

[7] O Alagoz, L M Maillart, A J Schaefer, and M S Roberts. The optimal timing of living-donor liver transplantation. *Management Science*, 50(10):1420–1430, 2004.

[8] O Alagoz, L M Maillart, A J Schaefer, and M S Roberts. Determining the acceptance of cadaveric livers using an implicit model of the waiting list. *Operations Research*, 55(1):24–36, 2007.

[9] C D Aliprantis and K C Border. *Infinite-dimensional Analysis: A Hitchhiker's Guide*. Springer-Verlag, Berlin, Germany, 1994.

[10] C F Allaart, Y P Goekoop-Ruiterman, J K de Vries-Bouwstra, F C Breedveld, B A Dijkmans, and FARR study group. Aiming at low disease activity in rheumatoid arthritis with initial combination therapy or initial monotherapy strategies: the best study. *Clinical and Experimental Rheumatology*, 24(6 (Supplement 43)):S77–S82, 2006.

[11] J Ananworanich, A Gayet-Ageron, M Le Braz, W Prasithsirikul, P Chetchotisakd, S Kiertiburanakul, W Munsakul, P Raksakulkarn, S Tansuphasawasdikul, S Sirivichayakul, M Cavassini, U Karrer, D Genne, R Nuesch, P Vernazza, E Bernasconi, D Leduc, C Satchell, S Yerly, L Perrin, A Hill, T Perneger, P Phanuphak, H Furrer, D Cooper, K Ruxrungtham, and B Hirschel. CD4-guided scheduled treatment interruptions compared with continuous therapy for patients infected with HIV-1: results of the Staccato randomised trial. *The Lancet*, 368(9534):459 – 465, 2006.

[12] J Ananworanich, U Siangphoe, P Cardiello, W Apateerapong, B Hirschel, A Mahanontharit, S Ubolyam, D Cooper, P Phanuphak, and K Ruxrungtham. Highly active antiretroviral therapy (HAART) retreatment in patients on CD4-guided therapy achieved similar virologic suppression compared with patients on continuous HAART - The HIV Netherlands Australia Thailand Research Collaboration 001.4 Study. *Journal of Acquired Immune Deficiency Syndromes*, 39(5):523–529, 2005.

[13] T M Apostol. *Mathematical Analysis*. Addison Wesley, Reading, Massachusetts, USA, 1974.

[14] T Ayer, O Alagoz, and N K Stout. A POMDP approach to personalize mammography screening decisions. *Operations Research*, 60:1017–1021, 2012.

[15] J Babb, A Rogatko, and S Zacks. Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistic in Medicine*, 17(10):1103–1120, 1998.

[16] P Bauer and J Roehmel. An adaptive method for establishing a dose-response relationship. *Statistics in Medicine*, 14:1595–1607, 1995.

[17] D S Bayard, M H Milman, and A Schumitzky. Design of dosage regimens: a multiple model stochastic control approach. *International Journal of Bio-Medical Computing*, 36:103–115, 1994.

[18] D R Beil. Analysis and comparison of multimodal cancer treatments. *Mathematical Medicine and Biology*, 18(4):343–376, 2001.

[19] D R Beil and L M Wein. Sequencing surgery, radiotherapy and chemotherapy: Insights from a mathematical analysis. *Breast Cancer Research and Treatment*, 74(3):279–286, 2002.

[20] R Bellman. A problem in the sequential design of experiments. *Sankhya: The Indian Journal of Statistics*, 16(3/4):221–229, 1956.

[21] A Ben-Tal, L El Ghaoui, and A Nemirovski. *Robust Optimization*. Princeton University Press, Princeton, NJ, USA, 2009.

[22] D A Berry. Modified two-armed bandit strategies for certain clinical trials. *Journal of the American Statistical Association*, 73(362):339–345, Jun 1978.

[23] D A Berry. Bayesian clinical trials. *Nature Reviews Drug Discovery*, 5:27–36, Jan 2006.

[24] D A Berry. Adaptive clinical trials: The promise and the caution. *Journal of Clinical Oncology*, 29(6):606–609, 2011.

[25] D P Bertsekas. *Dynamic Programming and Optimal Control*, volume 1 and 2. Athena Scientific, Nashua, NH, third edition, 2007.

[26] P Bolton and C Harris. Strategic experimentation. *Econometrica*, 67(2):349–374, 1999.

[27] B Bornkamp, F Bretz, H Dette, and J Pinheiro. Response-adaptive dose-finding under model uncertainty. *The Annals of Applied Statistics*, 5(2B):1611–1631, 2011.

[28] B Bornkamp, F Bretz, A Dmitrienko, G Enas, B Gaydos, C-H Hsu, F Koenig, M Krams, Q Liu, B Neuenschwander, T Parke, J Pinheiro, A Roy, R Sax, and F Shen. Innovative approaches for designing and analyzing adaptive dose-ranging trials. *Journal of Biopharmaceutical Statistics*, 17:965–95, 2007.

[29] S Boyd and L Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

[30] F C Breedveld and B Combe. Understanding emerging treatment paradigms in rheumatoid arthritis. *Arthritis Research & Therapy*, 13(Supplement 1):S3, 2011.

[31] F C Breedveld, M H Weisman, A F Kavanaugh, S B Cohen, K Pavelka, R van Vollenhoven, J Sharp, J L Perez, and G T Spencer-Green. The PREMIER study: A multicenter, randomized, double-blind clinical trial of combination therapy with adalimumab plus methotrexate versus methotrexate alone or adalimumab alone in patients with early, aggressive rheumatoid arthritis who had not had previous methotrexate treatment. *Arthritis and Rheumatism*, 54(1):26–37, 2006.

[32] Michael E Burton, L M Shaw, J J Schentag, and W E Evans. *Applied Pharmacokinetics & Pharmacodynamics : Principles of Therapeutic Drug Monitoring*. Lippincott Williams & Wilkins, Baltimore, Maryland, USA, 4th edition, 2006.

[33] W Burton, A Morrison, R Maclean, and E Ruderman. Systematic review of studies of productivity loss due to rheumatoid arthritis. *Occupational Medicine*, 56(1):18–27, 2006.

[34] P G Cardiello, E Hassink, J Ananworanich, P Srasuebkul, T Samor, A Mahanontharit, K Ruxrungtham, B Hirschel, J Lange, P Phanuphak, and D A Cooper. A prospective,

randomized trial of structured treatment interruption for patients with chronic hiv type 1 infection. *Clinical Infectious Diseases*, 40(4):594–600, 2005.

[35] J D Carroll. The 15 best-selling drugs of 2012. `http://www.fiercepharma.com/special-reports/15-best-selling-drugs-2012`, October 2012.

[36] T C Y Chan, T Craig, T Lee, and M B Sharpe. Generalized inverse multi-objective optimization with application to cancer therapy. *Operations Research*, 62(3):680–695, 2014.

[37] M Chang and S-C Chow. A hybrid bayesian adaptive design for dose response trials. *Journal of Biopharmaceutical Statistics*, 15:677–91, 2005.

[38] P Chaudhari. The impact of rheumatoid arthritis and biologics on employers and payers. *Biotechnology Healthcare*, 5(2):37–44, 2008.

[39] Y-F Chen, P Jobanputra, P Barton, S Jowett, S Bryan, W Clark, A Fry-Smith, and A Burls. A systematic review of the effectiveness of adalimumab, etanercept and infliximab for the treatment of rheumatoid arthritis in adults and an economic evaluation of their cost-effectiveness. *Health Technology Assessment*, 10(42):1–229, 2006.

[40] J Chhatwal, O Alagoz, and E S Burnside. Optimal breast biopsy decision making based on mammographic features and demographic factors. *Operations Research*, 58(6):1577–1591, 2010.

[41] S-C Chow and M Chang. Adaptive design methods in clinical trials - a review. *Orphanet Journal of Rare Diseases*, 3(11), 2008.

[42] EW St Clair, D van de Heijde, J S Smolen, R N Maini, J M Bathon, P Emery, E Keystone, M Schiff, J R Kalden, B Wang, K Dewoody, R Weiss, and D Baker. Active-controlled study of patients receiving infliximab for the treatment of rheumatoid arthritis of early onset study group: Combination of infliximab and methotrexate

therapy for early rheumatoid arthritis. *Arthritis and Rheumatism*, 50(11):3432–3443, 2004.

[43] B Combe and R F van Vollenhoven. Novel targeted therapies: the future of rheumatoid arthritis? mavrilumab and tabalumab as examples. *Annals of Rheumatoid Disease*, 72(9):1433–1435, 2013.

[44] C Danel, R Moh, A Minga, A Anzian, O Ba-Gomis, C Kanga, G Nzunetu, D Gabillard, F Rouet, S Sorho, M-L Chaix, Serge Eholie, H Menan, D Sauvageot, E Bissagnene, R Salamon, and X Anglaret. Cd4-guided structured antiretroviral treatment interruption strategy in hiv-infected adults in west africa (trivacan anrs 1269 trial): a randomised trial. *The Lancet*, 367(9527):1981 – 1989, 2006.

[45] G L Davis, J B Wong, J G McHutchison, M P Manns, J Harvey, and J Albrecht. Early virologic response to treatment with peginterferon alfa-2b plus ribavirin in patients with chronic hepatitis c. *Hepatology*, 38(3):645–52, Sep 2003.

[46] I Demonty, R T Ras, H C M van der Knaap, G Duchateau, L Meijer, P L Zock, J M Geleijnse, and E A Trautwein. Continuous dose-response relationship of the ldl-cholesterol–lowering effect of phytosterol intake. *Journal of Nutrition*, 139(2):271–284, 2009.

[47] A A den Broeder, M C W Creemers, A M van Gestel, and P L C M van Riel. Dose titration using the Disease Activity Score (DAS28) in rheumatoid arthritis patients treated with antiTNF. *Rheumatology*, 41(6):638–642, 2002.

[48] B T Denton, M Kurt, N D Shah, S C Bryant, and S A Smith. Optimizing the start time of statin therapy for patients with diabetes. *Medical Decision Making*, 29(3):351–67, 2009.

[49] B Van der Cruyssen, S Van Looy, B Wyns, R Westhovens, P Durez, F Van den Bosch, E M Veys, H Mielants, L De Clerck, A Peretz, M Malaise, L Verbruggen, N Vaste-

saeger, A Geldhof, L Boullart, and F De Keyser. DAS28 best reflects the physician's clinical judgment of response to infliximab therapy in rheumatoid arthritis patients: validation of the DAS28 score in patients under infliximab treatment. *Arthritis Res Ther*, 7(5):R1063–71, 2005.

[50] P Durez, F Van den Bosch, L Corluy, E M Veys, L De Clerck, A Peretz, M Malaise, J P Devogelaer, A Geldhof, and R Westhoven. A dose adjustment in patients with rheumatoid arthritis not optimally responding to a standard dose of infliximab of 3mg/kg every 8 weeks can be effective: a Belgian prospective study. *Rheumatology*, 44(4):465–468, 2005.

[51] S D Durham, N Flournoy, and W F Rosenberger. A random walk rule for phase I clinical trials. *Biometrics*, 53(2):745–760, 1997.

[52] L Eeckhoudt. *Risk and Medical Decision Making*. Kluwer Academic, Norwell, MA, USA, 2002.

[53] M Ehrgott, C Guler, H W Hamacher, and L Shao. Mathematical optimization in intensity modluated radiation therapy. *4OR*, 6:199–262, 2008.

[54] P Emery, A Sebba, and T W J Huizinga. Biologic and oral disease-modifying antirheumatic drug monotherapy in rheumatoid arthritis. `http://ard.bmj.com/content/early/2013/09/17/annrheumdis-2013-203485.full`, August 2013.

[55] Z Erkin, M D Bailey, L M Maillart, A J Schaefer, and M S Roberts. Eliciting patient's revealed preferences: an inverse Markov decision process approach. *Decision Analysis*, 7(4):358–365, 2010.

[56] S R Eussen, C J Rompelberg, O H Klungel, and J C van Eijkeren. Modelling approach to simulate reductions in ldl cholesterol levels after combined intake of statins and phytosterols/-stanols in humans. *Lipids in Health and Disease*, 10(187):1–10, 2011.

[57] Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). *Journal of the American Medical Association*, 285(19):2486–97, May 2001.

[58] PI Feder, DW Hobson, CT Olson, RL Joiner, and MC Matthews. Stagewise, adaptive dose allocation for quantal response dose-response studies. *Neuroscience and Biobehavioral Reviews*, 15:109–14, 1991.

[59] M Flendrie, M C W Creemers, and P L C M van Riel. Titration of infliximab treatment in rheumatoid arthritis patients based on response patterns. *Rheumatology (Oxford)*, 46(1):146–9, Jan 2007.

[60] U.S. Food and Drug Administration. Industry guideline: Dose-response information to support drug registration. `http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm073115.pdf`, November 1994.

[61] U.S. Food and Drug Administration. Guidance for industry: adaptive design clinical trials for drugs and biologics. `http://www.fda.gov/downloads/DrugsGuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf`, February 2010.

[62] J Fransen, G Stucki, and P L C M van Riel. Rheumatoid arthritis measures: Disease activity score (das), disease activity score-28 (das28), rapid assessment of disease activity in rheumatology (radar), and rheumatoid arthritis disease activity index (radai). *Arthritis Care & Research*, 49(S5):S214–S224, 2003.

[63] J Fransen and P L C M van Riel. The disease activity score and the eular response criteria. *Rheum Dis Clin North Am*, 35(4):745–57, vii–viii, Nov 2009.

[64] B Friedlin and EL Korn. Biomarker-adaptive clinical trial designs. *Pharmacogenomics*, 11(12):1679–82, 2010.

[65] M Gasparini and J Eisele. A curve-free method for phase i clinical trials. *Biometrics*, 56(2):609–15, Jun 2000.

[66] C Gaujoux-Viala, G Mouterde, A Baillet, P Claudepierre, B Fautrel, X Le Loet, and J-F Maillefert. Evaluating disease activity in rheumatoid arthritis: Which composite index is best? a systematic literature analysis of studies comparing the psychometric properties of the DAS, DAS28, SDAI and CDAI. *Joint Bone Spine*, 79(2):149 – 155, 2012.

[67] J C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, pages 148–177, 1979.

[68] AP Grieve and M Krams. Astin: a bayesian adaptive dose-response trial in acute stroke. *Clinical Trials*, 2(340-51), 2005.

[69] C Grigor, H Capell, A Stirling, A D McMahon, P Lock R Vallance, W Kincaid, and D Porter. Effect of a treatment strategy of tight control for rheumatoid arthritis (the TICORA study): a single-blind randomised controlled trial. *The Lancet*, 364(9830):17–23, 2004.

[70] SMART Study Group. Cd4+ count–guided interruption of antiretroviral treatment. *New England Journal of Medicine*, 355(22):2283–2296, 2006.

[71] LM Haines, I Perevozskaya, and WF Rosenberger. Bayesian optimal designs for phase I clinical trials. *Biometrics*, 59:591–600, Sep 2003.

[72] E J Hall and A J Giaccia. *Radiobiology for the Radiologist*. Lippincott Williams & Wilkins, Philadelphia, Pennsylvania, USA, 2005.

[73] D Hasenclever, M Loeffler, and V Diehl. Rationale for dose escalation of first line conventional chemotherapy in advanced hodgkin's disease. *Annals of Oncology*, 7(Supplement 4):S95–S98, 1996.

[74] M Hauskrecht and H Fraser. Planning treatment of ischemic heart disease with partially observable markov decision processes. *Artif Intell Med*, 18(3):221–44, Mar 2000.

[75] R A Hayward and H M Krumholz. Three reasons to abandon low-density lipoprotein targets: An open letter to the adult treatment panel iv of the national institutes of health. *Circulation: Cardiovascular Quality and Outcomes*, 5(1):2–5, 2012.

[76] R A Hayward, H M Krumholz, D M Zulman, J W Timbie, and S Vijan. Optimizing statin treatment for primary prevention of coronary artery disease. *Annals of Internal Medicine*, 152:69–77, 2010.

[77] PubMed Health. *Rheumatoid Artritis*, February 2012.

[78] J E Helm, M S Lavieri, M P Van Oyen, J D Stein, and D C Musch. Dynamic forecasting and control algorithms for glaucoma progression for clinical decision support. `http://sitemaker.umich.edu/jhelm/files/dynamic-forecasting-control-glaucoma-or.pdf`.

[79] C G Helmick, D T Felson, R C Lawrence, S Gabriel, R Hirsch, C K Kwoh, M H Liang, H M Kremers, M D Mayes, P A Merkel, S R Pillemer, J D Reveille, J H Stone, and National Arthritis Data Workgroup. Estimates of the prevalence of arthritis and other rheumatic conditions in the united states: Part i. *Arthritis and Rheumatism*, 58(1):15–25, 2008.

[80] R S Hogg, D Havlir, V Miller, and J Montaner. To stop or not to stop: That is the question, but what is the answer? *AIDS*, 16:787–789, 2002.

[81] C Hu, W S Lovejoy, and S L Shafer. Comparison of some suboptimal control policies in medical drug therapy. *Operations Research*, 44(5):696–709, 1996.

[82] K L Hyrich, D P Symmons, K D Watson, and A J Silman and. Comparison of the response to infliximab or etanercept monotherapy with the response to cotherapy with methotrexate or another disease-modifying antirheumatic drug in patients with rheumatoid arthritis: results from the british society for rheumatology biologics register. *Arthritis and Rheumatism*, 54(6):1786–1794, 2006.

[83] A Iasonos and J O'Quigley. Adaptive dose-finding studies: a review of model-guided phase I clinical trials. *Journal of Clinical Oncology*, 32(12), August 2014.

[84] Amgen Inc. Biologics and biosimilars. `http://www.amgen.com/pdfs/misc/Biologics_and_Biosimilars_Overview.pdf`, 2012.

[85] A Ivanova. A play-the-winner-type urn design with reduced variability. *Metrika*, 58:1–13, 2003.

[86] A Ivanova, JA Bolognese, and I Perevozskaya. Adaptive dose finding based on t-statistic for dose-response trials. *Stat Med*, 27(10):1581–92, May 2008.

[87] G Iyengar and W Kang. Inverse conic programming with applications. *Operations Research Letters*, 33(3):785–797, 2005.

[88] G N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[89] R W Jelliffe, P Maire, Fred Sattler, P Gomis, and B Tahani. Adaptive control of drug dosage regimens: basic foundations, relevant issues, and clinical examples. *International Journal of Bio-Medical Computing*, 36:1–23, 1994.

[90] N Joharatnam, D F McWilliams, D Wilson, M Wheeler, I Pande, and D A Walsh. A cross-sectional study of pain sensitivity, disease-activity assessment, mental health, and fibromyalgia status in rheumatoid arthritis. *Arthritis Research & Therapy*, 17(1):1–9, 2015.

[91] W Katchamart and C Bombardier. Systematic monitoring of disease activity using an outcome measure improves outcomes in rheumatoid arthritis. *Journal of Rheumatology*, 37:1411–1415, 2010.

[92] M N Katehakis and A F Veinott Jr. The multi-armed bandit problem: decomposition and computation. *Mathematics of Operations Research*, 12(2):262–268, 1987.

[93] A Kavanaugh, S Cohen, and J J Cush. The evolving use of tumor necrosis factor inhibitors in rheumatoid arthritis. *The Journal of Rheumatology*, 31(10):1881–1884, 2004.

[94] D M Keenan, R Basu, Y Liu, A Basu, G Bock, and J D Veldhuis. Logistic model of glucose-regulated c-peptide secretion: hysteresis pathway disruption in impaired fasting glycemia. *American Journal of Physiology - Endocrinology and Metabolism*, 303(3):E397–E409, 2012.

[95] G Keller and S Rady. Strategic experimentation with poisson bandits. *Theoretical Economics*, 5(2):275–311, 2010.

[96] G Keller, S Rady, and M Cripps. Strategic experimentation with exponential bandits. *Econometrica*, 73(1):39–68, 2005.

[97] M Kim, A Ghate, and M H Phillips. A markov decision process approach to temporal modulation of dose fractions in radiation therapy planning. *Physics in Medicine and Biology*, 54:4455–4476, 2009.

[98] M Kim, A Ghate, and M H Phillips. A stochastic control formalism for dynamic biologically conformal radiation therapy. *European Journal of Operational Research*, 219(3):541 – 556, 2012.

[99] S King. The best selling drugs of all time; humira joins the elite. `http://www.forbes.com/sites/simonking/2013/01/28/`

the-best-selling-drugs-of-all-time-humira-joins-the-elite/, January 2013.

[100] S Kirby, P Brain, and B Jones. Fitting emax models to clinical trial dose-response data. *Pharmaceutical Statistics*, 10(2):143–149, 2011.

[101] R Knevel, M Schoels, and T W J Huizinga et al. Current evidence for a strategic approach to the management of rheumatoid arthritis with disease-modifying antirheumatic drugs: a systematic literature review informing the eular recommendations for the management of rheumatoid arthritis. *Annals of Rheumatoid Disease*, 69:987–994, 2010.

[102] M Krams, KR Lees, W Hacke, AP Grieve, J-M Orgogozo, GA Ford, and ASTIN Study Investigators. Acute stroke therapy by inhibition of neutrophils. *Stroke*, 34:2543–8, Oct 2003.

[103] H M Krumholz and R A Hayward. Shifting views on lipid lowering therapy. *BMJ*, 341:c3531, 2010.

[104] TM Kuijper, FBG Lamers-Karnebeek, JWG Jacobs, JMW Hazes, and JJ Luime. Flare rate in patients with rheumatoid arthritis in low disease activity or remission when tapering or stopping synthetic or biologic dmard: a systematic review. *Journal of Rheumatology*, 42(11):2012–2022, 2015.

[105] P R Kumar. A survey of some results in stochastic adaptive control. *SIAM Journal on Control and Optimization*, 23(3):329–380, 1985.

[106] B Kuriya, E V Arkema, V P Bykerk, and E C Keystone. Efficacy of initial methotrexate monotherapy versus combination therapy with a biologic agent in early rheumatoid arthritis: a meta-analysis of clinical and radiographic remission. *Annals of Rheumatoid Disease*, 69(7):1298–1304, 2010.

[107] M Kurt, B T Denton, A J Schaefer, N D Shah, and S A Smith. The structure of optimal statin initiation policies for patients with type 2 diabetes. *IIE Transactions of Healthcare Systems Engineering*, 1(1):49–65, 2011.

[108] E B Laber, D J Lizotte, M Qian, W E Pelham, and S A Murphy. Dynamic treatment regimes: Technical challenges and applications. *Electronic Journal of Statistics*, 8(1):1225–1272, 2014.

[109] M Laurino and A Landi. A model predictive control strategy toward optimal structured treatment interruptions in anti-hiv therapy. *IEEE Transactions on Biomedical Engineering*, 57(5):1040–1050, 2010.

[110] B F Leeb, I Andel, J Sautner, T Nothnagl, and B Rintelen. The DAS28 in rheumatoid arthritis and fibromyalgia patients. *Rheumatology*, 43(12):1504–1507, 2004.

[111] W Lehmacher and G Wassmer. Adaptive sample size calculations in group sequential trials. *Biometrics*, 55:1286–90, Dec 1999.

[112] M Levi, S Grange, and N Frey. Exposure-exposure relationship of tocilizumab, an anti–il-6 receptor monoclonal antibody, in a large population of patients with rheumatoid arthritis. *The Journal of Clinical Pharmacology*, 53(2):151–159, 2013.

[113] T Lewens. Distinguishing treatment from research: a functional approach. *Journal of Medical Ethics*, 32(7):424–429, 2006.

[114] P E Lipsky, D M van der Heijde, E W St Clair, D E Furst, F C Breedveld, J R Kalden, J S Smolen, M Weisman, P Emery, M Feldmann, G R Harriman, and R N Maini. Anti-tumor necrosis factor trial in rheumatoid arthritis with concomitant therapy study group: Infliximab and methotrexate in the treatment of rheumatoid arthritis. *New England Journal of Medicine*, 343(22):1594–1602, 2000.

[115] L Lojo, G Bonilla, D Peiteado, A Villalba, C Plasencia, L Nuno, A Balsa, and E Martin-Mola. Down-titration and discontinuation of infliximab, adalimumab and etanercept

in established rheumatoid arthritis. *Annals of the Rheumatic Diseases*, 72(Supl 3):237, Jun 2013.

[116] F Lori and J Lisziewicz. Structured treatment interruptions for the management of HIV infection. *The Journal of the American Medical Association*, 286(23):2981–2987, 2001.

[117] E Louis, G Vernier-Massouille, J-C Grimaud, Y Bouhnik, D Laharie, J-L Dupas, H Pillant, L Picon, M Veyrac, M Flamant, G Savoye, R Jian, M De Vos, G Paintaud, E Piver, J-F Colombel, J-Y Mary, and M Lemann. Infliximab discontinuation in crohn's disease patients in stable remission on combined therapy with immunosuppressors: a prospective ongoing cohort study. *Gastroenterology*, 136(5 Suppl 1):A146, May 2009.

[118] W S Lovejoy. Computationally feasible bounds for partially observed markov decision processes. *Operations Research*, 39:162–175, 1991.

[119] W S Lovejoy. A survey of algorithmic methods for partially observed markov decision processes. *Annals of Operations Research*, 18:47–66, 1991.

[120] L Ma, P Ji, Y Wang, L Zhao, Y Xu, S Doddapaneni, and C G Sahajwalla. Exposure-response modeling and simulation of the efficacy endpoints in rheumatoid arthritis. In *American Conference on Pharmacometrics*, 2013.

[121] J Macdougall. Analysis of dose-response studies - Emax model. In N Ting, editor, *Dose Finding in Drug Development*, Statistics for Biology and Health, chapter 9, pages 127–145. Springer, New York, NY, USA, 2006.

[122] F Maggiolo, D Ripamonti, G Gregis, G Quinzan, A Callagaro, and F Suter. Effect of prolonged discontinuation of successful antiretroviral therapy on cd4 t cells: a controlled, prospective study. *AIDS*, 18:439–446, 2004.

[123] L M Maillart, J S Ivy, S Ransom, and K Diehl. Assessing dynamic breast cancer screening policies. *Operations Research*, 56(6):1411–1427, 2008.

[124] J W Mandema, D H Salinger, S W Baumgartner, and M A Gibbs. A dose–response meta-analysis for quantifying relative efficacy of biologics in rheumatoid arthritis. *Clinical Pharmacology & Therapeutics*, 90(6):828–835, 2011.

[125] J W Mandema, D H Salinger, S W Baumgartner, and M A Gibbs. A dose-response meta-analysis for quantifying relative efficacy of biologics in rheumatoid arthritis. *Clin Pharmacol Ther*, 90(6):828–35, Dec 2011.

[126] M Marino, E Zucchi, M Fabbro, I Lodolo, R Maieron, S Vadalà, and M Zilli. Outcome of infliximab discontinuation in ibd patients and therapy rechallenging in relapsers: Single centre preliminary data. *Journal of Crohn's and Colitis*, 8(Suppl 1), Feb 2014.

[127] A Markham and H M Lamb. Infliximab: a review of its use in the management of rheumatoid arthritis. *Drugs*, 59(6):1341–59, Jun 2000.

[128] H Markowitz. *Portfolio selection: efficient diversification of investments*. Wiley, New York, NY, USA, 1959.

[129] R E Marsh, J A Tuszynski, M B Sawyer, and K J E Vos. Emergence of power laws in the pharmacokinetics of paclitaxel due to competing saturable processes. *Journal of Pharmacy and Pharmaceutical Sciences*, 11(3):77–96, 2008.

[130] V Di Martino, C Richou, J-P Cervoni, J M Sanchez-Tapias, D M Jensen, A Mangia, M Buti, F Sheppard, P Ferenci, and T Thévenot. Response-guided peg-interferon plus ribavirin treatment duration in chronic hepatitis c: meta-analyses of randomized, controlled trials and implications for the future. *Hepatology*, 54(3):789–800, Sep 2011.

[131] J E Mason, B T Denton, N D Shah, and S A Smith. Optimizing the simultaneous management of blood pressure and cholesterol for type 2 diabetes patients. *European Journal of Operational Research*, 233(3):727–738, 2014.

[132] P Mease. Improving the routine management of rheumatoid arthritis: The value of tight control. *The Journal of Rheumatology*, 37(8):1570–1578, 2010.

[133] K Michaud, J Messer, H K Choi, and F Wolfe. Direct medical costs and their predictors in patients with rheumatoid arthritis. *Arthritis and Rheumatism*, 48(10):2750–2762, 2003.

[134] S Michelson and T Schofield. *The biostatistics cookbook*. Kluwer Academic, Boston, MA, USA, 1996.

[135] S A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society*, 65(2):331–366, 2003.

[136] S A Murphy, L M Collins, and A J Rush. Customizing treatment to the patient: adaptive treatment strategies. *Drug and Alcohol Dependence*, 88(Supplement 2):S1–S3, 2007.

[137] National Institutes of Health Office of the Director. NIH consensus statement on management of hepatitis c: 2002. *NIH Consens State Sci Statements*, 19(3):1–46, 2002.

[138] D Negoescu, K Bimpikis, M L Brandeau, and D A Iancu. Dynamic learning of patient response types: an application to treating chronic diseases. `http://www.isye.umn.edu/news/pdf/bandits_cronic_disease(1).pdf`, September 2014.

[139] A Nilim and L El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[140] J Nocedal and S J Wright. *Numerical Optimization*. Springer, New York, NY, USA, 2006.

[141] J R O'Dell, T R Mikuls, T H Taylor, V Ahluwalia, M Brophy, S R Warren, R A Lew, A C Cannella, G Kunkel, C S Phibbs, A H Anis, S Leatherman, , and E Keystone. Therapies for active rheumatoid arthritis after methotrexate failure. *New England Journal of Medicine*, 369(4):307–318, 2013.

[142] A Ogata, T Hirano, Y Hishitani, and T Tanaka. Safety and efficacy of tocilizumab for the treatment of rheumatoid arthritis. *Clinical medicine insights. Arthritis and musculoskeletal disorders*, 5:27–42, 2012.

[143] International Programme on Chemical Safery (IPCS). Principles for modeling dose-response for the risk assessment of chemicals. Technical report, World Health Organization, 2009.

[144] J O'Quigley, M Pepe, and L Fisher. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics*, 46(1):33–48, 1997.

[145] D Palmer and Y El Mledany. Treat-to-target: a tailored treatment approach to rheumatoid arthritis. *British Journal of Nursing*, 22(6):308–318, 2013.

[146] C H Papadimitriou and J N Tsitsiklis. The complexity of markov decision processes. *Mathematics of Operations Research*, 12:441–450, 1987.

[147] M A Peterzan, R Hardy, N Chaturvedi, and A D Hughes. Meta-analysis of dose-response relationships for hydrochlorothiazide, chlorthalidone, and bendroflumethiazide on blood pressure, serum potassium, and urate. *Hypertension*, 59(6):1104–1109, 2012.

[148] J Pinheiro. Evaluation and recommendations on adaptive dose-ranging trials: Highlights from the PhRMA adaptive dose-ranging studies working group. *The Journal of Clinical Pharmacology*, 50(S9):47S–49S, 2010.

[149] J Pinheiro, B Bornkamp, E Glimm, and F Bretz. Model-based dose finding under model uncertainty using general parametric models. *Statistics in Medicine*, 33(10):1646–1661, 2014.

[150] PRNewswire. Rheumatoid arthritis (RA): World drug market 2013-2023. `http://www.prnewswire.com/news-releases/`

rheumatoid-arthritis-ra-world-drug-market-2013-2023-210247851.html, June 2013.

[151] M Puterman. *Markov Decision Processes*. John Wiley and Sons, New Jersy, 1994.

[152] W Raasch. Artritis drug market. http://www.wikinvest.com/wiki/Arthritis_Drug_Market, 2009.

[153] M U Rahman, I Strusberg, P Geusens, A Berman, Yocum D, Baker D, C Wagner, J Han, and R Westhoven. Double-blinded infliximab dose escalation in patients with rheumatoid arthritis. *Annals of Rheumatoid Disease*, 66(9):1233–1238, 2007.

[154] H Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

[155] W F Rosenberger and L M Haines. Competing designs for phase i clinical trials. *Statistic in Medicine*, 21(18):2757–2770, 2002.

[156] B Sandikci, L M Maillart, A J Schaefer, O Alagoz, and M S Roberts. Estimating the patient's price of privacy in liver transplantation. *Operations Research*, 56(6):1393–1410, 2008.

[157] L Schacter, M Birkhofer, S Carter, R Canetta, S Hellmann, N Onetto, C Weil, B Winograd, and M Rozencweig. Anticancer drugs. In J O'Grady and P H Joubert, editors, *Handbook of Phase I/Phase II Clinical Trials*, pages 523–534. CRC Press, 1997.

[158] A J Schaefer, M D Bailey, S M Shechter, and M S Roberts. Modeling medical treatment using markov decision processes. In *Operations Research and Health Care*, volume 70 of *International Series in Operations Research & Management Science*, pages 593–612. Springer, Berlin, Germany, 2005.

[159] L G Schipper, W Kievit, A A den Broeder, M A van der Laar, E M M Adang, J Fransen, and P L C M van Riel. Treatment strategies aiming at remission in early rheumatoid

arthritis patients: starting with methotrexate monotherapy is cost-effective. *Rheumatology*, 50(7):1320–1330, 2011.

[160] L G Schipper, L T C van Hulst, R Grol, P L C M van Riel, M E J L Hulscher, and J Fransen. Meta-analysis of tight control strategies in rheumatoid arthritis: protocolized treatment has additional value with respect to the clinical outcome. *Rheumatology*, 49(11):2154–2164, 2010.

[161] M Schoels, R Knevel, and D Aletaha et al. Evidence for treating rheumatoid arthritis to target: results of a systematic literature search. *Annals of Rheumatoid Disease*, 69:638–643, 2010.

[162] A Schumitzky. Application of stochastic control theory to optimal design of dosage regimens. In D Z D'Argenio, editor, *Advanced methods of pharmacokinetic and pharmacodynamic systems analysis*, volume 1, chapter 13, pages 135–152. Plenum Press, 1991.

[163] C A Sesin and C O Bingham. Remission in rheumatoid arthritis: wishful thinking or clinical reality. *Seminars in Arthritis and Rheumatism*, 35(3):185–196, 2005.

[164] S M Shechter, M D Bailey, and A J Schaefer. A modeling framework for replacing medical therapies. *IIE Transactions*, 40(9):861–869, 2008.

[165] S M Shechter, M D Bailey, A J Schaefer, and M S Roberts. The optimal time to initiate hiv therapy under ordered health states. *Operations Research*, 56 (1), 2008.

[166] L Shen, S Peterson, A R Sedaghat, M A McMahon, M Callender, H Zhang, Y Zhou, E Pitt, K S Anderson, E P Acosta, and R F Siliciano. Dose-response curve slope sets class-specific limits on inhibitory potential of anti-hiv drugs. *Nat Med*, 14(7):762–6, Jul 2008.

[167] D M Shepard, M C Ferris, G H Olivera, and T R Mackie. Optimizing the delivery of radiation therapy to cancer patients. *SIAM Review*, 41(4):721–744, 1999.

[168] WJ Shih. Group sequential, sample size re-estimation and two-stage adaptive designs in clinical trials: a comparison. *Statistics in Medicine*, 25:933–41, 2006.

[169] R Simon. The use of genomics in clinical trial design. *Clin Cancer Res*, 14(19):5984–93, Oct 2008.

[170] J A Singh, D E Furst, A Bharat, J R Curtis, A F Kavanaugh, J M Kremer, L W Moreland, J O'Dell, K L Winthrop, T Beukelman, S L Bridges Jr., W W Chatham, H E Paulus, M Suarez-Almazor, C Bombardier, M Dougados, D Khanna, C M King, A L Leong, E L Matteson, J H Schousboe, E Moynihan, K S Kolba, A Jain, E R Volkmann, H Agrawal, S Bae, A S Mudano, N M Patkar, and K G Saag. 2012 update of the 2008 american college of rheumatology recommendations for the use of disease-modifying antirheumatic drugs and biologic agents in the treatment of rheumatoid arthritis. *Arthritis Care & Research*, 64(5):625–639, 2012.

[171] F Sivas, L A Aktekin, F Eser, F G Yurdakul, E Oksuz, K Ozoran, and H Bodur. Comparative results of das28 and quality of life in patients with rheumatoid arthritis and fibromyalgia. *Archives of Rheumatology*, 25(4):179–183, 2010.

[172] W Slob. Dose-response modeling of continuous endpoints. *Toxicological Sciences*, 66:298–312, 2002.

[173] R D Smallwood and E J Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 11:1071–1088, 1973.

[174] J Smolen, A Beaulieu, A Rubbert-Roth, C Ramos-Remus, J Rovensky, E Alecock, T Woodworth, R Alten, and OPTION trial investigators. Effect of interleukin-6 receptor inhibition with tocilizumab in patients with rheumatoid arthritis (option study): a double-blind, placebo-controlled, randomised trial. *The Lancet*, 371(9617):987–997, 2008.

[175] J S Smolen and D Aletaha. Monitoring rheumatoid arthritis. *Current Opinion in Rheumatology*, 23(3):252–258, 2011.

[176] J S Smolen, D Aletaha, J W J Bijlsma, F C Breedveld, D Boumpas G Boumpas, G Burmester, B Combe, M Cutolo, M de Wit, M Dougados, P Emery, A Gibofsky, J J Gomez-Reino, B Haraoui, J Kalden, E C Keystone, T K Kvien, I McInnes, E Martin-Mola, C Montecucco, M Schoels, and D van der Heijde. Treating rheumatoid arthritis to target: recommendations of an international task force. *Annals of Rheumatoid Disease*, 69(4):631–637, 2010.

[177] J S Smolen and P Emery. Infliximab: 12 years of experience. *Arthritis Research & Therapy*, 13(Supplement 1):S2, 2011.

[178] A Sofi, A Ali, S A Khuder, and A Nawras. Meta-analysis- maintenance of remission following discontinuation of infliximab in patients with crohn's disease. *Gastroenterology*, 144(5 Suppl 1):S637, May 2013.

[179] E J Sondik. *The optimal control of partially observable Markov processes*. PhD thesis, Stanford University, 1971.

[180] M Soubrier, X Puechal, J Sibilia, X Mariette, O Meyer, B Combe, R M Flipo, D Mulleman, F Berenbaum, C Zarnitsky, T Schaeverbeke, P Fardellone, and M Dougados. Evaluation of two strategies (initial methotrexate monotherapy vs its combination with adalimumab) in management of early active rheumatoid arthritis: data from the GUEPARD trial. *Rheumatology*, 48(11):1429–1434, 2009.

[181] N L Stokey, R E Lucas, and E C Prescott. *Recursive methods in economic dynamics*. Harvard University Press, Cambridge, Massachusetts, USA, 1989.

[182] A Talal, R M Ribeiro, K A Powers, M Grace, C Cullen, M Hussain, M Markatou, and A S Perelson. Pharmacodynamics of peg-ifn differentiate hiv/hcv coinfected sustained virological responders from nonresponders. *Hepatology*, 43(5):943–953, 2006.

[183] S ten Wolde, FC Breedveld, BAC Dijkmans, J Hermans, JP Vandenbroucke, MAFJ van de Laar, HM Markusse, M Janssen, and HR van den Brink. Randomised placebo-controlled study of stopping second-line drugs in rheumatoid arthritis. *The Lancet*, 347(8998):347–352, Feb 1996.

[184] P F M Teunis and A H Havelaar. The beta-Poisson dose-response model is not a single-hit model. *Risk Analysis*, 20(4):513–520, 2000.

[185] N Ting. *Dose Finding in Drug Development*. Statistics for Biology and Health. Springer, New York, NY, USA, 2006.

[186] D M Topkis. *Supermodularity and Complementarity*. Princeton University Press, Princeton, NJ, USA, 1998.

[187] K S Upchurch and J Kay. Evolution of treatment for rheumatoid arthritis. *Rheumatology*, 51(Supplement 6):28–36, 2012.

[188] A van der Maas, W Kievit, B J F van den Bemt, F H J van den Hoogen, P L C M van Riel, and A A den Broeder. Down-titration and discontinuation of infliximab in rheumatoid arthritis patients with stable low disease activity and stable treatment: an observational cohort study. *Annals of Rheumatoid Disease*, 71(11):1849–1854, 2012.

[189] A M van Gestel, C J Haagsma, and P L C M van Riel. Validation of rheumatoid arthritis improvement criteria that include simplified joint counts. *Arthritis and Rheumatism*, 41(10):1845–1850, 1998.

[190] A M van Gestel, M L L Prevoo, M A van Hof, M H van Rijswijk, L B A van de Putte, and P L C M van Riel. Development and validation of the european league against rheumatism response criteria for rheumatoid arthritis: Comparison with the preliminary american college of rheumatology and the world health organization/international league against rheumatism criteria. *Arthritis and Rheumatism*, 39(1):34–40, 1996.

[191] N van Herwaarden, AA den Broeder, W Jacobs, A van der Maas, JW Bijlsma, RF van Vollenhoven, and BJ van den Bemt. Down-titration and discontinuation strategies of tumor necrosis factor-blocking agents for rheumatoid arthritis patients with low disease activity. *Cochrane Database of Systematic Reviews*, 2014.

[192] R F van Vollenhoven. How to dose infliximab in rheumatoid arthritis: new data on a serious issue. *Annals of Rheumatoid Disease*, 68(8):1237–1239, 2009.

[193] K Visser and D van der Heijde. Optimal dosage and route of administration of methotrexate in rheumatoid arthritis: a systematic review of the literature. *Annals of Rheumatoid Disease*, 68(7):1094–1099, 2009.

[194] A Wailoo, A Brennan, N Bansback, R Nixon, F Wolfe, and K Michaud. Modeling the cost effectiveness of etanercept, adalimumab and anakinra compared to infliximab in the treatment of patients with rheumatoid arthritis in the medicare program. Technology assessment, Agency for Healthcare Research and Quality, 2006.

[195] W C Waterhouse. Do symmetric problems have symmetric solutions? *The American Mathematical Monthly*, 90(6):378–387, 1983.

[196] WebMD. Treating rheumatoid arthritis with DMARDs. `http://www.webmd.com/rheumatoid-arthritis/guide/dmard-rheumatoid-arthritis-treatment`, September 15 2013.

[197] L M Wein, S A Zenios, and M A Nowak. Dynamic multidrug therapies for hiv: A control theoretic approach. *Journal of Theoretical Biology*, 185(1):15 – 29, 1997.

[198] G Wells, J C Becker, J Teng, M Dougados, M Schiff, J Smolen, D Aletaha, and P L C M van Riel. Validation of the 28-joint Disease Activity Score (DAS28) and European League Against Rheumatism response criteria based on C-reactive protein against disease progression in patients with rheumatoid arthritis, and comparison with

the DAS28 based on erythrocyte sedimentation rate. *Annals of Rheumatoid Disease*, 68(6):954–960, 2009.

[199] J Whitehead and H Brunier. Bayesian decision procedures for dose determining experiments. *Statistic in Medicine*, 14(9):885–893, 1995.

[200] S Zeuzem, J-M Pawlotsky, E Lukasiewicz, M von Wagner, I Goulis, Y Lurie, E Gianfranco, J-M Vrolijk, J I Esteban, C Hezode, M Lagging, F Negro, A Soulier, E Verheij-Hart, B Hansen, R Tal, C Ferrari, S W Schalm, and A U Neumannand DITTO-HCV Study Group. International, multicenter, randomized, controlled study comparing dynamically individualized versus standard treatment in patients with chronic hepatitis C. *Journal of Hepatology*, 43(2):250–7, Aug 2005.

[201] J Zhang, B T Denton, H Balasubramanian N D Shah, and B A Inman. Optimization of prostate biopsy referral decisions. *Manufacturing and Service Operations Management*, 14(4):529–547, 2012.

[202] L-X Zhang, WS Chan, SH Cheung, and F Hu. A generalized drop-the-loser urn for clinical trials with delayed responses. *Statistica Sinica*, 17:387–409, 2007.

VITA

Jakob Kotas was born in Chicago, Illinois, and graduated from the Illinois Mathematics and Science Academy in 2005. He earned a Bachelor of Science Magna Cum Laude in Engineering Physics in 2009 and a Master of Engineering in Engineering Mechanics in 2011, both from Cornell University. He earned a Master of Science in Applied Mathematics in 2013 and Doctor of Philosophy in Applied Mathematics in 2016, both from the University of Washington–Seattle.

He can be reached at `jkotas@uw.edu`.

This manuscript was typed by the author.