OPTIMAL LEARNING VIA DYNAMIC RISK

BY CURTIS MCGINITY

A dissertation submitted to the
Graduate School—New Brunswick
Rutgers, The State University of New Jersey
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
Graduate Program in Operations Research

Written under the direction of
Andrzej Ruszczyński
and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2017

ABSTRACT OF THE DISSERTATION

Optimal Learning via Dynamic Risk

by Curtis McGinity

Dissertation Director: Andrzej Ruszczyński

We consider the dilemma of taking sequential action within a nebulous and costly stochastic system. In such problems, the decision–maker sequentially takes an action from a given set, then incurs a cost and observes a response depending stochastically on the action. Confronted with an unknown system, the decision–maker must learn about the system by experimenting with risky actions, thus enabling better decisions over time.

We thus consider the risk–averse optimal learning problem to dynamically choose actions to minimize the risk of the cumulative costs of learning. Motivated by problems in clinical trial design for novel pharmaceutical agents, we formulate the problem of Bayesian statistical inference under binary response as a Markov decision process with belief states. We formulate a certain class of standardized logistic models with quantile parameterizations and offer some general conditions under which belief states satisfy stochastic order and log–concavity under Bayesian dynamics. We also establish some stronger results under assumptions on the policy class.

We then introduce dynamic Markov risk measures, formulate dynamic programming equations, and discuss the challenges of their solution. We then offer an approximate DP (ADP) schema based on a coarse grid approximation within a parameterized distribution family utilizing log–concavity constraints. We also study risk–averse lookahead

policies, introducing a robust–response policy and a heuristic policy.

We compare the performance of the above policy classes to the state–of–the–art, and demonstrate its performance in computational experiments, including the design of dose–escalation policies for three chemotherapeutic agents (bleomycin, etoposide, 5–fluorouracil). The robust–response policy exhibits strong performance in the problem class, clarifying the role of risk measures under Bayesian belief dynamics and suggesting avenues of future research.

Acknowledgements

I would thank my thesis advisor Andrzej Ruszczyński for introducing me to the rich subject of dynamic risk and the potent simplicity of dynamic programming. Andrzej is a truly infinite source of patience, having now been verified by direct methods, whose masterful analyses have forever impressed upon me the power of high school mathematics, and whose laughter ever fills the soul.

I am also indebted to Endre Boros, Adi Ben-Israel, Darinka Dentcheva, and Michael Katehakis for their service on my thesis committee. I am grateful for their valuable comments, suggestions, and guidance over the years. I would express special thanks to Darinka Dentcheva for several constructive conversations, many revealing suggestions, and yet more charitable kindnesses. Her maternal spirit, prescient counsel, and incisive wit were indispensable in the course of this research.

I would also express sincere thanks to Fred Roberts for being a gracious mentor for many years. As the philosopher's wisdom, his sage advice, ever humbly delivered, has been an invaluable beacon of clean perspective untinged by trend. He remains a steadfast model of wearing many hats in academics—and wearing them well.

I would also express my gratitude to Larry Einhorn, a physician of the first order, foremost among that redeeming remnant true to the finest traditions of the profession. Undoubtedly laudable, his work has engendered yet thousands of remissions, but of note are his unremitting endeavors on the frontier of the treatment vs. experimentation dilemma that continue to produce protocol par excellence. Indeed, his work has imbued me with bone–deep inspiration for this research.

The research in this dissertation was supported by a RUTCOR Excellence Fellowship, National Science Foundation Award DMS–1312016, and an Instructorship in Management Science and Information Systems at Rutgers Business School, New Brunswick.

I would also acknowledge support through a Graduate Assistantship from the Command Control and Interoperability Center for Advanced Data Analysis (CCICADA) during my doctoral candidacy.

Dedication

To my parents. I love you both and owe you everything.

Table of Contents

Chapter 1

Introduction and Preliminaries

> *It is worth noting that the notation facilitates discovery. This, in a most wonderful way, reduces the mind's labor.*

Gottfried Wilhelm Leibniz, *1646–1716*

> *The impediment to action advances action. What stands in the way becomes the way.*

Marcus Aurelius, *121–180*

1.1 Introduction and Overview

1.1.1 Background and Motivations

Self–similarity is a fundamental concept observed in nature. The notion has been studied in a myriad of disciplines and depicted perhaps most famously in the uniquely exquisite imagery of fractals. Broadly speaking, its essence unites much of what is understood about such fundamental concepts as complexity, randomness, recursion, pattern, symmetry, coherence, meaning, beauty, harmony, emergence, equilibrium, and consciousness.

Learning, at essence, constitutes a recursion whereby one compares one's understanding (or belief) about how a system would have worked with how one observes the system did work, and in view of the differences, updates one's understanding with a new, similar understanding in more harmonious accord with the observed reality.

Table 1.1: Antithetical aphorisms intimate the depths of wisdom across all of life.

Seize the day.	*Patience is a virtue.*
He who hesitates is lost.	*Look before you leap.*
Don't sweat the small stuff.	*The devil is in the details.*
Actions speak louder than words.	*The pen is mightier than the sword.*
Idle hands are the devil's workshop.	*All work and no play makes Jack a dull boy.*
Where there's smoke, there's fire.	*You can't judge a book by its cover.*
Nothing ventured, nothing gained.	*Better safe than sorry.*
On the one hand...	*On the other hand...*

What is most fascinating about this framing of the intuitive phenomenon of learning is revealed when one poses the question: How to learn *best*? The more one ponders this, the ostensibly pedestrian notion of "best" becomes yet more subtle. What does best mean when I know that I do not yet know? As foreign or abstract as this may seem, we all negotiate this problem every day of our lives. Indeed, so it is that we find ourselves confronted by a mute and often perilous nature, desperately trying to ascertain what is going on, yet all the while needing to quest ever onward.

In Table 1.1 we have collected a sample of some collective wisdom expressed throughout the ages. Having survived the test of time, each colloquialism expresses a fundamental truth experienced in a myriad of disparate circumstances, and yet their contradictory sentiments cannot be denied. Rather than undermining their validity, however, their antithetical nature reveals a depth intrinsic to the complexity of life. Indeed, in any given scenario all of these are *simultaneously* valid, casting us into and out of various courses of action in a cognitive tempest. Choosing well, amid such waves of insight and squalls of the spurious, is verily the art of the wise. In this dissertation, we term this problem the *optimal learning problem*, and endeavor to make scientific progress on negotiating it. In this sense, we take up the quest for the mathematical and algorithmic abstraction of wisdom.

We begin by motivating our considerations with a panoply of examples. The general

problem is extremely broad, with potential applications in most corners of reality. In the next section we shall selfishly focus our motivations on *health optimization*, broadly defined, and later focus our investigations so as to make strides both theoretical and practical.

1.1.2 Survey of Applications

We focus on those of personal interest, particularly in medicine, nutrition, sport, and health optimization generally. See [106] for an extensive survey of practical applications.

Medicine and pharmaceuticals

Example 1 (**Distributed Clinical Trials**)**.** Clinical trials (CTs) are an essential means by which we control the quality of treatment protocols and drugs reaching human application and constitute a multi–billion USD industry—in operational overhead alone. In a broader sense, one would be hard pressed to argue that CTs operate on an efficient risk–reward frontier.

This is, however, rightfully so in the face of important ethical standards and the apparent requirements of statistical inference. Indeed, the state–of–the–art double–blind, randomized, controlled trials (DBRCTs) seek statistical power and unbiasedness. Yet this comes both at a direct cost of structuring DBRCTs, and at an increasingly alarming cost of applicability or scope. The latter has perhaps manifest in part with recent concerns on reproducibility in the medical literature. This particular problem is egregious in the dietary supplement industry, where compounds are consumed widely with little or no effect, ostensibly in refute of published studies. Perhaps the quotidian lifestyles of many functionally attenuate some significant fraction of published results valid in the context of a DBRCT. This would be entirely of a piece with observations of simple, low-dimensional behavior in complex systems. The emerging notion of personalized medicine is also pertinent in this context, although current methodologies have yet to lend themselves to this end. However, what is much less widely identified are the *indirect* costs of structuring DBRCTs, which is to say, the opportunity costs.

The optimal learning models below engender a vision of medicine wherein *treatment is synonymous with trial*. That is, CTs as currently formalized become a rarity and not distinguished from clinical treatment in general. Granted, we lose significant confidence in the empirical data, which becomes fraught by parts with uncertainty and confounding factors (to a greater degree than at present), but we gain a tremendous amount of data

with which to develop more sophisticated perspectives. Indeed, not unrealistic gains in the quantity of data could be as much as 3-5 orders of magnitude. Thus, we make a trade for robust and comprehensive knowledge laden with uncertainty over fragile and narrow knowledge with a high degree of confidence, as it were.

However, as a final point that we discuss in detail below, our exploration of the unknown possibilities is intrinsically characterized by the actions we take. It is for this reason that widely distributed, adaptive treatment–trials constitute a paradigm shift with exponential potential.

Example 2 (**Pharmacodynamic Tolerance (Intermittent Dosing)**)**.** The empirical phenomenon of drug *tolerance*, whereby physiological response to a given drug monotonically depends on its recent dosage history, is enough to conjecture the general suboptimality of the following class of dosage policies: Take x [quantity] every t [period of time], for constants $x, t \in \mathbb{R}_{\geq 0}$. This class of static dosage policies is nearly ubiquitous among prescription and over–the–counter pharmaceuticals, as well as dietary supplements, thus constituting a potentially massive collective loss.

The adaptive optimal learning models introduced below implicitly incorporate tolerance via empirical data. Moreover, insofar as tolerance may be cast in terms of coupled oscillators, tolerance models may be inferred in real time from the optimal dosage policies rather than explicitly (i.e., separately) studied by way of, e.g., formal clinical trials. This has recently emerged as of particular applicability, e.g., among celebrated inhibitors of PI3K, mTOR, etc., pathways; cf. [119, 113, 112, 138, 137].

Example 3 (**Hormone Replacement Therapy**)**.** It is well known that human hormone levels vary cyclically throughout the day, as well as various other timescales (e.g. weekly, monthly, etc.). These hormonal oscillations are often commonly referred to as Circadian rhythms. Perhaps surprisingly, however, virtually all hormone replacement therapy protocols neglect these phenomena in their dosage policies. Concerns about the risks of testosterone replacement therapy recently prompted updates to federal policy and a flurry of research, albeit without consideration of more sophisticated dosage recommendations; cf. [70, 96, 92, 49, 80].

Health and nutrition

Example 4 (**Dynamic Diet Problem & the Health Model Problem**)**.** Consider the problem of determining what to eat in order to not only survive, but to *thrive*, and to do so economically. The *diet problem* of determining what foods and/or supplements to eat in order to maximize a combination healthiness (by way of nutrition) and deliciousness, all while subject to budgetary constraints, is a concern familiar to most, in uncertain terms if not explicitly. Unfortunately, the link between health and nutrition is perhaps the most controversial and inflammatory health–related topic in public discourse, as evidenced by the "Diet & Nutrition" aisle at one's local bookstore, or worse still the "health segment" on one's local morning news. This is in no small part due to the widespread dogmatism plaguing nutrition science, engendering guru after guru and, ultimately, as a corrective reaction, independent non–profit organizations such as, e.g., the Nutrition Science Initiative (NuSI). Yet with each touted diet at odds with the next, and nutrition research delivering dubious, contradictory research, many have concluded that a healthy diet is largely a matter of personal preference, if not merely myth.

What is too often overlooked, perhaps as it lies in plain view, is that, as a matter of fact, *no functional model of health exists.* This is perhaps jarring at first brush, but observe that the only method currently in use is the disease model; that is, "unhealthy" is defined as "sufficiently correlated with disease", and "healthy" is implicitly defined as "not unhealthy".[1] However, identifying that which is unhealthy is not constructive in determining what is healthy. As an example, pointing out that consuming large amounts of sugar is unhealthy, because it is correlated with disease (e.g., Type II diabetes), says virtually nothing of what quantity of sugar one should consume to be healthy. At this point, one is tempted to invoke the aphorism "all things in moderation". Indeed, this is the very line taken by the Office of Disease Prevention and Health Promotion (ODPHP) in the only prescriptive healthy eating guideline of the 2015–2020 Dietary Guidelines

[1] Fully substantiating this claim is beyond the scope of this dissertation, but it is perhaps telling that `health.gov` is the URL of the Office of Disease Prevention and Health Promotion (ODPHP). Moreover, about the 2015-2020 Dietary Guidelines, ODPHP states that "its recommendations are ultimately intended to help individuals improve and maintain overall health...its focus is disease prevention."

for Americans [48]. However, this aphorism is often mistaken to mean that one should do all things, and do them in moderation. Rather, this pithy wisdom states that all things one should do should be done in moderation. What things one should be doing in the first place is thus revealed as the tacit premise, and on this point the wisdom is necessarily mute! This casts into sharp relief the widely held truism that a balanced diet is healthy. Sure, a certain balanced diet *is* healthy—but which one?

The void of a functional health model, once identified, is thus a keyhole through which to observe that nutrition science has suffered frequent encounter with Hume's guillotine,[2] which is to say the science has attempted to draw prescriptive conclusions from limited descriptive data.[3] This observation is pivotal: Lacking a workable definition of health and a viable model of functional nutrition *in principle*, it may come as no surprise that as a population we have defaulted to the natural dogmatic, anecdote–fueled paradigm, which we refer to as the guru model. Thus, any proposed definition of health stands to upset substantial portions of the relevant industries, and moreover cannot find any unassailable premise from which to gain purchase and dispense new ideas. The stakes are thus set for perfect inertia, and one cannot expect transformative progress in this area by the traditional model of (nutrition) science.

Perhaps surprisingly, however, general dynamic diets have not been considered in the literature, and only a few special cases have recently been proposed. It is not inconceivable that realizing sought–after health benefits would necessitate undulations between and among efficient points within several different, even antithetical dieting paradigms. A *dynamic diet*, which is to say a diet adaptively changing based on one's current physical state in order to improve the most informed, comprehensive measures of one's overall health as needs change, would naturally move into and out of otherwise disparate dieting paradigms. Presumably, prominent dieting paradigms would frequently arise within the methodology as real–time optimal solutions to improve health—or perhaps not, in which case one might rethink said prominence. Clearly, this process involves significant

[2] Hume's guillotine refers to what is also known as Hume's is–ought problem. Articulated first by the philosopher David Hume, the problem states that one cannot simply derive an ought from an is, i.e., it is not obvious how to derive normative from positive statements.

[3] Note that this logical fallacy is precisely that which the scientific endeavor aims to remedy.

uncertainties, e.g. the realized nutritional content of foods, one's unique, fluctuating nutritional needs, one's taste preferences, and the very definition of health, to name just a few.

From this vantage, we would therefore propose the alternative of a data–driven[4] paradigm of health and nutrition, to emerge organically from a distributed, robust, adaptive optimal learning methodology. Crucially, such a methodology allows for the simultaneous, systematic, scientific comparison of all proposed definitions of health by putting said definitions on equal footing. In particular, the methods entirely obviate any requirement for expert (at best, guru at worst) intervention. However, any expertise may of course be incorporated into the methodology, which is self–correcting and would serve only to substantiate truly expert proposals. The robust optimal learning framework proposed below is a technologically viable methodology for delivering real–time optimal solutions to the dynamic diet problem as a matter of course en route to learning a functional picture of health and nutrition from data.

Example 5 (**Intermittent Fasting**)**.** Apparent themes have emerged in research on intermittent fasting, i.e., oscillations between nutrient scarcity and surfeit. This is of a piece with several other branches of research in the life sciences, perhaps indicating that undulation between competing biochemical (sub–)processes is intrinsically healthy. Determining how long to fast, how much to consume when refeeding, etc., is a problem that may be effectively modeled squarely within the paradigm introduced below; cf. [72, 95, 33].

Training and sport

Example 6 (**Adaptive Protocol Design & Dynamic Workouts**)**.** Fitness regimens often suffer from a plateau of improvement, owing largely to their fundamentally static nature. More advanced strength and conditioning protocols, usually for serious and/or professional athletes, have long utilized some form of quasiperiodicity, colloquially known as cycling, that depends on the adaptation response of the athlete. However,

---

[4]Or, evidence–based, as coined by the NIH, ODPHP, etc.

similar to the scenario in Example 4, owing largely to the inability of exercise science to describe the subtle and sophisticated aspects of empirically successful training protocols, engineering an optimal training regimen has evolved in large part to the guru paradigm. In what is now our recurring theme, exercise science proceeds almost entirely by evaluating a set of protocols, which are provided externally, stemming from their prominence in practice, and thus driven by the guru model.

As before, we would propose an agnostic methodological framework for engendering an optimal training protocol. Just as in the case of health and nutrition, such a methodology would provide a scientifically optimal protocol which is, rather than constrained by expert preconceptions, instead fueled by their expertise. Indeed, expert ideas may be incorporated directly into the optimal learning methodology, which would self–correct and identify the best (combination of) protocols scientifically. Moreover, the optimal training protocol would be optimal for each athlete (i.e., personalized), and would be adaptive in the sense of dynamically constructing training, mobility, rehab, and recovery based on the athlete's performance. Importantly, the robust methodology introduced below is suited to the task of implementing such a distributed training protocol while maximally avoiding the potential issue of sacrificing some athletes short–term performance, or worse injuring an athlete, while experimenting with sub–optimal training protocols.

Example 7 (**Dynamic HRV Training**)**.** As an interesting example of the above, heart rate variability (HRV) training has been the foremost dynamic training protocol to emerge. This owes largely to the widespread availability of heart rate monitors, which has furthered research in the area. The concept behind using HRV as a guide to training is that HRV is a measure of the readiness of the nervous system, which in turn is related to an athlete's physical stress level, and ultimately, potential for adaptation and injury.

In HRV training, an athlete undulates between high-intensity and lower-intensity periods of work, as measured by predefined target HR. We propose dynamic HRV training, whereby the target HR levels would adapt in real time to the athlete's current performance and HRV, optimally leading the athlete along an adaptive, efficient

frontier. Moreover, the inherent stochastic element of the optimal learning algorithm could also offer an interesting training element, which may be of independent interest.

1.1.3 Organization of Dissertation

In Sections 1.2 to 1.4 we provide an overview of the preliminary and background material on which this dissertation builds.

Chapter 2 investigates the methodology of Bayesian inference in a general sense. We begin with an overview of logistic models, and in Section 2.1.2 we collect fundamental facts on logistic models relevant for future investigations. In Section 2.2 we introduce notions of stochastic order, and derive several results related to the structure of Bayesian belief dynamics for logistic models. Section 2.3 continues in this vein, deriving some conditions on the first integral arising in the dynamic setting. In Section 2.4 we turn to consider the generalized concavity of logistic models. We conclude with results on the log–concavity of Bayesian belief orbits in the case of quantile parameterizations.

Chapter 3 turns to the dynamic optimization problem, formulating the risk–aware optimal learning problem. In Section 3.1 we formulate active Bayesian sequential inference as a Markov decision process (MDP) with belief states. We then introduce risk, formulate the risk–aware MDP, and introduce a class of composite risk measures we call *robust response*. Finally, we formulate risk–neutral and risk–aware dynamic programming equations. In Section 3.2.4 we discuss the challenges associated with dynamic programming in the optimal learning setting.

In Section 3.3 we introduce the class of lookahead policies and study several prominent policy classes. Section 3.4 presents a discussion of some preliminary considerations material in future methods for augmenting Bayesian inference in this model class.

In Chapter 4 we introduce an approximate dynamic programming schema, namely approximation within the log-logistic distribution family. In Section 4.6 we present the associated ADP equations and discuss the difficulties surrounding their use.

In Chapter 5 we conduct a series of computational experiments to garner a sense of the effects of risk-aware optimal learning in practice. In Section 5.1 we present the results of a simulation study comparing three lookahead policies. In particular,

we demonstrate the robustness of risk-aware policies to spurious data in the optimal learning setting. Section 5.2 constitutes a corpus of case studies demonstrating the performance of the policies introduced above. In particular, we consider the design of dose-escalation policies in Phase I clinical trials for three chemotherapeutic agents: 5-fluorouracil (Section 5.2.1), bleomycin (Section 5.2.2), and etoposide (Section 5.2.3).

Finally, in Chapter 6 we elucidate several conclusions from our investigations on the role of dynamic risk in optimal learning. We conclude with a brief discussion of potential future directions.

1.2 Markov Decision Processes

Markov decision processes (MDPs) offer a general-purpose infrastructure for the modeling of sequential decision-making under uncertainty. Their straightforward premise, simple structure, and broad applicability have led both to their thorough study and practical application. Specifically, MDPs readily admit satisfactory modeling of stochastic elements and nonlinear dynamics, remarkably with a unified solution technique: dynamic programming. This is undoubtedly their principal strength and the source of their widespread application in practice.

In accord with the *no free lunch* principle, the amazing generality of optimization via dynamic programming (DP) is infamously foiled by the *curse of dimensionality*. That is, the generality of DP techniques comes part in parcel with a certain neglect of the particular problem structure, and therefore the computational complexity grows exponentially with the size of the problem. In this respect, proper DP solutions are largely precluded for all but the most simple problem instances, either of very small dimension or of very specific structure.

This fact has led to the development of approximate dynamic programming (ADP) solutions techniques, of which a myriad have been introduced in the literature. Broadly speaking, ADP techniques attempt to transform the problem via some approximation schema, so rendering the modified problem tractable by DP techniques. We discuss ADP in more detail in Chapter 4, but the standard reference would be [23].

Puterman has given a comprehensive treatment of discrete-time MDPs [109], and Bertsekas has also treated the topic authoritatively [24, 23, 25]. A generalization of the MDP framework is known as the adaptive MDP (AMDP), or in some disciplines as a hidden Markov model (HMM), which has been treated by Hernandez-Lerma [73]. Yet more generally, partially-observable MDPs (POMDPs) allow for uncertainty in the underlying state. The two models share some overlap, but the general POMDP model allows for observations to be uncertain as well (which translates to further uncertainty in the controlled transition kernel).

In the next Section, we formalize introduce the model by rigorously defining the controlled Markov process. Moreover, with a view to our purpose of studying optimal learning, for brevity we will develop the formal model sufficient also for an adaptive controlled Markov process. As we shall see, form the modeling perspective, this is simply a collection of classical models, although we defer the treatment of the optimization problem until Chapter 3. That is, the existence of the parameter space below is merely formal, and may be ignored at this moment.

1.2.1 Controlled Markov Model

The formalism of MDPs could perhaps be daunting for the uninitiated, although the essence of the methodology is rather intuitive. As such, we would take the opportunity to state the model plainly and give a sense of its character before introducing the mathematical elements. Put simply, an MDP describes the dynamic over time between a system and a decision-maker (DM). The system consists of a set of possible states, and at each point in time, the DM selects from a set of possible actions, and then the system moves stochastically from its current state to its next state. Given the current state, the transition to the next state occurs according to a given probability distribution that could depend on the current state, the action taken by the (DM), or both. In tandem to the system transitions, at each point in time, the DM collects a reward (or pays a cost) that depends on the current state and the selected action. The goal, then, is to maximize (minimize) the cumulative reward (cost) over the problem horizon, which may be finite or infinite.

As one can readily see, the MDP framework is exceedingly natural, and rather general, insofar as it imposes almost no restrictions on the kinds of system states, DM actions, the rewards (costs), or the nature of the transitions. We now introduce the formal elements of this powerful framework, in both the finite- and infinite-horizon cases, although we will focus on only the finite-horizon problem below. Our presentation will most closely follow that of [73].

The adaptive MDP framework consists of a tuple

$$\big(\mathcal{S}, \mathcal{U}, \Theta, Q_t(\cdot|s,u), c_t(s,u,\theta)\big),$$

where $\mathcal{S}$ is the system state space, $\mathcal{U}$ is the control (or action) space, Θ is the unknown-parameter space, Q_t is the controlled transition kernel at time t, and c_t is the one-step cost function of the controlled process, as described above. We now introduce several formal definitions:

Definition 1. *Adaptive Controlled Markov Model*

1. *The* state space $\mathcal{S}$ *is a Borel space (i.e. Polish space), a Borel subset of a complete separable metric space.*

2. *The* control space $\mathcal{U}$ *is also a Borel space. Introducing the measurable multifunction $U : \mathcal{S} \rightrightarrows \mathcal{U}$, for each state $s \in \mathcal{S}$, we denote the* admissible controls *by the non–empty, measurable set $U(s) \subseteq \mathcal{U}$. It is convenient to denote the graph of U by $U_g := \{(s,u) : s \in \mathcal{S}, u \in U(s)\}$.*

3. *The* parameter space Θ *is a Borel space. We assume there exists a unique "true" value denoted $\theta^* \in \Theta$, however this value is unknown.*

4. *The* controlled transition kernel *is a measurable mapping $Q : \mathcal{U}_g \times \Theta \to \mathcal{P}(\mathcal{S})$, where for $q \in Q$ and $S \in \mathcal{B}(\mathcal{S})$, $q\big(S \mid (s,u) \in U_g, \theta \in \Theta\big) = Pr\{s' \in S \mid (s,u) \in \mathcal{S} \times U(s), \theta \in \Theta\}$. Moreover, $\int_S v(s',\theta)\, q(ds' \mid s,u,\theta)$ is a continuous function of $u \in U(s)$ for all $s \in \mathcal{S}$, $\theta \in \Theta$, and all $v \in \mathcal{B}(\mathcal{S} \times \Theta)$.*

5. *The* one-step loss function *is a measurable function $c : U_g \times \Theta \to \mathbb{R}$ such that $|c(s_t,u_t,\theta)| \leq X < \infty$, for all $(s_t,u_t,\theta) \in U_g \times \Theta$, and $c(s,u,\theta)$ is a continuous function of $u \in U(s)$ for all $s \in \mathcal{S}$, $\theta \in \Theta$, and all $v \in \mathcal{B}(\mathcal{S} \times \Theta)$.*

1.2.2 Admissible Policies

For each $t = 1, 2, \cdots$, we denote the space of (admissible) state and control histories up to time t by $\mathcal{H}_t := U_g^t \times \mathcal{S}$, so that an element $h_t \in \mathcal{H}_t$ is given by the sequence $h_t = (s_1, u_1, s_2, u_2, \cdots, s_{t-1}, u_{t-1}, s_t)$. We can now introduce the set of (admissible) policies.

Definition 2. *History-dependent Policies*

1. *A* randomized policy *is a sequence* $\boldsymbol{\pi} = \{\pi_t, t = 1, 2, \cdots\}$*, of measurable functions* $\pi_t : \mathcal{H}_t \to \mathcal{P}(\mathcal{U})$*, such that each stochastic kernel* $\pi_t(\cdot \mid h_t)$ *is supported on the set of admissible controls; formally, we require* $\pi_t(U(s_t) \mid h_t) = 1$*, for all* $h_t \in \mathcal{H}_t$ *and* $t = 1, 2, \cdots$*. We denote the set of all history-dependent, randomized policies by* $\boldsymbol{\Pi}^{HR}$*.*

2. *A* deterministic policy *is a sequence* $\boldsymbol{f} = \{f_t, t = 1, 2, \cdots\}$*, of measurable functions* $f_t : \mathcal{H}_t \to \mathcal{U}$*, such that* $f_t(h_t) \in U(s_t)$*, for all* $h_t \in \mathcal{H}_t$ *and* $t = 1, 2, \cdots$*. Deterministic policies are clearly included in randomized policies, where for each* $h_t \in \mathcal{H}_t$ *and* $B \in \mathcal{B}(\mathcal{U})$*,* $\pi_t(B \,|\, h_t) = \mathbf{1}_B\big(f_t(h_t)\big)$*. We denote the set of all history-dependent, deterministic policies by* $\boldsymbol{\Pi}^{HD}$*.*

Intuitively, a Markov policy is simply a (randomized) policy depending only on the current state. The formal definitions are as follows.

Definition 3. *Markov Policies*

1. *We call the measurable function* $\pi_t : \mathcal{S} \to \mathcal{P}(\mathcal{U})$ *a Markov* decision rule *(or selector) at time* t *if* $\pi_t(s) \in \mathcal{P}(U(s))$ *for all* $s \in \mathcal{S}$*, and denote the set of all such decision rules by* Π_t^M*.*

2. *A (randomized)* Markov policy *is a sequence* $\boldsymbol{\pi} = \{\pi_t, t = 1, 2, \cdots\}$ *of measurable functions* $\pi_t \in \Pi_t$*. We call a Markov policy* deterministic *if for all* $s \in \mathcal{S}$ *and* $t = 1, 2, \cdots$*, the support of the measure* $\pi_t(\cdot \mid s)$ *is the singleton* $\{f_t(s)\} \subset U(s)$*. We denote the set of all Markov randomized and deterministic policies by* $\boldsymbol{\Pi}^{MR}$ *and* $\boldsymbol{\Pi}^{MD}$*, respectively.*

3. *We call a Markov policy* $\{\pi_t\}$ stationary *if there exists* $\pi \in \Pi$ *such that* $\pi_t(s) = \pi(s)$, *for all* $t = 1, 2, \cdots$, *and all* $s \in \mathcal{S}$.

Note that these policy classes satisfy $\mathbf{\Pi}^{MD} \subset \mathbf{\Pi}^{MR} \subset \mathbf{\Pi}^{HR}$, and $\mathbf{\Pi}^{MD} \subset \mathbf{\Pi}^{HD} \subset \mathbf{\Pi}^{HR}$.

1.2.3 Induced Markov Process

We now introduce the adaptive controlled Markov process (ACMP). We will not require the full generality of the model presented. For practical purposes, it is often sufficient to consider finite $\mathcal{S}$ and $\mathcal{U}$, in which case the below presentation proceeds analogously with some simplifications. See [109] for a presentation of the finite case, or [26] for a more rigorous presentation of the following.

Let Ω be the product space defined in a finite horizon model as

$$\Omega := \mathcal{S} \times \mathcal{U} \times \mathcal{S} \times \mathcal{U} \cdots \times \mathcal{S} = \big(\mathcal{S} \times \mathcal{U}\big)^{T-1} \times \mathcal{S},$$

and in an infinite horizon model as

$$\Omega := \mathcal{S} \times \mathcal{U} \times \mathcal{S} \times \mathcal{U} \cdots = \big(\mathcal{S} \times \mathcal{U}\big)^{\infty},$$

and denote the corresponding product σ-algebra by $\mathcal{F}$. When necessary, we will denote the corresponding σ-subalgebras by $\mathcal{F}_t$. Elements of Ω are (infinite) sequences of the form

$$\omega = (s_1, u_1, s_2, u_2, \cdots), \qquad s_t \in \mathcal{S}, u_t \in \mathcal{U} \text{ for all } t = 1, 2, \cdots,$$

and s_t, u_t are given by measurable coordinate mappings from Ω to $\mathcal{S}$ and $\mathcal{U}$, respectively.

The Ionescu-Tulcea theorem states that, for any given Markov policy $\{\pi_t, t = 1, 2, \cdots\} = \boldsymbol{\pi} \in \mathbf{\Pi}$, initial state $s_1 = s \in \mathcal{S}$, and $\theta \in \Theta$, there exists a unique probability measure $P_s^{\boldsymbol{\pi},\theta}$ on $(\Omega, \mathcal{F})$ given by

$$P_s^{\boldsymbol{\pi},\theta}(d\omega) = p_s(ds_1)\pi_1(du_1 \mid s_1) \prod_{t=2}^{\infty} q(ds_t \mid s_{t-1}, u_{t-1}, \theta)\pi_t(du_t \mid s_{t-1}, u_{t-1}, s_t), \quad (1.1)$$

and satisfying

$$P_s^{\boldsymbol{\pi},\theta}(\mathcal{H}_\infty) = 1, \tag{1.2}$$

$$P_s^{\boldsymbol{\pi},\theta}(s_1 = s) = 1, \tag{1.3}$$

$$P_s^{\boldsymbol{\pi},\theta}(u_t \in B \mid h_t) = \pi_t(B \mid h_t), \text{ for all } B \in \mathcal{B}(\mathcal{U}), h_t \in \mathcal{H}_t, t = 1, 2, \cdots, \tag{1.4}$$

$$P_s^{\boldsymbol{\pi},\theta}(s_t \in C \mid h_t, u_t) = q(C \mid s_t, u_t, \theta), \text{ for all } C \in \mathcal{B}(\mathcal{S}), h_t \in \mathcal{H}_t, t = 1, 2, \cdots. \tag{1.5}$$

In the case of Markov policies $\boldsymbol{\pi} \in \boldsymbol{\Pi}^{MR}$ and hence (1.4) becomes

$$P_s^{\boldsymbol{\pi},\theta}(u_t \in B \mid h_t) = \pi_t(B \mid s_t), \text{ for all } B \in \mathcal{B}(\mathcal{U}), h_t \in \mathcal{H}_t, t = 1, 2, \cdots. \tag{1.6}$$

Therefore, for each $\theta \in \Theta$, the induced stochastic process $\big(\Omega, \mathcal{F}, P_s^{\boldsymbol{\pi},\theta}, \{s_t\}\big)$ is a stationary controlled Markov process.

1.2.4 Performance Criteria

Suppose we have a fixed initial state $s_1 = s \in \mathcal{S}$ and $\theta \in \Theta$. Each policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}^{MR}$ generates the *Markov loss process* written as $\big(\Omega, \mathcal{F}, P_s^{\boldsymbol{\pi},\theta}, \{(s_t, c(s_t, u_t, \theta))\}\big)$. We will denote by $\mathcal{Z}_t$ the space of $\mathcal{F}_t$-measurable random variables on Ω. We now turn turn to the issue of evaluating the random loss (or cost) sequence given by

$$\{Z_t^\theta = c(s_t, u_t, \theta) \ : \ Z_t^\theta \in \mathcal{Z}_t, \ t = 1, 2, \cdots\}. \tag{1.7}$$

For all integrable functions $\boldsymbol{Z} : \Omega \to \mathbb{R}$, the expectation operator $\mathbb{E}_s^{\boldsymbol{\pi},\theta}$ with respect to the probability measure $P_s^{\boldsymbol{\pi},\theta}$ is given by

$$\mathbb{E}_s^{\boldsymbol{\pi},\theta}\big[\boldsymbol{Z}\big] = \int_\Omega \boldsymbol{Z}(\omega)\, P_s^{\boldsymbol{\pi},\theta}(d\omega). \tag{1.8}$$

1.2.5 Optimization Problems

According to the classical theory, one considers the *expected total discounted reward* performance criteria, given by

$$V(\boldsymbol{\pi}, s, \theta) := \mathbb{E}_s^{\boldsymbol{\pi},\theta}\Big[\sum_{t=1}^{\infty} \gamma_t\, c(s_t, u_t, \theta)\Big],$$

where γ_t are *discount factors*, scalars between zero and one. In the finite horizon case, for some time horizon $T \in \mathbb{N}$, we have the analogous performance criteria:

$$V(\boldsymbol{\pi}, s, \theta) := \mathbb{E}_s^{\boldsymbol{\pi},\theta}\Big[\sum_{t=1}^{T} \gamma_t \, c(s_t, u_t, \theta)\Big]. \tag{1.9}$$

In either case, the *optimal value function* v in this setting is therefore defined as

$$v(s, \theta) := \inf_{\boldsymbol{\pi} \in \boldsymbol{\Pi}^{MR}} V(\boldsymbol{\pi}, s, \theta). \tag{1.10}$$

As one can readily see, in this definition v depends on the unknown parameter $\theta \in \Theta$. Assuming the true parameter value $\theta^* \in \Theta$, the optimal *learning problem* is therefore to find an *optimal learning policy* $\boldsymbol{\pi}_\theta^*$. However, stated in this way the task may appear pedestrian, but as we will see in Chapter 3 significant considerations will need to be addressed.

Now, we continue to review the classical theory, focusing on the finite-horizon MDP. We will thus drop the dependence on the unknown parameter θ, to resume this line again below. We owe our presentation to lecture notes from a course on dynamic programming with Ruszczyński; any errors or lapses in rigor are surely our own.

Policy valuation

Suppose we are given a policy $\boldsymbol{\pi} = \{\pi_1, \pi_2, \cdots, \pi_{T-1}\}$, and we allow $\boldsymbol{\pi} \in \boldsymbol{\Pi}^{HR}$. Define the *value functions*

$$v_t^{\boldsymbol{\pi}}(h_\tau) \triangleq \mathbb{E}\left[\sum_{\tau=t}^{T-1} c_\tau(s_\tau, u_\tau) + c_{_T}(s_{_T}) \;\Big|\; h_\tau\right], \;\; t = 1, \cdots, T.$$

By the tower property of iterated conditional expectations,

$$\begin{aligned}
v_t^{\boldsymbol{\pi}}(h_\tau) &= \mathbb{E}\left[\mathbb{E}\left\{\sum_{\tau=t}^{T-1} c_\tau(s_\tau, u_\tau) + c_{_T}(s_{_T}) \;\Big|\; h_{t+1}\right\} \;\Big|\; h_t\right] \\
&= \mathbb{E}\left[\mathbb{E}\left\{c_t(s_t, u_t) + \sum_{\tau=t+1}^{T-1} c_\tau(s_\tau, u_\tau) + c_{_T}(s_{_T}) \;\Big|\; h_{t+1}\right\} \;\Big|\; h_t\right] \\
&= \mathbb{E}\left[c_t(s_t, u_t) + \mathbb{E}\left\{\sum_{\tau=t+1}^{T-1} c_\tau(s_\tau, u_\tau) + c_{_T}(s_{_T}) \;\Big|\; h_{t+1}\right\} \;\Big|\; h_t\right] \\
&= \mathbb{E}\left[c_t(s_t, u_t) + v_{t+1}^{\boldsymbol{\pi}}(h_{t+1}) \;\Big|\; h_t\right],
\end{aligned}$$

for all $t = 1, 2, \cdots, T-1$. Note that our policy $\boldsymbol{\pi}$ was arbitrary, and thus this derivation allows for the evaluation of any policy. Moreover, the value $v_1^{\boldsymbol{\pi}}$ coincides with the performance criterion (1.9).

Value functions

Given the policy $\boldsymbol{\pi}$, denote the partial policy by $\boldsymbol{\pi}_t = (\pi_t, \pi_{t+1}, \cdots, \pi_{T-1})$. Define now the *optimal value functions*

$$v_t^*(h_t) \triangleq \inf_{\boldsymbol{\pi}_t} \; v_t^{\boldsymbol{\pi}}(h_t), \quad t = 1, 2, \cdots, T. \tag{1.11}$$

Substituting in the recursion above, we obtain

$$\begin{aligned} v_t^*(h_t) &= \inf_{\boldsymbol{\pi}_t} \; \mathbb{E}\left[c_t(s_t, u_t) + v_{t+1}^{\boldsymbol{\pi}}(h_{t+1}) \;\middle|\; h_t \right] \\ &= \inf_{\pi_t} \; \mathbb{E}\left[c_t(s_t, u_t) + \inf_{\boldsymbol{\pi}_{t+1}} \; v_{t+1}^{\boldsymbol{\pi}}(h_{t+1}) \;\middle|\; h_t \right]. \end{aligned}$$

Just as before, we conclude the follow recursive system of equations is satisfied

$$v_t^*(h_t) = \inf_{\pi_t} \; \mathbb{E}\left[c_t(s_t, u_t) + v_{t+1}^*(h_{t+1}) \;\middle|\; h_t \right], \quad t = 1, 2, \cdots, T-1, \tag{1.12}$$

together with the final stage $v_{_T}^*(h_{_T}) = c_{_T}(s_{_T})$.

Deterministic Markov policies

Throughout the above, we have allowed the policy to depend on the entire history and to be randomized, i.e., $\boldsymbol{\pi} \in \boldsymbol{\Pi}^{HR}$. We now show that deterministic policies suffice. Rewriting (1.12) explicitly in in terms of the policy and again applying the tower property of conditional expectation, we observe that

$$\begin{aligned} v_t^*(h_t) &= \inf_{\pi_t} \; \mathbb{E}\left[\mathbb{E}\Big\{ c_t(s_t, u_t) + v_{t+1}^*(h_{t+1}) \;\Big|\; h_t, u_t \Big\} \;\middle|\; h_t \right] \\ &= \inf_{\pi_t} \; \mathbb{E}\left[c_t(s_t, u_t) + \mathbb{E}\Big\{ v_{t+1}^*(h_{t+1}) \;\Big|\; h_t, u_t \Big\} \;\middle|\; h_t \right] \\ &= \inf_{\pi_t} \; \int_{U_t(s_t)} c_t(s_t, u_t) + \mathbb{E}\Big\{ v_{t+1}^*(h_{t+1}) \;\Big|\; h_t, u \Big\} \, \pi_t(du|h_t) \\ &= \inf_{u \in U_t(s_t)} \; c_t(s_t, u_t) + \mathbb{E}\Big\{ v_{t+1}^*(h_t, u, s_{t+1}) \;\Big|\; h_t, u \Big\}, \end{aligned}$$

where the final equality stems from the fact that, under h_t, the integrand in the penultimate equality is a function of u only. Thus, insofar as π_t is a probability measure over the control set, it is optimal to concentrate mass on the value(s) of u yielding the smallest integrand, which is to say that deterministic policies are no worse than randomized policies.

Dynamic programming

Finally, we now show that only the current state, out of the entire history, is material in the computation of the optimal value functions. We proceed by backward induction, an argument attributed to Bellman [20].

Suppose for $t+1 \leq T$, we have $v^*_{t+1}(h_{t+1}) = v^*_{t+1}(s_{t+1})$. Proceeding from the last displayed equality, we therefore have

$$v^*_t(h_t) = \inf_{u \in U_t(s_t)} c_t(s_t, u_t) + \mathbb{E}\Big\{ v^*_{t+1}(s_{t+1}) \;\Big|\; h_t, u \Big\}.$$

For a fixed h_t and u, the conditional expectation coincides with the integral with respect to the distribution of the next state. That is,

$$\begin{aligned} \mathbb{E}\Big\{ v^*_{t+1}(s_{t+1}) \;\Big|\; h_t, u \Big\} &= \int_{\mathcal{S}} v^*_{t+1}(y) Q_t(dy | s_t, u) \\ &= \mathbb{E}\Big\{ v^*_{t+1}(s_{t+1}) \;\Big|\; s_t, u \Big\}. \end{aligned}$$

This follows essentially from the fact that the controlled transition kernel Q_t is Markov. We therefore conclude that

$$v^*_t(h_t) = \inf_{u \in U_t(s_t)} c_t(s_t, u) + \mathbb{E}\Big\{ v^*_{t+1}(s_{t+1}) \;\Big|\; s_t, u \Big\}.$$

But the RHS is a function of s_t only, and thus $v^*_t(h_t) = v^*_t(s_t)$, completing the inductive step. Given the fact that $v^*_{_T}(h_{_T}) = c_{_T}(s_{_T})$, we obtain by (backward) induction that the following *dynamic programming equations* must hold for the optimal value functions: For every state $s \in \mathcal{S}$,

$$v^*_{_T}(s) = c_{_T}(s), \tag{1.13}$$

$$v^*_t(s) = \inf_{u \in U_t(s)} c_t(s, u) + \mathbb{E}\Big\{ v^*_{t+1}(s_{t+1}) \;\Big|\; s_t = s, u_t = u \Big\},\ t = T-1, \cdots, 2, 1. \tag{1.14}$$

The DP equations may be solved backward in time for $v_1^*(\cdot)$, the solution of which coincides with the solution of the expected cost problem (1.9).

1.3 Dynamic Risk

Consider the probability space $(\Omega, \mathcal{F}, \mathcal{P})$, where Ω is a separable metric space, $\mathcal{F} = \mathcal{B}(\Omega)$ is the canonical Borel σ-algebra induced on Ω, and $\mathcal{P}$ is a probability measure on $\mathcal{F}$. Let the p-normed space $\mathcal{Z} = L_p(\Omega, \mathcal{F}, \mathcal{P})$, $1 \leq p < \infty$, denote the space of random variables with bounded moments with respect to P. We interpret each $Z \in \mathcal{Z}$ as a cost random variable, so that greater values are understood as unfavorable. In this context, we call any functional $\rho : \mathcal{Z} \to \mathbb{R}$ a *risk measure*.

1.3.1 Coherent and Convex Risk Measures

In order to standardize the measurement of risk, an axiomatic approach has been taken in the literature. In particular, the class of *coherent risk measures* introduced in [11] has been widely accepted and studied extensively in the finance, economics, and among others, operations research literature. Broadly speaking, coherent risk measures are characterized by four intuitive axioms that any sensible risk measure should satisfy:

Definition 4 (Coherent risk measure)**.** *A coherent risk measure is a functional* $\rho : \mathcal{Z} \to \mathbb{R}$*, satisfying for all* $Z, \tilde{Z} \in \mathcal{Z}$*,*

$$\text{Monotonicity:} \qquad \text{if } Z \geq \tilde{Z}, \text{ then } \rho(Z) \geq \rho(\tilde{Z}); \tag{1.15}$$

$$\text{Subadditivity:} \qquad \rho(Z + \tilde{Z}) \leq \rho(Z) + \rho(\tilde{Z}); \tag{1.16}$$

$$\text{Translation Equivariance:} \qquad \rho(a + Z) = a + \rho(Z); \tag{1.17}$$

$$\text{Positive Homogeneity:} \qquad \rho(\alpha Z) = \alpha \rho(Z), \ \forall\, \alpha > 0. \tag{1.18}$$

In some cases it may be desirable to consider a relaxation to the class of *convex risk measures*. It can be shown that (1.16) and (1.18) together imply convexity of coherent risk measures. However, convex risk measures generally do not satisfy positive homogeneity (1.18). Thus, coherent risk measures are often *equivalently* defined by

replacing (1.16) with the condition

$$\text{Convexity:} \qquad \rho(\alpha Z + (1-\alpha)\tilde{Z}) \leq \alpha\rho(Z) + (1-\alpha)\rho(\tilde{Z}), \tag{1.19}$$

for all $\alpha \in [0,1]$. The class of convex risk measures is defined by replacing both (1.16) and (1.18) by (1.19). In what follows, we will often refer to coherent and/or convex risk measures simply as "risk measures," although no confusion should arise. Additionally, we call a risk measure *law invariant* if $\rho(Z)$ depends only on the distribution of Z. For further details, see, e.g., [129, 125] and the references therein.

More than a few risk measures have been proposed in the literature. We present several prevailing examples now and discuss some of their properties below.

Example 8 (VaR_α)**.** *The value at risk at level $\alpha \in (0,1)$ is denoted VaR_α and given by*

$$\text{VaR}_\alpha(Z) = F_Z^{-1}(1-\alpha), \tag{1.20}$$

where F_Z^{-1} is the quantile function of Z. Note that VaR_α is not generally a coherent risk measure, as it violates (1.16)*. If the distribution function F_Z is log-concave, however, then VaR_α becomes coherent. Thus, e.g., VaR_α is coherent for exponential families of distributions.*

Example 9 (AVaR_α)**.** *The average value at risk at level $\alpha \in (0,1)$ is denoted AVaR_α and given by*

$$\text{AVaR}_\alpha(Z) = \frac{1}{\alpha}\int_0^\alpha F_Z^{-1}(1-\xi)\, d\xi, \tag{1.21}$$

where F_Z^{-1} is the quantile function of Z. This coherent risk measure is also often called the conditional value at risk ($CVaR_\alpha$), as is emphasized by writing (1.21) *in the alternative form*

$$\text{AVaR}_\alpha(Z) = \mathbb{E}[Z \mid Z \geq \text{VaR}_\alpha(Z)]. \tag{1.22}$$

Example 10 (MSD_κ)**.** *The mean upper-semideviation of order $q \in [1, p]$ at level $\kappa \in [0, 1]$ is denoted* MSD_κ *and given by*

$$\mathrm{MSD}_\kappa(Z) = \mathbb{E}[Z] + \kappa \mathbb{E}\left[(Z - \mathbb{E}[Z])_+^q\right]^{1/q}, \tag{1.23}$$

where $(\,\cdot\,)_+ \equiv \max\{\cdot, 0\}$ is the positive part function. One can readily verify that this is a coherent risk measure. Below we will always consider MSD_κ *of order $q = 1$, in which case the risk measure is notably (piecewise) linear.*

Example 11 (Entropic risk measure)**.** *The entropic risk measure at level $\theta > 0$ is denoted ρ^θ and given by*

$$\rho^\theta(Z) = \frac{1}{\theta} \log \mathbb{E}\left[e^{\theta Z}\right]. \tag{1.24}$$

The entropic risk measure is the prototypical example of a convex risk measure that is not generally coherent.

Example 12 ($\mathrm{EVaR}_{f,\beta}$)**.** *The f-entropic value at risk at level $\beta \geq 0$ is denoted* $\mathrm{EVaR}_{f,\beta}$ *and given by*

$$\mathrm{EVaR}_{f,\beta}(Z) = \max_{Q \in \mathcal{A}} \mathbb{E}_Q[Z], \tag{1.25}$$

where $\mathcal{A} = \left\{ Q \in \mathcal{P}(\Omega, \mathcal{F}) \;\middle|\; D_f(Q, P) \leq \beta \right\}$, and $D_f(\cdot, P)$ is the Csiszàr f-divergence defined in (1.38)*. This coherent risk measure was introduced in [2, 3] for the Kullback-Leibler divergence ($f(x) \equiv x \log x$). It shares connections with several important quantities germane to our theme and will be discussed further below.*

1.3.2 Representations of Risk Measures

We present the following consequence of Fenchel duality as a theorem without proof; the classical reference would be [116], and for further details, we refer the reader to Theorem 2.2 in [125].

Theorem 1 (Dual representation)**.** *Any coherent risk measure $\rho(Z)$ on the space $\mathcal{Z} = L_p(\Omega, \mathcal{F}, \mathcal{P})$, $1 \leq p < \infty$, may be written as*

$$\rho(Z) = \max_{\mu \in \mathcal{A}(\rho)} \mathbb{E}_\mu[Z], \tag{1.26}$$

where $\mathcal{A} \subset \mathcal{P}(\Omega, \mathcal{F})$ is a closed, convex set of probability measures given by $\mathcal{A}(\rho) = \partial\rho(0)$, the subdifferential *of the risk measure evaluated at zero.*

Theorem 2 (Kusuoka representation [88])**.** *Any law invariant, coherent risk measure $\rho(Z)$ on the space $\mathcal{Z} = L_p(\Omega, \mathcal{F}, \mathcal{P})$, $1 \leq p < \infty$, may be written as*

$$\rho(Z) = \max_{\mu \in \mathcal{M}} \int_0^1 \mathrm{AVaR}_\alpha(Z)\, d\mu(\alpha), \tag{1.27}$$

where $\mathcal{M}$ is a set of probability measures on the interval $(0, 1]$.

1.3.3 Dynamic Risk Measures

Let $T \in \mathbb{N}$ be fixed, $\mathcal{F}^T$ the canonical product σ-algebra with natural filtration $\{\mathcal{F}_t\}_{t=1}^T$, and $\mathcal{Z}_t$ the space of bounded $\mathcal{F}_t$-measurable random variables for all $t = 1, 2, \cdots, T$. Denote the natural product space by $\mathcal{Z}_{t,T} = \mathcal{Z}_t \times \cdots \times \mathcal{Z}_T$, for all $t = 1, 2, \cdots, T$.

Definition 5 (Conditional risk measure)**.** *By a* conditional risk measure *we understand a functional $\rho_{t,T} : \mathcal{Z}_{t,T} \to \mathbb{R}$ satisfying the* monotonicity property

$$\rho_{t,T}(Z_{t,T}) \leq \rho_{t,T}(\tilde{Z}_{t,T}), \tag{1.28}$$

for all $Z_{t,T}, \tilde{Z}_{t,T} \in \mathcal{Z}_{t,T}$ such that $Z_s \leq \tilde{Z}_s$, for all $s = t, t+1, \cdots, T$.

Definition 6 (Dynamic risk measure)**.** *By a* dynamic risk measure *we understand a collection $\rho = \{\rho_{t,T}\}_{t=1}^T$ of conditional risk measures.*

Notably, Ruszczyński has shown in [122] that from the notion of *time consistency* of dynamic risk measures one may construct an analogue to the tower property of conditional expectation. When such a dynamic risk measure $\rho = \{\rho_{t,T}\}_{t=1}^T$ may be represented in terms of the Markov transition kernel via transition risk mappings $\{\sigma_{t,T}\}_{t=1}^T$, we call it a *dynamic Markov risk measure*. We now develop this construction by introducing the requisite formal definitions:

Definition 7 (Time consistency)**.** *We call a dynamic risk measure* $\{\rho_{t,T}\}_{t=1}^{T}$ time-consistent*, if for all* $1 \leq i \leq k \leq T$ *and all cost sequences* Z, $\tilde{Z} \in \mathcal{Z}_{i,T}$*, the conditions*

$$\begin{cases} Z_j = \tilde{Z}_j, \qquad j = i, \cdots, k-1, \\ \rho_{k,T}(Z_k, \cdots, Z_{_T}) \leq \rho_{k,T}(\tilde{Z}_k, \cdots, \tilde{Z}_{_T}), \end{cases}$$

imply that

$$\rho_{i,T}(Z_i, \cdots, Z_{_T}) \leq \rho_{i,T}(\tilde{Z}_i, \cdots, \tilde{Z}_{_T}).$$

Stated plainly, if at some time in the future we evaluate Z as less risky than $\tilde{Z}$, yet between now and then the two are identical, then we should not now evaluate Z as any riskier than $\tilde{Z}$.

As mentioned above, Ruszczyński has shown in [122] that the *translation property*

$$\rho_{t,T}(Z_t, Z_{t+1}, \cdots, Z_{_T}) = Z_t + \rho_{t,T}(0, Z_{t+1}, \cdots, Z_{_T}),$$

and the *normalization property*,

$$\rho_{t,T}(0, \cdots, 0) = 0,$$

then it admits of a recursive structure reminiscent of the tower property of conditional expectation. Namely, the cumulative risk thus admits the form

$$\rho_{t,T}(Z_t, Z_{t+1}, \cdots, Z_{_T}) = Z_t + \rho_t\Big(Z_{t+1} + \rho_{t+1}\big(Z_{t+2} + \cdots + \rho_{T-1}(Z_{_T})\cdots\big)\Big), \quad (1.29)$$

Thus, a given time–consistent dynamic risk measure is completely characterized by a certain sequence of one–step conditional risk measures. Clearly, the form of (1.29) is tremendously more tractable than the generic expression with which we opened, yet still it may be further refined in the context of a controlled Markov process.

1.3.4 Risk–aware Optimization

Equipped with an understanding of dynamic risk preferences, we now consider the risk–averse control of MDPs. Specifically, consider a Markov loss process $\{(s_t, c(u_t; s_t))\}$,

with associated transition kernel Q_t. In particular, the finite–horizon problem has the goal of minimizing the cumulative risk:

$$v(\boldsymbol{\pi}; s_1) \triangleq \rho\left(\sum_{t=1}^{T-1} c(u_t; s_t) + c_{_T}(u_{_T})\right),$$

where we understand ρ is a time–consistent dynamic risk measure. Therefore, we know that the cumulative risk may be written as

$$v(\boldsymbol{\pi}; s_1) = c(u_1; s_1) + \rho_1\Bigg(c(u_2; s_2) + \rho_2\Big(c(u_3; s_3) + \cdots \\ \cdots + \rho_{T-1}\big(\, c(u_{_T}; s_{_T})\big) + \rho_T\big(\, c(u_{_{T+1}}; s_{_{T+1}})\big)\cdots\Big)\Bigg). \tag{1.30}$$

Moreover, if each of the conditional risk measures ρ_t, $t = 1, 2, \cdots, T-1, T$, in (1.29) admits the form

$$\rho_t(v(s_{t+1})) = \sigma_t(s_t, Q_t(\pi_t(s_t), v(s_t))), \tag{1.31}$$

then we call conditional risk mappings ρ_t a *Markov risk measure*, and $\rho = \{\rho_t\}_{t=1}^{T}$ a *dynamic Markov risk measure*. The mapping σ_t, $t = 1, 2\cdots, T-1, T$, is called a *transition risk mapping* associated with the controlled transition kernel Q_t, and it satisfies certain regularity conditions material in a general setting. (See [122] for a rigorous and general treatment.)

The relation in (1.31) crucially recasts the dynamic risk preferences in terms of the state space and the Markov transition kernel. This enables formulation of the risk–averse dynamic programming equations, of the familiar form:

$$v_{_{T+1}}(s) = \inf_{u\in\mathcal{U}} c(u; s), \tag{1.32}$$

$$v_t(s) = \inf_{u\in\mathcal{U}} \Big\{c(u; s) + \sigma_t\big(s,\, Q_t(\cdot|s, u),\, v_{t+1}(\cdot)\big)\Big\}, \quad t = 1, 2, \cdots, T. \tag{1.33}$$

1.4 Elements of Statistical Inference

Our goal will be to formulate the optimal learning problem in the framework of Markov decision processes, and thus we defer a discussion of the rich statistics literature until Chapter 3. Here we merely introduce some basic notation and collect fundamental results, and refer the reader to the burgeoning literature for a rigorous and complete treatment.

1.4.1 Statistical Manifolds

Let S be a *statistical manifold*, that is, a smooth, n–dimensional manifold where every point $p \in S$ is a probability distribution over an underlying space $\mathcal{X}$. We will be content to assume $\mathcal{X}$ is a finite set or finite-dimensional $\mathcal{X} \subseteq \mathbb{R}^n$. Let ξ be a coordinate system for S; that is, we have a coordinate mapping ϕ_ξ so that for any $p \in S$, $\xi = \phi_\xi(p) \in \Xi \subset \mathbb{R}^n$. The space Ξ is induced by ϕ_ξ according to $\Xi = \phi_\xi(S) = \{\phi_\xi(p) \mid p \in S\}$. We shall often express coordinates as a vector with components ξ^i, written as $\xi = [\xi^i]$, $i = 1, 2, \cdots, n$.

We take standard assumptions on the continuity and differentiability of coordinate mappings. In particular, we assume that for any two distinct coordinate mappings ϕ_ξ, ϕ_ζ, the coordinate transformation $\xi \mapsto \zeta$ given by $\zeta = \phi_\zeta \circ \phi_\xi^{-1}(\xi)$ is well-defined.

Let $p_\xi := p(x;\xi)$ denote a probability distribution on $\mathcal{X}$ corresponding to $\phi_\xi^{-1}(\xi) \in S$. Thus, viewing ξ as a parameter, we can express S as a family of distributions written as $S = \{p_\xi = p(x;\xi) \mid \xi \in \Xi \subset \mathbb{R}^n\}$. In this context, we call S a *statistical model*. We formalize this in the following definition.

Definition 8 (Statistical model)**.** *By a* statistical model *S, given by*

$$S = \{p_\xi = p(x;\xi) \mid \xi \in \Xi \subset \mathbb{R}^n\}, \tag{1.34}$$

we understand a statistical manifold and a C^∞ coordinate mapping ϕ_ξ such that $\phi_\xi(\Xi) = S$.

Example 13 (Arbitrary finite density)**.** *Let $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$, and ξ be a vector of probabilities. Then $S = \left\{p_\xi : \ p(x_i;\xi) = \xi^i, \ \xi \in \Xi\right\}$, where*

$$\Xi = \left\{\xi : \sum_{i=1}^{n} \xi^i = 1, \ \xi^i \geq 0, \ \forall \ i = 1, 2, \cdots, n\right\}.$$

Example 14 (Poisson Distribution)**.** *Let $\mathcal{X} = \{0, 1, 2, \cdots\}$, $n = 1$, and $\xi \in \mathbb{R}_{>0}$. Then*

$$S = \left\{p(x;\xi) = \frac{\xi^x}{x!} e^{-\xi}\right\}.$$

Example 15 (Normal Distribution)**.** *Let* $\mathcal{X} = \mathbb{R}$*,* $n = 2$*, and* $\xi = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_{>0}$*.*

$$S = \left\{ p(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \right\}.$$

Example 16 (Log-normal Distribution)**.** *Let* $\mathcal{X} = \mathbb{R}_{>0}$*,* $n = 2$*, and* $\xi = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_{>0}$*.*

$$S = \left\{ p(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log x - \mu)^2}{2\sigma^2} \right] \right\}.$$

Example 17 (Logistic Distribution)**.** *Let* $\mathcal{X} = \mathbb{R}$*,* $n = 2$*, and* $\xi = (\mu, \gamma) \in \mathbb{R} \times \mathbb{R}_{>0}$*.*

$$S = \left\{ p(x; \mu, \sigma) = \frac{\exp\left[-\frac{x-\mu}{\gamma} \right]}{\gamma \left(1 + \exp\left[-\frac{x-\mu}{\gamma} \right] \right)^2} \right\}.$$

Example 18 (Log-logistic Distribution)**.** *Let* $\mathcal{X} = \mathbb{R}_{>0}$*,* $n = 2$*, and* $\xi = (\hat{\mu}, \gamma) \in \mathbb{R} \times \mathbb{R}_{>0}$*.*

$$S = \left\{ p(x; \hat{\mu}, \gamma) = \frac{\gamma \hat{\mu}^{-\gamma} x^{\gamma-1}}{(1 + (x/\hat{\mu})^{\gamma})^2} \right\}.$$

1.4.2 Divergence

Let $(\mathbb{R}^n, \Omega, \mu)$ be a measure space, $L_w(\mu)$ a normed space, where $1 \leq w \leq \infty$ and μ is a Borel measure. Let p, q be two probability distributions absolutely continuous with respect to μ on Ω. Finally, let a statistical model S together with a coordinate system ξ be given.

Definition 9 (General divergence functions)**.** *By a* divergence *on a statistical model* S*, we understand a smooth function* $D : S \times S \to \mathbb{R}$ *satisfying*

$$D(p, q) \geq 0, \quad \forall\ (p, q) \in S \times S, \tag{1.35}$$

$$D(p, q) = 0, \quad \textit{iff}\ \ p = s, \tag{1.36}$$

and where the matrix given by

$$g_{ij}^{(D)} = -D(\partial_i, \partial_j) \tag{1.37}$$

is strictly positive definite on $\mathcal{T}(S) \times \mathcal{T}(S)$*.*

For any divergence D, one can associate a unique Riemannian metric $g^{(D)} = \langle \cdot, \cdot \rangle^{(D)}$ given by (1.37). We summarize the fundamental results in the following theorem, given without proof.

Theorem 3. *Any divergence D induces a torsion-free dualistic structure $(g, \nabla^{(D)}, \nabla^{(D^*)})$, and conversely, any given dualistic structure (g, ∇, ∇^*) is induced by some divergence D.*

Several important divergence classes have been introduced in the literature. In particular, the class of Csiszàr f-divergence and Bregman divergence have proven relevant for practical problems of interest. Both of these classes induce a dualistic structure with respect to the Fisher metric. Most notably, the class of α-divergence has been shown in [5] to be precisely the intersection of both classes on the space of positive measures. We first introduce these important examples, and then focus on the α-divergence class below.

Definition 10 (Csiszàr f-divergence)**.** *Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a convex function vanishing at unity.*

$$D_f(p, q) = \int_\Omega f\left(\frac{p(x)}{q(x)}\right) d\mu(x). \tag{1.38}$$

Another fundamental divergence class is that of the Bregman divergence [29]. The Bregman divergence is exceedingly natural, being derived from any smooth, convex function and a Legendre transformation. Perhaps unsurprisingly, any Bregman divergence introduces a dually flat structure.

Definition 11 (Bregman divergence)**.** *Let $\phi : L_w(\mu) \to \mathbb{R}$ be a smooth, convex functional. The Bregman divergence D_ϕ is given by*

$$D_\phi(p, q) = \phi(p) - \phi(q) + \big(\theta^i(q) - \theta^i(p)\big)\partial_i \phi(q), \tag{1.39}$$

where θ denotes the affine coordinate system with respect to the dual potential of ϕ.

We now introduce the important class of α-divergences.

Definition 12 (α-divergence)**.** *For $\alpha \in \mathbb{R}$, the α-divergence $D^{(\alpha)}$ is the Csiszàr $f^{(\alpha)}$-divergence $D_{f^{(\alpha)}} =: D^{(\alpha)}$, where*

$$f^{(\alpha)}(z) = \begin{cases} z \log z, & \text{if } \alpha = 1, \\ -\log z, & \text{if } \alpha = -1, \\ \dfrac{4}{1-\alpha^2}\left(1 - z^{(1+\alpha)/2}\right), & \text{otherwise.} \end{cases} \tag{1.40}$$

The class of α-divergences is in fact special, in that it uniquely belongs to both the Csiszàr f-divergence and Bregman divergence classes over the space of positive measures. Moreover, the α-divergences are *precisely* the canonical divergences corresponding to a dually flat geometrical structure on the space of positive measures. In particular, it can be shown that the Kullback-Leibler divergence is the unique member of all classes over the space of probability distributions [5], perhaps clarifying its persistence in the literature.

Example 19 (Hellinger distance)**.** *The $\alpha = 0$ divergence $D^{(0)}$ given by*

$$D^{(0)}(p,q) = 2\int_{\mathcal{X}} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx \tag{1.41}$$

is called the Hellinger distance*. Note that $D^{(0)}$ satisfies the axioms of distance, uniquely among all α-divergences.*

Example 20 (Kullback-Leibler divergence)**.** *The $\alpha = \pm 1$ divergence $D^{(\pm 1)}$ given by*

$$D^{(1)}(p,q) = D^{(-1)}(p,q) = \int_{\mathcal{X}} p(x) \log\left[\frac{p(x)}{q(x)}\right] dx \tag{1.42}$$

is called the Kullback-Leibler (KL) *divergence. This widely studied divergence is uniquely important in several ways. Namely, it uniquely satisfies a certain chain rule and additivity properties.*

We now collect the most fundamental properties of the divergence classes introduced above, with proof omitted.

Theorem 4. $D_f(p,q) \geq f\left(\int_{\mathcal{X}} p(x)\frac{q(x)}{p(x)}\right) \equiv 1$.

Theorem 5. *D_f is invariant under the affine transformation $f(z) \mapsto f(z) + c(z-1)$, $c \in \mathbb{R}$.*

Theorem 6 (Monotonicity)**.** *Let $K(y|x)$ be an arbitrary transition kernel on $y \in \mathcal{Y}$ for all $x \in \mathcal{X}$, and let $p_K(y) = \int K(y|x)p(x)dx$. Then*

$$D_f(p, q) \geq D_f(p_K, q_K). \tag{1.43}$$

Corollary 7. *D_f is invariant with respect to sufficient statistics $y = F(x)$.*

Corollary 8. *D_f satisfies joint convexity:*

$$D_f\big(\lambda p_1 + (1-\lambda)p_2, \lambda q_1 + (1-\lambda)q_2\big) \leq \lambda D_f(p_1, q_1) + (1-\lambda)D_f(p_2, q_2).$$

Chapter 2
Bayesian Learning Paradigm

> *What is not surrounded by uncertainty cannot be the truth.*
>
> Richard Feynman, *1976*

> *Intelligence consists of this: that we recognize the similarity of different things and the difference between similar ones.*
>
> Montesquieu, *1689–1755*

2.1 Overview of Logistic Models

Statistical inference is a central data analysis technique. However, before any techniques may be applied, a model of the problem at hand must be accepted, and on this point there is much debate in the literature. Arguably one of the most widely used models for statistical inference are *logistic models* [148, 19, 16, 90]. These include, for example, logistic regression and conditional random fields [89], among others. We would like to discuss the issues underlying the broad popularity of such models, particularly in classification and prediction problem settings, based on a survey of the literature.

First, we note that the popularity of logistic models as led to their relatively extensive study (see, e.g., [79]), which naturally begets their popularity. Logistic models do have some uniquely beneficial properties, as it happens. For one, the logistic function maps the extended real line into the unit interval, which allows for a probabilistic interpretation, where desired. For another, the fundamental assumption of logistic models

yields a certain computational efficiency through the employ of kernel methods, which is to say necessary quantities can be expressed as inner products between (canonical) parameters and data. Finally, logistic models also enjoy straightforward extensions of most theoretical results to more general (e.g., higher-dimensional) modeling architectures, owing largely to the linearity intimated above.

The history of the logistic function also partially explains its wide use. It played a pivotal role in the statistical literature on classification, as we will see below. It is the solution of a certain differential equation that arises naturally in applications, and this was one argument for its use. Within the domain of statistical inference and machine learning, in particular within the subfield of neural networks, employ of the logistic function met with improvements in learning performance. This was often attributed to the logistic distribution possessing "heavier tails" relative to the Normal distribution, affording a certain robustness in the task of learning. Even with the recently deified deep belief neural networks[1] and the more recent, and remarkably more interesting, dense nets[2] persist in utilizing logistic functions as activation potentials [74, 77]. Logistic models are thus the industry standard.

There is, in fact, a deeper reason why we observe empirical robustness in many practical cases with logistic models: The family of logistic models will be shown below to be precisely the family of mixtures of any exponential family of distributions. As many commonly observed distribution families are exponential families (e.g., Normal, Exponential, Gamma, Beta, Chi-squared, Poisson, etc.), logistic models find frequent applicability, and their representation as mixtures provides a richer family than any particular exponential family.

The richness of logistic models over exponential families naturally comes at a cost. In particular, certain theoretical results for exponential models do not hold for logistic models, and establishing others requires more sophisticated analysis. To glimpse this most fundamentally at a glance, note that logistic manifolds do not admit dually flat

[1]The term *deep belief* neural network simply indicates the usual neural network with more layers (deep) and Bayesian posterior densities (belief) between layers.

[2]Dense nets are the usual layered neural networks, but where each layer links also to the data of *every* preceding layer.

connections. Additionally, for practical problems wherein relevant or measurable control parameters do not coincide with canonical parameters, the computational demands increase significantly. To see this, one need look no further than to observe that the natural quantile parameterization has the form $1/(1+\exp[\pm 1/x])$, which admits no antiderivative in terms of elementary functions *or* known special functions. Thus, in such practical applications approximation is inherently more challenging for logistic relative to exponential models.

2.1.1 Origin of Logistic Models

For simplicity in the below discussion, but without loss of generality, consider the simple binary classification problem in which a given vector x (e.g., a vector of metrics, a feature vector of measurements, etc.) is labeled by observation of a random variable $r \in \mathcal{R} = \{0, 1\}$. A natural question to as is, given x what is the probability that $r = 1$? Bayes' Theorem tells us the probability we seek may be written as

$$\begin{aligned} P(1|x) &= \frac{p(x|1)P(1)}{P(x)} \\ &= \frac{p(x|1)P(1)}{p(x|1)P(1) + p(x|0)P(0)} \\ &= \frac{1}{1 + e^{-\log \frac{p(x|1)}{p(x|0)} - \log \frac{P(1)}{P(0)}}}, \end{aligned} \tag{2.1}$$

where the final equality is obtained by elementary machinations.

First, we note that (2.1) has the form of a logistic function, and the complementary probability $P(0|x)$ admits an analogous representation. Second, we emphasize that nothing special has been obtained by this manipulation, unless there were some simplifying or useful form of the marginal ratios. Indeed, logistic models, including logistic regression and conditional random fields, make the fundamental assumption that the exponent in (2.1) is affine in x. This is a defining characteristic of logistic models, and also the source of a wealth of theoretical results and a certain computational clemency.

It is important to ascertain the nature of this *logistic assumption* that the log-odds ratio of posterior distributions be affine. What is its significance? It is clear that (2.1) formally holds for any conditional density $p(x|r)$, and thus the question

becomes: Under what conditions on $p(x|r)$ do the assumption hold? In particular, one may inquire about the form of (2.1) for the important cases where $p(x|r)$ is Normal, Gamma, Beta, Dirichlet, Poisson, and so on. As it turns out, if $p(x|r)$ belongs to any single exponential family of distributions for all $r \in \mathcal{R}$, then it can be shown that $\log \frac{p(x|0)}{p(x|1)}$ is an affine function of x. At a glance this arises from the fundamental group homomorphism between multiplication and addition via the logarithm, and we examine this fundamental theme in detail in Section 2.4.1 in the context of generalized concavity.

Moreover, it can be shown that this correspondence is one-to-one [16]. That is, if the log-odds ratio is affine in x, then $p(x|r)$ belongs to an exponential family for all $r \in \mathcal{R}$. This fundamental fact could have important applications for ADP via approximation in policy space, although we will not consider such applications here. We first turn to elucidate the fundamental properties and challenges in this problem setting.

2.1.2 Fundamental Properties

Definition 13 (Logistic function)**.** *We define the logistic function* $f : \overline{\mathbb{R}} \to [0, 1]$ *as*

$$f(\varphi) = \frac{1}{1 + e^{-\varphi}}. \tag{2.2}$$

The logistic function possesses several fundamental (anti-)symmetries, verifiable by straightforward computation:

Lemma 9. *The logistic function* (2.2) *satisfies*

$$i.\ 0 < \|f\| < 1, \tag{2.3}$$

$$ii.\ 1 = f(\varphi) + f(-\varphi), \tag{2.4}$$

$$iii.\ \frac{\partial}{\partial \varphi} f(\varphi) = f(\varphi) f(-\varphi), \tag{2.5}$$

$$iv.\ \varphi = \log f(\varphi) - \log f(-\varphi). \tag{2.6}$$

Remark 1. *To ease the notation in what follows, we shall adopt the subscript convention to denote partial derivatives when convenient. Thus,* $\frac{\partial}{\partial x} f = \partial_x f = f_x$*. For higher-order derivatives, we may write* $f_{xx} = \partial_x^2 f$*,* $f_{xy} = \partial_{xy} f$*, etc. No confusion should arise.*

Definition 14 (Logistic model)**.** *By a logistic model, we understand a statistical model S_L on $\mathcal{R}$, together with a mixed coordinate system $(u, \eta) \in (\mathcal{U} \times \mathcal{N}) \subset \mathbb{R}^l \times \mathbb{R}^k$, $l, k \in \mathbb{N}$,*

$$S_L \triangleq \left\{ p(r; u, \eta) = \frac{1}{1 + \exp\left[\left\langle \alpha(r; \eta), u \right\rangle + \beta(r; \eta)\right]} \right\}.$$

For all distinct $r, \tilde{r} \in \mathcal{R}$, $u \in \mathcal{U}$ and $\eta \in \mathcal{N}$, we have

$$\log\left[\frac{\Phi(r; u, \eta)}{\Phi(\tilde{r}; u, \eta)}\right] = \left\langle a(\eta; r), u \right\rangle + b(\eta; r), \tag{2.7}$$

where $a(\eta; r) \in \mathbb{R}^l$, $\beta(\eta; r) \in \mathbb{R}$ for all $(\eta, r) \in \mathcal{N} \times \mathcal{R}$. That is, the log-odds ratio *of any two points in S_L is an affine function on $\mathcal{U}$.*

Example 21 (Binary classification)**.** *Let $\mathcal{R} = \{0, 1\}$ and consider the logistic model $S_L(\{0, 1\})$. Then immediately from* (2.6)*, we can see that each point in S_L is a logistic function with φ identically the RHS of* (2.7)*.*

Explicitly, we write the following: Without loss of generality, the condition (2.7) *becomes*

$$\log\left[\frac{\Phi(1; u, \eta)}{\Phi(0; u, \eta)}\right] = \left\langle a(\eta; 1), u \right\rangle + b(\eta; 1).$$

As S_L is a statistical model, $\Phi(\cdot; u, \eta)$ is a probability distribution, implying $1 = \Phi(0; u, \eta) + \Phi(1; u, \eta) = \Phi(0; u, \eta) + \Phi(0; u, \eta) e^{\langle a(\eta;1), u\rangle + b(\eta;1)}$. Therefore, we see that

$$\Phi(0; u, \eta) = \frac{1}{1 + e^{\langle a(\eta;1), u\rangle + b(\eta;1)}},$$

and analogously for $r = 1$. Hence, the r^{th} component of all $\Phi \in S_L$ has the form of a logistic function (2.2) *with $\varphi(r; u, \eta) = e^{\langle a(\eta;\check{r}), u\rangle + b(\eta;\check{r})}$.*

In anticipation of our principal application to clinical trials below, and to engender clarity in our investigations, we will devote special attention to the case of logistic regression with binary observations (i.e., binary classification). In particular, we introduce now what we call a Type I model. The motivations for this model class are elucidated in detail in Chapter 5, but at this moment we simply introduce the model without derivation to fix ideas for our subsequent investigations.

Definition 15 (Type I logistic model)**.** *By a* Type I (logistic) model *we understand a logistic model of Example 21, where in particular the parameter* $\eta \in \mathcal{N} \subset \mathbb{R}_{>0}$ *is defined as the* ν*–quantile parameter,* $\nu \in (0, 1/2)$*. That is,*

$$\Phi(1; \eta, \eta) \equiv \nu, \qquad \forall\, \eta \in \mathcal{N}.$$

Further, we may standardize the model to the unit interval, so that $\mathcal{U} \triangleq [0, 1]$*, and* $\mathcal{N} \triangleq [\epsilon, 1]$*,* $0 < \epsilon \ll 1$*. Finally, it is assumed that the probability of* $r = 1$ *is identically* p_ϵ *for* $u = 0$*, where* $0 < p_\epsilon < \nu$*. That is,*

$$\Phi(1; 0, \eta) \equiv p_\epsilon, \qquad \forall\, \eta \in \mathcal{N}.$$

In this case, the statistical model $\{\Phi(r; u, \eta),\ \forall\, (u, \eta) \in \mathcal{U} \times \mathcal{N}\}$ *thus takes the form*

$$\Phi(r; u, \eta) := \begin{cases} \dfrac{1}{1 + e^{-\varphi(u,\eta)}}, & r = 1, \\ \dfrac{1}{1 + e^{\varphi(u,\eta)}}, & r = 0, \end{cases} \tag{2.8}$$

where the discriminant φ *is given by*

$$\varphi(u, \eta) := \log\left[\frac{p_\epsilon}{1 - p_\epsilon}\right] + \frac{u}{\eta} \log\left[\frac{\nu\,(1 - p_\epsilon)}{p_\epsilon(1 - \nu)}\right]. \tag{2.9}$$

The standardized model thus admits the explicit form

$$\frac{1}{1 + \frac{1-p_\epsilon}{p_\epsilon}\left(\frac{(1-p_\epsilon)\nu}{p_\epsilon(1-\nu)}\right)^{-u/\eta}}. \tag{2.10}$$

For this case, we occasionally employ a simplified notation; specifically, we write

$$\Phi(u, \eta) \triangleq \Phi(1; u, \eta), \tag{2.11}$$

$$\varphi(u, \eta) \triangleq \varphi(1; u, \eta). \tag{2.12}$$

Additionally, we will maintain the convention of understanding r as an indicator of toxicity, whereby $r = 1$ means toxic, $r = 0$ means nontoxic. Thus, $u \mapsto \Phi(0; u, \eta)$ is a survival function, and its complement is a toxicity function. Formally, this is tantamount to the assumption

$$\partial_u \varphi(1; u, \eta) = \varphi_u = \alpha(\eta) > 0. \tag{2.13}$$

We have the following relations on the discriminant mapping $\varphi(u, \eta)$ of Type I models.

Assumption 1 (Type I Discriminant Assumptions)**.**

$$\begin{aligned} \text{Dose-response model:} \quad & 0 < \varphi_u = \alpha\,, \\ \text{Clinical application:} \quad & 0 < p_\epsilon < \nu < 1/2\,. \end{aligned}$$

Additionally, we introduce the following definitions, to more concisely represent certain quantities that feature frequently in calculations:

Definition 16 (Logarithmic quantities)**.**

$$\begin{aligned} \ell_\nu &:= \log\left(\frac{\nu}{1-\nu}\right), \\ \ell_{p_\epsilon} &:= \log\left(\frac{p_\epsilon}{1-p_\epsilon}\right), \\ \psi &:= \ell_\nu - \ell_{p_\epsilon} = \log\left(\frac{\nu}{1-\nu}\right) - \log\left(\frac{p_\epsilon}{1-p_\epsilon}\right), \\ \tilde{\psi} &:= \ell_\nu + \ell_{p_\epsilon} = \log\left(\frac{\nu}{1-\nu}\right) + \log\left(\frac{p_\epsilon}{1-p_\epsilon}\right). \end{aligned}$$

Finally, we collect some fundamental quantities for Type I models. We omit the proof, as all are verifiable by straightforward computation. Note that in a standardized model, $a = 0$.

Lemma 10 (Discriminant Relations)**.**

$$\begin{aligned} \text{(Logarithmic relations)} \quad & \ell_{p_\epsilon} < \ell_\nu < 0\,, \\ & \tilde{\psi} < 0 < \psi\,; \\ \text{(Derivative relations)} \quad & \varphi_u =: \alpha = \psi(\eta - a)^{-1} > 0, \\ & \varphi_\eta = \alpha_\eta u + \beta_\eta = \alpha_\eta(u-a) = \varphi_{u\eta}(u-a) < 0\,, \\ & \varphi_{u\eta} =: \alpha_\eta = -\psi(\eta-a)^{-2} < -\alpha < 0\,, \\ & \varphi_{u\eta\eta} =: \alpha_{\eta\eta} = 2\psi^2(\eta-a)^{-3} > \alpha > 0\,, \\ & \beta = (1/2)\left(\tilde{\psi} - \alpha(\eta+a)\right) < 0\,, \\ & \beta_\eta = -a\alpha_\eta > 0\,, \\ & 2\alpha_\eta^2 = \alpha\alpha_{\eta\eta}\,. \end{aligned}$$

2.2 Stochastic Order

We begin by introducing the central actor in this inference framework, the Bayes operator. Indeed, the Bayes operator may be viewed as the engine for updating belief in view of new information.

Definition 17 (Bayes operator)**.** *Let $s \in \mathcal{P}(\mathcal{N})$ be an arbitrary probability density on $\mathcal{N}$. By the* Bayes operator *we understand the operator $\Psi : \mathcal{R} \times \mathcal{U} \times \mathcal{P}(\mathcal{N}) \to \mathcal{P}(\mathcal{N})$ given by*

$$\Psi(r; s, u)(\cdot) \equiv \frac{\Phi(r; u, \cdot)s(\cdot)}{\int_{\mathcal{N}} \Phi(r; u, \eta)s(\eta)} \equiv \frac{\Phi(r; u, \cdot)s}{\big\langle \Phi(r; u, \cdot), s \big\rangle}. \tag{2.14}$$

Definition 18 (Normalization operator)**.** *By the* normalization operator *we understand the operator $Z : \mathcal{R} \times \mathcal{P}(\mathcal{N}) \times \mathcal{U} \to (0, 1]$ given by*

$$Z(r; s, u) = \int_{\mathcal{N}} \Phi(r; u, \eta)s(\eta) = \big\langle \Phi(r; u, \cdot), s \big\rangle. \tag{2.15}$$

The reason behind the nomenclature of Z is clear: It features as the denominator in the Bayes operator to ensure a probability distribution. However, this ostensibly pedestrian purpose obscures its pivotal role in the mechanics below.

Remark 2. *We note that oftentimes the explicit notation may be cumbersome. When a static setting is under consideration, which is to say the identity of s is contextually clear, we may write*

$$Z^r_u = Z(r; u) = Z(r; s, u), \text{ and similarly, } \Psi^r_u = \Psi(r; \cdot, u) = \Psi(r; s, u).$$

Additionally, in the binary case $\mathbb{R} = \{0, 1\}$, the form of Z simplifies so that $Z(0; , s, u) = 1 - Z(1; s, u)$. When convenient, we shall therefore write

$$z(u) = Z(1; s, u), \text{ whence, } 1 - z(u) = Z(0; s, u).$$

Theorem 11 (Monotonicity Properties)**.** *Let $\mathcal{R} = \{0, 1\}$. For all $u_i, u_j \in \mathcal{U}$, where $u_i < u_j$, and all $\eta \in \mathcal{N}$, $s \in \mathcal{P}(\mathcal{N})$, the operators Φ and Z satisfy the following*

monotonicity properties pointwise:

$$\prod_{r\in\mathcal{R}} \Big(\Phi(r;u_i,\eta) - \Phi(r;u_j,\eta)\Big) < 0,$$
$$\prod_{r\in\mathcal{R}} \Big(Z(r;u_i) - Z(r;u_j)\Big) < 0.$$

In the shorthand notation, we have explicitly

$$\Phi(u_i,\eta) < \Phi(u_j,\eta),$$
$$z(u_i) < z(u_j).$$

Proof. Let s be a given probability density on $\mathcal{N}$. Recall from (2.5) that for all $\eta \in \mathcal{N}$,

$$\frac{\partial}{\partial u}\Phi = \varphi_u \Phi(1-\Phi) > 0, \tag{2.16}$$

where the final inequality follows from (2.3) and assumption (2.13). It follows that $(u_j - u_i) > 0 \implies (\Phi(u_j,\eta) - \Phi(u_j,\eta)) > 0$, for all $\eta \in \mathcal{N}$. That is,

$$\big\langle u_j - u_i, \Phi(u_j,\eta) - \Phi(u_i,\eta)\big\rangle = (u_j - u_i)(\Phi(u_j,\eta) - \Phi(u_i,\eta)) > 0, \tag{2.17}$$

and we see that $\Phi : \mathcal{U} \to \mathscr{F}(\mathcal{N})$ is a monotonic operator in its first argument. The opposite result for $1-\Phi$ follows similarly.

We can establish the claims on the normalization operator Z by a similar computation.

$$\begin{aligned}
\frac{\partial}{\partial u} Z(1;u) &:= \partial_u \int_{\mathcal{N}} \Phi(u,\eta) s(\eta)\, d\eta \\
&= \int_{\mathcal{N}} \partial_u \Phi(u,\eta) s(\eta)\, d\eta \\
&= \int_{\mathcal{N}} \varphi_u \Phi(u,\eta)\big(1-\Phi(u,\eta)\big) s(\eta)\, d\eta \\
&> 0.
\end{aligned}$$

As before, the final inequality follows from (2.3), the assumption (2.13), and the fact that s is a probability density. Therefore,

$$(u_j - u_i)(Z(1;u_j) - Z(1;u_i)) > 0, \tag{2.18}$$

and we see that Z is a monotonic operator. The result for $Z(0;\cdot) = 1 - Z(1;\cdot)$ follows similarly. □

Theorem 12 establishes the respective monotonic movements of posterior distribution for each response. With respect to any prior (i.e., marginal) density $s \in \mathcal{P}(\mathcal{N})$, toxic response $r = 1$ shifts the posterior toward the origin, whereas a non–toxic response $r = 0$ shifts the posterior away from the origin, for any any dose.

Theorem 12 (Response dominance)**.** *Let the prior density $s \in \mathcal{P}(\mathcal{N})$ be given, and denote the random variable distributed according to s by η. For each response $r \in \mathcal{R} = \{0, 1\}$ and $u \in \mathcal{U}$, denote the random variable distributed according to the the posterior density $\Psi_u^r := \Psi(r; s, u)$ by η_u^r. Then for all $u \in \mathcal{U}$,*

$$\eta_u^1 \preceq_{\pentagon} \eta \preceq_{\pentagon} \eta_u^0, \tag{2.19}$$

where $\preceq_{\pentagon}$ denotes the increasing convex order*; that is,*

$$\mathbb{E}_{\Psi_u^1}\big[g(\eta_u^1)\big] \leq \mathbb{E}_s[g(\eta)] \leq \mathbb{E}_{\Psi_u^0}\big[g(\eta_u^0)\big]\,,$$

for all increasing, convex functions $g : \mathbb{R} \to \mathbb{R}$ such that the expectations exist.

Proof. This follows as a direct consequence of the monotonicity established in Theorem 11. We proceed to establish the assertion by way of an equivalent expression, namely

$$\mathbb{E}_{\Psi_u^1}\Big[\,\big(\eta_u^1 - x\big)_+\,\Big] \leq \mathbb{E}_s\Big[\,\big(\eta - x\big)_+\,\Big] \leq \mathbb{E}_{\Psi_u^0}\Big[\,\big(\eta_u^0 - x\big)_+\,\Big], \tag{2.20}$$

for all $x \in \mathcal{N}$. This is valid, since any increasing, convex function $g(\eta)$ may be arbitrarily closely approximated by a positive combination of functions $g_k(\eta) = a_k + (\eta - x_k)_+$, and therefore (2.20) holding for all $x \in \mathcal{N}$ implies the statement. (See [129] for a more general discussion of this technique.)

Without loss of generality, let $\mathcal{N} = [0, 1]$ and $F(x, \phi) := \int_x^1 (\xi - x)\phi(\xi)\, d\xi$. Note that $F(x, \phi) = \mathbb{E}_\phi[(\xi - x)_+] \geq 0$ by construction. Fix $u \in \mathcal{U}$. To begin, first observe that by

virtue of the natural complementarity, we have

$$\begin{aligned} F(x,\Psi_u^0) &= \int_x^1 (\xi - x)\Psi_u^0(\xi)\,d\xi \\ &= \frac{1}{1-z(u)}\int_x^1 (\xi - x)\big(1-\Phi(u,\xi)\big)s(\xi)\,d\xi \\ &= \frac{1}{1-z(u)}\left(\int_x^1 (\xi - x)s(\xi)d\xi - \int_x^1 (\xi - x)\Phi(u,\xi)s(\xi)\,d\xi\right) \\ &= \frac{1}{1-z(u)}\Big(F(x,s) - z(u)F(x,\Psi_u^1)\Big). \end{aligned}$$

Therefore, we arrive at a familiar identity:

$$F(x,s) = z(u)F(x,\Psi_u^1) + \big(1-z(u)\big)F(x,\Psi_u^0), \quad \forall\, x \in \mathcal{N}, \tag{2.21}$$

demonstrating $F(x,s)$ is a convex combination of the others.

We need now only to establish that $F(x,\Psi_u^1) \le F(x,\Psi_u^0)$. Consider

$$\begin{aligned} F(x,\Psi_u^0) - F(x,\Psi_u^1) &= \frac{1}{1-z(u)}F(x,s) - \frac{z(u)}{1-z(u)}F(x,\Psi_u^1) \\ &\le \frac{1}{z(u)(1-z(u))}F(x,s) - \frac{1}{1-z(u)}F(x,\Psi_u^1) \\ &\le \left(\frac{1}{z(u)(1-z(u))} - \frac{1}{1-z(u)}\right)F(x,s) \\ &= \frac{1-z(u)}{z(u)(1-z(u))}F(x,s) \\ &= \frac{1}{z(u)}F(x,s) \\ &< 0, \end{aligned}$$

where the inequalities follow by the monotonicity of Theorem 11 and the fact that $z(u) \in (0,1]$.

Note that all of the above holds pointwise for all $x \in \mathcal{N}$. Moreover, note that in all the above considerations, $u \in \mathcal{U}$ is arbitrary, and thus (2.20) holds for all $x \in \mathcal{N}$ and $u \in \mathcal{U}$, establishing the claim in (2.19).

□

Remark 3. *In the medical statistics literature, a* coherent *dosage policy is defined by the following property: If the previous patient was toxic, the next dose does not increase, whereas if the previous patient was non–toxic, the next dose does not decrease. Thus, this property is satisfied in particular whenever* (2.19) *holds.*

It is perhaps intuitive that posterior distributions shift in direct correspondence to the magnitude of the dose administered, preserving stochastic order relations. This property is generally not true, however, and in what follows we focus efforts on characterizing when this does hold. Perhaps interestingly, even through extensive empirical study, one is hard pressed to construct an instance where this property does not hold true. For example, in all but the most pathological states shown in Figure 3.4, the ordering holds. Yet proving this can be surprisingly be elusive, owing largely to diabolical normalization operators in the denominator.

We shall endeavor to make progress along these lines in the particular case of Type I logistic models, and defer general considerations to future work. We will first collect some facts about the relevant quantities and structure the analysis. We begin by recapitulating the setting:

Without loss of generality, let $\mathcal{N} = [0, 1]$, and $u_i, u_j \in \mathcal{U}$ be such that $u_i < u_j$. As before, for any $\phi : \mathcal{N} \to [0, 1]$, we define

$$F(x, \phi) \triangleq \int_x^1 (\xi - x)\phi(\xi)\, d\xi, \tag{2.22}$$

$$G(x, \phi) \triangleq \int_0^x (x - \xi)\phi(\xi)\, d\xi. \tag{2.23}$$

Just as before, note that $F(x, \phi) = \mathbb{E}_\phi[(\xi - x)_+] \geq 0$, and that $G(x, \phi) = \mathbb{E}_\phi[(x - \xi)_+] \geq 0$, by construction.

We focus on the case of $r = 1$, so that the relevant posteriors are Ψ_i^1, and Ψ_j^1. We employ the usual shorthand when needed as follows: Let $F_i^1(x)$ be given by $F_i^1(x) \triangleq F(x, \Psi_i^1) = F(x, \Phi(u_i)s/z(u_i))$, and similarly for $G_i^1(x)$; explicitly,

$$F(x, \Psi_u^1) = \frac{1}{z(u)} \int_x^1 (\xi - x)\Phi(u, \xi)s(\xi)\, d\xi,$$
$$G(x, \Psi_u^1) = \frac{1}{z(u)} \int_0^x (x - \xi)\Phi(u, \xi)s(\xi)\, d\xi.$$

For $r = 1$ the condition that the u_i– and u_j–posterior distributed random variables $\eta_i^1 \sim \Psi_i^1$, $\eta_j^1 \sim \Psi_j^1$ satisfy the increasing convex order $\eta_i^1 \preceq_{\text{cx}} \eta_j^1$ is equivalent to

$$F_i^1(x) \leq F_j^1(x), \quad \forall\, x \in \mathcal{N}. \tag{2.24}$$

An equivalent form of this condition can be made in terms of G^1. For any $r \in \{0,1\}$ and $u_k \in \mathcal{U}$, define the quantity

$$M_k^r(x) \triangleq F_k^r(x) + G_k^r(x), \quad x \in \mathcal{N}. \tag{2.25}$$

Thus, (2.24) is equivalent to

$$G_j^1(x) - G_i^1(x) \leq M_j^1(x) - M_i^1(x). \quad \forall\, x \in \mathcal{N}. \tag{2.26}$$

Our goal is to establish (2.24), (2.26), and it is clear that the conditions are dynamic, in the sense that they are focused on the change with respect to u and with x. Given our focus on Type I logistic models, which is to say we have some knowledge of the form of the general expressions, it is natural to use the calculus. Differentiating under the integral reveals that

$$\begin{aligned} \partial_x G_i^1(x) &= \partial_x \left[\frac{1}{z_i} \int_0^x (x-\xi)\Phi_i(\xi) s(\xi)\, d\xi \right] \\ &= \frac{1}{z_i} \int_0^x \Phi_i(\xi) s(\xi)\, d\xi \\ &=: \frac{1}{z_i} g_i^1(x). \end{aligned} \tag{2.27}$$

The complimentary result holds for $\partial_x F_i^1(x)$, namely

$$\partial_x F_i^1(x) =: -\frac{1}{z_i} f_i^1(x). \tag{2.28}$$

The form of the expressions (2.27), (2.28) intimates a simplectic structure reminiscent of a Hamiltonian system. Observe that the (r, u_k)–posterior mean median deviation MMD_k^r satisfies $\mathrm{MMD}_k^r = M_k^r(\hat{\mu}_k^r)$, where $\hat{\mu}_k^r$ denotes the (r, u_k)–posterior median. Specifically, the median is the minimizer of (2.25), so that we may define

$$\mathrm{MMD}_k^r \triangleq \min_{x \in \mathcal{N}} M_k^r(x), \tag{2.29}$$

$$\hat{\mu}_k^r \triangleq \operatorname*{argmin}_{x \in \mathcal{N}} M_k^r(x). \tag{2.30}$$

More to the point, the MMD_k^r is a *first integral* in the sense that it is constant along orbits of $M_k^r(x)$, which is to say it is independent of x. One can ask if there exist other

such constants. The relevant condition would be $\partial_x M_i^1 \equiv 0$, which by way of (2.27) and (2.28) implies

$$\begin{aligned} & g_i^1(x) = f_i^1(x) \\ \iff \quad & \int_0^x \Phi_i(\xi) s(\xi)\, d\xi = \int_x^1 \Phi_i(\xi) s(\xi)\, d\xi. \end{aligned}$$

This is precisely the defining characteristic of the median $\hat{\mu}_i^r$. Thus, we obtain the generalized coordinates $(\mathrm{MMD}_i^1, \hat{\mu}_i^1)$, which is to say that their knowledge is tantamount to knowledge of our system dynamics.

Let us now introduce a certain quantity, which we denote by $\Phi_{(\delta)}$ and define as

$$\Phi_{(\delta)} = \Phi_{(\delta)}(u_i, u_j; \eta) \triangleq \Phi(u_j, \eta) - \Phi(u_i, \eta), \ \forall\, \eta \in \mathcal{N}. \tag{2.31}$$

First, observe that by the linearity of the integral, for any density s we have

$$\int_{\mathcal{N}} \Phi_{(\delta)} s = \int_{\mathcal{N}} \Phi_j s - \int_{\mathcal{N}} \Phi_i s = z_j - z_i \triangleq z_{(\delta)}. \tag{2.32}$$

Pursuant to these definitions, we similarly define

$$\begin{aligned} G_{(\delta)}^1(x) \triangleq G(x, \Phi_{(\delta)} s) &= G\left(x, \frac{\Phi_j - \Phi_i}{z_j - z_i} s\right) \\ &= \left(\frac{z_j}{z_j - z_i}\right)\Big(G_j^1 - G_i^1\Big), \end{aligned} \tag{2.33}$$

$$F_{(\delta)}^1 \triangleq \left(\frac{z_j}{z_j - z_i}\right)\Big(F_j^1 - F_i^1\Big), \tag{2.34}$$

and finally, with $M_{(\delta)}^1 \triangleq F_{(\delta)}^1 + G_{(\delta)}^1$, it follows that

$$M_{(\delta)}^1 = \left(\frac{z_j}{z_j - z_i}\right)\Big(M_j^1 - M_i^1\Big). \tag{2.35}$$

As before, because of sufficient continuity and differentiability properties for the logistic model and the form of (2.36), $M_{(\delta)}^1$ inherits a unique first integral, denoted $\hat{\mu}_{(\delta)}^1$. In particular, we again have that this is the minimizer of $M_{(\delta)}^1$, so that

$$\mathrm{MMD}_{(\delta)}^1 \triangleq \min_{x \in \mathcal{N}} M_{(\delta)}^1(x) = M_{(\delta)}^1(\hat{\mu}_{(\delta)}^1). \tag{2.36}$$

In view of (2.26), we may yet again recast the condition for $\eta_i^1 \preceq_{\pentagon} \eta_j^1$ as

$$G_{(\delta)}^1(x) \leq M_{(\delta)}^1(x), \quad x \in \mathcal{N}. \tag{2.37}$$

This form simplifies harmoniously owing to the linearity of the normalization operator, but is not necessarily immediately more useful, as the first integral $\hat{\mu}^1_{(\delta)}$ is a somewhat subtle quantity. We suspect that it, and its dual counterpart $\mathrm{MMD}^1_{(\delta)}$, can be connected to the mutual information between η^1_i and η^1_j (and, in turn, the relative entropy between Ψ^1_i and Ψ^1_j, expressible in terms of divergences, etc.) in a very straightforward way. We leave these promising avenues for future work, and are content for our part to proceed to study this in our current context.

Let $\mathcal{X}_{(\delta)} \subset \mathcal{N}$ be defined by $\mathcal{X}_{(\delta)} \triangleq \{x \in \mathcal{N} \,|\, G^1_{(\delta)}(x) \leq \mathrm{MMD}^1_{(\delta)}\}$. Then it is clear from (2.36) that (2.37) is satisfied for all $x \in \mathcal{X}_{(\delta)}$. This more stringent condition offers a stronger version of the second-order dominance, insofar as it is restricted to the subspace $\mathcal{X}_{(\delta)}$, without diluting its guarantee. In the next Section 2.3, we consider the functional form of Type I models and show that $\mathcal{X}_{(\delta)}$ does, at least, admit a convex representation in terms of u_i, u_j. Along the way, we establish some relations on $\hat{\mu}^1_{(\delta)}$ in terms of $\hat{\mu}^1_i$, $\hat{\mu}^1_j$ that lend themselves to an interesting heuristic. First, we summarize the salient features of the above discussion in the following theorem.

Theorem 13 (Restricted dose dominance)**.** *Let the prior density $s \in \mathcal{P}(\mathcal{N})$ be fixed, denote the random variable distributed according to s by η, and let $r \in \mathcal{R}$ be given as $r = 1$. Let $\mathcal{X}_{(\delta)} \subset \mathcal{N}$ be given by*

$$\mathcal{X}_{(\delta)} = \left\{ x \in \mathcal{N} \,\middle|\, G^1_j(x) - G^1_i(x) \leq M^1_j(\hat{\mu}^1_{(\delta)}) - M^1_i(\hat{\mu}^1_{(\delta)}) \right\}, \tag{2.38}$$

where $\hat{\mu}^1_{(\delta)}$ is defined by (2.36)*.*

Then for all $u_i, u_j \in \mathcal{X}_{(\delta)}$ such that $u_i < u_j$, the $\Psi^1(u_i)$ and $\Psi^1(u_j)$ posterior–distributed random variables η^1_i and η^1_j satisfy the increasing convex order:

$$\eta^1_i \preceq_{\pentagon} \eta^1_j. \tag{2.39}$$

Proof. The claim follows from the above discussion by the unique invariance of the median as a first integral of the partial integral functional. Beginning from (2.32), we immediately have (2.36), which is equivalent to the condition for the increasing convex order by construction.

□

2.3 On the Monotonicity of the Median

We yet again begin by recapitulating the setting: Consider a standardized model, so that $\mathcal{N} = [0,1]$, fix a state $s \in \mathcal{P}(\mathcal{N})$, and let distinct points $u_i, u_j \in \mathcal{U}$ be given, such that

$$u_i < u_j. \tag{2.40}$$

For any response r, let $\hat{\mu}^r$ denote the posterior median operator, such that for any control u, $\hat{\mu}^r(u)$ denotes the median of the posterior distribution under control u and realized response r; see (2.30). Yet again, we employ the shorthand notation introduced above, so that $\hat{\mu}^r(u) = \hat{\mu}^r_u = \hat{\mu}(\Psi^r_u)$, etc.

We recall the following functions:

$$\begin{aligned}
F^1_u(x) = F(x, \Psi^1_u) &= \frac{1}{z(u)} \int_x^1 (\xi - x)\Phi(u,\xi)s(\xi)\, d\xi, \\
G^1_u(x) = G(x, \Psi^1_u) &= \frac{1}{z(u)} \int_0^x (x - \xi)\Phi(u,\xi)s(\xi)\, d\xi, \\
f^1_u(x) = f(x, \Psi^1_u) &= \int_x^1 \Phi(u,\xi)s(\xi)\, d\xi, \\
g^1_u(x) = g(x, \Psi^1_u) &= \int_0^x \Phi(u,\xi)s(\xi)\, d\xi.
\end{aligned}$$

The functions are related according to (2.27) and (2.28), namely

$$\begin{aligned}
\partial_x F^1_u(x) &= -\frac{1}{z(u)} f^1_u(x), \\
\partial_x G^1_u(x) &= \quad \frac{1}{z(u)} g^1_u(x).
\end{aligned}$$

Our point of departure is the consideration of two distinct points $u_i, u_j \in \mathcal{U}$, together with the assumption in (2.40). Our goal will be to ascertain the form of the space $\mathcal{X}_{(\delta)}$, and to clarify the role of the differenced quantities introduced in the previous section. First, recall that by the definition of $\Phi_{(\delta)}$ in (2.31) the normalization function $z_{(\delta)}$ in (2.32), we have the pair of identities

$$g^1_{(\delta)}(\hat{\mu}^1_{(\delta)}) = g^1_i(\hat{\mu}^1_i) - g^1_j(\hat{\mu}^1_j), \tag{2.41}$$

$$f^1_{(\delta)}(\hat{\mu}^1_{(\delta)}) = f^1_i(\hat{\mu}^1_i) - f^1_j(\hat{\mu}^1_j). \tag{2.42}$$

Note also that by linearity of these functions in the first argument, which is to say the linearity of the integral operator, (2.41) readily implies

$$\begin{aligned} & g_i^1(\hat{\mu}_{(\delta)}^1) - g_j^1(\hat{\mu}_{(\delta)}^1) = g_i^1(\hat{\mu}_i^1) - g_j^1(\hat{\mu}_j^1) \\ \iff \qquad & g_i^1(\hat{\mu}_{(\delta)}^1) - g_i^1(\hat{\mu}_i^1) = g_j^1(\hat{\mu}_{(\delta)}^1) - g_j^1(\hat{\mu}_j^1). \end{aligned}$$

That is, writing the last displayed equation explicitly and combining the integrals, we see more clearly that

$$\int_{\hat{\mu}_i^1}^{\hat{\mu}_{(\delta)}^1} \Phi_i s = \int_{\hat{\mu}_j^1}^{\hat{\mu}_{(\delta)}^1} \Phi_j s. \tag{2.43}$$

From (2.43) we may conclude the following: The positivity of the integrand immediately implies that if $\hat{\mu}_{(\delta)}^1 > \hat{\mu}_i^1$, then the LHS is necessarily positive and thus $\hat{\mu}_{(\delta)}^1 > \hat{\mu}_j^1$. By the same reasoning, $\hat{\mu}_{(\delta)}^1 < \hat{\mu}_i^1$ would imply the LHS is negative, and thus $\hat{\mu}_{(\delta)}^1 < \hat{\mu}_j^1$. The converse clearly holds in both cases, too. We summarize this result in the following Theorem 14.

Theorem 14. *For all $u_i, u_j \in \mathcal{U}$, $u_i < u_j$, the first integral $\hat{\mu}_{(\delta)}^1$ satisfies the following ordering*

$$\begin{cases} \hat{\mu}_{(\delta)}^1 > \hat{\mu}_i^1 \iff \hat{\mu}_{(\delta)}^1 > \hat{\mu}_j^1, \\ \hat{\mu}_{(\delta)}^1 < \hat{\mu}_i^1 \iff \hat{\mu}_{(\delta)}^1 < \hat{\mu}_j^1, \end{cases}$$

or, equivalently but more succinctly,

$$\left(\hat{\mu}_{(\delta)}^1 - \hat{\mu}_i^1\right)\left(\hat{\mu}_{(\delta)}^1 - \hat{\mu}_j^1\right) > 0. \tag{2.44}$$

Proof. This follows immediately from (2.43) and the positivity of the integrand $\Phi_u s > 0$ for all $u \in \mathcal{U}$, as described in the prior discussion. □

Corollary 15. *For any $u_i, u_j \in \mathcal{U}$, $u_i < u_j$, the following ordering holds:*

$$\left(\hat{\mu}_j^1 - \hat{\mu}_i^1\right)\left(\hat{\mu}_{(\delta)}^1 - \hat{\mu}_i^1\right)\left(\hat{\mu}_{(\delta)}^1 - \hat{\mu}_j^1\right) > 0. \tag{2.45}$$

Proof. Suppose $\hat{\mu}_i^1 < \hat{\mu}_j^1$. Then (2.44) implies

$$\hat{\mu}_i^1 < \hat{\mu}_j^1 < \hat{\mu}_{(\delta)}^1.$$

Conversely, suppose $\hat{\mu}_j^1 > \hat{\mu}_i^1$. Then (2.44) implies

$$\hat{\mu}_{(\delta)}^1 < \hat{\mu}_j^1 < \hat{\mu}_i^1.$$

The result (2.45) follows. □

Unfortunately, further assertions cannot be made from these elementary methods without further conditions on the functions involved. Employing our knowledge of the particular form of the logistic function in the case of Type I models, we can derive explicit functional forms characterizing when (2.45) holds with $\hat{\mu}^1(u_i) \leq \hat{\mu}^1(u_j)$. By implicit differentiation, we obtain

$$\frac{\partial^2 g_j^1(x)}{(\partial g_i^1(x))^2} = \frac{\Phi_j s}{(\Phi_i s)^2}\Big(\varphi_j'(1-\Phi_j) - \varphi_i'(1-\Phi_i)\Big). \tag{2.46}$$

Hence, in the case of a one-parameter model, we have

$$\begin{aligned}\varphi_j' &= -\psi(\eta - a)^{-2}(u_j - a)\\ \varphi_i' &= -\psi(\eta - a)^{-2}(u_i - a),\end{aligned}$$

from which we have

$$\varphi_j' - \varphi_i' = -\psi(\eta - a)^{-2}(u_j - u_i). \tag{2.47}$$

Note that (2.47) offers an alternative demonstration of the strong convexity of φ. Additionally, we have

$$\varphi_i'\Phi_i - \varphi_j'\Phi_j = -\psi(\eta - a)^{-2}\Big((u_i - a)\Phi_i - (u_j - a)\Phi_j\Big). \tag{2.48}$$

Considering the form of (2.46), the condition for our mapping $x \mapsto \Big(g_i^1(x), g_j^1(x)\Big)$ to be convex in fluid space (i.e., g_i^1-g_j^1 space) is that $0 \leq \varphi_j'(1-\Phi_j) - \varphi_i'(1-\Phi_i)$. Expanding this expression, we have

$$\begin{aligned}\varphi_j'(1-\Phi_j) - \varphi_i'(1-\Phi_i) &= \varphi_j' - \varphi_i' + \varphi_i'\Phi_i - \varphi_j'\Phi_j\\ &= -\psi(\eta - a)^{-2}\Big((u_j - u_i) + (u_i - a)\Phi_i - (u_j - a)\Phi_j\Big),\end{aligned}$$

and therefore our condition amounts to requiring that $0 < (u_j - u_i) + (u_i - a)\Phi_i - (u_j - a)\Phi_j$. Observing that the constant coefficients satisfy $u_j - a = (u_j - u_i) + (u_i - a)$, we

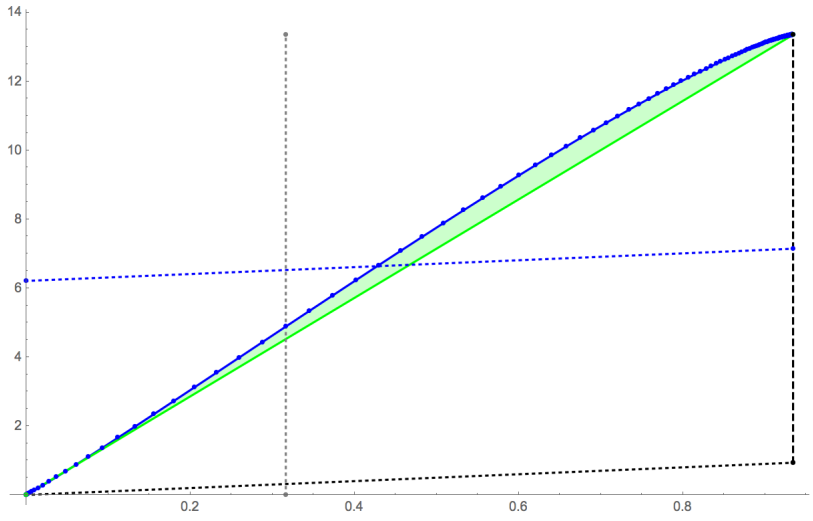


Figure 2.1: Visualizing the conditions (2.44), (2.45) for monotonicity of the myopic minimizer. Monotonicity requires that the dark blue line be below the green.

equivalently write this as

$$\begin{aligned} & 0 < (u_j - u_i) + (u_i - a)\Phi_i - (u_j - a)\Phi_j \\ \iff & 0 < (u_j - u_i) + (u_i - a)\Phi_i - (u_j - u_i)\Phi_j - (u_i - a)\Phi_j \\ \iff & 0 < (u_j - u_i)(1 - \Phi_j) - (u_i - a)\Phi_\delta \\ \iff & 0 < (1 - \Phi_j) - \gamma\Phi_\delta, \end{aligned}$$

where in the last expression we have let $\gamma := (u_i - a)/(u_j - u_i)$. Thus, we obtain the condition

$$\gamma\Phi_\delta < 1 - \Phi_j. \tag{2.49}$$

It is perhaps more instructive to view this as

$$\frac{(u_i - a)}{(u_j - u_i)} < \frac{1 - \Phi_j}{\Phi_j - \Phi_i}, \tag{2.50}$$

where given the bounds $a < u_i < u_j$, $\Phi_i < \Phi_j < 1$ we recognize this geometrically as the comparison in the following figure.

It can be shown that equality in (2.49) can occur at most once on the interior of $\mathcal{N}$, and thus it suffices to determine the point obtaining this equality. Enforcing equality (2.49) and expanding via the definition of Φ for a Type I model, we obtain the two

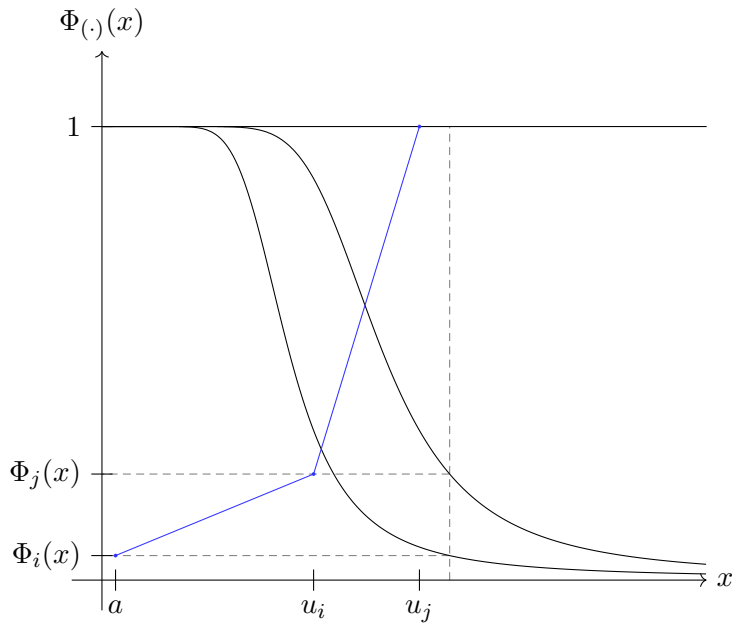


Figure 2.2: A visualization of the conditions for dose dominance in (2.50).

equivalent relations

$$(1+\gamma)e^{-\varphi(i)} = (2+\gamma)e^{-\varphi(j)} + e^{-(\varphi_i+\varphi_j)} \tag{2.51}$$

$$(1+\gamma)e^{\varphi(j)} = (2+\gamma)e^{\varphi(i)} + 1. \tag{2.52}$$

From these relations, we can extract simple relations on φ leading to the above convexity. The form of the relations proves interesting in view of our log-concavity considerations.

As it turns out, two results may be obtained from this analysis, although via slightly different methods. One approach is much simpler but yields only one of the results to be had, whereas the other requires more lengthy calculations but yields both results. We proceed first with the former, as the form of the analysis will be demonstrated yet the calculations may be done on fingers.

For simplicity, we make the following notational substitutions: Let $m = 1+\gamma$, and $n = 2+\gamma$. Beginning from (2.52), we solve for unity and multiply by $e^{\varphi_i+\varphi_j}$ to obtain

$$me^{\varphi_i+2\varphi_j} - ne^{2\varphi_i+\varphi_j} = e^{\varphi_i+\varphi_j}.$$

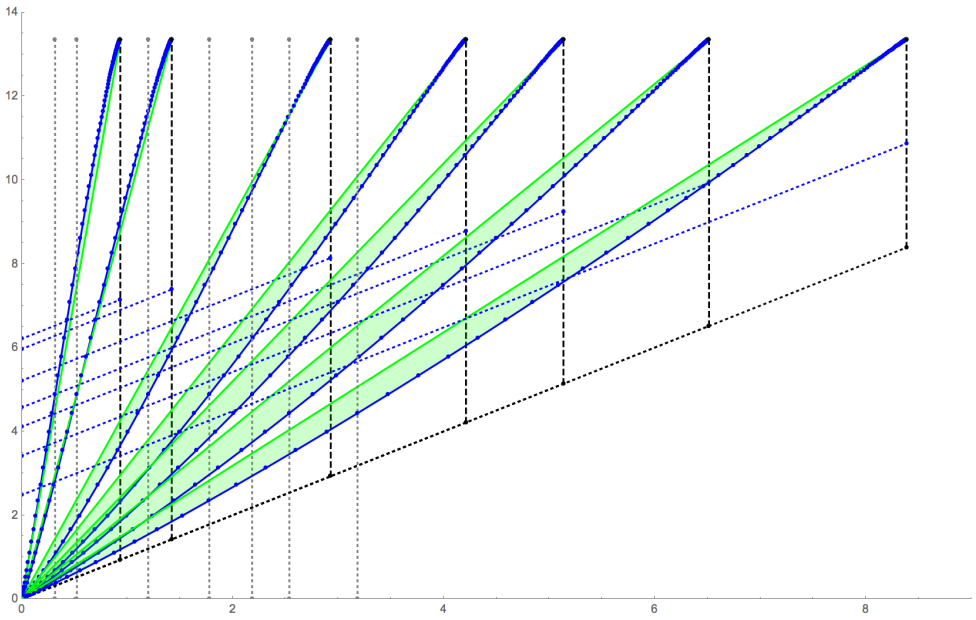


Figure 2.3: Visualizing the conditions (2.44), (2.45) for monotonicity of the myopic minimizer, for a spectrum of dosages u_i, u_j. Monotonicity requires that the dark blue line be below the green.

Comparing this to (2.51), we observe that

$$\left(me^{\varphi_i+2\varphi_j} - ne^{2\varphi_i+\varphi_j}\right)\left(me^{-\varphi(i)} - ne^{-\varphi(j)}\right) = 1. \tag{2.53}$$

After expanding and combining terms, (2.53) is equivalent to $\left(me^{\varphi(j)} - ne^{\varphi(i)}\right)^2 = 1$, from which we conclude

$$me^{\varphi(j)} - ne^{\varphi(i)} = \pm 1. \tag{2.54}$$

This expression is affine in exponentials, and shares connections to the conditions of log-concavity. Additionally, we write the relation in (2.54), which we will see below is equal to $+1$, in the illustrative form

$$\varphi_j = \log\left[\frac{1+m}{m}e^{\varphi(i)} + \frac{1}{m}\right], \tag{2.55}$$

and also $m(e^{\varphi(j)} - e^{\varphi(i)}) = 1 - e^{\varphi(i)}$. Finally, note that as $u_i \to u_j$, the quantity $m := 1 + \gamma \to \infty$, and therefore $\varphi_j \to \varphi_i$.

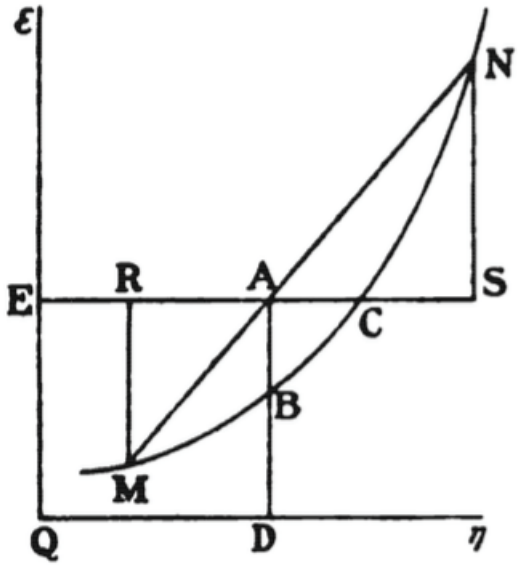


Figure 2.4: Interestingly, very similar graphs have been constructed by Gibbs in 1873 regarding thermodynamic properties of substances [65].

We now proceed with an alternative analysis along the same lines of inquiry. Beginning with (2.52), we take products with $e^{-\varphi(i)}$ and $e^{-\varphi(j)}$, respectively, to obtain

$$\begin{aligned}(1+\gamma)e^{-\varphi_i+\varphi_j} - e^{-\varphi(i)} &= (2+\gamma),\\ (1+\gamma) &= (2+\gamma)e^{\varphi_i-\varphi_j} + e^{-\varphi(j)}.\end{aligned}$$

Comparing these relations, we obtain

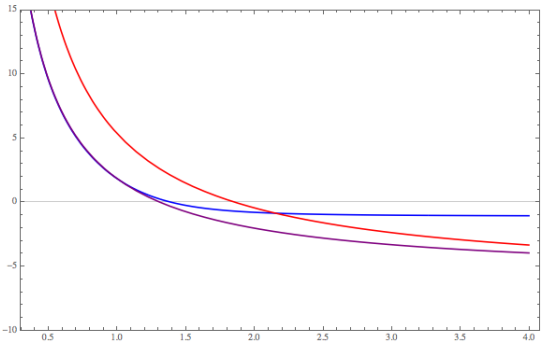
$$(1+\gamma)e^{-\varphi_i+\varphi_j} - e^{-\varphi(i)} = (2+\gamma)e^{\varphi_i-\varphi_j} + e^{-\varphi(j)} + 1. \tag{2.56}$$

As before, we take the product of this expression with $e^{\varphi_i+\varphi_j}$, yielding the relation

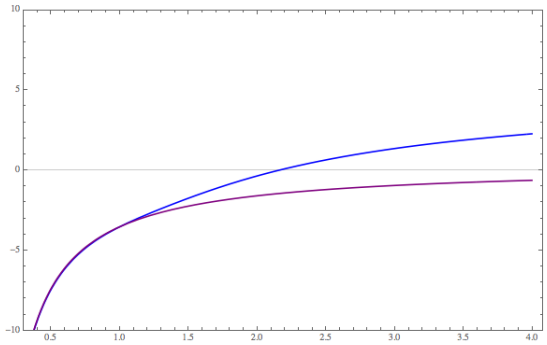
$$\left(me^{2\varphi(j)} - e^{\varphi(j)} - ne^{2\varphi(i)} - e^{\varphi(i)}\right)\left(me^{-\varphi(i)} - ne^{-\varphi(j)}\right) = 1. \tag{2.57}$$

To ease the calculation, we let $x := e^{-\varphi(i)}$ and $y := e^{-\varphi(j)}$. Expanding (2.57) with these substitutions, we obtain

$$\begin{aligned}& m^2\frac{x}{y^2} - m\frac{x}{y} - mn\frac{1}{x} - m - mn\frac{1}{y} + n + n^2\frac{y}{x^2} + n\frac{y}{x} - 1 = 0\\ \iff \qquad & m^2x^3 - mx^2y + n^2y^3 + nxy^3 - mnx^2y - mnxy^2 = 0.\end{aligned} \tag{2.58}$$



(a) Visualizing equality in (2.49). φ_j (red) crosses (2.60b) (blue) and (2.60a) (purple) out of the plot range.



(b) Depicting the same scenario, displaying the difference $\varphi_j - (2.60\text{b})$ (blue), and the same difference with (2.60a) (purple).

There exist two formal solutions to (2.58), given by

$$\begin{cases} y = \pm\sqrt{\dfrac{m}{n}}x \\ y = \dfrac{mx}{n+x}. \end{cases}$$

We therefore obtain conditions for equality to be achieved in (2.49)

$$\begin{cases} \varphi_j = \varphi_i + \dfrac{1}{2}\log\left[\dfrac{1+m}{m}\right] & \text{(2.60a)} \\ \varphi_j = \log\left[\dfrac{1+m}{m}e^{\varphi(i)} + \dfrac{1}{m}\right]. & \text{(2.60b)} \end{cases}$$

It is straightforward to observe that (2.60a) is strictly less than (2.60b), and thus the relevant condition for monotonicity of the operator $\hat{\mu}^1$ is that

$$\varphi_j(\hat{\mu}_i^1) < \log\left[\frac{1+m}{m}e^{\varphi_i(\hat{\mu}_i^1)} + \frac{1}{m}\right],$$

where $m = 1 + \gamma = 1 + (u_i - a)/(u_j - u_i)$.

2.4 On the Log-concavity of Belief in Logistic Models

Of central importance to the task of optimal learning in logistic models is the structure of belief states (i.e., posterior distributions) under Bayesian dynamics. Although a body of literature exists studying the asymptotic behavior of Bayesian orbits (i.e., sequences of Bayesian posterior distributions) under maximum likelihood control, considerably

less attention has been given to intermediate, non–equilibrium orbits and orbits generated by other control policies. In general, such orbits are *complex*, in the sense that their mathematical descriptions admit simplification only to a point; that is, the submanifold of Bayesian orbits under general control policies cannot be described more simply than by its outright computation. Contrast this circumstance for logistic models with that of a (Bayesian) conjugate model, wherein parameters of posterior distributions may be computed directly from known closed form expressions, opening the door to the operational calculus. The non–conjugacy of logistic models notwithstanding, the submanifold of Bayesian orbits is curved, its dynamics (in coordinates) thus generally nonlinear and non–convex, which is to say unwieldy.

In the face of these difficulties precluding a general characterization of Bayesian orbits for logistic models, one may grasp the issue from the other end by positing particular, desirable properties of the orbits and inquiring as to the form of logistic models exhibiting the properties. For example, in the application of clinical trial design considered in Chapter 5, it is natural to seek belief states that are *unimodal*. In a general sense, unimodality is desirable also because it guarantees global maximum likelihood estimators aiding convergent solution of the optimal learning problem, or at least its proof.

Establishing unimodality is generally challenging. The concepts of generalized concavity specifically *quasiconcavity*, extend the notion of unimodality to higher-dimensional spaces. Although characterization may still be difficult in many cases, we can more easily seek the more stringent requirement of a *strong unimodality* by establishing *log-concavity*. In a general sense, the concavity properties of a probability distribution materially describe the behavior of processes generated from it. This intuitive fact is doubly relevant when the process of interest is a controlled Markov process arising in an optimal learning context: The system costs recursively relate back to the underlying generator of the process, as we shall discuss in detail below. We first introduce the rudiments of the generalized concavity theory in the next Section 2.4.1.

2.4.1 Preliminaries

The *generalized concavity* theory is a generally useful basis from which to wield elements of convex analysis in probabilistic optimization. The theory is fundamentally built upon the *weighted means of order* α, authoritatively studied by Hardy, Littlewood, and Pólya in [71]:

$$m_\alpha(a,\lambda) = \left(\sum_{i=1}^{n} \lambda_i a_i^\alpha\right)^{1/\alpha},$$

where $\lambda \in \{\lambda \in \mathbb{R}^n | \lambda \geq 0,\ \sum_i \lambda_i = 1\}$ may be viewed as a vector of probabilities. For our purpose of developing generalized concavity of functions, we shall simply proceed with the case $n = 2$. In this case, the weighted mean of order α, or simply the *α-mean* m_α, is defined as follows.

Definition 19 (α-mean)**.** *For all $\alpha \in \overline{\mathbb{R}}$, $\lambda \in [0,1]$, and $a, b > 0$, as*

$$m_\alpha(a,b,\lambda) = \begin{cases} a^\lambda b^{1-\lambda}, & \text{if } \alpha = 0, \\ \max\{a,b\}, & \text{if } \alpha = \infty, \\ \min\{a,b\}, & \text{if } \alpha = -\infty, \\ (\lambda a^\alpha + (1-\lambda)b^\alpha)^{1/\alpha}, & \text{otherwise.} \end{cases} \tag{2.61}$$

Note that the familiar arithmetic, geometric, and harmonic means correspond to $\alpha = 1, 0$, and -1, respectively. As is well known, these means are fundamentally related via natural inequalities, establishing a natural hierarchy. The proverbial example in this hierarchy is perhaps the inequality of arithmetic and geometric means, better known simply as the *AM-GM inequality*.

One may build, in a rather straightforward way, a theory of concave functions on these means that inherits their natural hierarchy.

Definition 20 (α–concave function)**.** *A nonnegative function $f(x)$ defined on a convex set $\Omega \subset \mathbb{R}^n$ is* α-concave*, where $\alpha \in [-\infty, \infty]$, if for all $x, y \in \Omega$ and all $\lambda \in [0,1]$ we have*

$$f\big(\lambda x + (1-\lambda)y\big) \geq m_\alpha\big(f(x), f(y), \lambda\big). \tag{2.62}$$

Note that the familiar notions of concavity, log-concavity, and quasi–concavity correspond to $\alpha = 1, 0$, and $-\infty$, respectively. See [129] for a thorough treatment; cf. [107].

The role of the logarithm in the distinctive term for 0–concave functions is illustrative of the nature of the hierarchy alluded to above, and also deeply related to logistic models. Consider the AM-GM inequality, written in the vernacular of α–means (2.61) as $m_1(x, y, \lambda) \geq m_0(x, y, \lambda)$, or explicitly

$$\lambda x + (1 - \lambda)y \geq x^{\lambda} y^{(1-\lambda)}. \tag{2.63}$$

Recalling that the logarithm $x \mapsto \log(x)$ is monotonic for $x > 0$, its application will preserve the inequality. Observing also that $\log(x^{\lambda} y^{(1-\lambda)}) = \lambda \log(x) + (1 - \lambda) \log(y)$, we obtain that (2.63) is equivalent to

$$\log(\lambda x + (1 - \lambda)y) \geq \lambda \log(x) + (1 - \lambda) \log(y). \tag{2.64}$$

Moreover, in view of definition (2.62), we recognize (2.64) as a statement of the 1–concavity of the logarithm.

That is, the AM-GM inequality is equivalent to the concavity of the logarithm. By the same analysis, we see an arbitrary nonnegative function $f(\cdot)$ is 0–concave *if and only if* $\log f(\cdot)$ is a concave function. This is the origin of the term "log–concave" for 0–concave functions, but the implication here is also essential: Observe that if $f(\cdot)$ were 1–concave, then the 1–concavity of the logarithm implies $\log f(\cdot)$ is also 1–concave, and therefore $f(\cdot)$ is 0–concave. That is, the 1–concavity of a function implies its 0–concavity, or concavity of a function implies its log–concavity.

We have established this particular result because of its illustrative connections to the AM-GM inequality and the origin of the distinctive nomenclature "log–concave", but it can be shown that this hierarchy of α–concavity is a general property stemming from the natural ordering of α–means. We formalize this below without rigor, instead referring the reader to [129], p. 95, for a detailed proof.

Lemma 16 (Monotonicity of weighted means of order α)**.** *The mapping $\alpha \mapsto m_{\alpha}(a, b, \lambda)$ is nondecreasing and continuous.*

Proof. (See [129], p. 95.) □

Corollary 17 (Hierarchy of α-concave functions)**.** *Any α-concave function is also β-concave for all $\beta \leq \alpha$. In particular, α-concavity implies $(-\infty)$-concavity (quasiconcavity), for any α.*

Definition 21 (α-concavity of probability measures)**.** *A probability measure P defined on the Lebesgue measurable subsets of a convex set $\Omega \subset \mathbb{R}^n$ is α-concave, if for any Borel measurable sets $A, B \subset \Omega$ and for all $\lambda \in [0,1]$, we have*

$$P(\lambda A + (1-\lambda)B) \geq m_\alpha(P(A), P(B), \lambda), \tag{2.65}$$

where the sum is understood as the Minkowski sum.

Given a real-valued random vector Z, we say that Z has an α-concave distribution if the measure P_Z induced by Z is α-concave. We formalize this in the following lemma.

Lemma 18. *If a random vector Z induces an α-concave probability measure P_Z on $\mathbb{R}^n$, then the corresponding distribution function F_Z is an α-concave function.*

Proof. This follows directly from the definitions of F_Z and α-concavity of probability measures. Let $a, b \in \mathbb{R}^n$ be given. For any $\lambda \in [0,1]$, define $A := \{z \in \mathbb{R}^n \ : \ z \leq a\}$ and similarly $B := \{z \in \mathbb{R}^n \ : \ z \leq b\}$. Note that for all z in the set $\lambda A + (1-\lambda)B = \{z' \in \mathbb{R}^n \ : \ z' = \lambda a' + (1-\lambda)b', \ a' \in A, \ b' \in B\}$, it follows that $z \leq \lambda a + (1-\lambda)b$. Directly applying (2.65) and the definition of F_Z, we obtain $F_Z(\lambda a + (1-\lambda)b) \geq m_a(F_Z(a), F_Z(b), \lambda)$, thus satisfying definition 20. □

Definition 22 (α-concavity of discrete distributions)**.** *A distribution function F is called α-concave on the set $A \subset \mathbb{R}^n$, with $\alpha \in [-\infty, \infty]$, if*

$$F(z) = m_\alpha\left(F(x), F(y), \lambda\right)$$

for all $z, x, y \in A$, $\lambda \in (0,1)$ such that $z \geq \lambda x + (1-\lambda)y$.

For a given random vector Z, there naturally exist relations among the α–concavity of its induced probability measure, density, and distribution function. We refer the reader to [129, 107] for further details.

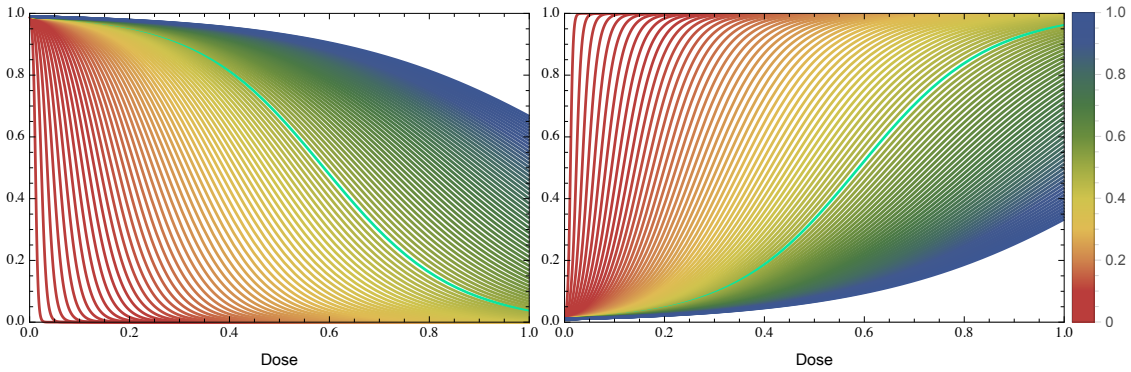


Figure 2.6: Example of a typical Type I model, $\{\Phi^0_\eta(u)$ (left),$\Phi^1_\eta(u)$ (right)$\}$, defined by: $\Phi^1_\eta(0) = p_\epsilon$, $\Phi^1_\eta(\eta) = \nu$. The domain $\mathcal{U}$; color spectrum denotes $\eta \in \mathcal{N}$.

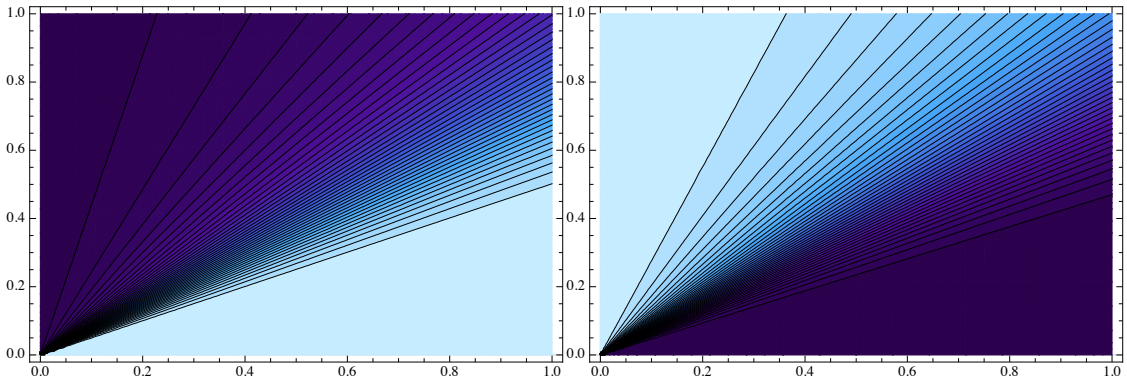


Figure 2.7: The obvious nonlinearity throughout belies a linear relation between the dose u and ν-quantile parameter η, as revealed by the contour plots.

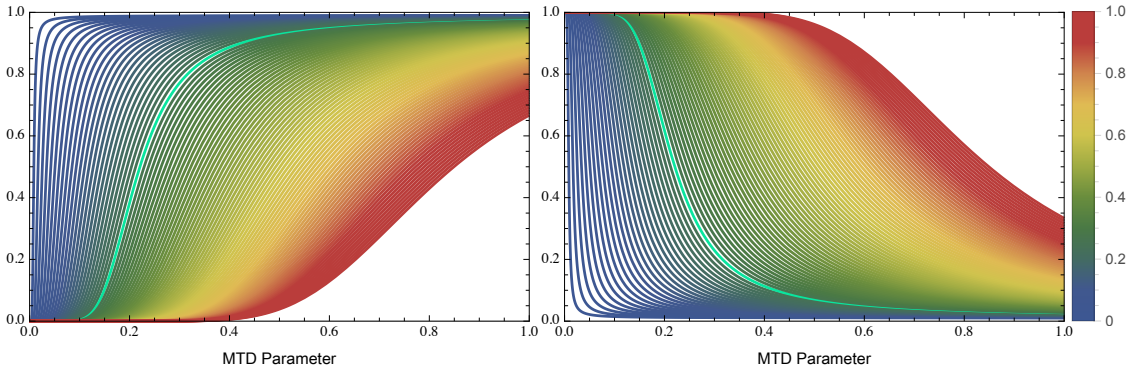


Figure 2.8: The same Type I model, over $\eta \in \mathcal{N}$. Color spectrum denotes $u \in \mathcal{U}$.

2.4.2 General Log-concavity

We emphasize a proof of the following result utilizing the log-concavity of the induced probability measure.

Theorem 19. *If for each $\eta \in \mathcal{N}$ the toxicity random variable $\tau \in \mathcal{U}$ has a log-concave density function ϕ of the form $\phi(\tau;\eta) = \phi(\tau - \eta)$, then the sequence $\{s = \Psi(s_t, u_t, r_t)\}_{t=1}^{T}$ of Bayesian posterior density functions is log-concave.*

Proof. For each $\eta \in \mathcal{N}$, let $\phi(\tau;\eta)$ be a log-concave probability density function of the toxicity random variable $\tau \in \mathcal{U}$, and let ϕ be given by

$$\phi(\tau;\eta) = \phi(\tau - \eta) \tag{2.66}$$

for all $\eta \in \mathcal{N}, \tau \in \mathcal{U}$. Therefore, for each $\eta \in \mathcal{N}$, ϕ induces a log-concave probability measure P_η on the Lebesgue measurable subsets of $\mathcal{U}$, and we have

$$P_\eta\{\tau \leq u\} = \int_{-\infty}^{u} \phi(\tau - \eta)\, d\tau.$$

We denote the distribution function $\Phi(u;\eta) := P_\eta\{\tau \leq u\}$, for all $\eta \in \mathcal{N}$ and $\tau \in \mathcal{U}$. On the other hand, applying the change of variables $\xi = u + \eta - \tau$, we see

$$\int_{-\infty}^{u} \phi(\tau - \eta)\, d\tau = \int_{\eta}^{\infty} \phi(u - \xi)\, d\xi, \tag{2.67}$$

for all $\eta \in \mathcal{N}$ and $u \in \mathcal{U}$.

Fixing $u \in \mathcal{U}$, we see that ϕ also induces a probability measure P_u on the Lebesgue measurable subsets of N that similarly inherits log-concavity. Indeed, defining the sets

$$A_\eta = [\eta, \infty],$$
$$B_\eta = [-\infty, \eta],$$

for all $\eta \in \mathcal{N}$, we obtain directly from (2.65) that for all fixed $u \in \mathcal{U}$

$$\Phi(u;\eta) = P_u\big(A_\eta\big) \tag{2.68}$$

$$1 - \Phi(u;\eta) = P_u\big(B_\eta\big) \tag{2.69}$$

are each a log-concave function of η.

We now establish log-concavity of the belief states defined as posterior probability density functions under a Bayesian update. Let $t \in \{1, 2, \cdots, T\}$ be fixed, and suppose the probability density function $s_t(\eta)$ is log-concave. After applying dose $u_t \in \mathcal{U}$ and observing response $r_t \in \{0, 1\}$, we obtain the posterior density $s_{t+1} = \Psi(s_t, u_t, r_t)$. Explicitly, we have

$$s_{t+1}(\eta) = \begin{cases} s_t(\eta)\big(1 - \Phi(u_t, \eta)\big)/Z_0(u_t; s_t), & r_t = 0, \\ s_t(\eta)\Phi(u_t, \eta)/Z_1(u_t; s_t), & r_t = 1, \end{cases} \tag{2.70}$$

where $Z_0, Z_1 \in \mathbb{R}$ are normalizing operators.

By (2.68) and (2.69), we see that in either case s_{t+1} is proportional to the product of two log-concave functions, and is therefore log-concave. Indeed, without loss of generality let $r_t = 1$ and observe that

$$\begin{aligned} \log s_{t+1} &= \log[s_t(\eta)\Phi(u_t, \eta)/Z_1(u_t; s_t)] \\ &= Z_1(u_t; s_t)^{-1}\big(\log s_t(\eta) + \log \Phi(u_t, \eta)\big) \end{aligned}$$

is a concave function of η by the log-concavity of s_t and $\Phi(u_t, \cdot)$. Thus, s_{t+1} is a log-concave function of η.

By supposition, the *a priori* density s_1 is log-concave, and thus by induction we obtain that the sequence

$$\Big\{ s = \Psi(s_t, u_t, r_t) \Big\}_{t=1}^{T}$$

is log-concave. □

2.4.3 Log-concavity of Logistic Quantile Parameterizations

Theorem 20. *Let S be a Type I logistic model over $\mathcal{R} = \{0, 1\}$, given by*

$$S_L = \left\{ \Phi(r; u, \eta) = \frac{1}{1 + e^{-\varphi(r;u,n)}},\ (u, \eta) \in \mathcal{U} \times \mathcal{N} \right\}, \tag{2.71}$$

where the parameter η is defined by the ν-quantile relation $\Phi(1; \eta, \eta) \equiv \nu$, $\nu \in (0, 1)$. Suppose the discriminant $\varphi_r := \varphi(r; \cdot, \cdot)$ is a bilinear mapping for each $r \in \mathcal{R}$.

Then S is necessarily a translation family of models; that is, we have the form

$$\varphi_1(u, \eta) = \alpha(u - \eta) + \beta, \tag{2.72}$$

together with $\varphi_0(u, \eta) = -\varphi_1(u, \eta)$, for constants $\alpha, \beta \in \mathbb{R}_{>0}$.

Proof. Beginning with the fact that φ_r is bilinear for all $r \in \mathcal{R}$, for some real scalars $\alpha, \beta, \gamma, \delta$, we can write φ_1 in the form

$$\varphi_1(u, \eta) = \alpha u + \gamma u\eta + \delta\eta + \beta. \tag{2.73}$$

Enforcing the fact that $\Phi(1; \eta, \eta) = \nu$, for all $\eta \in \mathcal{N}$, we see that there exists a constant $\tilde{\nu}$ such that $\varphi_1(\eta, \eta) = \tilde{\nu}$. That is,

$$\alpha\eta + \gamma\eta^2 + \delta\eta + \beta = \tilde{\nu}, \quad \text{for all } \eta \in \mathcal{N},$$

and differentiating immediately yields $\delta = -\alpha$, $\gamma = 0$. Hence φ_1 has the form

$$\begin{aligned} \varphi_1(u, \eta) &= \alpha u - \alpha\eta + \beta \\ &= \alpha(u - \eta) + \beta, \end{aligned}$$

as claimed. Moreover, S is a statistical model over $\mathcal{R}$, and so $1 = \Phi(1; u, \eta) + \Phi(0; u, \eta)$. Together with the form of the logistic model shown in (2.71), this implies $\varphi_0 = -\varphi_1$, which is to say α, β are constants not depending on r. Thus, S is a translation family. □

The log-concavity of this class of models leads to corresponding result for Bayesian posterior sequences.

Theorem 21. *Let s_1 be any log-concave prior density and consider the logistic model S_L given in* (2.71)*. Then for any dose-response sequences $(u, r) \in (\mathcal{U} \times \mathcal{R})^T$ and any $T > 0$, the sequence*

$$\left\{ s = \Psi(r_t; s_t, u_t) \right\}_{t=1}^{T}$$

of posterior densities is log-concave.

Proof. Recalling the the simplified notation from (2.11), (2.12) on page 36, we proceed without the explicit notation of r. We shall use the form (2.73) of the discriminant $\varphi(u,\eta)$ to establish the log-concavity. Specifically, we have that $\varphi(u,\eta) = \alpha(u-\eta)+\beta$, for some constants $\alpha, \beta \in \mathbb{R}_{>0}$.

Writing $\log \Phi(u,\eta) = -\log[1 + e^{-(\alpha(u-\eta)+\beta)}]$, we see that a sufficient condition for Φ to be log-concave is that the map $\eta \mapsto 1 + e^{-(\alpha(u-\eta)+\beta)}$ be log-convex for all $u \in \mathcal{U}$.

We now move to employ the fact that the sum of (sufficiently regular) log-convex functions is itself a log-convex function. Formally, let $\xi(\eta)$ denote the map $\eta \mapsto 1 + e^{-(\alpha(u-\eta)+\beta)}$, and observe that may be decomposed as $\xi(\eta) = f(\eta)+g(\eta)$, where $f(\eta) \equiv 1$, and $g(\eta) = e^{-(\alpha(u-\eta)+\beta)}$. It is straightforward to see that both f and g are log-convex. Indeed,

$$\begin{aligned} \log f(\eta) &\equiv \log[1] = 0, \\ \log g(\eta) &= -(\alpha(u-\eta)+\beta), \quad \forall\, u \in \mathcal{U}, \end{aligned}$$

and we see that both functions are continuous and linear in η and therefore convex. If we apply Hölder's inequality to Definition 20, we can show that the sum $\tilde{f} + \tilde{g}$ remains log–convex. To this end, letting $\eta_\lambda = \lambda\eta_1 + (1-\lambda)\eta_2$ for arbitrary points $\eta_1, \eta_2 \in \mathcal{N}$ and $\lambda \in (0,1)$, we have

$$\begin{aligned} f(\eta_\lambda) + g(\eta_\lambda) &\leq f^\lambda(\eta_1) f^{1-\lambda}(\eta_2) + g^\lambda(\eta_1) g^{1-\lambda}(\eta_2) \\ &\leq \big(f(\eta_1) + g(\eta_1)\big)^\lambda \big(f(\eta_2) + g(\eta_2)\big)^{1-\lambda}, \end{aligned}$$

where, again, the final inequality follows by Hölder's inequality. Thus, $\xi = f + g$ is log-convex. Therefore, $\eta \mapsto \Phi(u,\eta)$ is log-concave for all $u \in \mathcal{U}$.

The same result can be shown analogously for the complementary map $\eta \mapsto 1 - \Phi(u,\eta)$. First, observe that the following representation holds:

$$\begin{aligned} \log[1 - \Phi(u,\eta)] &= \log\left[\frac{e^{-(\alpha(u-\eta)+\beta)}}{1 + e^{-(\alpha(u-\eta)+\beta)}}\right] \\ &= \log\left[\frac{1}{1 + e^{\alpha(u-\eta)+\beta}}\right]. \end{aligned}$$

Therefore, we require only the modified decomposition $\xi(\eta) = f(\eta) + 1/g(\eta)$. Since $\log[1/g(\eta)] = -\log g(\eta)$ is linear, it remains convex, and we can see that the argument

proceeds exactly as before. It therefore follows that $\eta \mapsto 1 - \Phi(u, \eta)$ is log-concave for all $u \in \mathcal{U}$.

Finally, consider the Bayesian orbits beginning at $s_1(\eta)$. We proceed by induction. Let the stage t be given, and assume a log-concave density s_t is given. Then, for any $(u_t, r_t) \in \mathcal{U} \times \{0, 1\}$, we have the posterior density $s_{t+1} := \Psi(r_t; s_t, u_t)$ given explicitly by

$$s_{t+1}(\eta) = \begin{cases} s_t(\eta)\big(1 - \Phi(u_t, \eta)\big)/Z(0; s_t, u_t), & r_t = 0, \\ s_t(\eta)\Phi(u_t, \eta)/Z(1; s_t, u_t), & r_t = 1, \end{cases} \tag{2.74}$$

where Z is the normalization operator, to be defined formally in (2.15) below.

By assumption s_t is log-concave, and by the preceding arguments both $\Phi(u_t, \eta)$ and $1 - \Phi(u_t, \eta)$ are log-concave functions of η, for all $u_t \in \mathcal{U}$. Thus, for any $(u_t, r_t) \in \mathcal{U} \times \{0, 1\}$, s_{t+1} is proportional to the product of log-concave functions and therefore itself log-concave.

By supposition, the *a priori* density s_1 is log-concave, and thus by induction we obtain that the sequence

$$\Big\{ s = \Psi(r_t; u_t, s_t) \Big\}_{t=1}^{T}$$

is log-concave. □

We conclude with a simple alternative proof of the special case via differentiation:

Theorem 22. *The logistic function and its complement in* (2.4) *are simultaneously log-concave in* η *if and only if for all* $u \in \mathcal{U}$ *and* $\eta \in \mathcal{N}$*, the discriminant function* $\varphi(u, \eta)$ *satisfies*

$$\Phi - 1 < \frac{\varphi_{\eta\eta}}{\varphi_\eta^2} < \Phi, \quad \varphi_\eta \neq 0. \tag{2.75}$$

Proof. Differentiating twice with respect to η, we obtain

$$\partial_\eta^2 \log f(\varphi) = \varphi_{\eta\eta} f(-\varphi) - \varphi_\eta^2 f(\varphi) f(-\varphi),$$
$$\partial_\eta^2 \log f(-\varphi) = -\varphi_{\eta\eta} f(\varphi) - \varphi_\eta^2 f(\varphi) f(-\varphi).$$

Imposing the condition of log-concavity, the system becomes

$$\varphi_{\eta\eta} f(-\varphi) - \varphi_\eta^2 f(\varphi) f(-\varphi) < 0,$$
$$\varphi_{\eta\eta} f(\varphi) + \varphi_\eta^2 f(\varphi) f(-\varphi) > 0,$$

implying $\varphi_{\eta\eta} - f(\varphi)\varphi_\eta^2 < 0 < \varphi_{\eta\eta} + f(-\varphi)\varphi_\eta^2$. Provided $\varphi_\eta \neq 0$, we have

$$-f(-\varphi) < \frac{\varphi_{\eta\eta}}{\varphi_\eta^2} < f(\varphi),$$

and the more intelligible form in (2.75) follows by (2.4).

□

Corollary 23 (Linear Discriminant)**.** *For all $u \in \mathcal{U}$, let $\varphi(u, \eta)$ be affine in η. Then both $\Phi(u, \cdot)$ and $1 - \Phi(u, \cdot)$ are log-concave.*

Proof. The fact that φ is affine in η implies $\varphi_{\eta\eta} \equiv 0$, and (2.75) becomes

$$\Phi(u, \eta) - 1 < 0 < \Phi(u, \eta), \tag{2.76}$$

which clearly holds for all $(u, \eta) \in \mathcal{U} \times (-\infty, \infty)$. □

Chapter 3

Dynamic Risk and Sequential Inference

> *We know what we are, but know not what we may be.*

William Shakespeare, *1603*

> *Man cannot remake himself without suffering, for he is both the marble and the sculptor.*

Alexis Carrel, *Man, the unknown*, 1935

> *The understanding which we want is an understanding of an insistent present. The only use of a knowledge of the past is to equip us for the present... The present contains all that there is. It is holy ground; for it is the past, and it is the future.*

Albert North Whitehead, *1916*

As discussed above, the crux of the optimal learning problem is the fundamental self–reference resulting in a certain analytical recursion. To put it most simply, our belief is determined by our observations, which themselves are determined in part by our actions, yet which in turn are chosen according to our belief. The issues engendered by this recursion are intuitively familiar. Indeed, the colloquial notions of *self–fulfilling prophecy*, *confirmation bias*, *sunk-cost fallacy*, and *Russell conjugation*, among others, are rooted in this recursion. Against this backdrop, the particular optimal learning

problem we consider below amounts to optimally balancing the simultaneous (risk–adjusted) costs and benefits of learning, not unlike masterfully navigating stormy waters fraught with perilous waves—waves of information amid storms of the spurious.

Another central issue related to this fundamental recursion is well–known in the machine learning and multi–armed bandit (MAB) literature as the *exploration vs. exploitation* (EE) tradeoff, or yet more plainly as *learn vs. earn*. This tradeoff describes the simple phenomenon that the action which maximizes the profitable information we learn from its corresponding observation is generally different from the action with minimal (risk–adjusted) cost. To grasp this at a glance, observe simply that one cannot learn that which one believes. More generally, the EE tradeoff intimates the decomposition of the tradespace into the explicit cost and the implicit reward of learning, as extruded through action. Put simply, the EE tradeoff names the investment in learning vs. its opportunity cost.

The MAB boasts a rich literature, perhaps owing to the broad applicability of the problem class. The basic results surrounding the existence and form of dynamic allocation indices (DAI) is famously due to Gittins [66]. These so–called Gittins indices are notoriously difficult to compute in many settings. As Whittle humorously put it, the MAB problem "was formulated during the war, and efforts to solve it so sapped the energies and minds of Allied analysts that the suggestion was made that the problem be dropped over Germany, as the ultimate instrument of intellectual sabotage" [147]. Various reformulations of the indices have been made to render said indices more tractable, a seminal such decomposition method due to Katehakis and Veinott [84] has been studied in the literature. Other MAB formulations in terms of minimizing a measure of cumulative regret have seen extensive study, cf. [31], [83]. Recently, an information theoretic perspective was studied by Russo [120]. For a comprehensive treatment of MAB problems, see [67], and the references therein. Much of the MAB literature is concerned with aspects of the asymptotic optimality in problems with general information structures, whereas in our setting, we will be focused on relatively short time horizons and a particular information structure.

One fundamental issue emerging as most critical in our endeavor is quantifying the

value of learning, which is intrinsically a dynamic property. Foundational work in this vein is the subject of information theory, and the seminal notion of entropy was famously introduced in this context by Shannon in [127]. The dynamic counterpart to entropy is *relative entropy*, which has emerged as a popular method to quantify learning. Relative entropy is built upon the divergence functional, introduced in step by Csiszàr [38] and Bregman [29]; see also [63].

Recently, the fundamental duality running as a thread between and among relatively siloed branches of knowledge have been pointed out and observed more widely. For example, the branches of competitive games (i.e., game theory), Bayesian inference, machine learning, information theory, differential geometry, quantum mechanics, stochastic optimization, dynamic risk, etc., all share connections along these lines, as is perhaps well-known to the expert—and maddening to the student. For some recent work explicitly elucidating the duality connections among competitive games, the maximum entropy principle, and Bayesian inference, see [69]; for similar considerations with an emphasis on (Bayesian) machine learning, see the work by Reid and Williamson [111, 56], and the references therein. The dualistic connections between divergences as informations measures and stochastic optimization problems were studied extensively by Ben-Tal et al.; see [21], inter alia.

By way of, at bottom, the natural homomorphism between multiplication and addition, exponential distribution families have proved both interesting and tractable, and thus served as an indispensable bridge connecting the probability theory and its many subsidiaries to disciplines relatively more developed with respect to dynamics, the canonical examples of which being classical and quantum mechanics. Exponential families were studied authoritatively by Barndorff–Neilsen in the context of statistical inference, and paved the way for geometric considerations with the notion of the statistical manifold. This led to the introduction of dual connections and the general clarification of the role of divergences within the emerging subdiscipline of information geometry, largely due to Amari; see [6, 5] and the references therein.

3.1 Risk–sensitive MDP Belief Dynamics

The Markov property is perhaps the most fundamental premise underlying the theory of MDPs, and in Section 1.2 we recalled the classical theory. In Section 1.3 we introduced the fundamentals of dynamic risk, and in Section 1.3.4 we introduced risk–aware optimization. In this section, we formulate belief dynamics, i.e., learning, as a Markov process, whereby the optimal learning problem becomes an MDP. In particular, we introduce belief states and establish the existence of a stochastic kernel coinciding with the conditional state transition probability, thus demonstrating the essential Markov belief dynamics. The operator responsible for these Markov transitions may be formally constructed as a consequence of Bayes' Theorem, and hence we arrive at the term *Bayesian belief dynamics*. Additionally, we introduce dynamic risk in the optimal learning problem and investigate the prospects of the class of conditional risk measures that are composite in stochasticity and uncertainty.

For the purpose of balancing clear exposition and relevance to applications, we focus on a prototypical problem of logistic regression or, complementarily, classification. In particular, we shall study this problem in the context of optimal *clinical trial design* (CTD), which is an important motivation of our work. Specifically, we study the logistic regression problem of determining the optimal dosage policy for patients in a clinical trial of a novel pharmaceutical agent; equivalently, this may be viewed as a binary classification problem of dosage as either clinically toxic or non–toxic. Note that we study the formal problem in this context to facilitate acquisition for the uninitiated clinician and to foster intuition with respect to the general analysis. Thus, in what follows we introduce a notation with only mild limitations, although in examples and further discussion we shall focus on the case of binary classification with logistic regression.

3.1.1 Problem Formulation

We consider the challenge faced by a decision–maker[1] to choose actions u_t from the set $\mathcal{U}$ at sequential times[2] $t \in \{1, 2, \cdots, T-1, T\}$ over the finite time horizon $T \in \mathbb{N}$. At each time t, after taking action $u_t \in \mathcal{U}$, the decision–maker observes the response $r_t \in \mathcal{R}$. We call $\mathcal{R}$ the *response space*, and in general we assume only that $\mathcal{R}$ is a subset of a finite–dimensional Euclidean space; in what follows, we shall be content to further assume it to be a finite set, and we study in particular the case $\mathcal{R} = \{0, 1\}$.

For each action u, the response r occurs randomly according to the conditional probability distribution $\Phi^* = \Phi^*(r; u)$ on $\mathcal{R}$. However, this *true* (or *underlying*) controlled response distribution is itself unknown, but is assumed to reside in a statistical manifold S with mixed coordinate system $\xi := (u, \eta) \in \Xi = \mathcal{U} \times \mathcal{N} \subset \mathbb{R}^n$, for some finite $n \in \mathbb{N}$. We view ξ as a parameter, and the parameter space $\mathcal{N} \subset \mathbb{R}^k$, $k \in \mathbb{N}$, for $k \leq 2$. Moreover, we view u as the control parameter, whereas η is an unknown parameter. In this context, we thus consider the statistical model $S = \{\Phi_\xi(r) = \Phi(r; u, \eta) \mid (u, \eta) \in (\mathcal{U} \times \mathcal{N}) \subset \mathbb{R}^n\}$, with $\Phi^*_u(r) = \Phi(r; u, \eta^*) \in S$ for unknown, unique $\eta^* \in \mathcal{N}$.

We consider the regular probability space $(\Omega, \mathcal{F}, \mathcal{P})$, together with the filtration $\{\emptyset, \Omega\} = \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_{T-1} \subset \mathcal{F}_T = \mathcal{F}$. For each t, $\mathcal{F}_t$ denotes the canonical σ–algebra generated by the observed control–response sequence
$(u_i, r_i)^t = (u_1, r_1, u_2, r_2, \cdots, u_{t-1}, r_{t-1})$.

3.1.2 Markovian Dynamics and the Bayes Operator

The formulation of a Markovian system via the introduction of belief states together with learning dynamics is an intuitively appealing approach, and now a classical technique. In order to develop a tenable framework, "belief" and "learning" need to be rigorously defined. The preeminent approach is to understand belief probabilistically, that is, where one's current state of belief about a given system (or proposition) is understood as a probability distribution over an appropriate space. Viewing this as a

[1] or *agent*, *clinician*, etc.

[2] or, equivalently, *epochs*, *stages*, etc.

single entity, we call this characterization one's *belief state*. In this context, *learning*[3] is revealed as the *dynamics of belief*, the mechanism of mapping one's current belief state to a new belief state via new information.

Undoubtedly, the prime benefit of this framework is its inheritance of the rich, well–established probability theory. In the context of learning, as defined above, the definition of the conditional probability distribution is the principal actor. The central identity is essentially the result in Bayes' Theorem, which is the origin of the widely used term Bayesian inference (learning). The general theory of conditional probability distributions, under the mild condition of regularity, establishes that joint distributions may be resolved into the product of a transition kernel over the product space and marginal distribution. Bayes theorem then provides a method for obtaining a new, posterior marginal distribution given a prior marginal and a conditional observation.

Thus, given any a priori knowledge of the system, one may begin by positing a certain marginal distribution over the unknown parameter space. Even without a priori knowledge, one may still simply posit the existence of a non–vanishing marginal distribution, e.g., Jeffrey's prior, uniform distribution, etc., over the parameter space $\mathcal{N}$ and define the state space $\mathcal{P}(\mathcal{N})$ to be probability distributions on $\mathcal{N}$, from which an MDP may be formulated. Generally the latter construction impacts only the rate of, rather than the fact of, convergence in probability. However, we must mention that this degree of freedom has historically been a point of contention in the literature, the alternative to this "Bayesian" approach being the so–called "frequentist" approach. We refer the reader to the statistics literature for details on this debate.

We begin with the fundamentals central to clarifying these concepts. First, we recall the underlying statistical model $S(\mathcal{U}) = \{P(\tau;\xi)|\xi \in \Xi\}$ for the unobserved random variable $\tau \in \mathcal{U}$. For all $\xi \in \Xi$ and all Borel sets $A \in \mathcal{B}(\mathcal{U})$, we have $\mathbb{P}(\tau \in A|\xi) = \int_A P(\tau;\xi)d\tau$.[4] In our problem setting, the decision–maker is forced to operate within the *observed model* $S(\mathcal{R})$, whereby the observation mapping $\mathcal{H}_\xi : \tau \mapsto r$ transforms

[3]One might equivalently use the term *inference*, as in statistics vis–à–vis "statistical inference." The term learning aligns more with computer science vis–à–vis "machine learning."

[4]Assume henceforth that all equalities hold identically for all $\xi \in \Xi$, unless specified otherwise.

the underlying random variable $\tau \in \mathcal{U}$ into the observed random variable $r \in \mathcal{R}$. The distribution density $P(\tau;\xi)$ thus naturally induces the distribution $Q(r;\xi)$ of r on $\mathcal{B}(\mathcal{R})$, the Borel sets of (the image of $\mathcal{H}_\xi$ over $\mathcal{U}$ in) $\mathcal{R}$.

We assume below that $\mathcal{H}_\xi$ is surjective, so that for all $r \in \mathcal{R}$, $\mathcal{H}_\xi^{-1}(r) \subset \mathcal{U}$ and well–defined, where in particular $\mathcal{H}_\xi(\mathcal{U}) \cap \mathcal{R} = \mathcal{H}_\xi(\mathcal{U}) = \mathcal{R}$. Moreover, note that $\mathcal{H}_\xi$ is generally not one–to–one, however, we will be content to assume it is deterministic. That is, for any $\tau, \tau' \in \mathcal{U}$, $\tau = \tau'$ implies $\mathcal{H}_\xi(\tau) = \mathcal{H}_\xi(\tau')$. More formally, if we let $K(r|\tau;\xi)$ denote the regular conditional probability distribution (i.e., transition kernel) of r given τ, then we consider the case $K(r|\tau;\xi) = \delta_{\mathcal{H}_\xi(\tau)}(r)$, the Dirac delta.[5] centered at $\mathcal{H}_\xi(\tau)$

In deriving the observed model we shall preview the central techniques used below. We proceed by computing the induced distribution $Q(r;\xi) = \mathbb{P}(r|\xi)$. By definition, we have

$$\begin{aligned} Q(r;\xi) &= \int_{\mathcal{U}} \mathbb{P}\big(\tau, r|\xi\big)\, d\tau \\ &= \int_{\mathcal{U}} K(r|\tau;\xi) P(\tau;\xi)\, d\tau \\ &= \int_{\mathcal{U}} \delta\big(r - \mathcal{H}_\xi(\tau)\big) P(\tau;\xi)\, d\tau \\ &= \int_{A_\xi(r)} P(\tau;\xi)\, d\tau, \end{aligned}$$

where the last equality follows by the surjectivity of $\mathcal{H}_\xi$ and the properties of the Dirac δ distribution. Note that we introduce the set $A_\xi(r) = \mathcal{H}_\xi^{-1}(r) \subset \mathcal{U}$.

This general technique is thematic and will feature again below, but a few points need to be clarified for our problem setting. In particular, we consider a mixed coordinate system, in that $\xi = (u, \eta) \in \mathcal{U} \times \mathcal{N}$, and moreover

$$\mathcal{H}(\tau;\xi) \equiv \mathcal{H}(\tau;u),$$
$$\text{and } P(\tau;\xi) \equiv P(\tau;\eta).$$

Thus, letting $A_\xi(r) = A_u(r)$, we can more explicitly write $Q(r;\xi) \equiv \Phi(r;u,\eta)$ as

$$\Phi(r;u,\eta) = \int_{A_u(r)} P(\tau;\eta)\, d\tau. \tag{3.1}$$

[5]When $\mathcal{R}$ is discrete, this is understood as the Kronecker delta.

We now demonstrate the central role of conditional probability in this context. Applying the disintegration theorem to the joint probability $\mathbb{P}(r, \eta|u)$ in each of the first two arguments, respectively, we obtain the identity

$$\mathbb{P}(\eta|r, u)\mathbb{P}(r|u) = \mathbb{P}(r|u, \eta)\mathbb{P}(\eta|u). \tag{3.2}$$

Immediately, we make the following observations about the RHS, $\mathbb{P}(r|u, \eta)\mathbb{P}(\eta|u)$. First, $\mathbb{P}(r|u, \eta) = \Phi(r; u, \eta)$. Second, the mixed coordinates are independent, and therefore $\mathbb{P}(\eta|u) = \mathbb{P}(\eta)$. Let $s(\eta) \triangleq \mathbb{P}(\eta)$.

Turning now to the LHS, $\mathbb{P}(\eta|r, u)\mathbb{P}(r|u)$, we observe, by the same technique, that $\mathbb{P}(r|u) = \int_{\mathcal{N}} \mathbb{P}(r|u, \eta)\mathbb{P}(\eta|u)\, d\eta = \mathbb{E}_s[\Phi(r; u, \eta)]$. Letting $Z(r; u) \equiv \mathbb{P}(r|u)$, we have

$$Z(r; u) = \int_{\mathcal{N}} \Phi(r; u, \eta)s(\eta)\, d\eta. \tag{3.3}$$

Finally, letting $s_+(\eta; r, u) \equiv \mathbb{P}(\eta|r, u)$ and rearranging (3.2), we arrive at the conclusion of Bayes' theorem:

$$s_+(r; u, \eta) = \frac{\Phi(r; u, \eta)s(\eta)}{Z(r; u)}. \tag{3.4}$$

In light of the form of (3.3), the significance of (3.4) should be understood with respect to a fixed statistical model $\{\Phi(r; u, \eta)\}$, as follows: Given a belief state $s(\eta)$ and a control-response observation (u, r), the quantity in (3.4) characterizes a new belief state $s_+(\eta)$ incorporating the new information. Within a fixed model, it is *essential* that the belief state s_+ be determined only from the previous belief state s and the new observation (u, r). The initial belief state s is called the *prior* (or, a priori) distribution, and a realization of s_+ is called the *posterior*. This fact is paramount, and we reiterate it in remark 4.

Remark 4. *Sequential Bayesian inference is Markovian.*

The MDP framework is, obviously, built upon the Markov property; that is, the dynamics depend only on the current state and control. In order to recast the above precisely in this formalism, we recall the definition of the *Bayes operator*.

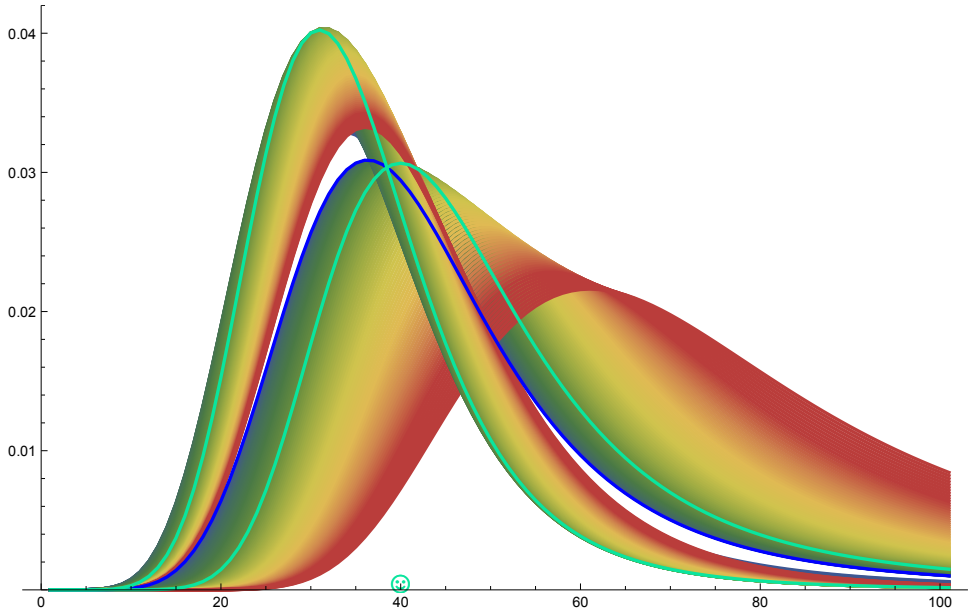


Figure 3.1: Visualizing Bayesian dynamics. Starting from the current belief state $s = s(\eta)$ (center, dark blue), all posterior states $\{s_+\}$ from (3.4) are shown, for each response $r = 0$ (moved right), $r = 1$ (moved left), and control u (spectrum). Also shown is the unknown optimal parameter η^* (☺), with its corresponding posteriors (bright green).

Definition 17 (Bayes operator)**.** *Let $s \in \mathcal{P}(\mathcal{N})$ be an arbitrary probability density on $\mathcal{N}$. By the* Bayes operator *we understand the operator $\Psi : \mathcal{R} \times \mathcal{U} \times \mathcal{P}(\mathcal{N}) \to \mathcal{P}(\mathcal{N})$ given by*

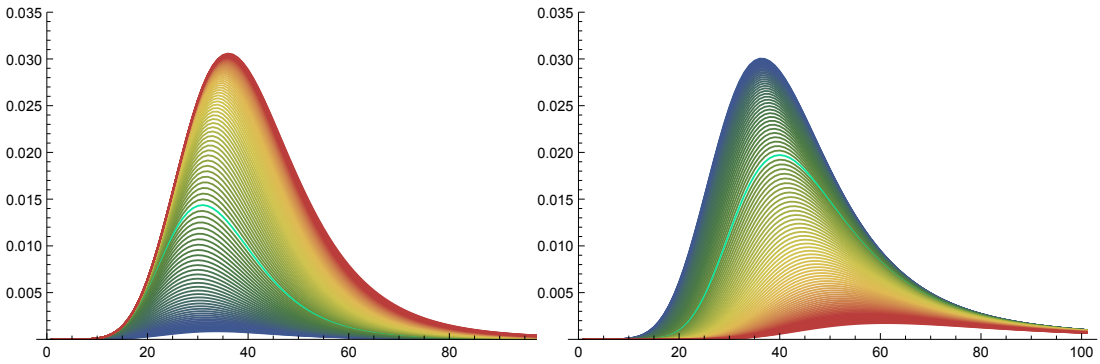
$$\Psi(r; s, u)(\cdot) \equiv \frac{\Phi(r; u, \cdot)s(\cdot)}{\int_{\mathcal{N}} \Phi(r; u, \eta)s(\eta)} \equiv \frac{\Phi(r; u, \cdot)s}{\langle \Phi(r; u, \cdot), s \rangle}. \tag{2.14}$$

The Bayes operator Ψ should be understood with respect to a fixed response r_t, as follows. Given the current state s_t and control u_t, the system transitions to the next state s_{t+1} according to

$$s_{t+1} = \Psi(r_t; s_t, u_t), \tag{3.5}$$

as given in definition 17 above and again in (3.4). We call Ψ_{r_t} the *state transition mapping*. Note that it is notationally convenient to write $\Psi_t(r_t) \triangleq \Psi(r_t; s_t, u_t)$, and we use this shorthand below.

Together with the state transition mapping, the MDP framework is centered around the Markov transition kernel, which is defined for all $C \in \mathcal{B}(\mathcal{S})$ as $\mathbb{P}(s_{t+1} \in C | s_t, u_t)$.



(a) Toxic response ($r = 1$) (b) Non-toxic response ($r = 0$)

Figure 3.2: All denormalized posteriors $\Phi(u;\eta)s(\eta)$. Spectrum denotes $u \in \mathcal{U}$.

By again employing the now familiar disintegration technique, we can write down the transition kernel explicitly. Formally, we have

$$\mathbb{P}(s_{t+1} \in C | s_t, u_t) = \sum_{r \in \mathcal{R}} \mathbf{1}_C\big[\mathbb{P}(s_{t+1}|r_t, s_t, u_t)\big]\mathbb{P}(r_t|s_t, u_t),$$

where $\mathbf{1}_C[\cdot]$ denotes the set indicator function of C. We recognize that the state transition mapping (3.5) together with the *observation kernel* $Z(r_t; s_t, u_t) \triangleq \mathbb{P}(r_t|s_t, u_t)$ is equivalent to (3.3), therefore define the *state transition kernel* $Q(C|s_t, u_t) \triangleq \mathbb{P}(s_{t+1} \in C|s_t, u_t)$ by

$$Q(C|s_t, u_t) = \sum_{r \in \mathcal{R}} \mathbf{1}_C\big[\Psi(r_t; s_t, u_t)\big] Z(r_t; s_t, u_t). \tag{3.6}$$

Note that Q depends only on the current state-control pair (s_t, u_t), and is therefore Markov. Moreover, it is also *stationary*, that is, Q is the same at each stage t.

It is convenient to identify for each stage t the σ-algebra $\mathcal{F}_t$ generated by the *information sequences*

$$(s_1, u_1, s_2, \cdots, u_t, s_{t+1}) = (s_1, u_1, \Psi_1(r_1), \cdots, u_t, \Psi_t(r_{t+1})).$$

Finally, we note that for any state control pair (s_t, u_t), the Bayes operator $\Psi_t(r_t)$ is injective, and clearly not surjective. Any injective mapping of a random variable trivially defines a sufficient statistic, and therefore it is formally correct to regard $s_{t+1} = \Psi_t(r_t)$ as a sufficient statistic for the statistical model S over $\mathcal{R}$, given $\mathcal{F}_t$. Of course, in our problem setting with finite response space $\mathcal{R}$, it is decidedly more profitable

to utilize this sufficiency in the converse. We shall take this route below to avoid the formality of writing such integrals as $\int_{\mathcal{S}} Q(s'|s_t, u_t)ds'$ over the infinite space $\mathcal{S} = \mathcal{P}(\mathcal{N})$, so that the equivalent expression $\sum_{r\in\mathcal{R}} Z(r_t; s_t, u_t)$ is primarily preferred. Thus, we prefer expressing the expected value of an $\mathcal{F}_t$-measurable random variable $f : \mathcal{S} \to \mathbb{R}$ as $\mathbb{E}[f(s')|\mathcal{F}_t] = \sum_{r_t\in\mathcal{R}} f(\Psi_t(r_t))Z(r_t; s_t, u_t)$.

3.1.3 Modified Cost Functional

Thus formally clad in an MDP modeling framework, we are now able to properly evaluate costs in this model. With each coordinate $(u, \eta) \in \mathcal{U} \times \mathcal{R}$, the decision-maker associates a cost $c(u; \eta)$, where the *cost function* $c : \Xi \to \mathbb{R}$ is known and fixed in time. Uncertainty in the random variable η propagates into uncertainty about the incurred

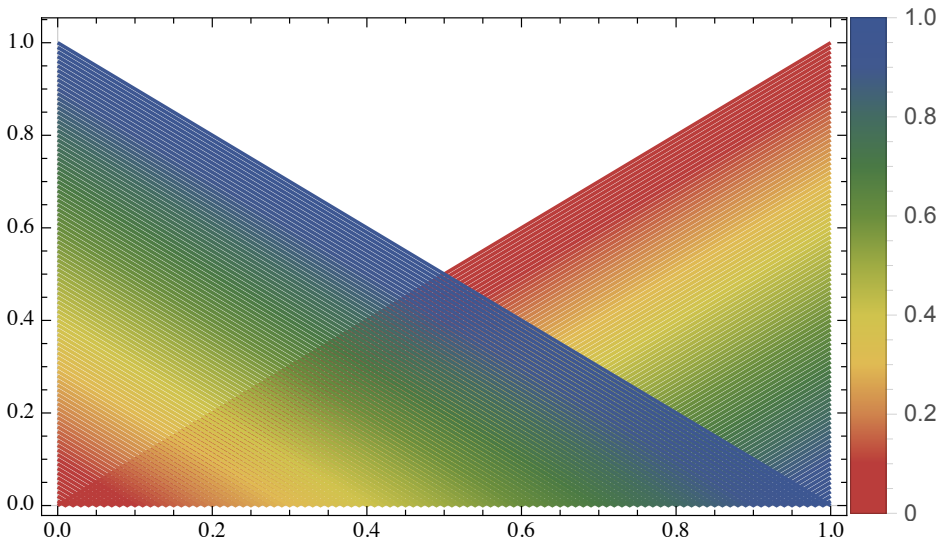


Figure 3.3: Example of a cost model for $c(u; \eta) = |u - \eta|$. Spectrum denotes $\mathcal{N}$.

cost of action u_t at any time t. This statement means that formally the cost of action at time t depends only on the current belief state s_t and the action u_t, in keeping with the theory of MDPs. For all times t, we therefore define the *expected cost* function $\bar{c}(u_t; s_t)$, given by

$$\bar{c}(u_t; s_t) := \mathbb{E}_t\big[c(u_t; \eta)\big] = \int_{\mathcal{N}} c(u_t; \eta)s_t(\eta)\, d\eta, \tag{3.7}$$

where the expectation operator $\mathbb{E}_t[\cdot] := \int_{\mathcal{N}} \cdot\, s_t(\eta)\, d\eta$. To ease notation, we will often write $c_t(u_t) \triangleq \bar{c}(u_t; s_t)$. Note that, in Chapters 4 and 5 below, we will focus on the cost function $c(u; \eta) := |u - \eta|$.

3.2 Dynamic Programming

3.2.1 Risk–neutral Dynamic Programming

The definition of the cost functional completes the formalization of our problem as a Markov Decision Problem. According to the classical theory, we would be concerned with the *expected system cost*

$$C_{t,T}(s; \boldsymbol{\pi}) := \mathbb{E}\Big[\sum_{t=1}^{T} \bar{c}(u_t; s_t)\Big],$$

where $s_1 = s$ and, at each stage t, the control $u_t = \pi_t(s_1, \ldots, s_t)$. Comprehensive literature is available on this topic. (See, among others, [109, 24, 23] and the references therein.) We briefly summarize the fundamental results.

Among all possible policies, including randomized and history-dependent, a Markov policy of the form $u_t = \pi_t(s_t)$ is best. It can be found by evaluating the *optimal value functions* v_t^* at stages $t = 1, \ldots, T$, which are defined as follows:

$$v_t^*(s) := \inf_{\boldsymbol{\pi} \in \boldsymbol{\Pi}} C_{t,T}(s; \boldsymbol{\pi}). \tag{3.8}$$

where Π is the set of Markov policies, adapted mappings from $\mathcal{S}^T$ into $\mathcal{U}$; specifically, $\Pi := \{\boldsymbol{\pi} = (\pi_t) \,|\, \pi_t(s_1, \cdots, s_t) = \pi_t(s_t) \in \mathcal{U},\ t = 1, 2, \cdots, T\}$. This can be accomplished by solving the dynamic programming equations through backward induction.

The following theorem can be formulated under rather general and abstract conditions, involving lower semicontinuity and weak continuity of the mappings involved [73]. We state it in the special case, which will play a major role in further considerations.

Theorem 24 (Risk–neutral dynamic programming equations)**.** *Suppose the sets $\mathcal{U}$ and $\mathcal{R}$ are finite. Then the optimal value function v_1^* satisfies the dynamic programming*

equations for all $s \in \mathcal{S}$:

$$v_{T+1}(s) = \min_{u \in \mathcal{U}} \bar{c}(u; s), \tag{3.9}$$

$$v_t(s) = \min_{u \in \mathcal{U}} \left\{ \bar{c}(u; s) + \int_{\mathcal{S}} v_{t+1}(y)\, Q(dy|s, u) \right\}, \quad t = 1, 2, \cdots, T. \tag{3.10}$$

Furthermore, the minimizers on the right-hand sides of the above equations define the optimal Markov policy $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_t^*)$.

The integration with respect to the transition kernel in (3.10) can be made more explicit by using (3.6), to obtain

$$v_t(s) = \min_{u \in \mathcal{U}} \left\{ \bar{c}(u; s) + \sum_{r \in \mathcal{R}} v_{t+1}\big(\Psi(r; s, u)\big) Z(r; s, u) \right\}, \tag{3.11}$$

$$\forall\, t = 1, 2, \cdots, T.$$

3.2.2 Risk–averse Dynamic Programming

In our Markov belief setting, we now consider a time-consistent dynamic Markov risk measure $\rho_{_T} := \{\rho_{t,T}\}_{t=1}^T$. We define the problem of choosing a policy $\boldsymbol{\pi} = \{\pi_t\}_{t=1}^T$ to minimize the accumulated *risk* over T stages:

Definition 23 (System Risk)**.** *For a dynamic risk measure $\rho_{_T} := \{\rho_{t,T}\}_{t=1}^T$, we define the system risk under Markov policy $\boldsymbol{\pi}$ from time t to T as follows*

$$R_{t,T}(s; \boldsymbol{\pi}) := \rho_{t,T}\Big(\big\{\bar{c}(u_i; s_i)\big\}_{i=t}^T\Big), \tag{3.12}$$

where $s_t = s$ and, at each stage i, the control $u_i = \pi_i(s_i)$.

The *optimal value function* v_t^* at stage t in this setting is therefore defined as

$$v_t^*(s) := \inf_{\boldsymbol{\pi} \in \boldsymbol{\Pi}} R_{t,T}(s; \boldsymbol{\pi}), \tag{3.13}$$

where Π is the set of Markov policies. Our goal is to specify an appropriate risk measure $\rho_{_T}$ and solve the problem (3.13) at $t = 1$. As in classical dynamic programming, this can be accomplished by finding all functions $v_t^*(\cdot)$ for $t = T, T-1, \cdots, 1$. We briefly outline this construction.

Let $Z_1, Z_2, \cdots, Z_{_T}$ be a sequence of random variables (understood as costs) adapted to the filtration $\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_{_T}$. According to the theory introduced in [122], a time-consistent dynamic risk measure satisfying the *translation property*:

$$\rho_{t,T}(Z_t, Z_{t+1}, \cdots, Z_{_T}) = Z_t + \rho_{t,T}(0, Z_{t+1}, \cdots, Z_{_T}),$$

and the *normalization property*,

$$\rho_{t,T}(0, \cdots, 0) = 0,$$

necessarily has the following structure:

$$\rho_{t,T}(Z_t, Z_{t+1}, \cdots, Z_{_T}) = Z_t + \rho_t\Big(Z_{t+1} + \rho_{t+1}\big(Z_{t+2} + \cdots + \rho_{T-1}(Z_{_T})\cdots\big)\Big), \quad (3.14)$$

where $\rho_t(Z_{t+1}) = \rho_{t,T}(0, Z_{t+1}, 0, \cdots, 0)$ are *one-step conditional risk measures* (Definition 5 on 23). Moreover, for the class of *Markov risk measures*, transition risk mappings[6] σ_t exist, such that for any Markov policy $\boldsymbol{\pi}$ the following equation is true for all $t = 1, \cdots, T-1$, and all $s \in \mathcal{S}$:

$$R_{t,T}(s; \boldsymbol{\pi}) = \bar{c}\big(\pi_t(s_t); s_t\big) + \sigma_t\Big(s_t,\, Q(\cdot|s_t, \pi_t(s_t)),\, R_{t+1,T}(\cdot; \boldsymbol{\pi})\Big). \quad (3.15)$$

This form allows for the development of *risk–averse* dynamic programming equations for our problem.

Theorem 25 (Risk–averse dynamic programming equations)**.** *The optimal value function v_1^* satisfies the dynamic programming equations for all $s \in \mathcal{S}$:*

$$v_{_{T+1}}(s) = \min_{u \in \mathcal{U}} \bar{c}(u; s), \quad (3.16)$$

$$v_t(s) = \min_{u \in \mathcal{U}} \Big\{\bar{c}(u; s) + \sigma_t\big(s,\, Q(\cdot|s, u),\, v_{t+1}(\cdot)\big)\Big\}, \;\; t = 1, 2, \cdots, T. \quad (3.17)$$

Proof. See Theorem 2, [122]. □

Transition risk mappings σ_t can be derived from law invariant measures of risk, by making their dependence on the probability measure explicit.

---

[6]We note that in the seminal literature [122], the term "risk transition mapping" was coined and quickly picked up. Shortly thereafter, it was changed to the more apt "transition risk mapping" [32, 58]. In our view, this simple transposition materially changes the literal interpretation. In non–technical language one might write "transition-risk mapping" (roughly, the risk in transitioning), more clearly conveying the interpretation of risk preferences as ambiguity in the transition kernel.

Example 22. *The mean–semideviation risk measure [99] has the following transition risk mapping counterpart:*

$$\sigma(s,q,v) = \int_{\mathcal{S}} v(s')\; q(ds') + \varkappa \int_{\mathcal{S}} \max\left(0, v(s') - \int_{\mathcal{S}} v(z)\, q(dz)\right)\; q(ds'), \tag{3.18}$$

where $\varkappa \in [0,1]$ is the risk–aversion parameter. In (3.17)*, we substitute $q = Q(s,u)$. Then, similar to* (3.11)*, we can use the specific structure of the kernel Q to write:*

$$\sigma(s,Q(s,u),v) = \bar{v}(s,u,v) + \varkappa \sum_{r\in\mathcal{R}} \max\Big(0, v(\Psi(r;s,u)) - \bar{v}(s,u,v)\Big)\; Z(r;s,u),$$

where the mean $\bar{v}(\cdot,\cdot,\cdot)$ has the form:

$$\bar{v}(s,u,v) = \sum_{r\in\mathcal{R}} v\big(\Psi(r;s,u)\big)\; Z(r;s,u).$$

Example 23. *The Average Value at Risk at level $\alpha \in (0,1]$, AVaR_α defined in* (1.21) *on page 21, has the following transition risk mapping counterpart:*

$$\sigma^\alpha(s,q,v) = \min_{\zeta\in\mathbb{R}} \left\{ \zeta + \frac{1}{\alpha} \int_{\mathcal{S}} \max\big(0, v(s') - \zeta\big)\; q(ds') \right\}, \tag{3.19}$$

After substituting $q = Q(s,u)$, we can use the specific structure of the kernel Q to write:

$$\sigma^\alpha(s,Q(s,u),v) = \min_{\zeta\in\mathbb{R}} \left\{ \zeta + \frac{1}{\alpha} \sum_{r\in\mathcal{R}} \max\big(0, v(\Psi(r;s,u)) - \zeta\big)\; Z(r;s,u) \right\}.$$

In particular, observe that for $\alpha = 1$, the risk transition mapping $\sigma^1(s,q,v) = \mathbb{E}_q[v|s]$. Similarly, as $\alpha \downarrow 0$, we obtain $\sigma^\alpha \to \max_r v(\Psi(r;s,u))$. In this problem setting, the efficacy of the Kusuoka representation, Equation (1.27), thus becomes clear: It is particularly efficient to represent any coherent, law invariant risk measure via combinations of $AVaR_\alpha$ in this one-dimensional setting.

3.2.3 Response-based Transition Risk Mappings

The examples discussed in the previous subsection suggest that we can model risk aversion in a compact form, by considering functions $v\big(\Psi(\cdot\,;s,u)\big)$ on the space of observations $r \in \mathcal{R}$, with probability measures $Z(\cdot\,;s,u)$. We thus focus on the following structure of the transition risk mapping:

$$\sigma_t(s,Q(s,u),v) = \sigma_t^r\big(s, Z(\cdot\,;s,u), v\big(\Psi(\cdot\,;s,u)\big)\big),$$

where $\sigma_t^r : \mathcal{S} \times \mathcal{P}(\mathcal{R}) \times \mathcal{V}(\mathcal{R}) \to \mathbb{R}$. This simplification is of great practical importance, because it reduces the risk model to considering finite distributions of possible responses.

The dynamic programming equations (3.16) simplify significantly:

$$v_{_T}(s) = \min_{u \in \mathcal{U}} \bar{c}(u; s), \tag{3.20}$$

$$v_t(s) = \min_{u \in \mathcal{U}} \Big\{ \bar{c}(u; s) + \sigma_t^r\big(s, Z(\cdot\,; s, u), v_{t+1}\big(\Psi(\cdot\,; s, u)\big)\big) \Big\}, t = 1, 2, \cdots, T. \tag{3.21}$$

The examples of the previous section are fully consistent with this setting.

In our application, we have only 2 possible response values: "toxic" and "nontoxic". In this case, it appears reasonable to consider two extreme cases of the response-based transition risk mapping: the expected value, and the worst case:

$$\sigma_t^r\big(s, Z(\cdot\,; s, u), v\big(\Psi(\cdot\,; s, u)\big)\big) = \max_{r \in \mathcal{R}} v_{t+1}\big(\Psi(r; s, u). \tag{3.22}$$

It completely eliminates the probability distribution $Z(\cdot\,; s, u)$ from the risk calculation: we treat the toxic and non–toxic case equally. We call this transition risk mapping *robust*.

Example 24 (MSD_κ DP equations)**.** *The* MSD_κ *is introduced in* (1.23) *admits analytic dual representation* (1.26)*, given by*

$$\sigma\big(r_t, Z_t(r_t; \pi_t), v_{t+1}\big) = \max_{\mu \in \mathcal{A}(\sigma)} \mathbb{E}_\mu[v_{t+1}],$$

where the subdifferential $\mathcal{A} := \partial(\sigma)|_0$ *is given by*

$$\mathcal{A} = \Big\{ \mu \in \Psi(\mathcal{R}) : \frac{\mu(r)}{Z(r; s, u)} = 1 + h(r) - \sum_{r \in \mathcal{R}} h(r) Z(r; s, u), \;\; h(r) \in [0, \kappa], \;\; \forall\, r \in \mathcal{R} \Big\}.$$

This formulation is linear, and thus for all $r \in \mathcal{R}$ *and* $u \in U$ *this may be solved by linear programming. For our purposes in* (3.17)*, however, is unfortunately not linear, or convex in the control. When log-concavity is satisfied, however, we may obtain a convex formulation in the second-stage. Unfortunately, dependence on the control renders the problem disjunctive, so that* min *and* max *operators may not be interchanged (see [122] p. 259).*

3.2.4 Challenges of Dynamic Programming

Dynamic programming (DP) is a powerful technique for obtaining solutions to problems that would otherwise be impossible. At its core, DP is built around the notion of the fixed point of an iterated mapping, which so long as the mapping is a contraction mapping in the underlying topology is guaranteed to exist and be unique. Indeed, this is the essence of the value iteration method of obtaining optimal solutions in the infinite–horizon problem. In finite horizon problems, one must instead solve a system of equations over the state space backward in time, known as backward induction. These techniques are guaranteed to lead to solutions under very general and abstract conditions.

Unfortunately, the techniques have computational complexity exponential in the state and control spaces. In many practical cases lacking an exploitable structure, the size of the state space and/or control spaces can render the approach utterly impossible. Bayesian inference with non–conjugate belief states constitutes such a case.

Rather than belabor this well–known phenomenon, we discuss the issues in our context by appeal to visual representation, which will be worth at least one thousand words indeed. Consider Figure 3.4, depicting all of the possible states arising from a very particular slice of the full optimal learning problem. Several observations may be made, but we first state some details about the image.

In Figure 3.4, we have begun from a fixed initial prior belief state s_1 to be uniform over $\mathcal{N}$. Note that this represents but one of very many candidate initial prior states. Second, we have begun with an initial starting dose fixed at the minimum feasible dose. That is, $u_1 \equiv 0$ for each orbit in the picture. Note that we have also discretized the dosage space to $|\mathcal{U}| = 100$ possible dosages, and so with this fixed choice of u_1 the picture represents roughly one part in one hundred of the "full picture". Finally, and certainly the most constraining, we have restricted to a one-step lookahead policy, in particular the vanilla expected value policy $\boldsymbol{\pi} \triangleq \boldsymbol{\pi}^{EV}$. The reduction in complexity from this restriction is hard to overstate.

With all of these restrictions, in Figure 3.4 we observe $2^{T-1} = 2^9 = 512$ belief orbits,

yielding $2^T = 2^{10} = 1024$ individual belief states. Supposing we still allow the initial prior and initial starting dose to be fixed, proper dynamic programming would require that we consider $(2 \cdot 100)^{10} = 1024 \cdot 10^{20}$ states! Thus, the picture depicts one part in 10^{20} of what is required, which is clearly computationally infeasible. Moreover, the significant overlap with respect to the densities suggests that there is little to be gained by forecasting too far into the future, and this naturally motivates consideration of lookahead policies. However, to investigate the prospects of forecasting, which is to say the prospects of backward induction, we shall investigate approximate DP techniques.

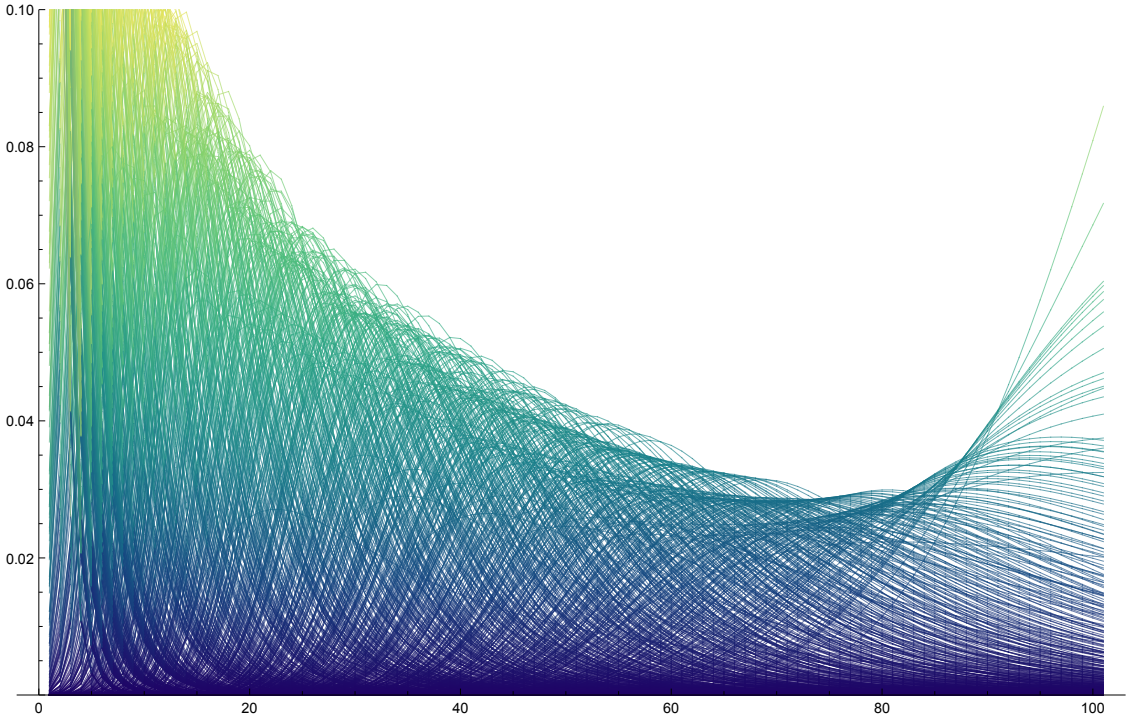


Figure 3.4: All 2^{T-1} possible states $\Psi^{\boldsymbol{\pi}}(s_1) = \{\Psi(r_t; \pi_t, s_t)\}$ after time $T = 10$, starting from s_1 (uniform), $u_1 \equiv u_{\min}$, and further restricted to myopic expected control $\boldsymbol{\pi}$; discretized to $|\mathcal{U}| = 100$ feasible controls. Note: Precisely T states are realized; moreover, proper DP would require backward induction on $(2 \cdot 100)^T = 200^{10}$ states!

3.3 The Class of Lookahead Policies

Arguably the simplest approach falling under the umbrella term approximate dynamic programming (ADP) is that of *lookahead policies*. The technique is straightforward:

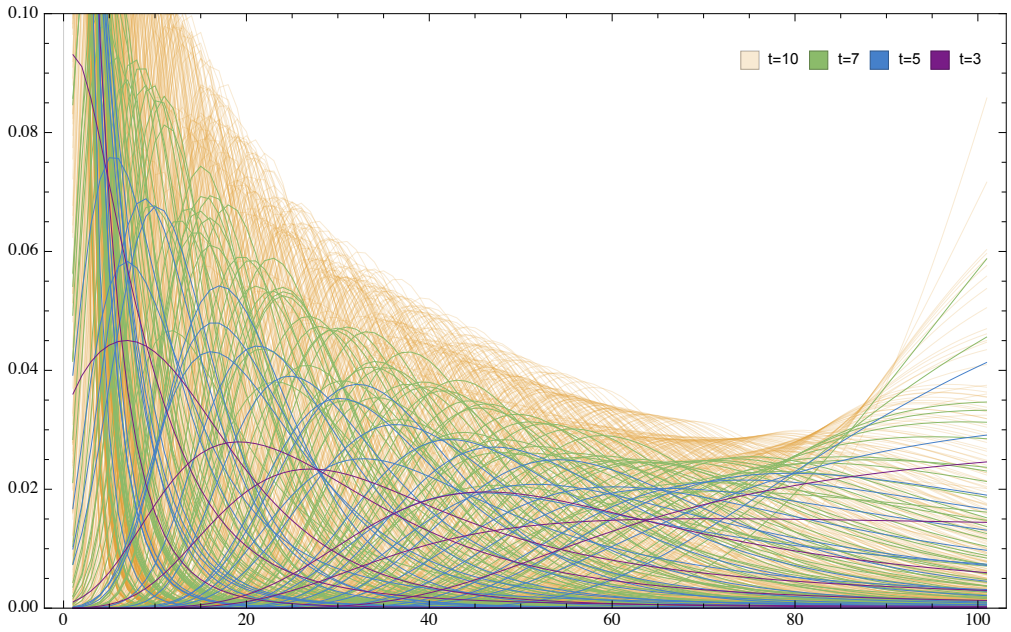


Figure 3.5: The same scenario shown in Figure 3.4, to give a sense of the thick fog inherent to dynamic programming with belief states.

Instead of solving the DP recursion (Bellman's equation), or in the finite-horizon case the system of DP equations, lookahead policies generally solve a truncated version of the problem. The k-step lookahead policy is to take the action at each time that appears optimal considering the next k time periods. Because uncertainty cascades systemically, 1-step lookahead (or *myopic*) policies are an attractive choice for balancing the improvements afforded by future considerations with their exponential computational demands. Owing to their simplicity, many classes of such problems have been studied, most notably in the online learning and multi-armed bandit (MAB) literature. Moreover, the employ of coherent risk measures is an intuitively appealing approach for controlling future risk without explicitly formulating the entire system of equations. Below, we present several of the foremost formulations and introduce the robust-response formulation.

3.3.1 Selected Prominent Policies

We first introduce some common notation used below. In what follows, we generally denote the myopic minimizer of the cost functional c with respect to a generic state s

by $\hat{\mu}$ and the associated value by $\hat{c}$. Formally, we write

$$\hat{\mu} \triangleq \operatorname*{argmin}_u \mathbb{E}[c(u;\eta)] = \operatorname*{argmin}_u \bar{c}(u;s),$$
$$\hat{c} \triangleq \min_u \mathbb{E}[c(u;\eta)] = \bar{c}(\hat{\mu};s).$$

In the lookahead context one frequently needs to consider the myopic minimizer of the next stage, which is, of course, stochastic with respect to the controlled response transition kernel. In the one-step lookahead setting, subscripting by time is unnecessary and becomes cumbersome. We thus continue our convention of overloading the operators to indicate the quantities one-step forward in time, as we have above. Specifically, given the generic state s, we write

$$\hat{c}(r;s,u) \triangleq \min_{x\in\mathcal{U}} \bar{c}(x;\Psi(r;s,u))$$
$$\text{(shorthand)} \qquad \hat{c}^r_u = \min_{x\in\mathcal{U}} \bar{c}(x;\Psi^r_u),$$

and similarly for the minimizer[7]

$$\hat{\mu}(r;s,u) \triangleq \operatorname*{argmin}_{x\in\mathcal{U}} \bar{c}(x;\Psi(r;s,u))$$
$$\text{(shorthand)} \qquad \hat{\mu}^r_u = \operatorname*{argmin}_{x\in\mathcal{U}} \bar{c}(x;\Psi^r_u).$$

Example 25 (Knowledge gradient)**.** *For any time t, denote by v_t^{KG} the optimal KG value function, given by*

$$v_t^{KG}(u) \triangleq \mathbb{E}_{Z^r_u}\left[\hat{c}(r;s_t,u)\right] \tag{3.23}$$
$$= \sum_{r\in\mathcal{R}} \hat{c}(r;s_t,u_t) Z(r;s_t,u)$$

The KG policy is then to choose a dose x at time t by solving

$$\pi_t^{KG} \triangleq \operatorname*{argmin}_{u\in\mathcal{U}} \bar{c}_t(u) + (T-t) v_t^{KG}(u). \tag{3.24}$$

---

[7]The intuition behind the notation is the following: Recall that we focus attention on the one-step cost function $c(u;\eta) = |u-\eta|$, and thus the minimizer is the posterior median.

Example 26 (Information-directed sampling)**.** *Let the information gain of dose $u \in \mathcal{U}$ be denoted $g_t(u)$, and given by*

$$g_t(u_t) := I_t(r, \eta) = D_{f^0}\big(\Phi(r_t; u_t, \eta)s_t(\eta), Z_t(r_t; u_t)\big),$$

which is more concisely expressed in the entropy-reduction form as

$$g_t(u_t) = \mathbb{E}[H^0(s_t) - H^0(\Psi(r_t; s_t, u_t))], \tag{3.25}$$

where H^0 is the Shannon entropy.

Letting $\Delta_t^{IDS}(u_t) = \mathbb{E}[c(u_t; \eta) - c(\hat{\mu}_t; \eta)]$, at time t the information-directed sampling (IDS) policy π_t^{IDS} is then to choose dose u_t according to

$$\pi_t^{IDS} = \operatorname*{argmin}_{u_t} \frac{(\Delta_t^{IDS}(u_t))^2}{g_t(u_t)} = \frac{\mathbb{E}[c(u_t; \eta) - c(\hat{\mu}_t; \eta)]^2}{\mathbb{E}[H^0(s_t) - H^0(\Psi(r_t; s_t, u_t))]}. \tag{3.26}$$

Several remarks may be made about the *information ratio* in (3.26). Note that the expectation in the numerator is with respect to $\eta \sim s_t$, whereas the expectation in the denominator is with respect to $r_t \sim Z(r_t; u_t)$ for all $u_t \in \mathcal{U}$. Thus, this quantity decouples, in a certain sense, expectations over the product space $\Omega = \mathcal{N} \times \mathcal{R}$, and it compares them numerically by the operation of division. The authors are able to provide theoretical bounds on the expected cost under this policy by way of entropic inequalities. However, they express ambiguity that it is "unclear if or when this is the right measure" [120].

3.3.2 Robust Response Policy

Given that the uncertainty in the model is compounded in the second stage by uncertainty both in the transition kernel and the resulting loss distribution, we posit risk measures with a composition structure. Specifically, for all t, we thus consider risk measures ρ_t of the form

$$\rho_t\big(c(u_t; s_t) + c(u_{t+1}, s_{t+1})\big) = \varrho_{t,s} \circ \varrho_{t,z}\big(c(u_t; s_t) + c(u_{t+1}, s_{t+1})\big) \tag{3.27}$$

$$= \varrho_{t,s}\Big(c(u_t; s_t) + \varrho_{t,z}\big(c(u_{t+1}, s_{t+1})\big)\Big). \tag{3.28}$$

Because in the case of a one-dimensional, compact control space, all measures of risk will necessarily be combinations of the expectation and the worst-case, it is intuitive to consider the robust approach wherein

$$\varrho_{t,z} \equiv \max_{r \in \mathcal{R}} c(\eta_{t+1}(r), u_{t+1}).$$

We thus formulate the robust lookahead policy.

Example 27 (Robust response)**.** *Let the time t be fixed, and denote by v_t^{RR} the optimal robust-response value function, given by*

$$v_t^{RR}(u_t) := \max_r \{\hat{c}_{t+1}(r; u_t)\}.$$

The robust response policy π_t^{RR} is to choose u_t according to

$$\pi_t^{RR} = \operatorname*{argmin}_{u_t} \mathbb{E}[c(u_t; \eta)] + \gamma_t v_t^{RR}(u_t), \tag{3.29}$$

where γ_t is a constant factor.

This policy has the unique benefit of decoupling the uncertainty in the optimal parameter η and the optimal transition kernel. Note that, in contrast to upper confidence bounding (UCB) or other robust approaches, the RR policy considers the average cost of worst-case response. Contrast this with, for example, the worst-case (over an uncertainty set) average of the average (over response) cost.

In the empirical study of Sections 5.1 and 5.2 below, we will observe that the RR policy exhibits strong performance in applications with relatively short horizon T, both in terms of statistical accuracy, precision, and accumulated cost. What is especially interesting about these, granted empirical, results is that this performance is achieved despite effectively disregarding the central conditional probability relation (3.2). Of course, this is a consequence of the fact that the controlled stochastic process is non-linear, whereas the conditional probability at any given time bears analogy to a local linearization. Indeed, the prefatory study in Section 5.1 illustrates the role of response risk measures as a local correction to this linearization, effectively tempering inference in a way analogous to inertial mass tempering acceleration. Indeed, we propose viewing

law-invariant response risk measures as *inertial belief*, an intrinsic property of a belief state. This motivates a more general investigation of the role of risk measures with respect to conditional probability in an adapted inference framework.

3.4 Augmenting Bayesian Inference

3.4.1 Information Loss and Monotone Insufficient Statistics

Statistical manifolds are classically studied under the assumption of a sufficient statistic, a collection of observable features that forms a basis for the underlying statistical manifold. Often times, however, in practice one is forced to work with an *insufficient statistic*. In this section, we offer some budding thoughts for augmenting Bayesian inference in the case of insufficient statistics satisfying stochastic ordering in the model class.

We focus on the case of binary observation as proxy to observation of a continuous random variable. For example, in the clinical trial design problem, instead of observing each patient's toxicity random variable τ, we instead observe the binary response $r \in \{0, 1\}$ to dose $u \in \mathcal{U}$. Intuitively, it is easy to see that a certain loss in information arises from the description of the real number τ by one of the statements "τ is greater than (less than) u."

Given that we know a priori the nature of information loss in this model, the idea is to account for the loss by augmenting our inference methods. Clearly, there exists a continuum of techniques for such an augmentation. We present three approaches.

Suppose we have the prior distribution $s(\eta)$, and consider applying dose u and observing response $r = 1$. Then by Bayes' theorem, we obtain that the posterior $s_+(\eta)$ is proportional to $\Phi(u;\eta)s(\eta)\ d\eta$. Suppose the realization τ is fixed. Then, in particular, for all $x \in [\tau, u]$ the same response $r = 1$ would be observed. By the monotonicity of $x \mapsto \Phi(x;\eta)$, there exists a natural stochastic ordering of posterior distributions depending on x. The value $x = \tau$ results in a posterior minimal with respect to this ordering. However, $x < \tau$ yields observation $r = 0$. Hence, the posterior distribution is proportional to $(1 - \Phi(u;\eta))s(\eta)$, which is *larger* with respect to the

ordering *for all* $x \in [\tau, u]$.

The above discussion develops the intuition that, ceteris paribus, those u near τ yield more information than those u far from τ.[8] However, it also illustrates a certain sensitivity, in that there exists a jump discontinuity in posterior distributions about the realization τ. Indeed, this may be formalized with an appropriate information metric in terms of divergences. Thus, for a fixed realization τ and dose u with corresponding response r, the posterior distributions corresponding to a notional dose $\hat{u} \in [\tau, u)$ (if $r = 1$), or $\hat{u} \in (u, \tau)$ (if $r = 0$), would result in greater shift of posterior from the prior *in the same direction*. If in the inference experiment observation $r = 1$ is assumed with $\hat{u} < \tau$, then large errors arise relative to what would be observed empirically, and analogously for $r = 0$ with $\hat{u} \geq \tau$. Thus, we observe a familiar risk–reward tradeoff in the inference tradespace: information recapture versus assumed observational error.

From this perspective, the maximally risk–averse approach would be to proceed with $\hat{u} = u$, which has vanishing probability of observational error. One alternative approach would be to employ a threshold probability level. Letting $\alpha \in [0, 1]$ be given, choose $\hat{u}$ to be the α-quantile of conditional distribution of τ, which we recognize as AVaR_α. Formally, $\hat{u}$ is the smallest dose satisfying $\mathbb{E}_{s(\eta)}[\Phi(\hat{u}; \eta)/\Phi(u, \eta)] \geq \alpha$. More generally, risk in the model uncertainty may be further controlled by employing other risk measures, all of which may be written as convex combinations of AVaR_α via the Kusuoka representation (1.27). Thus, the inference experiment would produce a notional dose $\hat{u}$, in the case $r = 1$, given by

$$\begin{aligned} &\underset{x \in \mathcal{U}, x \leq u}{\operatorname{argmin}}\ x \\ &\text{subject to } \ \rho\big(\Phi(x; \eta)/\Phi(u, \eta)\big) \geq \alpha. \end{aligned}$$

A Bayesian update would then result in the posterior

$$s_+(\eta) = \Phi(\hat{u}; \eta) s(\eta) / Z(\hat{u}; s, r).$$

We note that a more sophisticated version in the spirit of this approach would account for the differential information gain, movement of the posterior. The general

[8] Naturally, the information gain depends on the prior belief state.

idea is to balance the monotonic decrease in probability with the monotonic increase in information gain, although this is a well-known dilemma without established methodology. In fact, this dilemma is *precisely* that which we face in determining the dose u originally, although with respect to an objective and probability distribution with entirely different functional forms. As an example, suppose the value of information recapture and cost of error are quantified, respectively, with respect to a given divergence D_f as

$$\iota(\hat{u}) = D_f\big(\Psi(r; s, u), \Psi(r; \hat{u}, s)\big), \tag{3.30}$$

$$\ell(\hat{u}) = D_f\big(\Psi(1; \hat{u}, s), \Psi(0; \hat{u}, s)\big). \tag{3.31}$$

In this case, letting the conditional distribution function for τ be denoted $p(x; \eta) = \Phi(x; \eta)/\Phi(u, \eta)$, we have the corresponding expectation of total value of using x in place of u in this inference experiment:

$$I(x; \eta) = \iota(x)p(x; \eta) - \ell(x)\big(1 - p(x; \eta)\big). \tag{3.32}$$

The classical step would then be to choose x to maximize the expectation (or more generally the risk) with respect to uncertainty in the model: $\max_x \rho(I(x; \eta))$. We then propose proceeding with the Bayesian update using this maximizer.

Furthermore, we highlight the fact that this inference experiment occurs *after* observation. Therefore, the appropriate belief state with which to measure risk is the empirical posterior distribution $\Psi(r; s, u)$ obtained from the applied dose u. That is, in case particular case of expectation, i.e., $\rho \equiv \mathbb{E}$, we have

$$\hat{u} = \operatorname*{argmax}_{x \in \mathcal{U}, x \leq u} \int_{\mathcal{N}} I(x; \eta)\Psi(r; s, u)\, d\eta \tag{3.33}$$

$$= \operatorname*{argmax}_{x \in \mathcal{U}, x \leq u} \int_{\mathcal{N}} I(x; \eta)\Phi(u, \eta)s(\eta)\, d\eta, \tag{3.34}$$

and the ultimate posterior becomes $s_+(\eta) = \Psi(r; s, \hat{u})$.

As a derivative approach, again in this same spirit, we take inspiration from the evaluation of the cost of learning in IDS. Consider the following modification of the cost ℓ in (3.31). Given the notional response r is fixed, for any dose $\hat{u}$, we have the posterior $\Psi(r; s, \hat{u}) =: \hat{\Psi}^r(\hat{u})$. Again, the relevant cost is one of inferential error, wherein

$\hat{u}$ would in fact lead to $\tilde{r} \neq r$, with corresponding posterior $\Psi(\tilde{r}; s, \hat{u}) =: \hat{\Psi}^{\tilde{r}}(\hat{u})$. Suppose we evaluate this cost as

$$\mathbb{E}_{\hat{\Psi}^r(\hat{u})}[c(\hat{\mu}^{\tilde{r}}; \eta) - c(\hat{\mu}^r; \eta)]^2 = \left(\int \Big(c(\hat{\mu}^{\tilde{r}}; \eta) - c(\hat{\mu}^r; \eta) \Big) \Psi(r; \hat{u}, s)\, d\eta \right)^2,$$

where we let $\hat{\mu}^r$ and $\hat{\mu}^{\tilde{r}}$ denote the myopic minimizer with respect to posterior $\hat{\Psi}^r(\hat{u})$ and $\hat{\Psi}^{\tilde{r}}(\hat{u})$, respectively.

By writing the $f-$information gain in the entropy reduction form, we then obtain the modified quantities

$$\iota(\hat{u}) = H^f\big(\Psi(r; s, u)\big) - H^f\big(\Psi(r; s, \hat{u})\big), \tag{3.35}$$

$$\ell(\hat{u}) = \mathbb{E}_{\hat{\Psi}^r(\hat{u})}[c(\hat{\mu}^{\tilde{r}}; \eta) - c(\hat{\mu}^r; \eta)]^2. \tag{3.36}$$

Again, we have the conditional distribution function $p(\hat{u}; \eta) := \Phi(\hat{u}; \eta)/\Phi(u, \eta)$, and the corresponding inference value function $I(\hat{u}; \eta)$ is formally identical to (3.32) with $\hat{u}$ obtained as in (3.33) for some ρ.

Chapter 4
Approximate Dynamic Programming

> *Under the comb, the tangle and the straight path are the same.*
>
> Heraclitus, *535–475 BC*

> *Change your opinions, keep your principles; change your leaves, keep your roots.*
>
> Victor Hugo, *1907*

4.1 Overview

In the face of overwhelming complexity it is generally only possible to extract some approximation to the system as a proxy for analysis. In the throes of a complex system, one favors heuristic approaches, particularly simple, easy–to–recall calculations that give a sense of the magnitude and direction of a first–order correction—a rule of thumb. For example, when catching a flying object, say, a baseball, rather than attempt the slightest consideration of any equations of motion, instead one instinctively endeavors to maintain a constant viewing angle with the ball, moving forward or backward as necessary—simple, and effective. In a more considered setting, for example chess, often one relies on an ingenuity born of the quintessentially human capacity for creativity to distill complexity into its essence and extract the pertinent elements of the dynamic. Such heuristic approaches are surprisingly powerful, often mysteriously so.

Unfortunately, successful heuristics in one domain are generally not transferable, in an obvious or straightforward way, to problems in a different domain.[1] Moreover, and of increasing significance, the natural human bias toward simple heuristics squanders the prospects of machine learning and of delegating computational demands in general. Approximate dynamic programming (ADP) endeavors to combine the most effective aspects of the heuristics and of dynamic programming. In the previous Section 3.3 we investigated lookahead policies, which might be said to prioritize simplicity, insofar as the computational demands are relatively minor.

In this section, we introduce a more sophisticated approximation schema, which might be said to prioritize the use of computational resources to conduct the backward induction algorithm underlying dynamic programming. Specifically, we will investigate approximation in value space, and in this vein there exist many potential avenues. See [23] for a comprehensive treatment of various techniques. Inspired by a myriad of empirical observations indicating that complex, even chaotic, systems often exhibit low–dimensional behavior, we focus below on an approach essentially based on reducing the dimensionality by extracting the salient features of the problem.

4.2 Feature Selection

In view of our intended applications to clinical trial design (CTD) below, wherein the optimal control is given by a certain quantile of the toxicity distribution, we investigate structural feature sets arising naturally in this setting, in the following sense.

Definition 24 (Feature Vector and Extraction Mappings)**.** *Define the* feature extraction mapping $f : \mathcal{P}(\mathcal{N}) \to \mathbb{F} \subseteq \mathbb{R}^m$*, and denote its component mappings* $f_i : \mathcal{P}(\mathcal{N}) \to \mathbb{F}_i \subseteq \mathbb{R}$*,* $i = 1, 2, \cdots, m$*. For all states* $s \in \mathcal{S} \subseteq \mathcal{P}(\mathcal{N})$*, we associate the* feature vector $\phi = f(s)$*. We often write* $\phi = (\phi_i)$*,* $i = 1, 2, \cdots, m$*, and refer to its* i^{th} *component* $\phi_i = f_i(s)$ *as the* i^{th} feature *of* s*. In the following we study the CTD problem in the*

[1] However, we must note that doing so, cultivating such a *thematic interconnectedness*, might be regarded as the art of learning. Indeed, Waitzkin makes a strong case, as beautiful as it is compelling. See [145].

case of $m = 4$ features, with extraction mappings of the form

$$f_i := \operatorname*{argmin}_{u\in\mathcal{U}} \ \rho^i\big(c(u;\eta)\big), \qquad i = 1,2,3, \tag{4.1}$$

$$f_4 := \quad \min_{u\in\mathcal{U}} \ \rho^2\big(c(u;\eta)\big), \tag{4.2}$$

where each ρ^i_s is a one-step conditional risk measures.

It is perhaps natural, or at least efficient, to engineer features explicit in the dynamic programming equations (3.20)–(3.21). To this end, we proceed to investigate the case of $\rho^2 \equiv \mathbb{E}$. Moreover, in view of our efforts in the Chapter 5, we focus on cost functions of the form $c(u;\eta) = |\eta - u|$. In this case, recalling the expected cost functional $\bar{c}(u,s)$ defined in (3.7) on page 75, for all $s \in \mathcal{S}$ we set $\phi_4 \triangleq \min_{u\in\mathcal{U}} \bar{c}(u;s)$. It is well known that the minimizer of expected loss with respect to the L^1-norm is the median $\hat{\mu}$. Choosing $\phi_2 \triangleq \hat{\mu}$ to be this myopic minimizer, we obtain

$$\phi_2 = \operatorname*{argmin}_{u\in\mathcal{U}} \ \bar{c}(u;s) = \hat{\mu}(s) \tag{4.3}$$

$$\begin{aligned}\phi_4 = \quad & \min_{u\in\mathcal{U}} \ \bar{c}(u;s) = \bar{c}(\hat{\mu}(s);s) \\ = \ & \bar{c}(\phi_2;s).\end{aligned} \tag{4.4}$$

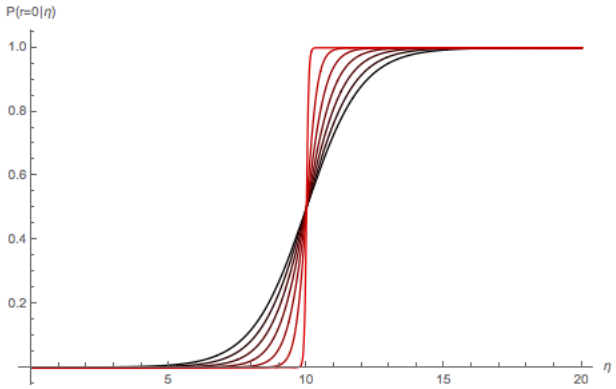


Figure 4.1: Sequence of logistic distribution functions tending (black to red) to a Heaviside function.

To motivate our choice of the remaining features ϕ_1, ϕ_3, consider the following idealized scenario. Fix the stage t and a state s_t, and consider the likelihood family given

by Heaviside functions $\{H(\eta,\cdot)\}_{\eta\in\mathcal{N}}$, so that $r_t = 0$, if $u_t < \eta$, and $r_t = 1$, if $u_t \geq \eta$. Let u_t be the myopic minimizer, i.e., the median, $\phi_2(s_t) = \hat{\mu}(s_t)$. As the posterior is given for $r_t = \{1, 0\}$, respectively, by $s_{t+1} = \{Z_t^{-1} s_t H(\cdot, \phi_2(s_t)), \tilde{Z}_t^{-1} s_t (1 - H(\cdot, \phi_2(s_t)))\}$, one can see from the properties of the Heaviside function that posteriors consist of normalized "halves" of the prior. Considering now $\phi_2(s_{t+1}) := \hat{\mu}(s_{t+1})$, it is clear that in this idealized scenario, the posterior median may be easily written in terms of quantiles of the prior:

$$\phi_2(s_{t+1}) = \begin{cases} q_{.25}(s_t), & \text{if } r_t = 1, \\ q_{.75}(s_t), & \text{if } r_t = 0. \end{cases}$$

Moreover, it can be shown that the family of Heaviside functions above can be expressed as limits of the logistic family, in the obvious way (see Figure 4.1). Therefore, for any family of logistic models, the .25- and .75-quantiles of any state s_t respectively serve as upper and lower bounds on $\phi_2(s_{t+1})$ conditioned on a myopic policy, in an almost-sure sense.

Motivated in this way, we define the remaining features as these quantile functions, and arrive at the feature set of feature extraction mappings:

$$\begin{cases} f_1 := q_{.25}(\cdot), \\ f_2 := q_{.50}(\cdot), \\ f_3 := q_{.75}(\cdot), \\ f_4 := \bar{c}(\hat{\mu}(\cdot), \cdot). \end{cases} \tag{4.5}$$

4.3 Feature Space Characterization

Construction of the feature space is central to the efficacy of ADP. On the one hand, minimal characterization fosters computational efficiency, while on the other hand completeness is required to accurately represent system dynamics.

We have natural relations on the components of the feature vector ϕ determined by (4.5). First and foremost, directly from the definition of the quantile function, ϕ_i,

$i = 1, 2, 3$, satisfy

$$\underline{\eta} \leq \phi_1 \leq \phi_2 \leq \phi_3 \leq \overline{\eta}. \tag{4.6}$$

Second, ϕ_i, $i = 1, 2, 3$, induce bounds on ϕ_4. An upper bound on ϕ_4 may be established by considering unimodality, and results in an elegant system of linear inequalities. The corresponding lower bound, however, requires additional considerations. In the case that log-concavity of states is relaxed, the lower bound may be obtained as the solution of a linear programming problem. The corresponding lower bound is universal but decidedly not sharp.

Sharp lower bounds may be obtained by employing shape constraints, for example, log-concavity, on the distribution s. A sharp bound, may be obtained by nonlinear optimization of the problem

$$\begin{aligned}
&\min_{s} \phi_4 \\
&\quad st \int_{\mathcal{N}(\phi_1)} ds = .25 \\
&\quad \int_{\mathcal{N}(\phi_2)} ds = .5 \\
&\quad \int_{\mathcal{N}(\phi_3)} ds = .75 \\
&\quad \int_{\mathcal{N}} ds = 1 \\
&\quad \log s_{\eta\eta} - 2 \log s_{\eta} + 2 \log s \geq 0 \\
&\quad s \geq 0.
\end{aligned}$$

The penultimate inequality enforces the log-concavity of the state s. When the problem is discretized, so that the states are probability vectors, Definition 22 may be profitably used. For any (agreeable) feature set, bounds obtained from general shape-constraints are necessarily valid for all distributions in the class and therefore offer a modicum of portability to alternative likelihood models. This fact perhaps justifies the additional efforts required to solve the nonlinear control problem.

4.4 Feature Disaggregation

When using projected or aggregation methods in ADP, the issue of determining feature transitions is problematic when the state transition kernel cannot be easily projected. In the CTD problem, and in the belief space of POMDPs generally, states transition according to a Bayes operator Ψ. Although the form of Ψ depends on the particular conditional likelihood model, for an arbitrary transition mapping $s_t \mapsto s_{t+1}(u, r)$, there exists a measurable kernel (family of joint conditional distributions) such that $s_{t+1}(\cdot, \cdot) = \Psi(\cdot, \cdot, s_t)$ is a Bayesian posterior of s_t with respect to the kernel. However, we cannot project the Bayes operator onto the features in an obvious way.

One approach for addressing this issue is mapping the feature space onto a set of probability distributions, from which the Bayes operator may be applied, and then projecting the posterior states back to feature space. That is, for any feature vector $\phi_t \mapsto \tilde{s}_t \in \Gamma$, whence $\tilde{s}_{t+1}(\cdot, \cdot) = \Psi(\cdot, \cdot, \tilde{s}_t)$ and $\phi_{t+1}(\cdot, \cdot) = \phi(\tilde{s}_{t+1}(\cdot, \cdot))$

$$(\phi_t, u, r) \mapsto \phi \circ \Psi(u, r, \cdot) \circ \Pi_\Gamma(\phi_t)$$

$$\phi_{t+1}(u, r) = \phi\Big(\Psi\Big(u, r, \Pi_\Gamma(\phi_t)\Big) \Big).$$

The situation will be elucidated much more clearly in Figures 4.3a and 4.3b below.

The central issue in the efficacy of such an approach is the ability for the set of distributions Γ to well approximate the original state space $\mathcal{S}$. There exist several challenges to this end. Note that $\mathcal{S}$ is in fact the space of all possible Bayesian sequences $\{\Psi_{s_1}^T(u^T, r^T)\}$, the elements of which are generally not in any well-known distribution family. For any given likelihood family, assumed not to be part of a conjugate pair, it is conceivable to construct a simplified family (or families), tailored to the likelihood family via some approximation schema. For example, one attractive approximation schema is defining families Γ_t, $t = 1, \cdots, T$, in terms of parameterized mixture distributions.

As it happens, for the case of logistic likelihood models with uniform prior s_1, elements of $\mathcal{S}$ find a close approximation in the 2-parameter family of log-logistic distributions.

Example 28 (Log-logistic Family of Distributions)**.** *We denote the (restricted) family of log-logistic distributions by* Γ_L*, given by*

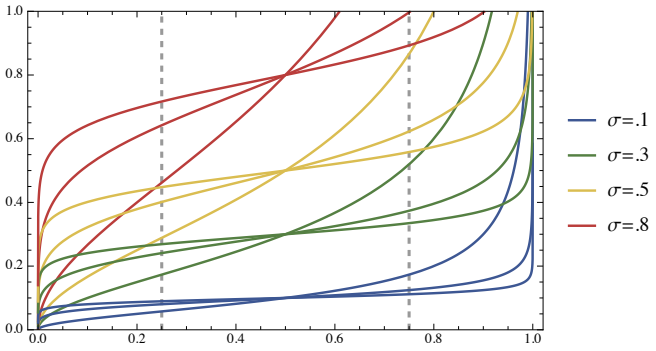
$$\Gamma_L = \left\{ s = g(\gamma, \sigma) \;\middle|\; g(\gamma, \sigma)(x) = \frac{\gamma \sigma^{-\gamma} x^{\gamma - 1}}{(1 + (x/\sigma)^{\gamma})^2}, \quad \gamma > 0, \sigma \in \mathcal{N} \right\}, \tag{4.7}$$

with support on $x \in [0, \infty)$*. Below, we shall denote this parameter space by* $\mathbb{R}^2_L := \{(\gamma, \sigma) \in (0, \infty) \times \mathcal{N}\}$*.*

We note that the support of this family is particularly appropriate in the case where the parameter η is a physical quantity, such as the MTD. The family Γ_L offers additional efficiencies in our setting. First, the scale parameter $\sigma > 0$ is, in fact, identically equal to the median on Γ_L, and therefore $\sigma \in \mathcal{N}$. In order to ease notation below, please note that we have enforced this condition in the definition (4.7). Additionally, for all $s = g(\gamma, \sigma) \in \Gamma_L$ we have

$$\phi_2 = \sigma. \tag{4.8}$$

Remark 5. *When there can be no confusion, we may employ a slight abuse of notation for the sake of clarity. For example, strictly speaking we have* $f_2(s) = \sigma(s)$*, for all* $s \in \Gamma_L$*, but writing* $\sigma = \sigma(s)$*, we formally obtain* $\phi_2 = f_2(s) = \sigma(s) = \sigma$*. Indeed,* (4.8) *suffers no loss in meaning.*



Second, elements of Γ_L possess an agreeable quantile function Q, given by

$$Q(p, s) = \sigma \Big(\frac{1 - p}{p} \Big)^{-\frac{1}{\gamma}},$$

$p \in [0,1], (\gamma, \sigma) \in \mathbb{R}^2_L$. Thus, we obtain the remaining quantile features

$$\begin{cases} \phi_1 = \sigma 3^{-\frac{1}{\gamma}} \\ \phi_3 = \sigma 3^{\frac{1}{\gamma}}. \end{cases} \tag{4.9}$$

Finally, ϕ_4 becomes

$$\begin{aligned} \phi_4(s_{\gamma,\sigma}) &= \int_{\mathcal{N}} |\eta - \sigma| \frac{\gamma \sigma^{-\gamma} \eta^{\gamma-1}}{(1+(\eta/\sigma)^\gamma)^2}\, d\eta \\ &= \sigma\left(1 - \frac{\underline{\eta}^\gamma}{\sigma^\gamma + \underline{\eta}^\gamma} - \frac{\overline{\eta}^\gamma}{\sigma^\gamma + \overline{\eta}^\gamma}\right) - \int_{\underline{\eta}}^{\sigma} \frac{\gamma(\eta/\sigma)^\gamma}{(1+(\eta/\sigma)^\gamma)^2}\, d\eta + \int_{\sigma}^{\overline{\eta}} \frac{\gamma(\eta/\sigma)^\gamma}{(1+(\eta/\sigma)^\gamma)^2}\, d\eta. \end{aligned} \tag{4.10}$$

The expression in (4.10) involves elliptic integrals and admits a closed form solution in terms of special functions, namely Gauss's hypergeometric function ${}_2F_1$, as shown in Figure 4.2. This computational clemency enables an efficient disaggregation schema.

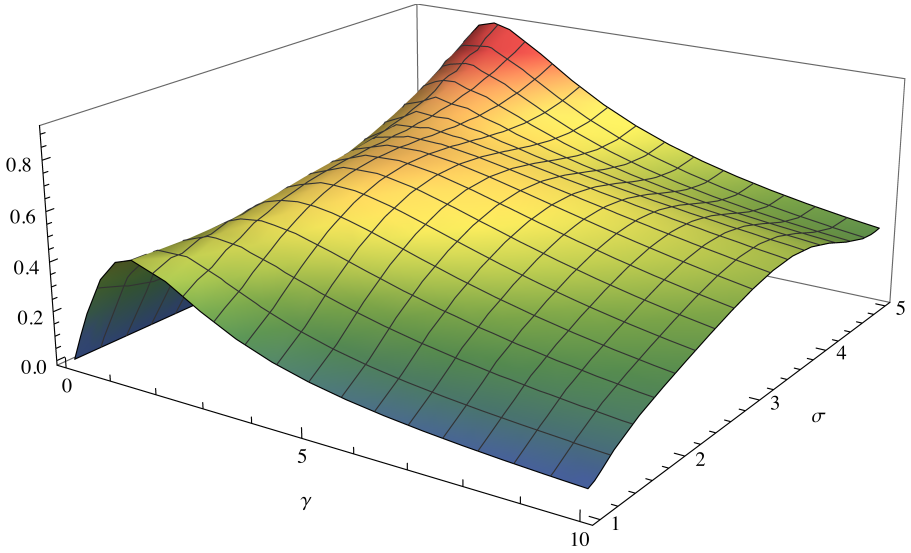


Figure 4.2: The mean median-deviation $\phi_4(s_{\gamma,\sigma})$, given in (4.10), over a subset of the log-logistic family given by $\{s_{\gamma,\sigma} \in \Gamma_L \mid (\gamma, \sigma) \in (0, 10] \times [1, 5]\}$.

Definition 25 (Feature Disaggregation Mapping)**.** *Given any (feature extraction) mapping* $\tilde{f} : X \to \mathbb{F}^X \subseteq \mathbb{F}$ *define the* feature disaggregation mapping $D(\tilde{f}, \cdot) : \mathbb{F}^X \to \tilde{f}^{-1}(\mathbb{F}^X) \subseteq X$*, where* $\tilde{f}^{-1}(\mathbb{F}^X)$ *denotes the preimage of* $\mathbb{F}^X$ *under* $\tilde{f}$*. For any feature vector* $\phi \in \mathbb{F}$*,* $D(\tilde{f}, \phi)$ *is given by the preimage of the least-squares projection of* ϕ *onto*

the image of $\tilde{f}$. That is,

$$D(\tilde{f}, \phi) := \underset{x \in \tilde{f}^{-1}(\mathbb{F}^X)}{\operatorname{argmin}} \left\| \phi - \tilde{f}(x) \right\|_2^2 . \tag{4.11}$$

Example 29 (Disaggregation into Distribution Family)**.** *Suppose $\tilde{f}_L = f|_\Gamma$, the restriction of f to an arbitrary family of distributions Γ. Then $X = \Gamma$, $\mathbb{F}^X = \mathbb{F}$, and for any $\phi \in \mathbb{F}$, we have the disaggregate state*

$$\tilde{s} = D(f|_\Gamma, \phi).$$

Example 30 (Parameterized Log-logistic Disaggregation)**.** *Let $\tilde{f} = f \circ g$. Then $X = \mathbb{R}^2_L$, $\mathbb{F}^X = \mathbb{F}$, and for any $\phi \in \mathbb{F}$, we have the log-logistic disaggregate state $\tilde{s} = g(\gamma^*, \sigma^*)$, where (γ^*, σ^*) are given by*

$$(\gamma^*, \sigma^*) = D(f \circ g, \phi). \tag{4.12}$$

Example 31 (Disaggregation with Higher-Order Aggregation)**.** *Consider the case of higher-order aggregation on features by, e.g., a course grid approximation. Letting the projection onto the grid be denoted by $h : \mathbb{F} \to \mathbb{F}^\Delta \subseteq \mathbb{F}$, we have $\tilde{f} = h \circ f \circ g$, $X = \mathbb{R}^2_L$, and $\mathbb{F}^X = \mathbb{F}^\Delta$. Proceeding just as before, we now obtain the log-logistic disaggregate state $\tilde{s} = g(\gamma^*, \sigma^*)$, where (γ^*, σ^*) are given by*

$$(\gamma^*, \sigma^*) = D(h \circ f \circ g, \phi). \tag{4.13}$$

Alternatively, consider two stages:

1. *$\tilde{f} = f \circ g$, $X = \mathbb{R}^2_L$, and $\mathbb{F}^X = \mathbb{F}$*
2. *$\tilde{f} = h$, $X = \mathbb{F}$, and $\mathbb{F}^X = \mathbb{F}^\Delta$*

$$(\gamma^*, \sigma^*) = D(f \circ g, \phi); \tag{4.14}$$

$$\phi^* = D(h, \phi^\delta). \tag{4.15}$$

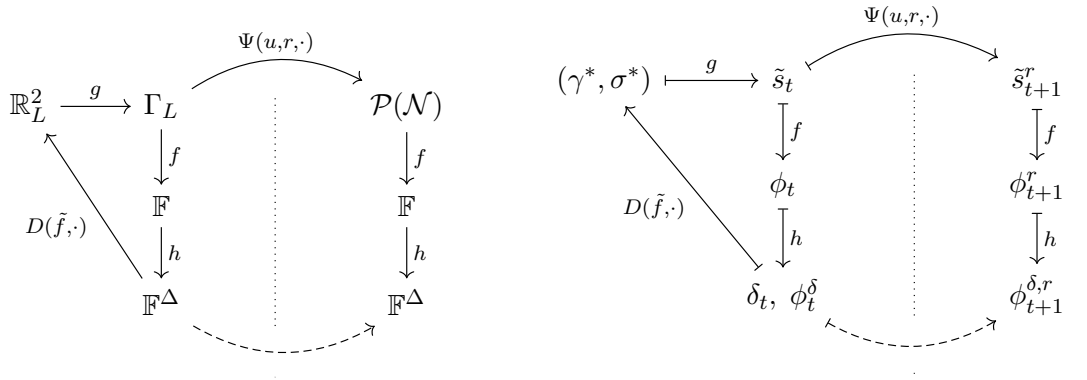
Explicitly, except for ϕ_4 given in (4.10), the error sum of squares in (4.12) has the form

$$\left(\delta_1 - \sigma 3^{-\frac{1}{\gamma}}\right)^2 + \left(\delta_2 - \sigma\right)^2 + \left(\delta_3 - \sigma\left(\frac{1}{3}\right)^{-\frac{1}{\gamma}}\right)^2 + \left(\delta_4 - f_4 \circ g(\gamma, \sigma)\right)^2. \tag{4.16}$$

4.5 Feature Aggregation Mappings

To this point, we have used projection methods to recast the original MDP involving an infinite-dimensional state space of probability distributions to an MDP involving a finite-dimensional vector space. However, in order to conduct the backward induction algorithm in this problem, we require a finite feature space. To this end, we take the standard approach of coarse grid approximation over the feature space. Again, we implement both projection and coarse grid approaches, rather than simply a coarse grid over $\mathcal{S}$, because efficient characterization of Bayesian orbits for logistic models, and non–conjugate likelihood models generally, is exceedingly challenging. Moreover, the observed orbits in any particular problem instance constitute an indefinitely minuscule subset of all possible orbits.

To formalize the coarse grid, let Δ denote discretization of feature space $\mathbb{F}$, with generic element $\delta \in \Delta$. We make the following definitions.



(a) Feature transitions via Log-logistic disaggregation $D(h \circ f \circ g, \cdot)$.

(b) The same feature transitions, illustrating mappings among elements.

Definition 26 (Feature Space Discretization)**.** *For a choice of M feature vectors $\phi^k \in \mathbb{F}$, $k = 1, 2, \cdots, M$, define $\Delta \subset \mathbb{F}$ as the set of these vectors. That is,*

$$\Delta := \Big\{ \delta = \phi^k,\ k = 1, 2, \cdots, M \Big\}. \tag{4.17}$$

The $\{\phi^k\}$ constituting the coarse grid may be chosen in various ways, e.g. with even spacing between adjacent elements, or a rescaled variant thereof. Denote by $\mathbb{F}^\Delta$ the

Table 4.1: Collection of the meaning and notation for each of the central mappings in the proposed ADP schema.

Mapping	**Notation**	**Usage**	**Ref.**
Feature extraction	$f : \mathcal{P}(\mathcal{N}) \to \mathbb{F}$	$\phi = f(s)$	(4.5)
Coarse grid projection	$h : \mathbb{F} \to \mathbb{F}^{\Delta} \subset \mathbb{F}$	$\phi^{\delta} = h(\phi)$	(4.19)
Feature disaggregation	$D(\tilde{f},\cdot) : \mathbb{F} \to \mathbb{R}^2_L$	$(\gamma^*, \sigma^*) = D(\tilde{f}, \phi)$	(4.16)
Log-logistic param'zn	$g : \mathbb{R}^2_L \to \Gamma_L$	$\tilde{s} = g(\gamma, \sigma)$	(4.7)

inclusion $\Delta \hookrightarrow \mathbb{F}$*, so that*

$$\mathbb{F}^{\Delta} := \Big\{ \phi^{\delta} = \iota(\delta) \in \mathbb{F} \;\Big|\; \delta \in \Delta \Big\}. \tag{4.18}$$

Note that this is merely a formal refinement between elements δ *of the coarse grid* Δ*, and the corresponding feature vectors* ϕ^{δ} *in feature vector space* $\mathbb{F}$*.*

Definition 27 (Feature Aggregation Mapping)**.** *Let* $h : \mathbb{F} \to \mathbb{F}^{\Delta}$ *denote projection onto the coarse grid embedding. That is, for all* $\phi \in \mathbb{F}$*,*

$$h(\phi) := \underset{\phi^{\delta} \in \mathbb{F}^{\Delta}}{\operatorname{argmin}} \; \Big\| \phi^{\delta} - \phi \Big\|_2^2. \tag{4.19}$$

4.6 Approximate Dynamic Programming Equations

Thus armed with the above approximation schema, we can now formulate ADP equations, analogous to the risk–neutral DP equations (3.9)–(3.10) and the risk–averse DP equations (3.16)–(3.17), or in the special case of binary response (3.20)–(3.21). One need not look long at the schema diagrams in Figures 4.3a and 4.3b to imagine that the formal ADP equations would appear rather complicated. We therefore state them here in their simplest, most intelligible form first and subsequently state the explicit form.

For all $\phi^{\delta} \in \mathbb{F}^{\Delta}$, and all $t = 1, 2, \cdots, T-1, T$,

$$v_{_T}(\phi^{\delta}) = \phi^{\delta}_4, \tag{4.20}$$

$$v_t(\phi^{\delta}) = \min_{u \in \mathcal{U}} \Big\{ \bar{c}(u; \tilde{s}(\phi^{\delta})) + \sigma^r_t\big(\tilde{s}(\phi^{\delta}), Z(\cdot\,; \tilde{s}(\phi^{\delta}), u), v_{t+1}\big(\Psi(\cdot\,; \tilde{s}(\phi^{\delta}), u)\big)\big) \Big\}. \tag{4.21}$$

Explicitly, the equations become, for all $\phi^\delta \in \mathbb{F}^\Delta$, and all $t = 1, 2, \cdots, T-1, T$,

$$v_{_T}(\phi^\delta) = \phi^\delta_4, \tag{4.22}$$

$$\begin{aligned} v_t(\phi^\delta) = \min_{u \in \mathcal{U}} \Bigg\{ & \bar{c}\Big(u; g \circ D(h \circ f \circ g, \phi^\delta)\Big) + \\ & + \sigma^r_t \Bigg(g \circ D(h \circ f \circ g, \phi^\delta), Z\Big(\cdot\,; g \circ D(h \circ f \circ g, \phi^\delta), u\Big), \ \ldots \\ & \ldots \ v_{t+1}\Big(h \circ f \circ \Psi(\cdot\,; g \circ D(h \circ f \circ g, \phi^\delta), u)\Big)\Bigg)\Bigg\}. \end{aligned} \tag{4.23}$$

Solving these ADP equations via backward induction provides an optimal policy $\boldsymbol{\pi}^* = (\pi^*_1, \pi^*_2, \cdots, \pi^*_{_{T-1}}, \pi^*_{_T})$, where each π^*_t may be viewed as a lookup table of optimal controls with the same dimensions as $\mathbb{F}^\Delta$. We note, however, that in the optimal learning setting, the optimal policy $\boldsymbol{\pi}^*$ is excessive, in the sense that very few approximate states are revisited over time, until convergence begins to occur. At the point of convergence, however, the optimal policy correspondingly converges to a stationary policy. Moreover, given initial conditions (s_1, u_1), only a very small number of states may be visited in the early stages. On the other hand, we cannot dispense with the non–stationarity of the policy, as this is the only way in which the optimal policy may treat the collisions that occur in the middle stages.

Thus, we see that in the optimal learning setting, any approximation schema will be excessive, in the above sense. This discussion motivates, in part, consideration of more simple approximation schema, especially those that are adaptive, in some sense. This naturally engenders lookahead policies (which we considered in the previous section) and methods of approximation in policy space (which we do not consider in this dissertation).

Chapter 5
Applications and Computational Experiments

> *As long as a branch of knowledge offers an abundance of problems, it is full of vitality.*
>
> David Hilbert, *1862–1943*

5.1 An Initial Comparative Study

In this section, we aim to provide a taste of the effect of risk measures in the optimal learning context. Specifically, we try to give a sense for the robustness engendered by risk measures, in the sense that they determine belief orbits that are less swayed by random perturbations. To this end, we conduct a simple simulation described in detail below. Before wading into the details, the essence of the experiment may be characterized thus: Suppose various lookahead policies all observe the exact same response data. How would each policy react to the same perturbation of one random response datum in each sample?

5.1.1 Simulation setup

Consider the finite-horizon problem introduced above, beginning from a uniform prior s_1 and initial control $\pi_1 \equiv u_1 = 0$, with horizon $T = 30$, and the three lookahead policy policies: vanilla expected-value (EV) policy, the knowledge gradient (KG) policy (3.24), and the robust-response (RR) policy (3.29).

Over $N = 100$ simulations, we generate a collection of random response orbits $\{\boldsymbol{r}^i = (r_1^i, r_2^i, \cdots, r_{T-1}^i, r_T^i)\}_{i=1}^N$, $r_t \in \{0, 1\}$, $\forall\, t$, by inverse transform sampling of the

true underlying logistic model of a greedy Bayesian policy with respect to a uniform distribution on the unit interval. That is, we set $r_t = 1$ when $x \geq \Phi^*(\hat{\mu}^1(s_t))$, and $r_t = 0$ otherwise, where x is sampled uniformly on $[0, 1]$ and $\hat{\mu}^1(s_t)$ is the median of the usual belief state s_t.

For each simulate we compute the belief orbit under policy $\boldsymbol{\pi} = (u_1, \pi_2, \pi_3, \cdots, \pi_{_T})$, where letting $s_t^i := \Psi(r_{t-1}^i; s_{t-1}, \pi_{t-1})$, $t = 2, 3, \cdots, T$, we denote the belief orbit by

$$\left\{ \boldsymbol{\Psi}^{\boldsymbol{\pi}}(\boldsymbol{r}^i) = (s_1, s_2^i, \cdots, s_{_{T-1}}^i, s_{_T}^i) \right\}_{i=1}^N.$$

For each policy $\boldsymbol{\pi} \in \{\boldsymbol{\pi}^{EV}, \boldsymbol{\pi}^{KG}, \boldsymbol{\pi}^{RR}\}$, we thus obtain the N belief orbits

$$\left\{ \boldsymbol{\Psi}^{EV}(\boldsymbol{r}^i), \boldsymbol{\Psi}^{KG}(\boldsymbol{r}^i), \boldsymbol{\Psi}^{RR}(\boldsymbol{r}^i) \right\}_{i=1}^N,$$

each generated from the same response sequence.

5.1.2 Perturbation Procedure

Then, for each simulate $i = 1, \cdots, N$, we pseudorandomly select $\tau(i) \in \{1, 2, \cdots, T-1, T\}$ and perturb response $r_{\tau(i)}^i$ by setting it to its complement in $\{0, 1\}$. Formally, denoting the complement of $r_{\tau(i)}^i$ by $\breve{r}_{\tau(i)}^i$, we set

$$r_{\tau(i)}^i \mapsto \breve{r}_{\tau(i)}^i.$$

For each policy and simulate i, the perturbed response generates a new orbit beginning from time $\tau(i)$. For each policy $\boldsymbol{\pi}$ we denote the perturbed orbit by

$$\left\{ \boldsymbol{\Psi}^{\boldsymbol{\pi}}(\breve{\boldsymbol{r}}^i) = (s_1, s_2^i, \cdots, \breve{s}_{\tau(i)}^i, \cdots, \breve{s}_{_{T-1}}^i, \breve{s}_{_T}^i) \right\}_{i=1}^N.$$

For each policy $\boldsymbol{\pi} \in \{\boldsymbol{\pi}^{EV}, \boldsymbol{\pi}^{KG}, \boldsymbol{\pi}^{RR}\}$, we thus obtain the N perturbed belief orbits

$$\left\{ \boldsymbol{\Psi}^{EV}(\breve{\boldsymbol{r}}^i), \boldsymbol{\Psi}^{KG}(\breve{\boldsymbol{r}}^i), \boldsymbol{\Psi}^{RR}(\breve{\boldsymbol{r}}^i) \right\}_{i=1}^N,$$

each generated from the same perturbed response sequence.

5.1.3 Comparison Metric

Finally, we measure the extent to which each policy is perturbed by computing the stagewise total variation norm. Specifically, for each policy $\boldsymbol{\pi}$ and simulate i, we compute the total variation norm $\boldsymbol{\delta}^{\boldsymbol{\pi},i}$ of each belief state in the orbit. Specifically, we

compute $\boldsymbol{\delta}^{\boldsymbol{\pi},i} = \left\{\delta_t^{\boldsymbol{\pi},i}\right\}_{t=1}^{T}$, where

$$\begin{aligned}\left\{\delta_t^{\boldsymbol{\pi},i}\right\}_{t=1}^{T} &\triangleq \left\{\max_{\eta\in\mathcal{N}}\left|\boldsymbol{\Psi}_t^{\boldsymbol{\pi}}(\boldsymbol{r}^i) - \boldsymbol{\Psi}_t^{\boldsymbol{\pi}}(\breve{\boldsymbol{r}}^i)\right|\right\}_{t=1}^{T} \\ &= \left\{0, 0, \cdots, \max_{\eta\in\mathcal{N}}\left|s^i_{\tau(i)} - \breve{s}^i_{\tau(i)}\right|, \cdots, \max_{\eta\in\mathcal{N}}\left|s^i_{T-1} - \breve{s}^i_{T-1}\right|, \max_{\eta\in\mathcal{N}}\left|s^i_{T} - \breve{s}^i_{T}\right|\right\}_{t=1}^{T}.\end{aligned}$$

We may therefore more succinctly write this as

$$\boldsymbol{\delta}^{\boldsymbol{\pi},i} = \left\{0, 0, \cdots, \delta^{\boldsymbol{\pi},i}_{\tau(i)}, \cdots, \delta^{\boldsymbol{\pi},i}_{T-1}, \delta^{\boldsymbol{\pi},i}_{T}\right\}. \tag{5.1}$$

For each policy $\boldsymbol{\pi} \in \{\boldsymbol{\pi}^{EV}, \boldsymbol{\pi}^{KG}, \boldsymbol{\pi}^{RR}\}$, we thus obtain the N series of T total variation norms:

$$\left\{\boldsymbol{\delta}^{EV,i}, \boldsymbol{\delta}^{KG,i}, \boldsymbol{\delta}^{RR,i}\right\}_{i=1}^{N}. \tag{5.2}$$

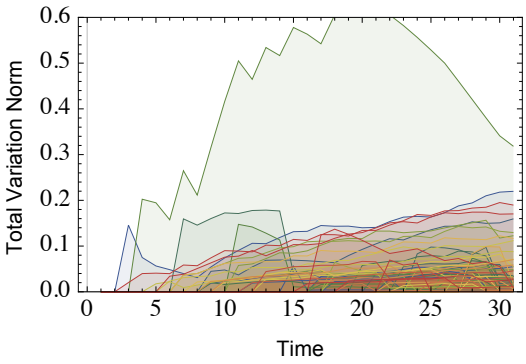
The results of this simple simulation study already indicate the effect of the robust-response policy in this setting. In particular, Section 5.1.3 demonstrates that the RR lookahead policy is more stable in view of potentially spurious information. As we will see in significantly more detail in Chapter 5, this hallmark property of risk aversion plays an even more profitable role in the optimal learning setting, owing to the fundamentally self-referential nature of the problem. In this sense, risk–averse policies, surprisingly, do not shoulder the burden of the proverbial risk–reward tradeoff. Rather, it would appear that over short time horizons risk–averse policies get to garner knowledge without the usual regret.

Optimal learning as a tool is applicable in a large and diverse collection of problem classes. In this Section, we focus on three case studies in the design of clinical trials. We conclude with a discussion of future applications in this domain.

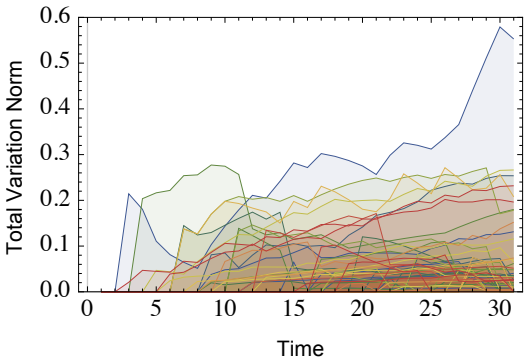
Arguably, the usual perverse incentives exist in medicine as a discipline, granted as in most professional industries, for treatments to gravitate naturally toward symbiotic agency from relatively passive fiduciary auspices.[1] As a discipline, medicine has moved,

[1] As an example from the financial industry, stock brokers are notoriously paid by commission *on transaction*, rather than on profits; the perverse incentives are now widely infamous. Consider, then, the medical insurance model that divulges capitation payments when medical treatment is *not* required, in stark contrast to the usual medical insurance model. The incentive structures are clearly dichotomous, although whether or not this would materially impact medical decisions is destined for debate—and far afield of the scope of this dissertation.

(a) Expected Value $\{\boldsymbol{\delta}^{EV,i}\}_{i=1}^{N}$



(b) Knowledge Gradient $\{\boldsymbol{\delta}^{KG,i}\}_{i=1}^{N}$



(c) Robust Response $\{\boldsymbol{\delta}^{RR,i}\}_{i=1}^{N}$

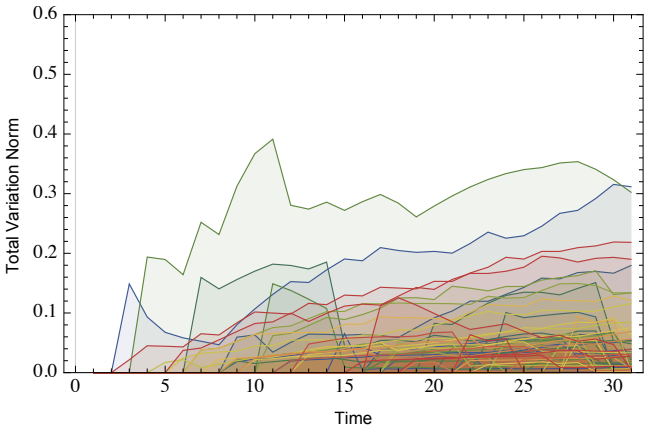


Figure 5.1: Stage-wise total variation norm $\{\boldsymbol{\delta}^{\boldsymbol{\pi},i}\}_{i=1}^{N}$ (5.1) of response-perturbed Bayesian orbits for the (a) expected-value, (b) knowledge gradient, and (c) robust-response policies (5.2). Note that the robust-response policy exhibits systemically smaller deviation over time.

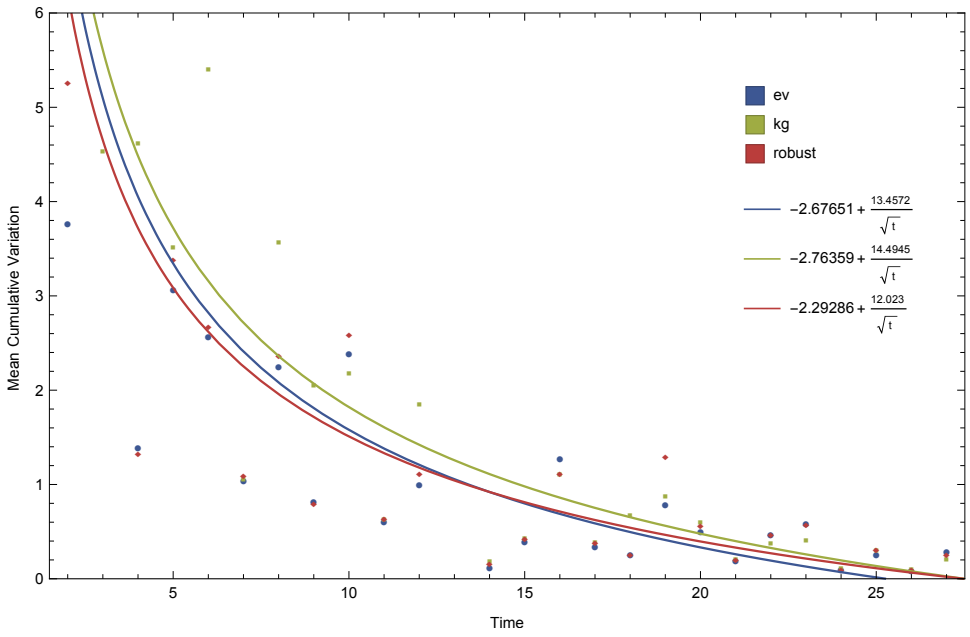


Figure 5.2: Mean total variation $\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\delta}^{\boldsymbol{\pi},i}$ of Bayesian orbits as a function of time.

likely by a natural industrial inertia, to prioritize quantity over quality, insofar as its operations may be better characterized as maximizing the number of patients routed to therapeutic endpoints subject to logistic, clinical and ethical constraints, rather than otherwise. Perhaps the most glaring example of which is the deficiency in iatrogenic considerations, which are inherently introspective. This is exacerbated to calamitous effect in cases of a nebulous or unavailable contrapositive, owing to the fallacy of conflating absence of evidence with evidence of absence. Such self–critical introspections are, at best, underappreciated and largely outside the paradigm of the trained physician.[2] These empirical observations belie the ostensibly scientific underpinnings of the discipline, wherein assumptions are continuously tested by failing to exclude them through experimentation.

This general discussion motivates a sound, practical methodology for iterating toward treatments of better and better quality while systematically tempering the experimentation in full view of iatrogenesis. To ground this abstract discussion, we focus on the case of cancer treatments. Cancer treatment, primarily chemotherapy and/or

[2]One need look no further than the recent alarm surrounding over-prescription (by 1-2 orders of magnitude!) of opiate pharmaceuticals in the United States.

radiation at the time of this writing, is perhaps the canonical instance of a fraught treatment regimen, wherein the overt toxic effects are justified only in the face of a prognosis beyond par. The legitimately grievous nature of cancer is only exacerbated by its colloquial perception, arguably leading to a "no holds barred" mentality and acceptance of adverse side effects, by way of a Hobson's choice if not otherwise. In this respect, it may be instructive to consider the extreme instance of childhood cancer, which has been proposed in this very context by Smith et al. [134] and recently seen a flurry of research in the literature cf. [10, 40, 153, 110].

Childhood cancer is universally held as tragic. It constitutes the case of maximal risk in all outcomes, which is to say the stakes could not be greater, forgoing the most life, in the worst case, or enduring the most time with long–term, generally accumulative side effects in the best case. By appeal to these sentiments, admittedly via some modicum of lurid Russell conjugation, the phenomenon of childhood cancer thus offers a uniquely potent vantage from which to motivate the general adoption of optimal learning in medicine. Along these lines, we offer some considerations as yet undetected in the medical literature.

Given the substantial development in cancer treatments since the 1970s, childhood cancer survivors constitute a novel, emerging cohort. That is, longitudinal observation of childhood cancer survivors is just beginning to emerge as a viable prospect. Compounding the paucity of feasible data, children are not "recruited" to trials, despite some evidence that "the practice is detrimental to their outcomes" [59, 108].

Moreover, and remarkably, childhood cancer survivors offer unparalleled comparison to the proper *contrapositives* of long–term iatrogenesis, which reasonably manifest as secondary and tertiary complications. That is, the peer (i.e., control) group to the cohort of childhood cancer survivors exhibits categorically lower all–cause morbidity, amounting to less statistical noise in the clinical study setting. Contrast this to adult cancer survivors, whose peer group exhibits statistically greater all–cause morbidity and mortality as a matter of course, by dint of natural aging if not otherwise. This obfuscates statistical analyses for the adult survivor cohort under general study settings and suggests childhood survivors offer the clearest picture of the long–term effects of

treatment. Of course, long–term effects are beyond the scope of the models presented in this dissertation; however, our intent with these considerations is merely to motivate endeavors to optimize the process, the prospects of long– and infinite–horizon optimization models notwithstanding. Against this backdrop, we therefore begin with a thorough study of the simplest conceivable problem along these lines.

5.2 Dose-finding in Clinical Trial Design

Phase I clinial trials are a crucial step in the development of treatment protocols utilizing novel pharmaceutical agents. Clinical research on human subjects is notoriously fraught with ethical concern, the prototypical example of which being early-stage trials for cytotoxic drugs. Generally, the goal of a Phase I trial is to determine a safe dosage of a pharmaceutical agent for subsequent use in a Phase II trial determining therapeutic dosage. However, in light of the very nature of cancer pathology, chemotherapy as a treatment has long been synonymous with cytotoxicity. Phase I cancer trials present a fundamentally more complicated scenario, embodying the quintessential safety versus treatment dilemma.

In this dissertation, we are motivated by the problem of optimal design of Phase I clinical trials for novel, cytotoxic pharmaceutical agents. The canonical example for such trials is a Phase I cancer trial of a novel chemotherapy drug. The primary goal of any Phase I trial is to assess the candidate drug with respect to safety by eliciting its toxicity as a function of dose. *Toxicity* is a binary classification (i.e., toxic, non–toxic), usually defined in terms of an ordinal toxicity grade, which is in turn defined with respect to a (set of) biomarker(s) related to undesirable side effects; the particular definitions and clinical implications of toxicity vary in relation to the drug under study. Various organizations, e.g. the National Cancer Institute and World Health Organization, disseminate clinical guidelines specifying toxicity grades for common toxicities [9, 142]. Prevalent dose-limiting toxicities in cancer trials include myelosuppression, neutropenia, anemia, nephrotoxicity, and hematologic toxicity.[3] We note, however,

[3]That is, toxicity of the bone marrow, white blood cells, red blood cells, and kidneys, respectively.

that reported toxicities are not fully standardized in the literature, and it is imperative that both the definition and interpretation of toxicities be scrupulously defined in each clinical application [30, 105]. Each trial participant generally tolerates a different dose level, and thus for an arbitrary participant drawn from the population the probability of toxicity as a function of dose is termed the *toxic response curve*[4].

For a general drug, the basic approach is to recruit volunteers for participation in the trial and then sequentially administer increasing dose levels until toxic or therapeutic response is observed. However, when a cytotoxic agent is under study, the setting materially changes for two important reasons, and this approach is rendered invalid. First, a cytotoxic agent increases the severity of toxic response, thus significantly increasing the risks to trial participants and engendering stringent ethical constraints on the eligibility of volunteers. Second, the mere candidacy of a potentially severely toxic drug in a human Phase I trial implies the relatively grievous prognosis of the disease(s) for which it is intended as treatment. Although perhaps more subtle, this is a pivotal point, revealing by mere participation in the trial, patients and their physicians implicitly evaluate the risks of the opportunity cost of treatment, as it were, as greater still. Hence, trial participants are in fact patients in need of treatment, rather than arbitrary, eligible volunteers, and we must hold their superlative treatment as a competing goal of the trial.

Thus, on the one hand, in its role as a scientific experiment, Phase I clinical trials for cytotoxic agents have the goal of eliciting the toxic dose-response curve to the benefit of potential future patients. Whereas, on the other hand, the trials must attain the best possible treatment for participating patients, in particular mitigating undue harm. Current trial designs attempt to balance these competing objectives by determining the *maximum tolerable dose* (MTD), the largest dose whose probability of toxicity within the patient population is a specified limit. Undoubtedly, the ethical considerations inherent in this dilemma run deep, however, equally clear is the fact that appropriately evaluating, balancing, and mitigating the inherent risks is of exceptional importance in

[4]The toxic response curve is interchangeably known as the tolerance distribution, toxicity distribution, dose-toxicity curve, and related variants.

such trials. Modern risk models described in Section 1.3 are particularly suited for this purpose, yet to our knowledge have largely not been used in the medical literature, or when used, employed only ad hoc rather than in the principled framework discussed above.

A notable body of literature exists in more technical disciplines, most notably statistics and operations research, of which medicine is a client. Seminal work on optimal sequential designs, including the original formulation of the MAB problem in the design of clinical trials, is largely due to Robbins [114]; see also Keifer and Wolfowitz [85], DeGroot [41] and Lai and Robbins [91]. Early work on this very topic[5] dates back also to Katehakis and Derman in [82], who formulated the problem as a multi-armed bandit (MAB) and showcased their methods in [84] for the efficient computation of optimal dosage policies within a conjugate Beta-Bernoulli model. The problem broad methodology of identifying the objective as the MTD and its association with the 1/3-quantile are due to Storer in [136], where in particular a non–parametric, heuristic approach is favored. Specifically, Storer demonstrated two-stage designs are more robust in small-sample settings, in that the maximum-likelihood estimation of the MTD has less overdose bias. Remarkably, it would appear that this result engendered the widespread adoption of the $3+3$ deigns and their variants, despite subsequent literature demonstrating benefits of more sophisticated statistical inference methods against the $3+3$. For example, the introduction of Bayesian sequential inference in the medical statistics literature was introduced the following year by O'Quigley et al. [103], and is known as the continual reassessment method (CRM). More recently, a body of work has emerged around demonstrating variants of the CRM approach and its application in particular settings; see, e.g., [93, 90] and the references therein.

We follow most closely the clinical modeling methodology presented in [14] and subsequently in [18]. We begin with a range $[u_{\min}, u_{\max}]$ of feasible dosages for the trial, as determined by clinicians from previous animal studies and clinical expertise.[6]

[5] In fact, the problem is cast in the converse: A treatment is either effective or not effective, rather than toxic or non–toxic. Clearly, the two are in tandem.

[6] We attempt to clarify this determination to some small measure below. In the process, we hope also to rectify, which is to say sidestep, a formal issue of singularity not explicitly treated in [14, 18].

It is understood that $u_{\min}$ is a conservative starting dose for the trial, and $u_{\max}$ is such that $u_{\min} \leq \text{MTD} \leq u_{\max}$ with a high degree of confidence. Thus, in a trial with T participants, the t^{th} participant, $t = 1, 2, \cdots, T$, would receive dose $u_t \in U$, and by convention $u_1 \triangleq u_{\min}$. Pursuant to administration of dose u_t, we observe the participant's response r_t as either *toxic* ($r_t = 1$) or *non–toxic* ($r_t = 0$).

We proceed with the "usual logistic model" [18] of toxic response as a function of dose, so that for all $t = 1, \cdots, T$,

$$\mathbb{P}(r_t = 1 \mid u_t) = \frac{1}{1 + e^{-(\alpha + \beta u_t)}}, \tag{5.3}$$

for canonical parameters $(\alpha, \beta) \in \mathbb{R}^2$. Recall from (2.13) that we assume $\alpha > 0$, so that (5.3) is monotonic increasing in dose u_t. A typical scenario is illustrated in Section 5.2.

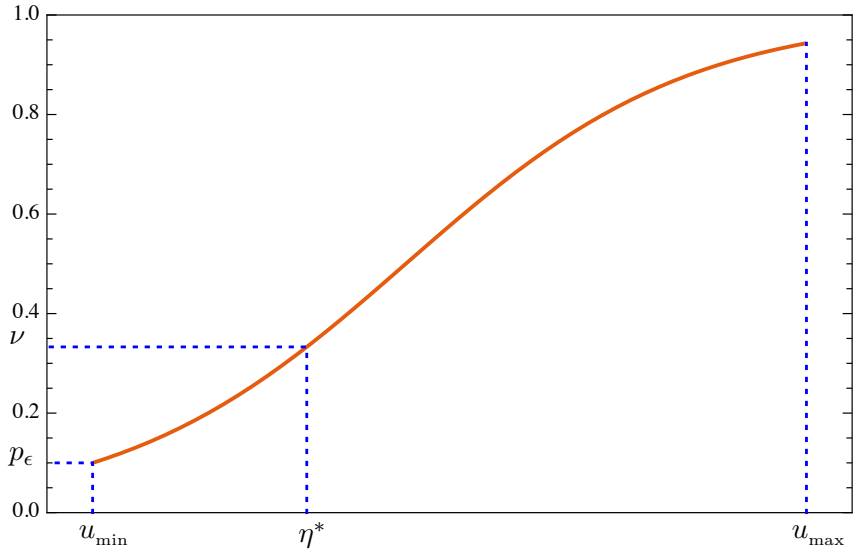


Figure 5.3: Probability of toxicity vs. dose, demonstrating the modeling methodology

Remark 6. *Although mathematically convenient, in that the discriminant in* (5.3) *is affine in* (α, β)*, these parameters lack an apparent clinical interpretation. In turn this obfuscates determination of a satisfactory prior distribution with sufficient confidence, especially for the uninitiated clinician. Indeed, clinical adoption of any such methodology depends critically on resolution this interpretability issue. We therefore seek a more interpretable (which is to say feasible, in this scenario) parameterization, albeit at the cost of some computational efficiency.*

We introduce the MTD as the principal model parameter, denoted by $\eta \in \mathcal{N} \triangleq [u_{\min}, u_{\max}]$. Formally, the MTD parameter is defined as the ν-quantile of the toxic response distribution; that is, the MTD is the dose for which the probability of toxic response is equal to a given probability ν. In practice, the choice of ν depends on the nature of toxicity in the trial, tending toward unity when toxicity is relatively mild and toward zero when relatively grievous. Originating in [136], the literature almost ubiquitously studies the case of $\nu = 1/3$, although one might argue that such an heuristic is widespread in practice [18]. In order to facilitate comparison, we continue this practice in our first case study below. Additionally, we conduct two experiments respectively prompting relatively small and large values of ν.

Additionally, we identify the second model parameter p_ϵ, defined as the probability of toxic response at the largest tolerable dose u_ϵ determined by previous clinical experience with the agent. Indeed, we assume the minimum feasible dose for the trial is chosen such that $u_{\min}$ differs from u_ϵ by some small quantity $\epsilon > 0$, understood to be on the order of one part in one hundred with respect to the length of the feasible dosage interval. That is, $u_{\min} = u_\epsilon + \epsilon$, for some $\epsilon \approx (u_{\max} - u_{\min})/100$.

However, in the experiments below, we simplify the two-parameter model in [14] by fixing $p_\epsilon \in [0, \nu)$ to a prescribed value, yielding a statistical model in terms of η alone. We find several reasons to implement this simplification: First, diligent examination of (5.5) reveals that relatively substantial simplification may be achieved by regarding p_ϵ as constant, as compared to other parameter η. Mitigation of the nonlinearity in (5.5) thus yields a significant reduction in computational complexity. Second, in an attempt to mitigate any undue advantage to Bayesian parametric methods when compared to the widely used non–parametric (which is to say heuristic) methods proposed by [136], it was pointed out in [14] that the proportion of patients overdosed in a Bayesian framework is maximized when p_ϵ is assumed known. Thus, such a scenario enables study of the worst-case performance, with respect to proportion of patients overdosed,

for Bayesian methods. From the two-parameter model (5.3), we therefore have

$$\begin{cases} p_\epsilon = \dfrac{1}{1+e^{-(\alpha+\beta u_\epsilon)}}, \\ \nu \ = \dfrac{1}{1+e^{-(\alpha+\beta\eta)}}, \end{cases} \implies \begin{cases} \alpha = \dfrac{\eta \log\left[\frac{p_\epsilon}{1-p_\epsilon}\right] - u_\epsilon \log\left[\frac{\nu}{1-\nu}\right]}{\eta - u_\epsilon}, \\ \beta = \dfrac{\log\left[\frac{\nu}{1-\nu}\right] - \log\left[\frac{p_\epsilon}{1-p_\epsilon}\right]}{\eta - u_\epsilon}. \end{cases} \tag{5.4}$$

Thus, the discriminant, φ of Section 3.1, in (5.3) is given by

$$\varphi(u,\eta) := \left(\frac{u-u_\epsilon}{\eta-u_\epsilon}\right)\log\left[\frac{\nu}{1-\nu}\right] - \left(\frac{u-\eta}{\eta-u_\epsilon}\right)\log\left[\frac{p_\epsilon}{1-p_\epsilon}\right], \tag{5.5}$$

where we reiterate that p_ϵ is viewed as a constant.

Finally, we note that without loss of generality, we may standardize the problem onto the unit interval via location-scale transformation. Specifically, we transform all dose variables via the mapping $x \mapsto (x-u_\epsilon)/(u_{\max}-u_\epsilon)$. This yields a nice simplification, wherein (5.5) simply becomes[7]

$$\varphi(u,\eta) := \log\left[\frac{p_\epsilon}{1-p_\epsilon}\right] + \frac{u}{\eta}\log\left[\frac{\nu\,(1-p_\epsilon)}{p_\epsilon(1-\nu)}\right], \tag{5.6}$$

the feasible dosage space $\mathcal{N} = [\epsilon, 1]$, and $\epsilon \approx 1/100$.

For each $r \in \mathcal{R} = \{0,1\}$ we thus consider the statistical model $\{\Phi(r;u,\eta),\ \forall\,(u,\eta) \in \mathcal{U}\times\mathcal{N}\}$, where

$$\Phi(r;u,\eta) := \begin{cases} \dfrac{1}{1+e^{-\varphi(u,\eta)}}, & r = 1, \\ \dfrac{1}{1+e^{\varphi(u,\eta)}}, & r = 0. \end{cases} \tag{5.7}$$

The standardized problem has the form

$$\frac{1}{1+\frac{1-p_\epsilon}{p_\epsilon}\left(\frac{(1-p_\epsilon)\nu}{p_\epsilon(1-\nu)}\right)^{-u/\eta}}\,. \tag{5.8}$$

We conduct several computational experiments to comprehensively evaluate performance for all methodologies under consideration. In the sequel, we consider several instances of the CTD optimal learning problem. Section 5.2.1 models a Phase I trial of

[7]For the sake of brevity, we make an innocuous abuse of notation by persisting with the nomenclature for all transformee variables.

5-fluorouracil, as previously studied in [14]. In Section 5.2.2, we consider a case of a very toxic drug, bleomycin, for treatment of germ cell tumors. In such cases, we additionally encounter a natural tendency toward clinical misspecification, in that the true MTD may tend to lie near the end of the feasible dosage range. Finally, in Section 5.2.3, we consider a case of high risk tolerance, insofar as the MTD quantile is chosen relatively large. This arises, in particular, in the face of dire prognosis, and we study the case of tandem dose escalation of etoposide and cyclophosphamide in a Phase II trial of the BEACOPP regimen for treatment of Hodgkin's lymphoma reported in [141, 55].

In all cases, we evaluate the prominent policies introduced above. Rather than evaluating policies in expectation, as is common in the optimal learning corpus[8], we evaluate policy performance *in distribution*. In particular, we evaluate distributions of the total regret, total cost, and the final recommendation $\hat{\eta}_T$, and compare policies in terms of common metrics, as well as dominance with respect to stochastic order on these distributions.

Moreover, in order to scrupulously evaluate policies for this problem, we compute the exact distributions for each policy, rather than approximations to these distributions via, for example, Markov Chain Monte Carlo (MCMC) methods. We shoulder this computational burden for the purpose of rigorously analyzing this initial work and guiding the future use of particular performance metrics, such as expected values, or more generally, (sets of) coherent risk measures. When focused on a particular metric, MCMC methods may then, of course, be used to more efficiently estimate the quantities of interest. However, the exact distributions are universal for the controlled Markov process, model class, and standardized problem instance, thus offering tremendous potential toward accelerating other numerical analyses.

5.2.1 5-Fluorouracil

We consider the design of a Phase I oncology trial to determine the MTD, with $\nu = 1/3$, of the antimetabolite 5-fluorouracil (5-FU) for the treatment of solid tumors in the colon

[8] We here use "optimal learning" broadly, to include reinforcement learning, sequential design theory, bandit problems, and the subset of MDPs concerned with HMM and POMDP modeling, etc.

via combination therapy with fixed levels of the agents leucovorin (20 $\mathrm{mg/m^2}$) and topotecan (0.5 $\mathrm{mg/m^2}$). Throughout the trial, toxic response constitutes observation of Grade IV hematologic or Grade III or IV nonhematologic toxicity within two weeks, as clinically defined by the National Cancer Institute Common Toxicity Criteria (CTC) [9].

Following the methodology of [14, 18], we first identify a domain $[u_{\min}, u_{\max}]$ of feasible dosages almost surely containing the MTD. Previous clinical studies of combination 5-FU and topotecan suggested a dose of $u_{\min} = 140\,\mathrm{mg/m^2}$ of 5-FU to be tolerable at a topotecan dosage of 0.5 $\mathrm{mg/m^2}$. Similarly, a previous trial of solely 5-FU concluded the MTD was 425 $\mathrm{mg/m^2}$, and therefore $u_{\max}$ was taken to be 425 $\mathrm{mg/m^2}$ for the combination trial since 5-FU has been empirically observed to be more toxic with topotecan. We consider $|\mathcal{U}| = 100$ dose levels, over the range indicated in Table 5.1. We formulate the logistic model (5.7) and together with a uniform prior distribution over $[u_{\min}, u_{\max}]$ for the MTD parameter η.

Table 5.1: Phase I dose escalation schema for antimetabolite 5-fluorouracil (5-FU) in combination therapy for treatment of solid tumors of the colon, adapted from [14]. Note: We consider $|\mathcal{U}| = 100$ dose levels, over the range 1–6 shown below.

5-FU	Dose Escalation ($\mathrm{mg/m^2}$)						
	Baseline	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Leucovorin	20						
Topotecan	0.5	Feasible Dosage Space					
5-Fluorouracil	125	140	150	175	200	225	425

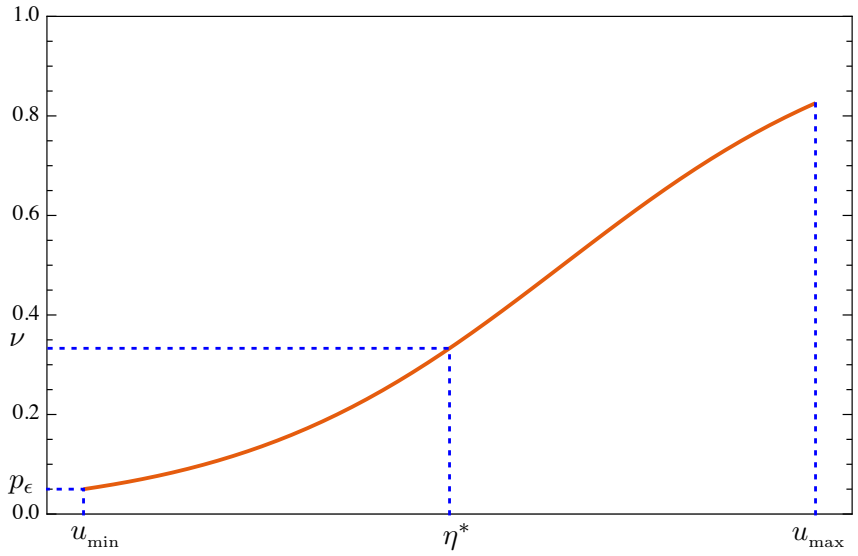


Figure 5.4: Probability of toxicity vs. dose in the 5-FU trial.

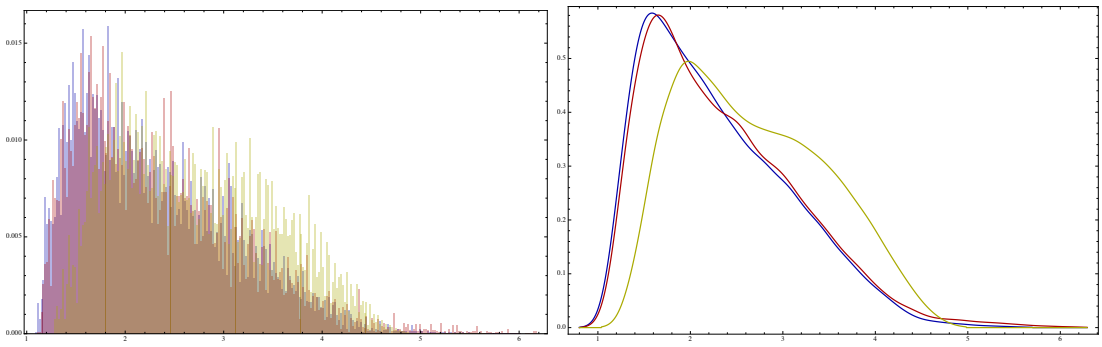


Figure 5.5: Distribution of cumulative cost $\sum_{t=1}^{T} c(\pi_t, \eta^*)$.

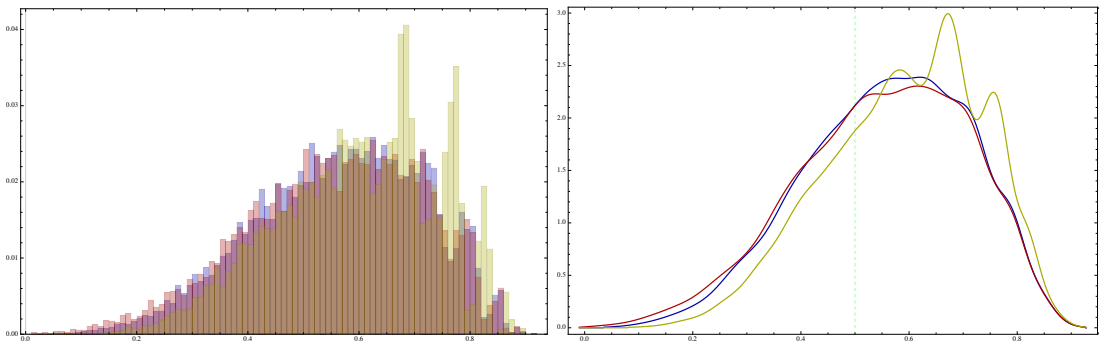


Figure 5.6: Distribution of final recommendation $\hat{\eta}_{T+1}$.

5.2.2 Bleomycin

As another example, we consider the design of a Phase I oncology trial to determine the MTD, of the polypeptide antibiotic antineoplastic agent bleomycin (B) for the treatment of germ-cell tumors (GCTs) via BEP combination therapy with fixed levels of the agents etoposide (100 mg/m^2) and cisplatin (20 mg/m^2). Throughout the trial, toxic response (clinically termed bleomycin pulmonary toxicity, or BPT) constitutes observation via thoracic computed tomography (CT) of any pulmonary fibroic changes, significant changes in pulmonary function test, and/or dyspnea commensurate with Grade III or IV pulmonary toxicity within two weeks, as clinically defined by the National Cancer Institute Common Toxicity Criteria (CTC) [9]; cf. [35, 133, 37, 10, 40, 153, 110, 54, 51, 52, 53].

The Phase I trial consists of $T = 13$ patients. We allow $|\mathcal{U}| = 100$ dose levels, with a minimum dose $u_{\min} = 5$ mg/m^2 and a maximum dose $u_{\max} = 35$ mg/m^2. Based on prior animal studies, it is estimated that $p_\epsilon = 1/100$ patients exhibit dose-limiting BPT at 5 mg/m^2. Bleomycin is known to be especially toxic, and so the MTD quantile is set to $\nu = 1/5$. The true, unknown MTD for the case study is $\eta^* = 14$ mg/m^2. These trial data are summarized in Table 5.2 and Figure 5.7.

Table 5.2: Phase II dose escalation schema of BEP combination therapy for treatment of advanced stage GCT in the Southeastern Cancer Study Group protocol, adapted from [54]. Note: We consider $|\mathcal{U}| = 100$ dose levels, over the range 1–6 shown below.

BEP	Dose Escalation (mg/m^2)						
	Baseline	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Etoposide	35						
Cisplatin	40	Feasible Dosage Space					
Bleomycin	5	10	15	20	25	30	35

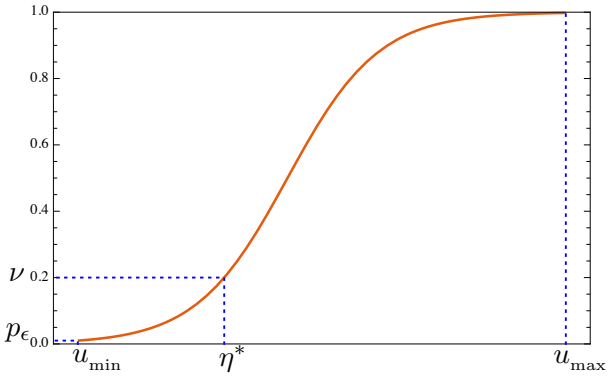


Figure 5.7: Probability of bleomycin pulmonary toxicity vs. dose in the BEP trial.

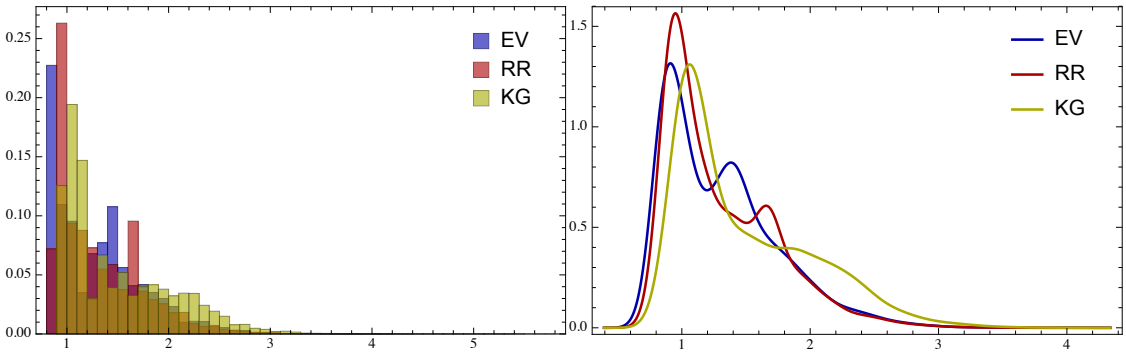


Figure 5.8: Distribution of cumulative cost $\sum_{t=1}^{T} c(\pi_t, \eta^*)$.

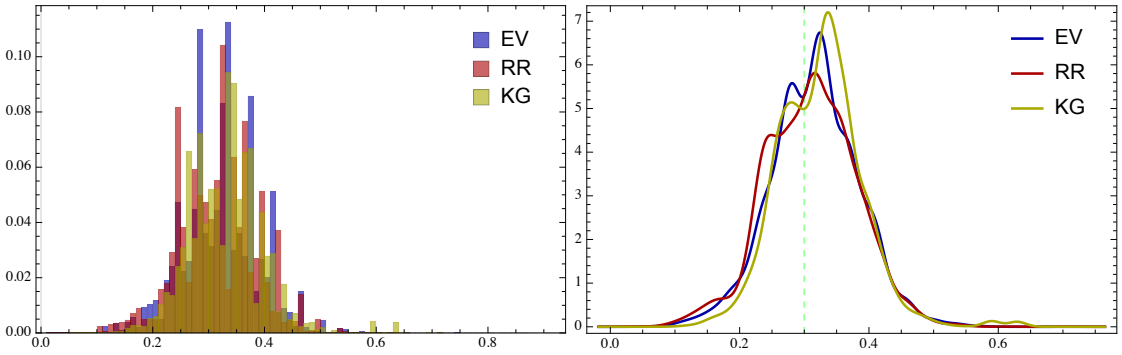
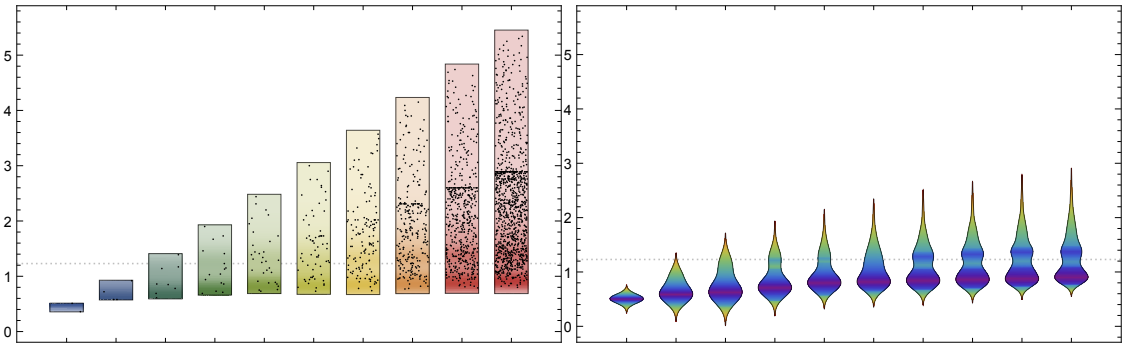
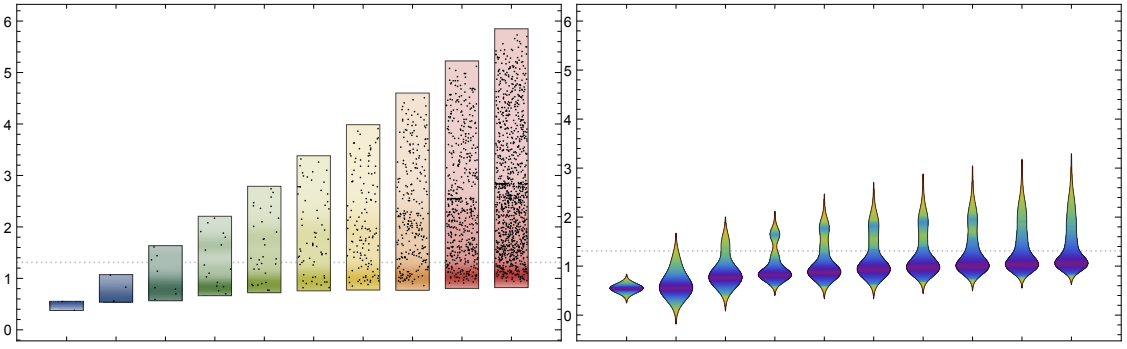


Figure 5.9: Distribution of trial MTD recommendation $\hat{\eta}_{T+1}$.

(a) Expected Value



(b) Knowledge Gradient



(c) Robust Response

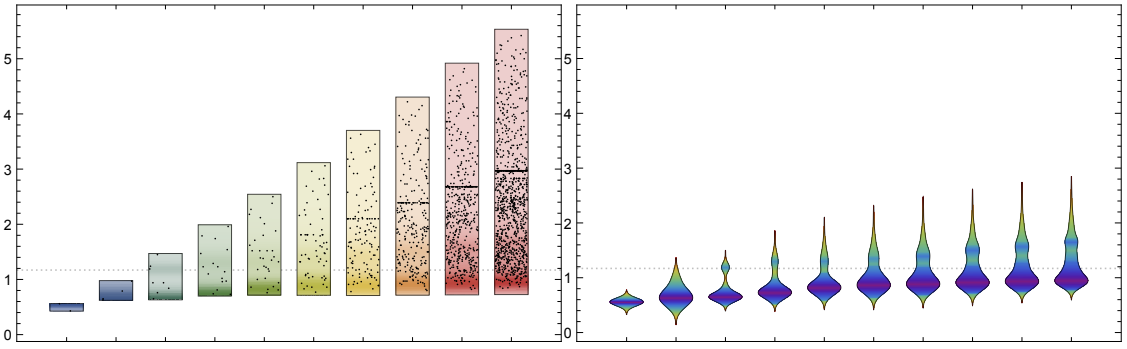
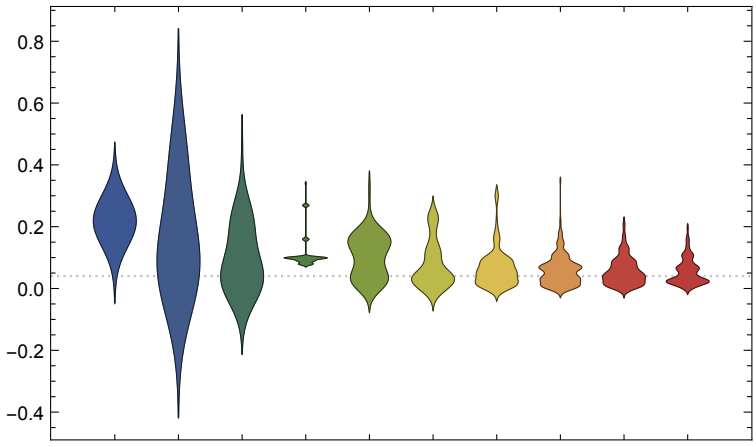


Figure 5.10: Stage-wise cumulative cost distributions in the BEP trial for (a) expected-value, (b) knowledge gradient, and (c) robust-response policies. Median of final-stage cost is shown dotted. Spectrum denotes stage $t = 1, 2, \cdots, T$; only 10 stages shown.

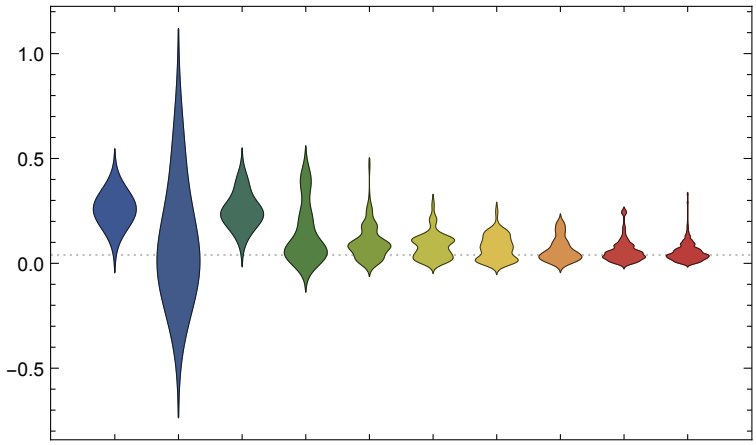
Table 5.3: Descriptive statistics for the stagewise distribution of cumulative cost $\sum_{\tau=1}^{t} c(\pi_\tau, \eta^*)$ in the BEP trial. We report the mean median deviation (MMD), root mean squared deviation (RMSD), median absolute deviation (MAD), and quantiles.

Patient t	$\boldsymbol{\pi}$	MMD_t^π	RMSD_t^π	MAD_t^π	$q_{.25}$	$q_{.5}$	$q_{.75}$
	RR	0.001	0.559	0.000	0.560	0.560	0.560
2	EV	0.002	0.509	0.000	0.510	0.510	0.510
	KG	0.002	0.548	0.000	0.550	0.550	0.550
	RR	0.062	0.704	0.000	0.630	0.630	0.630
3	EV	0.094	0.690	0.000	0.580	0.580	0.920
	KG	0.102	0.683	0.000	0.550	0.550	0.550
	RR	0.118	0.797	0.000	0.650	0.650	0.840
4	EV	0.163	0.809	0.000	0.610	0.610	1.140
	KG	0.146	0.945	0.000	0.790	0.790	0.790
	RR	0.143	0.908	0.000	0.730	0.730	1.020
5	EV	0.184	0.928	0.020	0.710	0.710	1.230
	KG	0.230	1.100	0.080	0.840	0.840	1.180
	RR	0.170	1.010	0.050	0.810	0.860	1.180
6	EV	0.212	1.030	0.150	0.750	0.880	1.250
	KG	0.281	1.210	0.050	0.870	0.920	1.360
	RR	0.222	1.100	0.110	0.860	0.880	1.320
7	EV	0.258	1.120	0.180	0.790	0.910	1.280
	KG	0.317	1.300	0.140	0.890	1.030	1.570
	RR	0.261	1.180	0.150	0.890	0.930	1.370
8	EV	0.291	1.190	0.220	0.810	0.960	1.330
	KG	0.352	1.380	0.130	0.940	1.050	1.670
	RR	0.290	1.240	0.170	0.910	1.030	1.450
9	EV	0.314	1.250	0.260	0.820	1.030	1.390
	KG	0.378	1.450	0.200	0.980	1.130	1.720
	RR	0.315	1.310	0.210	0.920	1.090	1.540
10	EV	0.331	1.320	0.290	0.880	1.120	1.460
	KG	0.400	1.520	0.270	1.010	1.230	1.780
	RR	0.333	1.370	0.240	0.940	1.170	1.620
11	EV	0.345	1.370	0.300	0.930	1.230	1.550
	KG	0.418	1.580	0.300	1.060	1.310	1.850

(a) Expected Value



(b) Knowledge Gradient



(c) Robust Response

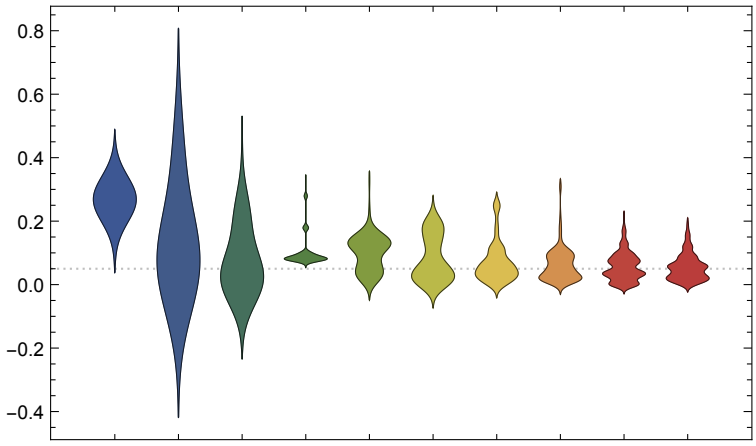
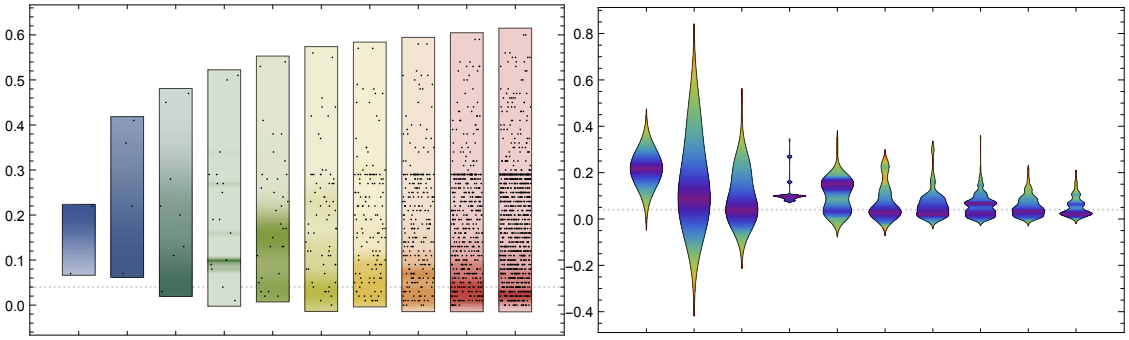
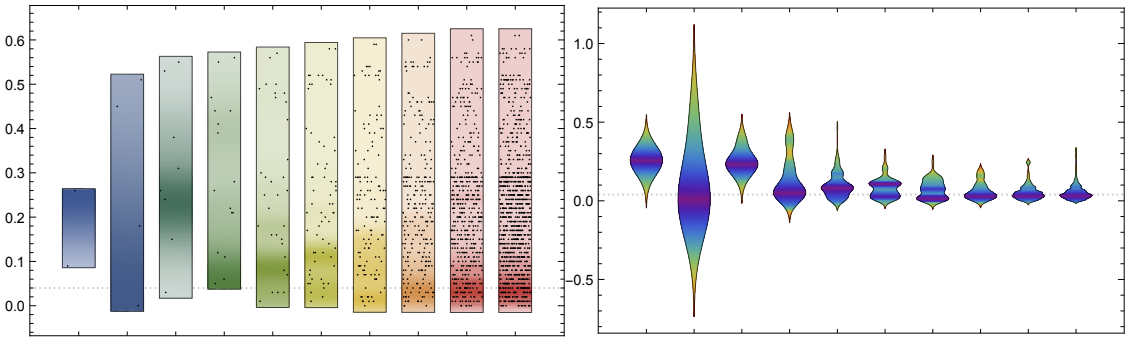


Figure 5.11: Stage-wise cost distributions in the BEP trial for (a) expected-value, (b) knowledge gradient, and (c) robust-response policies. Median of final-stage cost is shown dotted. Color spectrum denotes stage $t = 1, 2, \cdots, 10$.

(a) Expected Value



(b) Knowledge Gradient



(c) Robust Response

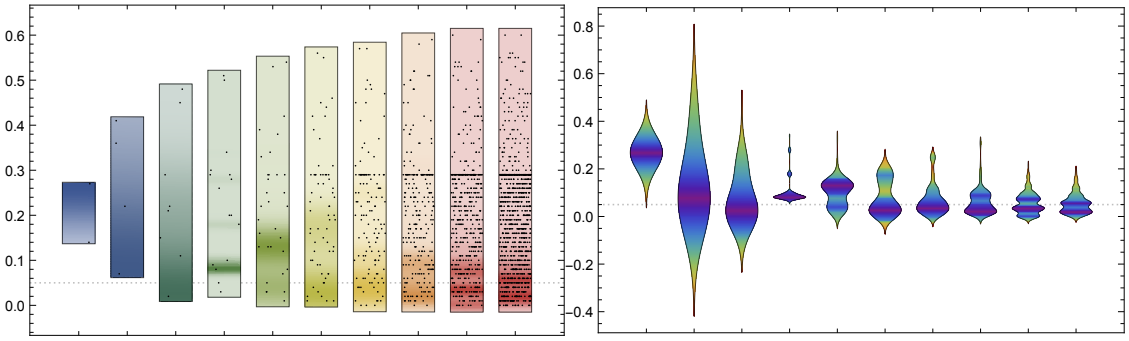
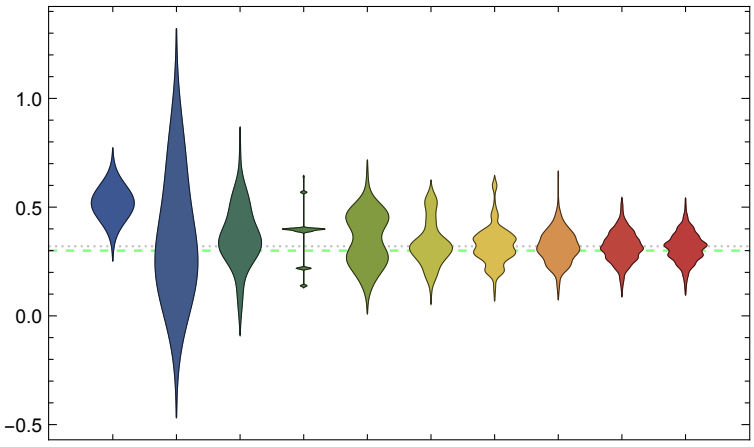


Figure 5.12: Stage-wise cost distributions in the BEP trial for (a) expected-value, (b) knowledge gradient, and (c) robust-response policies. Median of final-stage cost is shown dotted. Spectrum denotes stage $t = 1, 2, \cdots, T$; only 10 stages shown.

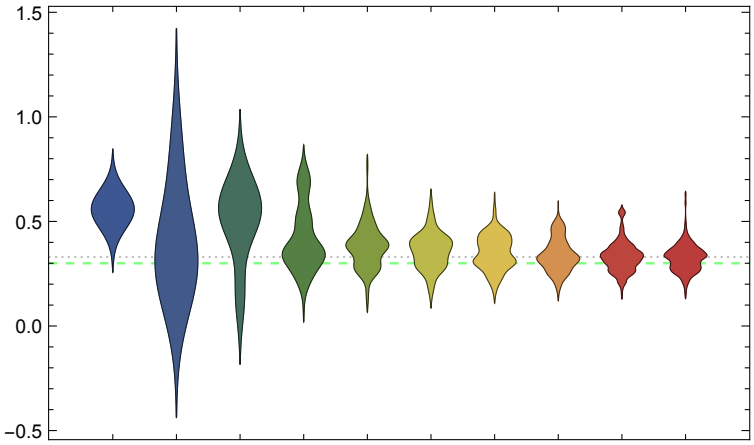
Table 5.4: Descriptive statistics for the stagewise distribution of cost $c(\boldsymbol{\pi}_t, \eta^*)$ in the BEP trial. We report the mean median deviation (MMD), root mean squared deviation (RMSD), median absolute deviation (MAD), and quantiles.

Patient t	$\boldsymbol{\pi}$	MMD_t^π	RMSD_t^π	MAD_t^π	$q_{.25}$	$q_{.5}$	$q_{.75}$
	RR	0.001	0.269	0.000	0.270	0.270	0.270
2	EV	0.002	0.219	0.000	0.220	0.220	0.220
	KG	0.002	0.259	0.000	0.260	0.260	0.260
	RR	0.064	0.188	0.000	0.070	0.070	0.070
3	EV	0.095	0.225	0.000	0.070	0.070	0.410
	KG	0.104	0.229	0.000	0.000	0.000	0.000
	RR	0.056	0.121	0.000	0.020	0.020	0.210
4	EV	0.069	0.139	0.000	0.030	0.030	0.220
	KG	0.044	0.264	0.000	0.240	0.240	0.240
	RR	0.025	0.118	0.000	0.080	0.080	0.100
5	EV	0.026	0.128	0.010	0.090	0.100	0.100
	KG	0.090	0.191	0.010	0.050	0.060	0.210
	RR	0.043	0.113	0.030	0.030	0.130	0.130
6	EV	0.059	0.124	0.060	0.020	0.110	0.170
	KG	0.053	0.127	0.050	0.030	0.080	0.110
	RR	0.058	0.108	0.040	0.020	0.050	0.140
7	EV	0.055	0.107	0.040	0.030	0.040	0.120
	KG	0.050	0.105	0.040	0.040	0.070	0.110
	RR	0.044	0.095	0.030	0.030	0.060	0.100
8	EV	0.041	0.090	0.040	0.020	0.050	0.090
	KG	0.049	0.097	0.050	0.020	0.070	0.110
	RR	0.037	0.080	0.030	0.020	0.050	0.090
9	EV	0.037	0.078	0.030	0.030	0.050	0.080
	KG	0.043	0.088	0.030	0.020	0.050	0.100
	RR	0.036	0.074	0.030	0.030	0.040	0.080
10	EV	0.034	0.072	0.030	0.020	0.050	0.080
	KG	0.036	0.080	0.030	0.030	0.050	0.080
	RR	0.032	0.070	0.030	0.020	0.050	0.080
11	EV	0.033	0.069	0.020	0.020	0.040	0.070
	KG	0.030	0.070	0.020	0.030	0.040	0.070

(a) Expected Value



(b) Knowledge Gradient



(c) Robust Response

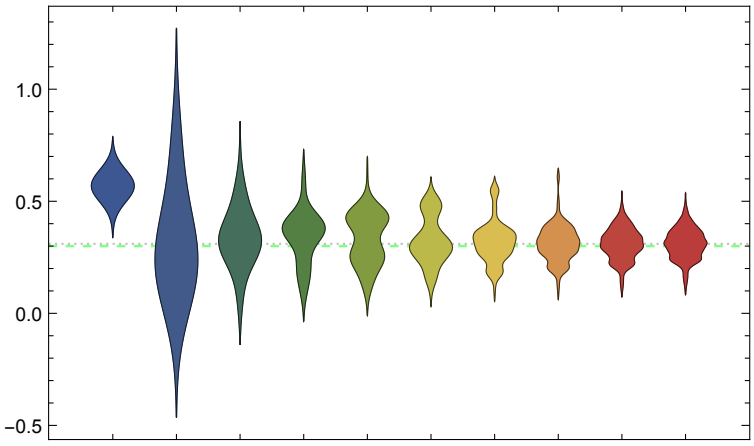
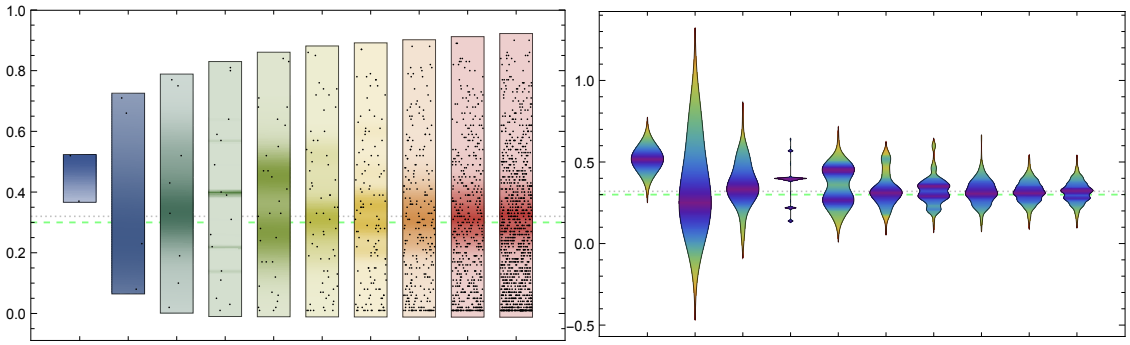
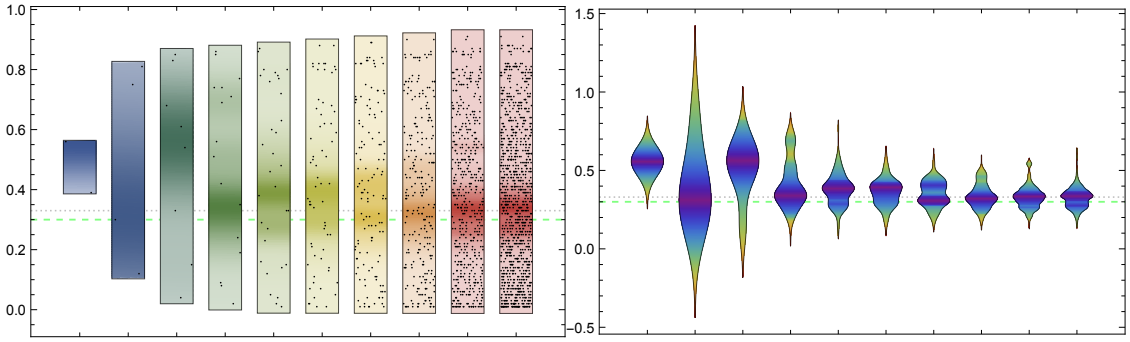


Figure 5.13: Stage-wise dosage distributions in the BEP trial for (a) expected-value, (b) knowledge gradient, and (c) robust-response policies. Final-stage median $\hat{\mu}_{_T}$ and actual MTD η^* are shown dotted. Color spectrum denotes stage $t = 1, 2, \cdots, 10$.

(a) Expected Value



(b) Knowledge Gradient



(c) Robust Response

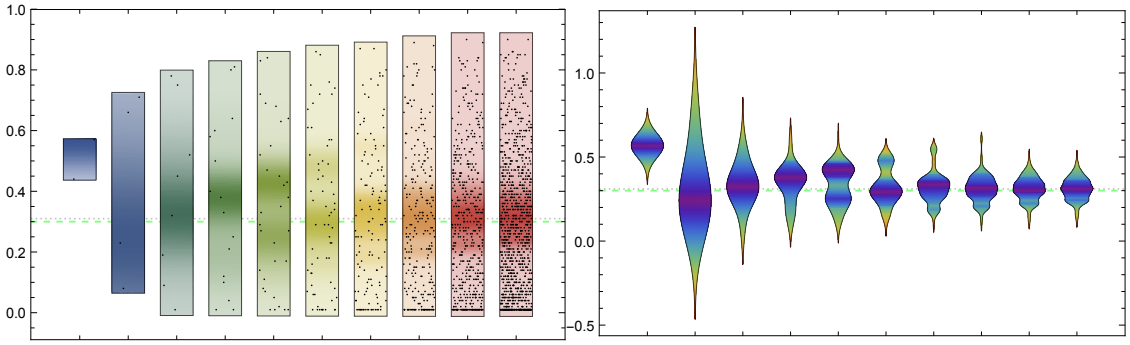


Figure 5.14: Stage-wise cost distributions in the BEP trial for (a) expected-value, (b) knowledge gradient, and (c) robust-response policies. Median of final-stage cost is shown dotted. Spectrum denotes stage $t = 1, 2, \cdots, T$; only 10 stages shown.

Table 5.5: Descriptive statistics for the stagewise distribution of dosage $c(\boldsymbol{\pi}_t, \eta^*)$ in the BEP trial. We report the mean median deviation (MMD), root mean squared deviation (RMSD), median absolute deviation (MAD), and quantiles.

Patient t	$\boldsymbol{\pi}$	MMD_t^π	RMSD_t^π	MAD_t^π	$q_{.25}$	$q_{.5}$	$q_{.75}$
	RR	0.001	0.569	0.000	0.570	0.570	0.570
2	EV	0.002	0.519	0.000	0.520	0.520	0.520
	KG	0.002	0.558	0.000	0.560	0.560	0.560
	RR	0.090	0.369	0.000	0.230	0.230	0.230
3	EV	0.134	0.422	0.000	0.230	0.230	0.710
	KG	0.104	0.452	0.000	0.300	0.300	0.300
	RR	0.060	0.357	0.000	0.320	0.320	0.320
4	EV	0.074	0.387	0.000	0.330	0.330	0.520
	KG	0.091	0.532	0.000	0.540	0.540	0.540
	RR	0.067	0.359	0.000	0.210	0.380	0.380
5	EV	0.068	0.380	0.010	0.390	0.400	0.400
	KG	0.106	0.444	0.110	0.350	0.350	0.510
	RR	0.098	0.350	0.100	0.230	0.330	0.430
6	EV	0.101	0.370	0.090	0.280	0.320	0.470
	KG	0.069	0.393	0.090	0.290	0.380	0.410
	RR	0.081	0.343	0.070	0.250	0.320	0.390
7	EV	0.074	0.349	0.070	0.260	0.330	0.380
	KG	0.067	0.371	0.050	0.290	0.370	0.410
	RR	0.069	0.333	0.050	0.270	0.310	0.370
8	EV	0.064	0.336	0.040	0.280	0.310	0.360
	KG	0.069	0.361	0.050	0.290	0.360	0.410
	RR	0.060	0.321	0.050	0.260	0.320	0.360
9	EV	0.059	0.328	0.040	0.260	0.330	0.370
	KG	0.061	0.349	0.050	0.280	0.330	0.390
	RR	0.057	0.320	0.040	0.260	0.300	0.360
10	EV	0.054	0.325	0.040	0.270	0.310	0.360
	KG	0.055	0.340	0.040	0.290	0.330	0.360
	RR	0.055	0.319	0.050	0.260	0.310	0.360
11	EV	0.052	0.324	0.040	0.280	0.320	0.370
	KG	0.049	0.332	0.040	0.280	0.330	0.360

5.2.3 Etoposide

As another case study, we consider the design of a Phase II oncology trial to assess introduction of the antineoplastic agent etoposide and tandem dose escalation of etoposide and cyclophasphamide (CP) for treatment of advanced-stage and/or persistent Hodgkin lymphoma (HL) via BEACOPP combination therapy, together with fixed levels of the agents bleomycin (10. $\mathrm{mg/m^2}$), adriamycin (35. $\mathrm{mg/m^2}$), vincristine (1.4 $\mathrm{mg/m^2}$), procarbazine (100. $\mathrm{mg/m^2}$), and prednisone (40. $\mathrm{mg/m^2}$). Throughout the trial, toxic response constitutes white blood count (WBC) less than $10.00\mathrm{e}2/\mu\mathrm{L}$ for more than 4 days, and/or platelet (PLT) count less than $2.500\,00 \times 10^5/\mu\mathrm{L}$ commensurate with grade III or IV pulmonary toxicity within two weeks, as clinically defined by the World Health Organization (WHO) [9].

The Phase II trial consists of $T = 13$ patients. We allow $|\mathcal{U}| = 100$ dose levels, with a minimum dose $u_{\min} = 100\,\mathrm{mg/m^2}$ and a maximum dose $u_{\max} = 250\,\mathrm{mg/m^2}$. Based on prior experience with the agents including analysis of the data in [141], it is estimated that $p_\epsilon = 1/50$ patients exhibit dose-limiting toxicity below $100\,\mathrm{mg/m^2}$. The MTD quantile is set to $\nu = 1/3$. The true, unknown MTD for the case study is $\eta^* = 175\,\mathrm{mg/m^2}$. These trial data are summarized in Table 5.6 and Figure 5.15.

Table 5.6: Phase II dose escalation schema of BEACOPP combination therapy for treatment of advanced Hodgkin lymphoma at the German Hodgkin's Lymphoma Study Group (GHSG), adapted from [141]. Note: We consider $|\mathcal{U}| = 100$ dose levels, over the range 1–6 shown below.

BEACOPP	Dose Escalation (mg/m^2)						
	Baseline	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Bleomycin	10						
Adriamycin	35						
Vincristine	1.4						
Procarbazine	100						
Prednisone	40	Feasible Dosage Space					
Cyclophos.	650	800	950	1100	1250	1400	1550
Etoposide	100	125	150	175	200	225	250

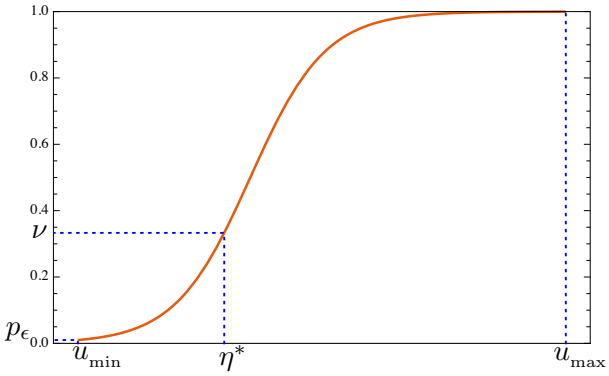


Figure 5.15: Probability of pulmonary toxicity vs. dose in the BEACOPP trial.

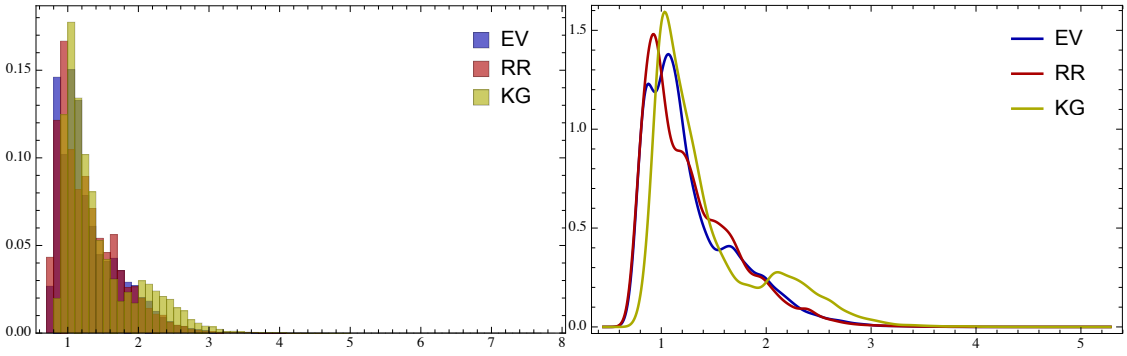


Figure 5.16: Distribution of cumulative cost $\sum_{t=1}^{T} c(\pi_t, \eta^*)$.

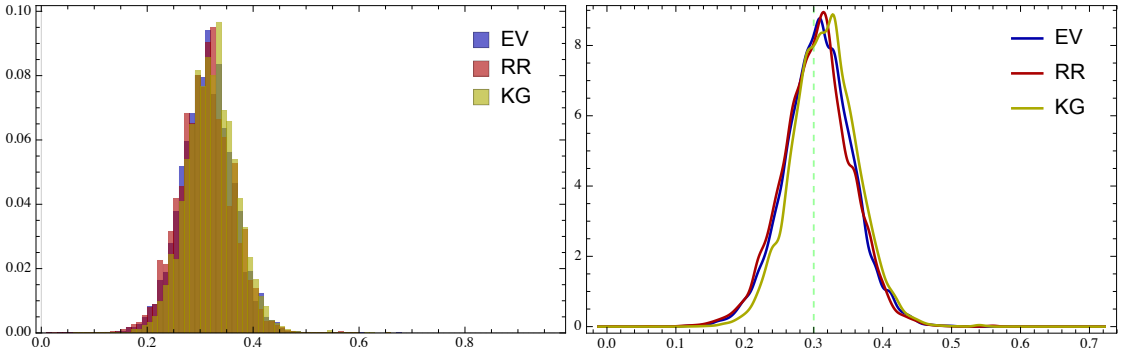
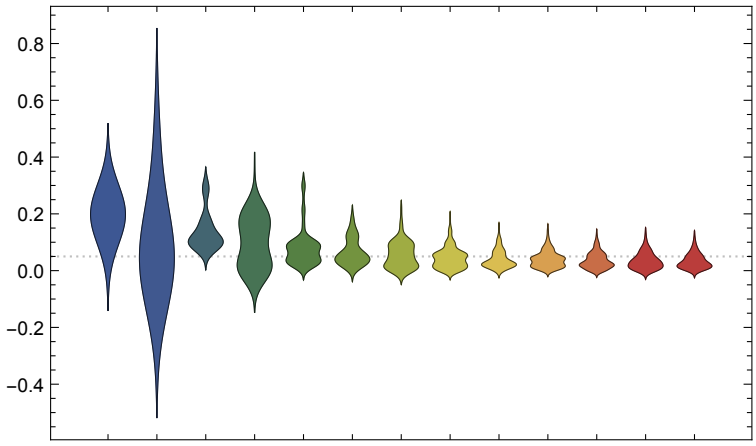
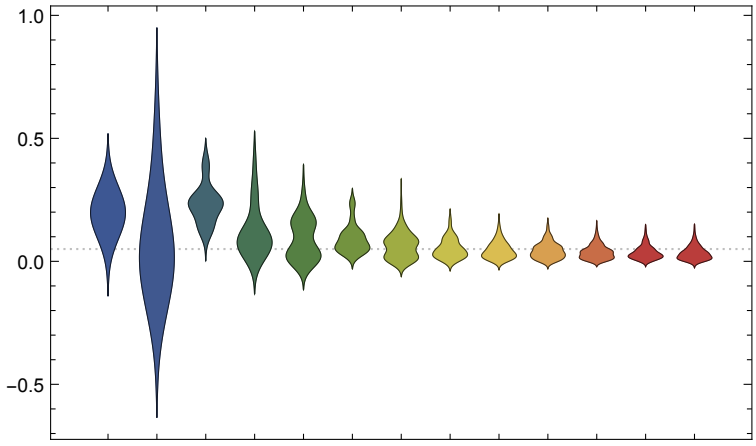


Figure 5.17: Distribution of trial MTD recommendation $\hat{\eta}_{T+1}$.

(a) Expected Value



(b) Knowledge Gradient



(c) Robust Response

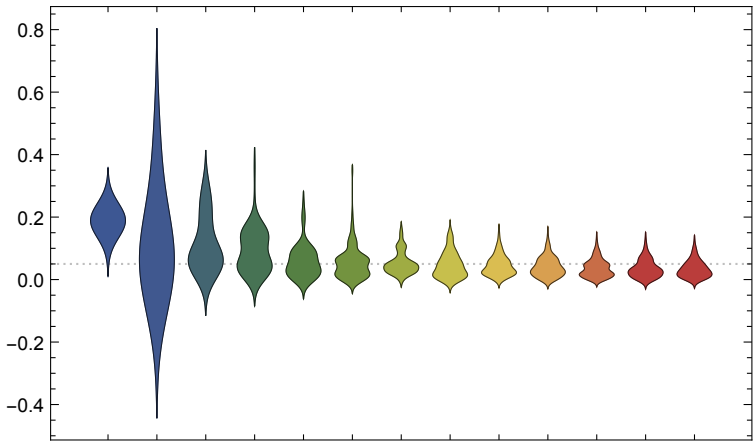
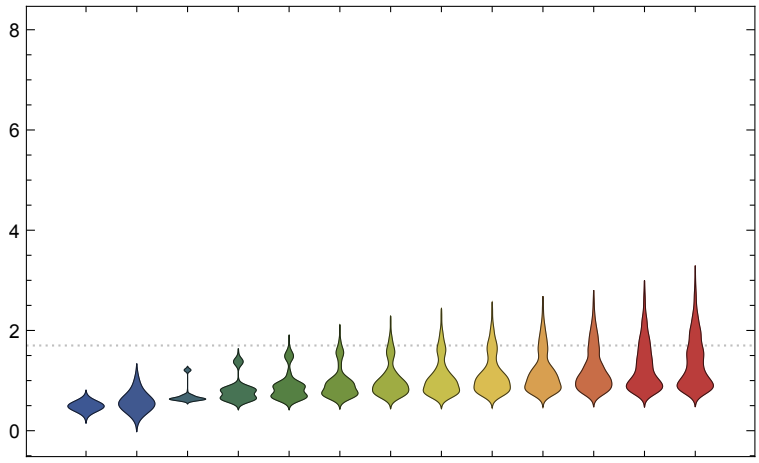
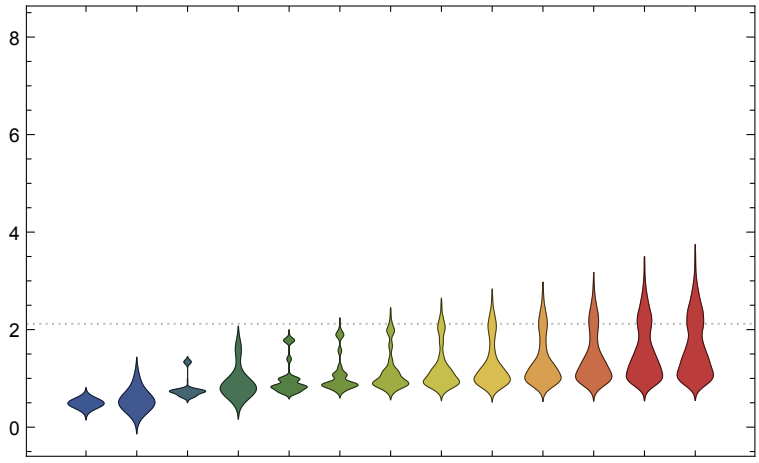


Figure 5.18: Stage-wise cost distributions in the BEACOPP trial for (a) expected-value, (b) knowledge gradient, and (c) robust-response policies. Median of final-stage cost is shown dotted. Color spectrum denotes stage $t = 1, 2, \cdots, 10$.

(a) Expected Value



(b) Knowledge Gradient



(c) Robust Response

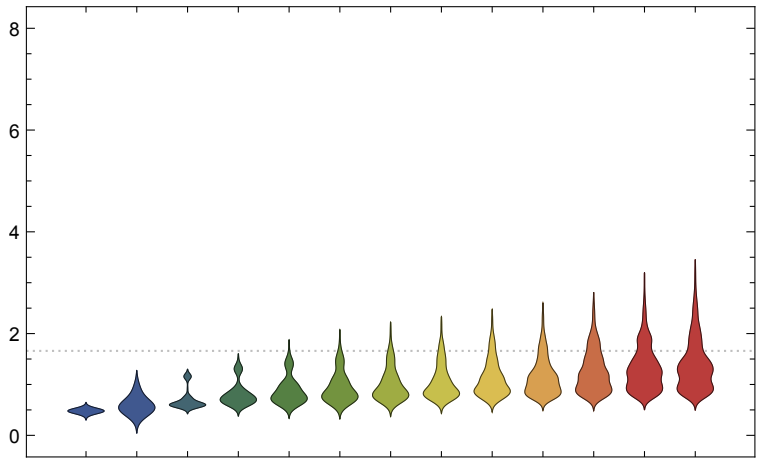
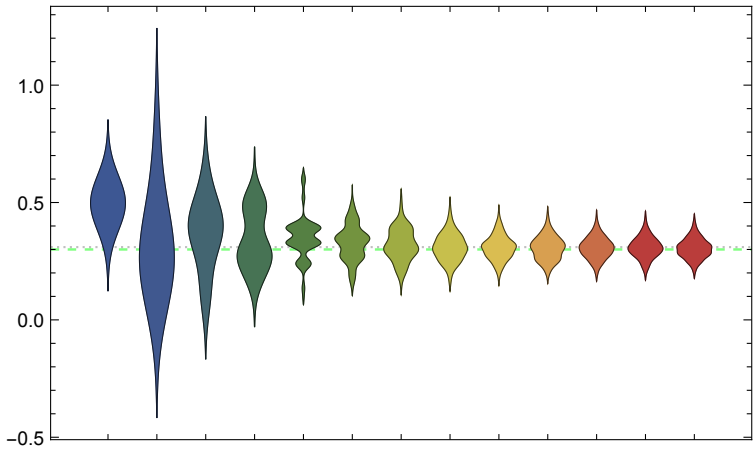
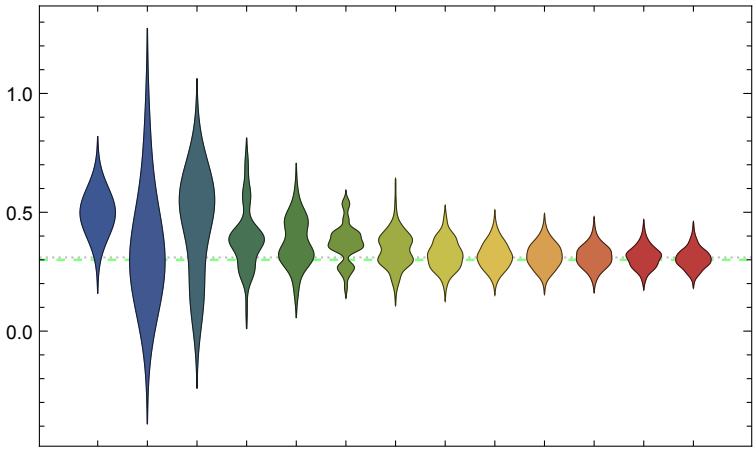


Figure 5.19: Stage-wise cost distributions in the BEACOPP trial for (a) expected-value, (b) knowledge gradient, and (c) robust-response policies. Median of final-stage cost is shown dotted. Color spectrum denotes stage $t = 1, 2, \cdots, 10$.

(a) Expected Value



(b) Knowledge Gradient



(c) Robust Response

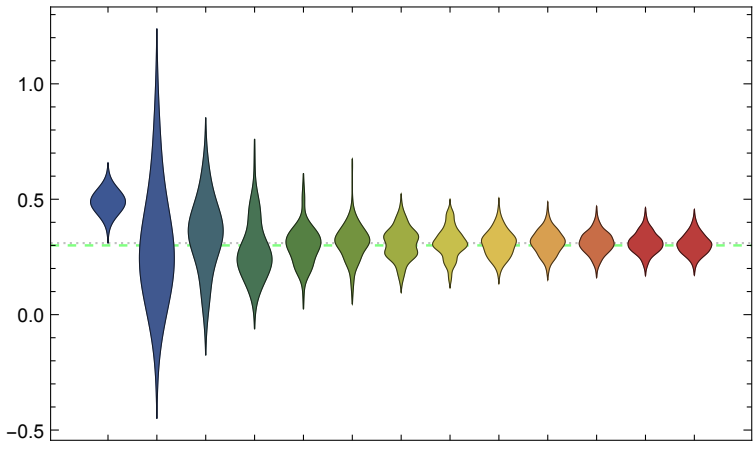


Figure 5.20: Stage-wise cost distributions in the BEACOPP trial for (a) expected-value, (b) knowledge gradient, and (c) robust-response policies. Median of final-stage cost is shown dotted. Color spectrum denotes stage $t = 1, 2, \cdots, 10$.

Chapter 6
Conclusions

> *Choose well. Your choice is brief, and yet endless.*
>
> Goethe, *1749–1832*

> *We are what we repeatedly do. Excellence, then, is not an act, but a habit.*
>
> Will Durant, *1926*

> *For these masters of living, presence to the day–to–day learning process is akin to that purity of focus others dream of achieving in rare climactic moments when everything is on the line...*
>
> *The secret is that everything is always on the line.*
>
> Josh Waitzkin, *2008*

6.1 Optimal Learning and Dynamic Risk

The continuous– and discrete–time controlled Markov systems have been the tremendously successful in modeling the dynamics of stochastic processes. By appeal to notions of classical mechanics, a stochastic process exhibiting the Markov property bears a certain analogy to a conservative field exhibiting path–independence of integrals. Given

that (stationary) Markov processes are recursive, in that the same underlying process repeatedly acts on the outcome of the previous action, the techniques of (approximate) dynamic programming (DP), in particular, are at once extremely natural and have proven powerfully effective in obtaining optimal solutions in this context under very general conditions. That is, again by appeal to classical notions, DP methods have been successful in obtaining optimal solutions without conditions, or constructs, invoking the differential and thus the methods of differential calculus. By way of analogy, this perhaps suggests that some analogue to the notion of differential is implicit in the process. Indeed, the concept of the fixed point of an iterated process is, in the deepest sense, the quintessence of knowledge, for what is a theorem if not a fixed point under iterated reasoning?

As alluded to above, optimal learning, which is to say controlled statistical inference or, plainly, intelligent experimentation, is intrinsically dynamic. Cast in the framework of a collection, or space, of Markov processes, optimal learning folds this paradigm back onto itself, inducing a non–stationarity in the realized process as one moves through the space of processes. This non–stationarity, this dynamic is of a fundamentally different nature than that of the stochastic process, one which is not unlike walking on a windy day or docking a ship in rough waters. The difference is characterized by agency, which is to say some measure of approximate control, in contrast to the external dynamic of stochastic process.

With respect to this active, controlled dynamic—learning—notions of the differential may again be useful. In this sense, the frameworks used in information geometry connect directly to those mathematical elements foundational in the classical mechanics, opening the door to a host of formal techniques developed through the study of the dynamics of more simple systems admitting description by, e.g., fiber bundles, and the differential calculus in a broad sense. However, initial pursuits in this direction indicate that the interpretations of statistical inference in terms of the differential geometry of statistical manifolds necessitate computation of fiendish quantities (such as, e.g., the deficiency of an efficient test, etc.) that have proven tremendously challenging, owing to the complicated geometry.

On the other hand, in a certain sense, risk measures capture the salient aspects of the convexity and nonlinearity of inference lurking in the current distribution itself, without recourse to parallel transport and computation of covariant derivatives. Put differently, each belief state constitutes a trove of information, not only about the history of the process, but also about the prospects of its future. In this intrinsically recursive setting, dynamic Markov risk measures thus offer one single lens through which both the stochastic dynamic and the learning dynamic may be coherently viewed, but which also lends itself directly to the recursive dynamic programming techniques inherent to the process. Thus, in a broad sense, the construction of dynamic Markov risk measures marks a momentous development in the advancement of controlled Markov processes generally, but especially to the theory (and practice) of optimal learning. Indeed, there exist many prospects for future developments, both theoretical and by way of applications.

References

[1] AM Abouammoh and AF Mashhour, *Variance upper bounds and convolutions of α-unimodal distributions*, Statistics & Probability Letters **21** (1994), no. 4, 281–289.

[2] Amir Ahmadi-Javid, *An information-theoretic approach to constructing coherent risk measures*, International Symposium on Information Theory (ISIT), IEEE, 2011, pp. 2125–2127.

[3] ———, *Entropic value-at-risk: A new coherent risk measure*, Journal of Optimization Theory and Applications **155** (2012), no. 3, 1105–1123.

[4] Bernard Altshuler, *Modeling of dose-response relationships*, Environmental Health Perspectives **42** (1981), 23–27.

[5] Shun-ichi Amari, *α-divergence is unique, belonging to both f-divergence and Bregman divergence classes*, IEEE Transactions on Information Theory **55** (2009), no. 11, 4925–4931.

[6] Shun-ichi Amari and Hiroshi Nagaoka; [translated from the Japanese by Daisha Harada], *Methods of information geometry*, Translations of Mathematical Monographs (Shoshichi Kobayashi and Masamichi Takesaki, eds.), vol. 191, American Mathematical Society, 1993.

[7] Els Ampe, Bénédicte Delaere, Jean-Daniel Hecq, Paul M Tulkens, and Youri Glupczynski, *Implementation of a protocol for administration of vancomycin by continuous infusion: pharmacokinetic, pharmacodynamic and toxicological aspects*, International Journal of Antimicrobial Agents **41** (2013), no. 5, 439–446.

[8] Jean-François Angers, *Fourier transform and Bayes estimator of a location parameter*, Statistics & Probability Letters **29** (1996), no. 4, 353–359.

[9] SG Arbuck, SP Ivy, Aea Setser, et al., *The revised common toxicity criteria: Version 2.0*, Cancer Therapy Evaluation Program. http://ctep. info. nih. gov (1999).

[10] Saro H Armenian, Wendy Landier, Liton Francisco, Claudia Herrera, George Mills, Aida Siyahian, Natt Supab, Karla Wilson, Julie A Wolfson, David Horak, et al., *Long-term pulmonary function in survivors of childhood cancer*, Journal of Clinical Oncology **33** (2015), no. 14, 1592–1600.

[11] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath, *Coherent measures of risk*, Mathematical Finance **9** (1999), no. 3, 203–228.

[12] Peter Auer, *Using confidence bounds for exploitation-exploration trade-offs*, Journal of Machine Learning Research **3** (2002), no. Nov, 397–422.

[13] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer, *Finite-time analysis of the multiarmed bandit problem*, Machine Learning **47** (2002), no. 2-3, 235–256.

[14] James Babb, André Rogatko, and Shelemyahu Zacks, *Cancer phase I clinical trials: efficient dose escalation with overdose control*, Statistics in Medicine **17** (1998), no. 10, 1103–1120.

[15] Arindam Banerjee, *On Bayesian bounds*, International Conference on Machine Learning, ACM, 2006, pp. 81–88.

[16] ———, *An analysis of logistic models: Exponential family connections and online performance*, International Conference on Data Mining, SIAM, 2007, pp. 204–215.

[17] OE Barndorff-Nielsen and PE Jupp, *Approximating exponential models*, Annals of the Institute of Statistical Mathematics **41** (1989), no. 2, 247–267.

[18] Jay Bartroff and Tze Lai, *Approximate dynamic programming and its applications to the design of phase I cancer trials*, Statistical Science (2010), 245–257.

[19] James Vere Beck and Kenneth J Arnold, *Parameter Estimation in Engineering and Science*, James Beck, 1977.

[20] Richard Bellman, *Dynamic programming and Lagrange multipliers*, Proceedings of the National Academy of Sciences **42** (1956), no. 10, 767–769.

[21] Aharon Ben-Tal, Adi Ben-Israel, and Marc Teboulle, *Certainty equivalents and information measures: duality and extremal principles*, Journal of Mathematical Analysis and Applications **157** (1991), no. 1, 211–236.

[22] Aharon Ben-Tal and Marc Teboulle, *An old-new concept of convex risk measures: the optimized certainty equivalent*, Mathematical Finance **17** (2007), no. 3, 449–476.

[23] Dimitri Bertsekas, *Dynamic Programming and Optimal Control*, 4 ed., vol. 2, Athena Scientific, Belmont, MA, 1995.

[24] ———, *Dynamic Programming and Optimal Control*, 2 ed., vol. 1, Athena Scientific, Belmont, MA, 1995.

[25] ———, *Abstract dynamic programming*, 2013.

[26] Dimitri Bertsekas and Steven Shreve, *Stochastic optimal control: The discrete time case*, vol. 23, Academic Press New York, 1978.

[27] Patrick Billingsley, *Probability and Measure*, John Wiley & Sons, 2008.

[28] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[29] Lev M Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR computational mathematics and mathematical physics **7** (1967), no. 3, 200–217.

[30] Michael D Brundage, Joseph L Pater, and Benny Zee, *Assessing the reliability of two toxicity scales: implications for interpreting toxicity data*, Journal of the National Cancer Institute **85** (1993), no. 14, 1138–1148.

[31] Apostolos N Burnetas and Michael N Katehakis, *Optimal adaptive policies for sequential allocation problems*, Advances in Applied Mathematics **17** (1996), no. 2, 122–142.

[32] Ozlem Cavus and Andrzej Ruszczynski, *Risk-averse control of undiscounted transient markov models*, SIAM Journal on Control and Optimization **52** (2014), no. 6, 3935–3966.

[33] Amandine Chaix, Amir Zarrinpar, Phuong Miu, and Satchidananda Panda, *Time-restricted feeding is a preventative and therapeutic intervention against diverse nutritional challenges*, Cell metabolism **20** (2014), no. 6, 991–1005.

[34] Kathryn Chapman, Stuart Creton, Hugo Kupferschmidt, G Randall Bond, Martin F Wilks, and Sally Robinson, *The value of acute toxicity studies to support the clinical management of overdose and poisoning: a cross-discipline consensus*, Regulatory Toxicology and Pharmacology **58** (2010), no. 3, 354–359.

[35] Jingyang Chen and JoAnne Stubbe, *Bleomycins: towards better therapeutics*, Nature Reviews Cancer **5** (2005), no. 2, 102–112.

[36] Edward Chu, Vincent T DeVita Jr, and Vincent T DeVita Jr, *Physicians' cancer chemotherapy drug manual 2016*, Jones & Bartlett Publishers, 2015.

[37] KR Cooper and WK Hong, *Prospective study of the pulmonary toxicity of continuously infused bleomycin*, Cancer Treatment Reports **65** (1980), no. 5-6, 419–425.

[38] I Csisz et al., *Information-type measures of difference of probability distributions and indirect observations*, Studia Sci. Math. Hungar. **2** (1967), 299–318.

[39] D Cunningham, Lesley McTaggart, M Soukop, J Cummings, GJ Forrest, and JFB Stuart, *Etoposide: a pharmacokinetic profile including an assessment of bioavailability*, Medical Oncology and Tumor Pharmacotherapy **3** (1986), no. 2, 95–99.

[40] Aliva De, Igor Guryev, Alejandro LaRiviere, Roberta Kato, Choo Phei Wee, Leo Mascarenhas, Thomas G Keens, and Rajkumar Venkatramani, *Pulmonary function abnormalities in childhood cancer survivors treated with bleomycin*, Pediatric blood & cancer **61** (2014), no. 9, 1679–1684.

[41] Morris H DeGroot, *Unbiased sequential estimation for binomial populations*, The Annals of Mathematical Statistics (1959), 80–101.

[42] Darinka Dentcheva, Spiridon Penev, and Andrzej Ruszczyński, *Statistical estimation of composite risk functionals and risk optimization problems*, Annals of the Institute of Statistical Mathematics (2015), 1–24.

[43] Darinka Dentcheva, András Prékopa, and Andrzej Ruszczyński, *Concavity and efficient points of discrete distributions in probabilistic programming*, Mathematical Programming **89** (2000), no. 1, 55–77.

[44] Darinka Dentcheva and Andrzej Ruszczyński, *Optimization with stochastic dominance constraints*, SIAM Journal on Optimization **14** (2003), no. 2, 548–566.

[45] ———, *Optimality and duality theory for stochastic optimization problems with nonlinear dominance constraints*, Mathematical Programming **99** (2004), no. 2, 329–350.

[46] ———, *Common mathematical foundations of expected utility and dual utility theories*, SIAM Journal on Optimization **23** (2013), no. 1, 381–405.

[47] ———, *Risk preferences on the space of quantile functions*, Mathematical Programming **148** (2014), no. 1-2, 181–200.

[48] KB DeSalvo, R Olson, and KO Casavale, *Dietary guidelines for americans*, Journal of the American Medical Association **315** (2016), no. 5, 457.

[49] Sandeep Dhindsa, Husam Ghanim, Manav Batra, Nitesh D Kuhadiya, Sanaa Abuaysheh, Kelly Green, Antoine Makdissi, Ajay Chaudhuri, and Paresh Dandona, *Effect of testosterone on hepcidin, ferroportin, ferritin and iron binding capacity in patients with hypogonadotropic hypogonadism and type 2 diabetes*, Clinical Endocrinology **85** (2016), no. 5, 772–780.

[50] Bradley Efron and Robert J Tibshirani, *An introduction to the bootstrap*, CRC press, 1994.

[51] Lawrence H Einhorn, *Treatment of testicular cancer: a new and improved model.*, Journal of clinical oncology **8** (1990), no. 11, 1777–1781.

[52] ———, *Curing metastatic testicular cancer*, Proceedings of the National Academy of Sciences **99** (2002), no. 7, 4592–4595.

[53] ———, *Salvage chemotherapy for patients with germ cell tumors: is there a best regimen?*, Journal of Clinical Oncology **30** (2012), no. 8, 771–772.

[54] Lawrence H Einhorn, Stephen D Williams, Patrick J Loehrer, Robert Birch, Ray Drasga, George Omura, and F Anthony Greco, *Evaluation of optimal duration of chemotherapy in favorable-prognosis disseminated germ cell tumors: a Southeastern Cancer Study Group protocol.*, Journal of Clinical Oncology **7** (1989), no. 3, 387–391.

[55] Andreas Engert, Volker Diehl, Jeremy Franklin, Andreas Lohri, Bernd Dörken, Wolf-Dieter Ludwig, Peter Koch, Mathias Hänel, Michael Pfreundschuh, Martin Wilhelm, et al., *Escalated-dose BEACOPP in the treatment of patients with advanced-stage Hodgkin's lymphoma: 10 years of follow-up of the GHSG HD9 study*, Journal of Clinical Oncology **27** (2009), no. 27, 4548–4554.

[56] Tim van Erven, Mark D Reid, and Robert C Williamson, *Mixability is bayes risk curvature relative to log loss*, Journal of Machine Learning Research **13** (2012), no. May, 1639–1663.

[57] International Warfarin Pharmacogenetics Consortium, et al., *Estimation of the Warfarin dose with clinical and pharmacogenetic data*, New England Journal of Medicine **2009** (2009), no. 360, 753–764.

[58] Jingnan Fan and Andrzej Ruszczyński, *Process-based risk measures for observable and partially observable discrete-time controlled systems*, arXiv preprint arXiv:1411.2675 (2014).

[59] Lorna A Fern and Jeremy S Whelan, *Recruitment of adolescents and young adults to cancer clinical trials? International comparisons, barriers, and implications*, Seminars in oncology, vol. 37, Elsevier, 2010, pp. e1–e8.

[60] Steven A Frank, *How to read probability distributions as statements about process*, Entropy **16** (2014), no. 11, 6059–6098.

[61] Steven A Frank and D Eric Smith, *Measurement invariance, entropy, and probability*, Entropy **12** (2010), no. 3, 289–303.

[62] Peter I Frazier, Warren B Powell, and Savas Dayanik, *A knowledge-gradient policy for sequential information collection*, SIAM Journal on Control and Optimization **47** (2008), no. 5, 2410–2439.

[63] Béla A Frigyik, Santosh Srivastava, and Maya R Gupta, *Functional Bregman divergence and Bayesian estimation of distributions*, IEEE Transactions on Information Theory **54** (2008), no. 11, 5130–5139.

[64] David R Gandara, Edith A Perez, WA Phillips, HJ Lawrence, and Michael DeGregorio, *Evaluation of cisplatin dose intensity: current status and future prospects*, Anticancer Research **9** (1988), no. 4, 1121–1128.

[65] Josiah Willard Gibbs, *A method of geometrical representation of the thermodynamic properties of substances by means of surfaces*, Connecticut Academy, 1873.

[66] J Gittins and D Jones, *A dynamic allocation index for the sequential allocation of experiments, in (j. gani, et al, eds.) progress in statistics*, 1974.

[67] John Gittins, Kevin Glazebrook, and Richard Weber, *Multi-armed bandit allocation indices*, John Wiley & Sons, 2011.

[68] Prem K Goel and Morris H DeGroot, *Comparison of experiments and information measures*, The Annals of Statistics (1979), 1066–1077.

[69] Peter D Grünwald and A Philip Dawid, *Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory*, Annals of Statistics (2004), 1367–1433.

[70] G. I. Hackett, *Testosterone replacement therapy and mortality in older men*, Drug Safety **39** (2016), no. 2, 117–130.

[71] Godfrey Harold Hardy, John Edensor Littlewood, and George Pólya, *Inequalities*, 2 ed., Cambridge University Press, 1952.

[72] Michelle N Harvie, Mary Pegington, Mark P Mattson, Jan Frystyk, Bernice Dillon, Gareth Evans, Jack Cuzick, Susan A Jebb, Bronwen Martin, Roy G Cutler, et al., *The effects of intermittent or continuous energy restriction on weight loss and metabolic disease risk markers: a randomized trial in young overweight women*, International Journal of Obesity **35** (2011), no. 5, 714–727.

[73] Onésimo Hernández-Lerma, *Adaptive Markov control processes*, Applied Mathematical Sciences (J.E. Marsden, L. Sirovich, and F. John, eds.), vol. 79, Springer, 1995.

[74] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh, *A fast learning algorithm for deep belief nets*, Neural computation **18** (2006), no. 7, 1527–1554.

[75] Sbeity Hode, Younes Rafic, Topsu Suat, and Mougharbel Imad, *Comparative study of the optimization theory for cancer treatment*, 4th International Conference on Biomedical Engineering and Informatics (BMEI), 2011.

[76] Douglas H Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid;[a Metaphoric Fugue on Minds and Machines in the Spirit of Lewis Carroll]*., Penguin Books, 1980.

[77] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten, *Densely connected convolutional networks*, arXiv preprint arXiv:1608.06993 (2016).

[78] Kiyomi Ito and J Brian Houston, *Prediction of human drug clearance from in vitro and preclinical data using physiologically based and empirical approaches*, Pharmaceutical Research **22** (2005), no. 1, 103–112.

[79] Michael I Jordan et al., *Why the logistic function? A tutorial discussion on probabilities and neural networks*, 1995.

[80] Steven D. Jones Jr., Thomas Dukovac, Premsant Sangkum, Faysal A. Yafi, and Wayne J.G. Hellstrom, *Erythrocytosis and polycythemia secondary to testosterone replacement therapy in the aging male*, Sexual Medicine Reviews **3** (2015), no. 2, 101–112.

[81] Edward L Kaplan and Paul Meier, *Nonparametric estimation from incomplete observations*, Journal of the American statistical association **53** (1958), no. 282, 457–481.

[82] Michael N Katehakis and Cyrus Derman, *Computing optimal sequential allocation rules in clinical trials*, Lecture notes-monograph series (1986), 29–39.

[83] Michael N Katehakis and Herbert Robbins, *Sequential choice from several populations.*, Proceedings of the National Academy of Sciences of the United States of America **92** (1995), no. 19, 8584.

[84] Michael N Katehakis and Arthur F Veinott Jr, *The multi-armed bandit problem: decomposition and computation*, Mathematics of Operations Research **12** (1987), no. 2, 262–268.

[85] Jack Kiefer and Jacob Wolfowitz, *Optimum designs in regression problems*, The Annals of Mathematical Statistics (1959), 271–294.

[86] Roger Koenker and Ivan Mizera, *Quasi-concave density estimation*, The Annals of Statistics (2010), 2998–3027.

[87] Rudolf Kulhavy, *Recursive nonlinear estimation: geometry of a space of posterior densities*, Automatica **28** (1992), no. 2, 313–323.

[88] Shigeo Kusuoka, *On law invariant coherent risk measures*, Advances in Mathematical Economics, Springer, 2001, pp. 83–95.

[89] John Lafferty, Andrew McCallum, and Fernando Pereira, *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*, International Conference on Machine Learning (ICML), vol. 1, 2001, pp. 282–289.

[90] Tze Leung Lai, *Sequential analysis*, Encyclopedia of Biostatistics (2001).

[91] Tze Leung Lai and Herbert Robbins, *Asymptotically efficient adaptive allocation rules*, Advances in Applied Mathematics **6** (1985), no. 1, 4–22.

[92] Michael S Lauer, *Testosterone replacement therapy: intent matters*, The Lancet Diabetes & Endocrinology **4** (2016), no. 6, 471–473.

[93] Christophe Le Tourneau, J Jack Lee, and Lillian L Siu, *Dose escalation methods in phase I cancer clinical trials*, Journal of the National Cancer Institute **101** (2009), no. 10, 708–720.

[94] Steve Levitt and Adi Ben-Israel, *On modeling risk in Markov decision processes*, Optimization and Related Topics, Springer US, 2001, pp. 27–40.

[95] Valter D Longo and Mark P Mattson, *Fasting: molecular mechanisms and clinical applications*, Cell metabolism **19** (2014), no. 2, 181–192.

[96] Ahmad Majzoub and Daniel A Shoskes, *A case series of the safety and efficacy of testosterone replacement therapy in renal failure and kidney transplant patients*, Translational Andrology and Urology **5** (2016), no. 6, 814–818.

[97] Maryann Mazer-Amirshahi, Gerald Sokol, John van den Anker, and Louis Cantilena, *A review of pharmaceutical labeling for overdose treatment and toxicity data*, Pharmacoepidemiology Drug Safety **22** (2013), no. 3, 319–323.

[98] Diana M Negoescu, Peter I Frazier, and Warren B Powell, *The knowledge-gradient algorithm for sequencing experiments in drug discovery*, INFORMS Journal on Computing **23** (2011), no. 3, 346–363.

[99] Włodzimierz Ogryczak and Andrzej Ruszczyński, *From stochastic dominance to mean-risk models: Semideviations as risk measures*, European Journal of Operational Research **116** (1999), no. 1, 33–50.

[100] _____, *On consistency of stochastic dominance and mean–semideviation models*, Mathematical Programming **89** (2001), no. 2, 217–232.

[101] Wlodzimierz Ogryczak and Andrzej Ruszczyński, *Dual stochastic dominance and quantile risk measures*, International Transactions in Operational Research **9** (2002), no. 5, 661–680.

[102] WLodzimierz Ogryczak and Andrzej Ruszczyński, *Dual stochastic dominance and related mean-risk models*, SIAM Journal on Optimization **13** (2002), no. 1, 60–78.

[103] John O'Quigley, Margaret Pepe, and Lloyd Fisher, *Continual reassessment method: a practical design for phase I clinical trials in cancer*, Biometrics (1990), 33–48.

[104] Carl C Peck, William H Barr, Leslie Z Benet, Jerry Collins, Robert E Desjardins, Daniel E Furst, John G Harter, Gerhard Levy, Thomas Ludden, John H Rodman, et al., *Opportunities for integration of pharmacokinetics, pharmacodynamics, and toxicokinetics in rational drug development*, Journal of Pharmaceutical Pciences **81** (1992), no. 6, 605–610.

[105] TJ Postma, JJ Heimans, MJ Muller, GJ Ossenkoppele, JB Vermorken, and NK Aaronson, *Pitfalls in grading severity of chemotherapy-induced peripheral neuropathy*, Annals of Oncology **9** (1998), no. 7, 739–744.

[106] Warren B Powell and Ilya O Ryzhov, *Optimal Learning*, vol. 841, John Wiley & Sons, 2012.

[107] András Prékopa, *Stochastic Programming*, Springer Science & Business Media, 1995.

[108] Kathy Pritchard-Jones and Darren Hargrave, *Declining childhood and adolescent cancer mortality: great progress but still much to be done*, Cancer **120** (2014), no. 16, 2388–2391.

[109] Martin L Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, 1994.

[110] Elizabeth Record, Rebecca Williamson, Karen Wasilewski-Masker, Ann C Mertens, Lillian R Meacham, and Jonathan Popler, *Analysis of risk factors for abnormal pulmonary function in pediatric cancer survivors*, Pediatric blood & cancer **63** (2016), no. 7, 1264–1271.

[111] Mark D Reid and Robert C Williamson, *Information, divergence and risk for binary experiments*, Journal of Machine Learning Research **12** (2011), no. Mar, 731–817.

[112] Nicholas Rensing, Lirong Han, and Michael Wong, *Intermittent dosing of rapamycin maintains antiepileptogenic effects in a mouse model of tuberous sclerosis complex*, Epilepsia **56** (2015), no. 7, 1088–1097.

[113] Victor M Rivera, Rachel M Squillace, David Miller, Lori Berk, Scott D Wardwell, Yaoyu Ning, Roy Pollock, Narayana I Narasimhan, John D Iuliucci, Frank Wang, et al., *Ridaforolimus (AP23573; MK-8669), a potent mtor inhibitor, has broad antitumor activity and can be optimally administered using intermittent dosing regimens*, Molecular cancer therapeutics **10** (2011), no. 6, 1059–1071.

[114] Herbert Robbins, *Some aspects of the sequential design of experiments*, Bulletin of the American Mathematical Society **58** (1952), no. 5, 527–535.

[115] Fred S Roberts, *Measurement Theory with applications to decisionmaking, utility, and the social sciences*, Encylcopedia of Mathematics and its Applications (Gian-Carlo Rota, ed.), vol. 7, Cambridge University Press, 1985.

[116] R Tyrrell Rockafellar, *Convex Analysis. Princeton landmarks in mathematics*, Princeton University Press, Princeton, NJ, 1997.

[117] R Tyrrell Rockafellar and Stan Uryasev, *The fundamental risk quadrangle in risk management, optimization and statistical estimation*, Surveys in Operations Research and Management Science **18** (2013), no. 1, 33–53.

[118] R Tyrrell Rockafellar, Stan Uryasev, and Michael Zabarankin, *Risk tuning with generalized linear regression*, Mathematics of Operations Research **33** (2008), no. 3, 712–729.

[119] Jordi Rodon, Rodrigo Dienstmann, Violeta Serra, and Josep Tabernero, *Development of PI3K inhibitors: lessons learned from early clinical trials*, Nature reviews Clinical oncology **10** (2013), no. 3, 143–153.

[120] Dan Russo and Benjamin Van Roy, *Learning to optimize via information-directed sampling*, Advances in Neural Information Processing Systems, 2014, pp. 1583–1591.

[121] Andrzej Ruszczyński, *Nonlinear Optimization*, Princeton University Press, 2006.

[122] ———, *Risk-averse dynamic programming for Markov decision processes*, Mathematical Programming **125** (2010), no. 2, 235–261.

[123] Andrzej Ruszczyński and Alexander Shapiro, *Stochastic Programming*, vol. 10, Elsevier Amsterdam, 2003.

[124] ———, *Conditional risk mappings*, Mathematics of Operations Research **31** (2006), no. 3, 544–561.

[125] ———, *Optimization of convex risk functions*, Mathematics of Operations Research **31** (2006), no. 3, 433–452.

[126] Ilya O Ryzhov, Warren B Powell, and Peter I Frazier, *The knowledge gradient algorithm for a general class of online learning problems*, Operations Research **60** (2012), no. 1, 180–195.

[127] Claude Elwood Shannon, *A mathematical theory of communication*, Bell System Technical Journal **27** (1948), no. 1, 379–423.

[128] Alexander Shapiro, *On Kusuoka representation of law invariant risk measures*, Mathematics of Operations Research **38** (2013), no. 1, 142–152.

[129] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński, *Lectures in Stochastic Programming, Modeling and Theory*, SIAM, 2009.

[130] J Siepmann and F Siepmann, *Mathematical modeling of drug delivery*, International Journal of Pharmaceutics **364** (2008), no. 2, 328–343.

[131] BI Sikic, JM Collins, EG Mimnaugh, and TE Gram, *Improved therapeutic index of bleomycin when administered by continuous infusion in mice*, Cancer Treatment Reports **62** (1978), no. 12, 2011–2017.

[132] R Simon, B Freidlin, L Rubinstein, S G Arbuck, J Collins, and M C Christian, *Accele titration designs for phase I clinical trials in oncology*, Journal of the National Cancer Institute **89** (1997), no. 15, 1138–1147.

[133] Stefan Sleijfer, *Bleomycin-induced pneumonitis*, Chest Journal **120** (2001), no. 2, 617–624.

[134] Malcolm Smith, Jeffrey Abrams, Edward L Trimble, and Richard S Ungerleider, *Dose intensity of chemotherapy for childhood cancers*, The oncologist **1** (1996), no. 5, 293–304.

[135] Jasper Snoek, Hugo Larochelle, and Ryan P Adams, *Practical Bayesian optimization of machine learning algorithms*, Advances in Neural Information Processing Systems, 2012, pp. 2951–2959.

[136] Barry E Storer, *Design and analysis of phase i clinical trials*, Biometrics (1989), 925–937.

[137] J Tabernero, C Saura, D Roda Perez, R Dienstmann, S Rosello, L Prudkin, JA Perez-Fidalgo, B Graña, C Jones, L Musib, et al., *First-in-human phase i study evaluating the safety, pharmacokinetics (PK), and intratumor pharmacodynamics (PD) of the novel, oral, ATP-competitive Akt inhibitor GDC-0068.*, Journal of Clinical Oncology **29** (2011), no. 15_suppl, 3022–3022.

[138] Josep Tabernero, Andres Cervantes, Michael S Gordon, Elena G Chiorean, Howard A Burris, Teresa Macarulla, Alejandro Perez-Fidalgo, Michael Martin, Katti Jessen, Yi Liu, et al., *Abstract CT-02: A phase i, open label, dose escalation study of oral mammalian target of rapamycin inhibitor INK128 administered by intermittent dosing regimens in patients with advanced malignancies*, 2012.

[139] Fabio Silvio Taccone, Maya Hites, Marjorie Beumier, Sabino Scolletta, and Frédérique Jacobs, *Appropriate antibiotic dosage levels in the treatment of severe sepsis and septic shock*, Current Infectious Disease Reports **13** (2011), no. 5, 406–415.

[140] Martin A Tanner, *Tools for Statistical Inference*, vol. 3, Springer, 1991.

[141] H Tesch, V Diehl, B Lathan, D Hasenclever, M Sieber, U Rüffer, A Engert, J Franklin, M Pfreundschuh, KP Schalk, et al., *Moderate dose escalation for advanced stage Hodgkin's disease using the bleomycin, etoposide, adriamycin, cyclophosphamide, vincristine, procarbazine, and prednisone scheme and adjuvant radiotherapy: a study of the German Hodgkin's Lymphoma Study Group*, Blood **92** (1998), no. 12, 4560–4567.

[142] Andy Trotti, Roger Byhardt, Joanne Stetz, Clement Gwede, Benjamin Corn, Karen Fu, Leonard Gunderson, Beryl McCormick, Mitchell Morris, Tyvin Rich, et al., *Common toxicity criteria: version 2.0. an improved reference for grading the acute effects of cancer treatment: impact on radiotherapy*, International Journal of Radiation Oncology* Biology* Physics **47** (2000), no. 1, 13–47.

[143] Gilles Vassal, C Michel Zwaan, David Ashley, Marie Cecile Le Deley, Darren Hargrave, Patricia Blanc, and Peter C Adamson, *New drugs for children and*

adolescents with cancer: the need for novel development pathways, The Lancet Oncology **14** (2013), no. 3, e117–e124.

[144] Miranda Verschraagen, Epie Boven, Rita Ruijter, Kasper Born, Johannes Berkhof, Frederick H Hausheer, and Wim JF Vijgh, *Pharmacokinetics and preliminary clinical data of the novel chemoprotectant BNP7787 and cisplatin and their metabolites*, Clinical Pharmacology & Therapeutics **74** (2003), no. 2, 157–169.

[145] Josh Waitzkin, *The Art of Learning: An inner journey to optimal performance*, Simon and Schuster, 2008.

[146] Stefan Weber, *Distribution-invariant risk measures, information, and dynamic consistency*, Mathematical Finance **16** (2006), no. 2, 419–441.

[147] Peter Whittle, *Discussion of Dr. Gittins' paper, Bandit processes and dynamic allocation indices*, Journal of the Royal Statistical Society. Series B (Methodological) **41** (1979), no. 2, 165–165.

[148] Shelemyahu Zacks, *The Theory of Statistical Inference*, vol. 34, Wiley New York, 1971.

[149] ———, *Parametric Statistical Inference: Basic theory and modern approaches*, vol. 4, Elsevier, 2014.

[150] Eberhard Zeidler, *Nonlinear Functional Analysis and Its Applications III: Variational Methods and Optimization*, Springer Science & Business Media, 1985.

[151] ———, *Applied Functional Analysis, Main principles and their applications*, Applied Mathematical Sciences (J.E. Marsden, L. Sirovich, and F. John, eds.), vol. 109, Springer, 1995.

[152] Arnold Zellner, *Bayesian estimation and prediction using asymmetric loss functions*, Journal of the American Statistical Association **81** (1986), no. 394, 446–451.

[153] Alexandra P Zorzi, Connie L Yang, Sharon Dell, and Paul C Nathan, *Bleomycin-associated lung toxicity in childhood cancer survivors*, Journal of pediatric hematology/oncology **37** (2015), no. 8, e447–e452.