Data-Driven Optimization for Individualized Medical Decision-Making in Cancer

by

Weiyu Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in the University of Michigan
2021

Doctoral Committee:

Professor Brian T. Denton, Chair
Professor Xiuli Chao
Associate Professor Ruiwei Jiang
Professor Ambuj Tewari

Weiyu Li

weiyuli@umich.edu

ORCID iD: 0000-0002-7325-3506

DEDICATION

To the Yellow River,
and my parents.

ACKNOWLEDGEMENTS

First, I would like to express my utmost gratitude to my Ph.D. advisor, Dr. Brian Denton. Dr. Denton's professional guidance and generous support have built the strongest fortress for my graduate studies. His diligence and conscientiousness as a researcher and advisor, and his warmth and generosity as a mentor and friend has been the morning star for my studies at the University of Michigan and my entire career beyond. I am extremely grateful to work with him over the last five years, and will never forget everything he has taught me.

Second, I would like to thank my committee members, Dr. Xiuli Chao, Dr. Ruiwei Jiang, and Dr. Ambuj Tewari, for their guidance in topics including, but not limited to, probability theory, stochastic process, stochastic programming and robust optimization, machine learning, and reinforcement learning, through graduate-level courses and our regular meetings. Moreover, their valuable comments and suggestions to my thesis have made it a stronger work.

I would also like to thank the people who have collaborated with me in my thesis. I am grateful for the guidance and support from Dr. Zheng Zhang, Dr. Todd Morgan, Dr. Lauren Steimle. I am also grateful to my collaborators from the Movember Foundation's Global Action Plan Prostate Cancer Active Surveillance (GAP3) consortium, which include Dr. Daan Nieboer, Dr. Jozien Helleman, Dr. Jonathan Olivier, and Dr. Arnauld Villers, and from the U.S. Department of Veterans Affairs, which include Dr. Rodney Hayward and Dr. Jeremy Sussman. In addition, I am thankful to many graduate and un-

dergraduate students that I have worked with, including Wantao Lin, Jiachen Wang, and Ryan Krueger. I would also like to appreciate the help I have received from Dr. Ji Zhu, Dr. Jian Kang, and Dr. Boang Liu during my master's studies in the Statistics department, which has built the foundation for my data science knowledge and skills.

Furthermore, I am grateful to all the faculty and staff members in the Industrial and Operations Engineering department at the University of Michigan, who have provided a spectacular research atmosphere and excellent working environment for all of us. I am also grateful for the friendships from my roommates, peers, and basketball teammates that have accompanied me during my graduate studies at the University of Michigan. They all have made me a better person.

Finally, I would like to thank my parents, for their unlimited love and support, without whom I am nothing.

This work is supported in part by the National Science Foundation through Grant Number CMMI 0844511. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This work was also supported by the Movember Foundation. The funders did not play any role in the study design, collection, analysis or interpretation of data, or in the editing of this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF ABBREVIATIONS

AS Active Surveillance

AS-POMDP Active Surveillance Partially Observable Markov Decision Process

CDC Centers for Disease Control and Prevention

DREs Digital Rectal Exams

EVPI Expected Value of Perfect Information

GAP3 Global Action Plan Prostate Cancer Active Surveillance

HMM Hidden Markov Model

JH Johns Hopkins

MDP Markov Decision Process

MPOMDP Multi-model Partially Observable Markov Decision Processes

MRI Magnetic Resonance Imaging

POMDP Partially Observable Markov Decision Process

PRIAS Prostate Cancer Research International Active Surveillance

PSA Prostate-specific Antigen

QALYs Quality Adjusted Life Years

UCSF University of California San Francisco

U of T University of Toronto

VSS Value of Stochastic solution

ABSTRACT

Cancer is one of the leading causes of death in many countries, including the United States. Medical decision-making in cancer detection and treatment is often a challenging engineering problem for three reasons: the unobservable nature of the cancer state, the trade-off between alternative detection and treatment policies, and the patient heterogeneity in disease progression and clinical effectiveness. In this thesis, we take a holistic approach on data-driven optimization methods for individualized medical decision-making in cancer via three studies, in the context of Active Surveillance (AS) for prostate cancer.

In the first study, we develop a Hidden Markov Model (HMM) to describe the stochastic process of cancer progression and diagnosis tests dynamics in four major studies. The model is subsequently used as the basis for simulation models to evaluate different published biopsy protocols. In the second study, we propose a finite-horizon Partially Observable Markov Decision Process (POMDP) to optimize the timing of biopsies for each individual patient. We develop two fast approximation algorithms to solve the proposed model, and show some important properties of the optimal biopsy policy. This study also considers the impact of parameter ambiguity caused by the variation across different clinical studies and patients' preferences. In the third study, we propose a new multi-model POMDP to address the issue of parameter ambiguity in POMDPs. We analyze the mathematical structure of the model, solution algorithms, and we present numerical results demonstrating the benefits of the Multi-model Partially Observable Markov Decision Processes (MPOMDP). Finally, we summarize the most important findings from this dissertation.

CHAPTER 1

Introduction

Cancer is one of the leading causes of death in many countries, including the United States. According to the Centers for Disease Control and Prevention (CDC), cancer accounted for 599,601 (out of 2,854,838) deaths in 2019 in the United States. The American Cancer Society estimates that, in 2021, there will be 1.9 million new cancer cases diagnosed and 608,570 cancer deaths in the United States. For most cancer diseases, early detection and treatment is the key to higher survival rates. A sizable portion of cancer patients are diagnosed with low-risk forms of cancer that require regular follow-up over time, known as *surveillance*, to monitor the patient for possible progression to a higher risk form of cancer. Examples of cancers for which surveillance may be relevant include breast cancer, colorectal cancer, lung cancer, and prostate cancer.

Optimizing medical decisions in cancer surveillance can be a challenging engineering problem for several reasons. First, patients' actual health states are typically not observable, and may transit stochastically over time. Second, the optimal design of cancer surveillance strategies requires balancing the benefits of early detection and treatment, and the harms of potential false test results and server side effects. There are often conflicts between alternative decisions that must be considered to not only prevent adverse health outcomes and extend life expectancy, but save costs and minimize potential side effects.

Third, patient heterogeneity in the disease progression and clinical effectiveness of cancer surveillance is prevalent in many settings.

This thesis studies data-driven stochastic modeling and optimization approaches for individualized medical decision-making in cancer. The methodological and theoretical contributions of this thesis are motivated by a real-world healthcare application in prostate cancer AS. Prostate cancer is the most common cancer in men globally. Patients with low-risk variants of prostate cancer are recommended to joining the AS, which monitors patients by medical tests until there is a sign of progression to a high-risk variant of prostate cancer, to avoid unnecessary invasive or harmful treatments like radiation therapy or surgery. The two most common medical tests in AS are the Prostate-specific Antigen (PSA) test and biopsy. The PSA test is a simple blood test with almost no direct harm. It measures the amount of PSA in blood serum. High PSA is associated with the presence of cancer, but there is also a high false-positive and false-negative rate. Biopsy is a much more accurate test, which samples the tissue with hollow-core needles. However, biopsy is still imperfect, with potential false-negative results caused by missing the tumor, often referred to as *under-sampling*. Moreover, biopsy is very painful and can cause infections and other harmful consequences for patients. Thus, it is critical to decide the optimal timing for biopsies for each patient in prostate cancer AS.

Deciding on the optimal biopsy policy is challenging because: 1) the patient's cancer state is not directly observable due to the inaccurate diagnostic tests; 2) cancer progression is a stochastic process for each patient; 3) patient preferences about how often to biopsy vary because the benefits and harms of biopsy versus early detection vary among patients. To resolve these challenges, this thesis provides a holistic approach on data-driven optimization for individualized medical decision-making in cancer via three main chapters. In Chapter 2, we present a new stochastic model to describe the process of prostate cancer

progression for men newly diagnosed with low-risk cancer. We then estimate the most important factors in prostate cancer AS, using data from several well-known prostate cancer studies, including the cancer progression rate, test accuracy, and reward mechanism. In Chapter 3, we use the models of Chapter 2 to formulate a POMDP to optimize whether and when to perform biopsies in AS of prostate cancer that balances the benefits and harms of biopsy, in light of the fact that the cancer states are not directly observable and can progress stochastically over time. In Chapter 4, we propose a new stochastic dynamic programming model, i.e., MPOMDP model, with hidden states that include two or more alternative models to address the issue of parameter ambiguity in POMDPs.

Our study achieves individualized medical decision-making from three aspects. First, the HMM described in Chapter 2 allows different patients to have different cancer progression paths. Further, the proposed POMDP optimization model in Chapter 3 generalizes the HMM by including decision-making with the goal of finding optimal biopsy policies for different individuals according to the estimates of their cancer progression paths. Second, the reward functions of the proposed POMDPs in Chapter 3 and MPOMDPs in Chapter 4 are defined upon a so-called "reward parameter". When applying our optimization models, the decision-makers can set the value of the reward parameter according to their own considerations of two competing objectives, which are to minimize the burden of cancer surveillance and to minimize the late detection to cancer progression. Third, when there are multiple plausible optimization models, the proposed MPOMDP in Chapter 4 can learn the model credibility for each patient according to his past surveillance actions and observations, to identify an individualized optimal policy.

Summary of main contributions. We provide a summary of the main contributions of each chapter as follows.

Chapter 2 focuses on the estimation of a stochastic model of prostate cancer for men

newly diagnosed with low-risk cancer, using the electronic health record data from several well-known prostate cancer studies. The main contributions of Chapter 2 are:

- We present a novel HMM to describe the stochastic system of prostate cancer AS, which involves partially observable cancer states, multiple kinds of cancer screenings, and stochastic relationships among them. The HMM has the flexibility to be applied to other types of cancer for which surveillance is relevant.
- Our results of estimating HMMs from longitudinal data provide important findings in the context of urology, including miss-classification error at diagnosis, biopsy undersampling error, PSA test accuracy, and prostate cancer progression rate, in four major prostate cancer AS studies worldwide: the Johns Hopkins (JH) Hospital, University of California San Francisco (UCSF) medical center, University of Toronto (U of T) medical center, and the Prostate Cancer Research International Active Surveillance (PRIAS) project. Moreover, we use the bootstrapping method to establish that the prostate cancer progression rate and the test accuracy in different study centers are statistically significantly different.
- We present a simulation study based on the HMMs to compare different published biopsy protocols across four study centers. The objective is to minimize the mean number of biopsies over a patient's lifetime and the mean delay time to detection of a prostate cancer progression. Our results show that there is no single best biopsy protocol for all patients in four studies because of the parameter ambiguity, thus providing evidence for the potential need to accommodate alternative surveillance protocols for different patients.

The results of the work presented in Chapter 2 were published in Li et al. (2020). The findings formed the basis of a case study for the next two chapters, which extend the

descriptive stochastic models to the optimization context.

Chapter 3 presents a POMDP to optimize prostate cancer AS in four study centers. The research objective of Chapter 3 is to determine the optimal strategy for AS, which balances the harm of diagnostic testing with the benefit of early detection of high-risk cancer. Moreover, we solve the POMDP model for each of the studies we considered in Chapter 2 to identify areas of ambiguity based on where the model recommendations differ. The main contributions of Chapter 3 are:

- We propose the first POMDP model to optimize the individualized biopsy policy for patients in the four prostate cancer AS studies of Chapter 2, which balances the harm of biopsy with the benefit of early detection of cancer progression. The POMDP model formulation is built on the descriptive HMM models of Chapter2.
- We analyze the structural properties of the proposed model to develop fast approximation methods and provide insight into the optimal policy. We also investigate the relationship between model-based dynamic policies that learn the optimal action based on observed data acquired as patients age, such as the POMDP model provides, and static pre-defined policies that have been recommended in the clinical literature.
- We evaluate the impact of ambiguity caused by variation across models fitted to different clinical studies, as well as variation in the reward criteria. We also use the models to estimate the implied reward parameters by inverse optimization, to establish the degree to which patient preferences vary with regard to the propensity for biopsies.
- We provide the first estimates – to our knowledge – of the potential impact of MRI as a means for early detection of cancer progression in the prostate cancer AS setting.

The work presented in Chapter 3 was submitted to an operations research journal, where the draft can be found in Li et al. (2021). The finding formed the basis of a case study for the next chapter, which extends the stochastic optimization model to account for parameter ambiguity.

Chapter 4 presents a new POMDP model that we refer to as a Multi-model POMDP MPOMDP. The research objectives are to address the issue of parameter ambiguity in POMDPs, and to study the benefit of accounting for parameter ambiguity in POMDP models. The main contributions of Chapter 4 are:

- We propose a new framework, i.e., MPOMDP, to address the issue of parameter ambiguity in POMDPs. Unlike other literature in robust optimization, the proposed MPOMDP model considers multiple credible POMDP models simultaneously, and seeks a single optimal policy based on learning the credibility of each POMDP model from the system outputs over time. We present numerical experiments that demonstrate some of the attractive properties of the MPOMDP policy, such as the robustness with respect to parameter ambiguity.
- We present some important structural properties of the proposed MPOMDP model, which not only motivate the solution methods, but also help analyze the effect of parameter ambiguity in POMDPs.
- We propose two fast approximation methods suited to solving MPOMDP models, which are shown to converge asymptotically. We use numerical experiments to demonstrate the trade-off between the solution quality and computation time.
- We apply the MPOMDP to the case study for prostate cancer AS optimization with ambiguity. Our results show that the MPOMDP model can find a biopsy policy that is only slightly worse than the optimal biopsy policy given by the correct POMDP

model of the same study center, but much better than the policies given by the wrong POMDP models or the mean-value POMDP model which uses the mean of model parameters. Given how the trade-off between the biopsy burden and late detection of a cancer progression by the decision-maker, the MPOMDP model achieved the minimum expected future costs when the true model was not known with certainty.

The thesis concludes with Chapter 5, which summarises the most important findings in Chapters 2-4 and some limitations of this thesis that lead to opportunities for future research.

CHAPTER 2

Comparison of Biopsy Under-sampling and Annual Progression Using Hidden Markov Models to Learn from Prostate Cancer Active Surveillance Studies

2.1 Introduction

Although early detection is key to preventing prostate cancer death, many patients are diagnosed with low-risk cancer that is unlikely to cause harm (Miller et al., 2006). Prostatectomy and radiation therapy are associated with potentially serious side effects, including incontinence, erectile dysfunction, and others (Anandadas et al., 2011). Therefore, definitive treatment of low-risk prostate cancer may cause more harm than good. AS is a form of expectant management, but in which a switch to curative treatment can be made as a result of tumor risk reclassification at any time. AS strategies involve monitoring patients through a combination of Digital Rectal Exams (DREs), PSA tests, selective use of imaging, and surveillance biopsies. AS defers or avoids definitive treatment until there is evidence of cancer misclassification or progression, thus reducing overtreatment of low-risk prostate cancer. PSA tests and DREs are minimally invasive, but they have poor predictive performance. Biopsy is the gold standard, but it involves sampling tissue from the prostate with hollow-core needles, which can be painful, costly, and may result

in infections. While PSA tests and DREs are routine elements of AS, they are far less informative than prostate biopsy for determining disease risk in this setting.

There are two main challenges when deciding the optimal biopsy plan for a given patient on AS. First, the true cancer state of each patient is not observable unless the patient is treated with radical prostatectomy, because biopsies are associated with under-sampling error. Second, patients who start AS may later progress from favorable to non-favorable risk over time due to cancer evolution. Moreover, the biopsy under-sampling errors and cancer progression rates are unknown and may vary among different cohorts. A related study estimated biopsy under-sampling error assuming no cancer progression during AS (Coley et al., 2017) and another study that estimated progression rate assuming perfect prostate biopsy (Inoue et al., 2018). There is one study (Barnett et al., 2018a) that considered biopsy under-sampling and prostate cancer progression simultaneously but it was based on a single very low-risk cohort and did not utilize PSA or treatment outcomes for model estimation. There is no study we are aware of that considers cancer progression and biopsy misclassification across multiple cohorts.

In this chapter, we estimated and compared the misclassification error of favorable risk cancer at diagnosis, subsequent cancer progression rate, biopsy sensitivity and specificity, and PSA distribution in four of the most well-known AS cohorts using the dataset (version 3.1) created by the Movember Foundation' Global Action Plan Prostate Cancer Active Surveillance (GAP3) (Bruinsma et al., 2018). We used an HMM to estimate the stochastic model that best describes the longitudinal observational data for each of the four cohorts. We further used the estimated models as the basis for a simulation model to compare previously published biopsy protocols across the four cohorts. We analyzed the differences in model estimates across the four cohorts and validated the results using bootstrapping. Finally, we compared the mean number of biopsies for patients on AS and the mean delay

Cohort	Number of Patients	Inclusion Criteria for AS	Biopsy Protocol
JH	1,434	clinical stage $\leq$ T1c, PSA density $\leq$ 0.15, Gleason score $\leq$ 6, total positive core $\leq$ 2, single core positivity $\leq$ 50%	Biopsy every year
UCSF	1,644	clinical stage T1-T2, PSA $\leq$ 10, Gleason score $\leq$ 6, total positive core $\leq$ 1/3 of total cores, single core positivity $\leq$ 50%	Biopsy 1 year after diagnosis, then every 1 to 2 years
U of T	1,243	clinical stage T1c/T2a, PSA $\leq$ 10, Gleason score $\leq$ 6	Biopsy 1 year after diagnosis, then every 3 years
PRIAS	4,700	clinical stage T1c/T2, PSA $\leq$ 10, PSA density $\leq$ 0.2, Gleason score $\leq$ 6, total positive core $\leq$ 2	Biopsy 1 year after diagnosis, then every 3 years

Table 2.1: The inclusion criteria and biopsy protocols of four major prostate cancer AS cohorts.

time to detection of non-favorable risk prostate cancer for the biopsy protocols previously proposed for each of these cohorts to assess variation in outcomes across cohorts.

2.2 Materials and Methods

2.2.1 Data

In 2014, the Movember Foundation launched the GAP3 plan initiative to create a global database tracking the selection and monitoring of men with low-risk prostate cancer on AS (Bruinsma et al., 2018). The database records the clinical and demographic characteristics of 20,652 patients on AS from 27 established cohorts worldwide (v3.1). In this study, we chose four cohorts including the two largest AS study cohorts in the USA: JH hospital (Tosoian et al., 2011) and UCSF medical center (Dall'Era et al., 2008), the largest AS study in Canada: U of T medical center (Klotz et al., 2010), and the largest AS study outside North America: the PRIAS project (Bul et al., 2013). These four cohorts not only include the greatest number of patients, but also have the most AS follow-up records over time. Importantly, these cohorts have different inclusion criteria and recommended surveillance strategies. Table 1 illustrates inclusion criteria and biopsy protocols in the four cohorts. The research was approved by the Institutional Review Board at the University of Michigan.

2.2.2 Natural History Models Based on HMMs

We formulated an HMM to determine the misclassification error of favorable risk cancer at diagnosis due to diagnosis test error, annual progression rate to non-favorable risk cancer, and follow-up biopsy under-sampling error for patients on AS in each of the four studies. HMMs are well suited to this analysis because prostate cancer progresses stochastically over time, and the true cancer state cannot be observed directly (it is hidden due to the imperfect accuracy of PSA testing and prostate biopsies) unless the patient is treated by radical prostatectomy. We defined the favorable risk cancer state as the cancer state that meets the inclusion criteria in each cohort in Table 2.1 and defined the non-favorable risk cancer state as any cancer state that does not meet the criteria, and thus represent cancer states for which patients may consider treatment rather than AS. Table 1 shows that the definitions of favorable and non-favorable risk cancer vary by cohort.

By definition of these cohorts, all patients were diagnosed with favorable risk prostate cancer and initiated AS as their initial management. However, due to the potential measurement error in DREs, PSA test, and biopsy, some patients starting AS were actually in the non-favorable risk cancer state at the time of diagnosis. We use the term misclassification at diagnosis to refer to instances where a patient with non-favorable risk prostate cancer is incorrectly diagnosed with favorable risk prostate cancer at the time of initiating AS. The probability of misclassification at diagnosis was estimated by the initial distribution of the HMM. Every year after initiating AS, patients may also progress from favorable risk cancer to non-favorable risk cancer with some annual progression rate, which determines the transition probability matrix in the proposed HMM. Figure 1 shows the state transition diagram of prostate cancer in the context of AS.

The observations used to fit the HMM were PSA level and biopsy. We did not consider other covariates including clinical stage, total positive cores in biopsy, single-core positiv-

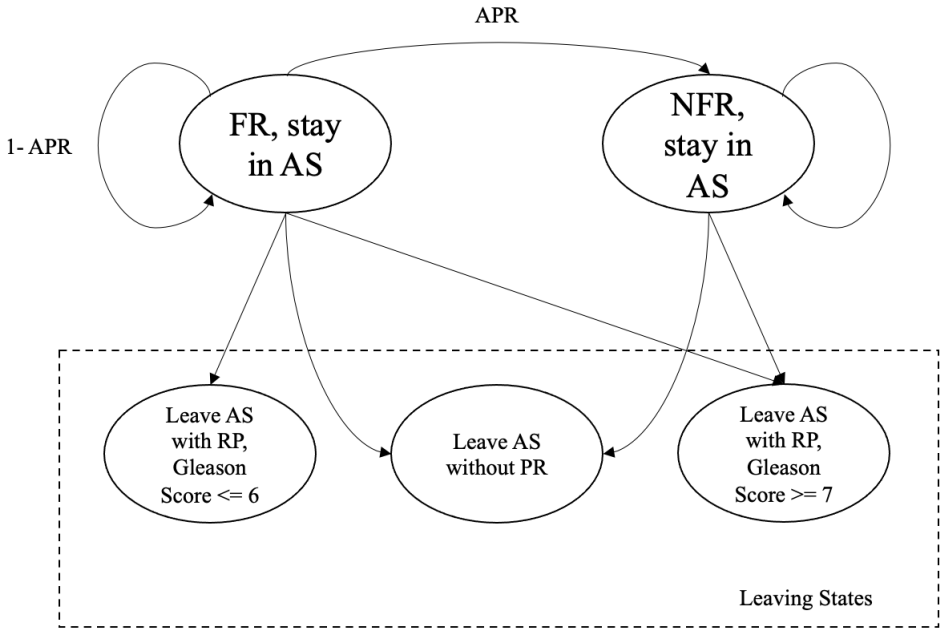

Figure 2.1: State transition diagram of prostate cancer in the context of AS. There are two hidden states and three observable states in the formulated HMM. Abbreviations: FR, favorable risk; NFR, non-favorable risk; APR, annual progression rate; RP, radical prostatectomy.

ity in biopsy, and Magnetic Resonance Imaging (MRI) scan because of the lack of data. In all four studies, PSA tests were routinely performed at office visits (every 3-6 months), while biopsies were performed at most once per year and often less frequently because of the design of biopsy protocols or other patient and clinical factors. Therefore, in our model, we set the frequency of test outcomes to be annual, which means we only used the most recent PSA test and one biopsy result at the end of each calendar year as observations for this annual time period. We also defined a null observation for instances of a missing test result. Given that biopsies are not perfect, we use the term biopsy under-sampling to denote the circumstance where there was a Gleason score 6 or lower biopsy result in a patient with (hidden) non-favorable risk cancer. The biopsy sensitivity (defined as rate of biopsy Gleason score 7 or higher while in the non-favorable risk cancer state) and specificity (defined as rate of biopsy Gleason score 6 or lower while in favorable risk cancer

state), and the distribution of the PSA testing result were estimated by the observation probability distributions in the HMM. Finally, every year, the patient might leave AS with or without treatment. If the patient left AS and underwent the radical prostatectomy, then his true cancer grade (Gleason score) was available based on post-radical prostatectomy pathology. Otherwise, the patient was assumed to leave AS without knowledge of the true cancer grade. Given this context, we defined the leaving states of the HMM as follows: 1) leaving AS with true Gleason score 6 or lower based on prostatectomy pathology, 2) leaving AS with true Gleason score 7 or higher based on prostatectomy pathology, and 3) leaving AS without radical prostatectomy. The probabilities of entering the leaving states were also elements of the transition probability matrix.

We used the Baum-Welch algorithm to fit the proposed HMM (Rabiner, 1989). The Baum-Welch algorithm is a special form of the standard EM (expectation-maximization) algorithm (Dempster et al., 1977), which iteratively updates the estimates of model parameters that locally maximize the likelihood function of given sequences of observations. To avoid local maxima, we randomly chose different starting points of the parameters before running the iterations, and then picked the set of estimated parameters with the largest likelihood function as the final estimates. For different cohorts, we fitted different HMMs with the same model structure but different parameters.

2.2.3 Solving the HMM

Given the observation sequences, our goal here is to estimate the parameters in the HMM which maximize the likelihood of the observation sequences. Since the likelihood function for the HMM is often irregular and non-convex, there is no known way to analytically solve the maximization problem. However, we can use an iterative procedure such as the Baum-Welch algorithm (Baum et al., 1970) to solve the parameters such that the

likelihood function is locally maximized. Next, we show how the classic work of Baum and his colleagues can be applied here to solve our proposed HMM.

The Baum-Welch algorithm is a special case of the general EM algorithm (Dempster et al., 1977), which can be described in the following steps: 1) initialize the model parameters; 2) calculate the likelihood function of the observation sequences given the initialized model parameters; 3) update the model parameters such that the calculated likelihood function is maximized; 4) go back to step 2 and recalculate the likelihood function using the updated model parameters; 5) repeat step 2 to 4 until convergence of the model parameters. We then describe how does each step work when solving our model.

The complete parameter set of our HMM is

$$\lambda = (b_0, P, F),$$

where b_0 is the initial belief (distribution) over all states, $P : S \times S$ is the state transition probability (cancer progression rate), and $F : S \times O$ is the observation probability. Here, since we have two types of observations from PSA test and biopsy, we can further write

$$F = F_x \times F_y,$$

where $F_x : S \times O_x$ is the PSA observation probability and $F_y : S \times O_y$ is the biopsy observation probability. Further, we use a Gaussian mixture model to describe the continuous distribution of PSA observation:

$$F_x(i,x) = \sum_{k=1}^{K} c_{ik} \mathcal{N}(x; \mu_k, \sigma_k), \ 1 \leq i \leq m$$

where c_{ik} is the mixture coefficeient for k^{th} mixture in state i, $\mathcal{N}(\mu_k, \sigma_k)$ is a Gaussian distribution with mean vector μ_k and variance σ_k, and K is the total number of mixtures. Here we choose $K = 2$.

Parameter Initialization. To initialize the parameter, we random sample each parameter from the uniform distribution. The other thing about the parameter initialization is that

many numerical experiments have shown that the performance of the EM (or equivalently Baum-Welch) algorithm is often sensitive to the choice of the initial parameters (Melnykov and Melnykov, 2012). So to avoid local maximization of the likelihood function (or equivalently sub-optimal estimations), we typically generate several different samples of the initial parameters, and run the algorithm separately before concluding the final result.

Likelihood Computation. Suppose in the dataset, we have N observation sequences denoted as

$$O = (O^1, ..., O^N),$$

where $O^n = (O_1^n, ..., O_{T_n}^n)$ for $n = 1, ..., N$ with T_n being the number of time epochs for the n^{th} observation sequence, and $O_t^n = (X_t^n, Y_t^n)$ being the observation at time t of the n^{th} observation sequence. To calculate the likelihood (i.e. probability) of the observation O given the model parameter λ, which can be written as $\mathbb{P}(O|\lambda)$, a straightforward way is to enumerate every possible state sequence for each observation sequence. However, such a straightforward method is very inefficient. The computation effort for enumerating all possible state sequence is a exponential function of the number of states and the length of each observation sequence. Here, we introduce another efficient method called the forward-backward procedure (Baum and Eagon, 1967), whose computation effort is polynomial in the number of states and the length of each observation sequence.

For each observation sequence O^n, consider the forward variable $\alpha_t^n(i)$ defined as

$$\alpha_t^n(i) := \mathbb{P}(O_1^n, ..., O_t^n, S_t^n = i|\lambda),\ t = 1, ..., T_n,\ n = 1, ..., N,\ i = 1, ..., m$$

which is the probability of the partial observation sequence $O_1^n, ..., O_t^n$ and state i at time t, given the model parameter λ. The forward variable can be solved inductively. For each observation sequence O^n, $n = 1, ..., N$:

1) initialization:

$$\alpha_1^n(i) = b_0(i)F_x(i,X_1^n)F_y(i,Y_1^n),\ i=1,...,m. \tag{2.1}$$

2) Induction:

$$\alpha_{t+1}^n(j) = [\sum_{i=1}^{m} \alpha_t^n(i)P(i,j)]F_x(j,X_t^n)F_y(j,Y_t^n),\ t=1,...,T_n-1,\ i=1,...,m. \tag{2.2}$$

3) Termination:

$$P_n := \mathbb{P}(O^n|\lambda) = \sum_{i=1}^{m} \alpha_{T_n}^n(i). \tag{2.3}$$

and at last, $\mathbb{P}(O|\lambda) = \prod_{n=1}^{N} P_n$. Notice here we consider two types of the observation. The derivation of equation 2.1, 2.2 and 2.3 is shown in Appendix.

Similarly, we can consider the backward variable $\beta_t^n(i)$ for each O^n defined as

$$\beta_t^n(i) := \mathbb{P}(O_{t+1}^n,...,O_{T_n}^n|S_t^n=i,\lambda),\ t=1,...,T_n-1,\ n=1,...,N,\ i=1,...,m$$

which is the probability of the partial observation sequence from time $t+1$ to the end, given state $S_t^n = i$ and the model parameter λ. The backward variable can be solved inductively. For each observation sequence O^n, $n=1,...,N$:

1) initialization:

$$\beta_{T_n}^n(i) = 1,\ i=1,...,m. \tag{2.4}$$

2) Induction:

$$\beta_t^n(i) = \sum_{j=1}^{m} P(i,j)F_x(j,X_{t+1}^n)F_y(j,Y_{t+1}^n)\beta_{t+1}(j),\ t=T_n-1,...,1,\ i=1,...,m. \tag{2.5}$$

Again, the derivation of equation 2.4 and 2.5 is shown in Appendix.

Parameter Update. The update formulas of the Baum-Welch algorithm can be derived directly by maximizing Baum's auxiliary function discussed in Baum and Sell (1968).

Here, we only provide the results of the update formulas, considering both continuous and discrete observations in the dataset. Here, we use the hat notation to denote the updated parameter.

1) Initial distribution b_0:

$$\begin{aligned}\hat{b}_0(i) &= \text{expected frequency in state i at time 1} \\ &= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{P_n}\alpha_1^n(i)\beta_1^n(i).\end{aligned}$$

2) Probability distribution matrix P:

$$\begin{aligned}\hat{P}(i,j) &= \frac{\text{expected number of transitions from state i to j}}{\text{expected number of transitions from state i}} \\ &= \frac{\sum_{n=1}^{N}\frac{1}{P_n}\sum_{t=1}^{T_n-1}\alpha_t^n(i)P(i,j)F_x(j,X_{t+1}^n)F_y(j,Y_{t+1}^n)\beta_{t+1}^n(j)}{\sum_{n=1}^{N}\frac{1}{P_n}\sum_{t=1}^{T_n-1}\alpha_t^n(i)\beta_t^n(i)}\end{aligned}$$

3) Discrete observation probability F_y:

$$\begin{aligned}\hat{F}_y(i) &= \frac{\text{expected number of times in state i with observation y}}{\text{expected number of times in state i}} \\ &= \frac{\sum_{n=1}^{N}\frac{1}{P_n}\sum_{t:Y_t=y}\alpha_t^n(i)\beta_t^n(i)}{\sum_{n=1}^{N}\frac{1}{P_n}\sum_{t=1}^{T_n}\alpha_t^n(i)\beta_t^n(i)}\end{aligned}$$

4) Continuous observation probability F_x: to estimate the parameters of F_x, we first define

$$\gamma_t(i,k) := \frac{\alpha_t\beta_t(i)}{\sum_i \alpha_t\beta_t(i)}\cdot\frac{c_{ik}\mathcal{N}(X_t|\mu_{ik},\sigma_{ik}^2)}{\sum_{k=1}^{K}c_{ik}\mathcal{N}(X_t|\mu_{ik},\sigma_{ik}^2)}$$

as the probability of being in state i at time t with the k^{th} mixture component accounting

for X_t. Then the update formulas can be written as:

$$\begin{aligned}\hat{c}_{ik} &= \text{expected frequency of the } k^{th} \text{ mixture accounting for } X \text{ in state i} \\ &= \frac{\sum_{n=1}^{N}\sum_{t=1}^{T_n}\gamma_t(i,k)}{\sum_{n=1}^{N}\sum_{n=1}^{T_n}\sum_{k=1}^{K}\gamma_t(i,k)}\end{aligned}$$

$$\begin{aligned}\hat{\mu}_k &= \text{empirical mean of } X \text{ in } k^{th} \text{ mixture component} \\ &= \frac{\sum_{n=1}^{N}\sum_{t=1}^{T_n}\sum_{s_i}\gamma_t(i,k)X_t}{\sum_{n=1}^{N}\sum_{t=1}^{T_n}\sum_{s_i}\gamma_t(i,k)}\end{aligned}$$

$$\begin{aligned}\hat{\sigma}_k^2 &= \text{empirical covariance of } X \text{ in } k^{th} \text{ mixture component} \\ &= \frac{\sum_{n=1}^{N}\sum_{t=1}^{T_n}\sum_{s_i}\gamma_t(i,k)(X_t-\mu_k)^2}{\sum_{n=1}^{N}\sum_{t=1}^{T_n}\sum_{s_i}\gamma_t(i,k)}\end{aligned}$$

Stopping Criteria. The Baum-Welch algorithm uses the above formulas to iteratively calculate the likelihood of the observation sequences and update the parameter λ. The stopping criteria of these iterative procedures can be specified by defining a tolerance parameter τ (e.g. $\tau = 10^{-6}$). We can stop the algorithm if the difference between the old and new likelihood is less than τ.

The Baum-Welch algorithm for solving the HMM is summarized in Algorithm 1.

Algorithm 1: Baum-Welch algorithm for HMMs.

Data: Independent observation sequences $O = (O^1, ..., O^N)$
initialization $r = 0$, $p^{old} = 0$, $\lambda^0 = (b_0^0, P^0, F^0)$;
compute the forward variable α^0 and backward variable β^0 using λ^0;
compute $p^{new} = \mathbb{P}(O|\lambda^0)$ using α^0 and β^0;
while $|p^{old} - p^{new}| > \tau$ **do**
  $r \leftarrow r+1$;
  $p^{old} \leftarrow p^{new}$;
  update $\lambda^r = (b_0^r, P^r, F^r)$ from α^{r-1}, β^{r-1} and λ^{r-1};
  compute the forward variable α^r and backward variable β^r using λ^r;
  compute $p^{new} = \mathbb{P}(O|\lambda^r)$ using α^r and β^r;
end

2.2.4 Statistical Analysis and Validation

To estimate confidence intervals (CIs) of the estimated model parameters in different cohorts, we used the non-parametric bootstrap method to compute the standard errors of estimated parameters (Efron, 1992). Specifically, for each cohort, we first randomly sampled patients with replacement. The number of sampled patients was equal to the sample size of the cohort. For each bootstrap sample, we then fitted an HMM using the observation sequences of the bootstrap sample. We drew 100 bootstrap samples and used the empirical standard errors and confidence intervals as the estimates of the standard errors and confidence intervals of the estimated parameters in this cohort. We repeated the same steps for all four cohorts.

We focused on internal validation in this study, because different cohorts had different study inclusion criteria. We validated the estimated models by comparing the observed and estimated distributions of the results of PSA test and biopsy. For PSA results, we compared the empirical and estimated distribution for both favorable risk cancer and non-favorable risk cancer patients. For biopsy results, we first simulated patients' underlying cancer states and biopsy observations (if the biopsy protocol suggested a biopsy) in each cohort using a simulation model (described in next) with the estimated model parameters. Then, we compared the observed and simulated biopsy positive rates at each biopsy time.

2.2.5 Biopsy Protocols Comparison by Simulation Model

We used the estimated HMMs to create a simulation model for each cohort to compare the mean number of biopsies performed while on AS and the mean delay in time to detection of non-favorable risk cancer by biopsy. Hypothetical patients in the simulation model were assumed to be diagnosed with favorable risk cancer at age 50 in different cohorts when using the different biopsy protocols described in Table 1. For each patient,

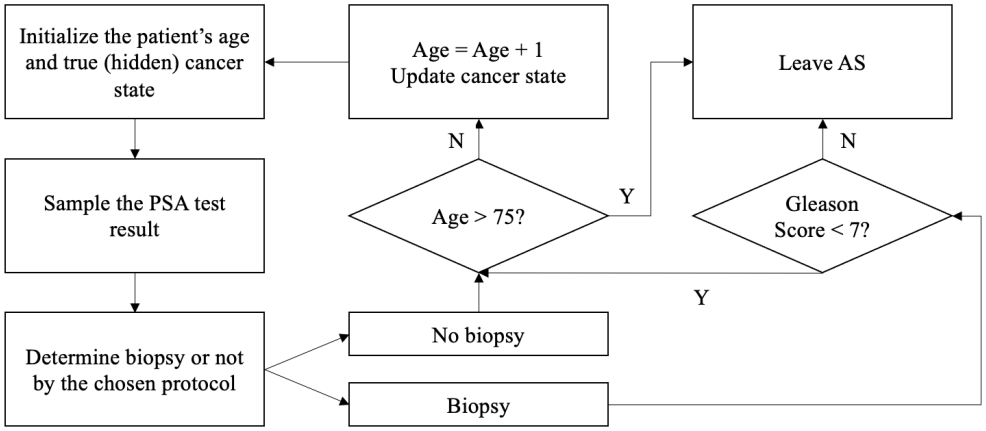

Figure 2.2: Simulation process flow for the proposed simulation model. The model parameters were determined by the estimates of the HMMs. Patients would leave the AS if they had a Gleason score 7 or higher biopsy, or they reached age 75.

we first sampled his initial cancer state when diagnosed with favorable risk cancer at age 50 according to the misclassification error at diagnosis as estimated by the HMM, and initiating AS. Second, for the next annual time-point, we simulated his new cancer state based on the previous cancer state and the estimated annual cancer progression rate. With the simulated cancer state, we then sampled the patient's PSA result using the estimated PSA probability density distribution. If a biopsy was indicated according to the chosen protocol, we sampled the biopsy result based on the estimated sensitivity and specificity of the biopsy obtained from the HMM for the cohort. If the sampled biopsy Gleason score was greater or equal to 7 at that time point, then the patient left AS; otherwise, the patient continued on AS for another year. Patients reaching age 75 were assumed to stop AS and transit to watchful waiting. The details of the simulation process flow are shown in Figure 2.

With the simulated true cancer states and biopsy results for all patients at all time periods, the mean number of biopsies performed while on AS was calculated as the average number of follow-up biopsies performed from initiating AS (age 50) to leaving AS (age

75 or a Gleason score 7 or higher biopsy), while the mean delay in time to detection of non-favorable risk cancer was calculated as the average difference between the time of the first sampled non-favorable risk cancer state and the time of a sampled Gleason score 7 or higher biopsy results for all patients. The number of sampled patients was set to 10,000 for each cohort and each protocol.

2.3 Results

2.3.1 Data

Table 2 summarizes patient characteristics at the time of diagnosis for patients with at least one follow-up year on AS. The means of age at diagnosis were similar in all four cohorts except UCSF, where patients were younger than compared to the other three cohorts. In terms of PSA levels and biopsy results, JH enrolled patients with lower PSA, lower maximum percentage of cancer in biopsy cores, and lower Gleason score than other three cohorts. UCSF and University of Toronto medical centers enrolled patients with the highest PSA level and percentage of patients with Gleason 3+4=7 or greater cancer. Additional information about patient characteristics at the time of each biopsy in AS can be found in Table S1-S4 in Appendix B.

As we can see from Table 2, some patients with medium/high-grade (non-favorable risk) cancer were also included in the AS. Those patients were generally older patients who continued on AS instead of moving on to treatment. For the purpose of our study, we removed those patients with medium/high-grade cancer at diagnosis when fitting the HMMs.

Cohort	JH	UCSF	U of T	PRIAS
Patients, n	1434	1644	1243	4700
Age at biopsy, year, mean (SD)	66 (6.1)	63 (7.6)	66 (8.1)	66 (6.9)
Months since diagnosis, month, mean (SD)	0 (0)	0 (0)	0 (0)	0 (0)
PSA, ng/mL, mean (SD)	5.2 (2.9)	6.4 (4.1)	6.2 (3.1)	5.9 (2.1)
No. of biopsy cores used, median (range)	12 (6-58)	14 (1-50)	10 (1-190)	12 (3-25)
Maximum % of cancer in any one core (SD)	10 (14.8)	26 (20.8)	21 (20)	NA (NA)
% of cores with cancer	12 (7.1)	17 (13.8)	23 (18.1)	13 (6.7)
ISUP grade group, # (%)				
No cancer	0 (0)	0 (0)	0 (0)	0 (0)
1 (3 + 3)	1428 (99.6)	1437 (87.4)	1104 (88.8)	4657 (99.1)
2 (3 + 4)	6 (0.4)	178 (10.8)	139 (11.2)	42 (0.9)
3 (4 + 3)	0 (0)	25 (1.5)	0 (0)	1 (0)
4 (4 + 4)	0 (0)	4 (0.2)	0 (0)	0 (0)
5 (9, 10)	0 (0)	0 (0)	0 (0)	0 (0)
NA	0 (0)	0 (0)	9 (0.7)	3 (0.1)
Medium/High-grade cancer (%)	6 (0.4)	207 (12.6)	139 (11.2)	43 (0.9)

Table 2.2: Patient Characteristics at the Time of Diagnosis. Abbreviations: SD, standard deviation; ISUP, International Society of Urologic Pathologists; NA, not available.

2.3.2 HMM Analysis and Validation

Table 3 and Figure 3 show the estimates of the most important HMM parameters for each cohort and the 95% confidence intervals estimated via the bootstrap method. The differences in the estimated annual cancer progression rates and biopsy sensitivities for distinct cohorts were statistically significant ($p<0.05$). The estimated annual progression rate from favorable risk cancer to non-favorable risk cancer was highest in UCSF and lowest in JH. Biopsy sensitivity was highest in PRIAS, with the highest proportion of non-favorable risk cancer patients correctly identified on biopsy; while JH had a slightly lower biopsy sensitivity than other three cohorts. In terms of misclassification errors at diagnosis, the proportion of patients considered to have non-favorable risk cancer at diagnosis was highest in UCSF and lowest in JH. All estimated biopsy specificities were close to 100%. In addition, based on the estimated 95% confidence intervals by bootstrapping, the estimated miss-classification errors at diagnosis, annual cancer progression rates, and (1- biopsy sensitivity)'s in the four cohorts are all statistically significantly greater than

Cohort	Number of Patients	Misclassification Error at Diagnosis	Annual Progression Rate	Biopsy Sensitivity	Biopsy Specificity
JH	1428	0.0583	0.0691	0.7184	0.9972
UCSF	1437	0.0809	0.1217	0.7431	0.9925
U of T	1104	0.0774	0.1016	0.7949	0.9962
PRIAS	4657	0.0653	0.0841	0.7614	0.9920

Table 2.3: Estimated Parameters by the HMMs for Different Cohorts.

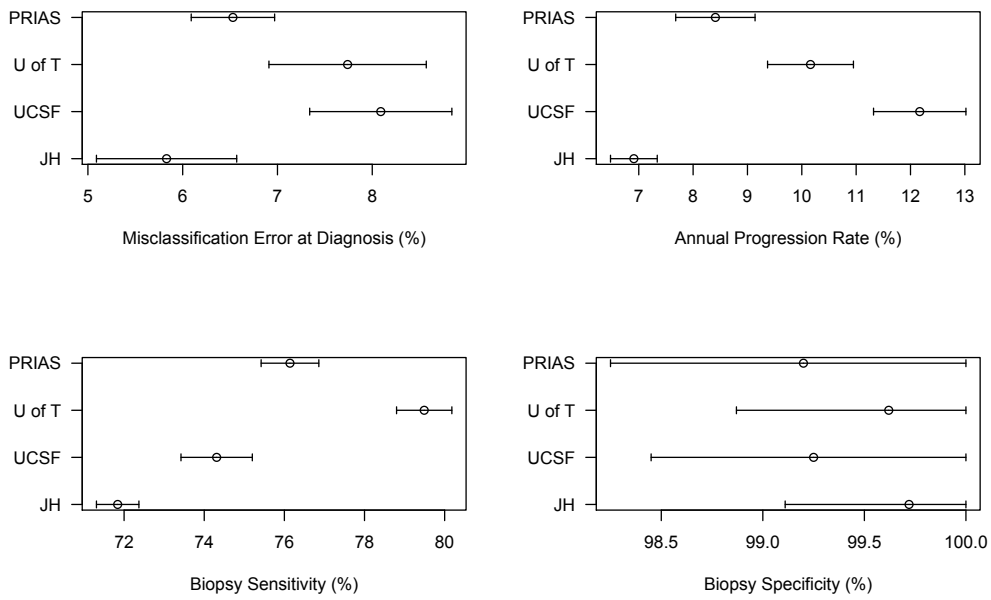

Figure 2.3: Estimated standard errors and 95% confidence intervals for the parameters in HMMs by the bootstrap method. All misclassification errors at diagnosis, annual cancer progression rates, and (1- biopsy sensitivity)'s are statistically significantly greater than 0.

zero.

For the estimates of the PSA distributions, we assumed that the logarithm of the PSA result follows a mixture of two Gaussian distributions. The details of the estimated parameters for the mixture distribution can be found in Table S5 and S6 in Appendix B.

We validated our models by comparing the biopsy positive rates and PSA probability density functions between observed and simulated data. Figure 4 shows the comparisons between observed and simulated biopsy positive rates for different cohorts, which were calculated as the number of patients with a positive biopsy (Gleason score 7 or higher

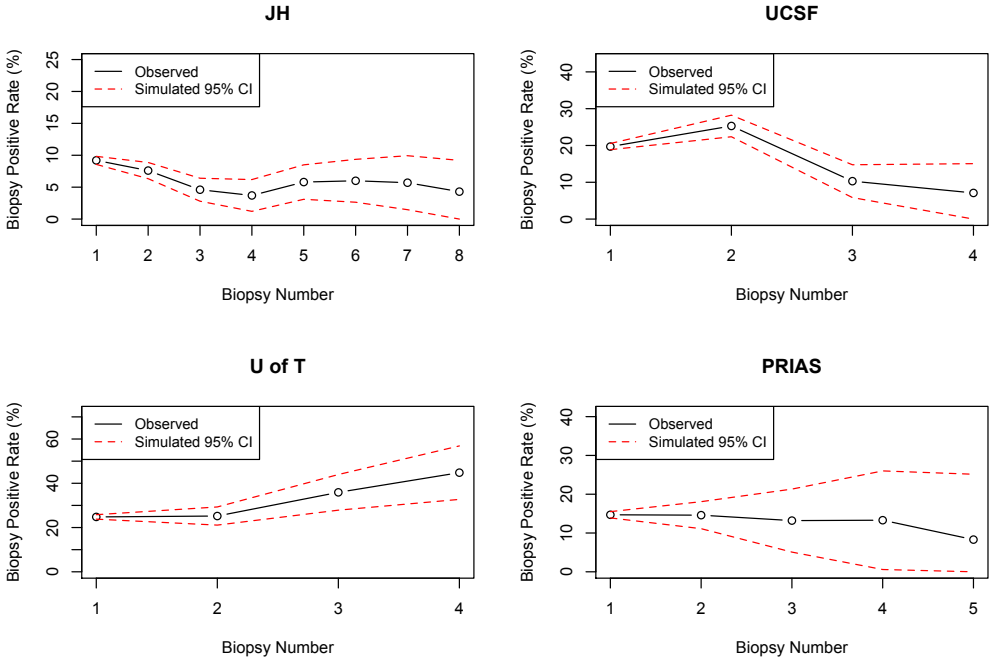

Figure 2.4: Comparison of observed and simulated biopsy positive rates at each biopsy time for different cohorts. All observed biopsy detection rates fell into the 95% CIs of the simulated detection rates.

Cohort	**JH**			**UCSF**			**U of T**			**PRIAS**		
Biopsy protocol	JH	UCSF	U of T	JH	UCSF	U of T	JH	UCSF	U of T	JH	UCSF	U of T
Mean number of biopsies	12.6	7.1	5.3	8.7	5.3	4.1	9.7	5.8	4.4	11.1	6.4	4.9
Average late detection time by biopsy (month)	4.5	13.9	22.9	5.0	15.2	26.0	3.8	12.7	21.7	4.0	13.2	22.3

Table 2.4: Comparisons of the mean number of biopsies used and average late detection time by biopsy between the time of diagnosis and the end of AS for different protocols in different cohorts by the proposed simulation model.

biopsy) results divided by the total number of patients who underwent biopsy in the observed and simulated datasets, at each biopsy time point. The observed positive biopsy rates all fell into the 95% confidence intervals of the simulated biopsy positive rates. The comparisons of PSA distributions are shown in Figure S1-S4 in Appendix B.

2.3.3 Comparison of Biopsy Protocols

We simulated a population of 5,000 patients for each cohort and each biopsy protocol using the simulation model described in Figure 2. Each patient was assumed to be diag-

nosed as favorable risk cancer and enter AS at age 50. We compared the frequency of biopsy and the mean delay in time to detection of non-favorable risk cancers between the time of diagnosis (age 50) and the end of AS. Table 4 shows the simulation results for all protocols in fours cohorts. In each cohort, the protocol employing fewer biopsies was associated with a longer late detection time on average. Also, if we compare the differences in the mean number of biopsies used and mean delay in detection by biopsy between different protocols, we can see the benefit from more frequent biopsies was diminishing.

2.4 Discussion

We estimated the misclassification error at diagnosis, the annual cancer progression rate, the sensitivity and specificity of biopsy, and the distribution of PSA in four prostate cancer AS cohorts part of the GAP3 consortium: JH, UCSF, U of T, and PRIAS. With the estimated HMMs, we then compared the mean number of biopsies performed versus late detection of cancer progression by biopsy when following different published biopsy protocols in four cohorts using a series of simulations. As expected, in each cohort, the biopsy protocol that recommended more frequent biopsies was associated with shorter time to reclassification. Our results show that because of the considerable variation in biopsy under-sampling error and annual progression rates across cohorts, there was no single best biopsy protocol that is optimal for all cohorts. Moreover, in each cohort, the biopsy protocol that recommended more frequent biopsies was associated with shorter time to reclassification, while the benefit from additional biopsies was diminishing.

Other studies have also tried to quantify the most important factors associated with testing errors and cancer progression rate on AS. Coley et al. (2017) proposed a Bayesian hierarchical model that included PSA and biopsy as covariates to predict the latent cancer state in the JH AS cohort. They estimated the misclassification error at diagnosis to

be between 20% and 31%, and the biopsy sensitivity to be 62%. The reason why their measurement error was much higher than ours was that they assumed there was no cancer progression during AS for any patient. For our fitted HMMs, we do see that estimates of both cancer progression rate and biopsy under-sampling error are statistically significantly greater than 0, as the bootstrapping 95% confidence intervals do not include 0. Thus, if we apply the bootstrap-t hypothesis test method discussed in Efron and Tibshirani (1994) to the estimates of both cancer progression rate and biopsy under-sampling error, we can reject the null hypothesis that the estimated parameter is equal to 0 with the type I error less than 5%. Also, the definition of the biopsy sensitivity in our study, is defined with respect to the non-favorable risk cancer state as defined in each of the studies as opposed to Gleason score alone, used by Coley and colleagues.

Another study by Barnett et al. (2018a), fit an HMM to estimate the cancer grade progression rate and biopsy under-sampling errors in the JH AS cohort only. They estimated the annual progression rate from Gleason score 6 cancer to Gleason score 7 or higher cancer to be 4.0%; then sensitivity and specificity of biopsy to be 61.0% and 98.6%. There are a number differences in their approach compared to our study. For example, they did not incorporate PSA observations or observations of radical prostatectomy or alternative treatment options, which can reveal the true cancer states, for patients to leave AS. Moreover, they considered only the JH cohort which was a very low risk patient cohort. Thus, we believe our model in this study was more informative than their model. A study by Inoue et al. (2018), which compared the biopsy upgrading rates in four prostate cancer AS cohorts including JH, UCSF, U of T, and Canary prostate cancer AS study cohorts found a statistically significant difference in biopsy upgrading risk for different cohorts. However, they did not account for possible biopsy Gleason score false-negative result and misclassification error.

In our results from estimating the HMMs for four different cohorts, based on the bootstrapped standard errors of the estimated parameters, all the mis-classification errors at diagnosis, annual cancer grade progression rates, and biopsy false-negative rates were statistically significantly greater than zero. This validates our assumptions about the non-zero progression rate in contrast to the above-referenced study by Coley et al. (2017) that assumes no progression, and the imperfect biopsy sensitivity in contrast to the study by Inoue et al. (2018) that assumes zero misclassification error and zero biopsy false-negative rate. All biopsy specificities were close to 100%, indicating it was very rare that a patient in favorable risk cancer state would have a biopsy Gleason sum 7 or higher. For mis-classification errors at the time of diagnosis and annual grade progression rates, we found that the estimates in the UCSF and U of T cohorts were greater than the estimates in JH and PRIAS cohorts. This was consistent with the fact that the UCSF and U of T cohorts included higher-risk patients than other two cohorts, which can also be seen in the summary statistics of PSA density, maximum percentage of cancer in any one core, and percentage of cores with cancer at the time of diagnosis in Table 2. For the biopsy sensitivities, we saw that JH cohort had the lowest estimate while the U of T cohort had the highest one. Our conjecture was that patients with lower risk had smaller tumors in general, so that they were harder to detect by biopsy if they were in non-favorable risk cancer state. Other possible reasons for such differences might include the different definition of favorable and non-favorable risk states, and the difference in the urologist practice when performing the tests in different cohorts. Our simulation study compared three published biopsy protocols in different cohorts. Within each cohort, the protocol that recommended more biopsies had less late detection years of non-favorable risk cancer by biopsy. However, we saw that the benefit in terms of early detection was diminishing along with the increasing number of biopsies. There was no single optimal protocol that recommended fewer biopsies but

could detect non-favorable risk cancer earlier, in any cohort. Two main reasons are: first, the model parameters estimated by the HMMs and used in the simulation model were statistically significantly different for different cohorts; second, there were two competing objectives when comparing the protocols that are minimizing the number of biopsies and minimizing the late detection time by biopsy.

There were some notable limitations in our study. First, we reduced a complex disease (prostate cancer) to a two-state (favorable and unfavorable risk) stochastic model with two outputs of the disease (results of PSA test and biopsy) as informative observations. However, although such models cannot capture all details about the disease, it consistently discriminates health states on the basis of the most significant factors defining study inclusion for each cohort. Second, our proposed HMM included the null observation of biopsy as non-informative missingness. In other words, we assumed no difference between a missed biopsy by the design of the study, and a missed biopsy result for other reasons (e.g. patient preference, data lost to follow-up). However, by using the null observation to denote the biopsy missingness in the HMM, we mitigated bias in our estimates of the model parameters. Finally, another way to monitor prostate cancer in recent AS protocols is by MRI scans, but it was not considered in this study due to the lack of sufficient longitudinal data.

The above limitations notwithstanding, our study quantified the most important factors in four prostate cancer AS cohorts, providing a number of insights into the role of different study designs and populations on AS. We found there was no single optimal biopsy protocol across cohorts and we provided evidence that there may be considerable variation in characteristics of prostate cancer across cohorts. This is likely explained by some combination of factors including: 1) differences in disease dynamics between the different cohorts due to variations in the inclusion criteria, and thus different definitions

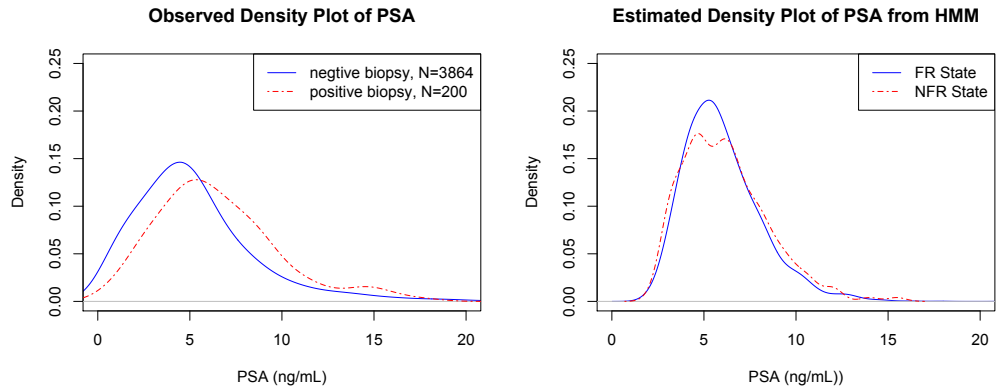

Figure 2.5: Observed and estimated density plots of the PSA in Johns Hopkins hospital.

of favorable vs. non-favorable risk prostate cancer. and 2) variation in healthcare delivery across health systems resulting from different practices in urology and pathology.

2.5 Appendix: Supporting Tables and Figures

	Biopsy								
Characteristics	Diagnosis	First	Second	Third	Fourth	Fifth	Sixth	Seventh	Eighth
Patients, n	1434	1229	776	524	349	224	134	88	47
Age at biopsy, year, mean (SD)	66 (6.1)	67 (6.2)	67 (6.1)	68 (5.5)	68 (5.4)	69 (5.3)	70 (4.7)	70 (4.2)	71 (4.2)
Months since diagnosis, month, mean (SD)	0 (0)	14 (13.2)	29 (16.3)	41 (15.4)	54 (14.4)	68 (15)	82 (15.5)	96 (16.1)	107 (14.8)
PSA, ng/mL, mean (SD)	5.2 (2.9)	5.3 (3.4)	5.4 (4.4)	5.4 (3.9)	5.3 (3.6)	5.6 (4.6)	5.7 (4.5)	5.3 (4.4)	4.7 (3.3)
No. of biopsy cores used, median (range)	12 (6-58)	12 (4-31)	12 (6-60)	12 (6-28)	12 (8-16)	14 (9-24)	14 (6-15)	14 (6-15)	14 (8-14)
Maximum % of cancer in any one core (SD)	10 (14.8)	22 (24.6)	17 (20.8)	13 (18.7)	18 (21.1)	17 (18.1)	14 (17.9)	18 (19.4)	15 (17.9)
% of cores with cancer	12 (7.1)	10 (12.4)	7 (10.3)	6 (8.9)	6 (8.9)	6 (8)	6 (8.7)	7 (8.7)	7 (8.9)
Gleason group, # (%)									
No cancer	0 (0)	519 (42.2)	386 (49.7)	289 (55.2)	191 (54.7)	127 (56.7)	72 (53.7)	43 (48.9)	23 (48.9)
1 (3 + 3)	1428 (99.6)	594 (48.3)	330 (42.5)	209 (39.9)	145 (41.5)	84 (37.5)	54 (40.3)	40 (45.5)	22 (46.8)
2 (3 + 4)	6 (0.4)	76 (6.2)	42 (5.4)	12 (2.3)	9 (2.6)	8 (3.6)	7 (5.2)	2 (2.3)	2 (4.3)
3 (4 + 3)	0 (0)	23 (1.9)	11 (1.4)	11 (2.1)	3 (0.9)	4 (1.8)	1 (0.7)	2 (2.3)	0 (0)
4 (4 + 4)	0 (0)	11 (0.9)	4 (0.5)	0 (0)	1 (0.3)	1 (0.4)	0 (0)	1 (1.1)	0 (0)
5 (9, 10)	0 (0)	3 (0.2)	2 (0.3)	1 (0.2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
NA	0 (0)	3 (0.2)	1 (0.1)	2 (0.4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Medium/High-grade cancer (%)	6 (0.4)	113 (9.2)	59 (7.6)	24 (4.6)	13 (3.7)	13 (5.8)	8 (6)	5 (5.7)	2 (4.3)

Table 2.5: Biopsy Characteristics for patients in Johns Hopkins hospital.

	Biopsy					
Characteristics	Diagnosis	First	Second	Third	Fourth	Fifth
Patients, n	1644	279	99	39	14	4
Age at biopsy, year, mean (SD)	63 (7.6)	64 (7.7)	65 (7.2)	67 (7.2)	69 (6.8)	69 (3.8)
Months since diagnosis, month, mean (SD)	0 (0)	25 (24.8)	42 (26.5)	63 (32.1)	82 (22.2)	109 (14.4)
PSA, ng/mL, mean (SD)	6.4 (4.1)	5.8 (5.8)	5.6 (3.5)	8.7 (18)	6 (4.8)	4.1 (2.5)
No. of biopsy cores used, median (range)	14 (1-50)	16 (2-31)	17 (4-25)	16 (5-27)	17 (14-26)	17.5 (14-22)
Maximum % of cancer in any one core (SD)	26 (20.8)	5 (7.3)	5 (6.9)	4 (4.9)	5 (5.1)	5 (5)
% of cores with cancer	17 (13.8)	17 (16.7)	17 (16.8)	14 (14.8)	15 (14.2)	16 (11.8)
Gleason group, # (%)						
No cancer	0 (0)	72 (25.8)	27 (27.3)	11 (28.2)	3 (21.4)	0 (0)
1 (3 + 3)	1437 (87.4)	152 (54.5)	47 (47.5)	24 (61.5)	10 (71.4)	3 (75)
2 (3 + 4)	178 (10.8)	39 (14)	20 (20.2)	2 (5.1)	1 (7.1)	1 (25)
3 (4 + 3)	25 (1.5)	12 (4.3)	3 (3)	2 (5.1)	0 (0)	0 (0)
4 (4 + 4)	4 (0.2)	3 (1.1)	1 (1)	0 (0)	0 (0)	0 (0)
5 (9, 10)	0 (0)	1 (0.4)	1 (1)	0 (0)	0 (0)	0 (0)
NA	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Medium/High-grade cancer (%)	207 (12.6)	55 (19.7)	25 (25.3)	4 (10.3)	1 (7.1)	1 (25)

Table 2.6: Biopsy Characteristics for patients in UCSF medical center.

	Biopsy					
Characteristics	Diagnosis	First	Second	Third	Fourth	Fifth
Patients, n	1243	911	385	131	29	4
Age at biopsy, year, mean (SD)	66 (8.1)	67 (8.2)	68 (7.6)	69 (7.4)	70 (7.3)	69 (10.4)
Months since diagnosis, month, mean (SD)	0 (0)	22 (17.5)	58 (25)	97 (31.7)	136 (39.2)	148 (47.2)
PSA, ng/mL, mean (SD)	6.2 (3.1)	7.7 (6.3)	10 (13.1)	9.3 (8.2)	11.7 (9.3)	NA (NA)
No. of biopsy cores used, median (range)	10 (1-190)	10 (2-27)	10 (3-250)	10 (5-170)	10 (5-13)	7 (6-14)
Maximum % of cancer in any one core (SD)	21 (20)	32 (26)	33 (25.6)	39 (27.1)	44 (23.9)	60 (NA)
% of cores with cancer	23 (18.1)	24 (24.5)	25 (28.9)	33 (31.2)	38 (38.5)	7 (14.3)
Gleason group, # (%)						
No cancer	0 (0)	226 (24.8)	128 (33.2)	29 (22.1)	8 (27.6)	3 (75)
1 (3 + 3)	1104 (88.8)	400 (43.9)	151 (39.2)	52 (39.7)	7 (24.1)	0 (0)
2 (3 + 4)	139 (11.2)	160 (17.6)	64 (16.6)	33 (25.2)	9 (31)	1 (25)
3 (4 + 3)	0 (0)	48 (5.3)	24 (6.2)	11 (8.4)	2 (6.9)	0 (0)
4 (4 + 4)	0 (0)	9 (1)	5 (1.3)	0 (0)	1 (3.4)	0 (0)
5 (9, 10)	0 (0)	9 (1)	4 (1)	3 (2.3)	1 (3.4)	0 (0)
NA	9 (0.7)	72 (7.9)	9 (2.3)	3 (2.3)	1 (3.4)	0 (0)
Medium/High-grade cancer (%)	139 (11.2)	226 (24.8)	97 (25.2)	47 (35.9)	13 (44.8)	1 (25)

Table 2.7: Biopsy Characteristics for patients in Toronto medical center.

	Biopsy						
Characteristics	Diagnosis	First	Second	Third	Fourth	Fifth	Sixth
Patients, n	4700	3535	1226	342	90	12	3
Age at biopsy, year, mean (SD)	66 (6.9)	67 (6.9)	68 (6.9)	69 (6.7)	70 (6.5)	72 (7.2)	70 (7.9)
Months since diagnosis, month, mean (SD)	0 (0)	14 (8)	41 (14.3)	63 (20.1)	77 (22.2)	84 (19.8)	87 (18.3)
PSA, ng/mL, mean (SD)	5.9 (2.1)	6.1 (3.3)	6.8 (3.7)	7.3 (4.2)	8.1 (4.3)	8.8 (3.4)	13.2 (1.6)
No. of biopsy cores used, median (range)	12 (3-25)	12 (3-25)	12 (2-25)	12 (3-25)	12 (6-25)	10 (8-12)	12 (10-12)
Maximum % of cancer in any one core (SD)	NA (NA)	NA (NA)	NA (NA)	NA (NA)	NA (NA)	NA (NA)	NA (NA)
% of cores with cancer	13 (6.7)	12 (14.8)	11 (13.5)	11 (15.6)	9 (10.9)	7 (11)	40 (19.2)
Gleason group, # (%)	NA	NA	NA	NA	NA	NA	NA
No cancer	0 (0)	1319 (37.3)	493 (40.2)	158 (46.2)	35 (38.9)	8 (66.7)	0 (0)
1 (3 + 3)	4657 (99.1)	1668 (47.2)	540 (44)	136 (39.8)	42 (46.7)	3 (25)	1 (33.3)
2 (3 + 4)	42 (0.9)	374 (10.6)	112 (9.1)	33 (9.6)	7 (7.8)	0 (0)	2 (66.7)
3 (4 + 3)	1 (0)	90 (2.5)	31 (2.5)	7 (2)	4 (4.4)	1 (8.3)	0 (0)
4 (4 + 4)	0 (0)	46 (1.3)	30 (2.4)	4 (1.2)	1 (1.1)	0 (0)	0 (0)
5 (9, 10)	0 (0)	9 (0.3)	6 (0.5)	1 (0.3)	0 (0)	0 (0)	0 (0)
NA	3 (0.1)	30 (0.8)	14 (1.1)	3 (0.9)	1 (1.1)	0 (0)	0 (0)
Medium/High-grade cancer (%)	43 (0.9)	519 (14.7)	179 (14.6)	45 (13.2)	12 (13.3)	1 (8.3)	2 (66.7)

Table 2.8: Biopsy Characteristics for patients in the PRIAS project.

Center	Prob. of C1 in FR state	Prob. of C1 in NFR state	Mean of C1	Mean of C2	SD of C1	SD of C2
JH	0.4730	0.3381	1.09	2.04	0.95	0.55
UCSF	0.0602	0.0434	1.15	2.15	1.20	0.45
Toronto	1.0000	0.2650	1.49	2.37	0.97	1.74
PRIAS	0.2238	0.1620	1.45	2.16	0.90	0.44

Table 2.9: Estimated of the mixture Gaussian distribution of the log(PSA) in different medical centers.

Range of PSA (ng/mL)		< 4	[4, 10]	>10
JH	FR Cancer	0.3552	0.4311	0.2137
	NFR Cancer	0.2868	0.4706	0.2426
UCSF	FR Cancer	0.0768	0.5680	0.3552
	NFR Cancer	0.0678	0.5736	0.3586
Toronto	FR Cancer	0.4573	0.3422	0.2005
	NFR Cancer	0.3312	0.2368	0.4320
PRIAS	FR Cancer	0.1361	0.5357	0.3282
	NFR Cancer	0.1094	0.5501	0.3405

Table 2.10: Estimated PSA distribution in different cohorts.

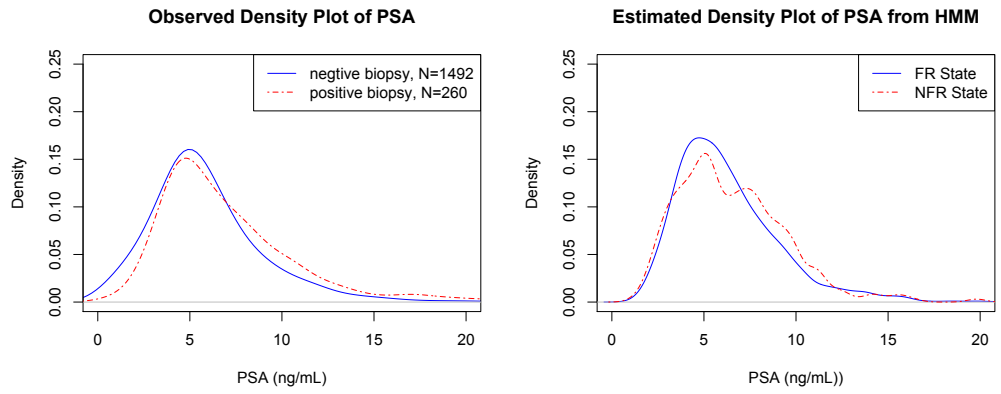

Figure 2.6: Observed and estimated density plots of the PSA in UCSF medical center.

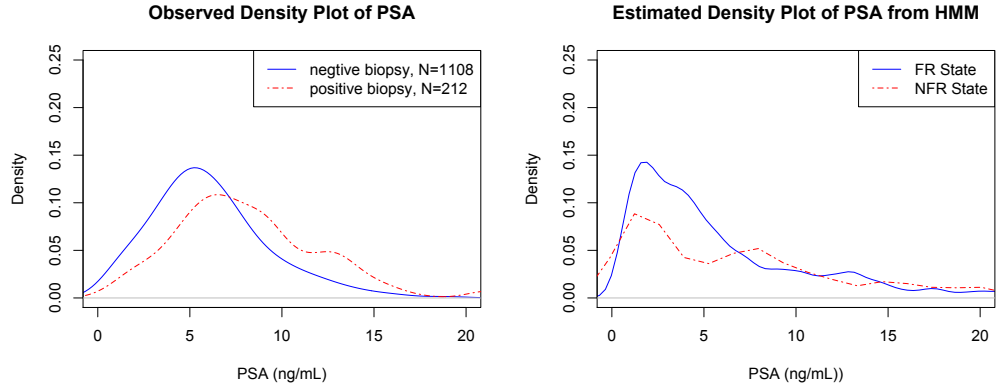

Figure 2.7: Observed and estimated density plots of the PSA in University of Toronto medical center.

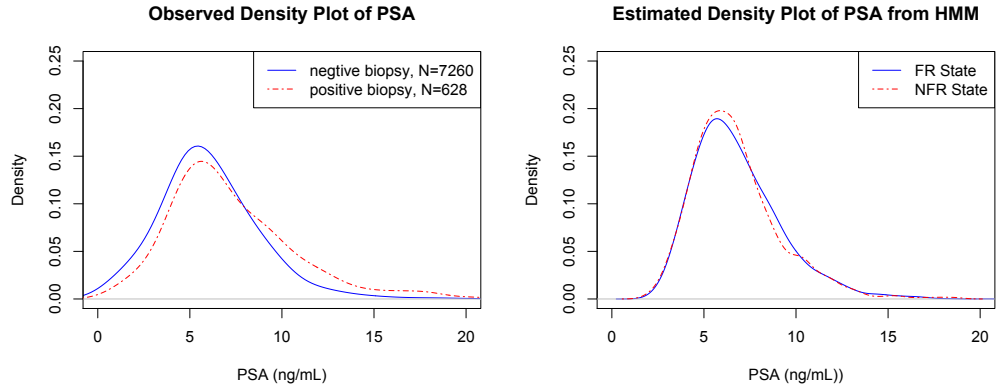

Figure 2.8: Observed and estimated density plots of the PSA in PRIAS dataset.

CHAPTER 3

Optimizing Active Surveillance for Prostate Cancer Using Partially Observable Markov Decision Processes

3.1 Introduction

Prostate cancer is the most common cancer in men. The American Cancer Society estimates that almost 250,000 new prostate cancer cases and more than 34,000 deaths will occur in the United States in 2021. Over the last decade, it has become clear that men with low-risk variants of prostate cancer can safely avoid major treatment like surgery and radiation therapy, which may have significant side effects including incontinence and erectile dysfunction (Anandadas et al., 2011). For this reason, AS has recently become the recommended approach for patients with low-risk prostate cancer. AS involves monitoring patients over time to test for evidence of cancer progression to a high-risk variant of the disease. This allows low-risk cancer patients to enjoy a higher quality of life and possibly avoid treatment altogether (Klotz, 2013).

AS involves regular testing to monitor a patient's health status. The PSA test is a simple blood test in AS that measures PSA amount in the blood serum. High PSA is associated with the presence of prostate cancer. Because the PSA test is very simple with almost no harm, it is commonly used; but high false-positive and negative rates make it unsuitable for

AS on its own. Prostate biopsy is the gold standard for AS, which involves sampling tissue with hollow-core needles during an outpatient procedure. Biopsy results are reported using the Gleason grading system, where a Gleason score is assigned by a pathologist to provide a measure of severity of the prostate cancer. Biopsy is much more accurate than the PSA test, but it is still prone to false-negative results if the extracted tissue samples miss the tumor. Biopsies are also very painful, and have potential side effects. Thus, decisions about when to perform biopsies are among the most important decisions for AS.

Unfortunately, there is a lack of consensus among urologists on the best biopsy policy. As one of the first healthcare centers to investigate AS, JH recommended annual biopsies for patients enrolled in AS (Tosoian et al., 2011). A more recent study conducted by the UCSF medical center recommends biopsy every two years after diagnosis (Dall'Era et al., 2008). A study at the U of T medical center and another European study, the PRIAS project, recommend biopsy every three years after diagnosis for patients enrolled in AS (Klotz et al., 2009; Bul et al., 2013).

Deciding the optimal biopsy policy is challenging because: 1) the patient's cancer state is not directly observable due to the inaccuracy of diagnostic tests; 2) cancer progression is a stochastic process; 3) patient preferences about how often to biopsy vary. To address these challenges, we formulated a finite-horizon two-state POMDP model to optimize the biopsy policy for AS using data from the four largest and most well known AS studies referenced above. POMDPs are well suited to this type of optimization problem because the decision-makers (physicians and patients) need to make decisions under conditions of uncertainty about the underlying health state, which progresses stochastically over time, and can only be partially observed from PSA test and biopsy results. Our model seeks to find the optimal biopsy policy that trades off two competing criteria: expected delays in detecting high-risk prostate cancer and the expected number of biopsies.

POMDP models are usually very hard to solve over long time horizons because of the *curse of history* (Pineau et al., 2003). Moreover, the model must be solved multiple times, as we will show in this study, to account for ambiguity in the reward function and the underlying stochastic system associated with each of the cohorts mentioned above. For this reason, we present fast approximation methods that can quickly compute near-optimal solutions. We compare our model-based policies, solved via the approximation methods, to established biopsy guidelines from the literature. We further use inverse optimization to estimate ranges of the implied decision-maker's weights on the two reward criteria. Finally, we combine the results for each of the cohorts to compute a risk-based policy region that partitions the region into three parts: 1) biopsy always recommended; 2) biopsy never recommended; 3) shared decision-making between the patient and physician is necessary to decide if a biopsy should be performed.

The remainder of this chapter is organized as follows. In Section 2, we review the relevant literature and describe our contributions to the literature. In Section 3, we formulate our Active Surveillance Partially Observable Markov Decision Process (AS-POMDP) model to optimize the biopsy policy in prostate cancer AS. We describe the exact solution method and two approximation methods for the AS-POMDP model in Section 4, and prove some structural properties of the AS-POMDP model in Section 5. In Section 6, we present the results of optimal policies in our case study. Finally, we conclude in Section 7 and discuss some potential directions for future research.

3.2 Literature Review

In this section, we briefly review the most relevant literature from the application and methodological perspectives. We then summarize our main contributions in light of the existing literature.

3.2.1 Applications

Much clinical research has been done in recent years to study prostate cancer AS. Several review articles, including Bastian et al. (2009); Klotz (2010); Dall'Era et al. (2012) and Thomsen et al. (2014), have discussed the clinical implication of prostate cancer AS with the focus on inclusion criteria, biopsy guidelines, patient outcomes, and future research needed. The urology community has largely converged on the appropriateness of AS for patients with low-risk cancer. However, different centers have proposed different AS guidelines, which vary most significantly in the recommended frequency of biopsies (Dall'Era et al., 2008; Klotz et al., 2009; Tosoian et al., 2011; Bul et al., 2013).

Epstein et al. (2012) presented results for predictive risk factors for outcomes of radical prostatectomy, which were instrumental in laying the framework for selection criteria for AS enrollment. More recently Coley et al. (2017) built a Bayesian hierarchical model to estimate the sensitivity and specificity of biopsy, and predict the latent cancer states in the JH study, while assuming no cancer progression. Barnett et al. (2018a) estimated a HMM to estimate the biopsy accuracy and cancer progression rate implied by observed data in the JH study. They further compared different biopsy guidelines using a simulation model based on the HMM. A recent study by Li et al. (2020) used a HMM to estimate the cancer progression rate, biopsy under-sampling error, and PSA distribution in the four largest AS cohorts, including JH hospital, UCSF medical center, U of T medical center, and the PRIAS project. The descriptive models given by this study provide the foundation for the prescriptive POMDP models we present here.

POMDP models have been found to be successful in recent decades for optimizing medical decisions when the health state is not directly observable. POMDP models applied to clinical decision-making include the study of screening based on mammography for breast cancer (Simmons Ivy et al., 2009; Ayer et al., 2012, 2016; Otten et al., 2020),

colonoscopy screening for colorectal cancer (Erenay et al., 2014), and liver transplantation decisions in the context of liver disease (Sandıkçı et al., 2013). The most related work to ours – and the only other work on POMDPs for prostate cancer that we are aware of – is that of Zhang et al. (2012a) and Zhang et al. (2012b), which used a POMDP model to optimize the one-time biopsy policy (i.e., the best timing for biopsy if only one biopsy is allowed) in prostate cancer screening, to maximize patients' Quality Adjusted Life Years (QALYs). Their work focused on screening of healthy patients who are asymptomatic, the vast majority of whom receive at most one biopsy. Thus, their model can be viewed as an *optimal stopping time problem*, as opposed to AS that involves a continuous process of sequential follow-up biopsies.

3.2.2 Methodology Literature

POMDP models were first studied by Åström (1965); Drake (1962) and Smallwood and Sondik (1973), and they have been applied in many contexts including machine maintenance (Ross, 1971), robot navigation (Cassandra et al., 1996), healthcare (Ayer et al., 2012; Zhang et al., 2012a; Erenay et al., 2014), and many others (see Cassandra (1998) for a survey). Smallwood and Sondik (1973) introduced the first exact solution method, referred to as the *one-pass algorithm*, for finite-horizon POMDP models. White (1991) and Littman et al. (1995) later proposed the more efficient *witness algorithm* that achieves computational efficiency through a refined approach for identifying the supporting hyperplanes that define the optimal value function. Zhang and Liu (1996) and Cassandra et al. (1997) introduced the *incremental pruning* algorithm, which has been found to be one of the most efficient exact algorithms for a number of problems. Despite its utility for real world applications, solving POMDP models exactly has been shown to be NP-hard, and in PSPACE (Vlassis et al., 2012), due to the so-called *curse of dimensionality* (Kaelbling

et al., 1998) and *curse of history* (Pineau et al., 2003).

Many approximation methods for the POMDP model have been studied in the past several decades. An early survey by Lovejoy (1991) discussed exact solution methods for finite-horizon POMDP models in theory, and their finite-memory and finite-grid approximations. Kaelbling et al. (1998) explored function-approximation methods for approximating the value function of POMDP models. Hauskrecht (2000) surveyed various value-function approximation methods for infinite-horizon problems in the application of agent navigation, analyzed their properties and relations, and also presented some novel approximation methods and refinements of existing methods. Unlike the finite-horizon problem, the infinite-horizon POMDP assumes a stationary (i.e., time-independent) value function, with the discounting factor for future rewards being strictly less than one. Pineau et al. (2003) formally defined the *point-based value iteration* (PBVI) algorithm for infinite-horizon POMDPs, and proved the estimation error is bounded. Spaan and Vlassis (2005) introduced the *Perseus* algorithm, which is closely related to PBVI. A more recent survey of point-based POMDP solvers for infinite-horizon problems was published by Shani et al. (2013). Although there were a number of instances where the existing approximation methods were found to be efficient, the issues of finding the best upper bound of the value function with a guaranteed error bound, especially in finite-horizon problems, can easily become intractable and remains unsolved.

Another topic of interest in the literature has been establishing monotonicity of optimal policies, since such policies can be easier to understand and implement, and maybe easier to solve. Ross (1971) first investigated the monotonicity of the optimal policy in a two-state production process described by a POMDP model. Albright (1979) proved the sufficient conditions for the monotonicity of the optimal policy in a two-state POMDP with the restriction that the actions are taken to improve, rather than investigate the system.

Other works include White (1979), Lovejoy (1987), and Miehling and Teneketzis (2020), which generalized this property to models with more than two states by defining the partial order of the belief.

References	Topic	How it differs from our work
Simmons Ivy et al. (2009)	Screening and treatment for breast cancer	Built a simulation method to evaluate policies using the POMDP model instead of solving for optimal policies.
Ayer et al. (2012)	Screening for breast cancer	The proposed POMDP model was to improve patients' QALYs. The optimal value function was monotone in belief. Only considered exact solution methods, which took more than 55 hours for a single model.
Zhang et al. (2012a)	Screening for prostate cancer	The objective was to improve patients' QALYs. Assumed one-time decision because patients could have at most one biopsy.
Sandıkçı et al. (2013)	Liver transplantation for liver disease	The objective was to improve patients' QALYs. The model had monotone optimal value function. Considered an approximate solution method that incorporated solving an LP at each decision epoch, without a bound on approximation error.
Erenay et al. (2014)	Screening for colorectal cancer	The proposed POMDP model was to improve patients' QALYs
Ayer et al. (2016)	Screening for breast cancer with imperfect adherence behavior	Similar model setting as in Ayer et al. (2012), but incorporates adherence behavior to policies. Only considered exact solution method, which took more than 153 hours for a single model.
Otten et al. (2020)	Post-treatment screening for breast cancer	Similar model setting as in Ayer et al. (2012), but the state space was continuous. Optimized the mammography decision within 10-year follow-up after treatment.

Table 3.1: Previous work on POMDP models for medical decision-making in different disease contexts. All have different model structures in terms of states, actions, and optimality equations.

3.2.3 Contributions to the Literature

Our work makes a novel contribution to the literature in several ways. First, we propose the first model to optimize individualized biopsy policies for prostate cancer AS patients. Our study has a model structure that differs from many previously formulated POMDPs, including −− but not limited to −− those arising in clinical contexts that are summarized in Table 3.1. We describe the approach we used to formulate this complex clinical problem, which is naturally expressed as a two-state POMDP, and then we evaluate the model using observational data from the four most well-known studies of AS thus far. Second, we analyze the model to provide theoretical insight into the structure of the optimal policy. We also discuss the relationship between model-based dynamic policies that learn based on observed data acquired as patients age, such as our POMDP model provides, and static pre-defined policies that have been recommended in the clinical literature. Third, we provide

a means to consider the impact of ambiguity caused by variation across models fitted to different clinical studies as well as variation in the reward criteria. Finally, our work collectively demonstrates the full spectrum of using clinical study data directly to estimate and solve POMDP models for an important medical decision-making problem affecting many men worldwide.

3.3 POMDP Model Formulation

In this section, we describe the discrete-time finite-horizon POMDP model we use to optimize the policy for prostate cancer AS. As noted in the introduction, the objectives are minimizing 1) expected delay in detection of high-risk prostate cancer; 2) expected number of lifetime biopsies. Clearly, it would be ideal to minimize both of these objectives, but that is not possible because they are competing; therefore, we settle for minimizing a weighted combination of the two criteria. We start by describing two main assumptions that form the basis for POMDP model formulation.

Assumption 3.1. *Prostate cancer progression can be described using a finite-state (two-state) Markov chain.*

Assumption 1 simplifies the stochastic process of prostate cancer progression to that of a first-order Markov chain. The finite state assumption naturally follows from the binary discrimination of health states on the basis of risk as determined by clinical thresholds using pathology information. We describe additional details about this when we discuss the model formulation.

Assumption 3.2. *The probability distributions of PSA test and biopsy results are conditionally independent given the current cancer state of the patient.*

Assumption 2 assumes conditional independence for different observations given the state

of the process. It is a common assumption in partially observable stochastic models that describes the causal relations between the underlying state and the associated observations, and can be adapted to the study of prostate cancer AS. Assumption 1-2 have been validated in a related study of HMMs for prostate cancer by Li et al. (2020).

With the main modeling assumptions established, we now define the elements of the proposed discrete-time finite-horizon AS-POMDP model. We also describe lesser but still important assumptions as part of the model description.

Decision Epochs. We index $t = 1, ..., T$ as the discrete-time periods (also referred to as decision epochs) at which the decision-maker can choose to biopsy, and the state transitions happen. In the AS-POMDP model, t is an annual epoch and the decision is made at the start of the epoch followed by the state transition. Epochs occur annually because this is an upper bound on the frequency of biopsies according to clinical guidelines, i.e., no guideline suggests biopsies more frequently than annually. Epoch $t = 1$ denotes the time of diagnosis and enrollment in AS, and epoch $t = T$ is the recommended stopping time for AS among patients who survive until age T, which is typically age 75 according to clinical guidelines due to increases in competing causes of death.

States. The set of states, S, contains two states: 1) low-risk prostate cancer state (LR); 2) high-risk prostate cancer state (HR). In reality, there are numerous health states defined by risk factors, including PSA and pathology from biopsies; however, urologists differentiate these states into two groups (LR and HR) to align clinical risk with treatment choices. Patients who are known to be in the LR state should continue AS, while those in the HR state should be treated (e.g., surgery or radiation therapy). We use s_t to denote the state of the system at time t for $t = 1, ..., T$.

Actions. The set of available actions, A, contains two elements: 1) defer biopsy; 2) conduct biopsy. As the PSA test is always done by default according to standard clinical

practice, the critical decision at each decision epoch is whether or not to conduct a biopsy. Note that in prostate AS, the action of conducting biopsy is to investigate, rather than to improve, the patient's cancer state. In other words, conducting a biopsy does not affect the stochastic process of cancer progression (unless the patient leaves AS for treatment because of a biopsy Gleason score upgrading defined later).

Transition Probabilities. At each decision epoch, the system undergoes state transitions according to transition probability P defined as follows:

$$P(i,j) := \mathbb{P}(s_{t+1} = j | s_t = i),\ \forall i,j \in S,\ \forall t = 1,...,T-1.$$

In our AS-POMDP model, the state can only progress from LR cancer to HR cancer, so that we use p to denote this annual progression rate.

Observations. At each decision epoch, after an action is taken, the PSA test and biopsy result (if conducted) will be observed. We denote O as the set of all possible observations, and $o_t \in O$ as the observation at time t for $t = 1,...,T$. By Assumption 2, at any decision epoch, given the state of the system, the observations of PSA test and biopsy are mutually independent. So, $O = O_{\text{PSA}} \times O_{\text{Biopsy}}$, where O_{PSA} is the observation space of the PSA test, and O_{Biopsy} is the observation space of the biopsy. We discretize the space of the measurement of PSA levels into three intervals, according to the widely used PSA cutoffs in clinical studies (Hoffman, 2011), so that O_{PSA} has three elements: $I_1 = [0,4]$, $I_2 = (4,10]$, and $I_3 = (10,\infty)$ (ng/mL). For biopsy, the elements in O_{Biopsy} are Gleason score upgrading (biopsy Gleason score greater or equal to 7), Gleason score not upgrading (biopsy Gleason score less or equal to 6), and null observation (biopsy not conducted). Such definition is based on the fact that the inclusion criteria for AS in all four study centers considered in this chapter require the biopsy Gleason score to be less or equal to 6 (Li et al., 2020). We now state the third assumption of the AS-POMDP model formulation as follows.

Assumption 3.3. *Patients leave AS immediately when a biopsy Gleason score upgrading is observed.*

Assumption 3 is reasonable because Gleason score upgrading is a common criterion for dropping from prostate cancer AS in practice, as well as in the studies used to parameterize and test our model. In some cases, the decision is nuanced, requiring a shared decision-making approach between the patient and physician because of considerations of age, comorbidities, and the patient's personal preferences. However, our AS-POMDP model assumes such patients leave the system and receive care that is specialized to their personal situation with the guidance of a urologist.

Observation Probabilities. The observation probability is defined as the probability of observing certain output given the state of the system and the action taken. In the AS-POMDP model, the observation probabilities $\mathbb{P}(o|s,a)$ for all $a \in A$ and $o = (x,y) \in O = O_{\text{PSA}} \times O_{\text{Biopsy}}$ are given by

$$\mathbb{P}((x,y)|s,a) = \begin{cases} q^{\text{LR}}(I_i), & a = \text{Defer Biopsy},\ s = \text{LR},\ y = \text{Null},\ x \in I_i,\ \forall i; \\ q^{\text{HR}}(I_i), & a = \text{Defer Biopsy},\ s = \text{HR},\ y = \text{Null},\ x \in I_i,\ \forall i; \\ q^{\text{LR}}(I_i), & a = \text{Conduct Biopsy},\ s = \text{LR},\ y = \text{Not Upgrading},\ x \in I_i,\ \forall i; \\ \gamma q^{\text{HR}}(I_i), & a = \text{Conduct Biopsy},\ s = \text{HR},\ y = \text{Not Upgrading},\ x \in I_i,\ \forall i; \\ (1-\gamma) q^{\text{HR}}(I_i), & a = \text{Conduct Biopsy},\ s = \text{HR},\ y = \text{Upgrading},\ x \in I_i,\ \forall i; \\ 0, & \text{otherwise}, \end{cases}$$

where q^{LR} and q^{HR} are probability mass functions of PSA in the LR and HR cancer states, and γ is the false-negative rate (1 - sensitivity) of biopsy defined as the probability of observing Gleason score not upgrading while in HR cancer state.

Here we assume that biopsies have perfect specificity, i.e., the probability of observing a Gleason score upgrading when in LR cancer state is zero. This is because biopsies involve sampling of prostate tissue, and thus sometimes may miss the tumor; however

when the tumor is sampled, the probability that it is identified by a qualified pathologist in nearly 1.

Reward Function. We let $r(s,a,o)$ denote the reward function when the system is in state $s \in \mathscr{S}$, action $a \in A$ is taken, and output $o = (x,y) \in O$ is observed at each decision epoch, which is given by

$$r(s,a,(x,y)) = \begin{cases} 0, & a = \text{Defer Biopsy},\ s = \text{LR}; \\ \theta, & a = \text{Defer Biopsy},\ s = \text{HR}; \\ \eta, & a = \text{Conduct Biopsy},\ s = \text{LR},\ y = \text{Not Upgrading}; \\ \eta, & a = \text{Conduct Biopsy},\ s = \text{HR},\ y = \text{Upgrading}; \\ \theta + \eta, & a = \text{Conduct Biopsy},\ s = \text{HR},\ y = \text{Not Upgrading}; \\ \text{Not Defined}, & \text{otherwise}, \end{cases}$$

where θ and η are non-positive scalars that denote the negative reward (cost) of one-year delayed detection of high-risk cancer and the burden of a biopsy, respectively. In the AS-POMDP model, we seek to minimize a weighted sum of the expected number of biopsies and years in late detection to cancer progression, so these are negative "rewards." Note that θ and η are pre-determined scalars that reveal the decision-maker's consideration of the two events. We set $\theta + \eta = -1$, so that varying θ and η allows computing the optimal policy for different patient preferences for the two criteria.

Figure 3.1 illustrates the stochastic control process of the proposed AS-POMDP model. At the beginning of each decision epoch, the decision-maker can choose the test action of whether to defer and conduct biopsy. Then, the test outcome is observed, which provides partial information about the underlying cancer state. Given the chosen action and test outcome, an immediate negative reward is assigned to the patient, which comes from the burden of test action and/or the penalty of failing to detect a cancer progression to the

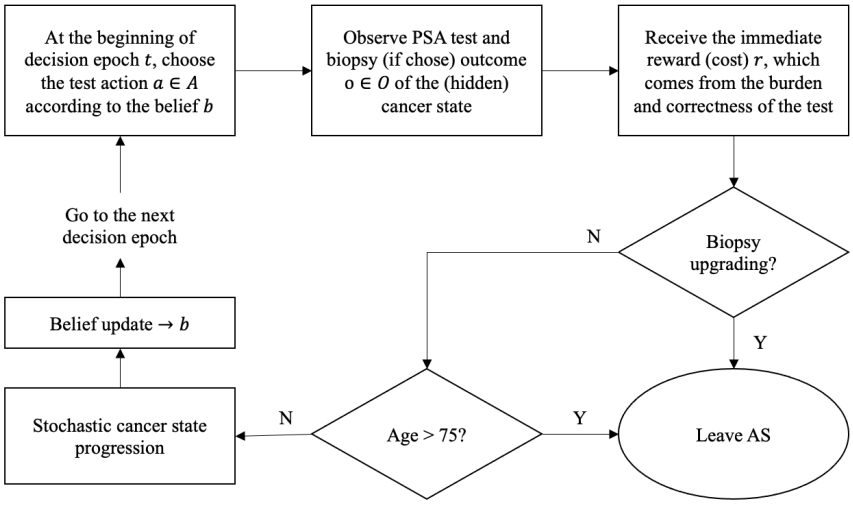

Figure 3.1: The stochastic control process of prostate cancer AS described by the proposed AS-POMDP model

HR state, if there was one. If the biopsy result shows Gleason score upgrading, then the patient will leave AS immediately. For the case that shows no upgrading or null biopsy, if the patient is older than age 75, he will also leave AS.

Belief state. We use b_t to denote the belief, i.e., the probability distribution, over the set of states, S, at the beginning of decision epoch t. In the AS-POMDP model, since there are only two states in S, the belief b_t only has one degree-of-freedom and can be represented by the probability of being in the HR cancer state, with $1 - b_t$ being the probability of being in the LR cancer state. In particular, for the starting time $t = 1$, b_1 is the probability that the patient who enters AS (because of being diagnosed with LR cancer) is actually in HR cancer state, i.e., misclassification error at diagnosis. The belief b_t at epoch t is well-known to be a sufficient statistic for the past sequence of actions and observations before time t. We use Λ to denote the Bayes updating formula from time t to $t+1$, i.e.,

$$b_{t+1} = \Lambda(b_t|a,o),\ 1 \le t \le T-1, \tag{3.1}$$

if action a is taken and output o is observed. The exact expression of Λ is given in the next section. Notice that sometimes we may drop the subscript of b_t when it is treated as the

argument of the value function defined later.

Policy. A policy $\pi = (\pi_1, ..., \pi_T)$ is defined as a set of functions from the belief space to the action space, where π_t specifies the actions to take for all possible belief states at decision epoch $t = 1, ..., T$.

Value Function. Given a policy π, we define the expected cumulative reward starting from time t_0 until the end of the time horizon T as:

$$V_{t_0}^{\pi}(b) := \mathbb{E}^{\pi}[\sum_{t=t_0}^{T} r(s_t, a_t, o_t)|b], \ \forall b, \forall t_0,$$

where the expectation is taken over all possible state, action, and observation trajectories following the policy π. For a fixed π and t_0, $V_{t_0}^{\pi}(b)$ is a function of the belief state b. Note that in the AS-POMDP model, if the patient leaves AS because of a Gleason score upgrading before time T, then the process will stop with no future reward.

As discussed in Smallwood and Sondik (1973), the POMDP model can be viewed as a continuous-state Markov decision process model with the state space being the space of all possible belief states. It follows immediately that there exists an optimal policy π^* that is deterministic and Markovian with respect to the belief, which maximizes the expected cumulative rewards at any time t:

$$V_t(b) := V_t^{\pi^*}(b) = \max_{\pi} V_t^{\pi}(b), \ \forall b,$$

which can be computed using the following optimality equations:

$$V_t(b) = \max_{a \in A} \{\sum_{s \in S} b(s) r(s,a) + \sum_{o \in O} \mathbb{P}(o|b,a) V_{t+1}(\Lambda(b|a,o))\}, \ \forall b, \ \forall t,$$

with the boundary condition

$$V_T(b) = \max_{a \in A} \sum_{s \in S} b(s) r(s,a), \ \forall b,$$

where $r(s,a) = \sum_{o \in O} \mathbb{P}(o|s,a) r(s,a,o)$ is the expected immediate reward when the system is in state s and action a is taken, and $\mathbb{P}(o|b,a) = \sum_{s \in S} b(s) \mathbb{P}(o|s,a)$ is the probability of observing o when the belief is b and action a is taken. By Assumption 3, in our AS-POMDP model, since the patient will leave AS upon receiving a biopsy that shows Gleason score upgrading, the optimality equation for $t < T$ should be modified as

$$V_t(b) = \max_{a \in A} \{ \sum_{s \in S} b(s) r(s,a) + \sum_{o \in O'} \mathbb{P}(o|b,a) V_{t+1}(\Lambda(b|a,o)) \}, \ \forall b, \ \forall t, \tag{3.2}$$

where $O' = O_{\text{PSA}} \times \{\text{Not Upgrading}, \text{Null}\}$ is a subset of O. Solving the optimality equations yields the optimal policy $\pi^* = (\pi_1^*, ..., \pi_T^*)$ as follows

$$\pi_t^*(b) := \arg\max_{a \in A} \{ \sum_{s \in S} b(s) r(s,a) + \sum_{o \in O'} \mathbb{P}(o|b_t,a) V_{t+1}(\Lambda(b|a,o)) \}, \ \forall b, \ \forall t < T,$$

and

$$\pi_T^*(b) := \arg\max_{a \in A} \sum_{s \in S} b(s) r(s,a), \ \forall b.$$

3.4 Solution Methods

In this section, we describe the approach we used to solve the AS-POMDP model formulated in Section 3.3. We start by describing an exact solution method, the classical *one-pass algorithm* of Smallwood and Sondik (1973), to set the foundation for describing our approach. Unfortunately, the one-pass algorithm is impractical for the AS-POMDP model, as the number of non-dominated α-vectors is growing exponentially in the size of the observation space at each time period. Because of the long time horizon and the fact that we intend to solve a number of different AS-POMDP model instances with different choices of model parameters, fast approximation methods are preferred over the exact method (which took more than 24 hours for a single set of model parameters using an Intel Core i7 2.6 GHz processor with 16 GB RAM). Therefore, we study two approximation

methods that give lower and upper bounds on the optimal value function, with bounded worst-case approximation errors. We further show in the numerical results that the gaps between the lower and upper bounds are very small so that our approximate solutions are accurate enough to be trusted.

3.4.1 Exact Solution Method

As shown in Smallwood and Sondik (1973), the optimal value function $V_t(b)$ is piecewise linear and convex in b, and can be written as

$$V_t(b) = \max_{\alpha \in \mathscr{A}_t} \alpha \cdot b, \ \forall b, \ \forall t,$$

where $\mathscr{A}_t$ is a set of linear functions, referred to as α-*vectors*, which be calculated by backward induction. Further, each α-vector in $\mathscr{A}_t$ corresponds to a decision tree that specifies the choices of action for all possible observations at each of the future decision epochs (see Kaelbling et al. (1998) for more details). It is easy to see that this property is also true in the AS-POMDP model, although equation (3.2) omits a part of *value-to-go* (which is linear in the belief) if the patient leaves the system before the end of AS due to observing a Gleason score upgrading. With this property, the AS-POMDP model can be solved by finding the set of α-vectors, $\mathscr{A}_t$, at each decision epoch t.

In our AS-POMDP model, since there are only two states, the belief can be represented by a scalar. We let b denote the belief in the high-risk cancer state, and thus the belief in the low-risk cancer state is $1-b$. Further, in the AS-POMDP model, each α-vector is a line, and can be determined by any two points on the line. For convenience, we use a vector, $(l(0), l(1))$, to represent the linear function $l(b)$, where $l(0)$ and $l(1)$ are the values of l at points $b = 0$ and $b = 1$, respectively. For models with more than two states, it is easy to generalize our results by using the extreme points of the belief simplex to represent α-vectors.

Starting with the boundary condition, the optimal value function at time T can be written as

$$V_T(b) = \max_{a \in A}\{(1-b)r(s_1,a) + br(s_2,a)\}, \forall b,$$

where $r(s_i,a) = \sum_y r(s_i,a,y)$ for $i = 1,2$. So,

$$\mathscr{A}_T = \{(r(s1,a_1), r(s_2,a_1)), (r(s_1,a_2), r(s_2,a_2)\}.$$

Now, given the set of α-vectors, $\mathscr{A}_{t+1}$ at time $t+1$, to derive the set of α-vectors, $\mathscr{A}_t$ at time t by way of backward induction, we can, for each α-vector in $\mathscr{A}_t$, find its values of at $b = 0$ and $b = 1$. In our AS-POMDP model, for each decision epoch, the belief update $b_{t+1} = \Lambda(b_t|a,o)$ is realized in two steps as follows,

$$b_t \xrightarrow{obs.} \tilde{b}_t \xrightarrow{trans.} b_{t+1}.$$

Specifically, suppose at time t, action a was taken and we observed o, then

$$\tilde{b}_t = \frac{\mathbb{P}(o|s_2,a)}{\mathbb{P}(o|b_t,a)} b_t,$$

and

$$b_{t+1} = \Lambda(b_t|a,o) = \tilde{b}_t + p(1-\tilde{b}_t) = \frac{1}{\mathbb{P}(o|b_t,a)}((1-p)\mathbb{P}(o|s_2,a)b_t + p\mathbb{P}(o|b_t,a)).$$

Then, by the optimality equations (3.2), for a specific action a, each α-vector α_t at time t can be be represented by

$$\alpha_t = (\alpha_t(0), \alpha_t(1))$$

where

$$\alpha_t(0) = r(s_1,a) + \sum_{o \in O'} \mathbb{P}(o|0,a)\alpha_{t+1,o}(\Lambda(0|a,o)) = r(s_1,a) + \sum_{o \in O'} \mathbb{P}(o|0,a)\alpha_{t+1,o}(p),$$

$$\alpha_t(1) = r(s_2,a) + \sum_{o \in O'} \mathbb{P}(o|1,a)\alpha_{t+1,o}(\Lambda(1|a,o)) = r(s_2,a) + \sum_{o \in O'} \mathbb{P}(o|1,a)\alpha_{t+1,o}(1)$$

for a specific set of choices of $\alpha_{t+1,o} \in \mathscr{A}_{t+1}$ for all $o \in O'$. Enumerating all such sets of choices of α-vectors at time $t+1$ and actions gives all α-vectors at time t, which we denote as $\tilde{A}_t$; however, some of the α-vectors in $\tilde{A}_t$ can be dominated by the others and thus can be *pruned* by solving a linear program (Smallwood and Sondik (1973)). Littman et al. (1995) and Zhang and Liu (1996) proposed the witness and incremental pruning algorithms that improve the pruning procedure and generate the minimal set of non-dominated α-vectors $\mathscr{A}_t$ at time t.

We let H denote the operator for the backward induction and pruning steps of V_t from V_{t+1} in the one-pass algorithm described above, and write the optimality equations as

$$V_t = HV_{t+1}, \quad t = T-1, ..., 1.$$

3.4.2 Point-based Approximation Method

The point-based approximation method is well-suited here, because instead of finding the set of all dominated α-vectors at each decision epoch, it only evaluates the value function at a set of sampled belief points to get an estimate of the value function. And by controlling the number of the sampled belief points, it limits the number of α-vectors to keep at each decision epoch. Different types of point-based value function approximation methods have been carefully studied in the surveys of Hauskrecht (2000), Pineau et al. (2003), and Shani et al. (2013) for infinite-horizon POMDPs, where the value function was assumed to be stationary (i.e., independent with time). We generalize their approach to our finite horizon non-stationary AS-POMDP model. In this section, we let B_t denote the sampled belief points at decision epoch t, and provide the methods for finding the lower and upper bounds of the value functions based on B_t for all $t = 1, ..., T$.

Lower Bound

At each decision epoch t, since the optimal value function can be written as the maximum of a set of linear functions in $\mathscr{A}_t$, a natural way to find a lower bound of V_t is to use a subset of $\mathscr{A}_t$. Starting from the optimal value function at the next decision epoch V_{t+1} and associated α-vectors, $\mathscr{A}_{t+1}$, we first derive the set of all (dominated and non-dominated) α-vectors $\tilde{\mathscr{A}_t}$ following the steps described in Section 4.1. Then, at each belief point, $b \in B_t$, we identify the supporting α-vectors in $\tilde{\mathscr{A}_t}$, resulting in $|B_t|$ α-vectors being selected from $\tilde{\mathscr{A}_t}$. We denote the set of selected α-vectors as $\hat{\mathscr{A}_t}$. Thus, at each decision epoch t, $\hat{V}_t$ defined as follows gives a lower bound of the true value function V_t:

$$\hat{V}_t(b) := \max_{\alpha \in \hat{\mathscr{A}_t}} \alpha \cdot b, \ \forall b.$$

The details of the lower bound approximation method is described in Algorithm 2.

Algorithm 2: Algorithm for approximate backward induction with operator L^B.

Input : V_{t+1}, B
Output: $\hat{V}_t$
Initialize $\hat{\mathscr{A}_t}$ as a empty set;
Let $\mathscr{A}_{t+1}$ as the set of α-vectors defining V_{t+1};
Find the set of all α-vectors at time t, $\tilde{\mathscr{A}_t}$ using $\mathscr{A}_{t+1}$ and backward induction;
for $b \in B$ **do**
    $\alpha_b \leftarrow \arg\max_{\alpha \in \tilde{\mathscr{A}_t}} \alpha \cdot b$;
    add α_b in $\hat{\mathscr{A}_t}$;
end
Define $\hat{V}_t(b) := \max_{\alpha \in \hat{\mathscr{A}_t}} \alpha \cdot b, \ \forall b$.

We let operator L^B denote approximate backward induction steps described in Algorithm 2. Note that L^B needs not to start from the exact optimal value function at the next decision epoch. If we start from any subset of $\mathscr{A}_{t+1}$, and the corresponding lower bound on V_{t+1}, then L^B will also provide a lower bound on V_t because the resulting $\hat{A}_t$ is always a subset of $\mathscr{A}_t$. In particular, if we start from the boundary condition V_T, with the sample

belief sets B_t for all $t = 1, ..., T-1$, then

$$\hat{V}_t = L^{B_t} L^{B_{t+1}} ... L^{B_{T-1}}(V_T), \forall t = 1, .., T-1 \tag{3.3}$$

is always a lower bound of V_t. The following theorem gives the error bound between $\hat{V}_t$ and V_t for each t, whose proof utilizes the triangle inequality and Holder's inequality and is adapted from Theorem 3.1 of Pineau et al. (2003). The proof of the Theorem 3.4 is in the Appendix.

Theorem 3.4. *Given the grids of the belief space at each decision epoch $B_t \subset [0,1]^{|S|}$ for all t, the error between the optimal value function V_t and approximated value function $\hat{V}_t$ given by (3.3) satisfies*

$$||V_t - \hat{V}_t||_\infty \leq \frac{(T-t)(T-t+1)}{2} ||r_{\max} - r_{\min}||_\infty \delta,$$

where

$$r_{\max}(s) := \max_{a \in A} \sum_{o \in O} \mathbb{P}(o|s,a) r(s,a,o), \; r_{\min}(s) := \min_{a \in A} \sum_{o \in O} \mathbb{P}(o|s,a) r(s,a,o), \; \forall s \in S,$$

and

$$\delta := \max_t \max_{b \in [0,1]^{|\mathscr{S}|}} \min_{b' \in B_t} ||b' - b||_1.$$

The bound in Theorem 3.4 tends to zero as $\delta \to 0$.

Upper Bound

Approaches to upper bound the optimal value function often involve solving many linear programs (Hauskrecht, 2000). Fortunately, for a two-state POMDP model such as the AS-POMDP model, the solution of the linear program can be given directly, which can further accelerate approximate backwards induction for our two-stage AS-POMDP model. At each decision epoch t, given the set of α-vectors A_{t+1} that defines V_{t+1} in the next decision epoch, to find the upper bound of V_t, we use the linear interpolation of

the sampled belief points and their values. Specifically, given the sampled belief set B_t, for each $b \in B_t$, we first calculate $u_t(b) := V_t(b)$ using the optimality equation. Then, as long as B contains the extreme points $b = 0$ and $b = 1$, for any belief point $b' \in [0,1]$, the solution of the following linear program will give the best linear interpolation for $V_t(b')$:

$$\begin{aligned} \bar{V}_t(b') := \min_\lambda \quad & \textstyle\sum_{b\in B} \lambda_b u_t(b) \\ \text{s.t.} \quad & \textstyle\sum_{b\in B} \lambda_b = 1, \\ & \lambda_b \geq 0, \quad \forall b \in B \\ & \textstyle\sum_{b\in B} \lambda_b b = b'. \end{aligned}$$

Further, $\bar{V}_t$ is an upper bound of V_t.

For a two-state POMDP model such as ours, the following results show that the optimal solution to the linear program is trivial, so that an upper bound of V_t can be obtained without resorting to solving linear programs.

Proposition 3.5. *In a two-state POMDP model, at time t, write the set of the sample belief point B_t as*

$$B_t = \{b^1, b^2, ..., b^{|B|}\}$$

such that $0 = b^1 < b^2 < ... < b^{|B|} = 1$. Then, for every $b' \in [0,1]$ such that $b^i \leq b' < b^{i+1}$, the optimal solution of the above linear program has only two variables λ_{b^i} and $\lambda_{b^{i+1}}$ being non-zero, and all others being zero.

Proof. Notice that the linear program has $|B|$ decision variables λ_b for $b \in B$ and $|B| + 2$ constraints. Then, the extreme point of the polyhedron defined by the constrains should satisfy $|B| - 2$ equations of $\lambda_b = 0$ for $b \in B$.

Now, for $b' \in [0,1]$ such that $b^i \leq b' < b^{i+1}$, suppose the extreme value $\bar{V}_t(b')$ is achieved with two λ_{b^j} and λ_{b^k} being non-zero, and all other decision variables being zero. Notice that to satisfy the first and last constraints, we can assume $b^j \leq b^i$ and $b^k \geq b^{i+1}$

without the loss of generality. Then, since V is convex, at b_t, the convex combination of b^j and b^k is greater than b^j and b^{i+1}, and the convex combination of b^j and b^{i+1} is greater than b^i and b^{i+1}. So, the optimal value $\bar{V}_t(b')$ is achieved with only λ_{b^i} and $\lambda_{b^{i+1}}$ being non-zero, and all other decision variables being zero. □

The above proposition shows that for belief points between b^i and b^{i+1}, $\bar{V}_t$ is defined by the line determined by two points $(b^i, u_t(b^i))$ and $(b^{i+1}, u_t(b^{i+1}))$, for all $i = 1, ..., |B| - 1$.The next proposition gives an expression of $\bar{V}_t$.

Proposition 3.6. *In a two-state POMDP model, at decision epoch t, denote β_i as the linear function determined by $(b^i, u_t(b^i))$ and $(b^{i+1}, u_t(b^{i+1}))$, for all $i = 1, ..., |B| - 1$, and let $\mathscr{B}$ be the set of all such linear functions:*

$$\mathscr{B} := \{\beta_1, ..., \beta_{|B|-1}\}.$$

Then,

$$\bar{V}_t(b_t) = \max_{\beta \in \mathscr{B}} \beta \cdot b_t, \ \forall b_t.$$

Proof. For each $i = 1, ..., |B| - 1$, since β^i is a line determined by

$$(b^i, u_t(b^i)) \text{ and } (b^{i+1}, u_t(b^{i+1})),$$

and since V_t is convex, then for $b \in (b^i, b^{i+1})$, $V_t(b) \leq \beta^i \cdot b$; for $b = b^i$ or $b = b^{i+1}$, $V_t(b) = \beta_i \cdot b$; and for $b \notin [b^i, b^{i+1}]$, $V_t(b) \geq \beta_i \cdot b$. By Proposition 1, for $b \in [b^i, b^{i+1}]$, $\bar{V}_t(b) = \beta_i \cdot b = max_{\beta \in \mathscr{B}} \beta \cdot b_t$. □

Algorithm 3 describes the steps for deriving the upper bound of the value function at each decision epoch by approximate backward induction. For convenience, we use operator U^B to denote Algorithm 3 for a given B. Note that the input of U^B can also be any upper bound of V_{t+1}, and the output $\bar{V}_t$ is always an upper bound of V_t because $u_t(b)$

is always greater than $V_t(b)$ for all $b \in B$. In particular, if we start from the boundary condition V_T, with the sample belief sets B_t for all $t = 1, ..., T-1$, then

$$\bar{V}_t = U^{B_t} U^{B_{t+1}} ... U^{B_{T-1}}(V_T), \forall t = 1, .., T-1. \tag{3.4}$$

is always an upper bound of V_t. The next theorem gives the error bound between $\bar{V}_t$ and V_t for each t. The proof of the Theorem is in the Appendix.

Algorithm 3: Algorithm for approximated backward induction U^B.

Input : V_{t+1}, B
Output: $\bar{V}_t$
Initialize $\mathscr{B}$ as a empty set;
Write $\mathscr{B} = \{b_1, ..., b_{|B|}\}$ such that $0 = b_1 < ... < b_{|B|=1}$;
for $b \in B$ **do**
  Calculate $u_t(b) := \max_a \{b \cdot r^a + \sum_o \mathbb{P}(o|b,a) V_{t+1}(U(b|a,o))\}$;
end
for $i = 1$ **to** $|B| - 1$ **do**
  Let β_i be the line determined by two points $(b_i, u_t(b_i))$ and $(b_{i+1}, u_t(b_{i+1}))$;
  Add β_i in $\mathscr{B}$;
end
Define $\bar{V}_t(b) = \max_{\beta \in \mathscr{B}} \beta \cdot b$ for all $b \in [0,1]$;

Theorem 3.7. *Given the grids of the belief space $B_t \subset [0,1]^{|S|}$ at each decision epoch t, the error between the optimal value function V_t and approximated value function $\bar{V}_t$ given by (3.4) satisfies*

$$||V_t - \bar{V}_t||_\infty \leq \frac{(T-t)(T-t+1)}{2} ||r_{\max} - r_{\min}||_\infty \delta, \ \forall t \leq T$$

where $r_{\max}$, $r_{\min}$, and δ are defined the same as in Theorem 3.4.

Remark 3.8. Later in Section 3.6, we show that the actual observed differences between the lower and upper bounds of the value functions in AS-POMDP all models were much smaller than the error bound given by Theorem 3.4 and 3.7. This is because in the AS-POMDP model, a patient will leave AS for treatment immediately after observing a Gleason score upgrading, with no future cost. As a result, the expected value-to-go

for conducting biopsy, as shown in the optimality equation (3.2), is shrunk by γ (biopsy false-negative rate). This further makes the error of the approximate value function much smaller than the worst case described in the proof of Theorem 3.4 and 3.7. However, since we do not know in advance what is the optimal action at each decision epoch, it is very difficult to improve the error bound. In the extreme case (e.g., always defer biopsy), it is possible that the error bound in Theorem 3.4 or 3.7 is achieved with equality. On the other hand, the results in Section 3.6 show that the proposed approximation methods work very well for the AS-POMDP model.

3.5 Structural Properties

In this section, we discuss some structural properties of the proposed AS-POMDP model to provide some insight into the results we present in Section 3.6.

3.5.1 Control-limit Type Policy

In Section 3.6 we will see the solution to the AS-POMDP model is a control-limit type policy, i.e., there is a threshold on the element of the belief vector that represents the probability of being in the high-risk state, below which it is optimal to defer biopsy, and above which it is optimal to conduct biopsy. There are many prior works that have discussed the existence of a control-limit type policy in a POMDP model. For example, White (1979) proved that the optimal replacement policy for the machine maintenance problem is a control-limit type policy. However, one of the distinctions of our model compared to the prior works is that our goal is to inspect and classify the system state (low-risk or high-risk cancer) rather than sequential system improvement, so that the optimal value function in our model is not monotone w.r.t. the belief anymore.

As in Section 3.4, we denote the set of non-dominated α-vectors at decision epoch t as

$\mathscr{A}_t = \{\alpha_1, ..., \alpha_n\}$, and write the optimal value function at time t as

$$V_t(b) = \max_{\alpha_i \in \mathscr{A}} \alpha_i(b), \ \forall b.$$

Then, it is easy to see that V_t has $n-1$ inflection points on $(0,1)$. The following lemma establishes a useful relationship among the positions of these $n-1$ inflection points, and the relationship between the slopes and endpoints of the non-dominated α-vectors.

Lemma 3.9. *For $\mathscr{A}_t = \{\alpha_1, ..., \alpha_n\}$, assume that slope$(\alpha_1) <$ slope$(\alpha_2) < ... <$ slope(α_n). Let the positions of the inflection points of V_t to be $b_1 < b_2 < ... < b_{n-1}$. Then, $(b_i, V_t(b_i))$ must be the intersection of α_i and α_{i+1}, $i = 1, ..., n-1$. Further,*

$$\alpha_1(0) > \alpha_2(0) > ... > \alpha_n(0),$$

and

$$\alpha_1(1) < \alpha_2(1) < ... < \alpha_n(1).$$

Proof. We prove the first part by contradiction. Suppose $(b_j, V_t(b_j))$ is the first inflection point of v_t such that it is not the intersection of α_j and α_{j+1}. Then, it should the intersection of α_j and α_k with $k > j+1$. So, $V_t(b) = \alpha_k(b)$ on $b \in (b_j, b_{j+1})$. Since α_{j+1} is not dominated, there must exists some $b_l \geq b_{j+1}$, such that $V_t(b) = \alpha_{j+1}(b)$ on $b \in (b_l, b_{l+1})$. Then, the slope of $V_t(b)$ is not increasing, which contradicts the convexity of V_t.

Now, choose $\beta = (\beta(1), \beta(2)) \in \mathscr{A}_t$ such that $\beta(2) = \min_{\alpha_i \in \mathscr{A}_t} \alpha_i(2)$. Then, it must be true that $\beta(1) = \max_{\alpha_i \in \mathscr{A}_t} \alpha_i(1)$; otherwise, β must be dominated by some α-vectors in $\mathscr{A}_t$. It is easy to see that the slope of β is the smallest in $\mathscr{A}_t$. So, $\beta = \alpha_1$. Remove α_1 from $\mathscr{A}_t$ and repeat the same steps until there is no element in $\mathscr{A}_t$ completes the proof. □

We now leverage the above lemma to provide a sufficient and necessary condition for the existence of a control-limit type policy in a two-dimension POMDP model.

Lemma 3.10. *For any time t, denote the set of non-dominated α-vectors at time t as $\mathscr{A}_t = \{\alpha_1, ..., \alpha_n\}$. Further, let $\mathscr{A}_t^1 = \{\alpha_1, ..., \alpha_m\}$ be the α-vectors corresponding to action "defer biopsy", and $\mathscr{A}_t^2 = \{\alpha_{m+1}, ..., \alpha_n\}$ be the α-vectors corresponding to action "conduct biopsy". We say $\mathscr{A}_t^1$ and $\mathscr{A}_t^2$ are separable at some $b \in [0,1]$, if at b all values of the α-vectors in $\mathscr{A}_t^1$ are greater or smaller than all values of the α-vectors in $\mathscr{A}_t^2$. Then, the optimal policy at time t is a control-limit type policy if and only if $\mathscr{A}_t^1$ and $\mathscr{A}_t^2$ are separable at $b = 0$, or equivalently, $\mathscr{A}_t^1$ and $\mathscr{A}_t^2$ are separable at $b = 1$.*

Proof. The existence of a control-limit type policy is equivalent to the existence of an inflection point $\bar{b}$ of $v_t(b)$, such that for $b \leq \bar{b}$, $V_t(b)$ is composed of the α-vectors in $\mathscr{A}_t^1$ and for $b > \bar{b}$, $V_t(b)$ is composed of the α-vectors in $\mathscr{A}_t^2$; Further, if there exists an inflection point $\bar{b}$ of $V_t(b)$, such that for $b < \bar{b}$, then the inflection points of $V_t(b)$ are the intersections between the α-vectors in $\mathscr{A}_t^1$; and for $b > \bar{b}$, the inflection points of $v_t(b)$ are the intersections between the α-vectors in $\mathscr{A}_t^2$. According to Lemma 3.9, the inflection points following the sequence of the slopes of the α-vectors, and the order of the slopes of the α-vectors is equivalent to the order of the values of the α-vectors at either endpoint. □

Focusing on our AS-POMDP model specifically, we let γ denote the false-negative rate of the biopsy, and note that the expected immediate reward for action "defer biopsy" at $b = 1$ is θ and the expected immediate reward for action "conduct biopsy" at $b = 1$ is $\eta + \gamma\theta$ $(= -1 - \theta + \gamma\theta)$. We only consider the case where $\eta + \gamma\theta$ is greater than θ, i.e., "conduct biopsy" is preferred to "defer biopsy" in HR cancer state. Using this notation, we now give a sufficient condition for which there exists a control-limit policy in this context.

Corollary 3.11. *Denote T as the end of time horizon. Suppose $\eta + \gamma\theta > \theta$, if*

$$(\gamma n - 1)\theta > (\eta + \gamma\theta)\frac{\gamma - \gamma^{n-1}}{1 - \gamma}$$

then there exists an optimal policy at time $T-n$ that is a control-limit type policy for $n = 1, ..., T-1$.

Proof. It is easy to calculate that at $t = T-n$, the smallest possible value at $b = 1$ of choosing "conduct biopsy" is $\eta + \gamma\theta + \gamma n\theta$, where the biopsy result shows not upgrading and "defer biopsy" will be chosen for all future times; the largest possible value at $b = 1$ of choosing "defer biopsy" is $\theta + \eta + \gamma\theta + \frac{1-\gamma^n}{1-\gamma}$, where "conduct biopsy" will be chosen for all future times with the observations all being not upgrading. If

$$\eta + \gamma\theta + \gamma n\theta > \theta + \eta + \gamma\theta + \frac{1-\gamma^n}{1-\gamma},$$

i.e., $(\gamma n - 1)\theta > (\eta + \gamma\theta)\frac{\gamma - \gamma^{n-1}}{1-\gamma}$, then at $b = 1$, the two sets of α-vectors corresponding to two actions are separable. By Lemma 3.10, we have the optimal policy for the two-state AS-POMDP model is a control-limit type policy. □

The existence of control-limit type policies in practical applications such as ours is a desirable feature since such policies conform well with the intuition of decision-makers. The sufficient condition in the above corollary holds for cases in which γ or θ approaches zero, for instance; however, we show that the existence of a control-limit type policy can be extended more broadly to a special (but not unrealistic) case of our model.

Proposition 3.12. *For the two-state AS-POMDP model, if decisions are made independent of the PSA test, the optimal policy is a control-limit type policy.*

Proposition 3.12 aligns well with clinical evidence that the PSA test is associated with high false-positive and false-negative errors and thus plays a limited role in making decisions about when to conduct routine biopsies. The proof of Proposition 3.12 is shown in the Appendix.

3.5.2 Static vs. Dynamic Policy

Our computational results in the next section show that although the optimal (dynamic) biopsy policies from the AS-POMDP model dominate the current (static) biopsy guidelines in the published literature, the difference is relatively small. Therefore, we conclude this section with some analysis to explain this by showing that eliminating the PSA test from the model makes it optimal to make biopsy decisions a priori without the need for dynamic decision making. In other words, the schedule of biopsies can be set at the time of diagnosis. Combining this with the fact that PSA is associated with high false-positive and false-negative rates and thus provides limited information for belief updating over time, suggests that the weakness of the PSA test limits the benefits of dynamic changes to the sequential decision to biopsy over time.

Theorem 3.13. *Consider a threshold-based biopsy policy for AS. If PSA test results are not used in cancer progression belief updates, then the threshold-based policy is equivalent to a static policy, in which the biopsy schedule is pre-determined at the time of diagnosis.*

Theorem 3.13 provides motivation for why the difference between dynamic and static policies is small, i.e., because the predictive value of the PSA test is weak. The proof of Theorem 3.13 is in the Appendix. Note that tests with better predictive performance than the PSA test, such as new molecular biomarker tests that are being developed (Barnett et al., 2018b), could lead to more significant benefits of dynamic over static policies. We revisit this in Section 3.6 with numerical experiments.

3.6 Results

In this section, we discuss the results of the AS-POMDP model for prostate cancer AS. We start by describing the model parameters. Next, we present the results for the

Center	misclassification error at diagnosis: b_1	Annual Cancer Progression rate: p	Biopsy Sensitivity: $(1-\gamma)$
JH	0.0583	0.0691	0.7184
UCSF	0.0809	0.1217	0.7431
U of T	0.0774	0.1016	0.7949
PRIAS	0.0653	0.0841	0.7614

Table 3.2: AS-POMDP model parameters in four study centers. Abbreviations: JH, Johns-Hopkins; UCSF, University of California-San Francisco; U of T, University of Toronto; PRIAS, Prostate Cancer Research International AS.

near-optimal value function and risk thresholds for the optimal biopsy policy given by the proposed AS-POMDP model using the algorithms in Section 3.4.2. These results also demonstrate the utility of the approximation methods we proposed. We also illustrate how the AS-POMDP model-based policy changes with respect to the reward parameters to understand how decisions might vary depending on patient preferences. Finally, we compare the near-optimal approximate policies with published guidelines.

3.6.1 Model Parameters

Tables 3.2 and 3.3 provide the As-POMDP model parameters for different centers that are computed using HMMs obtained in a previous study by Li et al. (2020). The PSA distributions were estimated by a mixture of two Gaussian distributions. In our AS-POMDP formulation, we discretized these continuous distributions using commonly used clinical thresholds, as shown in Table 3.3.

3.6.2 Optimal Biopsy policy Solved by AS-POMDP Model

The optimal policies of the AS-POMDP model vary across different centers, and reward parameters, which in turn depends on the decision-maker's preference. In our initial experiments, we set $\theta = \eta = -0.5$, which weighs the two criteria, i.e., expected delay in detection of high-risk cancer and expected number of biopsies, equally, and we evaluate the variation in policies across centers.

Center	Probability Mass Function of PSA (ng/mL): q			
	Cancer State	$I_1 = [0,4]$	$I_2 = (4,10]$	$I_3 = (10,\infty)$
JH	LR Cancer	0.3552	0.4311	0.2137
	HR Cancer	0.2868	0.4706	0.2426
UCSF	LR Cancer	0.0768	0.5680	0.3552
	HR Cancer	0.0678	0.5736	0.3586
U of T	LR Cancer	0.4573	0.3422	0.2005
	HR Cancer	0.3312	0.2368	0.4320
PRIAS	LR Cancer	0.1361	0.5357	0.3282
	HR Cancer	0.1094	0.5501	0.3405

Table 3.3: The probability mass functions of PSA in four study centers. Abbreviations: JH, Johns-Hopkins; UCSF, University of California-San Francisco; U of T, University of Toronto; PRIAS, Prostate Cancer Research International Active Surveillance; LR, low-risk; HR, high-risk.

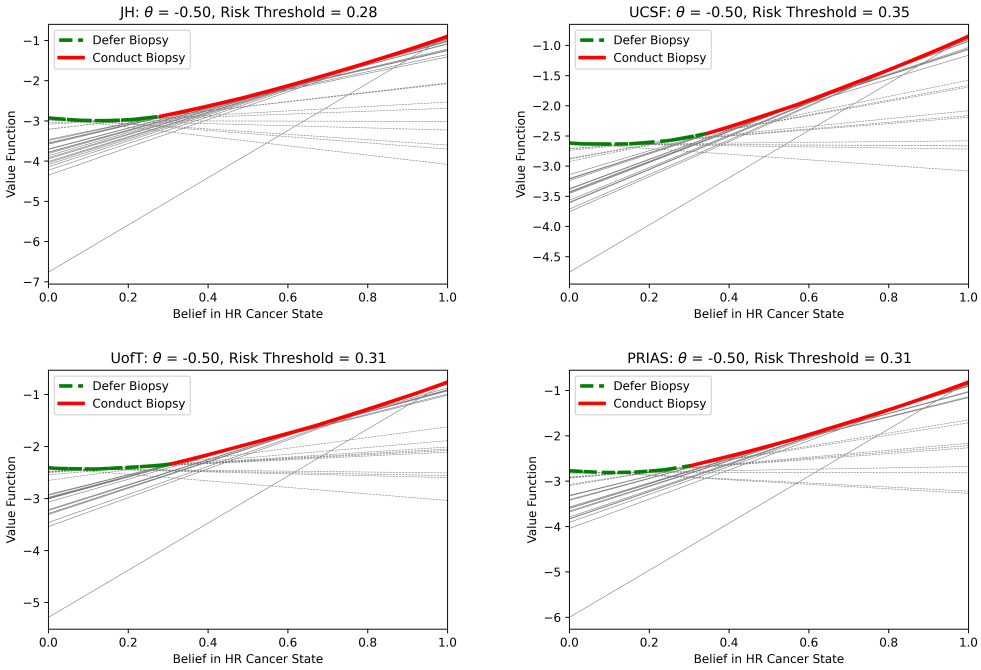

Figure 3.2: The (approximate) optimal value functions for a patient at age 50 in four different study centers when $\theta = -0.5$. All non-dominated hyperplanes, and their supremums are shown in the figure. The belief threshold for conducting a biopsy is indicated in the legend in each plot.

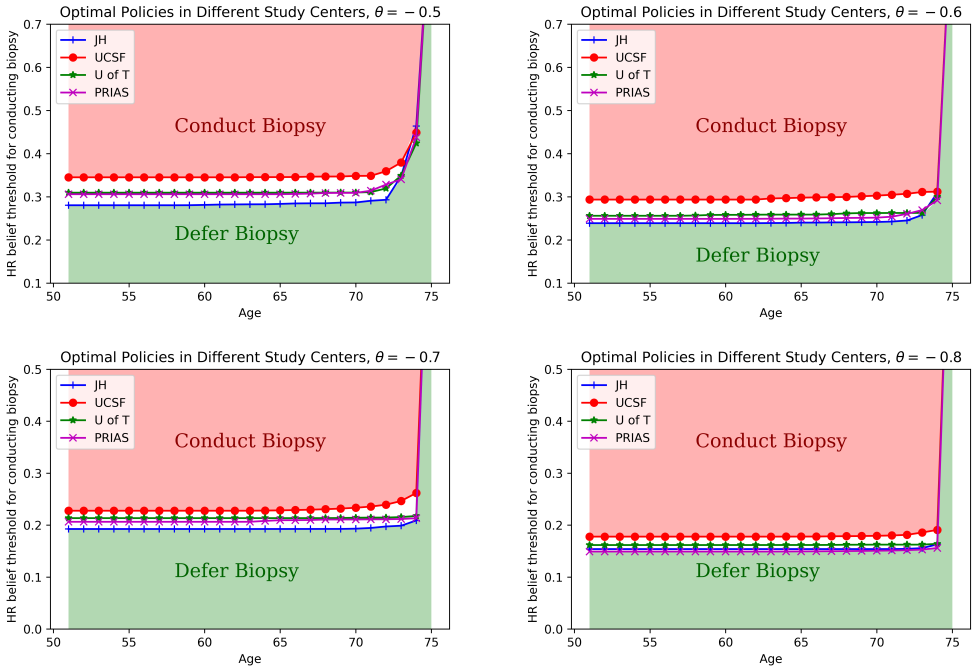

Figure 3.3: The (approximate) optimal belief thresholds for conducting biopsy in different AS studies when $\theta = -0.5, -0.6, -0.7, -0.8$

Figure 3.2 shows the approximate optimal value functions obtained by the method described in Section 3.4.2, for all four study centers assuming a patient at age 50. Here B_t is chosen to be $B_t = \{0, \frac{1}{30}, \frac{2}{30}, ..., \frac{29}{30}, 1\}$ with $|B_t| = 31$ for every $t \leq T$. As anticipated, the AS-POMDP model-based policies are all control-limit type policies. The risk threshold for triggering biopsy was the highest in the model generated from UCSF medical center data, and lowest in the model generated from JH hospital data, which is consistent with the difference in the annual cancer progression rates at those centers, and which in turn depends on the study admission criteria (JH study patients had more strict criteria for entry compared to UCSF patients). We further use Figure 3.3 to illustrate how AS-POMDP model-based policies differ across AS studies. For each risk-based policy, the range of risk threshold is relatively small.

As discussed in Section 3.3, our AS-POMDP model trades off the two competing criteria (delay in detection vs. harm from biopsies) based on the reward parameter θ. There-

fore, Figure 3.3 also shows how the optimal biopsy policies vary with respect to θ, which is the reward weights of the two criteria depending on individual patient's preference. Again, the closer the θ is to -1, the more the decision-maker weighs on the cost of delay in detection. As we change the value of θ parameter in the proposed AS-POMDP model, we observed that the optimal biopsy policy at each decision epoch is always a control limit type policy as discussed in relation to Proposition 3.12. Figure 3.3 also shows that the variation across models derived from the different AS studies decreases as theta decreases, i.e., as the weight on number of biopsies decreases. Moreover, the threshold for biopsy is consistently below 0.4 for all ages prior to 73.

3.6.3 Accuracy of Approximate Policies

To demonstrate that the approximated policies are very close to optimal, Table 3.4 provides the supremum norm of the difference between the lower and upper bounds of the optimal value function solved by the approximation methods in Section 3.4.2 using the uniform grid B_t with $|B_t| = 31$ for every $t \leq T$. As we can see from Table 3.4, the maximum relative error across all experiments is less than 0.55% of the value function, indicating the approximate policies are sufficiently accurate to be trusted. In terms of the running time, each experiment in Table 3.4 is completed within 30 seconds (compared with more than 24 hours for an exact solution) using an Intel Core i7 2.6 GHz processor with 16 GB RAM. Thus, the approximations enable the potential real-time implementation of the AS-POMDP model for shared patient/physician decision-making in clinical settings.

3.6.4 Implementation of Model-based Biopsy Policy in Practice

Before comparing different biopsy policies, we explain how the model-based policy can be used in practice to support decision-making in prostate cancer AS. In each study center, for each patient newly diagnosed with LR prostate cancer and admitted to AS,

Centers	$\|\|(\bar{V}-\hat{V})/\bar{V}\|\|_\infty \times 100\%$ at age 50 for different θ				
	-0.5	-0.6	-0.7	-0.8	-0.9
JH	0.27%	0.21%	0.15%	0.55%	0.28%
UCSF	0.15%	0.09%	0.08%	0.10%	$< 0.01\%$
U of T	0.19%	0.25%	0.16%	0.10%	$< 0.01\%$
PRIAS	0.18%	0.17%	0.25%	0.09%	0.01%

Table 3.4: The relative difference between $\bar{V}$ and $\hat{V}$ at age 50 for different θ in four AS studies.

his initial belief of being in HR cancer state is estimated by the misclassification error at diagnosis in Table 3.2. Subsequently, at each annual time period, the patient first receives a PSA test and the belief is updated using Equation 3.1. Next, the decision-maker decides whether to conduct or defer biopsy using an instance of the model based on the choice of the reward parameter θ that aligns with the patient's preferences, and the corresponding optimal HR belief threshold for triggering a biopsy base on the AS-POMDP model. If a biopsy is conducted, as shown in Figure 3.1, the patient will stay on AS if the result shows no biopsy upgrading and his age is less than 75 (the clinically recommended stopping time). The belief of HR cancer state is then updated again based on the annual cancer progression rate and biopsy sensitivity given by Table 3.2 using Equation 3.1; otherwise if the biopsy is deferred, then the HR cancer belief is updated only based on the annual cancer progression rate. Lastly, the patient will continue to the next time period, and follow the same steps as in the last time period until a biopsy upgrading is observed or age 75. We acknowledge that in practice, the decision of whether to conduct biopsy or not is often more nuanced, and requires a shared decision-making approach between the patient and physician. But our model-based biopsy policy can be used as a data-driven decision support tool to guide these decisions.

3.6.5 Comparison of Model-based Biopsy Policies vs. Current guidelines

Now, we compare the policies from solving the AS-POMDP model with published guidelines. The published guidelines include annual biopsy (JH guideline), biopsy every two years after diagnosis (UCSF guideline), biopsy every three years after diagnosis (PRIAS guideline, which is also implemented in the U of T study). We evaluate each policy for a simulated cohort of patients diagnosed with LR cancer who initiated AS at age 50. We first sample the initial cancer state at the starting time according to the misclassification error at diagnosis given in Table 3.2. Then, the patients will follow the process described in Figure 3.1, where at each decision epoch, the test action is given by the selected biopsy policy, the test results are sampled according to the observation probabilities, and the state transition is sampled according to the state transition probability. If a Gleason score upgrading is observed, the patient will leave AS immediately; otherwise, he continues to the next decision epoch, until age 75 when AS stops.

The number of hypothetical patients for the simulation is 10,000 for each study center and each biopsy policy. With the simulated true cancer states and biopsy results for all patients at all decision epochs, the expected number of biopsies performed while on AS is calculated as the average number of biopsies performed from initiating AS (age 50) to leaving AS (age 75 or a Gleason score upgrading), while the expected delay in time to detection of non-favorable risk cancer is calculated as the average difference between the time of the first sampled HR cancer state and the time of a Gleason score upgrading is observed for all patients.

Figure 3.4 illustrates the simulation results for different biopsy policies in four study centers. As we can see from Figure 3.4, in each center, for the optimal biopsy policies given by the AS-POMDP model, as the value of $|\theta|$ gets larger, the biopsy policy will

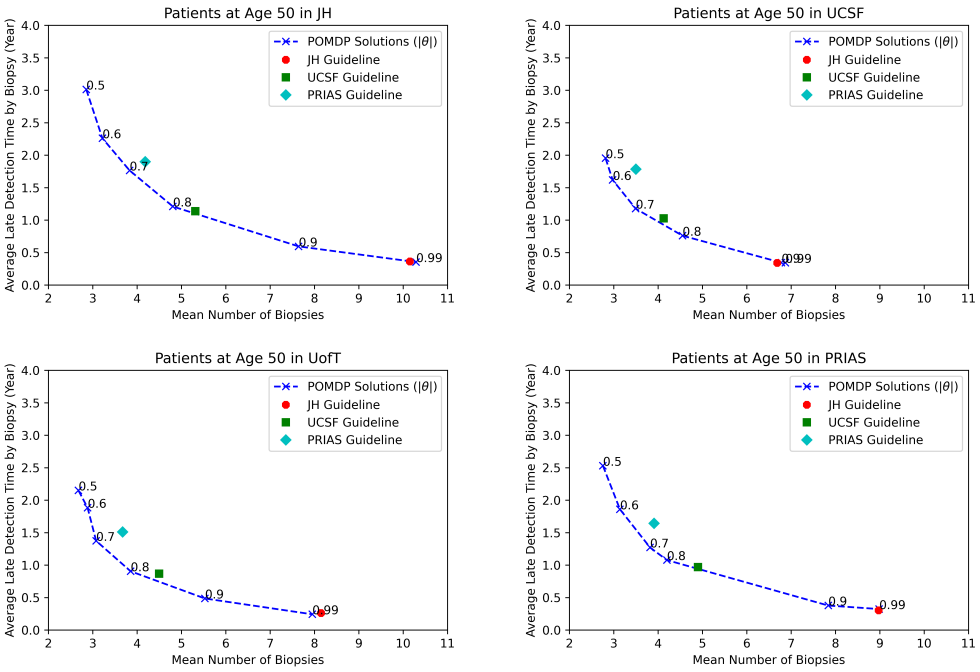

Figure 3.4: The comparison between policies given by the AS-POMDP model and current biopsy guidelines in different AS studies.

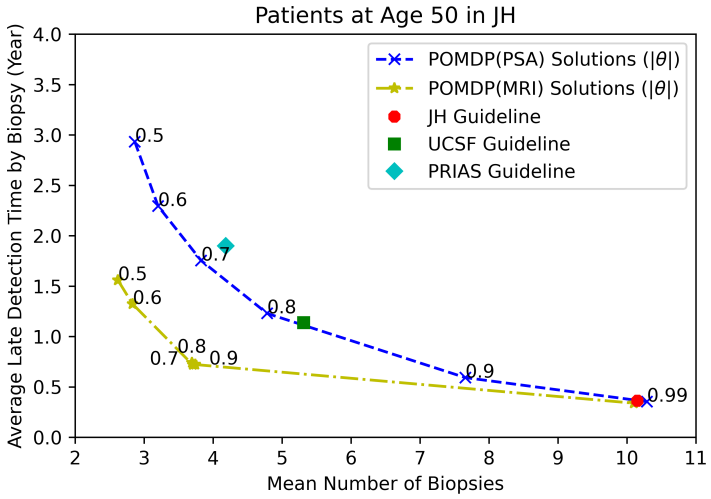

Figure 3.5: The comparison between policies given by two AS-POMDP (PSA and MRI) models and current biopsy guidelines in the JH center.

result in a greater number of expected biopsies and fewer years to the detection of cancer progression. Also, the optimal biopsy policies given by the AS-POMDP model are Pareto optimal compared with the static biopsy guidelines, i.e., they reduce the number of biopsies performed without increasing years in late detection to cancer progression.

3.6.6 Using MRI for AS

Since the PSA test has high false positive and negative rates, as previously noted, we do not observe a huge improvement in the policy given by the AS-POMDP model over current biopsy guidelines for each patient in Figure 3.4. Nevertheless, it is possible that more accurate bio-markers could lead to more significant improvement of the AS-POMDP model-based policies over current biopsy guidelines. One such approach to improving predictive performance that is receiving significant attention is MRI. Barnett et al. (2018b) showed the cost-effectiveness of MRI for early detection of prostate cancer. Motivated by their study, we conducted experiments using MRI as an alternative to the PSA test in the AS-POMDP model to show the potential benefit of the model-based policy. For MRI model parameters, we used the result from Grey et al. (2015), which estimated the sensitivity and specificity of MRI (using the prostate imaging reporting and data system score threshold of ≥ 4) to be 78.9% and 78.9%. Figure 3.5 shows the comparison among the policy given by the AS-POMDP model with either PSA test or MRI, and current biopsy guidelines for patients in the JH center. As we can see in Figure 3.5, as MRI is much more accurate than the PSA test, the benefit of the policies given by the AS-POMDP(MRI) model is more significant than it given by the AS-POMDP(PSA) model. Unfortunately, the study of Grey et al. (2015) was conducted on a different group of patients in the U.K., with a limited size of study population ($n = 201$), so that the result in Figure 3.5 is from a hypothetical experiment. We are looking forward to implement the MRI in the AS-POMDP

when more MRI data and studies become available.

3.6.7 Evaluating Implied Weights for Late Detection of Cancer Progression and Biopsy Burden

To understand how alternative policies trade-off between late detection to cancer progression and biopsy burden, we apply a simple *inverse optimization* (Ng et al., 2000) to estimate the reward function implied by each published biopsy guideline. Specifically, for a given biopsy guideline, denote $\pi = (\pi_1, ..., \pi_T)$ as the biopsy policy specified by the guideline. Since π is a static policy, then π_t is a constant action w.r.t. the belief state (either to defer biopsy or conduct biopsy) for all $t = 1, ..., T$. Now, denote $\bar{\pi}^t = (\bar{\pi}^t_1, ..., \bar{\pi}^t_T)$ as another static biopsy policy where

$$\pi_t \neq \bar{\pi}^t_t, \text{ and } \pi_k = \bar{\pi}^t_k, \ \forall k \neq t.$$

Further, define R_t as the set of reward functions such that $\bar{\pi}^t$ is dominated by π:

$$R_t := \{r : V^{\pi}_1(b_1) \geq V^{\bar{\pi}^t}_1(b_1)\}, \ \forall t = 1, ..., T,$$

where b_1 is the initial belief state. Notice that the reward function r is a function of θ, and the range of θ where the biopsy guideline π is the optimal static biopsy policy is given by

$$\theta \in R_1 \cap ... \cap R_T.$$

Table 3.5 shows the estimated range of θ implied by each guideline if applied to each center. As we can see from Table 3.5, all four study centers imply that avoiding delays in detecting high-risk prostate cancer is more important than avoiding biopsies; however, the relative weights vary significantly among the guidelines, which depend on the cancer progression rate and biopsy sensitivity in different study centers. Nevertheless, as some patients are highly averse to biopsies (Klotz, 2013), our study provides a solution to de-

Center	Range of θ implied by the biopsy guideline		
	JH guideline	UCSF guideline	PRIAS guideline
JH	$[-1, -0.93]$	$[-0.84, -0.83]$	$[-0.72, -0.71]$
UCSF	$[-1, -0.89]$	$[-0.75, -0.74]$	$[-0.68, -0.67]$
U of T	$[-1, -0.91]$	$[-0.83, -0.82]$	$[-0.78, -0.77]$
PRIAS	$[-1, -0.92]$	$[-0.83, -0.82]$	$[-0.71, -0.70]$

Table 3.5: Estimates of the range of θ implied by each published biopsy guideline in different AS study centers.

ciding the frequency of biopsy and a reference for the trade-off against the late detection to a cancer progression.

3.7 Conclusions

In this chapter, we proposed a finite-horizon two-state POMDP (AS-POMDP) model to optimize the biopsy policy in prostate cancer AS, where the objective is to minimize the number of biopsies and the delay in detection of high-risk cancer. Our study considered two kinds of parameter ambiguity: 1) heterogeneous transition and observation probabilities in different patient cohorts, and 2) variation in decision-maker's preferences as represented by reward functions. To evaluate alternative policies resulting from different parameters, it was necessary to solve many instances of the AS-POMDP model. To enable this, we introduced two fast approximation methods that are able to find the lower and upper bounds of the optimal value function of the AS-POMDP model. We compared the gap between the lower and upper bounds to show that our results were accurate enough for decision-making. Further, We discussed some structural properties of the AS-POMDP model that provide insight into the AS-POMDP model-based policies. We also discussed an explanation for why the dynamic biopsy policies given by the AS-POMDP model are similar to static policies recommended in the current biopsy guidelines, and we used inverse optimization to approximate how each guideline weighs biopsy burden versus late

detection of cancer progression.

In the computational result, we first presented the value functions and biopsy policies given by the AS-POMDP model in four different prostate cancer AS studies, if weighted equally on the burden of one biopsy and the penalty of one-year late detection to cancer progression. We observed that the optimal value function is not always monotone in the belief state. This is because the objective of the AS-POMDP model is to investigate rather than improve patients' cancer state, and patients may leave the system without any future cost if detected as high-risk cancer. Such models can be more straightforward for studies of medical testing, and more accurate, especially when other metrics such as QALYs are hard to estimate and too obscure for decision-making. Although the optimal value function is not monotone, we observe that the biopsy policies given by the optimal value function were monotone in the belief in high-risk cancer state, i.e., it would trigger a biopsy as long as the belief in the high-risk cancer state reached a threshold. The threshold of the optimal biopsy policy is dependent on the model parameters, which include cancer progression rate and biopsy sensitivity. In general, models with a higher cancer progression rate or lower biopsy sensitivity will give a lower belief threshold for conducting biopsy.

We then changed the reward weights in the reward function to see how does the model-based biopsy policy depends on the decision-maker's preference on biopsy burden and late detection time in each study center. We found that the more heavily the decision-maker weighs the late detection of cancer progression (the larger θ), the lower the belief threshold for triggering a biopsy in the optimal biopsy policy.

Finally, we compared the performance of the optimal biopsy policies given by the AS-POMDP model and current biopsy guidelines in four AS study centers by a simulation study. The model-based biopsy policies were all Pareto optimal. The policies based on published guidelines were close to the efficient frontier. We also ran a hypothetical

experiment using MRI in the AS-POMDP model, which showed the potential value of the AS-POMDP model with more accurate bio-markers than PSA. Lastly, we used an inverse optimization approach to estimate the reward weights implied by the current biopsy guidelines.

Besides the novelty of the application, our work also contributes to the POMDP literature. First, we introduced two fast approximation methods to quickly find the lower and upper bounds of the optimal value function of a finite-horizon POMDP model at each decision epoch. In particular, we showed that the best upper bound of the optimal value function at any belief point could be solved easily in a two-state model, without solving a large linear program as discussed in previous studies. We also provided the worst-case error bounds of the proposed approximation methods. Second, we showed that in extreme cases, the optimal biopsy policy given by the AS-POMDP model is a control-limit type policy, even if the optimal value function is not monotone in the belief state, which differs from all previous studies of the control-limit type policy. We discussed some intermediate results for the sufficient and necessary condition for the existence of a control-limit type policy in the POMDP model. We leave the statement for general cases as a conjecture in future research. Third, we showed that in the proposed AS-POMDP model, if the PSA test is not involved, then the optimal dynamic policy given by the model is equivalent to a static policy, in which the timing of conducting biopsy can be pre-determined. Further, we applied inverse optimization to approximate the value function implied by the current biopsy guidelines, which helped us understand how does each biopsy guideline weigh on late detection of cancer progression and biopsy burden.

There are also some limitations of our work, which could lead to opportunities for future research. First, we used a two-state POMDP model to approximate the stochastic system of prostate cancer AS, and only considered the information from PSA test and

biopsy. There might be other covariates in prostate cancer AS such as prostate volume, PSA doubling time, and the results of MRI scans that could be used to understand the underlying cancer state, but were not considered in this study. We look forward to improving our model by including these factors when more data becomes available. Second, the model parameters of the transition and observation probabilities are assumed to be stationary, i.e., independent of time, which may not be accurate in reality. However, incorporating time-dependent factors would require the estimates of the model parameters in pre-studies, and more computational effort to solve the model. Third, our results of the fast approximation method for finding the upper bound of the optimal value function, and the sufficient and necessary condition for the existence of a control-limit type policy only work in a two-state POMDP model. The generalization of these results to general POMDP models may not be trivial and is left for future studies. Although the focus of this chapter is on prostate cancer AS, our model formulation is flexible and could be applied to other medical decision-making problems in chronic disease management.

3.8 Appendix: Proofs

Proof of Theorem 3.4

First,

$$\begin{aligned}||V_t - \hat{V}_t||_\infty &= ||HV_{t+1} - L^B\hat{V}_{t+1}||_\infty \\ &= ||HV_{t+1} - H\hat{V}_{t+1} + H\hat{V}_{t+1} - L^B\hat{V}_{t+1}||_\infty \\ &\leq ||HV_{t+1} - H\hat{V}_{t+1}||_\infty + ||H\hat{V}_{t+1} - L^B\hat{V}_{t+1}||_\infty\end{aligned}$$

For the first term, $||HV_{t+1} - H\hat{V}_{t+1}||_\infty \leq ||V_{t+1} - \hat{V}_{t+1}||_\infty$. For the second term, let $b \in \mathscr{B}$ be the belief point where the point-based value approximation has the biggest error, and $\tilde{b} \in B$ be the closest sampled belief point to b. Also, let α be the vector that would be the

maximal at b, and $\tilde{\alpha}$ be the vector that is maximal at $\tilde{b}$, then $\tilde{\alpha}\cdot\tilde{b} \geq \alpha\cdot\tilde{b}$, and

$$\begin{aligned}
||HV_{t+1}^B - L^B V_{t+1}^B||_\infty &\leq \alpha\cdot b - \tilde{\alpha}\cdot b \\
&= \alpha\cdot b - \tilde{\alpha}\cdot b + (\alpha\cdot\tilde{b} - \alpha\cdot\tilde{b}) \\
&\leq \alpha\cdot b - \tilde{\alpha}\cdot b + (\tilde{\alpha}\cdot\tilde{b} - \alpha\cdot\tilde{b}) \\
&= (\alpha - \tilde{\alpha})\cdot(b - \tilde{b}) \\
&\leq ||\alpha - \tilde{\alpha}||_\infty ||b - \tilde{b}||_1
\end{aligned}$$

where the last step is by the Holder's inequality. Now, since each α-vector represents the cumulative reward from the current time until the end of time horizon followed by a policy specifying the choices of future actions for all possible observation sequences, then

$$||\alpha - \tilde{\alpha}||_\infty \leq (T-t)(r_{\max} - r_{\min}),$$

and

$$||HV_{t+1}^B - L^B V_{t+1}^B||_\infty \leq (T-t)(r_{\max} - r_{\min})\delta.$$

Repeat the steps above, we have

$$\begin{aligned}
||V_t - \hat{V}_t||_\infty &\leq ||V_{t+1} - \hat{V}_{t+1}||_\infty + (T-t)(r_{\max} - r_{\min})\delta \\
&\leq |V_{t+2} - \hat{V}_{t+21}||_\infty + [(T-t) + (T-(t+1))](r_{\max} - r_{\min})\delta \\
&\leq ... \\
&\leq \frac{(T-t)(T-t+1)}{2}||r_{\max} - r_{\min}||_\infty \delta.
\end{aligned}$$

Proof of Theorem 3.7

The proof is similar to the proof of Theorem 1. First,

$$\begin{aligned}
||V_t - \bar{V}_t||_\infty &= ||HV_{t+1} - U^B\bar{V}_{t+1}||_\infty \\
&= ||HV_{t+1} - H\bar{V}_{t+1} + H\bar{V}_{t+1} - U^B\bar{V}_{t+1}||_\infty \\
&\leq ||HV_{t+1} - H\bar{V}_{t+1}||_\infty + ||H\bar{V}_{t+1} - U^B\bar{V}_{t+1}||_\infty
\end{aligned}$$

For the first term, $||HV_{t+1} - H\bar{V}_{t+1}||_\infty \leq ||V_{t+1} - \bar{V}_{t+1}||_\infty$. For the second term, let $b \in \mathscr{B}$ be the belief point where the point-based value approximation has the biggest error, and $\tilde{b} \in B$ be the closest sampled belief point to b. Also, let α be the vector that would be the maximal at b, and $\tilde{\alpha}$ be the vector that is maximal at $\tilde{b}$, then $\tilde{\alpha} \cdot \tilde{b} \geq \alpha \cdot \tilde{b}$, and

$$||HV_{t+1}^B - U^B V_{t+1}^B||_\infty \leq \alpha \cdot b - \tilde{\alpha} \cdot b$$

The rest of the proof is exactly the same as the one for Theorem 1.

Proof of Proposition 3.12

First, it is easy to see that at each decision epoch, the α-vectors of the policy that always chooses "defer biopsy" is a non-dominated α-vector, which achieves a maximum value at $b = 0$ that can be denoted as $x_{t,0}$. Next, we prove by induction that at each decision epoch, all non-dominated α-vectors corresponding to action "defer biopsy" at current time must have their value at $b = 0$ being greater than $x_{t,0} + \eta$. If this statement is true, then the non-dominated α-vectors corresponding to action "defer biopsy" and the non-dominated α-vectors corresponding to action "conduct biopsy" are separable at $b = 0$. By Lemma 3.10, we have the optimal policy for the two-state AS-POMDP model is a control-limit type policy.

Now, at time T, the α-vectors corresponding to action "defer biopsy" is $(0, \theta)$, and the α-vectors corresponding to action "conduct biopsy" is $(\eta, \eta + \gamma\theta)$.

Assume that at time $t+1$, all non-dominated α-vectors corresponding to action "defer biopsy" have their value at $b = 0$ being greater that $x_{t+1,0} + \eta$, where $x_{t+1,0}$ is the value at $b = 0$ corresponding to the policy "no biopsy at all". At time t, denote the α-vectors corresponding to policy "no biopsy at all" as $(x_{t,0}, y_{t,0} + \theta)$. Suppose there exists a non-dominated α-vectors corresponding to action "defer biopsy", denoted as $(x_{t,1}, y_{t,1} + \theta)$, such that $x_{t,1} < x_{t,0} + \eta$. We are going to prove that $(x_{t,1}, y_{t,1} + \theta)$ is dominated by oth-

ers. Consider the α-vectors corresponding to policy "biopsy at time t and no biopsy afterwards", which is $(x_{t,0}+\eta, \gamma y_{t,0}+\eta+\gamma\theta)$. If $(x_{t,1}, y_{t,1}+\theta)$ is not dominated by the maximum of $(x_{t,0}, y_{t,0}+\theta)$ and $(x_{t,0}+\eta, \gamma y_{t,0}+\eta+\gamma\theta)$, then it must be true that the intersect of $(x_{t,0}, y_{t,0}+\theta)$ and $(x_{t,0}+\eta, \gamma y_{t,0}+\eta+\gamma\theta)$, denoted as b_1 is smaller than the intersection of $(x_{t,0}+\eta, \gamma y_{t,0}+\eta+\gamma\theta)$ and $(x_{t,1}, y_{t,1}+\theta)$, denoted as b_2. It is easy to calculate that

$$b_1 = \frac{\eta}{(1-\gamma)(y_{t,0}+\theta)},\ b_2 = \frac{\eta + x_{t,0} - x_{t,1}}{x_{t,0} - x_{t,1} + (y_{t,1}+\theta) - (\gamma\theta + \gamma y_{t,0})}.$$

By backward induction,

$$x_{t,0} = (1-p)x_{t+1,0} + p y_{t+1,0},\ x_{t,1} = (1-p)x_{t+1,1} + p y_{t+1,1}$$

if $x_{t,1} < x_{t,0}+\eta$, since $y_{t+1,0} < y_{t+1,1}$, then $x_{t+1,1} < x_{t+1,0}+\eta$. By assumption, the action at time $t+1$ corresponding to $(x_{t+1,1}, y_{t+1,1})$ is "conduct biopsy". So, for the α-vector $(x_{t,1}, y_{t,1}+\theta)$, its action at time t and time $t+1$ are "defer biopsy" and "conduct biopsy". Now, we consider an α-vector at time t, denoted as $(x_{t,2}, y_{t,2})$ whose action at time t and time $t+1$ are "conduct biopsy" and "defer biopsy", and actions after time $t+1$ are all same as the ones of $(x_{t,1}, y_{t,1}+\theta)$. We can calculate that

$$x_{t,2} = x_{t,1} + p(1-\gamma)(\theta + y_{t+1,2}),\ y_{t,2} = y_{t,1} + (\gamma-1)\theta.$$

Now, we are going to show that $(x_{t,1}, y_{t,1}+\theta)$ is dominated by the maximum of $(x_{t,2}, y_{t,2})$ and $(x_{t,0}+\eta, \gamma y_{t,0}+\eta+\gamma\theta)$. Denote the intersection between $(x_{t,2}, y_{t,2})$ and $(x_{t,0}+\eta, \gamma y_{t,0}+\eta+\gamma\theta)$ as b_3, then

$$b_3 = \frac{x_{t,0} + \eta - x_{t,2}}{x_{t,0} - x_{t,2} + y_{t,2} - \gamma y_{t,0} - \gamma\theta}.$$

Given $b_1 \le b_2$, it is easy to verify that $b_3 \le b_2$, which indicated that $(x_{t,1}, y_{t,1}+\theta)$ is dominated by the maximum of $(x_{t,2}, y_{t,2})$ and $(x_{t,0}+\eta, \gamma y_{t,0}+\eta+\gamma\theta)$. In other words,

if there exists an α-vector whose optimal action at time t and $t+1$ are "defer biopsy" and "conduct biopsy", then the α-vector should be dominated by another α-vector whose optimal action at time t and $t+1$ are "conduct biopsy" and "defer biopsy". This gives a conflict with the assumption that $(x_{t,1}, y_{t,1}+\theta)$ is non-dominated. As a result, we proved at time t, there is no non-dominated α-vector corresponding to action "defer biopsy" such that its value at $b=0$ is smaller than $x_{t,0}+\eta$.

To sum up, we have proved that at each decision epoch, the non-dominated α-vectors corresponding to action "defer biopsy" and the non-dominated α-vectors corresponding to action "conduct biopsy" are separable at $b=0$. By Lemma 3.10, we have the optimal policy for the two-state AS-POMDP model is a control-limit type policy.

Proof of Theorem 3.13

We use a straightforward induction argument to show that at each decision epoch, the belief the patient is in the high-risk cancer state can always be pre-calculated whether the biopsy is conducted or deferred at each decision epoch. In the beginning, the patient enters AS with a fixed initial belief of high-risk cancer state b_0. Now, suppose at time t, the patient stays in AS with a fixed belief of high-risk cancer state b_t, then the patient chooses to either choose to do biopsy according to the threshold-based biopsy policy or do nothing. If he chooses to do the biopsy, then he will stay in the AS until the next decision epoch only if the biopsy result is not Gleason score upgrading. So his belief in the high-risk cancer state at time $t+1$ can be calculated by the belief updating formula, which is a fixed value. Otherwise, if he does not perform the biopsy, then his belief of being in the high-risk cancer state at time $t+1$ can be calculated by the state progression formula, which is also fixed. Thus, at each decision epoch t, if the patient does biopsy according to the threshold-based biopsy policy, then his belief in high-risk cancer state is always fixed

so that the timing of biopsy is pre-determined.

CHAPTER 4

Multi-model Partially Observable Markov Decision Processes

4.1 Introduction

First introduced by Åström (1965); Drake (1962); Smallwood and Sondik (1973), POMDP models have been found successful in many problems including machine maintenance, robot navigation, healthcare, and others (see Cassandra (1998) for a survey). In Chapter 3, we used the POMDP model to optimize the medical decision-making in different prostate cancer AS studies, given the estimated model parameters in Chapter 2.

This chapter addresses the issue of *parameter ambiguity* in POMDP models defined as follows. In a POMDP model, the decision-maker can take actions to influence the transition dynamic, output, and reward from the system, such that the expectation of all future rewards are maximized. The transition, observation, and reward dynamics of a POMDP model are described by its model parameters. In practice, these model parameters are often estimated by pre-studies that fit machine learning models on historical observational data. As in Chapter 3, the input of the AS-POMDP model parameters were estimated by the HMM using the observational data in Chapter 2. A potential issue of this approach is that different studies can give different estimates of the model parameters. The difference in parameter estimates can arise from differences in the underlying study samples, study

designs, model formulations, or other factors. In the prostate cancer AS example, the difference in patients' cancer progression rate and biopsy accuracy may come from patients' heterogeneity and different clinical practices in different study centers. As a result, when applying the POMDP model for the optimization problem, there are multiple sets of model parameters that are all well-established. But the optimal value function and policy given by the POMDP model differ substantially for different model parameters. In this study, we call it the issue of *parameter ambiguity* in POMDP models.

In this chapter, we propose a new MPOMDP model to tackle the issue of parameter ambiguity. An MPOMDP model is a stochastic optimization and dynamic programming model that simultaneously considers multiple POMDP models, which have the same model structure but different model parameters. The goal is to find a single optimal policy that optimizes a "weighted" average of the value functions of all POMDP models. The model weight is given by the model belief vector, which can be interpreted as the importance and/or the probability of being the best model for each POMDP model, and is updated every time according to the information from system outputs. Traditionally, when it comes to the issue of parameter ambiguity, a decision-maker may randomly pick a single model, or take the average of multiple sets of model parameters. In this study, we will show that the proposed MPOMDP model outperforms the traditional methods by achieving a non-negligible Value of Stochastic solution (VSS), which is defined in Birge (1982). Our study also sheds light on the Expected Value of Perfect Information (EVPI) (Schlaifer and Raiffa, 1961), which may be relevant in situations where there are opportunities to collect additional information to resolve model uncertainty.

We describe several important properties of the proposed MPOMDP model, which not only show the benefits achieved by the MPOMDP model, but also motivate the solution method and fast approximation methods we propose to solve the MPOMDP model. First,

we show that an MPOMDP model can be reformulated as a special case POMDP model, with an enlarged state space. We discuss the existence and structure of the optimal policy of an MPOMDP model. Then, we show that the VSS and the EVPI are always non-negative under the MPOMDP model setting. After that, we describe an exact solution method and two fast approximation methods to the MPOMDP model. We also provide an example to illustrate the benefits of the proposed MPOMDP model, and how it can be applied to general problems.

As mentioned above, this work is motivated by the pre-study in prostate cancer AS in chapter 3. The objective is to find the optimal timing for biopsies in prostate cancer AS, such that the burden of biopsy and the delay in detecting cancer progression are minimized. We first estimated the cancer progression rates, biopsy under-sampling errors, and PSA distributions using an HMM in four major prostate cancer AS studies in the world, which include the JH hospital, the UCSF medical center, the U of T medical center, and the PRIAS project in Chapter 2. We also estimated the confidence intervals of all model parameters showing that the parameters were statistically significantly different in different studies. Based on that, we then used a finite-horizon POMDP model to find the optimal biopsy policy in each of the four major studies. The results in Chapter 3 show that the optimal policies solved by the POMDP model differ across AS studies and different settings of the reward functions. This result can be directly applied to the cases where the set of model parameters for the patients is known with certainty (e.g., finding the optimal biopsy policy for patients in the JH hospital). However, for a new patient without the knowledge of the best model describing his cancer dynamics, or for a newly initiated prostate cancer AS study seeking a biopsy protocol, it can be very risky to ignore the issue of parameter ambiguity and arbitrarily pick a single model for decision-making. We will show in the computational experiment in Section 6 that our proposed MPOMDP model can find a sin-

gle policy with the same complexity as the one given by a POMDP model, but achieves better overall performance in terms of minimizing the expected number of biopsies to conduct and the delay in detecting cancer progression over a patient's lifetime.

The rest of this chapter is organized as follows. In Section 2, we review the most related work in stochastic sequential decision-making under uncertainty and with parameter ambiguity, and summarize the main contribution of this work. In Section 3, we formally define the MPOMDP model and the optimal value optimization problem in an MPOMDP model. Then, in Section 4, we show some important structural properties of the MPOMDP and the weighted-value problem. In Section 5, we describe an exact solution method to the optimal value problem of an MPOMDP model. We also proposed two fast approximation methods for practical uses. We present the results of a toy example and a case study in prostate cancer AS in Section 6. Finally, we conclude this chapter and discuss potential future research in Section 7.

4.2 Literature Review

In this section, we first review the most closely related work in sequential decision-making under uncertainty and parameter ambiguity. Then, we describe the main contributions of this chapter with respect to the related literature.

As mentioned in Chapter 3, the partially observable Markov decision process was first introduced by Åström (1965); Drake (1962) and Smallwood and Sondik (1973). The POMDP model is a dynamic programming model for sequential decision-making, where the underlying system can be described by an HMM (Rabiner and Juang, 1986). On the one hand, the POMDP model subsumes the HMM in that it adds decision-making about what action to take at each time period, which will influence the transition, output, and reward dynamics of the system. The objective of a POMDP model is to find the policy

for actions to take at all time periods, such that the optimal cumulative reward is achieved. On the other hand, the POMDP model is a generalization of the Markov Decision Process (MDP) model (Puterman, 2014), where the underlying state is not observable and can only be inferred by the output of the system. POMDP models have found success in many problems including machine maintenance (Ross, 1971), robot navigation (Cassandra et al., 1996), healthcare (Ayer et al., 2012; Zhang et al., 2012a; Erenay et al., 2014), and many others (see Cassandra (1998) for a survey).

When applying the POMDP model to real-world problems, it often requires inputs of model parameters that include the initial distribution function, transition probabilities, observation probabilities, and reward function. However, the model parameters are usually borrowed from different studies that use statistical or machine learning methods to estimate the system dynamics. The estimation error and heterogeneity between different studies will further induce the parameter ambiguity in POMDP models. Li et al. (2021) used POMDP models to optimize AS strategies in prostate cancer, and showed that the optimal policies could vary considerably in different medical studies because of the difference in system dynamics revealed by model parameters. Saghafian (2018) proposed an ambiguous POMDP (APOMDP) model to address the issue of parameter ambiguity in the POMDP model. Boloori et al. (2020) then applied the APOMDP model in a study of post-transplant medication management, which improved the existing policies by considering variability among physicians' attitudes toward ambiguous outcomes and patients' progression dynamics. In contrast to the work in this chapter, in their proposed APOMDP model, the objective function is in a robust optimization setting, which weights the best-case and worst-case value functions across different sets of model parameters. Moreover, they assumed that the best and worst models were selected independently over time, which might violate the Markov property and induce inconsistency in model dynamics

across decision epochs. Nakao et al. (2021) described a distributionally robust Partially Observable Markov Decision Process (DR-POMDP), which estimates the distribution of the transition-observation probabilities using side information at the end of each period, to maximize the worst-case reward for any joint-distribution of the ambiguous model parameters. Different from their work, the study in the chapter seeks a single optimal policy that works well "on average", rather than optimizes the worst-case performance, when there are multiple credible POMDP models.

Despite the short history of the study of parameter ambiguity in POMDP models, there is a stream of research on parameter ambiguity in dynamic programming and MDP models over the last two decades. Nilim and El Ghaoui (2005) and Iyengar (2005) considered a robust formulation of an MDP to optimize the worst-case performance (referred to the "max-min" problem) of the model, while assuming a "rectangularity" property in ambiguity sets, i.e., the ambiguity in transition probabilities is independent with action, state, or time. They discussed the policy evaluation and other improved solution methods to the proposed robust MDP. Followed by their study, much of the research has focused on ways to construct ambiguity sets, to mitigate the rectangularity assumption on the ambiguity set, and to generalize the "max-min" objective function (Delage and Mannor, 2010; Xu and Mannor, 2012; Wiesemann et al., 2013; Delage and Iancu, 2015; Mannor et al., 2016). In contrast to these studies, our work in this paper addresses the issue of parameter ambiguity in a different manner. The MPOMDP model we proposed considers a weighted sum of value functions under different sets of model parameters, where the objective is to find a single policy that performs well overall possible models. Also, compared with the robust optimization formulation, our MPOMDP finds a less conservative policy that achieves the maximum of a weighted (by model belief) value function instead of the maximum worst-case value function. The most closed research to ours that we are aware of

is that of Steimle et al. (2021), which considered a multi-model Markov decision process (MMDP). They showed that any MMDP could be recast as a special case of a POMDP, as opposed to our MPOMDP formulation that is a generalization of POMDP with parameter ambiguity. In contrast to the previous work on parameter ambiguity in MDPs, the study in this chapter considers discrete ambiguity sets for the model parameters in POMDPs, and the objective is set to be optimizing the weighted value function.

To close this section, we describe the main contributions of this chapter to the literature. First, we address the issue of parameter ambiguity under the POMDP framework using the MPOMDP. Different from the work by Saghafian (2018), Nakao et al. (2021), and other literature in robust MDP, our model formulation considers the objective function to be a weighted sum of value functions given by the belief vector under different sets of model parameters. Such formulation allows inter-dependent mode transition, observation, and reward dynamics over time. Moreover, it provides less conservative policies than the robust optimization formulation, whose objective is to optimize the worst-case performance. Second, we study the structural properties of the proposed MPOMDP, which not only motive the solution methods, but also help analyze the effect of parameter ambiguity in POMDPs. Third, we describe the exact solution method and two different approximation methods to our model, which are shown to converge asymptotically and can provide near-optimal solutions in real-time. Finally, we present a case study for prostate cancer AS optimization, which illustrates how the MPOMDP can be applied in a real-world problem, and the benefit of the MPOMDP in stochastic sequential decision-making under parameter ambiguity.

4.3 MPOMDP Formulation

We start with the review of the formal definition of the POMDP, and then introduce the formulation of the MPOMDP, which generalizes the POMDP when there exists parameter ambiguity.

Definition 4.1. A POMDP model $\mathscr{M}$ is defined as a tuple (S, b_0, A, P, O, F, r), where S is the set of all states, b_0 is the initial distribution function over the set of states S, A is the set of all actions, $P : S \times A \times S \to [0,1]$ is the state transition probability distribution, O is the set of all observations, $F : S \times A \times O \to [0,1]$ is the observation probability distribution, and $r : S \times A \times O \to \mathbb{R}$ is the reward function.

Notice that in Definition 4.1, the state transition probability distribution and observation probability distribution, and the reward function are stationary, i.e., independent of time. A more general definition for the non-stationary model can be easily adapted using time-dependent model parameters. However, stationary models are usually more preferred than non-stationary models in practice because they are easier to understand and estimate model parameters.

POMDP models are widely used to solve stochastic sequential decision-making problems with partially observable states. We start by describing the finite-horizon problem. For a finite-horizon POMDP model, we can use $t = 0, 1, ..., T$ to denote its discrete time periods (also referred to as decision epochs), and b_t to denote the probability distribution over S (also referred to as belief vector) at time $t \leq T$. Then, given a policy $\pi = (\pi_0, ..., \pi_T)$, where each π_t is a mapping from the space of belief vector to A specifying the the action to choose for all possible belief states at time t, the value function of the policy π starting from belief state b at time t is defined as

$$V_t^{\pi}(b_t) := \mathbb{E}^{\pi}[\sum_{k=t}^{T} \gamma^{t-k} r(s_k, a_k, o_k) | b_t], \ \forall b_t, \forall t \leq T,$$

where $\gamma \in [0,1]$ is a discount factor that diminished the future rewards, s_k, a_k, and o_k are the state, action, and observation at time $k \leq T$, and the expectation is taken over all possible state, action, and observation trajectories following the policy π. Solving a POMDP model is equivalent to finding the optimal policy π^*, which achieves the maximum of the value function at any time t:

$$\pi^* := \arg\max_{\pi} V_t^{\pi^*}(b_t),\ \forall b_t, \forall t.$$

As shown in Smallwood and Sondik (1973), there always exists an optimal policy π^*, which is Markovian with respect to the belief vector. Starting from here, we may drop the subscript t of b_t in $V_t^{\pi}(b_t)$ when there is no confusion that V_t^{π} is the value function at time $t \leq T$. We may also substitute $V_t^{\pi^*}$ by V_t^* as a simplification for all $t \leq T$.

For the infinite-horizon POMDP, the policy π and value function V^{π} are defined to be stationary, i.e., independent with respect to time:

$$V^{\pi}(b) := \mathbb{E}^{\pi}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, o_t)|b],\ \forall b,$$

where the discount factor $\gamma \in [0,1)$ should be strictly less than 1. As discussed in Sondik (1978), for a infinite-horizon POMDP model, the sub-optimal value function and policy with an arbitrarily small error can be found via a *value iteration* algorithm.

In this chapter, we mainly focus on the finite-horizon problem for several reasons. First, finite-horizon models are more preferred than infinite-horizon models in healthcare applications and other applications where the survival time (length of decision epochs) can not be infinite. Second, although a finite-horizon POMDP model can be easily reformulated as an infinite-horizon POMDP model by appending the time index to the state definition, it does not automatically solve the problem as the computational complexity would increase along with the size of the state space. Further, focusing on the methodology for finite-horizon models can narrow down the recursion step for each state transition,

which further helps study the effect of parameter ambiguity.

As we can see from Definition 4.1, a POMDP model is defined upon a set of model parameters, which include the initial distribution, state transition probabilities, observation probabilities, and rewards. In practice, such parameters are often estimated from previous studies. As a result, it is common to see that different studies provide conflicting model parameters, which further give different optimal policies. For example, in Chapter 2, we used HMMs to estimate the cancer progression and biopsy under-sampling rates in four different prostate cancer AS studies, where the estimates in different studies were statistically significantly different because of the patient heterogeneity in different locations. Later on, in Chapter 3 we used a POMDP model to optimize the biopsy policy in prostate cancer AS, and showed that the optimal policies in different studies could disagree with each other under certain circumstances. The issue can arise when there is a new patient, and the physician can not decide which model to rely on for decision support.

The issue of parameter ambiguity motives the formulation of the MPOMDP in this work. Specifically, suppose there are M ($M < \infty$) different POMDP models, where all models share the same model structure of state space, action space, and observation space, but have different model parameters of initial distribution functions, transition probability and observation probability matrices, and reward functions. We assume that each model can possibly be the "right" model describing the underlying stochastic system to study. However, we are unable to pick a single model because of the lack of information on the best model. The way the MPOMDP model tackles this issue is to consider all different POMDP models simultaneously by assigning a weight to the objective function of each POMDP model according to the belief vector introduced later, and to optimize the weighted sum of the objective functions of all POMDP models. We will then argue that the MPOMDP model can find a single policy that works well "on average", which provides a

solution for conflicting model parameters. A formal definition of the MPOMDP model is given as follows.

Definition 4.2. An MPOMDP model $\mathscr{M}$ is defined as a tuple $(\mathscr{M}_1, ..., \mathscr{M}_M, \lambda)$, where M is the number of POMDPs, each $\mathscr{M}_m = (S, b_0^m, A, P^m, O, F^m, r^m)$ is a POMDP model as defined in definition 4.1 for $m = 1, ..., M$, and $\lambda = (\lambda_1, ..., \lambda_M)$ is a vector of the initial model weights for all M POMDP models such that

$$\lambda_m \in (0,1),\ \forall m = 1, ..., M, \text{ and } \sum_m^M \lambda_m = 1.$$

To understand the initial weight parameter vector λ in Definition 4.2, one can view each λ_m as the probability that the model $\mathscr{M}_m$ is the true model describing the underlying stochastic system to study at the starting time, for $m = 1, ..., M$. The initial λ vector is usually given by some prior knowledge about the relative importance and/or preference of each model, or set as a non-informative prior distribution. Then, every time when a system output is observed, the model and state probability distributions will be updated based on the information from the system output.

Before introducing the formal definitions of the optimal value problem, we first define the belief vector of an MPOMDP model.

Definition 4.3. (Belief Vector) For an MPOMDP model $\mathscr{M}$, the belief vector b_t of $\mathscr{M}$ at time t is defined as

$$b_t := (b_t^1, ..., b_t^M),$$

where each element is itself a vector

$$b_m^t = (b_m^t(s_1), ..., b_m^t(s_{|S|})),$$

and each $b_t^m(s_k)$ is the probability that the underlying model of the stochastic system is model $\mathscr{M}_m$ and the system is in state s_k at time t, for $m = 1, ..., M$, $t = 1, ..., T$, and all state

$s_k \in S$. Specially, at $t = 0$, the initial belief vector of $\mathscr{M}$ is defined as

$$b_0 := (b_0^1, ..., b_0^M) \circ \lambda,$$

where $b_0^1, ..., b_0^M$ are the initial belief vectors for models $\mathscr{M}_1, ..., \mathscr{M}_M$ respectively, λ is the initial belief weight, and $\circ$ is the Hadamard product.

To define the optimal value problem in an MPOMDP model $\mathscr{M}$, we first describe the process flow of the process. Initially, the underlying system is described by one of the given POMDP models, and is in one of the states in the state space. However, the decision-maker knows neither which of the given POMDP models is the best model nor the state of the system. Instead, the decision-maker obtains an initial weight parameter λ in advance based on prior knowledge and the estimate of the initial belief vector (i.e., the probability distribution over states) in each model. Then, at the beginning of each time period, with the estimate of the belief vector of the MPOMDP model, the decision-maker can take action to influence the dynamics of the underlying system. The system then generates an output according to the chosen action, the state of the system, and the observation probability function of the true underlying POMDP model. For the purpose of decision making, without the knowledge of the true model parameter, the decision-maker can approximate the observation probabilities by a adjusted observation probability function using the model belief, which will be discussed in detail in the next section. After observing the output, each POMDP model will calculate its own immediate reward according to the estimate of state distribution, the taken action, the output from the system, and its reward function. Lastly, the MPOMDP will update the belief vector, i.e. model and state distributions, at the beginning of the next time period according to the action taken and the transition probability functions. The objective of the optimal value problem is to optimize the expectation of the sum of the immediate rewards in all POMDP models until

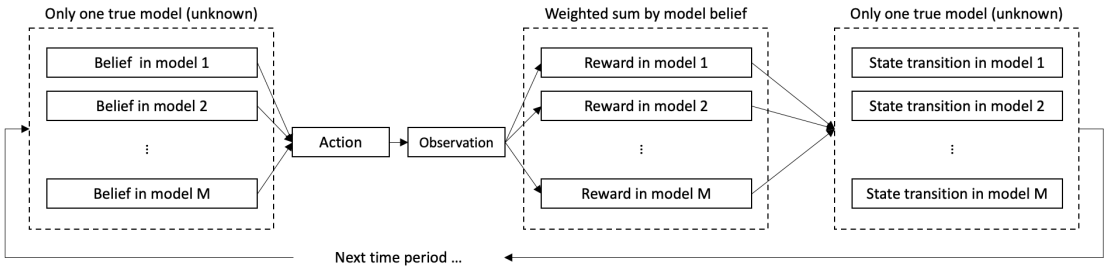

Figure 4.1: Illustration of the process flow of the optimal value problem in an MPOMDP.

the end of the time horizon. Figure 4.1 illustrates the process flow of the optimal value problem in an MPOMDP.

We now define the optimal value problem of an MPOMDP model $\mathscr{M}$ as follows.

Definition 4.4. For an MPOMDP model $\mathscr{M}$, the optimal value problem entails finding the optimal policy $\pi^* = (\pi_0^*, ..., \pi_T^*)$ that achieves the maximum value function defined as follows:

$$V_t^{\pi^*}(b_t) := \max_{\pi} \sum_{m=1}^{M} V_t^{m,\pi}(b_t^m),\ \forall b,\ \forall t,$$

where $b_t = (b_t^1, ..., b_t^M)$, b_t^m is the belief vector in $\mathscr{M}_m$, and $V_t^{m,\pi}(b_t^m)$ is the value function of policy π in $\mathscr{M}_m$ defined as

$$V_t^{m,\pi}(b_t^m) := \mathbb{E}^{m,\pi}[\sum_{k=t}^{T} \gamma^{t-k} r^m(s_k, a_k, o_k) | b_t^m],\ \forall b_t^m, \forall t \leq T,$$

with the expectation taken over all possible state, action, and observation trajectories following policy π in model $\mathscr{M}_m$ for $m = 1, ..., M$.

The definition of the optimal value problem is motivated by the case where there are a number of models that can possibly describe the underlying system and need to be considered simultaneously. Solving the optimal value problem yields a policy that achieves the maximum of the objective function defined as the weighted sum of the value functions of all possible models. From Definition 4.4, the optimal value problem of an MPOMDP

model $\mathscr{M}$ is defined upon the initial weight parameter vector λ, which is pre-specified in the definition of $\mathscr{M}$. Like the underlying state in an HMM, the underlying best model is not directly observable to the decision-maker. Instead, the decision-maker can only maintain an estimation of the probability of each POMDP model being the true model describing the stochastic system. First, at the starting time, there is an initial weight parameter vector λ for the probability distribution over all POMDP models, and an initial belief vector for the probability distribution over all states in all POMDP models. Next, at the beginning of each time period, according to the weight parameter vector and belief vectors, the decision-maker will choose an action. Then, the system will generate an output according to the distribution over states in all models and the observation probability functions of all models. There is an immediate reward specified by the reward functions in all models, and weighted by the probability distribution over models. At the end of each time period, the decision-maker can update the probability distribution over all models and all states in each model according to their posterior distributions given the prior action and the observed output from the system, and according to the state transition probability function in each model, until the end of the time horizon.

4.4 Model Properties

In this section, we discuss some structural properties of the proposed MPOMDP, which show how the model addresses the issue of parameter ambiguity in stochastic sequential decision-making, and motivate the solution methods introduced in the next section. We first provide the adjusted observation probability function and the belief update formula in the optimal value problem of an MPOMDP. Then, we show that the optimal value problem of an MPOMDP can be reformulated as a new POMDP model, so that all properties, especially the existence and structure of the optimal policy, and solution methods for

POMDP models will hold. We then discuss the effect of parameter ambiguity in POMDP model, and the VSS and EVPI under MPOMDP model settings.

We first provide the observation probability with respect to the belief vector in the optimal value problem. At each time, although the system output is generated according to the state, action, and observation probability function, the decision-maker has imperfect information about the true underlying model. Instead, given the belief vector, the decision-maker uses the following observation probability for the purpose of decision-making, as described in the following proposition.

Proposition 4.5. *Given an MPOMDP model $\mathscr{M}$, consider its optimal value problem defined in Definition 4.4. Then, at any time $t \geq T$, given the belief vector b, the probability of observing output o when action a is taken is*

$$\mathbb{P}(o|b,a) = \sum_{s,m} b^m(s) F^m(s,a,o), \ \forall o \in O \tag{4.1}$$

for all $o \in O$, belief vector b, and $a \in A$.

Proof.

$$\begin{aligned}
\mathbb{P}(o|b,a) &= \sum_{s,m} \mathbb{P}(o,(s,m)|b,a) \\
&= \sum_{s,m} \mathbb{P}((s,m)|b,a)\mathbb{P}(o|(s,m),b,a) \\
&= \sum_{s,m} b^m(s) F^m(s,a,o), \ \forall o \in O,
\end{aligned}$$

□

Given the adjusted observation probability of the optimal value problem, we then can show that the belief vector of the MPOMDP model is a sufficient statistic for decision-making at each time period. This property is important because it can help us keep track of the distribution over the state at each time period without requiring all historical information of actions and observations.

Proposition 4.6. *Given an MPOMDP model $\mathscr{M}$, consider its optimal value problem defined in Definition 4.4. Then, the belief vector b_t defined in Definition 4.3 is a sufficient statistic of the past sequence of actions and observations until time t for $t = 0, 1, ..., T$.*

Proof. Denote $I(t)$ as the total information available, i.e., historical actions and observations, at the end of time period t:

$$I(1) = \{a_1, o_1\}, I(t+1) = I(t) \cup \{a_{t+1}, o_{t+1}\}, \forall t \geq 1.$$

We are going to show that $b^m_{t+1}(s^m_{t+1})$ depends on $I(t)$ only through b_t for all $t \geq 1$, $s^m_{t+1} \in S$, and $m = 1, ..., M$:

$$\begin{aligned}
&b^m_{t+1}(s^m_{t+1}) \\
&= \mathbb{P}((s_{t+1}, m_{t+1})|a_t, o_t, I(t)) \\
&= \frac{\mathbb{P}((s_{t+1}, m_{t+1}), o_t|a_t, I(t))}{\mathbb{P}(o_t|a_t, I(t))} \\
&= \frac{\sum_{s_t \in S}\sum_{m_t} \mathbb{P}((s_{t+1}, m_{t+1}), (s_t, m_t), o_t|a_t, I(t))}{\mathbb{P}(o_t|a_t, I(t))} \\
&= \frac{\sum_{s_t \in S}\sum_{m_t} \mathbb{P}(o_t|(s_t, m_t), a_t, I(t))\mathbb{P}((s_{t+1}, m_{t+1})|(s_t, m_t), a_t, I(t))\mathbb{P}((s_t, m_t)|a_t, I(t))}{\mathbb{P}(o_t|a_t, I(t))} \\
&= \frac{\sum_{s_t \in S}\sum_{m_t} F^{m_t}(o, a, s_t)P^{m_t}(s_t, a_t, s_{t+1})b^m_t(s_t)}{\mathbb{P}(o_t|a_t, I(t))}.
\end{aligned}$$

Now, we can see the numerator of $b^m_{t+1}(s^m_{t+1})$ depends on $I(t)$ only through b_t, and the denominator is just the numerator summed over all possible values of s^m_{t+1}. Thus, b_t is a sufficient statistics of $I(t)$ for all $t = 1, ..., T$. □

We now can provide the belief update formula after taking action and observing an output at each time period in an MPOMDP model.

Proposition 4.7. *Consider the optimal value problem of an MPOMDP model $\mathscr{M}$. Suppose $b_t = (b^1_t, ..., b^M_t)$ is the belief vector of $\mathscr{M}$ at the beginning of time t, and observation o is*

observed after taking action a, then the belief vector $b_{t+1} = (b^1_{t+1}, ..., b^M_{t+1})$ *of* $\mathscr{M}$ *at the time* $t+1$ *is given by*

$$\mathbb{P}((s_{t+1}, m_{t+1})|o, b, a) = \frac{\sum_{s_t} F^{m_{t+1}}(s_t, a, o) P^{m_{t+1}}(s_{t+1}, a, s_t) b^{m_{t+1}(s_t)}}{\sum_{s_{t+1}, m_{t+1}} \sum_{s_t} F^{m_{t+1}}(s_t, a, o) P^{m_{t+1}}(s_{t+1}, a, s_t) b^{m_{t+1}(s_t)}}. \tag{4.2}$$

For simplicity, we use $b_{t+1} = \Lambda(b_t|a, o)$ *to denote the belief update formula given action a and observation o at time t for* $t = 0, 1, ..., T-1$.

Proof. First of all,

$$\mathbb{P}((s_{t+1}, m_{t+1})|o, b, a) = \frac{\mathbb{P}((s_{t+1}, m_{t+1}), o|b, a)}{\mathbb{P}(o|b, a)}.$$

For the numerator,

$$\begin{aligned}
&\mathbb{P}((s_{t+1}, m_{t+1}), o|b, a) \\
&= \sum_{s_t} \sum_{m_t} \mathbb{P}((s_{t+1}, m_{t+1}), o, (s_t, m_t)|b, a) \\
&= \sum_{s_t} \sum_{m_t} \mathbb{P}(o|(s_{t+1}, m_{t+1}), (s_t, m_t), b, a) \mathbb{P}(s_{t+1}, m_{t+1}), (s_t, m_t)|b, a) \\
&= \sum_{s_t} \sum_{m_t} \mathbb{P}(o|(s_t, m_t), a) \mathbb{P}(s_{t+1}, m_{t+1})|(s_t, m_t), a) \mathbb{P}((s_t, m_t)|b).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{P}((s_{t+1}, m_{t+1})|o, b, a) \\
&= \frac{\sum_{s_t} \mathbb{P}(o|(s_t, m_{t+1}), a) \mathbb{P}((s_{t+1}, m_{t+1})|(s_t, m_{t+1}), a) \mathbb{P}((s_t, m_{t+1})|b)}{\sum_{s_{t+1}, m_{t+1}} \sum_{s_t} \mathbb{P}(o|(s_t, m_{t+1}), a) \mathbb{P}((s_{t+1}, m_{t+1})|(s_t, m_{t+1}), a) \mathbb{P}((s_t, m_{t+1})|b)}.
\end{aligned}$$

□

In Proposition 4.7, the belief vector is updated using the Bayesian updating formula 4.2, which calculates a posterior distribution over models and states. In particular, even if none of the POMDPs considered in the MPOMDP is the true model for the study object, the belief update formula in Proposition 4.7 is still able to assign higher weights to the models with greater probability of generating the observed outputs.

Proposition 4.7 shows that the MPOMDP model is able to learn the model distribution, i.e., model credibility, over time from the past actions and observations. Propositions 4.6 and 4.7 also show that, an MPOMDP can be viewed as a continuous-state MDP when solving the optimal value problem, where the state is specified by the belief vector of the MPOMDP model. The state transition probabilities can be calculated by Proposition 4.7. Although the dimentionality of the state in such MDP can be very large, it helps us understand the structure of the MPOMDP. The following corollary shows that the optimal policy for the optimal value problem of an MPOMDP is deterministic and Markovian with respect to the belief vector, which is similar to it in MDPs.

Corollary 4.8. *For the optimal value problem of an MPOMDP model, there always exists an optimal policy that is deterministic and Markovian with respect to the belief vector at each time period.*

As in POMDP models, we also can show that the value function of the optimal value problem is piecewise linear and convex, which will be used as the basis for the solution methods that will be introduced in the next section.

Corollary 4.9. *Denote V_t as the optimal value function for the optimal value problem in Definition 4.4 at time $t = 0, 1, ..., T$. Then V_t is piecewise linear and convex in the belief vector b_t, and can be written as*

$$V_t(b_t) = \max_{\alpha \in \mathscr{A}} b_t \cdot \alpha, \ \forall b_t, \forall t.$$

Corollary 4.9 serves as the basis for the solution methods that will be introduced later. It also helps understand the effect of model ambiguity on the optimal value function and policy in POMDPs. A straightforward approach to solve the optimal value problem is to find the linear functions (refer to the "α-vectors") that determine the optimal value function at each time period. Further, calculating the difference in α-vectors among different sets

of model parameters can help infer that the optimal policy will change because of the parameter ambiguity. Specifically, suppose at time t, we have L numbers of non-dominated α-vectors for each POMDP model, denoted as

$$\begin{array}{ccc} \alpha_{t,1}^1, & ..., & \alpha_{t,L}^1 \\ \alpha_{t,1}^2, & ..., & \alpha_{t,L}^2 \\ ... & ... & ... \\ \alpha_{t,1}^M, & ..., & \alpha_{t,L}^M \end{array}$$

where $\alpha_{t+1,l}^m$ is the l^{th} non-dominated α-vector in the m^{th} model. So that the optimal value function for the MPOMDP model can be written as

$$v_t(b) = \max_{l \in [L]} \sum_{m=1}^{M} b^m \cdot \alpha_{t,l}^m$$

for all $b = (b^1, ..., b^M)$. Define

$$\delta_t(m_1, m_2) := \max_{l \in [L]} ||\alpha_{t,l}^{m_1} - \alpha_{t,l}^{m_2}||_\infty.$$

as the difference between the same α-vectors in two different models m_1 and m_2. The next proposition provides a bound on this difference. The proof is shown in the Appendix.

Proposition 4.10. *Define*

$$\xi^s(m_1, m_2) := \sum_z \sum_{s'} |P_{ss'}^{m_1} Q_{s'z}^{m_1} - P_{ss'}^{m_2} Q_{s'z}^{m_2}|,$$

and

$$\xi(m_1, m_2) := \max_{s \in \mathscr{S}} \xi^s(m_1, m_2)$$

then,

$$\delta_t(m_1, m_2) \leq \frac{\xi(m_1, m_2)}{2} \frac{(T-t)(T-t+1)}{2} \gamma^{T-t} |R_{\max} - R_{\min}|.$$

Proposition 4.10 bounds the difference between the same α-vectors in two different models. Using this result, we can then provide a sufficient condition such that for an

MPOMDP with two models, the non-dominated α-vectors in one model are also non-dominated in the other model, i.e., the parameter ambiguity does not affect the optimal value function and optimal policy.

Proposition 4.11. *For $M = 2$, suppose at time t, there is a non-dominated α-vector in model m_1 denoted as $\alpha_{t,l}^{m_1}$, such that there exist a belief vector b and $\varepsilon > 0$,*

$$\alpha_{t,l}^{m_1} \cdot b - \varepsilon \geq \max_{k \neq l} \alpha_{t,k}^{m_1} \cdot b.$$

Then $\alpha_{t,l}^{m_1} + \alpha_{t,l}^{m_2}$ is a a non-dominated α-vector in the MPOMDP model if $\varepsilon > \delta_t$, where δ_t is defined as in Proposition 4.10.

Proof. Since in model 1

$$\alpha_{t,l}^{1} \cdot b + \varepsilon \geq \max_{k} \alpha_{t,k}^{1} \cdot b,$$

and

$$\alpha_{t,l}^{2} \geq \alpha_{t,l}^{1} - \delta^t$$

also for all k,

$$\alpha_{t,k}^{2} \leq \alpha_{t,k}^{1} + \delta^t$$

then

$$\alpha_{t,l}^{1} + \alpha_{t,l}^{2} \geq 2\alpha_{t,l}^{1} - \delta^t$$

and for all $k \neq l$

$$\alpha_{t,k}^{1} + \alpha_{t,k}^{2} \leq 2\alpha_{t,k}^{1} + \delta^t \leq 2\alpha_{t,l}^{1} - 2\varepsilon + \delta^t$$

now if $\varepsilon > \delta_t$, then

$$\alpha_{t,k}^{1} + \alpha_{t,k}^{2} \leq 2\alpha_{t,l}^{1} - \delta^t \leq \alpha_{t,k}^{1} + \alpha_{t,k}^{2}$$

□

Proposition 4.11 provides a sufficient condition where the parameter ambiguity has no effect in a two-model MPOMDP. It requires calculating the α-vectors first to determine whether the optimal policies in different models are the same. We leave the generalization of Proposition 4.11 to the cases where the effect of parameter ambiguity can be determined before calculation, and to the cases with multiple POMDPs in further research.

The following proposition shows that the EVPI is always positive under the MPOMDP model setting.

Proposition 4.12. *Consider an MPOMDP model defined as* $\mathscr{M} = (\mathscr{M}_1, ..., \mathscr{M}_M, \lambda)$. *For each POMDP model* $\mathscr{M}_m$ *in* $\mathscr{M}$, *denote* π^m *and* V^m *as the optimal policy and optimal value function for* $\mathscr{M}_m$, $m = 1, ..., M$. *Then, the optimal value function* V *of the POMDP satisfies*

$$V_t(b) \leq \sum_{m=1}^{M} V_t^m(b^m), \forall b, \ \forall t.$$

Proof. At any time $t < T$,

$$\begin{aligned}
V_t(b) &= \max_{a \in A} \{r(b,a) + \sum_{o \in O} \mathbb{P}(o|b,a) V_{t+1}(\Lambda(b|a,o))\} \\
&= \max_{a \in A} \sum_{m=1}^{M} \{r^m(b^m,a) + \sum_{o \in O} \mathbb{P}(o|b^m,a) V_{t+1}^m(\Lambda(b^m|a,o))\} \\
&\leq \sum_{m=1}^{M} \max_{a \in A} \{r^m(b^m,a) + \sum_{o \in O} \mathbb{P}(o|b^m,a) V_{t+1}^m(\Lambda(b^m|a,o))\} \\
&= \sum_{m=1}^{M} \sum_{m=1}^{M} V_t^m(b^m)
\end{aligned}$$

for all belief vector b. And at time $t = T$,

$$V_T(b) = \max_{a \in A} r(b,a) = \max_{a \in A} \sum_{m=1}^{M} r^m(b^m,a) \leq \sum_{m=1}^{M} \max_{a \in A} r^m(b^m,a) = \sum_{m=1}^{M} \sum_{m=1}^{M} V_T^m(b^m)$$

for all belief vector b. □

4.5 Solution Methods

In this section, we discuss solution methods to the proposed MPOMDP model. We start with an exact solution method, which generalizes the one-pass algorithm by Smallwood and Sondik (1973) for POMDP models. However, because of the curse of dimensionality and the curse of history, the exact solution method can take a long time to run, even for small problems. Therefore, we introduce two approximation methods that can get near-optimal solutions efficiently. We also prove that the proposed approximation methods converge asymptotically. Finally, we compare the performance of the approximation methods in the next section.

4.5.1 Exact solution method

Recall the recursion formula, i.e., optimality equation, of the value function

$$V_t(b) = \max_{a \in A} \{ r(b,a) + \sum_{o \in O} \mathbb{P}(o|b,a) V_{t+1}(\Lambda(b|a,o)) \}, \ \forall b, \ \forall t,$$

with the boundary condition

$$V_T(b) = \max_{a \in A} r(b,a),$$

where

$$r(b,a) = \sum_{o \in O} \sum_{m} \sum_{s \in S} \mathbb{P}(o|b,a) r^m(s,a,o) b^m(s)$$

is the expected immediate reward, $\mathbb{P}(o|b,a)$ is the observation probability of output o, and $\Lambda(b|a,o))$ is the belief update formula provided by Proposition 4.7, given the current belief vector b and action a is taken for all possible belief b and action a.

As shown is Corollary 4.9, the optimal value function of an MPOMDP model $\mathscr{M}$ is piecewise-linear and convex in the belief vector b, and can be written as

$$V_t(b) = \max_{\alpha \in \mathscr{A}_t} \alpha \cdot b, \ \forall b,$$

where $\mathscr{A}_t$ is a set of $|S| \times M$-dimension vectors (also referred to as α-vectors) for all time periods t. Given this property, solving the optimal value problem of $\mathscr{M}$ is essentially to find the vector sets $\mathscr{A}_t$ for all time periods t. This can be done by the backward induction algorithm as follows. First, at the end of time horizon T, the set of α-vectors can be initiated as

$$\mathscr{A}_t = \{(\alpha^1_{T,a}, ..., \alpha^M_{T,a}) | a \in A\},$$

where

$$\alpha^m_{T,a}(s) = \sum_{o \in O} r^m(s,a,o) F^m(s,a,o), \ \forall m, \forall a, \forall s.$$

Then, given $\mathscr{A}_T$, each element in the set of α-vector $\mathscr{A}_{T-1}$ at time $T-1$ is given by $(\alpha^1_{T-1,l}, ..., \alpha^M_{T-1,l})$ where

$$\alpha^m_{T-1,l}(s) = \sum_{o \in O} r^m(s,a,o) F^m(s,a,o) + \sum_{o \in O} \gamma F^m(s,a,o) P(s,a,s') \alpha^m_{T,g_{T-1}(l,T)}(s'), \forall m, \forall s.$$

Notice that each subscript l of the α-vector $\alpha_{T-1,l}$ corresponds to a specific policy starting from time $T-1$. For convenience, we define a function $g_{T-1} : N \times T \to A$ with $g_{T-1}(l,t)$ being the action to take at time t that corresponds to the α-vector $\alpha_{T-1,l}$ at time $T-1$ for $l = 1, ..., L$ and $t \geq T-1$. The next step is to find the non-dominated α-vectors for the MPOMDP model at time $T-1$. Denote the new α-vectors (both dominated and non-dominated) as for M models as:

$$\begin{array}{ccc} \alpha^1_1, & ..., & \alpha^1_L \\ \alpha^2_1, & ..., & \alpha^2_L \\ ... & ... & ... \\ \alpha^M_1, & ..., & \alpha^M_L \end{array}$$

where α^m_l is the l^{th} α-vector in model m. Then, the non-dominated α-vectors for the MPOMDP model can be found by solving a linear program as follows: for a given sub-

script $l \in [L]$, if

$$
\begin{array}{lrcll}
\max & \delta & & & \\
\text{s.t.} & b^m \cdot \mathbf{1} & = & 1, & \forall m \in [M] \\
& b^m & \geq & 0, & \forall m \in [M] \\
& \sum_{m=1}^{M} \alpha_l^m \cdot b^m & \geq & \delta + \sum_{m=1}^{M} \alpha_h^m \cdot b^m, & \forall h \in [L] - \{l\} \\
& \delta & \geq & 0 &
\end{array}
$$

is feasible, then all α-vectors with subscript l in all models are non-dominated. Lastly, we can repeat the above two steps from time $t = T-1, ..., 0$ to find the optimal value function at the starting time.

For the exact solution method, since the number of new α-vectors (both dominated and non-dominated) L is growing exponentially in the number of actions and observations, the linear program can be too large to solve. Next, we are going to introduce the point-based approximation methods, which only find the non-dominated α-vectors at certain belief points, so that it can control the number of non-dominated α-vectors at each time period.

4.5.2 Sampling-based approximation methods

As we can see from the belief update formula, although the value function is a continuous function of the belief vector, there are only a finite number of reachable belief vectors at each time period if starting from a certain initial belief at the beginning. In other words, in order to find the optimal value function and the optimal policy of an MPOMDP model, it is sufficient to calculate the value function at all reachable belief vectors at each time period given a fixed initial belief vector, which can be done by backward induction starting from the end of time horizon. This makes the problem much easier than solving a large number of linear programs to find the exact value function. However, the number of all reachable belief vectors increases exponentially in the number of all possible actions

and observations along with the time. The ideal case is that we only need to know the value function at all reachable belief vectors under the optimal policy starting from the end of time horizon, and then use backward induction to calculate the value function at the previous time period until the initial time. Unfortunately, the optimal policy can not be determined without knowing the optimal value function.

As we have already shown, the optimal value function is piecewise-linear and convex, and can be represented by the supreme of a set of linear functions (α-vectors). Using this property, if one can identify the dominating α-vectors at some sampled reachable belief vectors, then their supremum also gives a lower bound approximation of the optimal value function over the entire belief vector space. A book chapter by Zhang and Denton (2018) has discussed some of the most recent approximation methods for the POMDP model. To summarize, the performance of such lower bound approximation, which can be defined as its distances to the exact optimal value function at all reachable belief vectors following the optimal policy, depends on whether it can correctly sample enough reachable belief vectors. On the one hand, we want to control the number of sampled belief vectors at each time period because of the consideration of computational complexity; on the other hand, we need to sample enough belief vectors to obtain a good estimation of the optimal value function, which helps identify the reachable belief vector following the optimal policy.

Next, we are going to introduce two sampling-based approximation methods for the proposed MPOMDP model. The first method uses a ε-greedy sampling method that balances exploitations and explorations of the reachable belief points based on the most recent estimate of the optimal value function, whose idea is similar to the ε-greedy algorithm for reinforcement learning problems as discussed in Sutton and Barto (2018). The second method is a tree-based branch-and-bound method, which improves the sampling efficiency of the first method by branching to the belief vector where the most recent estimate has

the largest error at each time period.

An ε-greedy sampling method

Denote $\mathscr{M}$ as the MPOMDP model to solve. To initialize, we sample a uniform grid of the entire space of the belief vector at each time period:

$$B_t^0 = \{0, \frac{1}{N}, \frac{2}{N}, ..., 1\}^M \subset [0,1]^M, \forall t = 1, ..., T,$$

where the superscript of B_t^0 denotes the number of iteration (here it is iteration 0), N controls the number of belief vectors and density of the uniform grid. With a finite set of grid points, an approximate backward induction works as follows. First, at the end of time horizon T, similar to Section 4.5.1, we calculate the set of α-vectors as

$$\mathscr{A}_T = \{(\alpha_{T,a}^1, ..., \alpha_{T,a}^M) | a \in A\},$$

where

$$\alpha_{T,a}^m(s) = \sum_{o \in O} r^m(s,a,o) F^m(s,a,o), \ \forall m, \forall a, \forall s.$$

Now, instead of keeping all α-vectors in $\mathscr{A}_T$, we only keep the ones that are non-dominated at the belief vectors in B_T^0, which gives $\hat{\mathscr{A}}_T$

$$\hat{\mathscr{A}}_T := \{\alpha \in \mathscr{A}_T | \alpha = \arg\max_{\alpha} \alpha \cdot b \text{ for some } b \in B_T^0\}.$$

Since $\hat{\mathscr{A}}_T \subset \mathscr{A}_t$, then $\hat{V}_T$ defined as

$$\hat{V}_T(b) := \max_{\alpha \in \hat{\mathscr{A}}_T} \alpha \cdot b, \ \forall b$$

gives a lower bound estimate of the optimal value function V_T at time T. Next, we go backward to time $T-1$. Similar as in Section 4.5.1, we can calculate the α-vectors at time $T-1$ using the optimality equation (1). But here, instead of using $\mathscr{A}_T$, we only use its subset $\hat{\mathscr{A}}_T$ to derive the set of α-vectors at time $T-1$, denoted as $\tilde{\mathscr{A}}_{T-1}$. It is easy to see

that $\tilde{\mathscr{A}}_{T-1}$ is a subset of $\mathscr{A}_{T-1}$, which is the set of all α-vectors at time $T-1$ if using $\mathscr{A}_T$ other than $\hat{\mathscr{A}}_T$ in backward induction. Again, instead of keeping all elements in $\tilde{\mathscr{A}}_{T-1}$, we only keep the ones that are dominating at the belief vectors in B^0_{T-1}, which gives $\hat{\mathscr{A}}_{T-1}$,

$$\hat{\mathscr{A}}_{T-1} := \{\alpha \in \tilde{\mathscr{A}}_{T-1} | \alpha = \arg\max_{\alpha} \alpha \cdot b \text{ for some } b \in B^0_{T-1}\}.$$

Since $\hat{\mathscr{A}}_{T-1} \subset \tilde{\mathscr{A}}_{T-1} \subset \mathscr{A}_{T-1}$, then $\hat{V}_{T-1}$ defined as

$$\hat{V}_{T-1}(b) := \max_{\alpha \in \hat{\mathscr{A}}_{T-1}} \alpha \cdot b, \ \forall b$$

gives a lower bound estimate of the optimal value function V_{T-1} at time $T-1$. We can keep going backward following the steps above until time $t=0$, which will give us the lower bound estimates of the value functions at all time period $\hat{V}_0, \hat{V}_1, ..., \hat{V}_T$.

The next step is to modify the grid of belief points $B^0_1, ..., B^0_T$ to improve the estimates of value functions. Starting from time $t=0$, denote b_0 as the initial belief vector. We can use the current estimate of the value function at time $t=1$ to find the optimal action to take under the current value function approximate at time $t=0$, which is given by

$$\hat{a} = \arg\max_{a} \sum_{m} \sum_{s \in S} b^m_0(s)\{r^m(s,a) + \sum_{o \in O} \sum_{s' \in S} \gamma F(s,a,o) P^m(s,a,s') \hat{V}^m_1(\Lambda(b^m_0|a,o))\}.$$

Notice that $\hat{a}$ may be sub-optimal, because it is selected using an approximation of the expected future value-to-go. Next, action $\hat{a}$ is selected with probability $1-\varepsilon$ and an alternative action with probability ε, for some $\varepsilon \in (0,1)$, to encourage the exploration of other actions that can potentially be better than $\hat{a}$. After taking the selected action, denoted as a_0, we then randomly sample an output of the system o_0 according to the observation probability matrix F. Given the action a_0 and observation o_0, the belief vector at time $t=1$ can be updated by

$$b_1 = \Lambda(b_0|a_0, o_0).$$

We then add b_1 into B_1^0 to get $B_1^1 = B_1^0 \cup \{b_1\}$. Now starting from belief b_1 at time $t = 1$, we repeat the steps above to sample the belief vectors $b_2, ..., b_T$ until the end of time horizon T, and get the new sets B_t^1 for $t = 2, .., T$. Collectively, the complete set of backward and forward step is one iteration of the ε-greedy sampling method.

In the next iteration, we conduct the backward induction steps on the new belief vector set $B_T^1, ..., B_1^1$, and then sample the new belief vectors to get the new sets B_t^2 for $t = 1, .., T$. We repeat these iterations until a stopping criterion is satisfied. For example, a stopping criterion could be that the size of the belief vector set is greater than some maximum number or the difference between the approximate value functions in two consecutive iterations is below some threshold. This completes the steps of our proposed approximation algorithm based on ε-greedy sampling. We summarize the complete algorithm in Algorithm 4.

As we can see from Algorithm 4, if we denote $\bar{V}_t^i$ for all t as the lower bound estimates of the optimal value functions after the i^{th} iteration, then $\bar{V}_t^i$ is determined by the set of sampled belief vectors B_t^i, which is generated by random sampling, for each time t. Next, we show that the lower bound estimates $\bar{V}_t^i$ converges to V_t in probability at all reachable belief vectors for all time periods t, as the number of iterations i goes to infinity.

Theorem 4.13. *For a given MPOMDP model $\mathscr{M}$, denote $\tilde{B}_t$ as the set of all reachable belief vectors at time $t \leq T$ starting from the initial belief vector b_0 following any policies. Denote V_t as the optimal value function at time $t \leq T$, and $\hat{V}_t^i$ as the lower bound estimate of the optimal value function at time $t \leq T$ given by the i^{th} iteration of Algorithm 4. Then for all $t \leq T$, for any $b \in \tilde{B}_t$,*

$$\hat{V}_t^i(b) \to V_t(b) \text{ in probability, as } i \to \infty.$$

Proof. We start from the end of time horizon $t = T$. For any reachable belief vector $b \in \tilde{B}_T$,

Algorithm 4: Approximation algorithm based on ε-greedy sampling.

Input : MPOMDP model $\mathscr{M}$, ε
Output: $\hat{V}_t$
Initialize B^0 as a uniform grid and $i = 0$;
repeat
  At time T, calculate $\mathscr{A}_T$;
  $\hat{\mathscr{A}_T} = \{\alpha \in \mathscr{A}_T | \alpha = \arg\max_\alpha \alpha \cdot b \text{ for some } b \in B_T^i\}$;
  $\hat{V}_T = (b) = \max_{\alpha \in \hat{\mathscr{A}_T}} \alpha \cdot b, \ \forall b$;
  for $t = T-1, ..., 0$ **do**
    Calculate the set of α-vectors $\tilde{\mathscr{A}_t}$ at time t by backward induction using $\hat{\mathscr{A}_{t+1}}$;
    $\hat{\mathscr{A}_t} = \{\alpha \in \tilde{\mathscr{A}_t} | \alpha = \arg\max_\alpha \alpha \cdot b \text{ for some } b \in B_t^i\}$;
    $\hat{V}_t(b) = \max_{\alpha \in \hat{\mathscr{A}_t}} \alpha \cdot b, \ \forall b$;
  end
  for $t = 0, ..., T-1$ **do**
    $\hat{a} = \arg\max_a (r(a) + \sum_{o \in O} \mathbb{P}(o|b_t, a) V_{t+1}(\Lambda(\hat{b_t}|a, o)))$;
    $a_t = \begin{cases} \hat{a}_t, & \text{with probability } 1-\varepsilon \\ \text{a random action}, & \text{with probability } \varepsilon \end{cases}$;
    Sample an output o_t according to b_t and F;
    $b_{t+1} = \Lambda(b_t | a_t, o_t)$;
    $B_t^{i+1} = B_{t+1}^i \cup \{b_{t+1}\}$;
  end
  $i = i + 1$;
until *some stopping criterion*;

we can show that in each iteration of Algorithm 4, the probability of sampling b at time $t = T$ is strictly greater than 0. Suppose b is reachable through the path

$$(b_0, a_0, o_0) \rightarrow (b_1, a_1, o_1), ..., \rightarrow (b_{T-1}, a_{T-1}, o_{T-1}) \rightarrow b_T = b.$$

If we denote f as the smallest non-zero element in F, then in i^{th} iteration of Algorithm 4 for any i,

$$\mathbb{P}(\{b \text{ is sampled in iteration } i\}) \geq (\varepsilon f)^T > 0.$$

From the definition of $\hat{V}_T$ in Algorithm 4 we can see,

$$\begin{aligned}
&\mathbb{P}(\{V_T(b) - \hat{V}_T^{i+1}(b) > 0\}) \\
\leq &\mathbb{P}(\{b \text{ is not in } B_T^i\}) \\
= &\mathbb{P}(\{\text{None of the first } i \text{ iterations has sampled } b\}) \\
\leq &(1 - (\varepsilon f)^T)^i \rightarrow 0, \text{ as } i \rightarrow \infty.
\end{aligned}$$

Thus, $\bar{V}_T^i(b)$ converges to $V_T(b)$ in probability for any $b \in \tilde{B}_T$.

Next, we use induction to show that $\hat{V}_t^i(b)$ converges to $V_t(b)$ in probability for any $b \in \tilde{B}_t$ for all $t \leq T$. In Algorithm 4, it is easy to see that, by applying the backward induction,

$$\hat{V}_t^i(b) = \max_a \sum_{s \in S} b(s)\{r(s,a) + \sum_{o \in O} \sum_{s' \in S} \gamma F(s,a,o) \hat{V}_{t+1}^i(\Lambda(b|a,o))\}, \ \forall b.$$

Then, at time $t < T$, for any belief vector $b \in \tilde{B}_t$, a sufficient condition such that $\hat{V}_t^i(b)$ converges to $V_t(b)$ will be $\hat{V}_{t+1}^i(\Lambda(b|a,o))$ converges to $V_{t+1}(\Lambda(b|a,o))$ for any action a and observation o, i.e., $\hat{V}_{t+1}^i(b')$ converges to $V_{t+1}(b')$ for all b' reachable at time $t+1$ from b at time t. Thus, we conclude that $\hat{V}_t^i(b)$ converges to $V_t(b)$ in probability for any $b \in \tilde{B}_t$ for all $t \leq T$. □

Although Theorem 4.13 shows that Algorithm 4 converges asymptotically to the true optimal value function, we found through experimentation that the value function approx-

imated at each iteration of Algorithm 4 is not monotone. In other words, the lower bound estimate of the optimal value function given by Algorithm 4 may not be monotone non-decreasing as we keep adding new reachable belief vectors to exploit. We use the next proposition to discuss this fact. The proof is by construction given in the Appendix.

Proposition 4.14. *Denote $\hat{V}_t^i$ as the lower bound estimate of the optimal value function at time $t \leq T$ given by the i^{th} iteration of Algorithm 4. Then, $\hat{V}_t^i$ is not monotone non-decreasing in i. In other words, there may exists an MPOMDP model $\mathscr{M}$ such that $\exists t$, $\exists b$, $\exists i$,*

$$\hat{V}_t^{i+1}(b) - \hat{V}_t^i(b) < 0.$$

It is possible this non-monotone behavior could slow the rate of convergence of Algorithm 4. However, some steps in Algorithm 4 can be modified to improve its efficiency. For example, instead of using a fixed ε, we may let the value of ε change adaptively over iterations; in line $13-15$, we can sample multiple outputs and append more than one belief vector to the belief vector set in each iteration; we could also come up with a certain rule to remove some existing belief vectors in the belief vector set. However, we did not observe a huge improvement by implementing these ideas in our computational experiment shown in the next section. However, we did observe that the random sampling of system outputs in line $13-15$ is with low efficiency. We then propose another approximation algorithm based on a branch-and-bound method, which improves the output sampling efficiency.

A Tree-based branch-and-bound method

Similar to the ε-greedy sampling method discussed above, we initially create an uniform grids of the entire space of the belief vector at all time period B_t^0 for $t = 1, ..., T$. Starting from the end of time horizon T, we first calculate $\mathscr{A}_T$ as the set of all α-vectors, and $\hat{\mathscr{A}}_T$ as the set of α-vectors that are dominating at the belief vectors in B_T^0. With $\hat{\mathscr{A}}_T$,

$\hat{V}_T$ gives a lower bound of V_T. Now, besides the lower bound estimate, we use B_T^0 to derive an upper bound of V_T at iteration 0 as follows. For each $b \in B_T^0$, calculate $v_T(b)$ as

$$v_T(b) = \max_{\alpha \in \mathscr{A}_T} \alpha \cdot b,$$

and define $v(B_T^0)$ as the set

$$v_T(B_T^0) := \{(b, v_T(b)) | b \in B_T^0\}.$$

Then, since V_T is a piecewise-linear and convex function, $v_T(B_T^0)$ can be used to find an upper bound $\bar{V}_T$ of V_T by the following linear program, where for all belief vector $b \in B_T^0$,

$$\begin{aligned} \bar{V}_t(b, v_T(B_T^0)) := \min_\lambda \quad & \textstyle\sum_{b' \in B_T^0} \lambda_{b'} v_T(b') \\ \text{s.t.} \quad & \textstyle\sum_{b' \in B_T^0} \lambda_{b'} = 1, \\ & \lambda_{b'} \geq 0, \quad \forall b' \in B_T^0 \\ & \textstyle\sum_{b' \in B_T^0} \lambda_{b'} b' = b. \end{aligned}$$

Next, at time $T-1$, similar to the procedure in Section 4.5.2, we use $\hat{\mathscr{A}}_T$ and $\hat{V}_T$ to derive the lower bound estimate $\hat{\mathscr{A}}_{T-1}$ and $\hat{V}_{T-1}$. For the upper bound estimate, for each $b \in B_{T-1}^0$, calculate $u_{T-1}(b)$ as

$$u_{T-1}(b) = \arg\max_a \sum_m \sum_{s \in S} b_0^m(s) \{r^m(s,a) + \sum_{o \in O} \sum_{s' \in S} \gamma F(s,a,o) P^m(s,a,s') \bar{V}_t^m(\Lambda(b_0^m | a, o))\},$$

and define $u_{T-1}(B_{T-1}^0)$ as the set

$$u_{T-1}(B_{T-1}^0) := \{(b, u_{T-1}(b)) | b \in B_{T-1}^0\}.$$

Then, since V_{T-1} is piecewise-linear and convex, the solution of $\bar{V}_{T-1}(b, u_{T-1}(B_{T-1}^0)))$ gives an upper bound of $V_{T-1}(b)$ for all b. We can repeat these steps for time $T-2, ..., 0$ to get the lower bound estimates $\hat{V}_{T-2}, ..., \hat{V}_0$ and upper bound estimates $\bar{V}_{T-2}, ..., \bar{V}_0$.

The next step is to modify the grid of belief vectors $B_1^0, ..., B_T^0$. Starting from time $t = 0$, denote b_0 as the initial belief vector. Similar to the ε-greedy sampling method, find the

currently best action $\bar{a}$ given by

$$\bar{a} = \arg\max_{a} \sum_{m} \sum_{s \in S} b_0^m(s)\{r^m(s,a) + \sum_{o \in O} \sum_{s' \in S} \gamma F(s,a,o) P^m(s,a,s') \bar{V}_1^m(\Lambda(b_0^m|a,o))\},$$

and take action a_0 to be $\bar{a}$ with probability $1-\varepsilon$ and one of other actions with probability ε, for some $\varepsilon \in (0,1)$. After taking action a_0, instead of randomly sampling a system output, in this case we select o_0 as follows

$$o_0 = \arg\max_{o \in O} (\bar{V}_1(\Lambda(b_0|a_0,o)) - \hat{V}_1(\Lambda(b_0|a_0,o))).$$

In other words, we select the system output where the current estimate of the value function has the largest error, so that it needs more exploitation in the next iteration. With a_0 and o_0, we then add the updated belief vector $b_1 = \Lambda(b_0|a_0,o_0)$ into B_1^0 to get $B_1^0 \cup \{b_1\}$, and similarly get B_t^1 for $t = 2,...,T$.

In the next iteration, we repeat all the steps above to get new estimates of the lower and upper bound of the value function, and new belief sets until a certain stopping criterion. The overall steps for the branch-and-bound approximation method are given in Algorithm 5. Notice that at any node of the scenario tree, if there exists another node at the same level (observation or action node) whose lower bound value is greater than the upper bound value of this selected node, then this node can be pruned. Note that We did not put the pruning steps in Algorithm 5 because the pruned node will not be selected in the future automatically.

As we can see from Algorithm 5, it accelerates the convergence rate of Algorithm 4 by sampling the action with the greatest upper-bound estimate, and the observation with the largest gap between the upper-bound and lower-bound estimates. So, the asymptotic convergence of Algorithm 5 is given as a corollary of Theorem 4.13.

Corollary 4.15. *For a given MPOMDP model $\mathscr{M}$, denote $\tilde{B}_t$ as the set of all reachable belief vectors at time $t \leq T$ starting from the initial belief vector b_0 following any policies.*

Algorithm 5: The tree-based branch-and-bound approximation method.

Input : MPOMDP model $\mathscr{M}$, ε
Output: $\hat{V}_t$
Initialize B^0 as a uniform grid and $i = 0$;
repeat
  At time T, calculate $\mathscr{A}_T$;
  $\hat{\mathscr{A}_T} = \{\alpha \in \mathscr{A}_T | \alpha = \arg\max_\alpha \alpha \cdot b \text{ for some } b \in B_T^i\}$;
  $v_T(B_T^i) := \{(b, v_T(b)) | b \in B_T^i\}$;
  for $t = T-1, ..., 0$ **do**
    Calculate the set of α-vectors $\tilde{\mathscr{A}_t}$ at time t by backward induction using $\hat{\mathscr{A}_t}$;
    $\hat{\mathscr{A}_t} = \{\alpha \in \tilde{\mathscr{A}_t} | \alpha = \arg\max_\alpha \alpha \cdot b \text{ for some } b \in B_t^i\}$;
    $\hat{V}_t(b) = \max_{\alpha \in \hat{\mathscr{A}_t}} \alpha \cdot b, \ \forall b$;
    $u_t(B_t^i) := \{(b, u_t(b)) | b \in B_t^i\}$;
    $\bar{V}_t(b) = \bar{V}_t(b, u_t(B_t^i)), \ \forall b$;
  end
  for $t = 0, ..., T-1$ **do**
    $\bar{a} = \arg\max_a (r(a) + \sum_{o \in O} \mathbb{P}(o|b_t, a) \bar{V}_{t+1}(\Lambda(b_t|a, o)))$;
    $a_t = \begin{cases} \bar{a}, & \text{with probability } 1-\varepsilon \\ \text{a random action}, & \text{with probability } \varepsilon \end{cases}$;
    $o_t = \arg\max_{o \in O} (\bar{V}_{t+1}(\Lambda(b_t|a_t, o)) - \hat{V}_{t+1}(\Lambda(b_t|a_t, o)))$.;
    $b_{t+1} = \Lambda(b_t|a_t, o_t)$;
    $B_t^{i+1} = B_{t+1}^i \cup \{b_{t+1}\}$;
  end
  $i = i + 1$;
until *some stopping criterion*;

Denote V_t as the optimal value function at time $t \leq T$, and $\hat{V}_t^i$ as the lower bound estimate of the optimal value function at time $t \leq T$ given by the i^{th} iteration of Algorithm 5. Then for all $t \leq T$, for any $b \in \tilde{B}_t$,

$$\hat{V}_t^i(b) \to V_t(b) \text{ in probability, as } i \to \infty.$$

There are two main differences between the ε-greedy sampling method and the tree-based branch-and-bound method. First, the branch-and-bound method samples the best action based on the current upper-bound estimate of the value function at each time period. This can accelerate the converge rate because exploiting a sub-optimal action will give a smaller upper bound estimate of its value function, so that it will quickly become dominated by other actions in future steps; but if exploiting a sub-optimal action based on the lower-bound estimate of the value function, as in the ε-greedy sampling methods, the lower-bound estimate of the value function will become larger in the next iteration, so that it might get stuck at a sub-optimal policy. Second, the branch-and-bound method samples the system output at each time period according to the gap between the upper and lower bound estimates at the resulting belief vector. Thus, the algorithm tends to modify the belief space grid in areas with the biggest estimation error. However, a drawback compared to the ε-greedy sampling algorithm is that the branch-and-bound method requires more computational effort for the upper bound estimate of the value function, which can be a problem when the number of the sampled belief vectors becomes large. Nevertheless, improving decisions about where to modify the belief space grid could lead to overall algorithm efficiency. In practice, we can use the branch-and-bound method to get a warm start, and then switch to the ε-greedy sampling method.

4.6 Computational Experiments

In this section, we describe two computational experiments to illustrate the application of the proposed MPOMDP. The first experiment is a toy example with two POMDPs, both of which have two states, two observations, and two actions. We use this toy example to visualize the value function and optimal policy of the MPOMDP model. We also show the value of the VSS and the EVPI in this context. Furthermore, we use the toy example to compare the performance of the two approximation methods introduced in Section 5. The second computational experiment is a case study in prostate cancer AS based on the POMDP models of Chapter 3. We use the proposed MPOMDP to find the optimal timing of biopsies in AS when the cancer progression rate and test accuracy are assumed to be uncertain because of the existence of multiple plausible selections of model parameters.

4.6.1 A two-model toy example

Suppose there are two POMDP models denoted as $\mathscr{M}_m = (S, b_0, A, P^m, O, F^m, r^m)$ for $m = 1, 2$, which have a same state space, observation space, and action space

$$S = \{s_1, s_2\},\ O = \{o_1, o_2\},\ A = \{a_1, a_2\}$$

but different transition and observation probabilities

$$P^1(a_1) = \begin{pmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{pmatrix} F^1(a_1) = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix},$$

$$P^1(a_2) = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} F^1(a_2) = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix},$$

$$P^2(a_1) = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} F^2(a_1) = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix},$$

$$P^2(a_2) = \begin{pmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{pmatrix} F^2(a_2) = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$$

and the reward function

$$a_1 : r(s_1, a_1, o_1) = 2, r(s_1, a_1, o_2) = 0, r(s_2, a_1, o_1) = 0, r(s_2, a_1, o_2) = 1$$

$$a_2 : r(s_1, a_2, o_1) = 1, r(s_1, a_2, o_2) = 0, r(s_2, a_2, o_1) = 0, r(s_2, a_2, o_2) = 2.$$

with the time horizon $t = 0, 1, 2, 3, 4, 5$.

We first solve the MPOMDP model using the exact solution method, and plot the exact value function. Figure 4.2 shows the value function $V_0(b)$ at time $t = 0$. Notice that the argument of the value function, which is the belief vector of the MPOMDP model, is a 4-dimension vector with three degree-of-freedom. Thus, we plot $V_0(b)$ for various choices of $b^2(s_1)$ to illustrate the 4-dimension function $V_0(b)$. As we can see from Figure 4.2, $V_0(b)$ is a piecewise linear and convex function in b, which is consistent with Corollary 4.9. When the belief vector lies in the blue region, then the optimal action to take at time $t = 0$ will be a_1; otherwise, if the belief vector lies in the yellow region, then the optimal action will be a_2.

Next, we show the value of the VSS achieved by the MPOMDP model, and the EVPI, as discussed in Proposition 4.12. We run a simulation study on a group of 10,000 samples where 50% of them are from model $\mathscr{M}_1$, and the other 50% are from model $\mathscr{M}_2$. We apply four different policies to the study group: (1) the optimal policy given by the POMDP model $\mathscr{M}_1$; (2) the optimal policy given by the POMDP model $\mathscr{M}_2$; (3) the optimal policy given by the mean-value POMDP model (i.e., the POMDP model with the model parameter being the mean parameter of $\mathscr{M}_1$ and $\mathscr{M}_2$); (4) the optimal policy given by the MPOMDP model $\mathscr{M} = (\mathscr{M}_1, \mathscr{M}_2, \lambda = 0.5)$. We also compare the results with the case where we have the perfect information, apply the optimal policy of $\mathscr{M}_1$ to patients

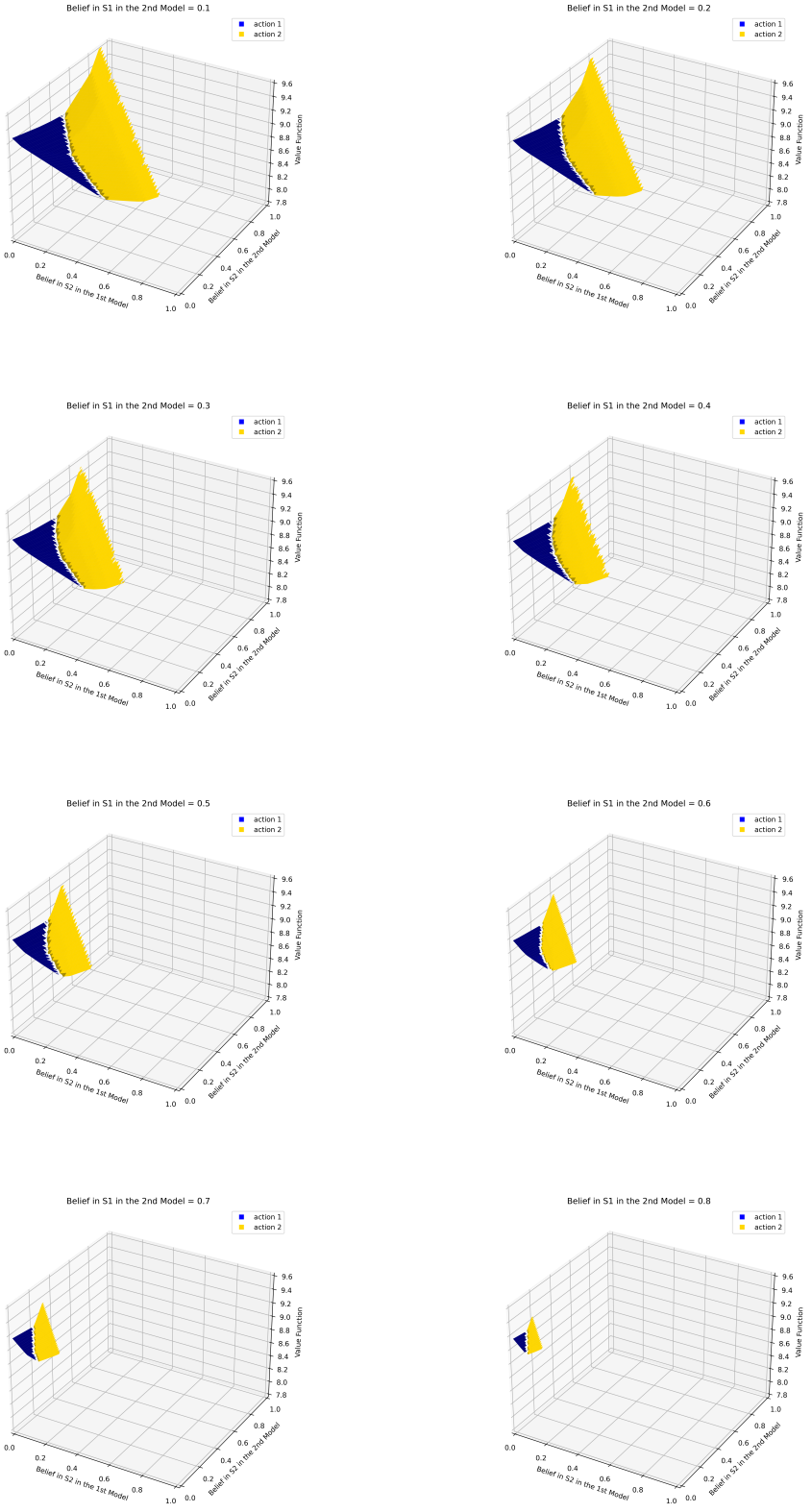

Figure 4.2: The value function $V_0(b)$ at time $t = 0$ for various choices of $b^2(s_1)$.

Belief vector b_0	Value of the optimal policy (Regret %)				
	Model $\mathscr{M}_1$	Model $\mathscr{M}_2$	Mean-value model	MPOMDP model	Perfect info.
(0.45, 0.05, 0.45, 0.05)	8.43 (9.52%)	7.81 (16.11%)	8.89 (4.55%)	9.08 (2.48%)	9.31 (0)
(0.45, 0.05, 0.25, 0.25)	8.10 (14.04%)	7.93 (15.82%)	9.03 (4.11%)	9.05 (3.95%)	9.42 (0)
(0.45, 0.05, 0.05, 0.45)	8.73 (8.40%)	8.03 (15.74%)	9.00 (5.54%)	9.18 (3.66%)	9.53 (0)
(0.25, 0.25, 0.45, 0.05)	8.29 (9.57%)	8.37 (8.65%)	8.24 (10.10%)	8.88 (3.18%)	9.17 (0)
(0.25, 0.25, 0.25, 0.25)	7.95 (14.28%)	8.46 (8.74%)	8.40 (9.42%)	8.95 (3.48%)	9.27 (0)
(0.25, 0.25, 0.05, 0.45)	8.60 (8.25%)	8.57 (8.60%)	8.51 (9.24%)	9.06 (3.38%)	9.38 (0)
(0.05, 0.45, 0.45, 0.05)	8.30 (9.65%)	8.86 (3.53%)	8.78 (4.40%)	9.07 (1.22%)	9.18 (0)
(0.05, 0.45, 0.25, 0.25)	7.96 (14.32%)	8.97 (3.41%)	8.92 (3.96%)	9.16 (1.41%)	9.29 (0)
(0.05, 0.45, 0.05, 0.45)	8.61 (8.39%)	9.12 (2.96%)	9.04 (3.80%)	9.27 (1.39%)	9.40 (0)

Table 4.1: The value function V_0 and the regrets at different initial belief vectors when applying different policies.

Belief vector b_0	$\lambda = (0.25, 0.75)$		$\lambda = (0.5, 0.5)$		$\lambda = (0.75, 0.25)$	
	VSS (%)	EVPI (%)	VSS (%)	EVPI (%)	VSS (%)	EVPI (%)
(0.45, 0.05, 0.45, 0.05)	0.11 (1.18%)	0.25 (2.71%)	0.19 (2.17%)	0.23 (2.55%)	0.58 (6.81%)	0.22 (2.39%)
(0.45, 0.05, 0.25, 0.25)	0.22 (2.41%)	0.24 (2.56%)	0.02 (0.17%)	0.37 (4.11%)	0.56 (6.65%)	0.32 (3.54%)
(0.45, 0.05, 0.05, 0.45)	0.43 (4.80%)	0.22 (2.29%)	0.18 (1.98%)	0.35 (3.80%)	0.34 (3.89%)	0.41 (4.53%)
(0.25, 0.25, 0.45, 0.05)	0.21 (2.40%)	0.38 (4.27%)	0.63 (7.70%)	0.29 (3.28%)	0.54 (6.45%)	0.20 (2.29%)
(0.25, 0.25, 0.25, 0.25)	0.05 (0.59%)	0.46 (5.18%)	0.55 (6.55%)	0.32 (3.61%)	0.69 (8.30%)	0.18 (2.03%)
(0.25, 0.25, 0.05, 0.45)	0.05 (0.50%)	0.46 (5.11%)	0.55 (6.45%)	0.32 (3.50%)	0.60 (7.10%)	0.12 (1.36%)
(0.05, 0.45, 0.45, 0.05)	0.13 (1.47%)	0.19 (2.10%)	0.29 (3.32%)	0.11 (1.23%)	0.73 (8.77%)	0.03 (0.36%)
(0.05, 0.45, 0.25, 0.25)	0.04 (0.47%)	0.22 (2.39%)	0.24 (2.66%)	0.13 (1.43%)	0.81 (9.64%)	0.00 (0.01%)
(0.05, 0.45, 0.05, 0.45)	0.10 (1.04%)	0.22 (2.34%)	0.23 (2.51%)	0.13 (1.41%)	0.73 (8.64%)	0.05 (0.59%)

Table 4.2: The VSS achieved by the MPOMDP and the EVPI for different initial belief vectors in the two-model example.

from $\mathscr{M}_1$ and the optimal policy of $\mathscr{M}_2$ to patients from $\mathscr{M}_2$.

Table 4.1 shows the values of V_0 at different initial belief vectors when applying different policies, and their regrets compared to the value function given by the optimal policy with perfect information. As we can see from Table 4.1, the optimal policy of the MPOMDP model $\mathscr{M}$ dominates the optimal policies of model $\mathscr{M}_1$, model $\mathscr{M}_2$, and the mean-value model. This says that when there exists parameter ambiguity, the MPOMDP model provides a better solution than ignoring the parameter ambiguity or averaging the model parameter.

Table 4.2 shows the VSS achieved by the MPOMDP and the EVPI for different initial belief vectors. For each initial belief vector, the VSS of the MPOMDP is calculated as the (relative) difference between the values of the mean-value POMDP model and the

Belief vector b_0	% of true optimal action over time				
	Model $\mathscr{M}_1$	Model $\mathscr{M}_2$	Mean-value model	MPOMDP model	Perfect info.
(0.45, 0.05, 0.45, 0.05)	59.70%	59.84%	79.33%	89.09%	100%
(0.45, 0.05, 0.25, 0.25)	57.91%	60.07%	85.38%	88.99%	100%
(0.45, 0.05, 0.05, 0.45)	72.79%	59.86%	86.31%	88.94%	100%
(0.25, 0.25, 0.45, 0.05)	59.40%	70.37%	61.50%	88.74%	100%
(0.25, 0.25, 0.25, 0.25)	57.60%	70.20%	67.76%	88.61%	100%
(0.25, 0.25, 0.05, 0.45)	72.51%	70.06%	69.18%	88.64%	100%
(0.05, 0.45, 0.45, 0.05)	60.14%	87.40%	79.42%	89.57%	100%
(0.05, 0.45, 0.25, 0.25)	58.76%	87.50%	85.45%	89.60%	100%
(0.05, 0.45, 0.05, 0.45)	73.70%	87.35%	86.81%	89.85%	100%

Table 4.3: The percentage of true optimal action over time compared to the optimal policy with the perfect information starting from different initial belief vectors for different policies.

MPOMDP model; and the EVPI is calculated as the (relative) difference between the values of the MPOMDP model and model with perfect information. As we can see from Table 4.2, in general, the VSS and EVPI are larger when the decision-maker is less certain about the model and state distribution. Table 4.3 also shows the percentage of true optimal action over time compared to the optimal policy when having the perfect information starting from different initial belief vectors.

Lastly, we compare the performance of the two approximation methods introduced in Section 5. We implement each approximation method with 100 iterations. Figure 4.3 reports the average error of V_0 in 20 runs. As we can see from Figure 4.3, both methods converge very fast at the beginning. However, the tree-based sampling method converges much faster after the first few iterations, with much smaller estimation errors. This is because, as discussed in Section 5, while both methods exploit the optimal action at each time period based on the current estimate of the value function, the tree-based sampling algorithm additionally calculates an upper bound estimate of the function to explore the scenarios where the current estimate has the largest error. This likely helps ensure more efficient exploration steps, and results in a faster overall convergence rate with respect to the number of iterations. On the other hand, in Table 4.4 we illustrate the computation

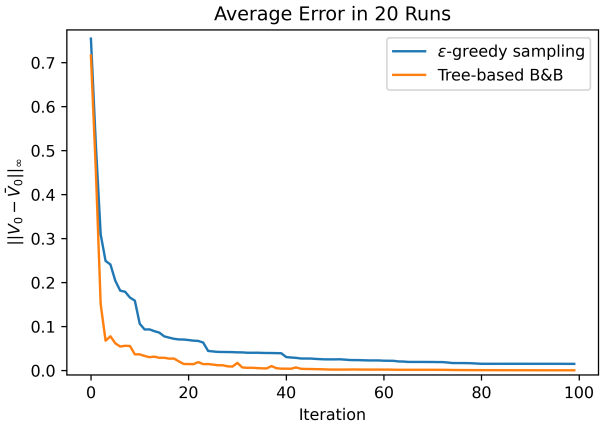

Figure 4.3: Comparisons of Algorithm 4 and 5 for the toy example of two-model POMDP.

	Exact method	ε-greedy sampling	Tree-based B&B
Mean time for each run	6836s	109s	551s
Number of α-vectors at $t = 0$	109	15	27

Table 4.4: Comparisons of the computational time and number of iterations of Algorithm 4 and 5 for the toy example of two-model POMDP.

time for each method on an Inter Core i7 2.6 GHz processor with 16 GB RAM. As we can see from Table 4.4, the tree-based sampling method takes more total computation time for each run than the ε-greedy sampling method. Thus, although the more judicious choice of belief grid modifications leads to fewer iterations for Algorithm 5, the shorter computation time per iteration of Algorithm 4 results in greater overall efficiency.

4.6.2 Case study: prostate cancer AS optimization

We implement the proposed MPOMDP model for optimizing AS for prostate cancer with imperfect information based on the POMDP models of Chapter 3. Prostate cancer is the most common cancer in men globally. Patients with low-risk variants of prostate cancer are recommended to join the AS, which monitors the patients by medical tests until there is a sign of progression to a high-risk variant of cancer, to avoid unnecessary treatments. The two most common medical tests in AS are PSA test and biopsy. As described in previous chapters, the PSA test is a simple blood test with almost no direct harm to patients. Biopsy

is a much more accurate diagnosis test, which samples the tissue with hollow-core needles to exam the severity of prostate cancer; however, biopsy is still imperfect, with potential false-negative results caused by miss-sampling. Moreover, biopsy is very painful and harmful to patients. Thus, it is critical to decide the optimal timings for biopsies for each patient in prostate cancer AS.

In Chapter 3, we used a finite-horizon two-state POMDP model to optimize the biopsy policy in each of four major prostate cancer AS study centers, which include the JH hospital, the UCSF medical center, the U of T medical center, and the PRIAS project. The objective of that study was to minimize the expected delay in the detection of high-risk prostate cancer and the expected number of lifetime biopsies. The result showed that, as different patient cohorts have heterogeneous cancer progression rates and test accuracy (model parameters), the optimal biopsy policies were different in the different study centers.

Our study in this chapter considers the case where the model parameters are not known with certainty, and we seek a single biopsy policy that works well in all four study centers. Examples may include optimizing the biopsy policy for a new patient who comes from a different area with an uncertain cancer progression rate, and for a newly initiated prostate cancer AS study that is unable to estimate the cancer dynamics (model parameters) because of a lack of data samples. For such new studies, a common strategy is to use the result from one of the previous studies as an approximate solution. The proposed MPOMDP model in this chapter allows the decision-maker to trade off all previous major studies instead of picking only one study ambitiously. The objective of the MPOMDP model, as in Chapter 3, is to minimize a weighted sum of the expected delays in the detection of high-risk prostate cancer and the expected numbers of lifetime biopsies in four study centers.

We first describe the MPOMDP model formulation $\mathscr{M}$ for optimizing prostate cancer AS adapted from the previous chapters. As introduced in Chapter 3, the decision epochs here are discrete annual time periods until age 75, which is the recommended stopping time for AS with the consideration of other cause mortality rates. There are two states in S, which are low-risk cancer state (LR) and high-risk cancer state (HR). The set A contains two actions that are "defer biopsy" and "conduct biopsy". At each decision epoch after taking action, there will be observations of PSA test and biopsy (if conducted). For the PSA test, we divide all possible outcomes into three intervals: $I_1 = [0,4]$, $I_2 = (4,10]$, and $I_3 = (10,\infty)$ (ng/mL); For biopsy, we list all possible outcomes as negative, positive, and null (not conducted) for simplicity. The transition and observation probabilities in the four different study centers were estimated in Chapter 2 and listed in Tables 4.5 and 4.6 for convenience. In Table 4.5, the misclassification error at diagnosis denotes the initial distribution b_0, the annual cancer progression rate denotes the transition probabilities, and the biopsy sensitivity denotes the observation probabilities for the biopsy. Table 4.6 denotes the observation probabilities for the PSA test. Lastly, the reward function $r(s,a,o)$ is defined as

$$r(s,a,o) = \begin{cases} 0, & a = \text{Defer Biopsy},\ s = \text{LR}; \\ \theta, & a = \text{Defer Biopsy},\ s = \text{HR}; \\ \eta, & a = \text{Conduct Biopsy},\ s = \text{LR},\ o = \text{Negative}; \\ \eta, & a = \text{Conduct Biopsy},\ s = \text{HR},\ o = \text{Positive}; \\ \theta + \eta, & a = \text{Conduct Biopsy},\ s = \text{HR},\ o = \text{Negative}; \\ \text{Not Defined}, & \text{otherwise}, \end{cases}$$

where θ and η are non-negative scalars that denote the cost of one-year delayed detection of high-risk cancer and the burden of a biopsy, respectively. We set $\theta + \eta = 1$, so that

Center	misclassification error at diagnosis: b_0	Annual Cancer Progression rate: p	Biopsy Sensitivity: $(1-\gamma)$
JH	0.0583	0.0691	0.7184
UCSF	0.0809	0.1217	0.7431
U of T	0.0774	0.1016	0.7949
PRIAS	0.0653	0.0841	0.7614

Table 4.5: The AS-POMDP model parameters in four study centers. Abbreviations: JH, Johns-Hopkins; UCSF, University of California-San Francisco; U of T, University of Toronto; PRIAS, Prostate Cancer Research International Active Surveillance.

Center	Probability Mass Function of PSA (ng/mL): q			
	Cancer State	$I_1 = [0,4]$	$I_2 = (4,10]$	$I_3 = (10,\infty)$
JH	LR Cancer	0.3552	0.4311	0.2137
	HR Cancer	0.2868	0.4706	0.2426
UCSF	LR Cancer	0.0768	0.5680	0.3552
	HR Cancer	0.0678	0.5736	0.3586
U of T	LR Cancer	0.4573	0.3422	0.2005
	HR Cancer	0.3312	0.2368	0.4320
PRIAS	LR Cancer	0.1361	0.5357	0.3282
	HR Cancer	0.1094	0.5501	0.3405

Table 4.6: The probability mass functions of PSA in four study centers. Abbreviations: JH, Johns-Hopkins; UCSF, University of California-San Francisco; U of T, University of Toronto; PRIAS, Prostate Cancer Research International Active Surveillance; LR, low-risk; HR, high-risk.

varying θ and η allows computing the optimal policy for different patient preferences for the two criteria. Here we choose $\theta = \eta = 0.5$ for simplicity.

Now, suppose that for a group of new patients, the decision-maker has no information about which model best describes the new patients. Traditionally, the decision-maker picks a single model based on their personal judgement/opinion about which is the best, and apply its optimal policy to new patients in practice. Here, our proposed MPOMDP model provides another solution to this problem. To show the benefit of the MPOMDP model, for each AS study, we compare the result of five different biopsy policies, which includes the four policies given by solving the JH, UCSF, U of T, and the PRIAS POMDP models, and the policy given by solving the MPOMDP model. For the MPOMDP model, we set a non-informative initial model weight $\lambda = (0.25, 0.25, 0.25, 0.25)$.

Center	Minimum cost of the optimal policy (regret %)				
	JH model	UCSF model	U of T model	PRIAS model	MPOMDP model
JH	2.74 (0)	2.92 (6.50%)	3.84 (40.42%)	3.01 (9.89%)	2.87 (4.80%)
UCSF	2.54 (5.35%)	2.41 (0)	2.95 (22.45%)	2.68 (11.33%)	2.49 (3.33%)
U of T	2.65 (12.34%)	2.42 (2.39%)	2.36 (0)	2.77 (17.54%)	2.40 (1.72%)
PRIAS	2.59 (4.19%)	2.63 (5.54%)	3.11 (24.71%)	2.49 (0)	2.54 (2.03%)

Table 4.7: The optimal value (minimum cost) function in different AS studies when applying different policies.

Table 4.7 shows the optimal value (minimum cost) function and the regret of each biopsy policy in each AS study center. The regret is calculated as the relative difference between the current policy and the best policy in each study center. As we can see from Table 4.7, the best policy in each study center is always the optimal policy given by the corresponding POMDP model. Moreover, the optimal policy given by the MPOMDP model is always better than policies from an inconsistent POMDP model in all four study centers. For each study center, the difference between the cost of the optimal policy given by the MPOMDP and a "wrong" model (different from the study center) is the VSS achieved by the MPOMDP model; and the difference between the MPOMDP and the "right" model is the EVPI. Figure 4.4 shows the comparison of the mean number of biopsies and average late detection time by biopsy in years in different AS studies when applying different policies in different models. Depending on how the decision-maker trades off between the mean number of biopsies and average late detection time by biopsy, the optimal policy given by the MPOMDP model is always the closest to the true optimal policy in each study center, compared with the policy given by a wrong POMDP model.

4.7 Conclusion

In this chapter, we introduced a new MPOMDP model to address the issue of parameter ambiguity in POMDP models. Motivated by the prostate cancer AS optimization problem

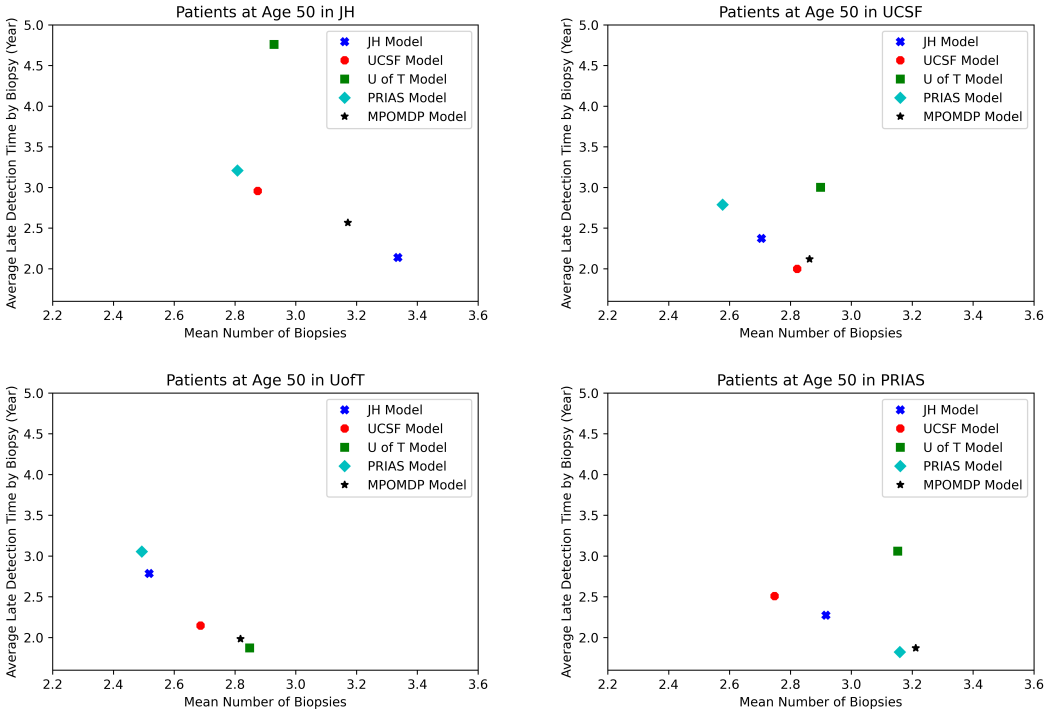

Figure 4.4: Comparisons of mean number of biopsies and average late detection time by biopsy in years in different AS studies when applying the optimal policies in different models. The reward parameter is set to be $\theta = \eta = 0.5$.

in Chapter 3, when there are multiple credible optimization models with the same structure but different model parameters, the proposed MPOMDP model can learn the model credibility from the system outputs over time, and seek a single optimal policy that maximizes the expected future rewards across models. We also discussed some structural properties of the proposed MPOMDP model, which not only reveal the benefit of the MPOMDP model by accounting for parameter ambiguity, but also motivate the solution methods to MPOMDP models. We then introduced an exact solution method and two fast approximation methods to MPOMDP models, which were shown to converge asymptotically, and compared their performance in a computational experiment. Lastly, we used the example of prostate cancer AS in Chapter 2 and 3 as a case study, to demonstrate how the MPOMDP model can be applied to a real-world problem to improve medical decision-making.

When applying the MPOMDP model to real-world problems, the model weight can be initialized by some prior knowledge or as a non-informative prior distribution over different POMDPs. Then, every time when there is new output from the system, the MPOMDP model can update the model belief so that more credible models will be assigned higher model weights. Notice that since the model weight is updated by the Bayesian formula, even if none of the POMDPs considered in the MPOMDP is the true model for the patient under consideration, the MPOMDP is still able to assign higher weights to the models with higher probability generating the observed outputs. We then showed that an MPOMDP could be reformulated as a new POMDP model with an extended state space, where a state in the new POMDP model is a combination of the current model and the state in the original POMDP model. Utilizing this property, we then derived the belief update formula for both the system state and model in an MPOMDP. Further, motivated by the one-pass algorithm for POMDP models, we introduced an exact solution method to the proposed MPOMDP model. However, because of the complexity of an MPOMDP model, even for

moderate size problems, the exact solution method is not feasible in a reasonable amount of time. We then introduced two fast approximation algorithms applying the ε-greedy and branch-and-bound sampling methods. The idea is that, instead of calculating the optimal value function of the MPOMDP over the entire belief space, we only evaluate the optimal value function on a subset of reachable belief points by sampling, and then approximate the value function on other places using the samples.

Compared with traditional robust optimization approaches, whose objectives are to optimize the worst-case performance, there are three main advantages of our MPOMDP model. First, in our MPOMDP model formulation, when considering the optimal value problem, there are some nice properties including that the belief vector is a sufficient statistic for the past information and the existence of a deterministic and Markovian optimal policy, which do not hold for robust optimization models. These properties are important because they help develop efficient solution methods so that the MPOMDP can be applicable to large real-world problems. Second, the MPOMDP model is able to learn the model credibility for each individual over time from past actions and observations, which is not the focus in robust optimization models. Third, the MPOMDP model that optimizes a weighted-average value function by the model belief usually results in a less conservative policy than the robust optimization models that optimize the worst-case value function. Therefore, on average, the MPOMDP achieves better performance than the robust optimization models.

In the computational experiments, we first used a toy example with two POMDPs to illustrate the use case of the proposed MPOMDP model. We formulated the MPOMDP with two POMDPs, solved for the optimal value function and policy exactly, and compared its performance with other traditional solutions. The results showed that the MPOMDP policy dominated the solution obtained by arbitrarily picking one POMDP model when the

wrong model was selected, and the mean-value POMDP model. This was because that the MPOMDP can consider the performance of both POMDPs according to the model weight learned from system outputs. We also used this example to compare the performance of two much faster approximation methods. The ε-greedy sampling method updated the lower bound estimate of the optimal value function in each iteration, which converged asymptotically over time. Compared with the ε-greedy sampling method, the branch-and-bound sampling method converged faster by maintaining an upper bound estimate of the value function. But it also required extra computation effort to calculate the upper bound estimates.

We further investigated the benefit of the MPOMDP model in a real case study of prostate cancer AS, which is the motivating example of this thesis discussed in Chapter 2 and 3. We showed that for a new patient starting prostate cancer AS, who may be best described by one of the models in the JH, UCSF, U of T, and PRIAS study centers, the MPOMDP model found a single optimal biopsy policy that is only slightly worse than the optimal biopsy policy given by the POMDP model of the true study center, but much better than the policies given by a wrong POMDP model and the mean-value POMDP model. Given the trade-off between the biopsy burden and late detection of a cancer progression by the decision-maker, the MPOMDP model achieved the minimum expected future costs when the true model was not known with certainty. Thus, the MPOMDP model appears to offer a robust policy that protects against uncertainty when the correct model is not known with certainty.

There are also some limitations of our work in this chapter, and opportunities for future research in parameter ambiguity in POMDP models. First, we only focused on the optimal value problem of an MPOMDP model in this chapter, where the objective was to maximize a weighted average of the value functions across different POMDPs accord-

ing to the model-state belief vector. There could be other objective functions, for example, maximizing the worst-case reward, minimize the conditional value-at-risk, and other probability measures that are widely used in stochastic programming and robust optimization problems. However, the potential issue for considering other objective functions is the existence of an optimal policy with a simple structure, for example, a deterministic and Markovian policy, that is practical for real-world problems. We leave the theoretical and methodological study of the extension to other objective functions to future research. Second, the proposed MPOMDP model only considers discrete uncertainty sets of model parameters, which assumes a limited number of the sets of the possible model parameter. This is different from many stochastic optimization works, where the uncertainty sets of model parameters are continuous. However, the work of MPOMDP in this chapter was motivated by the real-world application in prostate cancer AS, where there are several credible and competing well-established models. The framework we proposed can provide a valuable foundation for studying related problems that arise in other contexts.

4.8 Appendix: Proofs

Proof of Proposition 4.10

Recall the backward induction formula

$$\alpha_{t-1}(s) = r(a) + \gamma \sum_{z} \sum_{s'} \mathbb{P}(s'|s,a)\mathbb{P}(z|s',a) g_z^{\alpha_{t-1}}(s')$$

where $g_z^{\alpha_{t-1}}$ is a function mapping observation z to α-vectors in $\mathscr{A}_t$. We can also write this in matrix notation as follows

$$\alpha_{t-1}(s) = r(a) + \gamma \sum_{z} \sum_{s'} P_{ss'} Q_{s'z} g_z^{\alpha_{t-1}}(s').$$

Now, using the above equation,

$$\begin{aligned}
\alpha_{t-1}^{m_1}(s) - \alpha_{t-1}^{m_2}(s) =& \sum_z \sum_{s'} P_{ss'}^{m_1} Q_{s'z}^{m_1} g_z^{\alpha_{t-1}^{m_1}}(s') - \sum_z \sum_{s'} P_{ss'}^{m_2} Q_{s'z}^{m_2} g_z^{\alpha_{t-1}^{m_2}}(s') \\
=& \sum_z \sum_{s'} P_{ss'}^{m_1} Q_{s'z}^{m_1} (g_z^{\alpha_{t-1}^{m_1}}(s') - g_z^{\alpha_{t-1}^{m_2}}(s')) \\
&+ \sum_{(z,s'): P_{ss'}^{m_1} Q_{s'z}^{m_1} - P_{ss'}^{m_2} Q_{s'z}^{m_2} > 0} (P_{ss'}^{m_1} Q_{s'z}^{m_1} - P_{ss'}^{m_2} Q_{s'z}^{m_2}) g_z^{\alpha_{t-1}^{m_2}}(s') \\
&- \sum_{(z,s'): P_{ss'}^{m_1} Q_{s'z}^{m_1} - P_{ss'}^{m_2} Q_{s'z}^{m_2} < 0} (P_{ss'}^{m_1} Q_{s'z}^{m_1} - P_{ss'}^{m_2} Q_{s'z}^{m_2}) g_z^{\alpha_{t-1}^{m_2}}(s') \\
\leq& ||\alpha_{t-1}^{m_1} - \alpha_{t-1}^{m_2}||_\infty \\
&+ \sum_{(z,s'): P_{ss'}^{m_1} Q_{s'z}^{m_1} - P_{ss'}^{m_2} Q_{s'z}^{m_2} > 0} (P_{ss'}^{m_1} Q_{s'z}^{m_1} - P_{ss'}^{m_2} Q_{s'z}^{m_2}) g_z^{\alpha_{t-1}^{m_2}}(s') \\
&- \sum_{(z,s'): P_{ss'}^{m_1} Q_{s'z}^{m_1} - P_{ss'}^{m_2} Q_{s'z}^{m_2} < 0} (P_{ss'}^{m_1} Q_{s'z}^{m_1} - P_{ss'}^{m_2} Q_{s'z}^{m_2}) g_z^{\alpha_{t-1}^{m_2}}(s') \\
\leq& ||\alpha_t^{m_1} - \alpha_t^{m_2}||_\infty + \frac{\xi^s(m_1,m_2)}{2} \max_{(k,l)} \max_{s,s' \in \mathscr{S}} |\alpha_{t,k}^{m_2}(s) - \alpha_{t,l}^{m_2}(s')| \\
\leq& ||\alpha_t^{m_1} - \alpha_t^{m_2}||_\infty + \frac{\xi^s(m_1,m_2)}{2} (T-t) |R_{\max} - R_{\min}|
\end{aligned}$$

Proof of Proposition 4.14

Prove by construction. Consider an MPOMDP model $\mathscr{M} = (\mathscr{M}_1, \mathscr{M}_2, \lambda)$, where two POMDP models $\mathscr{M}_m = (S, b_0, A, P^m, O, F^m, r^m)$ for $m = 1,2$ have a same state space, observation space, and action space

$$S = \{s_1, s_2\},\ O = \{o_1, o_2\},\ A = \{a_1, a_2\}$$

but different transition and observation probabilities

$$P^1(a_1) = \begin{pmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{pmatrix} F^1(a_1) = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix},$$

$$P^1(a_2) = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} F^1(a_2) = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix},$$

$$P^2(a_1) = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} F^2(a_1) = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix},$$

$$P^2(a_2) = \begin{pmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{pmatrix} F^2(a_2) = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$$

and different reward functions

$$a_1 : r(s_1, a_1, o_1) = 2, r(s_1, a_1, o_2) = 0, r(s_2, a_1, o_1) = 0, r(s_2, a_1, o_2) = 1$$

$$a_2 : r(s_1, a_2, o_1) = 1, r(s_1, a_2, o_2) = 0, r(s_2, a_2, o_1) = 0, r(s_2, a_2, o_2) = 2.$$

We consider the time horizon to be $t = 0, 1$ and set the model weights to be $\lambda_1 = \lambda_2 = 0.5$. For any belief vector b of $\mathscr{M}$, we write

$$b = (1 - b^1, b^1, 1 - b^2, b^2),$$

where b^1 is the belief in state s_2 in $\mathscr{M}_1$ and b^2 is the belief in state s_2 in $\mathscr{M}_2$. For any α-vector α_t at time $t = 0, 1$, we write

$$\alpha_t = (\alpha_t^1, \alpha_t^2) = (\alpha_t^1(0), \alpha_t^1(1), \alpha_t^2(0), \alpha_t^2(1))$$

where α_t^1 is the α-vector in $\mathscr{M}_1$, $\alpha_t^1(0)$, $\alpha_t^1(1)$ are the values of α_t^1 at $b^1 = 0$ and $b^1 = 1$; and similarly for α_t^2. We can use the exact solution method to find the set of all α-vectors at time $t = 1$:

$$\mathscr{A}_1 = \{(1.6, 0.8, 1.2, 0.6), (0.7, 1.4, 0.9, 1.8)\}.$$

Now, apply Algorithm 4. Suppose in the first iteration we sample two belief vectors at time $t = 0$, which are $b_0^1 = (0.25, 0.25, 0.25, 0.25)$ and $b_0^2 = (0.5, 0, 0, 0.5)$; and then we sample action a_1 and observation o_1, resulting two belief vectors at time $t = 1$, which are $b_1^1 = (0.13, 0.37, 0.29, 0.21)$ and $b_1^2 = (0.07, 0.6, 0.03, 0.3)$.

We then can identify $(0.7, 1.4, 0.9, 1.8) \in \mathscr{A}_1$ as the only non-dominated α-vector at time $t = 1$, and two non-dominated α-vectors

$$(2.019, 2.631, 2.529, 3.021),\ (2.22, 2.28, 1.56, 2.94)$$

at time $t = 0$. In the next iteration, suppose we sample one more belief vector $b_0^3 = (0, 0.5, 0.5, 0)$ at time $t = 0$ and $b_1^3 = (0.45, 0.05, 0.45, 0.05)$ at time $t = 1$ following action a_1 and observation o_1. Then, using three sampled belief vectors b_1^1, b_1^2, b_1^3, we can identify all two α-vectors in $\mathscr{A}_1$ as non-dominated α-vectors at time $t = 1$.

At time $t = 0$, using three sampled belief vectors b_0^1, b_0^2, b_0^3, we can find three non-dominated α-vectors at time $t = 0$, which are

$$(2.019, 2.631, 2.529, 3.021), (2.93, 1.57, 2.19, 2.31), (2.48, 2.32, 2.34, 1.26).$$

In other words, in the first iteration, we have

$$\hat{\mathscr{A}}_0^1 = \{(2.019, 2.631, 2.529, 3.021), (2.22, 2.28, 1.56, 2.94)\}$$

and in the second iteration, we have

$$\hat{\mathscr{A}}_0^2 = \{(2.019, 2.631, 2.529, 3.021), (2.93, 1.57, 2.19, 2.31), (2.48, 2.32, 2.34, 1.26)\}.$$

Now, consider the belief point $b = (0.45, 0.05, 0, 0.5)$ at time $t = 0$:

$$\hat{V}_0^2(b) = 2.552 < 2.583 = \hat{V}_0^1(b),$$

i.e., the lower bound estimate of V_0 after the second iteration is smaller than the estimate after the first iteration at b.

CHAPTER 5

Summary and Conclusions

In this thesis, we took a holistic approach on data-driven optimization for individualized medical decision-making in cancer. We used a healthcare application in optimizing prostate cancer AS as the motivating example. First, we built an HMM to describe the stochastic system of prostate cancer AS and estimate the model dynamics, including cancer progression rate, test accuracy, and reward mechanism, by fitting the historical observational data. Second, with the estimated stochastic system, we developed a POMDP model to optimize medical decision-making in cancer that balances the benefits and harms in light of the fact that the health states are not directly observable and can progress stochastically over time. Third, we studied the issue of parameter ambiguity in POMDP models by proposing a new stochastic dynamic programming model, which is the MPOMDP model. We discussed the structural properties and solution methods for the proposed MPOMDP model, and showed its benefits in case studies. The individualized medical decision-making was achieved from three aspects: (1) individualized cancer progression paths estimated by the HMM; (2) individualized reward function definition in the POMDP and MPOMDP optimization models; (3) individualized model credibility learned by the MPOMDP from past actions and observations.

In Chapter 2, we fitted HMMs to estimate the misclassification error at diagnosis, the

annual cancer progression rate, the sensitivity and specificity of biopsy, and the distribution of PSA in four prostate cancer AS cohorts part of the GAP3 consortium: JH, UCSF, U of T, and PRIAS. With the estimated HMMs, we compared the mean number of biopsies performed versus late detection of cancer progression by biopsy when following different published biopsy protocols in four cohorts using a series of stochastic simulations. As expected, in each cohort, the biopsy protocol that recommended more frequent biopsies was associated with a shorter time to reclassification. Our results showed that no single biopsy protocol was optimal for all cohorts because of the considerable variation in biopsy under-sampling error and annual progression rates across cohorts. Moreover, in each cohort, the biopsy protocol that recommended more frequent biopsies was associated with a shorter time to reclassification, while the benefit from additional biopsies was diminishing.

From our estimates of HMMs in four different cohorts, based on the bootstrapped standard errors of the estimated parameters, all the misclassification errors at diagnosis, annual cancer grade progression rates, and biopsy false-negative rates were statistically significantly greater than zero. All biopsy specificities were close to 100%, indicating it was very rare that a patient in the low-risk cancer state would have a biopsy Gleason sum 7 or higher. For misclassification errors at the time of diagnosis and annual grade progression rates, we found that the estimates in the UCSF and U of T cohorts were greater than the estimates in JH and PRIAS cohorts. This was consistent with the fact that the UCSF and U of T cohorts included higher-risk patients than the other two cohorts. For the biopsy sensitivities, we saw that the JH cohort had the lowest estimate while the U of T cohort had the highest one. We conjectured that patients with lower risk had smaller tumors in general, so that they were harder to detect by biopsy if they were in the high-risk cancer state. Other possible reasons for such differences might include the different definitions of low and high-risk states, and the difference in the urologist practice when

performing the tests in different cohorts.

The simulation study in Chapter 2 compared three published biopsy protocols in four different cohorts. Within each cohort, the protocol that recommended more biopsies had fewer late detection years of high-risk cancer by biopsy. However, we saw that the benefit of early detection was diminishing along with the increasing number of biopsies. There was no single optimal protocol that recommended fewer biopsies but could detect high-risk cancer earlier, in any cohort. Two possible reasons for this are: first, the model parameters estimated by the HMMs and used in the simulation model were statistically significantly different for different cohorts; second, there were two competing objectives of minimizing the number of biopsies and minimizing the late detection time by biopsy when comparing the protocols.

In Chapter 3, we proposed a finite-horizon two-state AS-POMDP model to optimize the timing of biopsies in prostate cancer AS, where the objective is to minimize the number of biopsies and the delay in detection of high-risk cancer for each patient. Our study also considered two kinds of parameter ambiguity: 1) heterogeneous transition and observation probabilities in different patient cohorts, and 2) variation in decision-maker's preferences as represented by reward functions. To solve many instances of the AS-POMDP model and evaluate alternative policies resulting from different parameters, we introduced two fast approximation methods that were able to find the lower and upper bounds of the optimal value function of the AS-POMDP model. We compared the gap between the lower and upper bounds to show that our results were accurate enough for decision-making. Further, We discussed some structural properties of the AS-POMDP model that provide insight into the AS-POMDP model-based policies. We also discussed an explanation for why the dynamic biopsy policies given by the AS-POMDP model are similar to static policies recommended in the current biopsy guidelines, and we used inverse optimization

to approximate how each guideline weighs biopsy burden versus late detection of cancer progression.

In the computational results of Chapter 3, we first presented the value functions and biopsy policies given by the AS-POMDP model in four different prostate cancer AS studies, if weighted equally on the burden of one biopsy and the penalty of one-year late detection to cancer progression. We observed that the optimal value function was not always monotone in the belief state. This was because the objective of the AS-POMDP model was to investigate rather than improve patients' cancer state, and patients may leave the system without any future cost if detected as high-risk cancer. Such models can be more straightforward for studies of medical testing, and more accurate, especially when other metrics such as QALYs are hard to estimate and too obscure for decision-making. We also observed that the biopsy policies given by the optimal value function were monotone in the belief in high-risk cancer state, i.e., it would trigger a biopsy as long as the belief in the high-risk cancer state reached a threshold. The optimal biopsy policy threshold would depend on the model parameters, including cancer progression rate and biopsy sensitivity. In general, models with a higher cancer progression rate or lower biopsy sensitivity would give a lower belief threshold for conducting a biopsy. We then changed the reward weights in the reward function to see how the model-based biopsy policy depends on the decision-maker's preference on biopsy burden and late detection time in each study center. We found that the more heavily the decision-maker weighs the late detection of cancer progression, the lower the belief threshold for triggering a biopsy in the optimal biopsy policy.

Lastly, we compared the performance of the optimal biopsy policies given by the AS-POMDP model and current biopsy guidelines in four AS study centers by a simulation study. The model-based biopsy policies were all Pareto optimal. The policies based

on published guidelines were close to the efficient frontier. We also ran a hypothetical experiment using MRI in the AS-POMDP model, which showed the potential value of the AS-POMDP model with more accurate bio-markers than PSA. Lastly, we used an inverse optimization approach to estimate the reward weights implied by the current biopsy guidelines.

In Chapter 4, we introduced a new POMDP model that we referred to as an MPOMDP that addresses the issue of parameter ambiguity in POMDP models. Motivated by the prostate cancer AS optimization problem in Chapter 3, when there are multiple credible optimization models with the same structure and different model parameters, the proposed MPOMDP model can learn the model credibility from the system outputs over time, and seek a single optimal policy that maximizes the expected weighted future rewards across models. We also discussed some structural properties of the proposed MPOMDP model, which not only reveal the benefit of the MPOMDP model by accounting parameter ambiguity, but also motivate the solution methods to MPOMDP models. We then introduced an exact solution method and two approximation methods suited for MPOMDP models, and compared their performance in terms of the quality of solutions and computation times using computational experiments. Finally, we used the example of prostate cancer AS in Chapter 2 and 3 as a case study, to demonstrate the potential impact of the MPOMDP model when applied to a real-world sequential decision making problem in the context of medical decision-making.

The MPOMDP model considered multiple POMDP models simultaneously using the weight parameter. The model weight can be interpreted as the model importance, or the probability that each model being the best model describing the object to study. When applying the MPOMDP model to real-world problems, the model weight can be initialized by some prior knowledge or as a non-informative prior distribution over different POMDP

models. Then, every time when there is new output from the system, the MPOMDP model can update the model weight so that more credible models will be assigned higher model weights. We also showed that an MPOMDP could be reformulated as a new POMDP model with an extended state space, where a state in the new POMDP model is a combination of the current model and the state in the original POMDP model. Utilizing this property, we then derived the belief update formula for both the system state and model in an MPOMDP. Further, motivated by the one-pass algorithm for POMDP models, we introduced an exact solution method to the proposed MPOMDP model. However, because of the complexity of an MPOMDP model, even for moderate size problems, the exact solution method is not feasible in a reasonable amount of time. We then introduced two fast approximation algorithms applying the ε-greedy and branch-and-bound sampling methods, and showed the asymptotic convergence of each approximation method. Compared with the robust optimization approach, there are three main advantages of the MPOMDP model. First, the MPOMDP model formulation has some nice properties, including that the belief vector is a sufficient statistic for the past information and the existence of a deterministic and Markovian optimal policy. Second, the MPOMDP model is able to learn the model credibility for each individual over time from past actions and observations. Third, the MPOMDP model provides a less conservative policy that maximizes the average of the value functions weighted by the model belief.

In the computational experiments of Chapter 4, we first used a toy example with two POMDPs to illustrate the use case of the proposed MPOMDP model. We formulated the optimization problem of two ambiguous POMDPs as an MPOMDP, solved the optimal value function and policy exactly, and compared its performance with other traditional solutions. The result showed that when there were two ambiguous POMDPs, the MPOMDP solution dominated the solution by randomly pick one POMDP model, or the mean-value

POMDP model. This was because that the MPOMDP can consider the performance of both POMDPs according to the model weight learned from system outputs. From the result we can see, the VSS and EVPI are larger when the decision-maker is less certain about the model and state distribution. We also used this example to compare the performance of two fast approximation methods. The ε-greedy sampling method updated the lower bound estimate of the optimal value function in each iteration, which converged asymptotically over time. Compared with the *epsilon*-greedy sampling method, the branch-and-bound sampling method converged faster by maintaining an upper bound estimate of the value function. But it also required extra computation effort to calculate the upper bound estimates. Lastly, we demonstrated the benefit of the MPOMDP model in the case study of prostate cancer AS, which is the motivating example of this thesis in both Chapter 2 and 3. We showed that for a new patient in prostate cancer AS, who may be best described by either of the models in the JH, UCSF, U of T, and PRIAS study centers, the MPOMDP model could find a single optimal biopsy policy that is slightly worse than the optimal biopsy policy given by the POMDP model of the true study center, but much better than the policy given by a wrong POMDP model or the mean-value POMDP model. Given the trade-off between the biopsy burden and late detection of a cancer progression by the decision-maker, the MPOMDP model achieved the minimum expected future costs when the true model was not known with certainty.

There are also some limitations of the work in this thesis, which lead to opportunities for future research. In Chapter 2, we reduced a complex disease (prostate cancer) to a two-state (low-risk and high-risk cancer states) stochastic model with two outputs of the disease (results of PSA test and biopsy) as informative observations. Although such models cannot capture all details about the disease, it consistently discriminates health states on the basis of the most significant factors defining study inclusion for each cohort. We are looking

forward to improving our model formulation when more data is available. Second, our proposed HMM included the null observation of biopsy as non-informative missingness. In other words, we assumed no difference between a missed biopsy by the design of the study, and a missed biopsy result for other reasons (e.g., patient preference, data lost to follow-up). However, by using the null observation to denote the biopsy missingness in the HMM, we mitigated bias in our estimates of the model parameters. Finally, another way to monitor prostate cancer in recent AS protocols is by MRI scans, but it was not considered in this study due to the lack of sufficient longitudinal data to date.

In Chapter 3, we used a two-state POMDP model to approximate the stochastic system of prostate cancer AS, and only considered the information from PSA test and biopsy. There might be other covariates in prostate cancer AS such as prostate volume, PSA doubling time, and the results of MRI scans that could be used to understand the underlying cancer state, but were not considered in this study. We look forward to improving our model by including these factors when more data becomes available. Second, the model parameters of the transition and observation probabilities are assumed to be stationary, i.e., independent of time, which may not be accurate in reality. However, incorporating time-dependent factors would require the estimates of the model parameters in pre-studies, and more computational effort to solve the model. Third, our results of the fast approximation method for finding the upper bound of the optimal value function, and the sufficient and necessary condition for the existence of a control-limit type policy only work in a two-state POMDP model. The generalization of these results to general POMDP models may not be trivial and is left for future studies. Although the focus of this article is on prostate cancer AS, our model formulation is flexible and could be applied to other medical decision-making problems in chronic disease management.

In Chapter 4, we mainly focused on the optimal value problem of an MPOMDP model

in this chapter, where the objective was to maximize the value function across all different POMDPs. Although the weight parameter given by the model belief is adaptive in the sense that it is learned from the system outputs over time, it may not capture all purposes of a decision-maker. There could be other objectives in ambiguous POMDPs, for example, maximizing the worst-case reward, minimize the conditional value-at-risk, and other probability measures that are widely used in robust optimization. However, a potential issue for considering other objective functions will be the existence of an optimal policy that is not overly complex and practical for real-world problems, for example, a deterministic and Markovian policy. We leave the theoretical and methodological study of the extension of the optimal value problem to future research. Second, the proposed MPOMDP model only considers discrete uncertainty sets of model parameters, which assumes a limited number of possible model parameters. This is different from many stochastic optimization works, where the uncertainty sets of model parameters are often continuous. However, the work of MPOMDP in this chapter was motivated by the issue when there were multiple credible but competing well-established models for the optimization problem. We look forward to generalizing the work of the MPOMDP model to continuous uncertainty sets in other applications where there are strong needs and better fits.

In summary, this thesis presented data-driven stochastic sequential decision-making approaches with a focus on cancer screening applications. We formulated stochastic and statistical models to describe and estimate the cancer screening system, which were able to account for the stochastic progression, partial observability, and patients' heterogeneity of the cancer disease. We also developed stochastic optimization models for individualized sequential decision-making in cancer screening, with the extension to address the issue of parameter ambiguity. There are several opportunities for future research on this topic including the following. First, the methodology we applied to this thesis was from

a descriptive analysis of estimating the system of cancer surveillance using observational data, to a prescriptive analysis of optimizing sequential decision-making using optimization models built upon the descriptive models. The potential issues of this methodology could be the lack of observational data, and the delay in model updating for new observations. Development of new online learning approaches that can not only update the models with new observations in real time, but also balance the optimal strategy learned from the existing data with policy implications of newly acquired data that may reflect changes to the underlying system over time. Second, for the topic of uncertainty in POMDP models, we focused on a specific case of parameter ambiguity where there are a finite number of credible POMDPs have the same structure but different parameters, in other words, the model parameters have discrete uncertainty sets. Our work could be extended to the cases where the uncertainty sets of the model parameters are continuous, and especially following certain distributions. Third, from a theoretical perspective, when using the POMDP and MPOMDP for sequential decision-making optimization, the optimal value functions are given by the backward induction approach, so that the structures of the optimal value function and policy are obscure. For the structure of the optimal value function, the best we know is that it is piecewise-linear and convex in the belief vector; and for the optimal policy, we often want to show it is a control-limit type policy for the ease of use and explanation. Finally, it could be worthwhile to investigate structural properties of the optimal value function and policy given by backward induction, which would in turn motivate more efficient solution and approximation methods to the POMDP and MPOMDP models, and perhaps help to popularize them for real-world applications. The work in this thesis helps provide a foundation for these and other possible future directions of research.

BIBLIOGRAPHY

Albright, S. C. (1979). Structural results for partially observable markov decision processes. *Operations Research*, 27(5):1041–1053.

Anandadas, C. N., Clarke, N. W., Davidson, S. E., O'Reilly, P. H., Logue, J. P., Gilmore, L., Swindell, R., Brough, R. J., Wemyss-Holden, G. D., Lau, M. W., et al. (2011). Early prostate cancer–which treatment do men prefer and why? *BJU international*, 107(11):1762–1768.

Åström, K. J. (1965). Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205.

Ayer, T., Alagoz, O., and Stout, N. K. (2012). Or forum—a pomdp approach to personalize mammography screening decisions. *Operations Research*, 60(5):1019–1034.

Ayer, T., Alagoz, O., Stout, N. K., and Burnside, E. S. (2016). Heterogeneity in women's adherence and its role in optimal breast cancer screening policies. *Management Science*, 62(5):1339–1362.

Barnett, C. L., Auffenberg, G. B., Cheng, Z., Yang, F., Wang, J., Wei, J. T., Miller, D. C., Montie, J. E., Mamawala, M., and Denton, B. T. (2018a). Optimizing active surveillance strategies to balance the competing goals of early detection of grade progression and minimizing harm from biopsies. *Cancer*, 124(4):698–705.

Barnett, C. L., Davenport, M. S., Montgomery, J. S., Wei, J. T., Montie, J. E., and Denton, B. T. (2018b). Cost-effectiveness of magnetic resonance imaging and targeted fusion biopsy for early detection of prostate cancer. *BJU international*, 122(1):50–58.

Bastian, P. J., Carter, B. H., Bjartell, A., Seitz, M., Stanislaus, P., Montorsi, F., Stief, C. G., and Schröder, F. (2009). Insignificant prostate cancer and active surveillance: from definition to clinical implications. *European urology*, 55(6):1321–1332.

Baum, L. E. and Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73(3):360–363.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1):164–171.

Baum, L. E. and Sell, G. (1968). Growth transformations for functions on manifolds. *Pacific Journal of Mathematics*, 27(2):211–227.

Birge, J. R. (1982). The value of the stochastic solution in stochastic linear programs with fixed recourse. *Mathematical programming*, 24(1):314–325.

Boloori, A., Saghafian, S., Chakkera, H. A., and Cook, C. B. (2020). Data-driven management of post-transplant medications: an ambiguous partially observable markov decision process approach. *Manufacturing & Service Operations Management*, 22(5):1066–1087.

Bruinsma, S. M., Zhang, L., Roobol, M. J., Bangma, C. H., Steyerberg, E. W., Nieboer, D., Van Hemelrijck, M., consortium, M. F. G. A. P. P. C. A. S. G., Trock, B., Ehdaie, B., et al. (2018). The movember foundation's gap3 cohort: A profile of the largest global prostate cancer active surveillance database to date. *BJU international*, 121(5):737–744.

Bul, M., Zhu, X., Valdagni, R., Pickles, T., Kakehi, Y., Rannikko, A., Bjartell, A., Van Der Schoot, D. K., Cornel, E. B., Conti, G. N., et al. (2013). Active surveillance for low-risk prostate cancer worldwide: the prias study. *European urology*, 63(4):597–603.

Cassandra, A., Littman, M. L., and Zhang, N. L. (1997). Incremental pruning: A simple, fast, exact method for partially observable markov decision processes. In *In Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 54–61. Morgan Kaufmann Publishers.

Cassandra, A. R. (1998). A survey of pomdp applications. In *Working notes of AAAI 1998 fall symposium on planning with partially observable Markov decision processes*, volume 1724.

Cassandra, A. R., Kaelbling, L. P., and Kurien, J. A. (1996). Acting under uncertainty: Discrete bayesian models for mobile-robot navigation. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS'96*, volume 2, pages 963–972. IEEE.

Coley, R. Y., Fisher, A. J., Mamawala, M., Carter, H. B., Pienta, K. J., and Zeger, S. L. (2017). A bayesian hierarchical model for prediction of latent health states from multiple data sources with application to active surveillance of prostate cancer. *Biometrics*, 73(2):625–634.

Dall'Era, M. A., Cooperberg, M. R., Chan, J. M., Davies, B. J., Albertsen, P. C., Klotz, L. H., Warlick, C. A., Holmberg, L., Bailey Jr, D. E., Wallace, M. E., et al. (2008). Active surveillance for early-stage prostate cancer: review of the current literature. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 112(8):1650–1659.

Dall'Era, M. A., Albertsen, P. C., Bangma, C., Carroll, P. R., Carter, H. B., Cooperberg, M. R., Freedland, S. J., Klotz, L. H., Parker, C., and Soloway, M. S. (2012). Active surveillance for prostate cancer: a systematic review of the literature. *European urology*, 62(6):976–983.

Delage, E. and Iancu, D. A. (2015). Robust multistage decision making. In *The operations research revolution*, pages 20–46. INFORMS.

Delage, E. and Mannor, S. (2010). Percentile optimization for markov decision processes with parameter uncertainty. *Operations research*, 58(1):203–213.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Drake, A. W. (1962). *Observation of a Markov process through a noisy channel*. PhD thesis, Massachusetts Institute of Technology.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Epstein, J. I., Feng, Z., Trock, B. J., and Pierorazio, P. M. (2012). Upgrading and downgrading of prostate cancer from biopsy to radical prostatectomy: incidence and predictive factors using the modified gleason grading system and factoring in tertiary grades. *European urology*, 61(5):1019–1024.

Erenay, F. S., Alagoz, O., and Said, A. (2014). Optimizing colonoscopy screening for colorectal cancer prevention and surveillance. *Manufacturing & Service Operations Management*, 16(3):381–400.

Grey, A. D., Chana, M. S., Popert, R., Wolfe, K., Liyanage, S. H., and Acher, P. L. (2015). Diagnostic accuracy of magnetic resonance imaging (mri) prostate imaging reporting and data system (pi-rads) scoring in a transperineal prostate biopsy setting. *BJU international*, 115(5):728–735.

Hauskrecht, M. (2000). Value-function approximations for partially observable markov decision processes. *Journal of artificial intelligence research*, 13:33–94.

Hoffman, R. M. (2011). Screening for prostate cancer. *New England Journal of Medicine*, 365(21):2013–2019.

Inoue, L. Y., Lin, D. W., Newcomb, L. F., Leonardson, A. S., Ankerst, D., Gulati, R., Carter, H. B., Trock, B. J., Carroll, P. R., Cooperberg, M. R., et al. (2018). Comparative analysis of biopsy upgrading in four prostate cancer active surveillance cohorts. *Annals of internal medicine*, 168(1):1–9.

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

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134.

Klotz, L. (2010). Active surveillance for prostate cancer: a review. *Current urology reports*, 11(3):165–171.

Klotz, L. (2013). Prostate cancer overdiagnosis and overtreatment. *Current Opinion in Endocrinology, Diabetes and Obesity*, 20(3):204–209.

Klotz, L., Zhang, L., Lam, A., Nam, R., Mamedov, A., and Loblaw, A. (2009). Clinical results of long-term follow-up of a large, active surveillance cohort with localized prostate cancer. *Journal of Clinical Oncology*, 28(1):126–131.

Klotz, L., Zhang, L., Lam, A., Nam, R., Mamedov, A., and Loblaw, A. (2010). Clinical results of long-term follow-up of a large, active surveillance cohort with localized prostate cancer. *Journal of Clinical Oncology*, 28(1):126–131.

Li, W., Denton, B. T., and Morgan, T. (2021). Optimizing active surveillance for prostate cancer using partially observable markov decision processes. *Optimization Online*.

Li, W., Denton, B. T., Nieboer, D., Carroll, P. R., Roobol, M. J., Morgan, T. M., and Movember Foundation's Global Action Plan Prostate Cancer Active Surveillance consortium (2020). Comparison of biopsy under-sampling and annual progression using hidden markov models to learn from prostate cancer active surveillance studies. *Cancer Medicine*.

Littman, M. L., Cassandra, A. R., and Kaelbling, L. P. (1995). Efficient dynamic-programming updates in partially observable markov decision processes. Technical report, Brown University.

Lovejoy, W. S. (1987). Some monotonicity results for partially observed markov decision processes. *Operations Research*, 35(5):736–743.

Lovejoy, W. S. (1991). A survey of algorithmic methods for partially observed markov decision processes. *Annals of Operations Research*, 28(1):47–65.

Mannor, S., Mebel, O., and Xu, H. (2016). Robust mdps with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509.

Melnykov, V. and Melnykov, I. (2012). Initializing the em algorithm in gaussian mixture models with an unknown number of components. *Computational Statistics & Data Analysis*, 56(6):1381–1395.

Miehling, E. and Teneketzis, D. (2020). Monotonicity properties for two-action partially observable markov decision processes on partially ordered spaces. *European Journal of Operational Research*, 282(3):936–944.

Miller, D. C., Gruber, S. B., Hollenbeck, B. K., Montie, J. E., and Wei, J. T. (2006). Incidence of initial local therapy among men with lower-risk prostate cancer in the united states. *Journal of the National Cancer Institute*, 98(16):1134–1141.

Nakao, H., Jiang, R., and Shen, S. (2021). Distributionally robust partially observable markov decision process with moment-based ambiguity. *SIAM Journal on Optimization*, 31(1):461–488.

Ng, A. Y., Russell, S. J., et al. (2000). Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2.

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

Otten, M., Timmer, J., and Witteveen, A. (2020). Stratified breast cancer follow-up using a continuous state partially observable markov decision process. *European journal of operational research*, 281(2):464–474.

Pineau, J., Gordon, G., Thrun, S., et al. (2003). Point-based value iteration: An anytime algorithm for pomdps. In *IJCAI*, volume 3, pages 1025–1032.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Rabiner, L. and Juang, B. (1986). An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Ross, S. M. (1971). Quality control under markovian deterioration. *Management Science*, 17(9):587–596.

Saghafian, S. (2018). Ambiguous partially observable markov decision processes: Structural results and applications. *Journal of Economic Theory*, 178:1–35.

Sandıkçı, B., Maillart, L. M., Schaefer, A. J., and Roberts, M. S. (2013). Alleviating the patient's price of privacy through a partially observable waiting list. *Management Science*, 59(8):1836–1854.

Schlaifer, R. and Raiffa, H. (1961). *Applied statistical decision theory*.

Shani, G., Pineau, J., and Kaplow, R. (2013). A survey of point-based pomdp solvers. *Autonomous Agents and Multi-Agent Systems*, 27(1):1–51.

Simmons Ivy, J., Black Nembhard, H., and Baran, K. (2009). Quantifying the impact of variability and noise on patient outcomes in breast cancer decision making. *Quality Engineering*, 21(3):319–334.

Smallwood, R. D. and Sondik, E. J. (1973). The optimal control of partially observable markov processes over a finite horizon. *Operations research*, 21(5):1071–1088.

Sondik, E. J. (1978). The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations research*, 26(2):282–304.

Spaan, M. T. and Vlassis, N. (2005). Perseus: Randomized point-based value iteration for pomdps. *Journal of artificial intelligence research*, 24:195–220.

Steimle, L. N., Kaufman, D. L., and Denton, B. T. (2021). Multi-model markov decision processes. *IISE Transactions*, pages 1–39.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Thomsen, F. B., Brasso, K., Klotz, L. H., Røder, M. A., Berg, K. D., and Iversen, P. (2014). Active surveillance for clinically localized prostate cancer—-a systematic review. *Journal of surgical oncology*, 109(8):830–835.

Tosoian, J. J., Trock, B. J., Landis, P., Feng, Z., Epstein, J. I., Partin, A. W., Walsh, P. C., and Carter, H. B. (2011). Active surveillance program for prostate cancer: an update of the johns hopkins experience. *J Clin Oncol*, 29(16):2185–2190.

Vlassis, N., Littman, M. L., and Barber, D. (2012). On the computational complexity of stochastic controller optimization in pomdps. *ACM Transactions on Computation Theory (TOCT)*, 4(4):1–8.

White, C. C. (1979). Optimal control-limit strategies for a partially observed replacement problem. *International Journal of Systems Science*, 10(3):321–332.

White, C. C. (1991). A survey of solution techniques for the partially observed markov decision process. *Annals of Operations Research*, 32(1):215–230.

Wiesemann, W., Kuhn, D., and Rustem, B. (2013). Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183.

Xu, H. and Mannor, S. (2012). Distributionally robust markov decision processes. *Mathematics of Operations Research*, 37(2):288–300.

Zhang, J. and Denton, B. T. (2018). Partially observable markov decision processes for prostate cancer screening, surveillance, and treatment: a budgeted sampling approximation method. *Decision Analytics and Optimization in Disease Prevention and Treatment*, pages 201–222.

Zhang, J., Denton, B. T., Balasubramanian, H., Shah, N. D., and Inman, B. A. (2012a). Optimization of prostate biopsy referral decisions. *Manufacturing & Service Operations Management*, 14(4):529–547.

Zhang, J., Denton, B. T., Balasubramanian, H., Shah, N. D., and Inman, B. A. (2012b). Optimization of psa screening policies: a comparison of the patient and societal perspectives. *Medical Decision Making*, 32(2):337–349.

Zhang, N. L. and Liu, W. (1996). Planning in stochastic domains: Problem characteristics and approximation. Technical report, Technical Report HKUST-CS96-31, Hong Kong University of Science and Technology.