Modeling and Optimizing Resource Allocation Decisions through Multi-model Markov Decision Processes with Capacity Constraints

by

Onur Demiray

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Industrial Engineering

KOÇ ÜNİVERSİTESİ

November 19, 2020

Modeling and Optimizing Resource Allocation Decisions through Multi-model Markov Decision Processes with Capacity Constraints

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Onur Demiray

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Prof. Lerzan Örmeci (Advisor)

_____

Assoc. Prof. Evrim Didem Güneş [Co-Advisor]

_____

Assoc. Prof. Simge Küçükyavuz

_____

Assist. Prof. Barış Yıldız

Date: _____

To my beloved family

ABSTRACT

Modeling and Optimizing Resource Allocation Decisions through Multi-model Markov Decision Processes with Capacity Constraints
Onur Demiray
Master of Science in Industrial Engineering
November 19, 2020

This thesis proposes a new formulation for the dynamic resource allocation problem, which converts the traditional MDP model with known parameters and no capacity constraints to a new model with uncertain parameters and a resource capacity constraint. Our motivating example comes from a medical resource allocation problem: patients with multiple chronic diseases can have either regular or special care, where the capacity of special care is limited due to financial or human resources. In such systems, it is difficult, if not impossible, to generate good estimates for the disease evolution for each patient. We formulate the problem as a two-stage stochastic integer program. However, it becomes easily intractable in larger instances of the problem for which we propose and test a parallel approximate dynamic programming algorithm. We show that commercial solvers are not capable of solving the problem instances with a large number of scenarios. Nevertheless, the proposed algorithm provides a solution in seconds even for very large problem instances. In our computational experiments, it finds the optimal solution for 42.86% of the instances. On aggregate, it achieves 0.073% mean gap value. Finally, we estimate the value of our contribution for different realizations of the parameters. Our findings show that there is a significant amount of additional utility contributed by our model.

ÖZETÇE

Kaynak Dağıtımı Kararlarının Kapasite Kısıtlı Çok Modelli Markov Karar Süreçleri ile Modellenmesi ve Eniyilenmesi
Onur Demiray
Endüstri Mühendisliği, Yüksek Lisans
19 Kasım 2020

Bu tez dinamik kaynak dağıtımı problemi için yeni bir formülasyon önermektedir. Bu formülasyon, parametrelerin bilindiği varsayılan ve kapasite kısıtları içermeyen geleneksel MDP modelini parametrelerdeki belirsizliklerin ve kapasite kısıtlarının da dahil edildiği yeni bir modele dönüştürmektedir. Bu formülasyon, birden fazla kronik hastalığı olan hastalara kullanımı kısıtlı olan özel bakım kaynaklarının nasıl paylaşılacağının çalışıldığı bir medikal kaynak dağıtımı probleminden esinlenerek geliştirilmektedir. Bu sistemlerde herbir hasta için sağlık durumunun gelişimini doğru tahmin etmek oldukça güçtür. Bu çalışmada bu problem 2 aşamalı stokastik tamsayılı programlama ile modellenmektedir. Fakat, bu model büyük ölçekli problem örnekleri için çözülmesi zor bir duruma gelebilmektedir. Bundan dolayı, bu çalışmada bir paralel yaklaşık dinamik programlama algoritması önerilmektedir. Tezde hesaplamalı deneylerle ticari eniyileme yazılımlarının yüksek sayıda senaryo içeren problem örneklerinde yetersiz kaldığı gösterilmektedir. Bununla birlikte, önerilen algoritma çok büyük ölçekli problem örneklerinde bile saniyeler içerisinde çözümler üretmektedir. Önerilen algoritma, hesaplamalı deneylerde kullanılan problem örneklerinin %42.86'sında en iyi çözümü bulmaktadır. Bütün problem örnekleri hesaba katıldığında, önerilen algoritma ortalama olarak en iyi çözümden %0.073 uzaklıkta sonuçlar üretmektedir. Son olarak, çeşitli parametreler altında modelin değeri tahmin edilmektedir. Seçilen örnek çalışmada yapılan hesaplamalı deneyler göstermektedir ki yeni geliştirilen model önemli ölçüde ek fayda sunmaktadır.

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my master thesis advisors Prof. Lerzan Örmeci and Assoc. Prof. Evrim Didem Güneş. It would not be possible to have such a productive year without their diligent guidance and constant support. I would like to especially thank them for being more than a thesis advisor.

I would like to thank the rest of my dissertation committee members, Assoc. Prof. Simge Küçükyavuz and Assist. Prof. Barış Yıldız for sparing their valuable time. I wish also thank Assoc. Prof. Hakan Gültekin, Assist. Prof. Gültekin Kuyzu, and Assist. Prof. Eda Yücel for inspiring me from the first years of my undergraduate education.

I would like to especially thank my family for all of their sacrifices, continuous support, and unconditional love. I would also like to thank my lovely girlfriend Ekin for her infinite support.

There are also three special names, with whom I started this journey. First, I would like to thank my dear childhood friend Yasin. We had a lot of fun in our apartment with our friends, and I will always cherish our memories. I would also thank to Baturhan and Sude for all the times we spent together. We confront many difficulties together both at TOBB ETU and Koç University.

I would like to thank Hakan and Rabia for their feedbacks on my research.

I would like to thank Aysima, Beyza, Cem, Doğancan, Faruk, Feray, Merdan, Mücahit, and Özgür for making this process enjoyable.

Lastly, I would like to thank my close friends Alperen, Batu, Bilgehan, Gazi, Hilal, Mücahit, and Pınar.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

LP	Linear Programming
SP	Stochastic Programming
EF	Extensive Form
MDP	Markov Decision Process
MRP	Markov Reward Process
PMF	Probability Mass Function
SDP	Stochastic Dynamic Programming
VSS	Value of Stochastic Solution
VPI	Value of Perfect Information
PAM	Patient Activation Measure
CDC	Centers for Disease Control and Prevention
EVSS	Expected Value of Stochastic Solution
EVPI	Expected Value of Perfect Information
MMDP	Multi-model Markov Decision Process
POMDP	Partially Observable Markov Decision Process

Chapter 1

INTRODUCTION

Markov Decision Processes (MDPs) are successfully used to find optimal policies in sequential decision making problems under uncertainty. The application areas of MDPs vary from inventory management, finance, robotics, telecommunication to humanitarian logistics [Altman, 1999, Boucherie and Van Dijk, 2017, Meraklı and Küçükyavuz, 2019, Puterman, 2014]. It is also extensively employed in medical decision making literature. Practices in this literature range from determining the initiation time of a drug to scheduling patients for a treatment [Alagoz et al., 2004, Denton et al., 2009, Denton, 2018, Shechter et al., 2008]. We refer to [Alagoz et al., 2010] for a comprehensive review on the medical practices of MDPs. However, they have certain limitations, as they assume that both transition probability matrices for all actions and the rewards incurred for all state-action pairs are known. In practice, these parameters are generally estimated by maximum likelihood estimators derived using observational data [Zhang et al., 2017a]. However, having the right data set and deriving reliable estimators are notoriously hard and it is prone to errors which may cause suboptimal solutions [Denton, 2018, Grand-Clement et al., 2020, Mannor et al., 2016, Mannor et al., 2007, Meraklı and Küçükyavuz, 2019, Zhang et al., 2017a, Zhang et al., 2017b]. Hence, in recent years, there are studies on MDP models with uncertain parameters, such as [Steimle et al., 2018]. Another important feature that traditional MDP models generally ignore is the capacity constraints, which has also received increasing attention in the recent years [Ayvaci et al., 2012, Cevik et al., 2018, Deo et al., 2013]. Our work combines these two emerging areas, as we consider a system with uncertain transition probability matrices and rewards, which operate under a resource constraint.

Most of the studies in the literature employ robust optimization or stochastic programming to deal with parameter uncertainty in MDPs. However, all of them, to the best of our knowledge, address only the unconstrained version of the model. Nevertheless, some healthcare applications employing MDPs necessitate taking capacity constraints into account due to existance of resource scarcity [Ayvaci et al., 2012, Cevik et al., 2018, Deo et al., 2013]. [Deo et al., 2013] proposed a constrained MDP model scheduling patients for a community-based chronic care delivery problem. [Ayvaci et al., 2012] and [Cevik et al., 2018] developed constrained MDP and partially observable MDP (POMDP) models of diagnostic and screening decisions for breast cancer. However, none of these studies incorporated parameter uncertainty which may cause suboptimal solutions for such medical resource allocation problems.

With a motivation from medical resource allocation problems, we, in this paper, introduce a capacity constrained, finite-horizon, and finite-state MDP model with uncertain transition probabilities and rewards. In addition to the finite number of non-absorbing states, we consider only one absorbing state to reflect all conditions that stop the decision process. Under the healthcare setting, this absorbing state corresponds to all the terminating conditions such as death, stroke, or hearth failure. The another assumption we make in our model is that the action space consists of two actions, which is consistent considering the nature of the resource allocation problems (allocate versus don't allocate or accept versus reject). Though we restrict the model to have two actions in each decision epoch, the formulation can be extended easily to the cases with more complicated action spaces. Moreover, the model can represent any capacity constrained problem using MDPs, i.e., its scope is larger than medical resource allocation problems. For example, our approach can be applied to the dynamic product portfolio management problem [Seifert et al., 2016] under a financial constraint, where the probabilities associated with product life cycle transitions are uncertain. The case of a varying level of working capital over different business cycles can also be addressed. Thus the proposed model can be used under different domains spanning from finance, inventory management to marketing.

We adopt a multi-model MDP approach to solve our model. In other words, we provide an extensive-form formulation that is an approximation of the underlying two-stage stochastic integer program with the help of multiple MDP representations, i.e., scenarios. In this respect, our model can be considered as an extension of the model developed by [Steimle et al., 2018] with capacity constraints. They formulated their models based on the primal linear program (LP) developed for finding optimal policies for an unconstrained MDP. However, we build our formulation based on the dual LP which is the version allowing us to incorporate additional constraints to the MDP framework. To the best of our knowledge, this is the first study combining capacity constraints and parameter uncertainties in MDPs.

The extensive-form formulation, i.e., the deterministic equivalent of the two-stage stochastic programming formulation, can easily become intractable because of capacity constraints, scenarios, and decision epochs. With this in mind, we first prove results leveraging the problem structure, and then develop a parallel approximate dynamic programming algorithm based on these results. This algorithm makes us capable of solving the model efficiently even if we have a large number of scenarios and decision epochs. Lastly, the proposed model and the corresponding solution approaches are applied to a chronic care delivery problem, which also forms the basis for the computational experiments. Through computational studies, we compare different solution approaches. We also measure the value of perfect information and stochastic solution introduced by taking transition probability and reward uncertainties into account. Then, we address specific managerial aspects such as price of fairness, value of flexibility, and capacity management. It is worth to mention that all experiments are conducted for a fixed state space which provides a reasonable number of states for our analysis. Thus, our algorithm reflects to the challenges for increasing number scenarios and decision epochs.

Our main contributions can be summarized as follows:

- We introduced a new model incorporating both parameter uncertainty and capacity constraints in MDPs. To the best of our knowledge, this is the first study considering these two dimensions together.

- We prove problem specific structural results, and propose a novel network representation, which together enable a parallel approximate dynamic programming algorithm to solve larger instances of the problem.

- We conduct extensive computational experiments to measure the value of embracing parameter uncertainty and to compare the solution approaches using the chronic care delivery problem setting as an example.

- We provide managerial insights for potential practitioners with a special focus on the price of fairness, the value of flexibility, and capacity management.

The remainder of this thesis is organized as follows. A detailed literature review is made in Chapter 2, and the necessary background is provided in Chapter 3. The problem is formally described and mathematically modeled in Chapter 4. Chapter 5 is devoted to explain the proposed solution approach. Then, a comprehensive computational study based on the chronic care delivery problem is provided in Chapter 7, after the problem is introduced in Chapter 6. Finally, Chapter 8 concludes the thesis and draws directions for future research.

Chapter 2

RELATED WORK

This study mainly contributes to the literature addressing parameter uncertainties within MDP components. Hence, Section 2.1 is devoted to analyze this literature. In addition to this, both our motivation and the illustrative case study belong to the medical decision-making literature utilizing MDPs. In this regard, Section 2.2 studies the existing body of the literature employing MDPs as a modeling tool in medical decision making problems. Then, Section 2.3 specifically focuses on the studies in medical decision making literature both using MDPs and taking parameter uncertainties into account. Lastly, Section 2.4 addresses the corresponding part of the literature where capacity constraints are taken into account.

2.1 Parameter Uncertainty in MDPs

Robust optimization is extensively employed to deal with parameter uncertainty in MDPs [Grand-Clement et al., 2020, Meraklı and Küçükyavuz, 2019, Zhang et al., 2017a]. Robust optimization is one of the widely used techniques for optimization under uncertainty where no probability distribution is assumed for uncertain data. It is based on the worst-case realization of the parameters from a set of alternatives what is so-called uncertainty set [Ben-Tal et al., 2009, Ben-Tal and Nemirovski, 2002, Bertsimas et al., 2011, Gorissen et al., 2015]. A number of studies approached the parameter uncertainty in MDP problem employing robust optimization with a polyhedral uncertainty set due to its properties allowing tractable solution algorithms [Givan et al., 2000, Satia and Lave Jr, 1973, Tewari and Bartlett, 2007, White III and Eldeib, 1994], whereas another group modeled the problem with the rectangularity assumption which ensures the independence of the rows in transition probability matrices [Iyengar, 2005, Nilim and El Ghaoui, 2004, Sinha

and Ghate, 2016]. [Wiesemann et al., 2013] studied robust MDPs by relaxing this assumption.

Parameter uncertainty in MDPs is also examined with stochastic programming where different types of Bayesian approaches are adopted [Meraklı and Küçükyavuz, 2019]. In this context, [Steimle et al., 2018] propose a multi-model MDP (MMDP) approach for a finite-horizon MDP, where transition probabilities and rewards can come from different models. They further claim that having a single policy optimized with respect to all models can be a good idea to capture variations within MDP parameters. An MMDP model can be seen as a two-stage stochastic programming formulation where different models correspond to different scenarios. [Meraklı and Küçükyavuz, 2019] provided a stochastic programming model for an infinite-horizon MDP model under a risk-averse setting, and applied it to a humanitarian inventory management problem.

2.2 Medical Decision Making Applications of MDPs

As policy makers discover the potential impact, we have witnessed a growing interest in operations research methodologies, particularly MDP models, for medical decision-making problems in recent years. Within this context, [Kurt et al., 2011] developed a model to find optimal statin initiation policies where statin is useful for lipid abnormalities, but dangerous due to its side effects. Moreover, [Alagoz et al., 2004] developed a discrete-time, infinite-horizon, discounted MDP model which determines the optimal timing of living-donor liver transplantation by maximizing patients' quality-adjusted life expectancy. Then, [Shechter et al., 2008] decided when to initiate HIV treatment through an MDP model by considering the trade-off between negative side effects and potential damages caused by a delayed therapy. In a similar manner, [Denton et al., 2009] studied when and which patients with particular cardiovascular risk scores and attributes should be targeted for statin therapy that is effective for the patients with diabetes. [Denton, 2018] also examined the optimal clinical management of chronic diseases through the lens of MDP and POMDP models. We refer to [Alagoz et al., 2010] for further healthcare applications of MDPs.

2.3 Medical Decision Making Applications of MDPs with Parameter Uncertainty

MDPs with transition probability and/or reward uncertainties are also studied in the field of medical decision making. [Zhang et al., 2017a] is the first study addressing this issue in the literature. They retained the rectangularity assumption. Even though this allowed them to develop a tractable solution approach, it also led to conservative solutions as a result of taking extremely rare conditions into account. In order to overcome this challenge, they used the concept of budget uncertainty which controls the conservatism level by including an upper bound on the deviations from the nominal model [Bertsimas and Thiele, 2006a, Bertsimas and Thiele, 2006b]. They applied their robust MDP framework to a treatment planning problem where an agent decides when to initiate medications for glycemic control which is vital for the patients with type 2 diabetes. Then, [Grand-Clement et al., 2020] developed a robust MDP framework in order to produce sound strategies for proactive intensive care unit transfers. They used factor matrix uncertainty model instead of rectangularity assumption to avoid conservative solutions. The factor matrix uncertainty set which relaxes the independence of rows of transition probabilities assumes that transition probability rows are jointly varied. This is first proposed by [Goh et al., 2018] and further theoretically analyzed by [Goyal and Grand-Clement, 2018].

2.4 Medical Decision Making Applications of Capacity Constrained MDPs

MDPs are not only employed for individualistic models, but also extensively used in population level medical decision making problems where a fixed amount of a resource is tried to be efficiently allocated to the patients in a population. In other words, they are kind of public health problems in the form of resource allocation framework. This framework exclusively focuses on the allocation of a medical resource to a subset of individuals at certain periods due to scarcity, where operational (capacity constraints) and medical (clinical outcomes) aspects mutually effect each other [Deo et al., 2013]. From this perspective, [Deo et al., 2013] is the first study

developing this framework. In this study, the authors present a model of a healthcare delivery system specialized for the patients with chronic diseases. In this model, the social planner solves a scheduling problem in which different patient groups are assigned to the service under a capacity constraint in each decision period. Moreover, they consider both medical and operational aspects. After this study, [Ayvaci et al., 2012] and [Cevik et al., 2018] studied the allocation of preventative instruments for breast cancer which causes the majority of cancer deaths for women. [Cevik et al., 2018] studied the allocation of mammography screening tools among women, whereas [Ayvaci et al., 2012] had focused on the diagnostic process happened after a mammography screening activity. Both have studied their problems under limited capacity due to labor and technology constraints, and also allowed the model to consider the progress of women's health status, while assuming the parameters of the problem are known.

Chapter 3

BACKGROUND

The basis of the model in this study is an MDP. All extensions arising from the novelty of the problem are built on this model. For this reason, it is important to wise up to the essentials of an MDP model. In this regard, Section 3.1 is devoted to remark the fundamentals of MDPs. Then, Section 3.2 explains the solution techniques developed for finding optimal policies to them. Finally, Section 3.3 makes reference to the 2-stage SP with which we are able to solve the optimization problems where there exists an uncertainty within problem parameters. The last section is required for the complete comprehension of the proposed solution technique in the study.

3.1 Markov Decision Processes

MDPs, also referred to as SDPs have accomplishedly utilized for finding optimal policies for sequential decision making problems under uncertainty [Ross, 2014b]. They have been applied to vast range of areas such as biology, economics, and engineering [Puterman, 2014]. As the name suggests, it satisfies the Markov property defined in Definition 3.1 [Ross, 2014a].

Definition 3.1 *Consider a stochastic process* $\{S_n : n \in \mathbb{N} \cup \{0\}\}$ [1] *and let* S_n *be the state of the process at time* $n \in N \cup \{0\}$*. Then, Markov property states that the future of the process is independent of how it has reached its current state. It just depends on the present state, i.e.,*

$$\mathbb{P}\{S_{n+1} = s_{n+1} | S_0 = s_0, \ldots, S_n = s_n\} = \mathbb{P}\{S_{n+1} = s_{n+1} | S_n = s_n\} \tag{3.1}$$

for all $s_0, \ldots, s_{n+1} \in \mathcal{S}$ *and* $n \geq 0$ *where* $\mathcal{S}$ *is the set of the states.*

[1] Here we focus on a discrete-time stochastic process.

An MDP is a decision process in which the underlying evolution of states is endowed with the Markov property and is dependent on decisions.

Definition 3.2 *An MDP is a tuple* $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$ *where*

- $\mathcal{S}$ *and* $\mathcal{A}$ *are the finite sets of states and actions, respectively.*
- $\mathcal{P} = \{P_{iaj}\}_{(i,j)\in\mathcal{S}\times\mathcal{S}, a\in\mathcal{A}}$ *is the transition probabilities, i.e.,*

$$P_{iaj} \triangleq \mathbb{P}\{S_{t+1} = j | S_t = i, A_t = a\}, \quad \forall (i,j) \in \mathcal{S}\times\mathcal{S}, a \in \mathcal{A} \tag{3.2}$$

  where A_t *denotes the action taken at time* t*, and* P_{iaj} *represents the probability that the process will reach to state* $j \in \mathcal{S}$ *just after state* $i \in \mathcal{S}$ *when the course of action is* $a \in \mathcal{A}$*.* $\mathcal{P}$ *has to satisfy the following conditions:*

$$P_{iaj} \geq 0, \qquad \forall (i,j) \in \mathcal{S}\times\mathcal{S}, a \in \mathcal{A} \tag{3.3a}$$

$$\sum_{j\in\mathcal{S}} P_{iaj} = 1, \qquad \forall i \in \mathcal{S}, a \in \mathcal{A} \tag{3.3b}$$

- $\mathcal{R} = \{R_{ia}\}_{i\in\mathcal{S}, a\in\mathcal{A}}$ *is the set of immediate rewards where* R_{ia} *shows the immediate reward (utility) obtained by taking action* $a \in \mathcal{A}$ *when the process is in state* $i \in \mathcal{S}$
- $\gamma \in [0,1]$ *is a discount factor. [Sutton et al., 1998]* [2]

The behavior of an agent [3] is characterized by what is called *policy*. A policy is denoted by π and is nothing but a conditional probability mass function over the set of actions for a given state. That is,

$$\pi(a|i) = \mathbb{P}\{A_t = a | S_t = i\}, \qquad \forall a \in \mathcal{A}, i \in \mathcal{S} \tag{3.4a}$$

$$\sum_{a\in\mathcal{A}} \pi(a|i) = 1, \qquad \forall i \in \mathcal{S} \tag{3.4b}$$

Moreover, a policy may be deterministic. If so, it becomes a function $\pi : \mathcal{S} \to \mathcal{A}$ which maps each state in the state space into an action.

[2] We slightly change the notation to be compatible with the rest of the study.

[3] Agent, policy-maker, and decision-maker are all used interchangeably.

The value function $v(i)$ is the total expected reward starting from state $i \in \mathcal{S}$ and acting optimally afterwards. It has to satisfy the Bellman optimality conditions. That is,

$$v(i) = \max_{a \in \mathcal{A}} \Big\{ R_{ia} + \gamma \sum_{j \in \mathcal{S}} P_{iaj} v(j) \Big\}, \quad \forall i \in \mathcal{S} \tag{3.5}$$

So far, we reconnoitred infinite-horizon MDPs in which an agent always follows the same policy π^* whenever she is required to take an action[4]. However, some problems require separate decision blocks for which the agent may apply different policies. More specifically, a finite-horizon problem, in addition to the components listed in Definition 3.2, consists of decision epochs denoted by $\mathcal{T} = \{1, \ldots, T\}$. At each time $t \in \mathcal{T} \setminus \{T\}$, the agent is expected to apply π^{t*}. The collection of policies applied at each decision epoch constitutes the strategy denoted by $\Pi^* = \{\pi^{t*}\}_{t \in \mathcal{T} \setminus \{T\}}$.

In finite-horizon problems, the value function $v_t(i)$ becomes the total expected reward starting from state $i \in \mathcal{S}$ at time $t \in \mathcal{T}$ and acting optimally afterwards. Then, Bellman operator implies the following optimality conditions.

$$v_T(i) = R_i, \qquad \forall i \in \mathcal{S} \tag{3.6a}$$

$$v_t(i) = \max_{a \in \mathcal{A}} \Big\{ R_{ia} + \sum_{j \in \mathcal{S}} P_{iaj} v_{t+1}(j) \Big\}, \qquad \forall t \in \mathcal{T} \setminus \{T\}, i \in \mathcal{S} \tag{3.6b}$$

Where R_i shows the final stage reward of state $i \in \mathcal{S}$, which indicates the immediate reward obtained by finalizing the decision process in state $i \in \mathcal{S}$. It is also worth to mention that we are no longer required to use the discount factor γ.

3.2 Solving Markov Decision Processes

For an agent, the important thing is to find the optimal policy and thus the optimal value function. For finite-horizon problems, there exists a polynomial time algorithm called backward induction. For finite-horizon problems, three efficient methods draw the attention of the researchers: Value iteration, policy iteration, and

[4]We accept the optimization principle indicating that a rational economic agent always performs the best possible option [Varian, 1996].

linear programming. We refer to [Sutton et al., 1998] and [Ross, 2014b] for the methods apart from linear programming which we explain in this section. The reason why we only focus on the LP approach stems from the fact that our mathematical model is built on the dual LP since it both allows us to express the parameter uncertainties and capacity constraints, which is not possible through other approaches. Although we show the LP model for the infinite-horizon case, it is slightly extended to the finite-horizon problems by adding additional index for the decision epochs as we carry out in the study.

$$\begin{aligned} \text{(Primal LP)} \quad & \underset{\mathbf{v}\in\mathbb{R}^{|\mathcal{S}|}}{\textbf{minimize}} \quad \mathbf{w}^T\mathbf{v} \\ & \textbf{subject to;} \quad v(i) \geq R_{ia} + \gamma \sum_{j\in\mathcal{S}} P_{iaj} v(j), \quad \forall i \in \mathcal{S}, a \in \mathcal{A} \end{aligned} \tag{3.7}$$

The primal LP in (3.7) gives us the optimal value function and therefore the optimal policy. Here, $\mathbf{w}$ is a $|\mathcal{S}|$ dimensional vector indicating the prior probabilities. In other words, w_i shows the probability that the process starts the decision process in state $i \in \mathcal{S}$. $\mathbf{1}^T\mathbf{w} = 1$ has to be satisfied since $\mathbf{w}$ must constitute a valid PMF.

Although (3.7) could be easily solved due to its pure LP structure, this formulation, with its current form, is not handy for the constrained MDP models since it is not obvious how to define the constraints. To tackle this challenge, the dual form of the primal LP is introduced as follows.

$$\begin{aligned} \text{(Dual LP)} \quad & \underset{X\in\mathbb{R}_+^{|\mathcal{S}|\times|\mathcal{A}|}}{\textbf{maximize}} \quad R \odot X \\ & \textbf{subject to;} \quad \sum_{a\in\mathcal{A}} X_{ja} - \gamma \sum_{i\in\mathcal{S}} \sum_{a\in\mathcal{A}} X_{ia} P_{iaj} = w_j, \quad \forall j \in \mathcal{S} \end{aligned} \tag{3.8}$$

Where X is called *occupancy measure matrix*.

Constraints in (3.8) can be interpreted as flow conservation constraints. In other words, what is done in dual LP is nothing but just maximizing the total expected reward under the condition of probability flow conservation. This interpretation will be useful to develop the model for the finite case in this study.

3.3 Two-Stage Stochastic Programming

In two-stage SPs, the decision is divided into two parts separated by an intervening stochastic event represented by a random data vector $\tilde{\omega}$ [Küçükyavuz and Sen, 2017]. A general form of the mathematical formulation for a two-stage SP is represented as follows:

$$\begin{aligned} \text{(2-SP)}\ \textbf{minimize} \quad & c^T x + \mathbb{E}[h(x, \tilde{\omega})] \\ \textbf{subject to;} \quad & Ax \geq b \\ & x \in \mathcal{X} \end{aligned} \tag{3.9}$$

where for a particular scenario ω of $\tilde{\omega}$, $h(x, \omega)$, what is so-called recourse function, is defined as follows:

$$\begin{aligned} h(x, \omega) \triangleq \textbf{minimize} \quad & g(w)^T y \\ \textbf{subject to;} \quad & W(\omega)y \geq r(\omega) - T(\omega)x \\ & y \in \mathcal{Y} \end{aligned} \tag{3.10}$$

where $x \in \mathbb{R}^{n_1}, c \in \mathbb{R}^{n_1}, A \in \mathbb{R}^{m_1 \times n_1}, b \in \mathbb{R}^{m_1}, y \in \mathbb{R}^{n_2}, q(\omega) \in \mathbb{R}^{n_2}, W(\omega) \in \mathbb{R}^{m_2 \times n_2}, r(\omega) \in \mathbb{R}^{m_2}$, and $T(\omega) \in \mathbb{R}^{m_2 \times n_1}$.

In (2-SP), the first-stage decisions, x, must be made before any realization of random data vector $\tilde{\omega}$. The second-stage decisions, namely the recourse action, denoted by y is taken once nature reveals a realization ω of $\tilde{\omega}$. Here, it is important to notice that the recourse function is another LP which depends on x and ω, and the aim is to minimize the sum of the cost caused by the first-stage decisions and the expected cost of the second-stage decisions computed through the known distribution of $\tilde{\omega}$.

In practice, it is intractable to compute the expectation in (3.9) for high dimensional problems. For this reason, the general approach is to approximate the true distribution of $\tilde{\omega}$ with a set of scenarios $\Omega \triangleq \{1, \ldots, \omega, \ldots, |\Omega|\}$ by assuming that sufficiently large number of scenarios becomes sufficient to represent the underlying distribution of $\tilde{\omega}$ [Kall et al., 1994]. Each scenario $\omega \in \Omega$ is associated with a probability measure λ_ω so that $\sum_{\omega \in \Omega} \lambda_\omega = 1$.

The two-stage formulation in (3.9) and (3.10) can be expressed by single deterministic mathematical program called extensive form (EF)[5] under the given distribution of $\tilde{\omega}$ [Shapiro et al., 2014]. The monolithic representation given in (3.11) is obtained by creating copies of the second-stage variables and constraints for each scenario.

$$\begin{aligned} \textbf{minimize} \quad & c^T x + \sum_{\omega \in \Omega} \lambda_\omega g(\omega)^T y_\omega \\ \textbf{subject to;} \quad & Ax \geq b \\ & T(\omega)x + W(\omega)y_\omega \geq r(\omega), \quad \forall \omega \in \Omega \\ & x \in \mathcal{X}, y \in \mathcal{Y} \end{aligned} \tag{3.11}$$

where y_ω encodes the second-stage variables under scenario $\omega \in \Omega$.

[5]Sometimes it is called deterministic equivalent formulation (DEF).

Chapter 4

PROBLEM DEFINITION AND FORMULATION

We consider a capacity allocation problem in a system with N homogeneous individuals who can be in different states throughout the planning horizon. The actions are taken at the population level to respect a capacity constraint, whereas the individuals' states evolve depending on their initial states and on the actions. We first describe the process for an individual, and then relate it with the process at the population level.

The state of an individual evolves according to an uncertain transition probability matrix. In addition, the rewards collected are also uncertain. Hence, her evolution can be represented by a Markov decision process with uncertain parameters, which corresponds to a multi-model Markov decision process (MMDP) with the corresponding tuple $(\mathcal{T}, \mathcal{S}, \mathcal{A}, \Omega, \Lambda)$ (see e.g., [Steimle et al., 2018]). Here, $\tilde{\mathcal{T}} \triangleq \{1, \ldots, T-1\}$ is the set of decision epochs and $\mathcal{T} = \tilde{\mathcal{T}} \cup \{T\}$ is the set of all periods; an individual can be in state $s \in \mathcal{S}$ in time period t, where $\mathcal{S}$ is the set of all states, $\tilde{\mathcal{S}}$ is the set of non-absorbing states, and D is the single absorbing state, so that $\mathcal{S} = \tilde{\mathcal{S}} \cup \{D\}$; $\mathcal{A} = \{0, 1\}$ is the set of action space, where $a = 1$ corresponds to special service with a limited capacity, and $a = 0$ to regular service with unlimited capacity ; Ω is the set of models or scenarios that specify different transition probability distributions and rewards; and $\Lambda = (\lambda_\omega)_{\omega \in \Omega}$, where $\lambda_\omega > 0$ is the probability that scenario ω is in effect and $\sum_{\omega \in \Omega} \lambda_\omega = 1$.

Scenario ω specifies two transition probability matrices; one among the non-absorbing states, $\mathcal{P}^\omega \triangleq \{P^\omega_{iaj}\}_{(i,j) \in \tilde{\mathcal{S}} \times \tilde{\mathcal{S}}, a \in \{0,1\}}$, and the other from non-absorbing states to the absorbing state D, $\mathcal{Q}^\omega \triangleq \{Q^\omega_{ia}\}_{i \in \tilde{\mathcal{S}}, a \in \{0,1\}}$. In words, when action $a \in \mathcal{A}$ is taken in state $i \in \tilde{\mathcal{S}}$ under scenario $\omega \in \Omega$, P^ω_{iaj} is the probability that an individual in state i moves to state $j \in \tilde{\mathcal{S}}$, whereas Q^ω_{ia} is the probability that the individual reaches the absorbing state. For each scenario $\omega \in \Omega$, $\mathcal{P}^\omega$ and $\mathcal{Q}^\omega$ satisfy

the following inequalities:

$$P^{\omega}_{iaj} \geq 0 \qquad \forall (i,j) \in \tilde{\mathcal{S}} \times \tilde{\mathcal{S}}, a \in \mathcal{A}, \omega \in \Omega \tag{4.1a}$$

$$Q^{\omega}_{ia} \geq 0 \qquad \forall i \in \tilde{\mathcal{S}}, a \in \mathcal{A}, \omega \in \Omega \tag{4.1b}$$

$$\sum_{j \in \tilde{\mathcal{S}}} P^{\omega}_{iaj} + Q^{\omega}_{ia} = 1 \qquad \forall i \in \tilde{\mathcal{S}}, a \in \mathcal{A}, \omega \in \Omega \tag{4.1c}$$

For all non-absorbing states $i \in \tilde{\mathcal{S}}$, the reward gained by action a under scenario ω is denoted by r^{ω}_{ia}, whereas the reward of being in state i at the final stage T is given by R^{ω}_{i}. We further assume that the reward of visiting the absorbing state, R^{D}, is fixed for all scenarios.

In general, the objective of MMDPs is to find a strategy over the entire planning horizon, denoted Π, that performs well with respect to different scenarios accounted in MMDP. One of the widely-used techniques to solve such problems is to employ a two-stage stochastic program, where the first stage determines the strategy Π and the second stage evaluates the performance of Π. The discrete set of scenarios called *solution sample* allows developing a mixed integer program that solves for both stages at the same time, which is also called *extensive form of MMDP*. Our solution approach also uses this technique; however it has to incorporate the additional capacity constraint at the population level, as we describe below.

The initial population consists of N individuals, where n_i of them are in state i so that $N = \sum_{i \in \tilde{\mathcal{S}}} n_i$. Consequently, the initial distribution of states is specified by $\theta_i = \frac{n_i}{N}$. The state of the system in period t is given by $\mathbf{n}(t) = (n_i(t))_{i \in \tilde{\mathcal{S}}}$, where $N - \sum_{i \in \tilde{\mathcal{S}}} n_i(t)$ individuals are in the absorbing state at time t. Individuals receive either regular or special service, where the special service needs additional resources. Hence, the problem is a capacity allocation problem, where scarce resources are rationed among the individuals in different states in each decision period to maximize the total expected reward over T periods. We represent the scarcity of the resources by a constraint that limits the expected number of individuals who receive special service by C_t at decision epoch $t \in \tilde{\mathcal{T}}$. This constraint requires tracking the evolution of all individuals through the states as well as of the actions over time.

The system manager decides on the group(s) of individuals who will receive

special service in each period, i.e., she chooses a subset of $\tilde{\mathcal{S}}$ who are entitled to use the additional resources under this constraint. Hence, the action space in all states is given by $\mathcal{D} = (0,1)^{|\tilde{\mathcal{S}}|}$. We set $\pi_i^t = 1$ if the resource is used for the individuals in state i at decision epoch t, and $\pi_i^t = 0$ otherwise, where $i \in \tilde{\mathcal{S}}$. The policy in decision period t is then given by $\pi^t = (\pi_i^t)_{i\in\tilde{\mathcal{S}}}$, and the strategy over the planning horizon by $\Pi \triangleq (\pi^t)_{t\in\tilde{\mathcal{T}}}$. This study aims to find a deterministic strategy Π that maximizes the total expected reward over T periods, while satisfying the capacity constraints in each decision period under all scenarios. Hence, the optimal strategy Π, which is independent of the scenarios, determines the groups that will receive special service in all periods $t \in \tilde{\mathcal{T}}$.

To account for the expected rewards and to respect the capacity constraints, we represent the evolution of the system in the model by defining the so-called *occupancy measures* for each scenario ω:

- $\mathcal{X}_{ia}^{\omega t}$: probability of being in state i under action a at decision epoch t
- $\mathcal{Z}^{\omega t}$: probability of visiting the absorbing state at period t
- $\mathcal{Y}_i^{\omega}$: probability of finalizing the decision process in non-absorbing state i

For a given strategy Π, the total expected reward is computed as follows:

$$\mathcal{U}(\Pi) = \underbrace{\underbrace{N\left[\sum_{\omega\in\Omega}\lambda_\omega\Big(\overbrace{\sum_{t\in\mathcal{T}\setminus\{1\}}\mathcal{Z}^{\omega t}R^D}^{\text{component 1}} + \overbrace{\sum_{t\in\tilde{\mathcal{T}}}\sum_{i\in\tilde{\mathcal{S}}}\mathcal{X}_{i\pi_i^t}^{\omega t}r_{i\pi_i^t}^{\omega}}^{\text{component 2}} + \overbrace{\sum_{i\in\tilde{\mathcal{S}}}\mathcal{Y}_i^{\omega}R_i^{\omega}}^{\text{component 3}}\Big)\right]}_{\text{expected reward of an individual under scenario } \omega\in\Omega}}_{\text{expected reward of an individual}}}_{\text{total expected reward of all individuals}}. \tag{4.2}$$

The quantity in the brackets corresponds to the expected reward of an individual under all scenarios, which gives the total expected reward of the system when multiplied by the total population N. The expected reward of an individual under scenario $\omega \in \Omega$ has three components: Component 1 computes the expected reward over all periods due to visits to the absorbing state. Component 2 accounts for the expected reward of an individual in the non-absorbing states throughout all stages $t \in \tilde{\mathcal{T}}$. Finally, component 3 determines the expected reward in the final period, T.

At this point, we point out certain characteristics of the problem: (1) Actions π_i^t can be either 0 or 1, so that either all individuals in state i use the additional resources or all are excluded from the special service. We can easily relax this condition to allow partial coverage, as demonstrated in some of our numerical experiments in Section 5. (2) Policy π^t changes with respect to time t, but not with respect to the state of the system, $\mathbf{n}(t)$. More explicitly, we determine the strategy Π for the entire planning horizon at time 0 and commit to it till the end of period T. This allows us to account for the future effects of our current actions in expectation. Note that in practice, the realized number of individuals to receive the special service may exceed the capacity. However in this paper we aim to support strategic-level decisions, whereas such violations which may be encountered and dealt with at the operational level are ignored. Solving this model on a rolling-time horizon basis by replacing the initial population with the current number of individuals in each period will account for the dynamic change of the system state and ensure that the capacity constraint is never violated.

The mixed integer programming (MIP) formulation demonstrated through (4.3)-(4.11) finds Π^* which maximizes the total expected reward of the population without violating the capacity constraints for the MMDP model. For the remainder of the paper, we call it (MIP-MMDP). It is the deterministic equivalent of the underlying two-stage stochastic integer programming formulation. In this formulation, $\Pi = \{\pi_i^t\}_{i\in\tilde{\mathcal{S}},t\in\tilde{\mathcal{T}}}$ form the first-stage decision variables that need to be made before nature reveals the uncertainty. Hence, they are independent of scenarios. The second-stage problem, on the other hand, determines the values of the occupancy measures after values of the uncertain parameters are realized by the scenarios. As mentioned above, these two stages can be combined in a single mixed-integer program when solving for MMDPs.

$$\textbf{max} \quad \mathcal{U}(\Pi) \tag{4.3}$$

$$\textbf{st;} \quad \sum_{a\in\mathcal{A}} \mathcal{X}_{ia}^{\omega,1} = \theta_i \qquad \forall i \in \tilde{\mathcal{S}}, \omega \in \Omega \tag{4.4}$$

$$\sum_{i\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ia}^{\omega,t-1} P_{iaj}^{\omega} = \sum_{a\in\mathcal{A}} \mathcal{X}_{ja}^{\omega,t} \qquad \forall t \in \tilde{\mathcal{T}} \setminus \{1\}, j \in \tilde{\mathcal{S}}, \omega \in \Omega \tag{4.5}$$

$$\sum_{i\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ia}^{\omega,T-1} P_{iaj}^{\omega} = \mathcal{Y}_{j}^{\omega} \qquad \forall j \in \tilde{\mathcal{S}}, \omega \in \Omega \tag{4.6}$$

$$\sum_{i\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ia}^{\omega,t-1} Q_{ia}^{\omega} = \Delta\mathcal{Z}^{\omega,t} \qquad \forall t \in \mathcal{T} \setminus \{1\}, \omega \in \Omega \tag{4.7}$$

$$N\sum_{i\in\tilde{\mathcal{S}}} \mathcal{X}_{i1}^{\omega,t} \leq C_t \qquad \forall t \in \tilde{\mathcal{T}}, \omega \in \Omega \tag{4.8}$$

$$\mathcal{X}_{i1}^{\omega,t} \leq \pi_i^t \qquad \forall i \in \tilde{\mathcal{S}}, t \in \tilde{\mathcal{T}}, \omega \in \Omega \tag{4.9}$$

$$\mathcal{X}_{i0}^{\omega,t} \leq (1-\pi_i^t) \qquad \forall i \in \tilde{\mathcal{S}}, t \in \tilde{\mathcal{T}}, \omega \in \Omega \tag{4.10}$$

$$\pi_i^t \in \{0,1\} \qquad \forall t \in \tilde{\mathcal{T}}, i \in \tilde{\mathcal{S}} \tag{4.11}$$

$$\mathcal{X}_{ia}^{\omega,t} \geq 0 \qquad \forall (t,i,a,\omega) \in (\tilde{\mathcal{T}}, \tilde{\mathcal{S}}, \mathcal{A}, \Omega) \tag{4.12}$$

$$\mathcal{Z}^{\omega,t} \geq 0 \qquad \forall t \in \mathcal{T} \setminus \{1\}, \omega \in \Omega \tag{4.13}$$

$$\mathcal{Y}_i^{\omega} \geq 0 \qquad \forall i \in \tilde{\mathcal{S}}, \omega \in \Omega \tag{4.14}$$

where $\Delta\mathcal{Z}^{\omega,t} = \mathcal{Z}^{\omega,t} - \mathcal{Z}^{\omega,t-1} \geq 0$, and $\mathcal{Z}^{\omega,1} = 0$, for all $\omega \in \Omega$.

In (MIP-MMDP), the objective function is given by (4.3), which maximizes the total expected reward obtained by all individuals, as defined by Equation (4.2). Equations in (4.4)-(4.7) are called *forward equations*, and they recursively compute the values of the occupancy measures. From a different viewpoint, they balance the probability flows in the model. Furthermore, (4.8) ensures that capacity constraints are satisfied for each scenario. Particularly, the expected number of individuals entitled to special service, i.e., those in states with action $a = 1$, cannot exceed the allowed capacity C_t at decision epoch $t \in \tilde{\mathcal{T}}$ for each scenario. Constraints (4.9) and (4.10) establish the link between the first stage decisions Π and the second stage variables $\mathcal{X}$. Specifically, if the strategy Π does not take action a for the group of individuals in state i at time t, then the corresponding occupancy measures has to be equal to 0. Otherwise, they can take values up to 1. Constraints (4.11) ensure

that all individuals in state i receive either regular or special service.

We now provide two valid inequalities for the (MIP-MMDP) formulation. Their proofs are presented in Appendix A. Proposition 4.1 is a simple consequence of transition probability distributions, while Proposition 4.2 is achieved by aggregating the capacity constraints over all decision epochs.

Proposition 4.1 *Hyperplanes in (4.15) are valid inequalities for (MIP-MMDP).*

$$1 = \begin{cases} \sum_{i\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ia}^{\omega,t}, & \text{if } t = 1 \\ \sum_{i\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ia}^{\omega,t} + \mathcal{Z}^{\omega,t}, & \text{if } t \in \{2,\dots,T-1\} \\ \sum_{i\in\tilde{\mathcal{S}}} \mathcal{Y}_{i}^{\omega} + \mathcal{Z}^{\omega,t}, & \text{if } t = T \end{cases} \quad \forall \omega \in \Omega \tag{4.15}$$

Proposition 4.2 $\sum_{t\in\tilde{\mathcal{T}}} \Big[\sum_{i\in\tilde{\mathcal{S}}} \mathcal{X}_{i0}^{\omega,t} + \mathcal{Z}^{\omega,t} \Big] \geq T - \frac{\sum_{t\in\tilde{\mathcal{T}}} C_t + N}{N}$ *defines a valid inequality for each scenario* $\omega \in \Omega$ *for (MIP-MMDP).*

Chapter 5

SOLUTION APPROACH

Global optima may be obtained in reasonable times for (MIP-MMDP) with small and medium-size problem instances by using commercial solvers. However, reaching the optimal solution(s) for larger instances of the problem is typically impossible. Nevertheless, the applicability of the model for real settings requires a huge number of parameters, since a large number of scenarios should be used to satisfy the convergence of the extensive-form formulation. Therefore, we propose a parallel approximate dynamic programming (PADP) algorithm that provides either optimal solutions or solutions which are very close to global optima. To that end, our approach is based on certain structural properties of the (MIP-MMDP) that allows us to represent it as a specific network problem. Hence, we first prove these structural properties, and then present our algorithm.

5.1 Structural Properties

In this section, we first introduce Proposition 5.1, observing that the problem is reduced to a Markov Reward Process (MRP) when a strategy is fixed. Then, its immediate consequences are asserted in Corollary 5.1, followed by Corollary 5.2.

Proposition 5.1 *A strategy* $\hat{\Pi}$ *yields a unique set of values of the occupancy measures for each scenario in* Ω*.*

The observation stated by Proposition 5.1 leads to our first result. The values of the occupancy measures can be computed by a strategy evaluation algorithm, such as Algorithm 4 (see Appendix C). Algorithm 4 utilizes the forward equations in (4.4)-(4.7) to find the values of the occupancy measures.

Corollary 5.1 *The feasibility of a strategy* $\hat{\Pi}$ *can be determined by Algorithm 4.*

Corollary 5.1 is implied by Proposition 5.1 as it is shown in Appendix B.2.

Corollary 5.2 *Consider two strategies* $\hat{\Pi}$ *and* $\bar{\Pi}$ *with the following property: There exists* $t^* \in \tilde{\mathcal{T}} \setminus \{T-1\}$ *such that* $\hat{\pi}_i^t = \bar{\pi}_i^t$ *for all* $i \in \tilde{\mathcal{S}}, t \leq t^*$*, and* $\hat{\pi}_{t^*+1}^i \neq \bar{\pi}_{t^*+1}^i$ *for some* $i \in \tilde{\mathcal{S}}$*. Even though* $\hat{\Pi} \neq \bar{\Pi}$*, the occupancy measures until* t^* *under* $\hat{\Pi}$ *and* $\bar{\Pi}$ *are equal to each other for each scenario.*

Corollary 5.2 stems from the fact that Algorithm 4 computes the occupancy measures in a forward direction. Thus, any change in policy π^t impacts only the occupancy measures related to decision epoch t and afterwards. This idea is the cornerstone of the proposed algorithm, as it enables us to express the problem in a network structure.

5.2 The Parallel Approximate Dynamic Programming Algorithm

5.2.1 The Network Representation

In this section, we provide the components of the network structure used by the PADP algorithm.

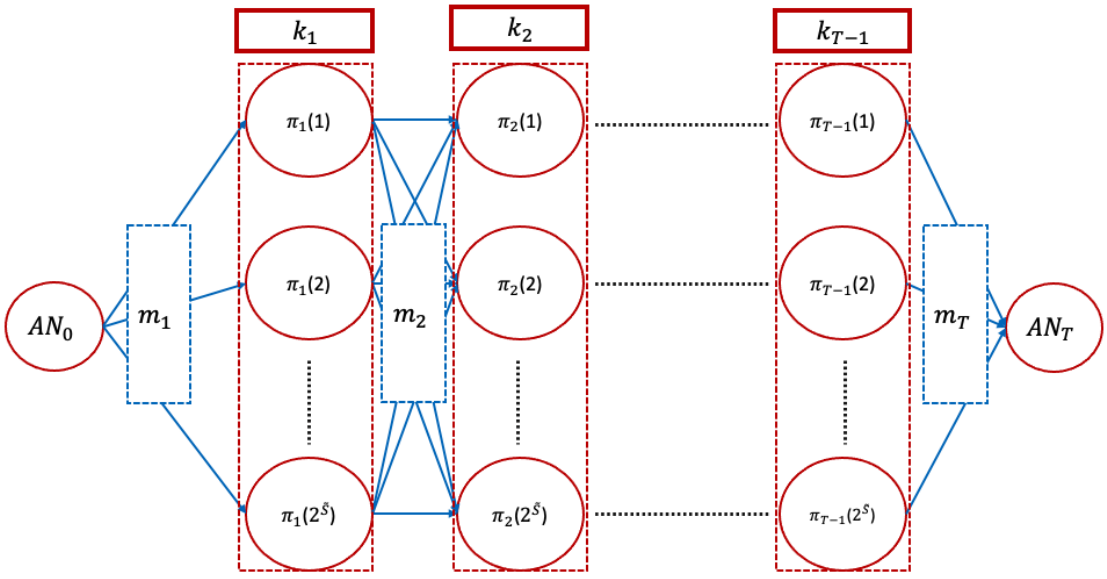
policies: There are $2^{|\tilde{\mathcal{S}}|}$ different possible policies to follow at a decision epoch $t \in \tilde{\mathcal{T}}$, considering that the action space consists of only two actions[1]. Then, $\mathcal{O} \triangleq \{1, \ldots, o, \ldots, 2^{|\tilde{\mathcal{S}}|}\}$ is the set of all policy options. Each option $o \in \mathcal{O}$ characterizes a unique policy π^t for $t \in \tilde{\mathcal{T}}$.

nodes: There is a node for each pair of policy option and decision epoch. In other words, a node is generated for each element in $\tilde{\mathcal{T}} \times \mathcal{O}$. This way of creating nodes allows us to represent all possible strategies in a single network. The node associated with decision epoch $t \in \tilde{\mathcal{T}}$ and option $o \in \mathcal{O}$ is represented by $\pi_t(o)$, which implies that the policy represented by o is accepted for t. The set of nodes related to $t \in \tilde{\mathcal{T}}$ is shown by $k_t \triangleq \{\pi_t(o)\}_{o \in \mathcal{O}}$, and $\mathcal{K} \triangleq \{k_t\}_{t \in \tilde{\mathcal{T}}} \cup \{AN_0, AN_T\}$ becomes the set of nodes, where AN_0 and AN_T are the artificial nodes taking place for $t = 0$ and $t = T$, respectively. They are called artificial since they do not involve an action.

[1]More generally, there are $|\mathcal{A}|^{|\tilde{\mathcal{S}}|}$ different possible policies.

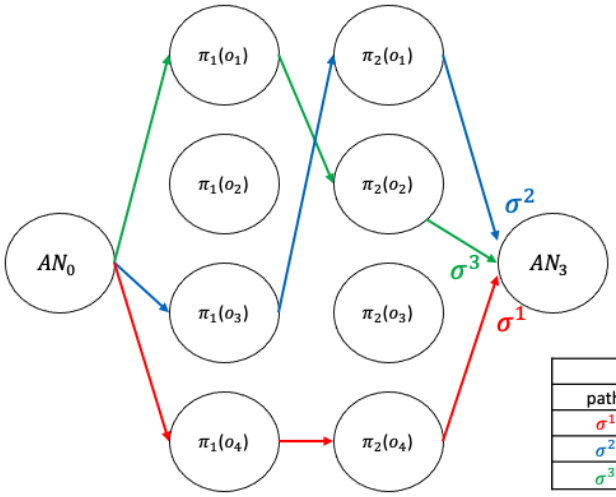
arcs: Arcs take place only between the nodes of two consecutive decision epochs. An arc from $\pi_t(\hat{o}) \in k_t$ to $\pi_{t+1}(\tilde{o}) \in k_{t+1}$ means that the strategy follows the policies in $\hat{o} \in \mathcal{O}$ and $\tilde{o} \in \mathcal{O}$ for decision epochs t and $t+1$, respectively. The set of arcs coming to the nodes in k_t is denoted by $m_t \triangleq \{(\pi_{t-1}(\hat{o}), \pi_t(\tilde{o})) : (\hat{o}, \tilde{o}) \in \mathcal{O} \times \mathcal{O}\}$, where $t \in \mathcal{T} \setminus \{1, T\}$. Then, $\mathcal{M} \triangleq \{m_t\}_{t \in \mathcal{T} \setminus \{1,T\}} \cup \{m_1, m_T\}$ becomes the set of arcs, where $m_1 \triangleq \{(AN_0, \pi_1(o)) : o \in \mathcal{O}\}$ and $m_T \triangleq \{(\pi_{T-1}(o), AN_T) : o \in \mathcal{O}\}$ are the arcs for $t = 1$ and $t = T$, respectively. Figure 5.1 provides an illustration of the network representation. It can be also seen from the figure that the state space grows exponentially, as the number of non-absorbing states and/or actions increases. Therefore, in this study, we draw our attention to the challenges posed by large numbers of scenarios and decision epochs.

Figure 5.1: Network Representation



strategy path: A strategy path, denoted by σ, is an ordered list of the nodes in $\mathcal{K}$ so that there is exactly one node from $k_t, \forall t \in \tilde{\mathcal{T}}$. Since it involves exactly one node, thus policy for each decision epoch, it defines a strategy. Figure 5.2 provides an illustrative example with 2 non-absorbing states and 3 periods. In this simple example, we have the following set of policy options: $\mathcal{O} = \{(0,0), (0,1), (1,0), (1,1)\}$,

Figure 5.2: A Simple Illustrative Example with 3 Different Strategic Paths



a) 3 different strategy paths in the network

Option	State 1	State 2
o_1	0	0
o_2	0	1
o_3	1	0
o_4	1	1

b) All policy options

	t=1		t=2	
path	state 1	state 2	state 1	state 2
σ^1	1	1	1	1
σ^2	1	0	0	0
σ^3	0	0	0	1

c) Strategies within the paths

as we have 2 non-absorbing states. Each option in $\mathcal{O}$ implies a policy as shown in Figure 5.2b. Furthermore, σ^1, σ^2, and σ^3 all define a strategic path in Figure 5.2a, equivalently a strategy for the problem. Figure 5.2c also illustrates how σ^1, σ^2, and σ^3 represent strategies. Since a path defines a strategy, it is endowed with Proposition 5.1. That is, there is a unique set of values of the occupancy measures, found by Algorithm 4, for a given strategy path. Figure 5.3 provides a simple example with $T = 3$ to show how it is achieved for a path $\sigma = (\pi^1, \pi^2)$ under scenario $\omega \in \Omega$. It shows how an arc $\ell \in m_t$ specifies the values of the occupancy measures for $t \in \mathcal{T}$. Besides, Figure 5.4 shows our parallel processing scheme that makes computations with scenario-based decomposition.

the length of an arc: The length of ℓ, denoted by η_ℓ, is then determined as the additional expected reward obtained by the occupancy measures of period t.

$$\eta_\ell = \begin{cases} \sum_{\omega\in\Omega} \lambda_\omega \big(\mathcal{Z}^{\omega t} R^D + \sum_{i\in\tilde{\mathcal{S}}} \sum_{a\in\{0,1\}} \mathcal{X}_{ia}^{\omega t} r_{ia}^{\omega} \big), & \text{if } t \in \tilde{\mathcal{T}} \\ \sum_{\omega\in\Omega} \lambda_\omega \big(\mathcal{Z}^{\omega t} R^D + \sum_{i\in\tilde{\mathcal{S}}} \mathcal{Y}_i^{\omega} r_i^{\omega} \big), & \text{if } t = T \end{cases} \tag{5.1}$$

Figure 5.3: An Example Computation of Occupancy Measures

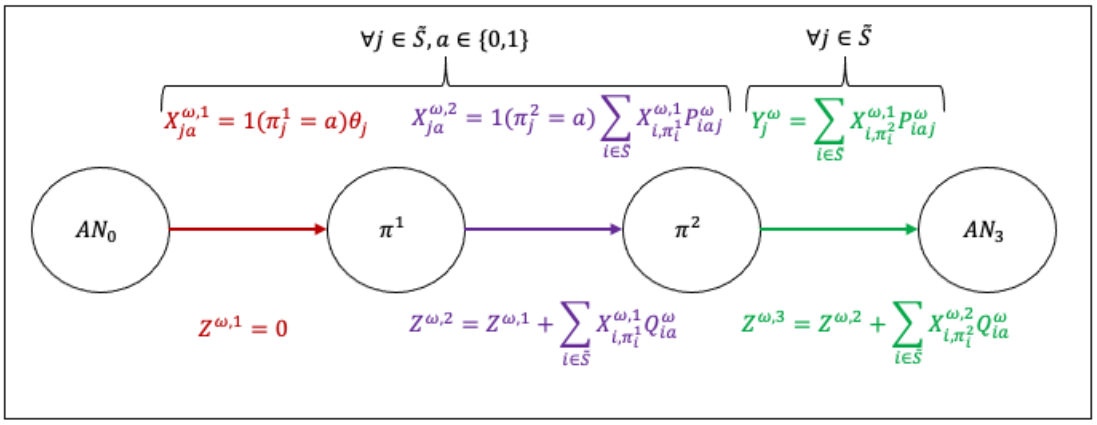
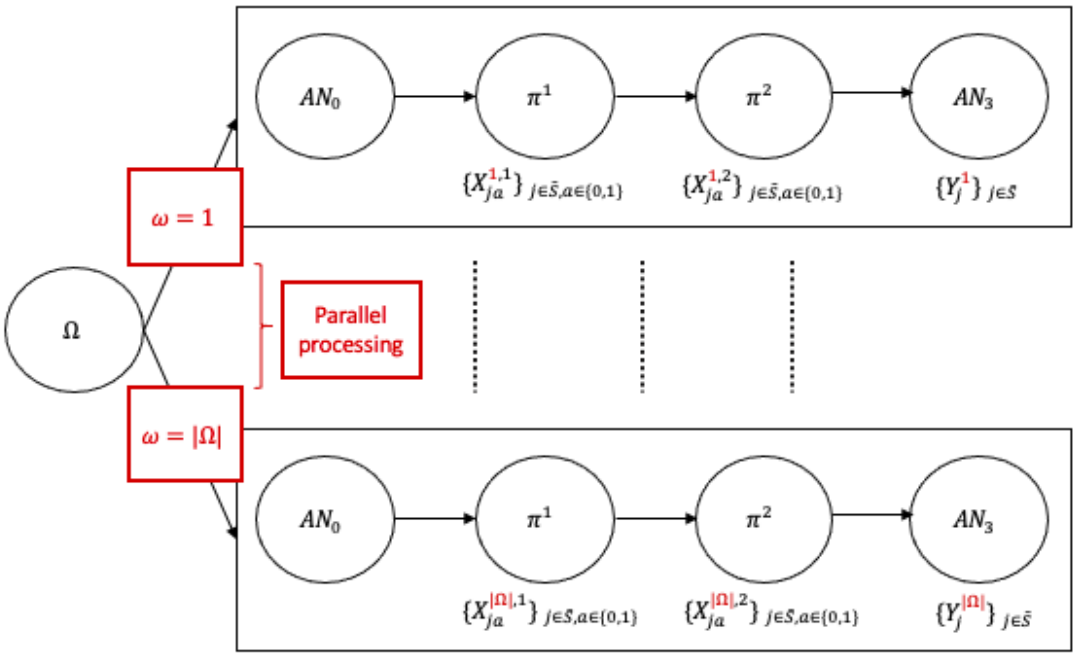


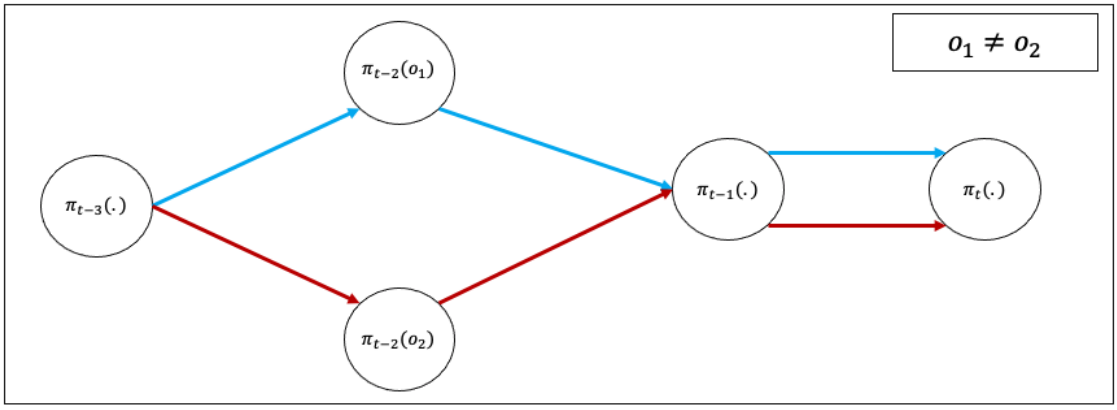
Figure 5.4: An Illustration of the Parallel Processing Scheme



Remark 5.1 *For an arc $\ell \in m_t$, η_ℓ is not static but changes depending on the nodes associated with $t = 1, \ldots, t-1$ in the path.*

Figure 5.5 is provided in order to motivate the remark above. Consider two different paths in the figure: the blue path and the red path. Notice that the only difference between them is the nodes they visit at period $t-2$. As a result, $\eta_{(\pi_{t-1}(.),\pi_t(i))}$ may be different within these two paths since they may specify different values of occupancy measures for period $t-1$, thus period t. This simple example motivates the remark, and further demonstrates the additional computational complexity we encounter.

Figure 5.5: A Sample Case to Motivate Remark 5.1



5.2.2 The Algorithm

In Section 5.2.1, the length of an arc is defined in a way that the strategy Π^* which provides the maximum total expected reward without violating the capacity constraints is equivalent to σ^*. Then, our ultimate goal is to find σ^*, i.e., the feasible longest strategy path. In this regard, we develop a parallel approximate dynamic programming (PADP) algorithm. The path found by the PADP algorithm does not guarantee the optimality since it only searches the feasible nodes in k_{t-1} for the node $\pi_t(o)$. However, the algorithm verifies the feasibility of a node, as shown

in Algorithm 2, based only on the best found path to it, where another path that makes it feasible could be more promising, considering the subsequent iterations. Hence, the path obtained as a result of the PADP algorithm is an approximation of σ^*.

We now introduce the notation that helps us to understand the fundamentals of the proposed PADP algorithm. These are as follows:

- $\hat{\mathcal{X}}_{ia}^{\omega t}(\pi_t(o))$: The value of $\mathcal{X}_{ia}^{\omega t}$ when the policy within $\pi_t(o)$ is accepted for period t, and the best-found feasible longest path to $\pi_t(o)$ is followed.

- $\hat{\mathcal{Z}}^{\omega t}(\pi_t(o))$: The value of $\mathcal{Z}^{\omega t}$ when the policy within $\pi_t(o)$ is accepted for period t, and the best-found feasible longest path to $\pi_t(o)$ is followed.

- $\hat{f}_{\pi_t(o)}$: The total length obtained from the first decision epoch to period t when the policy within $\pi_t(o)$ is accepted for period t, and the best-found feasible longest path to $\pi_t(o)$ is followed.

Following remarks are essential in order to comprehend the notation, and the logic behind the algorithm.

- The values of the occupancy measures of periods $t' = 1, \ldots, t-1$ are known by Proposition 5.1 for the best-found feasible longest path to $\pi_t(o)$. They can be obtained by calling Algorithm 4 and terminating it at the end of period $t-1$.

- We do not need to know the values of the occupancy measures of periods $t' = 1, \ldots, t-2$ by Corollary 5.2 in order to compute $\hat{\mathcal{X}}_{ia}^{\omega t}(\pi_t(o))$. It is enough to have $\{\hat{\mathcal{X}}_{ia}^{\omega,t-1}(\pi_{t-1}(o'))\}_{i\in\tilde{\mathcal{S}},a\in\{0,1\},\omega\in\Omega,o'\in\mathcal{O}}$.

- For the node $\pi_1(o) \in k_1$, $\{\hat{\mathcal{X}}_{ia}^{\omega,1}(\pi_1(o))\}, \{\hat{\mathcal{Z}}^{\omega,1}(\pi_1(o))\}$[2], and $\hat{f}_{\pi_1(o)}$ are found as indicated by Algorithm 1. It first finds the occupancy measure values for

[2] $\{\hat{\mathcal{X}}_{ia}^{\omega t}(\pi_t(o))\}_{i\in\tilde{\mathcal{S}},a\in\{0,1\},\omega\in\Omega}$ is shortly represented by $\{\hat{\mathcal{X}}_{ia}^{\omega t}(\pi_t(o))\}$, whereas $\{\hat{\mathcal{Z}}^{\omega t}(\pi_t(o))\}$ denotes $\{\hat{\mathcal{Z}}^{\omega t}(\pi_t(o))\}_{\omega\in\Omega}$.

$t = 1$, and then checks the feasibility based on these values. If the feasibility is violated, it sets $\hat{f}_{\pi_1(o)}$ to $-\infty$, and occupancy measure values are determined as null[3]. Otherwise, the computed values are assigned as shown in the pseudocode. This assignment does not involve an optimization since there is only one arc in m_1

- For the node $\pi_t(o) \in k_t$, $\{\hat{\mathcal{X}}_{ia}^{\omega t}(\pi_t(o))\}, \{\hat{\mathcal{Z}}^{\omega t}(\pi_t(o))\}$, and $\hat{f}_{\pi_t(o)}$ are found as shown in Algorithm 2. It basically searches all the nodes in k_{t-1}. It computes the occupancy measure values of period t for each node in k_{t-1}. Then, it selects the node which provides the maximum $\hat{f}_{\pi_t(o)}$ value among the ones which does not violate the feasibility. If there is no node satisfying the feasibility, then $\hat{f}_{\pi_t(o)}$ is set to $-\infty$.

- For the node $\pi_{T-1}(o) \in k_{T-1}$, $\{\hat{\mathcal{X}}_{ia}^{\omega,T-1}(\pi_{T-1}(o))\}, \{\hat{\mathcal{Z}}^{\omega,T-1}(\pi_{T-1}(o))\}$, and $\hat{f}_{\pi_{T-1}(o)}$ are found by following Algorithm 2 with a replacement which changes line 9 of it as *if* $f_{\tilde{\pi}} + \eta_{(\tilde{\pi},\pi_t(o))} + \eta_{(\pi_t(o),AN_T)} > f_{\pi_t(o)}$ *then*. Consequently, $f_{\pi_t(o)}$ in line 10 is updated to $f_{\tilde{\pi}} + \eta_{(\tilde{\pi},\pi_t(o))} + \eta_{(\pi_t(o),AN_T)}$.

- Our algorithm uses parallel processing techniques while decomposing the scenarios in Ω whenever a computation or a feasibility check is required.

Our PADP algorithm starts with $t = 1$, and finds $\{\hat{\mathcal{X}}_{ia}^{\omega,1}(\pi_1(o))\}, \{\hat{\mathcal{Z}}^{\omega,1}(\pi_1(o))\}$, and $\hat{f}_{\pi_1(o)}$ for each node $\pi_1(o)$ in k_1 by using Algorithm 1. Then, iterations are performed in a forward direction, i.e., $t = 2, \ldots, T-1$. For period t, it finds $\{\hat{\mathcal{X}}_{ia}^{\omega t}(\pi_t(o))\}$, $\{\hat{\mathcal{Z}}^{\omega t}(\pi_t(o))\}$, and $\hat{f}_{\pi_t(o)}$ by using Algorithm 2. Then, the objective function value found by the algorithm is determined as the maximum $\hat{f}$ value[4] tackled by the nodes within k_{T-1}, i.e., $f^{PADP} \triangleq \max_{i \in k_{T-1}} \hat{f}_i$.

[3]Basically, we leave them empty since an infeasible node is not taken into account by the upcoming periods.

[4]Here, we use the term $\hat{f}$ with abuse of notation in order to make the explanation convenient.

Algorithm 1 Evaluation and Feasibility for the First Period

Require: $\pi_1(o)$

$\pi^1 \leftarrow \pi_1(o)$

$\mathcal{X}_{ia}^{\omega,1} \leftarrow 1(\pi_i^1 = a)\theta_i,\ \forall i \in \tilde{\mathcal{S}}, a \in \{0,1\}, \omega \in \Omega$

$\mathcal{Z}^{\omega,1} = 0,\ \forall \omega \in \Omega$

if $\exists \omega \in \Omega : N \sum_{i \in \tilde{\mathcal{S}}} \mathcal{X}_{i1}^{\omega,1} > C_1$ **then**

  $f_{\pi_1(o)} \leftarrow -\infty$

  $\{\hat{\mathcal{X}}_{ia}^{\omega,1}(\pi_1(o))\} \leftarrow \emptyset$

  $\{\hat{\mathcal{Z}}^{\omega,1}(\pi_1(o))\} \leftarrow \emptyset$

else

  $f_{\pi_1(o)} \leftarrow \eta_{(AN_0,\pi_1(o))}$

  $\hat{\mathcal{X}}_{ia}^{\omega,1}(\pi_1(o)) \leftarrow \mathcal{X}_{ia}^{\omega,1},\ \forall i \in \tilde{\mathcal{S}}, a \in \{0,1\}, \omega \in \Omega$

  $\hat{\mathcal{Z}}^{\omega,1}(\pi_1(o)) \leftarrow \mathcal{Z}^{\omega,1},\ \forall \omega \in \Omega$

end if

Ensure: $\{\hat{\mathcal{X}}_{ia}^{\omega,1}(\pi_1(o))\}, \{\hat{\mathcal{Z}}^{\omega,1}(\pi_1(o))\}, \hat{f}_{\pi_1(o)}$

Algorithm 2 Evaluation and Feasibility for the Periods $t \in \{2, \ldots, T-2\}$

Require: $\pi_t(o)$

$\pi^t \leftarrow \pi_t(o)$

$f_{\pi_t(o)} \leftarrow -\infty$

$\{\hat{\mathcal{X}}_{ia}^{\omega t}(\pi_t(o))\} \leftarrow \emptyset$

$\{\hat{\mathcal{Z}}^{\omega t}(\pi_t(o))\} \leftarrow \emptyset$

for $\tilde{\pi} \in k_{t-1}$ **do**

  $\mathcal{X}_{ia}^{\omega t} \leftarrow 1(\pi_i^t = i) \sum_{h \in \tilde{\mathcal{S}}} \sum_{a \in \{0,1\}} \hat{\mathcal{X}}_{ha}^{\omega,t-1}(\tilde{\pi}) P_{hai}^{\omega},\ \forall i \in \tilde{\mathcal{S}}, a \in \{0,1\}, \omega \in \Omega$

  $\mathcal{Z}^{\omega t} \leftarrow \hat{\mathcal{Z}}^{\omega,t-1}(\tilde{\pi}) + \sum_{h \in \tilde{\mathcal{S}}} \sum_{a \in \{0,1\}} \hat{\mathcal{X}}_{ha}^{\omega,t-1}(\tilde{\pi}) Q_{ha}^{\omega},\ \forall \omega \in \Omega$

  if $\forall \omega \in \Omega : n \sum_{i \in \tilde{\mathcal{S}}} \mathcal{X}_{i1}^{\omega t} \leq C_t$ **then**

    if $f_{\tilde{\pi}} + \eta_{(\tilde{\pi},\pi_t(o))} > f_{\pi_t(o)}$ **then**

      $f_{\pi_t(o)} \leftarrow f_{\tilde{\pi}} + \eta_{(\tilde{\pi},\pi_t(o))}$

      $\hat{\mathcal{X}}_{ia}^{\omega t}(\pi_t(o)) \leftarrow \mathcal{X}_{ia}^{\omega t},\ \forall i \in \tilde{\mathcal{S}}, a \in \{0,1\}, \omega \in \Omega$

      $\hat{\mathcal{Z}}^{\omega t}(\pi_t(o)) \leftarrow \mathcal{Z}^{\omega t},\ \forall \omega \in \Omega$

    end if

  end if

end for

Ensure: $\{\hat{\mathcal{X}}_{ia}^{\omega t}(\pi_t(o))\}, \{\hat{\mathcal{Z}}^{\omega t}(\pi_t(o))\}, \hat{f}_{\pi_t(o)}$

Chapter 6

CASE STUDY: CHRONIC CARE DELIVERY PROBLEM

U.S. National Center for Health Statistics defines a chronic condition as a disease which lasts three months or more, and they further claim that there is not any vaccine, medication, or cure that can stop it, or it does not disappear spontaneously. Therefore, it is assumed that recovery is not possible when a patient has a chronic condition. CDC demonstrates stroke, cancer, and diabetes as major chronic diseases. [Bernell and Howard, 2016] further include hypertension, pulmonary conditions, and mental illnesses to the list of chronic diseases.

According to the records of National Health Insurance in the United States, 65% of the population have multimorbidity, i.e., they have more than one chronic diseases. From societal perspective, a huge economic burden, approximately 80% of Medicare spending, is caused by the patients with 4 or more chronic diseases [Wolff et al., 2002]. In addition, multimorbidity is more common in disadvantaged groups, which causes more social inequality [Britt et al., 2008]. Consequently, we are witnessing a changing paradigm called *chronic care management* where polycronic patients' physical, mental and social needs are managed in one center. As in all systems, this approach has trade-offs since this system requires both fixed and variable costs. In this regard, the effective use of this resource is important. Therefore, the question of which patients should be targeted for chronic care delivery under capacity constraints becomes important. Here we will assume a hypothetical chronic care management system with a constrained resource. In this setting, the patient classes that will be eligible for the service for each period in the planning horizon should be determined at the system design stage.

The most distinctive and challenging part of designing a system for polychronic patients is that the underlying mechanism for the progress of health conditions may significantly vary from disease to disease. In the case of multimorbidity, the

number of diseases, their severity, and which combinations of diseases occur can vary dramatically for each patient. For instance, a patient may have type-2 diabetes and breast cancer, whereas another patient may have hypertension, asthma, and a cardiovascular disease. In this perspective, it is necessary to deal with transition probability and reward uncertainties, considering the range of chronic conditions, their breakdowns, and various levels of severity for the diseases.

Patient targeting problem for chronic care delivery model is a suitable case study to test our model with its capacity constraints and inherently appeared parameter uncertainty within the system. In this regard, the associated MDP model is introduced in Section 6.1.

6.1 MDP Model

In this section, we conceptualize the dynamics of a patient in regard of chronic care delivery. In the rest of this section, we explain the components of the underlying MDP model.

6.1.1 States

A state involves two components. The first one is *health status* which can be, without loss of generality, any measure indicating the physical well-being of a patient. In this study, we discretized this component into 3 categories: *simple, moderate*, and *complex*. The severity of conditions is getting worse from simple to complex.

The second component of a state includes the *engagement* of a patient due to the statistics revealing that about 50% of premature deaths are caused by behaviors that could be changed [Council et al., 2015]. Moreover, [Meng et al., 1999] states that chronic diseases have very close link with lifestyle habits such as excessive alcohol use, poor nutrition, lack of physical activity, and tobacco use. In this regard, the term engagement corresponds to all behavioral aspects required to maintain patients' own physical well-being. It is discretized by two alternatives: *low* and *high*. The consciousness of patients increases from low to high.

In summary, we have six states coming from the pairwise combination of the

physical and behavioral components: Low-Simple (0), Low-Moderate (1), Low-Complex (2), High-Simple (3), High-Moderate (4), High-Complex (5). In addition to these states, we also include the state *death (6)* as an absorbing state.

6.1.2 Actions

For each patient with a health status and a behavioral characteristic, an agent chooses either *regular care* or *special care*. Thus, the action space consists of two elements: 0, for regular care and 1, for special care. The action space is not binding for those reaching to the absorbing state.

6.1.3 Transition Probabilities and Immediate Rewards

Transition probabilities are subject to uncertainty due to highly combinatorial nature of multimorbidity. Hence, it is not possible to have data representing all possible transitions to obtain significant statistical inferences. Even if we have such data, the estimations are prone to large statistical errors because of limited or missing data for each possible transition. Nevertheless, it is possible to approximate the probability distributions by using expert opinions. We demonstrate one approach to employ these opinions. In this respect, we first generate hierarchical rules, i.e., relations among transition probabilities based on the prior beliefs (see Appendix D). These are the rules that can be generated by any healthcare analyst who is familiar with the system. The base model, the so-called *nominal model*, assumes that the parameters take the values that a Monte Carlo (MC) algorithm depicted in Algorithm 3 converges to. The points where Monte Carlo converges may be seen as a center of the polyhedral set in which the transition probabilities lie for our nominal model. Then, all scenarios are generated based on this center by adding uniform random noises around it. A similar approach is carried out for the immediate rewards with their own rules. Detailed description of scenario generation is explained in Section 7.1. Our MC algorithm starts with empty set of patients. A patient is randomly generated following the hierarchical rules in each iteration until reaching to the desired number of iterations. Then, the parameters of the nominal model are

Algorithm 3 Monte Carlo Approach for Transition Probability Estimation

Require: number of iterations, expert opinions

nIteration $\leftarrow 1$
$\bar{\mathcal{P}} \leftarrow \emptyset$
$\bar{\mathcal{Q}} \leftarrow \emptyset$
for nIteration $\leq$ number of iterations **do**
    $\mathcal{P}'' \leftarrow$A random realization which satisfies the required rules
    $\mathcal{Q}'' \leftarrow$A random realization which satisfies the required rules
    $\bar{\mathcal{P}} \leftarrow \bar{\mathcal{P}} \cup \{\mathcal{P}''\}$
    $\bar{\mathcal{Q}} \leftarrow \bar{\mathcal{Q}} \cup \{\mathcal{Q}''\}$
    nIteration $\leftarrow$ nIteration $+ 1$
end for
$\hat{\mathcal{P}} \leftarrow \text{mean}(\bar{\mathcal{P}})$
$\hat{\mathcal{Q}} \leftarrow \text{mean}(\bar{\mathcal{Q}})$

Ensure: $\hat{\mathcal{P}}, \hat{\mathcal{Q}}$

determined as the mean of the corresponding parameters of the generated patients.

Chapter 7

COMPUTATIONAL EXPERIMENTS

In this section, we conduct computational experiments to obtain an understanding of (i) the contribution of the proposed model; (ii) how effective our solution approach is. All experiments are conducted on a computer with an Intel® Core ™ i7-6500U 2.39 GHz processor and 64 GB of RAM, with the Windows 10 operating system. In this context, Section 7.1 explains the way problem instances are generated. Then, Section 7.2 is devoted to investigate the impact of valid inequalities, whereas Section 7.3 examines the computational aspects of the proposed algorithm. Section 7.5 analyzes the value of perfect information after Section 7.4 studies the value of stochastic solution. Finally, Section 7.6 provides managerial insights for potential practitioners with a special focus on the price of fairness, the value of flexibility, and capacity management.

7.1 Instance Generation

For chronic care delivery problem, it is not possible to have a sufficiently large dataset that includes longitudinal data capturing all possible transitions in sufficient numbers. This is due to the large number of chronic conditions each of which has a variety of types. Furthermore, patients may have different numbers of chronic conditions of varying severity. In this respect, it becomes difficult to have reliable time-series data of patients as in specialized services such as diabetes, HIV, and cancer treatments. Nevertheless, we have some prior beliefs about the relations among the uncertain parameters. These are not complicated medical consequences, but simple justifiable facts allowing us to come up with a polyhedral uncertainty set for the nominal model. Point estimations for the so-called nominal model, are generated through Monte Carlo approach depicted in Algorithm 3. Then, different

scenarios are generated by adding uniform noise to the nominal model. We refer to [Hörmann et al., 2013] and [Zhang et al., 2017b] for different scenario generation methodologies.

Let $\bar{\omega}_x$ denotes the nominal value of uncertain parameter x, whereas ω_x denotes the corresponding parameter value in scenario $\omega \in \Omega$[1]. Given nominal model $\bar{\omega}$, scenario ω is generated as follows:

$$\omega \triangleq \left\{ \omega_x : \omega_x \sim \mathit{UNIF}\Big((1-\epsilon)\bar{\omega}_x, (1+\epsilon)\bar{\omega}_x \Big) \right\}, \quad \epsilon \in (0,1) \tag{7.1}$$

Then, elements of transition probability matrices are normalized so that they satisfy the conditions in (4.1a)-(4.1c).

A problem instance is uniquely characterized by 4 parameters. (i) $|\Omega|$, the number of scenarios in the solution sample; (ii) T, the number of decision epochs; (iii) $c = C/n$, the proportion of the population that can utilize the resource at a particular decision epoch; (iv) ϵ, the maximum allowed variation from the nominal model. Without loss of generality, we assume that the capacities allocated per decision epoch are equal. Then, a problem instance with the given parameters is denoted by $\mathcal{I}(|\Omega|, T, c, \epsilon)$. Furthermore, prior probabilities are assumed to be equal. That is, the probability for a new patient to start the decision process in a state is equal for each non-absorbing state in $\tilde{\mathcal{S}}$, i.e., $\mathbb{P}\{s_1 = i\} = \frac{1}{|\tilde{\mathcal{S}}|}, \forall i \in \tilde{\mathcal{S}}$. Moreover, the probability of each scenario is equal to each other, i.e., $\lambda_\omega = \frac{1}{|\Omega|}, \forall \omega \in \Omega$. Lastly, the reward obtained after visiting the absorbing state, R^D, is set to zero.

7.2 Impact of Valid Inequalities

There are three components affecting the computational complexity of the problem: the number of decision epochs, scenarios, and states. In the case study, i.e., chronic care delivery problem, the size of the state space is fixed so that we can embrace the problems with a large number of scenarios and/or stages. Therefore, we generated problem instances for a varying number of scenarios and decision epochs. To this end, we created 35 problem instances, each with the values of $c = 0.4$ and

[1]Nominal value corresponds to the expected value of a parameter.

$\epsilon = 0.25$. While generating them, we use the set of different numbers of scenarios $\{5, 10, 25, 50, 100, 250, 500\}$ and the set of various numbers for decision epochs $\{5, 10, 20, 30, 40\}$.

In order to prepare an experiment analyzing the impact of the valid inequalities, we first solve (MIP-MMDP) by using the well-known commercial solver CPLEX 12.10 for all problem instances. Then, each valid inequality is separately included in the formulation, and solved again by the solver. All computational runs are imposed a time limit of 4 hours. In this time horizon, the solver could find the optimal solutions only for 21 instances among the 35 instances we generated. Table 7.1 provides required solution times in seconds and relative percentage gap values for each instance and setting. We compare the required solution times for the 21 instances where the optimal solutions could be found, whereas the relative percentage gap values are compared for the rest of the instances.

When we focus on the instances with zero gap values[2], we realize that valid inequality 1 provides the best solution time in 33.3% of the instances, whereas this statistic becomes 15.15% for valid inequality 2. In the case where we look at the remaining instances, valid inequality 2 provides the best gap values for almost half of the instances, whereas valid inequality 1 provides it only for 2 problem instances.

7.3 Computational Performance of the Parallel Approximate Dynamic Programming Algorithm

Table 7.1 shows that commercial solvers are not capable of solving problem instances with the desired number of scenarios for convergence. Therefore, an efficient algorithm providing solutions of high quality in a short span of time is essential for the applicability and the value of the developed model. In this context, a high quality solution refers to the ones that are optimal or very close to it and found within an acceptable amount of time.

In Section 5, a parallel approximate dynamic programming algorithm is proposed as an alternative to solvers to provide an efficient algorithm required for the model.

[2]Valid inequality 1 and 2 refer to Propositions 4.1 and 4.2, respectively.

In this section, we test its capabilities in terms of time and optimality. In accordance with this purpose, we first use the 21 problem instances for which the solver could provide the optimal solutions in Section 7.2. The proposed algorithm, PADP, is run for these problem instances, and the corresponding solution times in seconds and the percentage gap values are presented through Table 7.2. PADP finds the optimal solutions for 42.857% of the problem instances. For the remainder of them, the maximum deviation from the optimal solution is only 0.24% which is quite good for an algorithm trying to find an approximate solution. When we focus on the aggregate level performance metrics, we see an outstanding approximation so that the mean percentage gap value is only 0.073%. Another remarkable fact about the superiority of the algorithm is that it achieves these approximations in considerably short span of times. It finds the solution about 1000 times faster than the solver.

In Table 7.3, we compare the capabilities of CPLEX 12.10 and PADP for the instances the solver could not find the optimal solution in 4 hours in Section 7.2. Recall that the solver loses its power for the instances with a large number of scenarios. The column *improvement(%)* shows the relative percentage improvement provided by PADP. A positive value in the column shows that the algorithm provides a better solution than the solver with 4 hours time limitation. On the contrary, the negative value indicates that the solution found by the solver in 4 hours provides a better objective function than the algorithm. Results suggest that the algorithm produces better results than the solver when the number of scenarios exceeds 250, i.e., for more realistic instances. The amount of improvement becomes more significant as the number of scenarios increases.

7.4 Value of Stochastic Solution

If agents do not take parameter uncertainty into account then the actions will become sub-optimal. With this in mind, in this section, we investigate the value of incorporating stochastic solution. In other words, we estimate the potential loss in objective function value, if we do not consider the transition probability and reward uncertainties. There is certainly a strong relation between the value of stochastic

solution and the way we generate the problem instances. The main assumption we make during the generation of data is the maximum variation from the nominal model. Thus, we conduct sensitivity analysis for different values of ϵ. Particularly, we choose three values of ϵ for the experiments: $0.10, 0.25$, and 0.50.

We also postulate uniform distribution when we consider the deviations from the nominal model, and by force of uniform distribution, it is not possible to observe any realization for the parameters outside of the sphere determined by ϵ. However, observing lots of outliers which lie in the exterior region of the sphere is very likely, considering the large number of combinations within the nature of the problem. Thus, the problem includes more randomness than our problem instances. Nevertheless, there are two advantages of this approach. Firstly, the value of stochastic solution asserted for these problem instances intuitively provides a lower bound for the actual value since we do not allow outliers. Therefore, if we find a considerable value of stochastic solution, then we can assert that the actual value is expected to be higher than it, which makes our contribution realistic. The second advantage is that this way of generating data materializes itself without loss of generality. That is, results associated with other experiments about the complexity of the model and the efficiency of our algorithm are not affected by the assumptions made during the generation of the problem instances.

Expected value of stochastic solution, EVSS(%), under solution sample $\Omega = \{1, \ldots, \omega, \ldots, |\Omega|\}$ is computed as follows:

$$\text{EVSS}(\%) \triangleq \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \frac{f^{\textit{MIP-MMDP}} - f^{\omega}(\Omega)}{f^{\omega}(\Omega)} \times 100 \tag{7.2}$$

where $f^{\textit{MIP-MMDP}}$ is the total expected reward of the optimal strategy found for Ω, i.e., the objective function value of (MIP-MMDP) for Ω, and $f^{\omega}(\Omega)$ is the total expected reward of following the optimal strategy for scenario ω, where the realization of nature is still explained by all scenarios in Ω. It is clear that $f^{\omega}(\Omega)$ yields a lower bound for $f^{\textit{MIP-MMDP}}$, i.e., $f^{\omega}(\Omega) \leq f^{\textit{MIP-MMDP}} \forall \omega \in \Omega$, due to the sub-optimal behavior of the strategy governed by a single scenario. Thus, $\frac{f^{\textit{MIP-MMDP}} - f^{\omega}(\Omega)}{f^{\omega}(\Omega)}$ shows the relative value of stochastic solution compared to the deterministic solution found for

scenario ω. Then, $EVSS(\%)$ gives the expected value of stochastic solution. Appendix E shows how $f^{\omega}(\Omega)$ is evaluated for scenario $\omega \in \Omega$.

Table 7.4 shows the expected value of stochastic solution for different values of ϵ and T when the number of scenarios is 200. Results suggest that expected value is increasing as the value of ϵ increases. Another important observation is that it also increases with the growing number of decision epochs which implies that our approach matters more for long-term problems.

7.5 Value of Perfect Information

Value of perfect information (VPI) is defined as the maximum cost a policy-maker is willing to pay in order to obtain the perfect information about the uncertain parameters. Since we are dealing with a maximization problem, the objective function of single scenario deterministic problem, denoted by f^{ω}, becomes an upper bound for $f^{\text{MIP-MMDP}}$, i.e., $f^{\text{MIP-MMDP}} \leq f^{\omega}, \forall \omega \in \Omega$. Then, the expected VPI (EVPI) under the set of scenarios Ω is computed as follows:

$$\text{EVPI} = \frac{\sum_{\omega \in \Omega} f^{\omega}}{|\Omega|} - f^{\text{MIP-MMDP}} \tag{7.3}$$

Then, the percentage expected VPI, denoted by EVPI(%) is determined as $\frac{\text{EVPI}}{f^{\text{MIP-MMDP}}} \times 100$. Table 7.5 illustrates the expected value of perfect information for different values of ϵ and T when the solution sample consists of 200 scenarios. We find EVPI(%) for pairwise combinations of $\epsilon = 0.1, 0.25, 0.5$ and $T = 5, 10, 20, 30$. The average EVPI(%) value is 7.88%. They also indicate that EVPI(%) increases with the increasing number of ϵ and T.

7.6 Managerial Insights

In this section, we investigate managerial aspects of applying (MIP-MMDP). In this context, we first examine the price of fairness, that is, the loss in total expected reward caused by ethical issues. Then, we focus on the value of having flexibility in operations. Lastly, we analyze the impact of different capacity levels. It is important to notice that each part, except the one measuring the effect of different capacity

levels, requires certain modifications in the model. For these parts, we change the required components of the model, and then solve them with the solver. That is, we do not modify our algorithm for the new cases, which is beyond the scope of our study. Thus, the analyzes are limited with the problem instances solved by the solver in 4 hours, i.e., small-medium problem instances. However, we think that even this can be useful to understand the potential values and drawbacks of the model for practical issues.

7.6.1 Price of Fairness

(MIP-MMDP) is limited to deterministic policies, even a randomized policy can promise a better objective function value. Nevertheless, we require deterministic policies due to the fairness concerns raised in healthcare operations [Cevik et al., 2018, Steimle et al., 2018]. This is also the case for other domains such as humanitarian relief operations [Meraklı and Küçükyavuz, 2019]. The term fairness relates to the ethical issues which oblige policy-makers to follow the same policy for all patients in the same state. Hence, (MIP-MMDP) with randomized policies, i.e., without (4.9) - (4.11), becomes a relaxation of the original formulation. Then, *price of fairness (%)*, determined as $\frac{f^r - f^{\text{MIP-MMDP}}}{f^r} \times 100$, corresponds to the loss in the objective function value because of the constraints which eliminate randomized policies, where f^r denotes the optimal objective function value of the model allowing randomized policies.

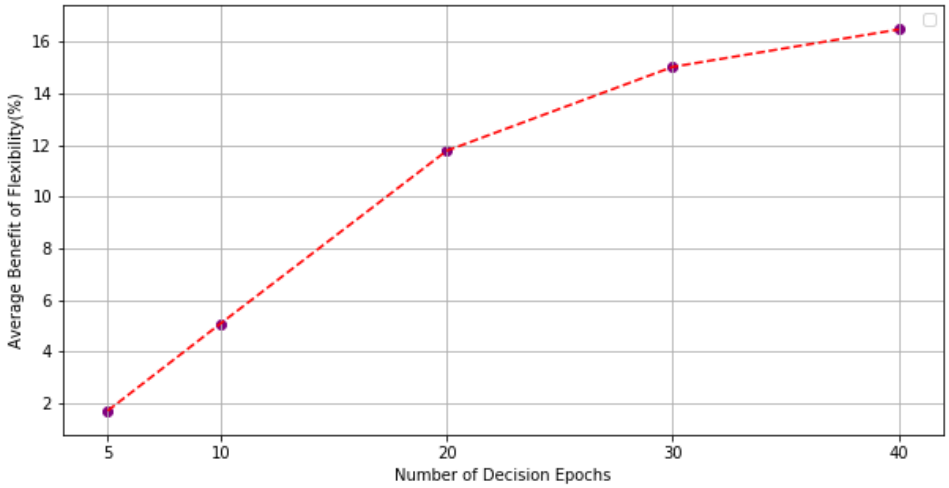
Price of fairness (%) for each small-medium problem instance is demonstrated in Table 7.6. The average loss caused by the ethical concerns is 5.75%. It stems from the fact that a portion of the capacity at each decision epoch becomes idle due to the deterministic policy constraints. Allowing randomized policies enables the model to fulfill the capacity, even if only a subset of patients in a state utilize the resource. Results also show that as the number of scenarios increases, price of fairness tends to increase.

7.6.2 Value of Flexibility

We have so far assumed flexibility in actions, where in each period the actions are independent of the actions taken in other periods. That is, one state may receive special care in stage t and not accepted in stages $t' \in \tilde{\mathcal{T}} \setminus \{t\}$. In practice, policy makers may not have this flexibility; a state is either entitled to the resource for all decision epochs from the beginning or not at all. Therefore, here we consider the case in which once action is 1 for a state in a particular decision epoch, then it must also be 1 for other decision epochs. Mathematically, this corresponds to adding the following constraint to (MIP-MMDP): $\pi_i^t - \pi_i^{t-1} = 0, \forall i \in \tilde{\mathcal{S}}, t \in \tilde{\mathcal{T}} \setminus \{1\}$. It is clear that this narrows the feasible region so that the previous objective function value becomes an upper bound for the new model, i.e., $\tilde{f} \leq f^{\text{MIP-MMDP}}$ where $\tilde{f}$ is the optimal objective function value under the same policy approach. In this study, flexibility implies that policy-makers can change their course of actions over time. Since (MIP-MMDP) allows it, *value of flexibility (%)* is determined as $\frac{f^{\text{MIP-MMDP}} - \tilde{f}}{\tilde{f}} \times 100$.

Value of flexibility (%) is computed for each small-medium problem instance, and the values are represented in Table 7.7. It shows that 7.98% additional value can be obtained by (MIP-MMDP) instead of using the same policy approach. In other words, embracing flexibility yields almost 7.98% additional total expected reward. Furthermore, average values for $T = 5, 10, 20, 30$, and 40 are computed and summarized in Figure 7.1. It demonstrates that as the number of decision epochs increases, the value of embracing flexibility also increases dramatically since the suboptimal behavior is retained for longer periods. This also shows the relative importance of our model for long-term planning problems compared to the ones for short-terms.

Figure 7.1: Average Value of Flexibility for $T = 5, 10, 20, 30, 40$



7.6.3 Capacity Management

In this section, we analyze the additional total expected reward that can be obtained by adding one more capacity at each period. To this end, we find total expected rewards for each instance from $\bigcup_{c=0.2}^{0.8} \mathcal{I}(200, t, c, 0.25)$ for $t \in \{10, 20, 30, 40\}$. In words, for $t \in \{10, 20, 30, 40\}$, we start with stage capacity 200 and gradually increase it until 800 for a fixed solution sample which consists of 200 scenarios. Figure 7.2 and Figure 7.3 illustrate the results of the experiment for the cases with deterministic policies and randomized policies, respectively. The main difference between deterministic and randomized policies become more clear when we compare the figures. Figure 7.3 clearly shows that the graph tends to perform concave characteristics. That is, the marginal utility of adding one unit of stage capacity is decreasing, as the value of stage capacity increases. Unlike the graph in Figure 7.3, we do not observe smooth increases within the graph in Figure 7.2. That is, increasing the capacity for deterministic policies does not necessarily improve the current conditions since the model still may not find an available group to fulfill the remaining capacity. Thus, system managers embracing deterministic policies must take this situation into account when they need to decide how much to increase the capacity.

Figure 7.2: Total Expected Rewards under Different Stage Capacities and Decision Epochs for the Model Restricted with Deterministic Policies

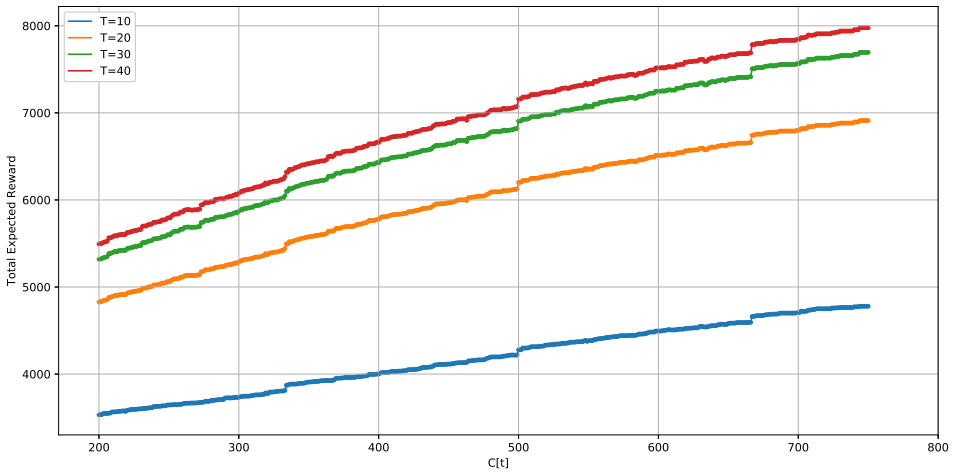


Figure 7.3: Total Expected Rewards under Different Stage Capacities and Decision Epochs for the Model with Randomized Policies

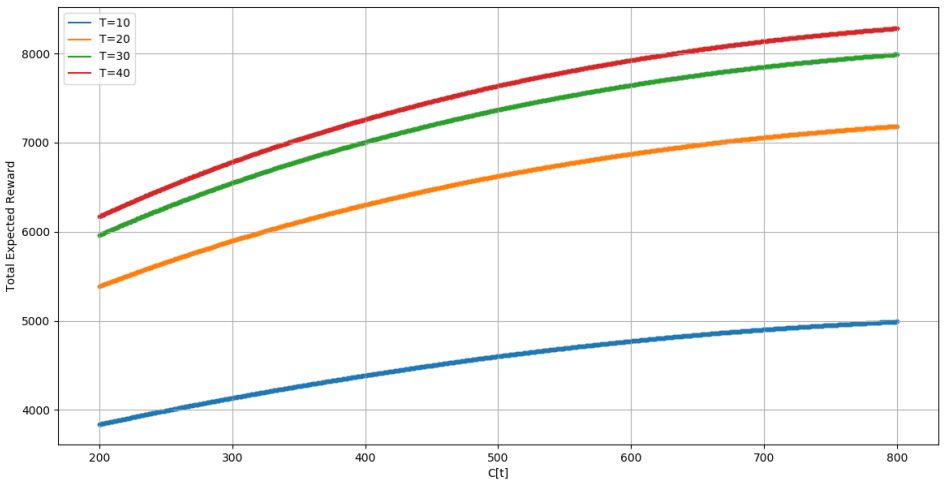


Table 7.1: Impact of Valid Inequalities in 35 Problem Instances

		(MIP-MMDP) Model		Valid Inequality 1		Valid Inequality 2	
nScenario	nStage	Time (sec)	Gap (%)	Time (sec)	Gap (%)	Time (sec)	Gap (%)
5	5	0.2	0	0.247	0	0.965	0
	10	2.371	0	2.36	0	2.538	0
	20	8.241	0	8.95	0	7.27	0
	30	10.31	0	9.99	0	10.818	0
	40	9.63	0	9.788	0	10.232	0
10	5	0.225	0	0.221	0	0.21	0
	10	6.8	0	6.845	0	8.441	0
	20	53.848	0	80.211	0	116.207	0
	30	101.423	0	162.75	0	132.885	0
	40	138.449	0	157.163	0	170.913	0
25	5	0.511	0	0.532	0	0.47	0
	10	256.734	0	229.86	0	251.009	0
	20	1891.068	0	1515.171	0	1368.60	0
	30	1425.692	0	1193.41	0	1783.103	0
	40	3015.04	0	2818.66	0	3052.309	0
50	5	9.664	0	9.486	0	9.846	0
	10	3643.571	0	1176.563	0	2355.285	0
	20	14401.456	0.6	14401.213	0.9	14401.537	0.8
	30	14401.767	0.5	14439.982	0.01	14402.806	0.6
	40	14470.37	1.1	14429.526	1.4	14430.794	1.2
100	5	106.115	0	196.078	0	96.072	0
	10	9669.398	0	13778.265	0	14425.731	0.1
	20	14404.549	2.0	14406.66	1.9	14409.823	2.3
	30	14421.458	2.1	14406.278	2.1	14432.652	1.9
	40	14406.975	1.8	14434.273	2.1	14411.393	1.9
250	5	193.304	0	339.143	0	272.681	0
	10	14405.094	1.9	14407.678	1.8	14407.396	1.7
	20	14406.464	5.4	14407.621	5.4	14409.304	4.8
	30	14404.798	6.4	14405.341	6.0	14405.035	5.3
	40	14406.866	5.7	14407.235	6.1	14403.145	5.5
500	5	742.221	0	861.041	0	1103.164	0
	10	14407.241	4.5	14402.248	5.4	14406.788	5.3
	20	14406.308	7.0	14407.461	7.1	14410.242	6.9
	30	14406.978	7.7	14405.62	7.6	14406.41	7.7
	40	14405.674	8.9	14409.774	8.9	14407.159	9.2

Table 7.2: Computational Comparison of PADP and CPLEX 12.10 for Small-Medium Problem Instances

Instance	Time (MIP)	Time (PADP)	GAP(%)
$\mathcal{I}(5, 5, 0.4, 0.25)$	0.2	0.084	0
$\mathcal{I}(5, 10, 0.4, 0.25)$	2.371	0.204	0
$\mathcal{I}(5, 20, 0.4, 0.25)$	8.241	0.683	0.13
$\mathcal{I}(5, 30, 0.4, 0.25)$	10.31	1.127	0.16
$\mathcal{I}(5, 40, 0.4, 0.25)$	9.63	1.567	0.16
$\mathcal{I}(10, 5, 0.4, 0.25)$	0.225	0.115	0
$\mathcal{I}(10, 10, 0.4, 0.25)$	6.8	0.36	0
$\mathcal{I}(10, 20, 0.4, 0.25)$	53.848	1.041	0.02
$\mathcal{I}(10, 30, 0.4, 0.25)$	101.423	1.742	0.05
$\mathcal{I}(10, 40, 0.4, 0.25)$	138.449	2.437	0.06
$\mathcal{I}(25, 5, 0.4, 0.25)$	0.511	0.149	0
$\mathcal{I}(25, 10, 0.4, 0.25)$	256.734	0.466	0
$\mathcal{I}(25, 20, 0.4, 0.25)$	1891.068	1.332	0.14
$\mathcal{I}(25, 30, 0.4, 0.25)$	1425.692	2.248	0.2
$\mathcal{I}(25, 40, 0.4, 0.25)$	3015.04	3.176	0.22
$\mathcal{I}(50, 5, 0.4, 0.25)$	9.664	0.17	0.02
$\mathcal{I}(50, 10, 0.4, 0.25)$	3643.571	0.555	0.24
$\mathcal{I}(100, 5, 0.4, 0.25)$	106.115	0.245	0
$\mathcal{I}(10, 10, 0.4, 0.25)$	9669.398	0.802	0.13
$\mathcal{I}(250, 5, 0.4, 0.25)$	193.304	0.304	0
$\mathcal{I}(500, 5, 0.4, 0.25)$	742.221	0.434	0
min	0.2	0.084	0
max	9669.398	3.176	0.24
mean	1013.563	0.916	0.073

Table 7.3: Computational Comparison of PADP and CPLEX 12.10 for Large Problem Instances

Instance	CPLEX Time(sec)	CPLEX Gap(%)	PADP Time(sec)	Improvement(%)
$\mathcal{I}(50, 20, 0.4, 0.25)$	14401.456	0.6	1.628	-0.33
$\mathcal{I}(50, 30, 0.4, 0.25)$	14401.767	0.5	2.699	-0.36
$\mathcal{I}(50, 40, 0.4, 0.25)$	14470.37	1.1	4.025	-0.37
$\mathcal{I}(100, 20, 0.4, 0.25)$	14404.549	2	2.251	0.05
$\mathcal{I}(100, 30, 0.4, 0.25)$	14421.458	2.1	3.635	0.01
$\mathcal{I}(100, 40, 0.4, 0.25)$	14406.975	1.8	4.9	-0.14
$\mathcal{I}(250, 10, 0.4, 0.25)$	14405.094	1.9	0.98	0.09
$\mathcal{I}(250, 20, 0.4, 0.25)$	14406.464	5.4	3.005	0.66
$\mathcal{I}(250, 30, 0.4, 0.25)$	14404.798	6.4	5.658	0.87
$\mathcal{I}(250, 40, 0.4, 0.25)$	14406.866	5.7	7.724	0.13
$\mathcal{I}(500, 10, 0.4, 0.25)$	14407.241	4.5	1.502	0.62
$\mathcal{I}(500, 20, 0.4, 0.25)$	14406.308	7	5.662	1.13
$\mathcal{I}(500, 30, 0.4, 0.25)$	14406.978	7.6	8.255	1.48
$\mathcal{I}(500, 40, 0.4, 0.25)$	14405.674	8.9	10.203	2.66

Table 7.4: Expected Value of Stochastic Solution for Different ϵ and T Values with 200 Scenarios

ϵ	nStage	Value of stochastic solution(%)
0.1	5	2.97
	10	4.52
	20	4.47
	30	4.53
0.25	5	4.03
	10	5.70
	20	6.83
	30	7.31
0.5	5	3.95
	10	5.89
	20	7.80
	30	9.30

Table 7.5: Expected Value of Perfect Information for Different ϵ and T Values with 200 Scenarios

ϵ	nStage	Value of perfect information(%)
0.1	5	1.72
	10	2.14
	20	1.84
	30	1.76
0.25	5	5.62
	10	7.09
	20	5.81
	30	5.17
0.5	5	12.44
	10	16.36
	20	17.66
	30	16.92

Table 7.6: Price of Fairness for Small-Medium Problem Instances

nScenario	nStage	Price of Fairness (%)	nScenario	nStage	Price of Fairness (%)
5	5	4.10	25	5	5.96
5	10	4.66	25	10	6.98
5	20	4.15	25	20	6.45
5	30	3.89	25	30	6.31
5	40	3.80	25	40	6.27
10	5	4.19	50	5	6.81
10	10	5.02	50	10	7.99
10	20	4.89	100	5	7.13
10	30	4.82	100	10	8.39
10	40	4.80	250	5	7.08
			500	5	7.07

Table 7.7: Value of Flexibility for Small-Medium Problem Instances

nScenario	nStage	value of Flexibility (%)	nScenario	nStage	value of Flexibility (%)
5	5	2.53	25	5	2.12
5	10	5.38	25	10	5.34
5	20	10.24	25	20	13.33
5	30	12.39	25	30	17.61
5	40	13.25	25	40	19.59
10	5	2.16	50	5	1.55
10	10	5.88	50	10	4.73
10	20	11.81	100	5	1.23
10	30	15.11	100	10	4.15
10	40	16.63	250	5	1.21
			500	5	1.25

Chapter 8

CONCLUSION

In this study, we introduce a capacity-constrained multi-model Markov decision process model, and illustrate its use in medical resource allocation problems. Multi-model approach implies the relaxation of the traditional assumption which assumes perfect knowledge of transition probabilities and rewards. Capacity constraints pose limitations on the selection of particular actions. To the best of our knowledge, the proposed model is the first constrained MMDP model.

We develop a mixed integer programming (MIP) formulation for the model in order to find the strategy which maximizes the total expected reward without violating capacity constraints. It corresponds to an extensive-form formulation for the underlying two-stage stochastic integer program. The scenarios used in the extensive-form formulation basically corresponds to different MDP models. The MIP formulation becomes easily intractable as the size of the problem increases. In this regard, we propose a parallel approximate dynamic programming algorithm leveraging the problem structure. We also propose two valid inequalities in the hope that they can strengthen the formulation.

We use the chronic care delivery problem as an example to test our model and algorithm. We adopt a Monte Carlo approach which uses prior beliefs about system dynamics to generate problem instances. Extensive computational experiments are then performed to test the computational aspects of the proposed algorithm as well as the value of our model. We show that the algorithm works pretty fast so that it generates solutions in seconds even for very large problem instances. It also offers very high quality solutions in terms of optimality. However, we have a limitation that our algorithm embraces the cases with a large number of scenarios and stages, excluding the last level of complexity which is the state space.

There are several potential approaches to complement our work. First, a new

algorithm can be proposed that also works fast for large state space problems. In this context, it may be promising to address constrained reinforcement learning models. Secondly, some medical problems such as breast cancer make it impossible to have perfect information of the health status of patient [Cevik et al., 2018]. To deal with these kind of challenges, our model can be extended for POMDP models.

BIBLIOGRAPHY

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

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

[Altman, 1999] Altman, E. (1999). *Constrained Markov decision processes*, volume 7. CRC Press.

[Ayvaci et al., 2012] Ayvaci, M. U., Alagoz, O., and Burnside, E. S. (2012). The effect of budgetary restrictions on breast cancer diagnostic decisions. *Manufacturing & Service Operations Management*, 14(4):600–617.

[Ben-Tal et al., 2009] Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust optimization*, volume 28. Princeton University Press.

[Ben-Tal and Nemirovski, 2002] Ben-Tal, A. and Nemirovski, A. (2002). Robust optimization–methodology and applications. *Mathematical programming*, 92(3):453–480.

[Bernell and Howard, 2016] Bernell, S. and Howard, S. W. (2016). Use your words carefully: what is a chronic disease? *Frontiers in public health*, 4:159.

[Bertsimas et al., 2011] Bertsimas, D., Brown, D. B., and Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM review*, 53(3):464–501.

[Bertsimas and Thiele, 2006a] Bertsimas, D. and Thiele, A. (2006a). Robust and data-driven optimization: modern decision making under uncertainty. In *Models, methods, and applications for innovative decision making*, pages 95–122. INFORMS.

[Bertsimas and Thiele, 2006b] Bertsimas, D. and Thiele, A. (2006b). A robust optimization approach to inventory theory. *Operations research*, 54(1):150–168.

[Boucherie and Van Dijk, 2017] Boucherie, R. J. and Van Dijk, N. M. (2017). *Markov decision processes in practice*, volume 248. Springer.

[Britt et al., 2008] Britt, H. C., Harrison, C. M., Miller, G. C., and Knox, S. A. (2008). Prevalence and patterns of multimorbidity in australia. *Medical Journal of Australia*, 189(2):72–77.

[Cevik et al., 2018] Cevik, M., Ayer, T., Alagoz, O., and Sprague, B. L. (2018). Analysis of mammography screening policies under resource constraints. *Production and Operations Management*, 27(5):949–972.

[Council et al., 2015] Council, N. R., on Population, C., et al. (2015). *Measuring the Risks and Causes of Premature Death: Summary of Workshops*. National Academies Press.

[Denton, 2018] Denton, B. T. (2018). Optimization of sequential decision making for chronic diseases: From data to decisions. In *Recent Advances in Optimization and Modeling of Contemporary Problems*, pages 316–348. INFORMS.

[Denton et al., 2009] Denton, B. T., Kurt, M., Shah, N. D., Bryant, S. C., and Smith, S. A. (2009). Optimizing the start time of statin therapy for patients with diabetes. *Medical Decision Making*, 29(3):351–367.

[Deo et al., 2013] Deo, S., Iravani, S., Jiang, T., Smilowitz, K., and Samuelson, S. (2013). Improving health outcomes through better capacity allocation in a community-based chronic care model. *Operations Research*, 61(6):1277–1294.

[Givan et al., 2000] Givan, R., Leach, S., and Dean, T. (2000). Bounded-parameter markov decision processes. *Artificial Intelligence*, 122(1-2):71–109.

[Goh et al., 2018] Goh, J., Bayati, M., Zenios, S. A., Singh, S., and Moore, D. (2018). Data uncertainty in markov chains: Application to cost-effectiveness analyses of medical innovations. *Operations Research*, 66(3):697–715.

[Gorissen et al., 2015] Gorissen, B. L., Yanıkoğlu, İ., and den Hertog, D. (2015). A practical guide to robust optimization. *Omega*, 53:124–137.

[Goyal and Grand-Clement, 2018] Goyal, V. and Grand-Clement, J. (2018). Robust markov decision process: Beyond rectangularity. *arXiv preprint arXiv:1811.00215*.

[Grand-Clement et al., 2020] Grand-Clement, J., Chan, C. W., Goyal, V., and Escobar, G. (2020). Robust policies for proactive icu transfers. *arXiv preprint arXiv:2002.06247*.

[Hörmann et al., 2013] Hörmann, W., Leydold, J., and Derflinger, G. (2013). *Automatic nonuniform random variate generation*. Springer Science & Business Media.

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

[Kall et al., 1994] Kall, P., Wallace, S. W., and Kall, P. (1994). *Stochastic programming*. Springer.

[Küçükyavuz and Sen, 2017] Küçükyavuz, S. and Sen, S. (2017). An introduction to two-stage stochastic mixed-integer programming. In *Leading Developments from INFORMS Communities*, pages 1–27. INFORMS.

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

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

[Mannor et al., 2007] Mannor, S., Simester, D., Sun, P., and Tsitsiklis, J. N. (2007). Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322.

[Meng et al., 1999] Meng, L., Maskarinec, G., Lee, J., and Kolonel, L. N. (1999). Lifestyle factors and chronic diseases: application of a composite risk index. *Preventive medicine*, 29(4):296–304.

[Meraklı and Küçükyavuz, 2019] Meraklı, M. and Küçükyavuz, S. (2019). Risk aversion to parameter uncertainty in markov decision processes with an application to slow-onset disaster relief. *IISE Transactions*, pages 1–21.

[Nilim and El Ghaoui, 2004] Nilim, A. and El Ghaoui, L. (2004). Robustness in markov decision problems with uncertain transition matrices. In *Advances in neural information processing systems*, pages 839–846.

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

[Ross, 2014a] Ross, S. M. (2014a). *Introduction to probability models*. Academic press.

[Ross, 2014b] Ross, S. M. (2014b). *Introduction to stochastic dynamic programming*. Academic press.

[Satia and Lave Jr, 1973] Satia, J. K. and Lave Jr, R. E. (1973). Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740.

[Seifert et al., 2016] Seifert, R. W., Tancrez, J.-S., and Biçer, I. (2016). Dynamic product portfolio management with life cycle considerations. *International Journal of Production Economics*, 171:71–83.

[Shapiro et al., 2014] Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2014). *Lectures on stochastic programming: modeling and theory*. SIAM.

[Shechter et al., 2008] Shechter, S. M., Bailey, M. D., Schaefer, A. J., and Roberts, M. S. (2008). The optimal time to initiate hiv therapy under ordered health states. *Operations Research*, 56(1):20–33.

[Sinha and Ghate, 2016] Sinha, S. and Ghate, A. (2016). Policy iteration for robust nonstationary markov decision processes. *Optimization Letters*, 10(8):1613–1628.

[Steimle et al., 2018] Steimle, L. N., Kaufman, D. L., and Denton, B. T. (2018). Multi-model markov decision processes. *Optimization Online URL http://www. optimization-online. org/DB_FILE/2018/01/6434. pdf*.

[Sutton et al., 1998] Sutton, R. S., Barto, A. G., et al. (1998). *Introduction to reinforcement learning*, volume 135. MIT press Cambridge.

[Tewari and Bartlett, 2007] Tewari, A. and Bartlett, P. L. (2007). Bounded parameter markov decision processes with average reward criterion. In *International Conference on Computational Learning Theory*, pages 263–277. Springer.

[Varian, 1996] Varian, H. R. (1996). Intermediate microeconomics, w. w.

[White III and Eldeib, 1994] White III, C. C. and Eldeib, H. K. (1994). Markov decision processes with imprecise transition probabilities. *Operations Research*, 42(4):739–749.

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

[Wolff et al., 2002] Wolff, J. L., Starfield, B., and Anderson, G. (2002). Prevalence, expenditures, and complications of multiple chronic conditions in the elderly. *Archives of internal medicine*, 162(20):2269–2276.

[Zhang et al., 2017a] Zhang, Y., Steimle, L., and Denton, B. (2017a). Robust markov decision processes for medical treatment decisions. *Optimization online*.

[Zhang et al., 2017b] Zhang, Y., Wu, H., Denton, B. T., Wilson, J. R., and Lobo, J. M. (2017b). Probabilistic sensitivity analysis on markov decision processes with uncertain transition probabilities: an application in evaluating treatment decisions for type 2 diabetes.

Appendix A

PROOF OF VALID INEQUALITIES

A.1 Proposition 4.1

- for $t = 1$:

  We are given that $\sum_{a\in\mathcal{A}} \mathcal{X}_{ia}^{\omega,1} = \theta_i$ for each $i \in \tilde{\mathcal{S}}$ and $\omega \in \Omega$ by (4.4). Then,

$$\begin{aligned}\sum_{i\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ia}^{\omega t} &= \sum_{i\in\tilde{\mathcal{S}}} \theta_i, \qquad \forall\omega\in\Omega \\ &= 1, \qquad \forall\omega\in\Omega \quad (\textit{Since } \theta \textit{ is a valid PMF over } \tilde{\mathcal{S}})\end{aligned}$$

- for $2 \le t \le T-1$:

  We are given that $\sum_{a\in\mathcal{A}} \mathcal{X}_{ia}^{\omega t} = \sum_{h\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ha}^{\omega,t-1} P_{hai}^{\omega}$ for each $i \in \tilde{\mathcal{S}}$ and $\omega \in \Omega$ by (4.5). Moreover, we can re-write $Z^{\omega t}$ as shown in (4.7). Then, following equations are satisfied for each scenario $\omega \in \Omega$:

$$\begin{aligned}\sum_{i\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ia}^{\omega t} + \mathcal{Z}^{\omega t} &= \sum_{i\in\tilde{\mathcal{S}}}\sum_{h\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ha}^{\omega,t-1} P_{hai}^{\omega} + \sum_{h\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ha}^{\omega,t-1} Q_{ha}^{\omega} + \mathcal{Z}^{\omega,t-1} \\ &= \sum_{h\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ha}^{\omega,t-1} \Big[\underbrace{\sum_{i\in\tilde{\mathcal{S}}} P_{hai}^{\omega}}_{1,\forall(h,a)\in(\tilde{\mathcal{S}},\mathcal{A})} + Q_{ha}^{\omega} \Big] + \mathcal{Z}^{\omega,t-1} \\ &= \sum_{h\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ha}^{\omega,t-1} + \mathcal{Z}^{\omega,t-1}\end{aligned}$$

  Now, we should prove that $\sum_{h\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ha}^{\omega,t-1} + \mathcal{Z}^{\omega,t-1}$ must be equal to 1. To that end, we utilize *proof by induction* as follows:

  - *Base Case:* for $t = 2$, $\sum_{h\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ha}^{\omega,t-1} + \mathcal{Z}^{\omega,t-1} \stackrel{?}{=} 1, \forall\omega\in\Omega$

For $t = 2$, the lhs is reduced to $\sum_{h\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ha}^{\omega,1} + \mathcal{Z}^{\omega,1}$. We have already proved that $\sum_{h\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ha}^{\omega,1} = 1, \forall\omega \in \Omega$. We also know that $\mathcal{Z}^{\omega,1} = 0, \forall\omega \in \Omega$. Thus, base case is satisfied.

– *Induction Step:* $\sum_{h\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ha}^{\omega,t-1} + \mathcal{Z}^{\omega,t-1} \stackrel{?}{=} 1$ given that $\sum_{h\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ha}^{\omega,t-2} + \mathcal{Z}_{\omega,t-2} = 1, \forall\omega \in \Omega$

We are given that $\sum_{a\in\mathcal{A}} \mathcal{X}_{ha}^{\omega,t-1} = \sum_{i\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ia}^{\omega,t-2} P_{iah}^{\omega}, \forall\omega \in \Omega$ for each $h \in \tilde{\mathcal{S}}$ by (4.5). Moreover, we can re-write $Z^{\omega,t-1}$ as expressed in (4.7). Then following equations are satisfied for each scenario $\omega \in \Omega$,

$$\begin{aligned}
\sum_{h\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ha}^{\omega,t-1} + \mathcal{Z}^{\omega,t-1} &= \sum_{i\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ia}^{\omega,t-2}\Big[\underbrace{\sum_{h\in\tilde{\mathcal{S}}} P_{iah}^{\omega} + Q_{ia}^{\omega}}_{1,\forall(i,a)\in(\tilde{\mathcal{S}},\mathcal{A})}\Big] + \mathcal{Z}^{\omega,t-2} \\
&= \sum_{i\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ia}^{\omega,t-2} + \mathcal{Z}^{\omega,t-2} \\
&= 1
\end{aligned}$$

- for $t = T$:

We are given that $\mathcal{Y}_i^{\omega} = \sum_{h\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ha}^{\omega,T-1} P_{hai}^{\omega}$ for each $i \in \tilde{\mathcal{S}}$ and $\omega \in \Omega$. Moreover, we can re-write $Z^{\omega,T-1}$ as depicted in (4.7). Then, following equations are satisfied for each scenario $\omega \in \Omega$:

$$\begin{aligned}
\sum_{i\in\tilde{\mathcal{S}}} \mathcal{Y}_i^{\omega} + \mathcal{Z}^{\omega,T} &= \sum_{i\in\tilde{\mathcal{S}}}\sum_{h\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ha}^{\omega,T-1} P_{hai}^{\omega} + \sum_{h\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ha}^{\omega,T-1} Q_{ha}^{\omega} + \mathcal{Z}^{\omega,T-1} \\
&= \sum_{h\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ha}^{\omega,T-1}\Big[\underbrace{\sum_{h\in\tilde{\mathcal{S}}} P_{hai}^{\omega} + Q_{ha}^{\omega}}_{1,\forall(h,a)\in(\tilde{\mathcal{S}},\mathcal{A})}\Big] + \mathcal{Z}^{\omega,T-1} \\
&= \sum_{h\in\tilde{\mathcal{S}}}\sum_{a\in\mathcal{A}} \mathcal{X}_{ha}^{\omega,T-1} + \mathcal{Z}^{\omega,T-1} \\
&= 1
\end{aligned}$$

A.2 Proposition 4.2

$$\sum_{i\in\tilde{\mathcal{S}}} \mathcal{X}_{i,1}^{\omega t} \leq \frac{C_t}{N} \qquad \forall t \in \tilde{\mathcal{T}}, \omega \in \Omega$$

$$\sum_{t\in\tilde{\mathcal{T}}}\sum_{i\in\tilde{\mathcal{S}}} \mathcal{X}_{i,1}^{\omega t} \leq \frac{\sum\limits_{t\in\tilde{\mathcal{T}}} C_t}{N} \qquad \forall \omega \in \Omega$$

$$\sum_{t\in\tilde{\mathcal{T}}} \Bigg[\underbrace{\sum_{i\in\tilde{\mathcal{S}}} \Big[\mathcal{X}_{i,1}^{\omega t} + \mathcal{X}_{i,0}^{\omega t}\Big] + \mathcal{Z}^{\omega t}}_{\text{1 by (4.1)}} \Bigg] \leq \frac{\sum\limits_{t\in\tilde{\mathcal{T}}} C_t}{N} + \sum_{t\in\tilde{\mathcal{T}}} \Big[\sum_{i\in\tilde{\mathcal{S}}} \mathcal{X}_{i,0}^{\omega t} + \mathcal{Z}^{\omega t} \Big] \qquad \forall \omega \in \Omega$$

$$T - 1 \leq \frac{\sum\limits_{t\in\tilde{\mathcal{T}}} C_t}{N} + \sum_{t\in\tilde{\mathcal{T}}} \Big[\sum_{i\in\tilde{\mathcal{S}}} \mathcal{X}_{i,0}^{\omega t} + \mathcal{Z}^{\omega t} \Big] \qquad \forall \omega \in \Omega$$

$$T - \frac{\sum\limits_{t\in\tilde{\mathcal{T}}} C_t + N}{N} \leq \sum_{t\in\tilde{\mathcal{T}}}\sum_{i\in\tilde{\mathcal{S}}} \mathcal{X}_{i,0}^{\omega t} + \mathcal{Z}^{\omega t} \qquad \forall \omega \in \Omega$$

Appendix B

PROOF OF STRUCTURAL PROPERTIES

B.1 Proposition 5.1

- for $t = 1$:

  It is known that either $\mathcal{X}_{i,0}^{\omega,1} = 0$ or $\mathcal{X}_{i,1}^{\omega,1} = 0$ by (4.9) and (4.10) for each $i \in \tilde{\mathcal{S}}, \omega \in \Omega$. Since we are given the value of π_i^1, we know which one of them equals zero, i.e., $\mathcal{X}_{i,0}^{\omega,1} = 0$, if $\pi_i^1 = 1$; otherwise, $\mathcal{X}_{i,1}^{\omega,1} = 0$. For each $i \in \tilde{\mathcal{S}}, \omega \in \Omega$, $\mathcal{X}_{i,0}^{\omega,1} + \mathcal{X}_{i,1}^{\omega,1} = \theta_i$ by (4.4). Since we know the one which equals to zero in the left-hand-side, and the value of θ_i, there is only one value for $\mathcal{X}_{i,0}^{\omega,1}$ and $\mathcal{X}_{i,1}^{\omega,1}$, and can be found by the equality ensured by (4.4).

- for $2 \leq t \leq T-1$:

  Equations for $2 \leq t \leq T-1$ are governed by (4.5) and (4.7). Let's first focus on (4.5). We know the exact value of the first term in the equation since we know the values of the occupancy measures for $t = 1$ and equations are updated recursively. Thus, we know the value for the following summation: $\mathcal{X}_{j,0}^{\omega t} + \mathcal{X}_{j,1}^{\omega t}$. Again, we know which one of $\mathcal{X}_{j,0}^{\omega t}$ and $\mathcal{X}_{j,1}^{\omega t}$ equals zero by π_j^t. Hence, we know the value of the other occupancy measure. Similar reasoning can be made for the equality in (4.7) by using the fact that $\mathcal{Z}^{\omega,1} = 0, \forall \omega \in \Omega$.

- for $t = T$:

  Related equations are governed by (4.6) where left-hand-side specifies a unique value for each $\omega \in \Omega, j \in \tilde{\mathcal{S}}$. Since there is just one term for each of them in the right-hand-side of the equation (R_i^ω), again we can find a unique value for each of them.

B.2 Corollary 5.1

By Proposition 5.1, the values of the corresponding occupancy measures for a given strategy $\hat{\Pi}$ are known. It is also known that the resulting values definitely satisfy all the constraints in (MIP-MMDP) except (4.8). Hence, the feasibility check is reduced to verify whether (4.8) is satisfied or not. Since the left-hand-side of the inequalities are now known, we can easily verify that whether (4.8) is satisfied or not, thus the feasibility of $\hat{\Pi}$.

Appendix C

PSEUDOCODE OF THE STRATEGY EVALUATION ALGORITHM

Algorithm 4 Strategy Evaluation Algorithm

Require: $\hat{\Pi}$: the given strategy

for $t \in \mathcal{T}$ **do**
  if $t = 1$ **then**
    for $j \in \tilde{\mathcal{S}}$ **do**
      if $\hat{\pi}_j^1 = 1$ **then**
        $\mathcal{X}_{j,0}^{\omega,1} \leftarrow 0$ **and** $\mathcal{X}_{j,1}^{\omega,1} \leftarrow \theta_j \quad \forall \omega \in \Omega$
      else
        $\mathcal{X}_{j,0}^{\omega,1} \leftarrow \theta_j$ **and** $\mathcal{X}_{j,1}^{\omega,1} \leftarrow 0 \quad \forall \omega \in \Omega$
      end if
      $\mathcal{Z}^{\omega,1} \leftarrow 0 \quad \forall \omega \in \Omega$
    end for
  end if
  if $1 < t < T$ **then**
    for $j \in \tilde{\mathcal{S}}$ **do**
      if $\hat{\pi}_j^t = 1$ **then**
        $\mathcal{X}_{j,0}^{\omega t} \leftarrow 0$
        $\mathcal{X}_{j,1}^{\omega t} \leftarrow \sum_{i \in \tilde{\mathcal{S}}} \mathcal{X}_{i\pi_i^{t-1}}^{\omega,t-1} P_{iaj}^{\omega} \quad \forall \omega \in \Omega$
      else
        $\mathcal{X}_{j,1}^{\omega t} \leftarrow 0$
        $\mathcal{X}_{j,0}^{\omega t} \leftarrow \sum_{i \in \tilde{\mathcal{S}}} \mathcal{X}_{i\pi_i^{t-1}}^{\omega,t-1} P_{iaj}^{\omega} \quad \forall \omega \in \Omega$
      end if
      $\mathcal{Z}^{\omega t} \leftarrow \mathcal{Z}^{\omega,t-1} + \sum_{i \in \tilde{\mathcal{S}}} \mathcal{X}_{i\pi_i^{t-1}}^{\omega,t-1} Q_{ia}^{\omega} \quad \forall \omega \in \Omega$
    end for
  end if
  if $t = T$ **then**
    for $j \in \tilde{\mathcal{S}}$ **do**
      $\mathcal{Y}_j^{\omega} \leftarrow \sum_{i \in \tilde{\mathcal{S}}} \mathcal{X}_{i\pi_i^{t-1}}^{\omega,t-1} P_{iaj}^{\omega} \quad \forall \omega \in \Omega$
    end for
  end if
end for

Ensure: $(\mathcal{X}, \mathcal{Y}, \mathcal{Z})^{\hat{\Pi}}$: the resulting occupancy measures

Appendix D

HIERARCHICAL RULES FOR TRANSITION PROBABILITIES AND REWARDS

D.1 Transition Probabilities

1. Consider two patients in the same engagement level. The patient with worse health status is more likely to worsen.

$$P_{1,a,2} \geq P_{0,a,1} \qquad \forall a \in \{0,1\}$$
$$P_{4,a,5} \geq P_{3,a,4} \qquad \forall a \in \{0,1\}$$

2. Consider two patients in the same complexity level. The patient with the higher engagement level is less likely to worsen.

$$P_{0,a,1} \geq P_{3,a,4} \qquad \forall a \in \{0,1\}$$
$$P_{1,a,2} \geq P_{4,a,5} \qquad \forall a \in \{0,1\}$$

3. Consider two patients in the same complexity and engagement levels. The patient taking regular care is more likely to worsen than the patient taking special care.

$$P_{0,0,1} \geq P_{0,1,1}$$
$$P_{1,0,2} \geq P_{1,1,2}$$
$$P_{3,0,4} \geq P_{3,1,4}$$
$$P_{4,0,5} \geq P_{4,1,5}$$

4. Probability of making a transition from low engagement level to high engagement level are equal regardless of the health status for all patients under same

action. Moreover, special care is increasing this probability.

$$P_{0,a,3} = P_{1,a,4} = P_{2,a,5} \qquad \forall a \in \{0,1\}$$
$$P_{i,1,i+3} \geq P_{i,0,i+3} \qquad \forall i \in \{0,1,2\}$$

5. Probability of making a transition from high engagement level to low engagement level are equal regardless of the health status for all patients under same action. Moreover, special care is decreasing this probability.

$$P_{3,a,0} = P_{4,a,1} = P_{5,a,2} \qquad \forall a \in \{0,1\}$$
$$P_{i,0,i-3} \geq P_{i,1,i-3} \qquad \forall i \in \{3,4,5\}$$

6. A patient cannot change its health status and engagement level in one decision epoch. That is, diagonal arcs in the network is not possible for all patients under any course of action. Therefore, related transition probabilities equal to zero.

$$P_{0,0,4} = P_{1,0,5} = P_{3,0,1} = P_{4,0,2} = P_{4,0,0} = P_{5,0,1} = P_{1,0,3} = P_{2,0,4} = 0$$
$$P_{0,1,4} = P_{1,1,5} = P_{3,1,1} = P_{4,1,2} = P_{4,1,0} = P_{5,1,1} = P_{1,1,3} = P_{2,1,4} = 0$$

7. There is not any direct transition from *simple* to *complex*.

$$P_{0,a,2} = P_{3,a,5} = 0 \qquad \forall a \in \{0,1\}$$

8. Recovery is not possible for all patients under any course of action. Therefore, transition probabilities associated with the arcs to the left equal to zero.

$$P_{1,a,0} = P_{2,a,1} = P_{4,a,3} = P_{5,a,4} = 0 \qquad \forall a \in \{0,1\}$$

9. Following relations related to the death probabilities have to be satisfied:

$$Q_{2,a} \geq Q_{5,a} = Q_{1,a} \geq Q_{4,a} = Q_{0,a} \geq Q_{3,a} \qquad \forall a \in \{0,1\}$$

10. Special care reduces the death probabilities

$$Q_{i,1} \leq Q_{i,0} \qquad \forall i \in \{0,\ldots,5\}$$

11. There is an upper bound for death probabilities

$$Q_{ia} \leq 0.20 \qquad \forall i \in \{0, \ldots, 5\}, a \in \mathcal{A}$$

D.2 Rewards

1. Following relations have to be satisfied:

$$r_{2,a} \geq r_{5,a} = r_{1,a} \geq r_{4,a} = r_{0,a} \geq r_{3,a} \qquad \forall a \in \{0, 1\}$$

2. Special care is more desirable than regular care.

$$r_{i,1} \geq r_{i,0} \qquad \forall i \in \{0, \ldots, 5\}$$

3. Final stage rewards are computed as follows:

$$R_i \triangleq \frac{\sum\limits_{a \in \mathcal{A}} r_{ia}}{|\mathcal{A}|} \qquad \forall i \in \{0, \ldots, 5\}$$

4. Domain of rewards is $[100, 1000]$ except R^D where $R^D = 0$.

$$100 \leq r_{ia} \leq 1000 \qquad \forall i \in \{0, \ldots, 5\}, a \in \mathcal{A}$$

Appendix E

EVALUATING $F^{\omega}(\Omega)$

An optimal strategy for a single scenario does not guarantee the feasibility under a large number of scenarios. Because the occupancy measures implied by the strategy of a single scenario may violate the capacity constraints for one of the other scenarios. Thus, we embrace the following approach to evaluate $f^{\omega}(\Omega)$.

Let $\bar{\Pi}^{\omega} \triangleq \{\bar{\pi}_i^{\omega t}\}_{i\in\tilde{\mathcal{S}},t\in\tilde{\mathcal{T}}}$ be the optimal strategy, and $\bar{\pi}_i^{\omega t}$ be the optimal policy at stage $t \in \tilde{\mathcal{T}}$ for state $i \in \tilde{\mathcal{S}}$ under scenario ω. We first evaluate $\bar{\Pi}^{\omega}$ under Ω by using Algorithm 4. If it does not involve any infeasibility, the total expected reward characterized by $\bar{\Pi}^{\omega}$ becomes $f^{\omega}(\Omega)$. Otherwise, $\bar{\Pi}^{\omega}$ consists of inapplicable policies. Thus, we find the nearest feasible strategy Π^{*} to $\bar{\Pi}^{\omega}$, and evaluate $f^{\omega}(\Omega)$ as the total expected reward characterized by Π^{*}. Here, the nearest feasible strategy refers to the one that entails feasible policies with a minimum number of changes in the original policy. The distance is specified based on the changes in the policy since we assume that policy makers tend to apply as few as possible changes in the original plan as inconveniences appear during the execution. In this regard, the nearest strategy Π^{*} to $\bar{\Pi}^{\omega}$ is determined by the following mixed-integer program.

$$\textbf{min} \quad \sum_{t\in\tilde{\mathcal{T}}}\sum_{i\in\tilde{\mathcal{S}}} \beta_i^t \tag{E.1}$$

$$\textbf{st;} \quad (4.4)-(4.11) \tag{E.2}$$

$$\pi_i^t - \bar{\pi}_i^{\omega t} \leq \beta_i^t \qquad \forall t \in \tilde{\mathcal{T}}, i \in \tilde{\mathcal{S}} \tag{E.3}$$

$$\bar{\pi}_i^{\omega t} - \pi_i^t \leq \beta_i^t \qquad \forall t \in \tilde{\mathcal{T}}, i \in \tilde{\mathcal{S}} \tag{E.4}$$

$$\beta_i^t \in \{0,1\} \qquad \forall t \in \tilde{\mathcal{T}}, i \in \tilde{\mathcal{S}} \tag{E.5}$$