Monotonicity of Value Function and Optimal Policy in Cross-Layer Design of Communication Systems

Borna Sayedana

Electrical and Computer Engineering group
Engineering department
McGill University
Montreal, Canada

July 2019

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Engineering.

2019/07/31

To my parents ...

Abstract

Markov decision theory has been widely used to model engineering setups as sequential optimization problems. Using Markov decision models and underling techniques in this theory let the engineers improve the performance of the system. Dynamic programming and approximate dynamic programming are the main tools to find optimal policies; however, just finding the optimal policy numerically does not add anything to our engineering intuition about the physical system. In order to grasp a deeper understanding of the underlying physical process, designer is interested in investigating qualitative properties of the optimal value function and optimal policy. These qualitative results not only help researchers to understand the behavior of the physical system better, but also let designers simplify their implementation. Knowing the structure of optimal policy can reduce the implementation of an optimal policy from a look up table to a sparse matrix or just set of thresholds. Markov decision theory has been widely used in queuing problems related to communication systems. In these problems a transmitter is dealing with transmitting a stream of data packets queued in a buffer, over a physical channel. Transmitter should not only deal with stochasticity in the arriving data but also it should manage the physical layer constraints.

These type of problems usually are categorized as queuing problems, and in the literature, the solution to these problems are investigated with tools in both queuing theory and Markov decision theory. In this thesis, our emphasis is on investigating monotonicity property of the optimal strategy in such models. We start with brief introduction to Markov decision processes and common techniques and tools in Markov decision theory to prove monotonicity. We then investigate a classic result in this area. We present simpler proofs for the existing results and try to generalize the idea in two directions. First, we try to establish monotonicity property when transmitter has access to an ACK/NACK feedback channel. In the second approach, we try to show these properties in an energy harvesting scenario.

Résumé

La théorie de la décision de Markov a été largement utilisée pour modéliser les configurations d'ingénierie sous forme séquentielle problème d'optimisation. L'utilisation de modèles de décision de Markov et de techniques sous-jacentes de cette théorie a permis aux ingénieurs d'améliorer les performances du système. Une programmation dynamique approximative est le principal outil pour trouver des politiques optimales. Cependant le fait de trouver numériquement la politique optimale n'ajoute rien à notre ingénierie Intuition sur le système physique; Afin de mieux comprendre le déséquilibre du processus physique dérivé on pourrait être intéressé à étudier les propriétés qualitatives de la fonction valeur optimale et de la politique optimale. Ces résultats qualitatifs aident non seulement les chercheurs à mieux comprendre le comportement du système physique, mais également à permettre aux concepteurs de simplifier leur mise en oeuvre. Connaître la structure d'une politique optimale peut réduire celle d'une table de consultation à une matrice clairsemée ou à un ensemble de seuils. La théorie de la décision de Markov a été largement utilisée dans les problèmes de file d'attente liés à des systèmes de communication. Ces problèmes concernent un émetteur en train de transmettre un flux de paquets de données en file d'attente dans une mémoire tampon sur un canal physique. L'émetteur ne devrait pas traiter uniquement avec la stochasticité dans les données à venir mais il devrait aussi gérer la physique des contraintes desouche. Ce type de problème est généralement classé en tant que problème de file d'attente et, dans la littérature technique, la solution est étudiée à l'aide d'outils de la théorie de la file d'attente et de la théorie de la décision de Markov. Dans ma thèse l'accent est mis sur l'investigation de la monotonie. Propriété de la stratégie optimale dans de tels modèles. Nous commençons par une brève introduction à Markov concernant le processus de décision et les techniques et outils communs dans sa théorie pour prouver la monotonicité. Nous étudions ensuite un résultat classique dans ce domaine. Nous présentons des preuves plus simples pour les résultats existants et essayons de généraliser l'idée dans deux directions. Tout d'abord nous essayons d'établir la propriété de monotonicité lorsque l'émetteur a accès à un retour ACK / NACK canal. Dans la seconde approche nous essayuns de montrer ces propriétés dans un système de récupération d'énergie. Scénario

Acknowledgments

First of all, I would like to express gratitude to my supervisor Prof. Aditya Mahajan. I appreciate him for admitting me to the Masters program at McGill University and trusting in my research abilities. His broad and meanwhile deep knowledge of Control theory, Communication theory, Information theory and the history of each field always impressed me. I appreciate his patience when I was learning scientific writing, his passion when I was struggling with the proofs, and his intuition whenever I got stuck in the equations. He taught me academic research, scientific writing, and clean coding, and I 'm grateful for all of these valuable skills.

I would like to thank Prof. Peter Caines for his trust in my teaching skills and his valuable help in these two years. I also like to thank Prof. Farzad Parvaresh. He taught me first principles of Information theory and encouraged me to work in this field when I was an undergraduate student.

I would like to thank my lab mates for these two years of wonderful experiences. Especially I would like to thank Mohammad for all the interesting and educational discussion we had, Raihan and Jayakumar for their help in coding.

I would like to thank my wonderful friends in Canada, Hadi, Farhad, Nima, and Mohammad for their company and my old friends Hafez and Pedram for their support. In addition, I would like to thank Katia for her help in translation of Resume.

At last I want to express my deepest gratitude toward my parents. Their unbounded, unconditional, and immeasurable love and support has always been the main motivation of my life. Their belief in my abilities along with their day to day motivating advices let me overcome all the difficulties in my life. None of my achievement would have been possible without their help and encouragement. I would like to thank them for all the sacrifices they made in life so that I can pursue my dreams.

Contents

List of Figures

List of Acronyms

MDP	Markov Decision Processes
PDF	Probability Density function
PMF	Probability Mass Function
CDF	Cumulative Density function
ACK	Acknowledgment
NACK	Negative Acknowledgment
AWGN	Additive White Gaussian Noise

Chapter 1

Introduction

1.1 Project Overview

The goal of this thesis is to investigate the qualitative properties of the optimal transmission policies for dynamic resource allocation in communication networks. In the literature, Markov decision theory is the main tool to solve such problems. In this thesis, we use conventional tools in Markov decision theory to analyze such dynamic allocation scenarios. Using the results in that domain, we try to find qualitative properties of the optimal value function and transmission policy.

We start by reviewing basic concepts and notions in Markov decision theory and common techniques to prove monotonicity in the structure of optimal value function and optimal policy. We follow by presenting a classic model and results which formulate power-delay trade-off in fading channels. We generalize this result in two independent directions. First, we try to establish the result in the presence of an ACK/NACK feedback channel. Second, we try to formulate the power delay trade off when an energy harvesting transmitter is dealing with stochastic energy and packet arrivals. In both of these cases, we establish monotonicity properties of the optimal value function. In the feedback case, we are able to prove the monotonicity of optimal policy. In the case of energy harvesting transmitter we present counterexamples for the monotonicity of the optimal policy.

1.2 Thesis Structure

This thesis is structured as follows:

Chapter 2 introduces preliminary concepts and notions of Markov decision theory. Most of the results are taken from [1].

Chapter 3 introduces common techniques in proving qualitative properties of the optimal value function and optimal policy along with minor justification of these results for the case of queuing problems where action space is a function of the state. Most results are taken from [1] and [2].

Chapter 4 examines the classical application of Markov decision theory in proving monotonicity of optimal policy and optimal value function for queuing models. Most of the results are taken from [3].

Chapter 5 investigates a general model of transmission of bursty traffic over fading channel in presence of an adaptive decision feedback. The results in this chapter are original contribution of the writer.

Chapter 6 describes an energy harvesting communication system for which counter-intuitively, optimal policy is not monotone in the state of queue nor amount of energy in the battery. The results in this chapter are original contribution of the writer.

Chapter 2

Markov Decision Theory

This chapter focuses on reviewing basic results and notions from Markov decision theory. All the results in this chapter are from [1]. The goal of this chapter is familiarize the reader to basic notions and concepts in Markov decision theory.

2.1 Introduction

Markov decision theory is widely used in engineering applications. In the environments where sequential decision making is involved, one can use this theory to model the physical systems and find an optimal solution. Dynamic allocation of resources, controlling systems which are following Markovian dynamics, managing traffic in wireless networks are just some applications of Markov decision theory. This theory helps the designer to assign a mathematical model to the physical system, and then uses the existing results in Markov decision theory to find optimal or close to optimal solutions.

In this chapter, we summarize the main results in Markov decision theory which are used in following chapters. Since these are standard results in Markov decision theory, proof of these results are not presented. Reader may refer to the references for detailed proofs.

Markov decision theory in general concerns with sequential decision making in the presence of uncertainty in the underlying environment. The environment is assumed to be a state evolving system with a known Markovian dynamics. The state of the system is a function of agent's decision, the current state of the system, and the dynamic of the system. The dynamic of the system can be deterministic or stochastic. Depending on the

model, agent's decision can have two implications. An instantaneous cost(or reward) and determining the future evolution of the system. As a result of the second implication, it is not necessarily optimal for the agent to choose its action myopically. As a result decision maker has to consider all the sample paths along the horizon and consider the one which maximizes his measure of reward. In most cases, this measure of reward is cumulative rewards over all decision epochs.

2.2 Model Formulation

A Markov decision process is a tuple of following components:

1. Set of decision epochs: This set can be finite $T = \{1, 2, 3 \ldots, N\}$, or infinite $T = \{1, 2, \ldots\}$. We denote the elements of decision epochs by n. If the set T is finite, the model is called a finite horizon MDP, otherwise, it is called infinite horizon MDP.

2. Set of possible states of the system: We denote this set by $\mathcal{S}$. We denote the state of the system at time n, by $s_n \in \mathcal{S}$.

3. Set of possible actions in each state: We denote the set of feasible action in state s by $\mathcal{A}_s$. We show action of the agent at time n, by $a_n \in \mathcal{A}_s$.

4. Set of cost functions: Cost function $c_n : \mathcal{S} \times \mathcal{A}_s \to \mathbb{R}$, is defined as the function of immediate cost that agent receives as a function of state and action for all epochs $n \in \{1, 2, \ldots, N-1\}$. The final cost term is defined as $c_N : \mathcal{S} \to \mathbb{R}$

5. Dynamics of the system: Transition probability function $p_n(s_{n+1} \mid s_n, a_n)$, which gives the probability distribution of next states given the current state s_n and the agent's action a_n.

Note that tuple of these 5 elements completely defines the underlying evolutionary process. In order to reach the Markov decision process, we have to impose an additional assumption. We assume, set of available actions, rewards, and transition probabilities are only a function of current state and not the past states and past actions [1]. The last assumption is imposed to make the evolution of the model Markovian. As a result of this assumption, each induced sample path of the process becomes a Markov chain.

Remark 2.2.1. In the special case when the system is time-homogeneous, i.e., the cost function and the dynamics do not depend on time n, we omit the subscripts and denote c_n and p_n by c and p. we assume c_n, and p_n are not functions of decision epochs n. Hence, we do not include subscript n in the arguments.

Remark 2.2.2. Instead of cost $c(\cdot,\cdot)$, one can formulate the problem based on immediate reward $r(\cdot,\cdot)$, and the intention of the decision maker to maximize the cumulative $r(\cdot,\cdot)$. For the purpose of this thesis it is more suitable to formulate the problem using notion of cost.

Remark 2.2.3. Notice that in general $\mathcal{A}_s$ can be fixed for all the states in the system. In this thesis, usually $\mathcal{A}_s$ denotes the number of feasible packets to transmit as a function of number of packets in the queue, available energy, or state of the channel defined as states of the system. As a result, $\mathcal{A}_s$ the set of feasible actions, usually depends on the current state s.

After defining all these components, we refer to the tuple $(T, \mathcal{S}, \mathcal{A}_s, p(.|s,a), c(s,a))$ as a Markov decision process and we denote it by the abbreviation MDP. The ultimate goal of solving an MDP model is to find the optimal strategy which minimizes the cost measure. It is usually convenient to focus on decision rules at each decision epoch rather than the entire strategy. In the following, we define different types of decision rules.

A decision rule can generally be deterministic or stochastic, Markovian or history dependent. As it is proved in [1], without loss of optimality, we can restrict our attention to Markovian, deterministic decision rules.

***Definition* 2.2.1.** Deterministic decision rule $f_n: \mathcal{S} \to \mathcal{A}$, where $f_n(s) \in \mathcal{A}_s, \forall s \in \mathcal{S}$, is a function which prescribes agent's action at time n.

***Definition* 2.2.2.** We define the policy as the set of all decision rules as following:

$$\pi = (f_1, f_2, \ldots).$$

2.3 Induced stochastic process of an MDP

In order to use the results of probability theory in the theory of MDPs, we have to define the induced probability space of an MDP. This section is devoted to definitions of sample

space, σ-algebra, and probability measure for MDPs. These definitions are important for evaluating performance of a policy using conditional expectation. We define corresponding sample space of an MDP as following:

***Definition* 2.3.1.** Sample space of a finite and infinite MDP is define as the Cartesian product of state and action spaces. Particularly:

$$\Omega_{fin} = \mathcal{S} \times \mathcal{A} \times \mathcal{S} \ldots \mathcal{A} \times \mathcal{S} = \{\mathcal{S} \times \mathcal{A}\}^{N-1} \times \mathcal{S},$$
$$\Omega_{inf} = \{\mathcal{S} \times \mathcal{A}\}^{\infty}.$$

Where by Ω_{fin}, we mean sample space of a finite horizon MDP, and by Ω_{inf}, we mean sample space of an infinite horizon MDP.

Element $w \in \Omega$ consists of a sequence of states and actions, and is defined as a sample path of the induced process. Immediately, we can define σ-algebra as following:

***Definition* 2.3.2.** We define corresponding σ-algebras of the process as following:

$$B(\Omega_{fin}) = B(\{\mathcal{S} \times \mathcal{A}\}^{N-1} \times \mathcal{S}),$$
$$B(\Omega_{inf}) = B(\{\mathcal{S} \times \mathcal{A}\}^{\infty}).$$

where $B(\cdot)$ denotes the Borel set.

With defined sample space of the process and corresponding σ-algebra, we can talk about probability measure induced by fixing policy π. By fixing the policy π, a probability measure will be induced on $B(\Omega_{fin})$, as a result, we can define following set of random variables.

***Definition* 2.3.3.** Random variables S_n, and A_n denote state and action at time n.

$$S_n(\omega) = s_n,$$
$$A_n(\omega) = a_n.$$

We denote these two random variables by S_n, A_n.

Since S_n and A_n are random variables, cost function $c(S_n, A_n)$ is also a random variable. We also can define the history process H_n, as following:

***Definition* 2.3.4.** We define history process as:

$$H_1(\omega) = s_1,$$
$$H_n(\omega) = (s_1, a_1, \ldots, s_n).$$

Suppose W denotes a random variable defined on probability space $\{\Omega, B(\Omega), P^\pi\}$. We define induced expectation of random variable W for a fixed policy, as following:

$$\mathbb{E}^\pi(W) = \sum_{\omega\in\Omega} W(\omega).P^\pi\{\omega\} = \sum_{w\in\mathbb{R}} w.P^\pi\{\omega : W(\omega) = w\}.$$

We can define W as cost of a sample path as following:

$$c(s_1, a_1, \ldots, s_N) = \sum_{n=1}^{N-1} c_n(s_n, a_n) + c_N(s_N).$$

As a result, we can define the expected cost of policy π as following:

$$\mathbb{E}^\pi\Big[c(S_1, A_1, S_2, A_2, \ldots, S_N) \mid S_1 = s_1\Big] = \mathbb{E}^\pi[\sum_{n=1}^{N-1} c_n(S_n, A_n) + c_N(S_N) \mid S_1 = s_1].$$

This defines an evaluation metric for policy π. We use following notation for this evaluation metric:

$$V_N^\pi(s) = \mathbb{E}\Big[\sum_{n=1}^{N-1} c_n(S_n, A_n) + c_N(S_N)\Big| S_0 = s\Big].$$

This expression is normally the optimality metric in Markov decision processes. Based on this definition, we can define the optimal policy π^* as following:

***Definition* 2.3.5.** We define the policy π^*, to be optimal if we have:

$$V_N^{\pi^*}(s) \leq V_N^\pi(s) \quad \forall\pi, \quad s \in \mathcal{S}.$$

In other words, we define the optimal policy to be the solution of following functional optimization problem:

$$\inf_{\pi\in\Pi} V_N^\pi(s).$$

Where by Π, we mean the space of all possible policies.

By the convention of [1], we define the value of an MDP to be

$$V_N^*(s) = \inf_{\pi \in \Pi} V_N(s),$$

immediately we can reason:

$$V_N^{\pi^*}(s) = V_N^*(s),$$

when the inf value is achievable. We can recursively define value of a policy for different decision epochs. As a result, definition of value function can be extended to all the decision epochs by cumulating the onward costs from a decision epoch. We usually refer to this function as cost to go function.

2.4 Backward induction algorithm

In this section, we present Backward induction algorithm. This algorithm is based on recursively solving optimality equations. This algorithm is the basis of all structural proofs in chapters 2, 3, and 4. In most of these proofs we assume certain property holds for the cost to go function at the horizon and using induction properties, we show it also holds for cost to go function in other decision epochs. The algorithm is as following: Let $V_N^*(s) = C_N(s)$. For $n \in \{N-1, \ldots, 1\}$, recursively, define:

$$\begin{aligned} Q_n(s,a) &= c(s,a) + \sum_{s'} V_{n+1}(s') p(s' \mid s, a), \\ V_n &= \min_{a \in \mathcal{A}_s} Q_n(s,a). \end{aligned} \tag{2.1}$$

Let $f_n^*(s)$ denote any $\arg\min$ of the RHS of 2.1. Let $\pi^* = (f_1^*, \ldots, f_N^*)$, then: $\pi^* \in \Pi$ is optimal policy and satisfies:

$$V_N^{\pi^*}(s) = \inf_{\pi \in \Pi} V_N^{\pi}(s), \quad \forall s \in \mathcal{S}$$

and

$$V_n^{\pi^*}(s_n) = V_n^*(s_n), \quad \forall s_n \in S$$

For all $n = 1, 2, \ldots, N$.

The detailed proof of this theorem can be found in [1, Vol I, Theorem 4.5.1]. This

theorem not only states that backward induction algorithm finds the optimal policy, but also it evaluates the optimal policy. In other words this algorithm returns the optimal policy along with its value.

2.5 Infinite horizon Markov decision problems

In the previous section we covered the results for finite horizon MDP problems. We also need to refer to some of the basic results in infinite horizon setup which we will use in following chapters. For the purpose of this thesis we only introduce discounted Markov decision problems and we do not cover average cost problems

2.5.1 Problem formulation and optimality criteria

As stated in previous sections, infinite horizon problem is an MDP setup with $N = \infty$, as a result, we can define the value function of an MDP as following:

$$V^{\pi}(s) = \mathbb{E}^{\pi}\left[\sum_{n=1}^{\infty} \beta^{n-1} c(S_n, A_n) \mid S = s\right],$$

where $\beta \in (0, 1)$ is the discount factor.

***Assumption* 2.5.1.** *We start by imposing following set of assumptions :*

1. *Both cost and transition probabilities are stationary. In other words $c(s, a)$, and $p(s'|s, a)$, are not a function of decision epochs.*
2. *We assume costs are bounded.*

$$|c(s, a)| < M, \quad \forall s \in S, \quad a \in A.$$

3. *Future costs are discounted according to a discounting factor β, where $0 \leq \beta < 1$.*
4. *We assume state space $\mathcal{S}$ is discrete.*

Under these assumptions we have:

***Theorem* 2.5.1.** *For each $s \in \mathcal{S}$, there always exists a Markovian and deterministic optimal policy.*

This theorem is proved in [1, Vol I, Theorem 6.2.7].

We define the matrix representation for value function. This matrix representation helps us prove the monotonicity in infinite horizon setup. P_f is $|\mathcal{S}|\times|\mathcal{S}|$ matrix and $c_f(s)$ is a $|\mathcal{S}|$ dimensional vector with following definitions:

$$C_f(s) = c(s, f(s)),$$
$$P_f(s, j) = p_f(j|s).$$

Using matrix representation of cost function and transition probability matrix, we can express the value function as following:

$$V^\pi(s) = \sum_{n=1}^{\infty} \beta^{n-1} P_\pi^{n-1} C_\pi(s).$$

Where $\pi = (f, f, \ldots)$. Note that due to [1, Vol I, Theorem 6.2.7], there always exists an optimal Markovian deterministic policy, and as a result, we can restrict our attention only to this family of policies. Following this notation we can express the value function of any Markovian deterministic policy as:

$$V = C_f + \beta.P_f.V.$$

One can easily show that [1, Vol I, Theorem 6.1.1] this equation is equivalent to :

$$V = C_f(I - \beta.P_f)^{-1}.$$

The matrix representation of the value function along with the following definition of Bellman equations are useful in stating the results in following chapters.

Using the optimality equation on value function in finite horizon MDPs along with taking the limit as horizon goes to ∞, we can show that the following optimality equation for infinite horizon holds:

$$V(s) = \inf_{a\in\mathcal{A}_s} \Big[c(s, a) + \sum_{s'\in\mathcal{S}} \beta p(s'|s, a)V(s')\Big].$$

Suppose we define $\mathcal{V}$ as the space of all value functions $V : \mathcal{S} \to \mathbb{R}$. For $V \in \mathcal{V}$, we define

nonlinear Bellman operator $\mathscr{B}$, on the space $\mathcal{V}$, as following:

$$\mathscr{B}V = \inf_f \{c_f + \beta.P_f.V\}.$$

Using this definition of Bellman operator, we can rewrite value of an MDP as the function V which satisfies following equation:

$$V = \mathscr{B}V.$$

Similar to the case of finite-horizon, we define action-value function:

$$Q(s,a) = \Big[c(s,a) + \sum_{s' \in \mathcal{S}} \beta p(s'|s,a)V(s')\Big].$$

In the following, we state some of the properties of optimal value and policy for the infinite horizon case.

***Theorem* 2.5.2.** *Suppose $V^* \in \mathcal{V}$ is the optimal value function, and $V \in \mathcal{V}$, where $\mathcal{V}$ is the space of all value functions, then:*

- *If $V \geq \mathscr{B}V$, then $V \geq V^*$.*
- *If $V \leq \mathscr{B}V$, then $V \leq V^*$.*
- *$V = \mathscr{B}V$, then V is the only element of $\mathcal{V}$, with this property and $V = V^*$.*

Proof. This theorem is proved in [1, Vol I, Theorem 6.2.2]. □

Next we state an important theorem in functional analysis which helps us find fixed point of the Bellman operator.

***Definition* 2.5.1.** We say operator T on Banach space U is a contraction if there exists an integer J and a scalar λ', $0 \leq \lambda' < 1$, such that, for all u and v in U,

$$||T^J u - T^J v|| \leq \lambda' ||u - v||$$

***Theorem* 2.5.3.** *(Banach fixed point theorem) Suppose U, is a Banach space and $T : U \to U$, is a contraction mapping. Then*

- *There exists a unique V^* in U, such that $TV^* = V^*$; and*
- *For arbitrary V^0 in U, the sequence $\{V^n\}$ defined by :*

$$V^{n+1} = TV^n = T^{n+1}V^0,$$

converges to V^*.

This theorem is proved in [1, Vol I, Theorem 6.2.3]. In the following theorem, we state the implication of Banach fixed point theorem in MDPs.

***Theorem* 2.5.4.** *For the discounted MDP model defined above, with bounded cost $c(s, a)$, we have following statements :*

- *Operator $\mathscr{B}$ is a contraction mapping on the space $\mathcal{V}$.*
- *There exists a unique $V^* \in \mathcal{V}$ satisfying the fixed point equation $\mathscr{B}V^* = V^*$.*
- *For each Markovian policy π, there exits a unique $V \in \mathcal{V}$, satisfying $\mathscr{B}^\pi V = V$. Moreover, V is unique.*

This theorem is proved in [1, Vol I, Theorem 6.2.4-6.2.5].
At last we state following theorem regarding optimal policy:

***Theorem* 2.5.5.** *A policy $\pi^* \in \Pi$, is optimal if and only if V^{π^*} is the solution of Bellman equation.*

This theorem is proved in [1, Vol I, Theorem 6.2.6].

2.5.2 Value Iteration

Most of the results in infinite horizon MDPs rely on properties of ϵ-optimal value functions or policies. In this section, we briefly review some of these results.

***Definition* 2.5.2.** A policy π^*_ϵ is ϵ-optimal, for fixed ϵ and for all $s \in \mathcal{S}$ if we have:

$$V_\beta^{\pi^*} \leq V_\beta^*(s) + \epsilon.$$

Following algorithm finds a stationary ϵ-optimal policy f_ϵ^∞, and an approximation to its value.

Algorithm 1 Value iteration algorithm

1: Set $n = 0$
2: Select $V^0 \in \mathcal{V}$
3: **for** $s \in \mathcal{S}$ **do**
4: $V^{n+1}(s) = \min_{a \in \mathcal{A}_s} \left[c(s,a) + \sum_{s' \in \mathcal{S}} \beta.p(s'|s,a).V^n(s') \right]$
5: **end for**
6: **if** $||V^{n+1} - V^n|| \leq \frac{\epsilon.(1-\beta)}{2\beta}$ **then**
7: go to step 11.
8: **else**
9: $n \leftarrow n+1$, go to step 3.
10: **end if**
11: **for** $s \in \mathcal{S}$ **do**
12: $f_\epsilon(s) \in \arg\min_{a \in \mathcal{A}_s} \left[c(s,a) + \sum_{s' \in \mathcal{S}} \beta.p(s'|s,a).V^{n+1}(s) \right]$
13: **end for**

This algorithm is called Value Iteration. We define infinity norm as following:

$$||V||_\infty = \max_s V(s).$$

Following theorem states some basic results regarding this algorithm.

***Theorem* 2.5.6.** *Let $V^0 \in \mathcal{V}$, $\epsilon > 0$, and let $\{V_n\}$ satisfies update rule $V^{n+1} = \mathscr{B}V^n$, for $n \geq 1$, then*

- $$\lim_{n \to \infty} ||V_n - V^*||_\infty = 0$$

  .

- *The stationary policy which results from the algorithm is ϵ-optimal.*

- *For $n \geq N$, we have:*
$$||V^{n+1} - V^*||_\infty \leq \frac{\epsilon}{2}$$

Proof of this can be found in [1, Vol I, Theorem 6.3.1].

2.6 Conclusion

In this chapter, we reviewed some of the basic definitions and results in Markov decision theory.

We rigorously defined a Markov decision process, and deterministic Markovian policies. We then provided the definition of optimal value function, and optimal policy along with an algorithm to find these two functions. At the end, we provided dual of these results for the case of infinite horizon MDPs.

Chapter 3

Monotonicity in Markov decision processes

3.1 Introduction

In this chapter we review some of the basic techniques and results to prove structural properties such as monotonicity of optimal value function and optimal policy in Markov decision processes. Monotonicity of optimal value function and optimal policy not only bring more insight about the physical system but also it helps the designer to implement optimal or near optimal strategies computationally efficient. Due to this reason, structural, and qualitative results are always of interest when a physical system can be formulated via MDP theory.

In this chapter some conventional results for proving monotonicity in MDPs are gathered. Since the goal of this research project is the investigation of these techniques and their applications in different models, proof of all of these results are brought. Preliminary results in section 3.2 are from [1], however, Lemma 3.2.1, and 3.2.2 are proved differently. The general results in 3.2.2 and 3.2.3 are stated in [1] without proof. To the best of our knowledge, these generalizations along with the proof using defined notion of restricted submodularity which are stated in this chapter are not provided in the literature. Dual of preliminary results in section 3.3 are provided in [2]. To the best of our knowledge, Theorems 3.3.4, 3.3.5 and their MDP duals, Theorems 3.3.9, and 3.3.10 are original contribution of the writer.

3.2 Preliminaries

We start by defining some mathematical notions which are important in our proofs.

***Definition* 3.2.1.** Let $\mathcal{X}$ and $\mathcal{Y}$ be totally ordered sets and $g(x, y)$ be a real valued function on $\mathcal{X} \times \mathcal{Y}$. We say that $g(x, y)$ is submodular if for $x^+ \geq x^-$ in $\mathcal{X}$ and $y^+ \geq y^-$ in $\mathcal{Y}$, we have:

$$g(x^+, y^-) + g(x^-, y^+) \geq g(x^+, y^+) + g(x^-, y^-).$$

If the reverse inequality holds then we say $g(x, y)$ is supermodular. Minimization operator on supermodular functions has the following property:

***Lemma* 3.2.1.** *If $g(\cdot, \cdot)$ is submodular function on $\mathcal{X} \times \mathcal{Y}$ and $\forall x \in \mathcal{X}$, $\min_{y \in \mathcal{Y}} g(x, y)$ exists, then $f(x) = \min\{y' \in \arg\min_{y \in \mathcal{Y}} g(x, y)\}$ is increasing in x.*

Proof. We prove this lemma by contradiction. Suppose

$$x^+ \geq x^- \implies f(x^-) > f(x^+),$$

by submodularity assumption we have:

$$g(x^+, f(x^+)) + g(x^-, f(x^-)) \geq g(x^-, f(x^+) + g(x^+, f(x^-)). \tag{3.1}$$

Now by the definition of $f(x)$ we know

$$\Big[g(x^-, f(x^+)) - g(x^-, f(x^-))\Big] \geq 0, \tag{3.2}$$

hence from (3.1) and (3.2) we deduce:

$$g(x^+, f(x^+)) \geq g(x^+, f(x^-)) + \Big[g(x^-, f(x^+)) - g(x^-, f(x^-))\Big],$$

hence:

$$g(x^+, f(x^+)) \geq g(x^+, f(x^-)),$$

which is a contradiction since $f(.)$ is defined as minimizing function. As a result, our

conjecture is wrong and we have :

$$x^+ \geq x^- \implies f(x^+) \geq f(x^-).$$

□

Similar result exists for supermodular functions. This result is stated in the following lemma.

***Lemma* 3.2.2.** *If $g(\cdot,\cdot)$ is supermodular function on $\mathcal{X} \times \mathcal{Y}$ and for $x \in \mathcal{X}$, $\min_{y \in Y} g(x,y)$ exists, then $f(x) = \min\{y' \in \arg\min_{y \in \mathcal{Y}} g(x,y)\}$ is decreasing in x.*

Proof is skipped. It is exactly similar to proof of previous lemma. In the following, we define a slightly different notion of submodularity which is useful in our models.

***Definition* 3.2.2.** Let $\mathcal{X}$ and $\mathcal{Y}$ be totally ordered sets and for each $x \in \mathcal{X}$, $\mathcal{Y}_x$ is a subset of $\mathcal{Y}$ such that $\mathcal{Y}_x$ is increasing in x. Let $\mathcal{Z} = \{(x,y) : x \in \mathcal{X}, y \in \mathcal{Y}_x\}$. A function $f : \mathcal{Z} \to \mathbb{R}$ is called restricted submodular if for arbitrary x^+ and x^- such that $x^+ \geq x^-$, and $y^+, y^- \in \mathcal{Y}_{x^-}$ such that $y^+ \geq y^-$, we have:

$$f(x^+, y^+) + f(x^-, y^-) \leq f(x^+, y^-) + f(x^-, y^+).$$

We can establish similar results to 3.2.2, for restricted submodular functions.

***Lemma* 3.2.3.** *If*

1. *function $f(\cdot,\cdot)$ is a restricted submodular function on (x,y),*
2. *$\forall y' \in \mathcal{Y}_{x^+} \setminus \mathcal{Y}_{x^-}$, we have:*
$$y \leq y' \quad \forall y \in \mathcal{Y}_{x^-},$$

Then function $\arg\min_{y \in \mathcal{Y}_x} f(x,y)$ is increasing function of x. Where by $\arg\min_{y \in \mathcal{Y}_x} f(x,y)$ we mean the smallest element $y^* \in \arg\min_{y \in \mathcal{Y}_x} f(x,y)$.

Proof. Notice that we can write the claim as following:

$$\arg\min_{y \in \mathcal{Y}_{x^-}} f(x^-, y) \leq \arg\min_{y \in \mathcal{Y}_{x^+}} f(x^+, y) \tag{3.3}$$

First we prove following claim:

$$\arg\min_{y\in\mathcal{Y}_{x^-}} f(x^-,y) \leq \arg\min_{y\in\mathcal{Y}_{x^-}} f(x^+,y). \tag{3.4}$$

We prove this claim by contradiction:

Contradictory assume,

$$\arg\min_{y\in\mathcal{Y}_{x^-}} f(x^-,y) > \arg\min_{y\in\mathcal{Y}_{x^-}} f(x^+,y).$$

We denote $\arg\min_{y\in\mathcal{Y}_{x^-}} f(x^-,y)$, by y^+, and $\arg\min_{y\in\mathcal{Y}_{x^-}} f(x^+,y)$, by y^-. Since $y^+, y^- \in \mathcal{Y}_{x^-}$, and $y^+ > y^-$, we can write the equation of restricted submodularity as following:

$$f(x^+,y^+) + f(x^-,y^-) \leq f(x^+,y^-) + f(x^-,y^+),$$

if and only if

$$f(x^+,y^+) + \Big[f(x^-,y^-) - f(x^-,y^+)\Big] \leq f(x^+,y^-).$$

Since we defined y^+ as $\arg\min_{y\in\mathcal{Y}_{x^-}} f(x^-,y)$, we infer

$$\Big[f(x^-,y^-) - f(x^-,y^+)\Big] \geq 0.$$

And hence:

$$f(x^+,y^+) \leq f(x^+,y^-).$$

Which is a contradiction since y^+ was defined to by $\arg\min_{y\in\mathcal{Y}_{x^-}} f(x^+,y)$. As a result we prove claim (3.4). Now to prove the claim (3.3), we consider two cases.

- Case (1):

  $$\arg\min_{x\in\mathcal{Y}_{x^+}} f(x^+,y) \in \mathcal{Y}_{x^-}$$

  Then by claim (3.4), we infer:

  $$\arg\min_{y\in\mathcal{Y}_{x^-}} f(x^-,y) \leq \arg\min_{y\in\mathcal{Y}_{x^+}} f(x^+,y)$$

- Case(2):

$$y^* = \arg\min_{y \in \mathcal{Y}_{x^+}} f(x^+, y) \notin \mathcal{Y}_{x^-} \Rightarrow y^* \in \mathcal{Y}_{x^+} \setminus \mathcal{Y}_{x^-}$$

then by the assumptions, we know

$$y \leq y^*, \quad \forall y \in \mathcal{Y}_{x^-},$$

and hence :

$$\arg\min_{y \in \mathcal{Y}_{x^-}} f(x^-, y) \leq \arg\min_{y \in \mathcal{Y}_{x^+}} f(x^+, y).$$

And this concludes the proof.

□

Next we state an important lemma for the increasing sequences. This lemma helps us show the properties of the expectation term in value function.

***Lemma* 3.2.4.** *Let $\{x_i\}, \{x'_i\}$ be real-valued non-negative sequences satisfying*

$$\sum_{i=k}^{\infty} x_i \geq \sum_{i=k}^{\infty} x'_i, \quad \forall k > 0.$$

Also suppose $v_{j+1} \geq v_j$ for $j = 0, 1, \ldots,$ then

$$\sum_{i=0}^{\infty} x_i v_i \geq \sum_{i=0}^{\infty} x'_i v_i.$$

Proof.

$$\begin{aligned}
\sum_{j=0}^{\infty} v_j x_j &= \sum_{j=0}^{\infty} x_j \sum_{i=0}^{j} (v_i - v_{i-1}) = \sum_{j=0}^{\infty} (v_j - v_{j-1}) \sum_{i=j}^{\infty} x_i \\
&= \sum_{j=1}^{\infty} (v_j - v_{j-1}) \sum_{i=j}^{\infty} x_i + v_0 \sum_{i=0}^{\infty} x_i \geq \sum_{j=1}^{\infty} (v_j - v_{j-1}) \sum_{i=j}^{\infty} x'_i + v_0 \sum_{i=0}^{\infty} x'_i \\
&= \sum_{j=0}^{\infty} v_j x'_j.
\end{aligned}$$

□

3.2.1 Montonicity of optimal policy and value function in Markov decision problems

We start by stating the basic result in monotonicity of optimal policy. For this result, we assume that:

$$\mathcal{A}_s = \mathcal{A}, \quad \forall s \in \mathcal{S}.$$

In other words, we assume action space is not a function of current state. Also we define reverse CDF function as following:

$$q(k|s,a) = \sum_{s'=k}^{\infty} p(s'|s,a).$$

This quantity is the probability of reaching a state greater than k at decision epoch $n+1$ while agent is at state s, and chooses action a.

Following theorem states the main result to prove the monotonicity of the value function in MDP problems, when $\mathcal{A}_s = \mathcal{A}, \forall s \in \mathcal{S}$.

***Theorem* 3.2.1.** *Suppose that for $n = \{1, 2, \ldots, N\}$, we have following conditions:*

1. *$c(s,a)$ is an increasing function in s, $\forall a \in \mathcal{A}$.*
2. *$q(k \mid s,a)$ is increasing in s, $\forall a \in \mathcal{A}$ and for all $k \in S$.*
3. *$c_N(s)$ is increasing in s.*

Then

$$V_n^*(s) = \min_{a \in \mathcal{A}} \left[c(s,a) + \sum_{s'=0}^{\infty} p(s' \mid s,a) V_{n+1}^*(s') \right]$$

is increasing in s.

Proof. We prove this theorem by induction:

For $n = N$ we know $V_n^*(s) = c_N(s)$ which is by third assumption an increasing function of s. Now we assume for $t = n+1, n+2, n+3, \ldots, N$ the theorem holds and for $t = n$ we

have:

$$V_n^*(s) = \min_{a \in \mathcal{A}} \left[c(s,a) + \sum_{s'=0}^{\infty} p(s' \mid s,a) V_{n+1}^*(s') \right].$$

We assume $V_n^*(s)$ exists and is attained by the action called a_s^*. Now we consider two states called s^+ and s^- for which we have $s^+ \geq s^-$. By induction hypothesis we have $V_{n+1}^*(s^+) \geq V_{n+1}^*(s^-)$ hence we can write the following inequalities:

$$\begin{aligned} V_n^*(s^+) =& c(s^+, a_{s^+}^*) + \sum_{s'=0}^{\infty} p(s' \mid s^+, a_{s^+}^*) V_{n+1}^*(s') \\ \overset{a}{\geq}& c(s^-, a_{s^+}^*) + \sum_{s'=0}^{\infty} p(s' \mid s^-, a_{s^+}^*) V_{n+1}^*(s') \\ \overset{b}{\geq}& \min_{a \in A'} \left[c(s^-, a) + \sum_{s'=0}^{\infty} p(s' \mid s^-, a) V_{n+1}^*(s') \right] = V_n^*(s^-). \end{aligned}$$

Where (a) follows directly from the assumptions and lemma 3.2.4 and (b) follows from the minimum function properties. □

***Corollary* 3.2.1.** *Suppose that for $n = \{1, 2, \ldots, N\}$ we have following conditions:*

1. *$c(s,a)$ is an increasing function in s, $\forall a \in \mathcal{A}$.*
2. *$q(k \mid s,a)$ is increasing in s, $\forall a \in \mathcal{A}$ and for all $k \in S$.*
3. *$c_N(s)$ is increasing in s.*

Then

$$V_n^*(s) = \min_{a \in \mathcal{A}} \left[c(s,a) + \sum_{s'=0}^{\infty} p(s' \mid s,a) V_{n+1}^*(s') \right]$$

is increasing in s.

Proof. Proof is exactly similar to Theorem 3.2.1. □

Now we try to generalize this result for the case in which $\mathcal{A}$ is not necessarily a constant set and can be a function of state s. Following theorem states sufficient conditions under which monotonicity of the value function is achievable for the cases in which action set is not a constant set and is a function of the state.

***Theorem* 3.2.2.** *Suppose for $n = 1, 2, \ldots, N$ we have following set of assumptions :*

1. *$c(s, a)$ is increasing in s for all $a \in \mathcal{A}_s$.*
2. *$q(k \mid s, a)$ is increasing in s, $\forall a \in \mathcal{A}_s$ and for all $k \in S$.*
3. *$c_N(s)$ is increasing in s.*
4. *for each pair of s^+ and s^- in which $s^+ \geq s^-$ we have following properties:*
   (a) *$\mathcal{A}_{s^-} \subseteq \mathcal{A}_{s^+}$,*
   (b) *if $\forall a' \in \mathcal{A}_{s^+} \setminus \mathcal{A}_{s^-}$, we have:*

$$a \leq a' \quad \forall a \in \mathcal{A}_{s^-}.$$

   (c) *For any action $a_{s^+} \in \mathcal{A}_{s^+} \setminus \mathcal{A}_{s^-}$, there exists an action $a_{s^-} \in \mathcal{A}_{s^-}$ for which we have the following properties:*
      i. *$c(s^+, a_{s^+}) \geq c(s^-, a_{s^-})$,*
      ii. *$q(k \mid s^+, a_{s^+}) \geq q(k \mid s^-, a_{s^-})$ for all $k \in S$.*

then

$$V_n^*(s) = \min_{a \in \mathcal{A}_s} \left[c(s, a) + \sum_{s'=0}^{\infty} p(s' \mid s, a) V_{n+1}^*(s') \right]$$

is increasing in s.

Proof. We prove by induction. For $n = N$ we know $V_n^*(s) = c_N(s)$. Now suppose theorem holds for $n = t+1, t+2, \ldots, N$. For $t = n$ consider two arbitrary states s^+ and s^- such that $s^+ \geq s^-$. Now we assume that for $t = n$ the

$$V_n^*(s^+) = \min_{a \in \mathcal{A}_{s^+}} \left[c(s^+, a) + \sum_{s'=0}^{\infty} p(s' \mid s^+, a) V_{n+1}^*(s') \right]$$

exists and is attained by the action called $a_{s^+}^*$.
We consider two cases:

1. If $a^*_{s^+} \in \mathcal{A}_{s^-}$ then we have the following chain of inequalities:

$$\begin{aligned} V^*_n(s^+) &= c(s^+, a^*_{s^+}) + \sum_{s'=0}^{\infty} p(s' \mid s^+, a^*_{s^+}) V^*_{n+1}(s') \\ &\overset{a}{\geq} c(s^-, a^*_{s^+}) + \sum_{s'=0}^{\infty} p(s' \mid s^-, a^*_{s^+}) V^*_{n+1}(s') \\ &\overset{b}{\geq} \min_{a \in \mathcal{A}_{s^-}} \left[c(s,a) + \sum_{j=0}^{\infty} p(s' \mid s, a) V^*_{n+1}(s') \right] = V^*_n(s^-). \end{aligned}$$

   Where (a) follows from the theorem assumptions and Lemma 3.2.4, (b) follows from property of minimum function.

2. If $a^*_{s^+} \notin \mathcal{A}_{s^-}$ we deduce that $a^*_{s^+} \in \mathcal{A}_{s^+} \setminus \mathcal{A}_{s^-}$. Now from the hypothesis 4, we get that there exists an action $a'_{s^-} \in \mathcal{A}_{s^-}$ for which we have:

   (a) $c(s^+, a^*_{s^+}) \geq c(s^-, a'_{s^-})$,

   (b) $q(k \mid s^+, a^*_{s^+}) \geq q(k \mid s^-, a'_{s^-})$ for all $k \in \mathcal{S}$.

   So we can write the following chain of inequalities:

$$\begin{aligned} V^*_n(s^+) &= c(s^+, a^*_{s^+}) + \sum_{s'=0}^{\infty} p(s' \mid s^+, a^*_{s^+}) V^*_{n+1}(s') \\ &\overset{c}{\geq} c(s^-, a'_{s^-}) + \sum_{s'=0}^{\infty} p(s' \mid s^-, a'_{s^-}) V^*_{n+1}(s') \\ &\overset{d}{\geq} \min_{a \in \mathcal{A}_{s^-}} \{ c(s,a) + \sum_{s'=0}^{\infty} p(s' \mid s, a) v^*_{n+1}(s') \} = V^*_n(s^-). \end{aligned}$$

   Where c follows from the hypothesis 4 and lemma 3.2.4, and d follows from the minimum function property.

□

Previous two theorems are our main tools to prove monotonicity of the value function in MDP problems. Although monotonicity of the value function can bring valuable insight about the structure of the problem and environment, these results do not provide any insight about the structure of optimal policy. In the rest of this section, we bring main

theorems to prove monotonicity of the optimal policy. Monotonicity of the optimal policy is more desired and interesting property because it let the designer of the system implement simpler strategies. In order to prove the structure of optimal policy for models with variable action spaces, we need to define notion of restricted submodular functions.

***Theorem* 3.2.3.** *Suppose that we have following assumptions:*

1. $c(s,a)$ *is increasing in* s *for all* $a \in \mathcal{A}_s$.
2. $q(k \mid s,a)$ *is increasing in* s *for all* $a \in \mathcal{A}_s$ *and for all* $k \in \mathcal{S}$.
3. *for each pair of* s^+ *and* s^- *in which* $s^+ \geq s^-$ *we have the following properties*
   (a) $\mathcal{A}_{s^-} \subseteq \mathcal{A}_{s^+}$,
   (b) *if* $\forall a' \in \mathcal{A}_{s^+} \setminus \mathcal{A}_{s^-}$, *we have:*
   $$a \leq a' \quad \forall a \in \mathcal{A}_{s^-}.$$
   (c) *For any action* $a_{s^+} \in \mathcal{A}_{s^+} \setminus \mathcal{A}_{s^-}$, *there exists an action* $a_{s^-} \in \mathcal{A}_{s^-}$ *for which we have the following properties:*
      i. $c(s^+, a_{s^+}) \geq c(s^-, a_{s^-})$,
      ii. $q(k \mid s^+, a_{s^+}) \geq q(k \mid s^-, a_{s^-})$ *for all* $k \in \mathcal{S}$.
4. $c(s,a)$ *is a restricted submodular function.*
5. $q(k \mid s,a)$ *is a restricted submodular function for all* $k \in \mathcal{S}$.
6. $c_N(s)$ *is increasing in* s.

then there exists an optimal policy which is increasing in s.

Proof. Our goal is to prove that action-value function $Q(s,a)$ is restricted submodular in (s,a). From conditions (1)-(3), and (6), and theorem 3.2.2 we infer $V_n(s)$ is increasing in s. From condition (5), we infer $q(k \mid s,a)$ is restricted submodular in (s,a) for all $k \in \mathcal{S}$. As a result, for two fixed states s^+ and s^-, and actions $a^+, a^- \in \mathcal{A}_{s^-}$, we have:

$$\sum_{s'=0}^{\infty} p(s' \mid s^+, a^+) + \sum_{s'=0}^{\infty} p(s' \mid s^-, a^-) \leq \sum_{s'=0}^{\infty} p(s' \mid s^-, a^+) + \sum_{s'=0}^{\infty} p(s' \mid s^+, a^-).$$

Since terms $p(s' \mid s^+, a^+) + p(s' \mid s^-, a^-)$ and $p(s' \mid s^-, a^+) + p(s' \mid s^+, a^-)$ are non-negative and function $V(s)$, is an increasing function in s, by Lemma 3.2.4, we infer

$$\sum_{s'=0}^{\infty} p(s' \mid s^+, a^+)V(s') + \sum_{s'=0}^{\infty} p(s' \mid s^-, a^-)V(s') \leq \sum_{s'=0}^{\infty} p(s' \mid s^-, a^+)V(s') + \sum_{s'=0}^{\infty} p(s' \mid s^+, a^-)V(s').$$

As a result, the term

$$\sum_{s'=0}^{\infty} p(s' \mid s^+, a^+)V(s')$$

is restricted submodular. By condition 4, we know $c(s, a)$ is restricted submodular, as a result, action-value function

$$Q(s, a) = c(s, a) + \sum_{s'=0}^{\infty} p(s' \mid s^+, a^+)V(s')$$

is restricted submodular in (s, a). By this fact, conditions $3(b)$, and Lemma 3.2.3, we infer there exists an optimal policy which is increasing in s. □

***Corollary* 3.2.2.** *If the conditions of Theorem 3.2.3 are satisfied and instead of conditions 4 and 5, we have that $Q(s, a)$ is restricted submodular on (s, a), then there exists an optimal policy which is increasing in s.*

From Theorem 3.2.3, we can deduce following corollary. In this corollary we assume

$$\mathcal{A}_s = \mathcal{A}, \quad \forall s \in \mathcal{S}.$$

***Corollary* 3.2.3.** *Suppose that we have following conditions:*

1. *$c(s, a)$ is increasing in s for all $a \in \mathcal{A}$.*
2. *$q(k \mid s, a)$ is increasing in s for all $a \in \mathcal{A}$.*
3. *$c(s, a)$ is a submodular function on $\mathcal{S} \times \mathcal{A}$.*
4. *$q(k \mid s, a)$ is submodular function on $\mathcal{S} \times \mathcal{A}$ for all $k \in \mathcal{S}$.*
5. *$c_N(s)$ is increasing in s.*

Then there exists an optimal policy which is increasing in s.

Proof. Conditions 1-2 are similar to Theorem 3.2.3, condition 3 of theorem 3.2.3 is satisfied since:

1. $\mathcal{A} \subseteq \mathcal{A}$,
2. $\forall a' \in \mathcal{A}_{s^+} \setminus \mathcal{A}_{s^-} = \emptyset$.

And conditions 4,5 of Theorem 3.2.3 are satisfied since submodular functions are restricted submodular. As a result of theorem 3.2.3, we conclude the proof.

□

Notice that the key idea in this proof is submodularity of $Q_n(s, a)$. As a result, if we can reach the submodularity of $Q_n(s, a)$ in (s, a) with a different approach, then we can use the result of this theorem.

***Corollary* 3.2.4.** *If $Q_n(s, a)$ is submodular in (s, a), then optimal policy is increasing in s.*

***Theorem* 3.2.4.** *Suppose that we have following conditions:*

1. *$c(s, a)$ is increasing in s for all $a \in \mathcal{A}$.*
2. *$q(k \mid s, a)$ is increasing in s for all $a \in \mathcal{A}$ and for all $k \in \mathcal{S}$.*
3. *$c(s, a)$ is a supermodular function on $\mathcal{S} \times \mathcal{A}$.*
4. *$q(k \mid s, a)$ is supermodular function on $\mathcal{S} \times \mathcal{A}$ for all $k \in \mathcal{S}$.*
5. *$c_N(s)$ is increasing in s.*

Then there exists an optimal policy which is decreasing in s.

Proof. Proof is similar to Corollary 3.2.3. □

3.3 Structural Properties of optimal policy in multi-dimensional state and action spaces

3.3.1 Preliminary Definitions

In this section, we generalize the results in the previous section to the MDP models in which state space or action space are in multi-dimensional spaces. In a multi-dimensional

space, there might not exists a total order. As a result, we have to resort to definitions of partial order and techniques in lattice programming to establish our results. Dual of most of the results in this section exist in [2]. We start by defining basic notions and results in lattice programming. We then use these results in MDP context.

Consider a set X. A binary relation $\lesssim$ is a relation between the elements $x', x'' \in X$, where the statement $x' \lesssim x''$ is either true or false. A binary relation $\lesssim$ is called reflexive if $\forall x \in X$, $x \lesssim x$. A relation is antisymmetric if $x' \lesssim x''$, and $x'' \lesssim x'$ imply $x' = x''$, and if $x' \lesssim x''$ and $x'' \lesssim x'''$ imply $x' \lesssim x'''$.

Using these properties, we define partial order as following :

***Definition* 3.3.1.** A partially ordered set is a set X on which a binary relation $\lesssim$ is defined and it satisfies reflexivity, anti-symmetricity, and transitivity properties.

Two elements x' and x'' of a partially ordered set are ordered if either $x' \lesssim x''$ or $x'' \lesssim x'$ are true statements; otherwise, x' and x'' are unordered. A partially ordered set is a chain if it does not contain an unordered pair of elements.
Assume that X is a partially ordered set and $X' \subseteq X$. If $x' \in X$ and $x \lesssim x', \forall x \in X'$, then x' is an upper bound for X'. If $x' \in X$ is an upper bound for X', then x' is the greatest element of X'.

***Definition* 3.3.2.** If two elements, x' and x'', of a partially ordered set X have a least upper bound (greatest lower bound) in X, it is their join (meet) and is denoted by $x' \vee x''$ ($x' \wedge x''$). A partially ordered set that contains the join and the meet of each pair of its elements is a lattice.

If X' is a subset of a lattice X and X' contains the join and meet of each pair of elements of X', then X' is a sublattice of X.

***Definition* 3.3.3.** A function $f(x)$ from a partially ordered set X to a partially ordered set Y is increasing if $x' \lesssim x''$ implies $f(x') \lesssim f(x'')$ in Y. Similarly, $f(x)$ is decreasing if $x' \lesssim x''$ implies $f(x') \gtrsim f(x'')$ in Y.

***Definition* 3.3.4.** A function $f(x)$ from a partially ordered set X to a partially ordered set Y is non-decreasing if $x' \lesssim x''$ implies $f(x') \not\succ f(x'')$ in Y. Similarly, $f(x)$ is non-increasing if $x' \lesssim x''$ implies $f(x') \not\prec f(x'')$ in Y.

Remark 3.3.1. Notice that in a multi-dimensional space where partial order $\lesssim$ is not necessarily a total order, non-decreasing, and non-increasing properties are not equivalent to weakly-increasing and weakly-decreasing properties. Since

$$a_1 \not\precsim a_2 \quad \xRightarrow{\text{does not necessarily}} \quad a_1 \gtrsim a_2.$$

Now we define two notions of decreasing difference and submodularity in multi-dimensional spaces.

***Definition* 3.3.5.** Suppose that X and T are partially ordered sets and $f(x,t)$ is a real valued function on a subset S on $X \times T$. For $t \in T$, let S_t denote the section of S at t. If $f(x,t'') - f(x,t')$ is decreasing in x on $S_{t'} \cap S_{t''}$ for all $t' \lesssim t''$, then $f(x,t)$ has decreasing differences in (x,t) on S. In other words, if for $t' \lesssim t''$, and $x' \lesssim x''$, we have the following expression, we say $f(x,t)$ has decreasing difference in (x,t).

$$f(x',t'') - f(x',t') \geq f(x'',t'') - f(x'',t').$$

If the inequality is reversed, then we define $f(x,t)$, to have increasing difference on (x,t).

Remark 3.3.2. Notice that the notion of decreasing/increasing difference is independent of the element order.

$$\begin{aligned} & f(x',t'') - f(x',t') \geq f(x'',t'') - f(x'',t') \\ \Leftrightarrow\ & f(x'',t') - f(x',t') \geq f(x'',t'') - f(x',t'') \end{aligned}$$

***Definition* 3.3.6.** If for all x' and x'' on a lattice X, and real valued function $f(x)$, we have:

$$f(x') + f(x'') \leq f(x' \vee x'') + f(x' \wedge x''),$$

then we call this function a supermodular function. If $-f(x)$ is supermodular, then $f(x)$ is submodular.

Similar to the case of one-dimensional case, minimization operator on submodular functions has an important property. First we show that $\arg\min$ forms a lattice.

3.3.2 Preliminary results

***Theorem* 3.3.1.** *If $f(x)$ is submodular on a lattice X, then* $\arg\min_{x\in X} f(x)$ *is a sublattice of X.*

Proof. Pick any $x', x'' \in \arg\min_{x\in X} f(x)$. Because $f(x)$ is submodular on X and $x', x'' \in \arg\min_{x\in X} f(x)$, we can write:

$$0 \overset{a}{\leq} f(x' \wedge x'') - f(x'') \overset{b}{\leq} f(x') - f(x' \vee x'') \overset{c}{\leq} 0, \tag{3.5}$$

where (a) follows from the fact that x'' belongs to $\arg\min_{x\in X} f(x)$, (b) follows from submodularity of function $f(x)$ in x, and (c) follows from the fact that x' belongs to $\arg\min_{x\in X} f(x)$. As a result, we have:

$$\begin{aligned} &0 \overset{a}{\leq} f(x' \wedge x'') - f(x'') \overset{b}{\leq} f(x') - f(x' \vee x'') \overset{c}{\leq} 0, \\ \Rightarrow\ &0 = f(x' \wedge x'') - f(x'') = f(x') - f(x' \vee x'') = 0. \end{aligned}$$

Hence, $(x' \wedge x''), (x' \vee x'') \in \arg\min_{x\in X} f(x)$. As a result, the set function $\arg\min_{x\in X} f(x)$ is a sublattice of X, since existence of any two elements x', x'' in the set implies existence of $(x' \wedge x''), (x' \vee x'')$. □

Following lemma, states an important result in optimization over lattices. This lemma is the basic of most of our results in multi-dimensional action spaces.

***Lemma* 3.3.1.** *If*

- *X is a lattice.*
- *T is a partially ordered set.*
- *$S_t \subseteq X$.*
- *S_t is increasing in t, and*
- *$\forall t', t''$ such that $t' \lesssim t''$, $x' \in S_{t'}$ and $x'' \in S_{t''}$, we have:*

$$f(x' \wedge x'', t') + f(x' \vee x'', t'') \leq f(x', t') + f(x'', t''),$$

then $\arg\min_{x\in S_t} f(x,t)$ *is increasing in* t *on* $\{t : t \in T, \arg\min_{x\in S_t} f(x,t) \textit{is non empty}\}$.

Proof. By previous equation and theorem 3.3.1, we know $\arg\min_{x\in S_t} f(x,t)$ is a sublattice of X for each $t \in T$. Pick $t' \lesssim t''$. Then pick $x' \in \arg\min_{x\in S_{t'}} f(x,t')$ and $x'' \in \arg\min_{x\in S_{t''}} f(x,t'')$. Because $S_{t'} \subseteq S_{t''}$, $x' \wedge x'' \in S_{t'}$ and $x' \vee x'' \in S_{t''}$. Then:

$$0 \overset{a}{\leq} f(x' \wedge x'', t') - f(x', t') \overset{b}{\leq} f(x'', t'') - f(x' \vee x'', t'') \overset{c}{\leq} 0,$$

where (a) follows from the fact that $x' \in \arg\min_{x\in S_{t'}} f(x,t')$, (b) follows from lemma's assumption, and (c) follows from $x'' \in \arg\min_{x\in S_{t''}} f(x,t'')$. Thus equality holds and $x' \wedge x'' \in \arg\min_{x\in S_{t'}} f(x,t')$ and $x' \vee x'' \in \arg\min_{x\in S_{t''}} f(x,t'')$. Hence, since $x' \wedge x'' \lesssim x' \vee x''$, we infer $\arg\min_{x\in S_t} f(x,t)$ is increasing in t. □

Remark 3.3.3. When we state $\arg\min_{x\in S_t} f(x,t)$ in increasing/decreasing in t, we mean there exists elements in the set $\arg\min_{x\in S_t} f(x,t)$ which is increasing/decreasing in t. For example, in the previous theorem, we showed $x' \wedge x'' \in \arg\min_{x\in S_{t'}} f(x,t')$, and $x' \vee x'' \in \arg\min_{x\in S_{t''}} f(x,t'')$. By the existence of these two elements in sets $S_{t'}$, and $S_{t''}$, we reason $\arg\min_{x\in S_t} f(x,t)$ is increasing in t.

Following lemma, establishes sufficient conditions to prove that $\arg\min_{x\in S_t} f(x,t)$ is decreasing in t.

***Lemma* 3.3.2.** *If*

- X *is a lattice.*
- T *is a partially ordered set.*
- S_t *is a subset of* X *for each* t *in* T.
- S_t *is increasing in* t *in* T.
- $\forall t', t''$, *such that* $t' \lesssim t''$, $x' \in S_{t'}$ *and* $x'' \in S_{t''}$, *we have:*

$$f(x' \wedge x'', t'') + f(x' \vee x'', t') \leq f(x', t') + f(x'', t''),$$

then $\arg\min_{x\in S_t} f(x,t)$ *is decreasing in* t *on* $\{t : t \in T, \arg\min_{x\in S_t} f(x,t) \textit{is non empty}\}$.

Proof. By previous equation and theorem 3.3.1, we know $\arg\min_{x\in S_t} f(x,t)$ is a sublattice of X for each $t \in T$. Pick $t' \lesssim t''$. Then pick $x' \in \arg\min_{x\in S_{t'}} f(x,t')$ and $x'' \in \arg\min_{x\in S_{t''}} f(x,t'')$. Because $S_{t'} \subseteq S_{t''}$, $x' \wedge x'' \in S_{t'}$ and $x' \vee x'' \in S_{t''}$ Then:

$$0 \overset{a}{\leq} f(x' \wedge x'', t'') - f(x'', t'') \overset{b}{\leq} f(x', t') - f(x' \vee x'', t') \overset{c}{\leq} 0.$$

Where (a) follows from the fact that $x'' \in \arg\min_{x\in S_{t''}} f(x,t'')$, (b) follows from lemma's assumption, and (c) follows from $x' \in \arg\min_{x\in S_{t'}} f(x,t')$. Thus equality holds and $x' \wedge x'' \in \arg\min_{x\in S_{t''}} f(x,t)$ and $x' \vee x'' \in \arg\min_{S_{t'}} f(x,t'')$. Hence, since $x' \wedge x'' \lesssim x' \vee x''$ we infer $\arg\min_{x\in S_t} f(x,t)$ is decreasing in t. □

Following theorem provides easier to check sufficient conditions for function $f(x,t)$ to prove monotonicity in $\arg\min$ argument. This conditions, provide independent conditions on joint behavior of optimization parameters as well as each of them to prove monotonicity in $\arg\min$. These behaviors get translated to lemma 3.3.1.

***Theorem* 3.3.2.** *If:*

1. *X is a lattice.*
2. *T is a partially ordered set.*
3. *$S_t \subseteq X, \forall t \in T$.*
4. *S_t is increasing in t.*
5. *$f(x,t)$ has decreasing difference in (x,t) on $X \times T$.*
6. *$f(x,t)$ is submodular in x, $\forall t \in T$.*

then $\arg\min_{x\in S_t} f(x,t)$ *is increasing in t, when:*

$$\{t : t \in T, \arg\min_{x\in S_t} f(x,t) \textit{is non empty}\}.$$

Proof. It is sufficient to prove that the equation in the hypothesis of lemma 3.3.1 is true.

Choose $t', t'' \in T$, $x' \in S_{t'}$ and $x'' \in S_{t''}$, then following chain of inequalities hold:

$$f(x' \wedge x'', t') - f(x', t') \overset{a}{\leq} f(x'', t') - f(x' \vee x'', t') \overset{b}{\leq} f(x'', t'') - f(x' \vee x'', t'')$$
$$\Rightarrow f(x'', t') - f(x', t') \leq f(x'', t'') - f(x', t'').$$

Where (a) follows from the submodularity of function $f(., t)$ in x, for fixed t, and (b) follows from decreasing difference property. Inequality (b) can be translated to standard form of definition of decreasing difference as following:

$$f(x'', t') - f(x' \vee x'', t') \overset{b}{\leq} f(x'', t'') - f(x' \vee x'', t'') \Leftrightarrow$$
$$f(x' \vee x'', t'') - f(x' \vee x'', t') \leq f(x'', t'') - f(x'', t').$$

As a result, we verify the hypothesis of the lemma 3.3.1 and conclude the proof. □

***Theorem* 3.3.3.** *If:*

1. *X is a lattice.*
2. *T is a partially ordered set.*
3. *$S_t \subseteq X, \forall t \in T$.*
4. *S_t is increasing in t.*
5. *$f(x, t)$ has increasing difference in (x, t) on $X \times T$.*
6. *$f(x, t)$ is submodular in x, $\forall t \in T$.*

then $\arg\min_{x \in S_t} f(x, t)$ *is decreasing in t, when:*

$$\{t : t \in T, \arg\min_{x \in S_t} f(x, t) \textit{is non empty}\}$$

Proof. It is sufficient to prove that the equation in the hypothesis of lemma 3.3.2 is true. choose $t', t'' \in T$, $x' \in S_{t'}$ and $x'' \in S_{t''}$, then following chain of inequalities hold:

$$f(x' \wedge x'', t'') - f(x'', t'') \overset{a}{\leq} f(x'', t'') - f(x' \vee x'', t'') \overset{b}{\leq} f(x', t') - f(x' \vee x'', t')$$
$$\Rightarrow f(x'', t') - f(x', t') \leq f(x'', t'') - f(x', t'')$$

Where (a) follows from the submodularity of function $f(.,t)$ in x, for fixed t, and (b) follows from decreasing difference property. Inequality (b) can be translated to standard form of definition of decreasing difference as following:

$$f(x'',t') - f(x' \vee x'',t') \overset{b}{\leq} f(x'',t'') - f(x' \vee x'',t'') \Leftrightarrow$$
$$f(x' \vee x'',t'') - f(x' \vee x'',t') \leq f(x'',t'') - f(x'',t')$$

As a result, we verify the hypothesis of lemma 3.3.2 and conclude the proof.

□

Previous two theorems provide sufficient conditions to prove the structure of arg min functions. Dual of these theorems are provided for arg max functions in [2]. We try to translate these results and provide sufficient conditions for MDP problems later in this chapter. Following two theorems provide weaker results about the structure of arg min. They provide certain restrictions for the structure of arg min functions. Notice that theorems 3.3.2, 3.3.3 only provide sufficient conditions that there exists elements in arg min which satisfies certain properties. Following two theorems are stronger in the sense that, they establish that no element in arg min exists with certain properties. To the best of our knowledge these results, are not established in the literature before.

***Theorem* 3.3.4.** *If:*

1. *X is a lattice.*
2. *T is a partially ordered set.*
3. *$S_t \subseteq X, \forall t \in T$.*
4. *S_t is increasing in t.*
5. *$f(x,t)$ has increasing difference in (x,t) on $X \times T$.*

Then $\arg\min_{x \in S_t} f(x,t)$ *is non-increasing in* t *on* $\{t : t \in T, \arg\min_{x \in S_t} f(x,t) \textit{is non empty}\}$.

Proof. From increasing difference property, for $t' \leq t''$, $x' \lesssim x''$, we have:

$$f(x',t'') - f(x',t') \leq f(x'',t'') - f(x'',t'),$$

if and only if:

$$f(x', t'') + f(x'', t') - f(x'', t'') \leq f(x', t'). \tag{3.6}$$

In order to prove the claim of theorem, we use contradiction. Suppose we use following notation:

$$x^{*'} = \arg\min_{x \in S_{t'}} f(x, t') \quad , \quad x^{*''} = \arg\min_{x \in S_{t''}} f(x, t'').$$

Now contradictory to the claim, we assume:

$$t' \lesssim t'' \quad \text{and} \quad x^{*'} < x^{*''}.$$

Using this assumption and equation 3.6 we can write:

$$f(x^{*'}, t'') + \Big[f(x^{*''}, t') - f(x^{*''}, t'') \Big] \leq f(x^{*'}, t').$$

Now, due to property of min function, we know

$$\Big[f(x^{*''}, t') - f(x^{*''}, t'') \Big] \geq 0 \Rightarrow f(x^{*'}, t'') < f(x^{*'}, t').$$

Which is a contradiction, since $x^{*'} = \arg\min_{x \in S_{t'}} f(x, t')$ and we complete the proof. □

Similarly, we can establish following theorem:

***Theorem* 3.3.5.** *If:*

1. *If X is a lattice.*
2. *T is a partially ordered set.*
3. *$S_t \subseteq X, \forall t \in T$.*
4. *S_t is increasing in t.*
5. *$f(x, t)$ has decreasing difference in (x, t) on $X \times T$.*

Then $\arg\min_{x \in S_t} f(x, t)$ *is non-increasing in* t *on* $\{t : t \in T, \arg\min_{x \in S_t} f(x, t) \textit{is non empty}\}$.

Proof. Proof is exactly similar to theorem 3.3.5. □

3.3.3 Monotonicity results in multi-dimensional state and action MDP problems

In the this section, we investigate the implication of previous results on proving monotonicity in Markov decision process. In all of our results sufficient conditions are provided using the definition of reverse CDF function $q(.)$; however, these type of sufficient conditions on the distribution of stochasticity of the MDP models usually are stated differently in the literature. For the sake of completeness, we briefly review these style of defining these properties. A more detailed explanation can be found in [2].

Consider a collection of distribution$\{F_t(w) : t \in T\}$ on $\mathbb{R}^n$ that are parameterized by $t \in T$, where $T \in \mathbb{R}^n$. If $\int_S F_t(w)$ is an increasing function of $t \in T$, for each set S, then we define $F_t(w)$ to be stochastically increasing in t. If $T \in \mathbb{R}^n$ is a lattice and $\int_S F_t(w)$ is submodular function of t for each increasing set $S \in \mathbb{R}^n$, then $F_t(w)$ is stochastically submodular in t.

It is well-known that a collection of distribution functions $\{F_t(w) : t \in T\}$ on $\mathbb{R}^1$ is stochastically increasing in $t \in T \subset \mathbb{R}^m$ if and only if $1 - F_t(w)$ is increasing in $t \in T$ for each $w \in \mathbb{R}$. Since the source of stochasticity in my thesis is usually one dimensional distribution, we provide all of our sufficient conditions based on definition of $q(.)$ function.

The following theorem, provides the sufficient conditions to prove monotonicity of value function in the state s.

***Theorem* 3.3.6.** *Suppose we define a total order on state space $\mathcal{S}$, and a partial order $\overset{a}{\lesssim}$ on action space $\mathcal{A}$. Also suppose we have following set of assumptions:*

1. *$c(s, a)$ is increasing in s for all $a \in \mathcal{A}_s$.*
2. *$q(k \mid s, a)$ is increasing in s, $\forall a \in \mathcal{A}_s$ and for all $k \in S$.*
3. *$c_N(s)$ is increasing in s.*
4. *for each pair of s^+ and s^- in which $s^+ \geq s^-$ we have following properties:*
   - *(a) $\mathcal{A}_{s^-} \subseteq \mathcal{A}_{s^+}$,*
   - *(b) if $\forall a' \in \mathcal{A}_{s^+} \setminus \mathcal{A}_{s^-}$, we have:*

$$a \lesssim a' \quad \forall a \in \mathcal{A}_{s^-}.$$

(c) *For any action* $a_{s^+} \in \mathcal{A}_{s^+} \setminus \mathcal{A}_{s^-}$*, there exists an action* $a_{s^-} \in \mathcal{A}_{s^-}$ *for which we have the following properties:*

i. $c(s^+, a_{s^+}) \geq c(s^-, a_{s^-})$,

ii. $q(k \mid s^+, a_{s^+}) \geq q(k \mid s^-, a_{s^-})$ *for all* $k \in S$.

then

$$V_n^*(s) = \min_{a \in \mathcal{A}_s} \left[c(s,a) + \sum_{s'=0}^{\infty} p(s' \mid s, a) V_{n+1}^*(s') \right]$$

is increasing in s.

Proof. This theorem exactly follows from theorem 3.2.1. The partial order on the action space does not change any steps of the proof.

□

Following theorem states sufficient conditions to prove the monotonicity of the optimal policy when action space is multi-dimensional, a partial order $\overset{a}{\lesssim}$ is defined on action space and induces a lattice, also action space $\mathcal{A}_s$ is not a constant function of state s.

***Theorem* 3.3.7.** *Suppose* $\mathcal{A}_s$ *is a sublattice of action spaces for each* s*. State space is ordered with a total order, and following conditions are satisfied:*

1. $c(s,a)$ *is increasing in* s *for all* $a \in \mathcal{A}_s$.

2. $q(k \mid s, a)$ *is increasing in* s, $\forall a \in \mathcal{A}_s$ *and for all* $k \in S$.

3. $c_N(s)$ *is increasing in* s.

4. *for each pair of* s^+ *and* s^- *in which* $s^+ \geq s^-$ *we have following properties:*

   (a) $\mathcal{A}_{s^-} \subseteq \mathcal{A}_{s^+}$,

   (b) *if* $\forall a' \in \mathcal{A}_{s^+} \setminus \mathcal{A}_{s^-}$*, we have:*

   $$a \lesssim a' \quad \forall a \in \mathcal{A}_{s^-}.$$

   (c) *For any action* $a_{s^+} \in \mathcal{A}_{s^+} \setminus \mathcal{A}_{s^-}$*, there exists an action* $a_{s^-} \in \mathcal{A}_{s^-}$ *for which we have the following properties:*

   i. $c(s^+, a_{s^+}) \geq c(s^-, a_{s^-})$,

ii. $q(k \mid s^+, a_{s^+}) \geq q(k \mid s^-, a_{s^-})$ *for all* $k \in S$.

5. $c(s,a)$ *is submodular in* a*, for fixed* s.

6. $q(s,a)$ *is submodular in* a*, for fixed* s.

7. $c(s,a)$ *has decreasing difference in* $\mathcal{S} \times \mathcal{A}$*(with partial order* $\overset{a}{\lesssim}$*).*

8. $q(s,a)$ *has decreasing difference in* $\mathcal{S} \times \mathcal{A}$*(with partial order* $\overset{a}{\lesssim}$*).*

Then there exist an optimal policy which is increasing in s *with partial order* $\overset{a}{\lesssim}$*, on actions.*

Proof. From conditions 1-4 and theorem 3.3.6, we infer value function $V^*(s)$ is increasing in s. First we prove for arbitrary states s^+ and s^-, such that $s^+ \geq s^-$, induced action space $\mathcal{A}_{s^-}$, and all epochs $n = 1, 2, \ldots, N$ we have:

$$\arg\min_{a \in \mathcal{A}_{s^-}} Q_n(s^-, a) \leq \arg\min_{a \in \mathcal{A}_{s^-}} Q_n(s^+, a). \tag{3.7}$$

By equation (3.7), we mean there exists an element in $\arg\min_{a \in \mathcal{A}_{s^-}} Q_n(s^-, a)$, which is less than or equal to an element in $\arg\min_{a \in \mathcal{A}_{s^-}} Q_n(s^+, a)$. Suppose $a^+, a^- \in \mathcal{A}_{s^-}$ such that $a^+ \geq a^-$. From assumption (8), we know $q(\cdot, \cdot)$ function has decreasing difference in (s, a), hence we can write:

$$\sum_{s'=0}^{\infty} p(s' \mid s^+, a^+) + \sum_{s'=0}^{\infty} p(s' \mid s^-, a^-) \leq \sum_{s'=0}^{\infty} p(s' \mid s^-, a^+) + \sum_{s'=0}^{\infty} p(s' \mid s^+, a^-).$$

We know $V^*(s)$ is increasing function of s. Right hand side and left hand side of previous equation are non-negative, hence, as a result of lemma 3.2.4, we infer

$$\sum_{s'=0}^{\infty} p(s' \mid s, a) V^*(s'),$$

also has decreasing difference in (s, a). From assumption (7), we know $c(s, a)$ also has decreasing difference in (s, a), hence :

$$Q_n(s, a) = c(s, a) + \sum_{s'=0}^{\infty} p(s' \mid s, a) V^*_{n+1}(s')$$

has decreasing difference in (s, a). Moreover, from assumptions 5 and 6, we infer $Q_n(s, a)$ is submodular in a, for fixed s. Also notice that $\mathcal{A}_{s^-} \subseteq \mathcal{A}_{s^+}$ by assumptions. Hence, by using theorem 3.3.2, we infer optimal policy is increasing in the state s.
In order to conclude the proof we consider two cases. Consider two states s^+, s^-, induced action spaces $\mathcal{A}_{s^-}$, and $\mathcal{A}_{s^+}$. Suppose we denote $\arg\min_{a \in \mathcal{A}_{s^+}} Q_n(s^+, a)$ by a^*. If $a^* \in \mathcal{A}_{s^-}$, by equation 3.7, we deduce the claim of theorem. If $a^* \in \mathcal{A}_{s^+} \setminus \mathcal{A}_{s^-}$, we deduce the claim of the theorem by condition $3(b)$ and this concludes the proof. □

In the following theorem, we state sufficient conditions under which optimal policy is decreasing in the state s, for the defined partial order $\overset{a}{\lesssim}$. Since we use this theorem for the conditions in which $\mathcal{A}_s$ is not a function of s ,and $\mathcal{A}_s = \mathcal{A}, \quad \forall s \in \mathcal{S}$, we only state the theorem for fixed action space $\mathcal{A}$. The generalization of this theorem for $\mathcal{A}_s$ is also achievable.

***Theorem* 3.3.8.** *Suppose $\mathcal{A}$ is a sublattice of action spaces for all states s. State space is ordered with a total order, and following conditions are satisfied:*

1. *$c(s, a)$ is increasing in s for all $a \in \mathcal{A}$.*
2. *$q(k \mid s, a)$ is increasing in s, $\forall a \in \mathcal{A}$ and for all $k \in S$.*
3. *$c_N(s)$ is increasing in s.*
4. *$c(s, a)$ is submodular in a, for all s.*
5. *$q(s, a)$ is submodular in a, for all s.*
6. *$c(s, a)$ has increasing difference in $\mathcal{S} \times \mathcal{A}_{s^-}$ (with respect to partial order $\overset{a}{\lesssim}$).*
7. *$q(s, a)$ has increasing difference in $\mathcal{S} \times \mathcal{A}_{s^-}$ (with respect to partial order $\overset{a}{\lesssim}$).*

Then there exist an optimal policy which is decreasing in s with partial order $\overset{a}{\lesssim}$, on actions.

Proof. The proof is exactly similar to Theorem 3.3.7. □

In the final two theorems of this chapter, we state two weaker results related to structural results, in multi-dimensional action and state spaces. These theorems tell us the structure of optimal policy cannot by in a restricted area. One can use these last two theorems when increasing/decreasing difference on $\mathcal{S} \times \{\mathcal{A}\}$ exists; however, submodularity in the action space $\mathcal{A}$ is not easily verifiable.

***Theorem* 3.3.9.** *Suppose $\mathcal{A}$ is a sublattice of action spaces for all states s. State space is ordered with a total order, and conditions* (1)-(3)*, and* (6)-(7) *of theorem 3.3.8 , are satisfied. Then there exist an optimal policy which is non-increasing in s with partial order $\overset{a}{\lesssim}$, on actions.*

Proof. It is shown in theorem 3.3.8 that conditions (1)-(3) and (6)-(7), are sufficient to show $Q(s, a)$ has increasing difference. The result of theorem follows from this property and theorem 3.3.4. □

***Theorem* 3.3.10.** *Suppose $\mathcal{A}_s$ is a sublattice of action spaces for each s. State space is ordered with a total order, and conditions* (1)-(4)*, and* (6)-(7) *of theorem 3.3.7 are satisfied. Then there exist an optimal policy which is non-decreasing in s with partial order $\overset{a}{\lesssim}$, on actions.*

Proof. It is shown in theorem 3.3.7 that conditions (1)-(4) and (6)-(7), are sufficient to show, $Q(s, a)$ has decreasing difference. The result of the theorem follows from this property and theorem 3.3.5. □

Chapter 4

Power Delay Tradeoff in Wireless Communication Systems

In this chapter, we use the techniques of the previous chapter in a simple model of wireless systems. We model a physical channel along with a buffer with stochastic arrival. The cost function is defined as the combination of power consumption to transmit packets and the delay incurred by the existing packets in the buffer. Most of the results in this chapter are from [3] but the proofs are different. We prove monotonicity of value function and optimal policy as a function of number of existing packets in the buffer. These results are the classic application of structural results of MDP theory in communication systems.

4.1 Introduction

In wireless communication systems, networks are designed in layers mainly to reduce designing complications. Recently, a number of interesting models have been proposed to optimize transmission of data between layers in networks. Most of these models can be summarized into a queue being feed by higher layers sporadic data arrivals along with a physical model for the channel. In most of these scenarios, transmission scheduling to control the buffer is the main concern of designers. In this chapter, we present one of the first cross layer queuing models which used Markov decision processes to find the optimal policy. These results originally appeared in [3]. In addition, the structural results in MDP theory is used to prove the monotonicity of optimal value function and optimal policy.

4.2 Problem Formulation

We begin by defining the queue dynamics. Suppose N_k denotes the number of existing packets in the buffer at time k, U_k denotes number of transmitted bits at time k, and buffer's capacity is equal to L. Also suppose there exists an arrival process where the number of arriving packets at time k is denoted by D_k. Given these notations we can formulate the queue dynamics as following:

$$N_{k+1} = \min\Big\{ \max\{N_k + D_{k+1} - U_k, D_{k+1}\}, L\Big\}.$$

This dynamic mimics the possibility of depleting the buffer completely and the possibility of overflow. We denote this function with following notation:

$$N_{k+1} = [N_k - u + D_k]_L$$

Notice that L might be infinite. The state of this buffer is a function of arriving process and control action U_k. Transmitter decides on the rate U_k based on the state of queue N_k, state of the channel H_k and arriving processes D_k. We show the space of these variables as $(N_k, H_k, D_k) \in \mathcal{N} \times \mathcal{H} \times \mathcal{D}$. We assume $\{H_k\}$ and $\{D_k\}$ are independent and ergodic Markov chains and transmitter does not have any control over these processes. Each transmission rate incurs a power consumption. Power consumption is a function of chosen control action and state of the channel. We denote power function by $P(h_k, u_k)$. Furthermore, we assume power function is of the form:

$$P(h, u) = \frac{\tilde{P}(u)}{|h|^2}.$$

where $\tilde{P}(u)$ is a monotonically increasing convex function of u.

Additionally at each time k a buffer cost $c(n_k)$ is incurred. We assume that $c(n)$ depends only on n and is an increasing, convex function of n.

The control sequence $f_k : \mathcal{N} \times \mathcal{H} \times \mathcal{D} \to \mathcal{U}$ determines the transmission rate at time k. The goal of transmitter is to determine $U_k = f_k(N_k, H_k, D_k)$ such that cumulative cost is minimized. For a transmission policy, the discounted transmission power is:

$$\lim_{M\to\infty} \sup \mathbb{E}\Big[\beta^M \sum_{k=1}^{M} P\Big(H_k, f_k(N_k, H_k, D_k)\Big)\Big].$$

Likewise, the discounted delay cost of the problem is defined as:

$$\limsup_{M\to\infty} \mathbb{E}\Big[\beta^M \sum_{k=1}^{M} c(N_k)\Big].$$

The designer might be interested in minimizing both of these quantities; however, these two objectives are conflicting with one another. In fact, there is a tradeoff between average power usage and average delay incurred by the system. In order to characterize this tradeoff better, we define our objective as the combination of power and delay as following:

$$J(f) = \limsup_{M\to\infty} \mathbb{E}^f\Big[\beta^M \sum_{k=1}^{M} P(H_k, f_k(N_k, H_k, D_k)) + \lambda.c(N_k)\Big]. \tag{4.1}$$

Where $\lambda > 0$ is the Lagrange multiplier associated with a constraint on the average delay incurred by the buffer. Given these definitions, we can rigorously formulate the model as an MDP problem as following:

Problem 1. *Given the buffer length L, power cost $P(\cdot)$, delay cost $c(\cdot)$, PMF of the arrival process, PMF of channel processes, Lagrange factor λ, and the discount factor β, choose a feasible scheduling policy f to minimize the performance $J(f)$ given by* (4.1).

We impose following assumptions on the model. First the arrival process is always limited by queue capacity L, $A \subset \{0, \ldots, L\}$, and transmission rates are also upper bounded by queue capacity $\mathcal{U} = \{0, \ldots, L\}$. We assume $H \in \mathbb{C}$, and $|H| \leq \infty$. We also use the $P_{d,d'}$, and $P_{h,h'}$ for data arriving and fading distributions.

$$P_{d,d'} = Pr\Big(D_{n+1} = d' | D_n = d\Big)$$

$$P_{h,h'} = Pr\Big(H_{n+1} = h' | H_n = h\Big)$$

By Theorem 2.5.1, we know for such formulation there exists an optimal Markovian deterministic policy.

In the $\beta-$discounted problem, the optimal cost depends on the initial state. Let $J^*_\beta(n, h, d)$ denote this cost when $N_0 = n, H_0 = h, D_0 = d$. We write the optimality

equation for model as following:

$$V^*(n,h,d) = P(f(n,h,d),h) + \lambda.c(n) + \beta \sum_{h' \in \mathcal{H}, d' \in \mathcal{D}} P_{d,d'} P_{h,h'} V^*([N_k - u + D_k]_L, h', d')$$
$$= \inf_{u \in \mathcal{U}} \Big(P(u,h) + \lambda.c(n) + \beta \sum_{h' \in \mathcal{H}, d' \in \mathcal{D}} P_{d,d'} P_{h,h'} V^*([N_k - u + D_k]_L, h', d') \Big).$$

Following the notation in chapter 2, we define the nonlinear Bellman operator $\mathscr{B}$ as the right hand side of the previous equation. Let $\mathscr{B}^k$ denote composition of $\mathscr{B}$, for k times. Then using theorem 2.5.6, we can find optimal value as following:

$$\lim_{k \to \infty} \mathscr{B}^k(V) = V^*_\beta$$

Where V is an arbitrary initialization in the space of value functions. We know V^* is the optimal value of the MDP and unique fixed point of the the equation:

$$V^*_\beta = \mathscr{B} V^*_\beta.$$

Consider two random variables defined on a common probability space. X is said to be *stochastically larger* than Y if $Pr(X > a) \geq Pr(y > a)$ for all $a \in \mathbb{R}$. It can be easily proved that X is stochastically greater than Y if and only if $\mathbb{E}(f(X)) \geq \mathbb{E}(f(Y))$ holds for all non-decreasing functions $f : \mathbb{R} \to \mathbb{R}$. A real valued Markov chain $\{X_n\}$ is defined to be stochastically monotone if $Pr(X_n > a | X_{n-1} = x)$ is a non-decreasing function of x for all n,and all a. We will refer to this property for the fading process. We define $\{H_n\}$ to be stochastically monotone if $\{|H_n|\}$ is. Intuitively this means that the probability of deep fade is no higher, given that the channel is good, than when the channel is bad. Notice that memoryless channels are stochastically monotone.

***Assumption* 4.2.1.** *Process $\{H_n\}$ is a stochastically monotone process.*

4.3 Monotonicity of value function

In this chapter we establish the monotonicity of value function in this problem as a function of different parameters. More specifically we show $V^*(n,h,d)$ is increasing in n, and d, and it is decreasing in $|h|$. These results correspond to the intuition that it is less desirable

for the transmitter to have large queue occupancy, and have a deep fading in the channel. Transmitter also prefers to start with a lower initial arrival rate.

***Lemma* 4.3.1.** *$V^*(n,h,d)$ is increasing in n for all $h \in \mathcal{H}$ and all $d \in \mathcal{D}$.*

Proof. First we solve the problem in finite horizon setting. We use theorem 3.2.1 to prove this property in finite horizon setup. If we assume states h, and d are fixed, cost function $P(u,h) + \lambda c(n)$ is an increasing function in state n. Moreover transmission probabilities are i.i.d. and satisfy condition 2 in theorem 3.2.1. Finally, we can assume the final cost is increasing in the n. As a result, we prove $V^*_\beta(n,h,d)$ is increasing in n, for fixed h,d. For the infinite horizon case, as shown above, we know operator $\mathscr{B}$ keeps the monotonicity property. As a result in each iteration:

$$V^{n+1} = \mathscr{B}V^n,$$

monotonicity of value function get preserved. It is sufficient to choose V^0 in the monotone value functions and since monotonicity property get preserved under the limit we have:

$$V^* = \lim_{n\to\infty} \mathscr{B}^n V,$$

is also monotone. □

***Lemma* 4.3.2.** *$V^*(n,h,d)$ is weakly decreasing in $|h|$.*

Proof. We use Corollary 3.2.1 to prove this property. We start with the first and last conditions. For fixed n, and d, the cost function is decreasing in h, moreover the final cost can assume to be constant or decreasing in $|h|$. Since $\{H_n\}$ is stochastically monotone, we satisfy the second condition on theorem 3.2.1. As a result V^* is decreasing in $|h|$. In the infinite horizon case, we know if similar conditions are satisfied then the Bellman operator preserves the monotonicity property. As a result, if we start by an initial value function V^0 which is decreasing in $|h|$, since the Bellman operator preserves the monotonicity property and monotonicity get preserved under the limit, we know:

$$V^* = \lim_{n\to\infty} \mathscr{B}^n V,$$

also is decreasing in $|h|$. □

***Lemma* 4.3.3.** *If the arrival process, $\{D_n\}$ is stochastically monotone, then V^* is non-decreasing in d for all $n \in \mathcal{N}$ and all $h \in \mathcal{H}$.*

Proof. First we prove the lemma for the finite horizon case. We use theorem 3.2.1. Notice that there is no dependencies to d in the cost terms, Suppose the the final cost function is monotone decreasing in d. Process $\{D_n\}$ being stochastically monotone is sufficient to satisfy the second condition in theorem 3.2.1. As a result, V^* is decreasing in d. The extension to infinite horizon case also follows exactly similar to the proof of the previous lemma. □

So far, using the results in the earlier chapters we have shown that optimal value function is an increasing function in state of the queue and is a decreasing function in the fading level of the channel. This result is matching with our intuition that less number of buffered packets in the queue is more desirable while transmitter prefers channel with lower levels of fading. In the following section, we try to extend these results more, and try to find the structural properties of the optimal policy. The proof of monotonicity of the optimal policy in the state of the queue is following from convexity of the value function. In the last result of this section, we bring the proof of convexity of the value function as the state of the queue. First we review definition and some properties of convex functions.

***Definition* 4.3.1.** A function $f : \mathcal{S} \to \mathbb{R}$ is defined to be *convex* if for all $s \in \{1, \ldots, L-1\}$,

$$f(s+1) - f(s-1) \geq 2f(s).$$

***Lemma* 4.3.4.** *Let $f : \mathcal{S} \to \mathbb{R}$ be a function defined on $S = \{0, \ldots, L\}$; then following statements are equivalent:*

1. *f is convex.*
2. *$f(s+1) - f(s)$ is non-decreasing in s.*
3. *$f(s) + f(t) \geq f(\left\lceil \frac{s+t}{2} \right\rceil) + f(\left\lfloor \frac{s+t}{2} \right\rfloor)$ for all $s, t \in \mathcal{S}$.*

Proof. Assume 1 is true, then by the definition and reordering we get:

$$f(s+1) - f(s) \geq f(s) - f(s-1) \quad \forall s \in \{0, \ldots, L\}.$$

Which is equivalent to the second statement. We prove 3 by assuming 2 is true and induction. 3 is obviously true with equality if $|s-t| \leq 1$. Now we use an induction argument to prove this for any arbitrary integer d. By induction hypothesis, assume 3 holds for all $s,t \in \mathcal{S}$, when $|s-t| \leq d$, and $d \in \{1,\ldots,L\}$, we then show that 3 must be true for any $s,t \in \mathcal{S}$ where $|s-t| = d+1$. Assume $s',t' \in \mathcal{S}$, and $s' = t' + (d+1)$, then from 2, we have:

$$f(s') - f(s'-1) \geq f(t'+1) - f(t'),$$

and thus,

$$f(s') + f(t') \geq f(t'+1) + f(s'+1).$$

Now since $|s'-1-(t'+1)| = d-1$ and thus by induction hypothesis we have:

$$f(t'+1) - f(s'-1) \geq f(\left\lceil \frac{s'+t'}{2} \right\rceil) + f(\left\lfloor \frac{s'+t'}{2} \right\rfloor).$$

□

***Lemma* 4.3.5.** *If $L = \infty$, then for all $h \in \mathcal{H}, d \in \mathcal{D}$, $V^*(n,h,d)$ is convex in n.*

Proof. We first prove that convexity is preserved through Bellman operation and consequently we prove value function is convex in n, for the finite and infinite horizon cases. Suppose the value function at iteration n is convex. Also suppose the optimal action at state $n+1$, is u^*_{n+1}, and the optimal action at state $n-1$ is u^*_{n-1}. Then we have the following:

$$V_{n+1}(n+1,h,d) + V_{n+1}(n-1,h,d) = P(h,u^*_{n+1}) + P(h,u^*_{n-1}) + \lambda.c(n+1) + \lambda.c(n-1) + \beta. \sum_{h' \in \mathcal{H}, d' \in \mathcal{D}} P_{d,d'} P_{h,h'} \Big(V_n([n+1-u^*_{n+1}+d']_L, h', d')) + V_n(([n-1-u^*_{n-1}+d']_L, h', d')) \Big).$$

Now we show the convexity term by term. Since the power function is convex, we have:

$$P(h,u^*_{n+1}) + P(h,u^*_{n-1}) \geq P(h, \left\lceil \frac{u^*_{n+1}+u^*_{n-1}}{2} \right\rceil) + P(h, \left\lfloor \frac{u^*_{n+1}+u^*_{n-1}}{2} \right\rfloor).$$

Also we have the convexity of the delay function:

$$\lambda.c(n+1) + \lambda.c(n-1) \geq 2\lambda c(n).$$

And since we assume that value function at time n is convex, we have the following:

$$\beta. \sum_{h' \in \mathcal{H}, d' \in \mathcal{D}} P_{d,d'} P_{h,h'} \Big(V_n([n+1-u^*_{n+1}+d']_L, h', d')) + V_n(([n-1-u^*_{n-1}+d']_L, h', d'))\Big) \geq$$
$$\beta. \sum_{g \in \mathcal{G}, a \in \mathcal{A}} P_{d,d'} P_{h,h'} \Bigg(V_n\Big(\Big\lceil \frac{[n+1-u^*_{n+1}+d']_L + [n-1-u^*_{n-1}+d']_L}{2} \Big\rceil, h', d'\Big) +$$
$$V_n\Big(\Big\lfloor \frac{[n+1-u^*_{n+1}+d']_L + [n-1-u^*_{n-1}+d']_L}{2} \Big\rfloor, h', d'\Big)\Bigg).$$

Notice that since we have assumed that $L = \infty$, we have:

$$\Big\lfloor \frac{[n+1-u^*_{n+1}+d']_L + [n-1-u^*_{n-1}+d']_L}{2} \Big\rfloor = n + d' - \Big\lfloor \frac{u^*_{n+1}+u^*_{n-1}}{2} \Big\rfloor.$$

And also we have:

$$\Big\lceil \frac{[n+1-u^*_{n+1}+d']_L + [n-1-u^*_{n-1}+d']_L}{2} \Big\rceil = n + d' - \Big\lceil \frac{u^*_{n+1}+u^*_{n-1}}{2} \Big\rceil.$$

As a result of this we get:

$$V_{n+1}(n+1,h,d) + V_{n+1}(n-1,h,d) \geq$$
$$P(h, \Big\lceil \frac{u^*_{n+1}+u^*_{n-1}}{2} \Big\rceil) + P(h, \Big\lfloor \frac{u^*_{n+1}+u^*_{n-1}}{2} \Big\rfloor) + 2\lambda c(n) + \beta. \sum_{g \in \mathcal{G}, a \in \mathcal{A}} P_{d,d'} P_{h,h'}$$
$$\Big(V_n\Big(n + d' - \Big\lceil \frac{u^*_{n+1}+u^*_{n-1}}{2} \Big\rceil, h', d'\Big) + V_n\Big(n + d' - \Big\lfloor \frac{u^*_{n+1}+u^*_{n-1}}{2} \Big\rfloor, h', d'\Big)\Big) \geq$$
$$2.V_n(n,h,d).$$

□

This completes the proof and shows that Bellman operator preserves the convexity of value function. Now notice that since Bellman operator preserves the convexity property, in finite horizon setup if we start with a convex final cost and in infinite horizon if we start by a convex initialization of the value function in value iteration algorithm, we can reason that value function in general is convex. Convexity of value function can help us prove submodularity of value-action function. This fact is stated in the following lemma:

***Lemma* 4.3.6.** *Suppose function $V(n,s)$ is convex in n, then the function $V(n-u,s)$ is submodular in arguments (n,u).*

Proof. Suppose $n^+ \geq n^-$ and $u^+ \geq u^-$, for $V(n-u)$ to be submodular we should have:

$$V(n^+ - u^+) + V(n^- - u^-) \leq V(n^+ - u^-) + V(n^- - u^+).$$

Which is equivalent to:

$$V(n^- - u^-) - V(n^- - u^+) \leq V(n^+ - u^-) - V(n^+ - u^-).$$

The Quantity $(n^+ - u^+) - (n^+ - u^-) = (n^- - u^+) - (n^- - u^-)$ is fixed, $V(n-u^-) - V(n-u^+) \geq 0$, and $V(\cdot, s)$ is convex and increasing in n, then by the lemma 4.3.4 part 2, we get the last inequality. As a result, $V(n-u)$ is submodular in (n,u). □

We can extend the submodularity in previous lemma to action-value function. This fact is stated in the following lemma:

***Lemma* 4.3.7.** *Action-value function denoted by $Q(n,u,h)$ is submodular in (n,u) for fixed h.*

Proof. We start by denoting the action-value function as following:

$$Q(n,h,u) = P(h,u) + \lambda.c(n) + \beta. \sum_{h' \in \mathcal{H}, d' \in \mathcal{D}} P_{d,d'} P_{h,h'} \left(V_n([n+1-u^*_{n+1}+d']_L, h', d') \right).$$

Notice that the summation term is submodular by the previous lemma and the fact that submodularity get preserved under summation and weightings. Now consider the cost part. We want to show:

$$Q(n^+, h, u^+) + Q(n^-, h, u^-) \leq Q(n^+, h, u^-) + Q(n^-, h, u^+).$$

Notice that the cost part in submodularity definition is equal in both sides.

$$P(h,u^+) + \lambda.c(n^+) + P(h,u^-) + \lambda.c(n^-) = P(h,u^-) + \lambda.c(n^+) + P(h,u^+) + \lambda.c(n^-).$$

Also note that we have proved submodularity in $V(n,u)$, and as a result, the term

$$\sum_{h'\in\mathcal{H},d'\in\mathcal{D}} P_{d,d'}P_{h,h'}\left(V_n([n+1-u^*_{n+1}+d']_L,h',d')\right)$$

is also submodular. Hence $Q(n,h,u)$ is submodular in (n,u). □

4.4 Structural properties of the Optimal policies

In this section we try to use the results in previous section along with techniques in previous chapter to prove the monotonicity of the optimal policy for this model. We start by proving the monotonicity of optimal policy as a function of number of packets in the queue.

***Theorem* 4.4.1.** *Assuming that buffer has infinite capacity, there exists an optimal policy $f(n,h,d)$ which is increasing in n, for fixed $d\in\mathcal{D}$, $h\in\mathcal{H}$.*

Proof. We use lemma 4.3.7, to prove the monotonicity of the value function. Notice that we have shown $Q(n,u,h)$ is submodular function in (n,u), for fixed h, as a result of corollary 3.2.2 we infer there exists an optimal policy which is increasing n. □

***Theorem* 4.4.2.** *There exists an optimal policy which is increasing in h.*

Proof. In order to prove this we have to show that $Q(n,h,u)$ is submodular in (h,u) for fixed n. We start by writing the definition fo submodularity and expanding the action-value function.

$$Q(n,h^+,u^+)+Q(n,h^-,u^-)\leq Q(n,h^+,u^-)+Q(n,h^-,u^+).$$

We write the cost term of action-value function at the beginning:

$$P(h^+,u^+)+\lambda.c(n)+P(h^-,u^-)+\lambda.c(n)\leq P(h^+,u^-)+\lambda.c(n)+P(h^-,u^+)+\lambda.c(n)\Leftrightarrow$$
$$P(h^+,u^+)+P(h^-,u^-)\leq P(h^-,u^+)+P(h^-,u^+).$$

Remember that we had:

$$P(h,u)=\frac{\tilde{P}(u)}{|h|^2}.$$

Hence we have:

$$\frac{\tilde{P}(u^+)}{|h^+|^2} + \frac{\tilde{P}(u^-)}{|h^-|^2} \leq \frac{\tilde{P}(u^+)}{|h^-|^2} + \frac{\tilde{P}(u^-)}{|h^+|^2} \Leftrightarrow$$
$$\frac{\tilde{P}(u^+)}{|h^+|^2} - \frac{\tilde{P}(u^-)}{|h^+|^2} \leq \frac{\tilde{P}(u^+)}{|h^-|^2} - \frac{\tilde{P}(u^-)}{|h^-|^2}.$$

The last inequality is true due to convexity of $\tilde{P}$ function. Now notice that state variable is not a part of value function $V(n, h, u)$, as a result, value function $V(n, h, u)$ satisfies the submodularity inequality with equality, and hence based on corollary 3.2.2, we get the optimal policy is increasing in $|h|$. □

4.5 Conclusion

This chapter investigates the qualitative properties of the optimal value and policy corresponding to a simple cross layer design model. This is one of the earliest applications of Markov Decision theory techniques in cross layer design problems. In the following chapters, we investigated more complicated scenarios in cross layer designs using the same tools.

Chapter 5

Optimal Transmission Strategy for Bursty Traffic and Adaptive Decision Feedback

5.1 Introduction

In this section, we generalize the model in the previous chapter to the scenarios with ACK-/NACK feedback channels. The presence of feedback channel can reduce the probability of error at the cost of retransmission of data. Receiver has the option to choose retransmission when the probability of error is high; however, this retransmission increases both the power consumption and the incurred delay by other packets in the queue. In order to rigorously depict this trade-off we formulate the problem as MDP model and try to find qualitative properties of optimal policies. This framework tries to integrate three concepts of queuing delay, transmission power, and decoding error probability. At each transmission epoch, receiver has the option to ask for a retransmission from the transmitter. This decision is made based on a power threshold for declaring an erasure. When the threshold is chosen to by zero, receiver never declares an erasure and the feedback channel will never be used. Choosing a larger threshold improves the error exponent; however, the power consumption and delay will be increased.

5.2 Model and Problem Formulation

5.2.1 Modeling assumptions

The model and problem formulation is due to [4]. Consider a communicatio system depicted in Figure 5.1. The higher layer of network is modeled with a source which generates bursty traffic. These packets are stored in the queue. The transmitter chooses the rate and encodes the message at the beginning of an epoch. Receiver, decodes the message at the end of the epoch and uses the feedback channel if it is needed. The frequency of transmissions are chosen such that state of the channel is assumed to by constant during each epoch. Channel is assumed to be block fading channel, as a result, in each decision epoch transmitter is dealing with a fixed fading level. Number of existing packets in the queue along with current state of the channel is assumed to be known for both of the transmitter and receiver.

At each transmission epoch, transmitter chooses the transmission rate based on current queue size and state of the channel. Choosing the rate, it encodes the message and send it over the fading channel. At the end of decision epoch, receiver decodes the received message, and based on the channel state and queue size, decides on utilizing the decision feedback. This decision is being made based on a power threshold function, queue state and channel state. If the receiver has asked for retransmission, transmitted packets stay in the queue to be transmitted in future transmission epochs, otherwise, they will be removed from the queue.

There are three control variables for the transmitter to choose. The number of bits to transmit (or equivalently, the transmission rate), the transmission power , and the threshold for declaring an erasure. There are also three performance metrics which are of interest for the designer of the system. Average transmitted power, average delay, and the average probability of block error. In our model, we assume that we fix number of transmitted bits and erasure threshold, then the transmitted power is chosen such that the probability of block error is less than a certain value. Using this setup, we restrict our attention to choosing two control variables, number of bits to transmit and erasure threshold. The performance metrics that we consider is a Lagrangian multiplier of average transmitted power and the average delay. We now describe the notation used in the rest of the chapter.

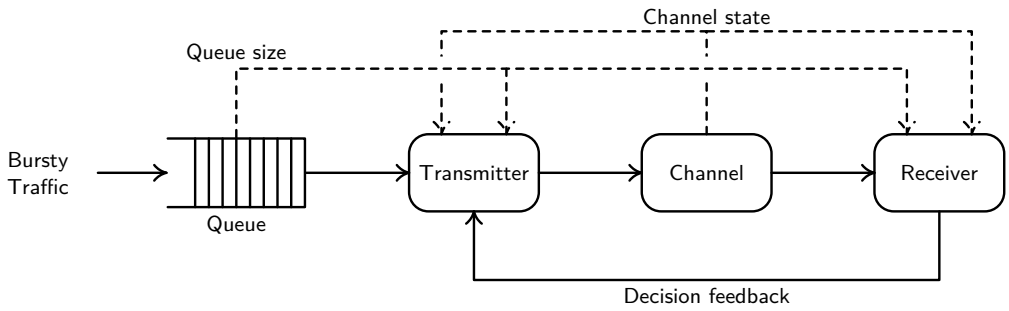

Fig. 5.1 Model of a transmitter with decision feedback

- $A_n \in \mathbb{N} = \{0, 1, \dots, \}$ denotes number of arriving packets at epoch n.
- $S_n \in \mathcal{S}$ denotes the state of the channel during slot n. We assume that $\mathcal{S}$ is finite and a lower value of s implies a better channel state *e.g.*, S_n may be taken to be the reciprocal of the fading gain.
- $Q_n \in \mathbb{N}$ denotes the number of bits in the queue at the beginning of slot n.
- $U_n \in \mathbb{N}$ denotes the number of bits transmitted in slot n. This is one of the control variables and it should satisfy $U_n \leq Q_n$. We also denote the set of feasible number of bits to transmit by $F(q)$.
- $T_n \in \mathcal{T}$ denotes the erasure threshold at the end of slot n. We assume that $\mathcal{T}$ is either a convex or a finite subset of $\mathbb{R}^+$. In both cases $0 \in \mathcal{T}$.

The delay at time n depends on the queue size at the end of slot n. We assume that the delay is of the form $d(Q_{n+1} - A_n)$, where $d(q)$ is an increasing and convex function of q.

The transmitted power at each time epoch is a function of transmitted bits U_n, the erasure threshold T_n, and the channel state S_n. We assume the transmitted power is of the form $\pi(U_n, T_n)h(S_n)$ where $\pi(u, t)$ is strictly increasing in u, strictly decreasing in t, and $h(s)$ is strictly increasing and convex in s. We denote probability of error by $\mathfrak{E}(T_n)$. Note that by fixing the erasure threshold t, the erasure probability will be determined by $\mathfrak{E}(t)$. We assume $\mathfrak{E}(t)$ is increasing and convex in t.

5.2.2 Problem formulation

First we introduce the dynamics of the queue. Suppose the initial state of the queue Q_0 is set to 0. The random arrival at epoch 0 is denoted by A_0, as a result, $Q_1 = A_0$. From $n = 1$ onward,

$$Q_{n+1} = \begin{cases} Q_n - U_n + A_{n+1}, & \text{with prob. } (1 - \mathfrak{E}(T_n)); \\ Q_n + A_{n+1} & \text{with prob. } \mathfrak{E}(T_n). \end{cases}$$

Furthermore, we impose following assumptions on the fading and arriving processes.

- $\{A\}_n$ and $\{S\}_n$ are independent and identically distributed.
- $\{A\}_n \sim P_A$ and $\{S\}_n \sim P_S$
- $\{A\}_n$ and $\{S\}_n$ are independent from each other.
- PMF of arriving process $\{A\}_n$ is convex and decreasing.

We choose control actions U_n and T_n according to transmission policy $g_n(.,.)$ and feedback policy $f_n(.,.)$ where:

$$U_n = g_n(Q_n, S_n) \quad \text{and} \quad T_n = f_n(Q_n, S_n).$$

We define a policy to be feasible, if both transmission and feedback policies satisfy following set of constraints:

$$\begin{aligned} g_n(q, s) &\in F(q), \quad \forall q \in N, s \in S, \\ f_n(q, s) &\in \mathcal{T}, \qquad \forall q \in N, s \in S. \end{aligned}$$

We impose a terminal cost on the model $d_N(q, s)$. We assume $d_N(q, s)$ is increasing in q, for fixed s, and constant function of s for each q. At the end of each slot, two types of costs are incurred. We define a cost due to transmitted power, which is given by

$$\pi(U_t, T_n)h(S_n),$$

and a cost due to delay, which is given by

$$\begin{cases} d(Q_n - U_n), & \text{with prob. } (1 - \mathfrak{E}(T_n)), \\ d(Q_n), & \text{with prob. } \mathfrak{E}(T_n). \end{cases}$$

We further assume function $\pi(\cdot,\cdot)$ is an increasing function in u and decreasing function in t. Probability function $\mathfrak{E}(t)$ is an increasing function of threshold t, and delay function $d(\cdot)$ is a convex increasing function of its argument. We are interested in two optimization problems. The first problem is to choose feasible transmitting and feedback policies to minimize

$$\sum_{n=1}^{N} \mathbb{E}\Big[\pi(U_n, T_n) h(S_n)\Big]$$

$$\text{such that} \quad \sum_{n=1}^{N-1} \mathbb{E}\{d(Q_{n+1} - A_n)\} + d_N(Q_N) \leq D.$$

The second problem is the dual of the first.

Both these problems are constrained optimization problems. As such, they can be solved by considering a Lagrangian relaxation [5]. We are interested in deriving structural properties of optimal policies for both problems. For this, we consider a combined Lagrangian relaxation of both problems, *i.e.*, choose transmission and feedback policies to minimize

$$\sum_{n=1}^{N-1} \mathbb{E}\Big[\lambda_1 \pi(U_n, T_n) c(S_n) + \lambda_2 d(Q_{n+1} - A_n)\Big] + \lambda_2 d_N(Q_N),$$

where λ_1 and λ_2 are positive constants. We define following optimization problem on the model:

Problem 2. *Choose feasible transmission and feedback policies to minimize*

$$\sum_{n=1}^{N-1} \mathbb{E}\Big[\lambda_1 \pi(U_n, T_n) c(S_n) + \lambda_2 d(Q_{n+1} - A_n)\Big] + \lambda_2 d_N(Q_N),$$

where λ_1 and λ_2 are pre-specified positive constants.

5.3 Dynamic programming decomposition

Problem 2 is a Markov decision process. As a result, we can formulate the value function at each state $q \in \mathbb{N}$ and $s \in S$ as following :

$$V_N(q,s) = d_N(q)$$

For $n = N-1, N-2, \ldots, 1$, recursively define

$$V_n(q,s) = \min_{u \in F(q), t \in T} \left[\mathbb{E}^{f,g}_{Q,S} \left[c(q,u,s,t) + V_{n+1}(Q,S) \middle| Q_n = q, S_n = s \right] \right],$$

.

where by $\mathbb{E}^{f,g}_{a,s}$ we mean the expectation on joint probability measure on (Q,S) when policies (f,g) are chosen. By $c(q,u,s,t)$, we mean per step cost. Notice that we can simplify this term as following:

$$c(q,u,s,t) = \lambda_1 \pi(u,t) h(s) + \lambda_2 \mathfrak{E}(t) d(q) + \lambda_2 (1 - \mathfrak{E}(t)) d(q-u).$$

Following simplification is also possible for value function:

$$\mathbb{E}^{f,g}_{Q,S} \left[V_{n+1}(Q,S) \right] = \mathbb{E}^{f,g}_{A,S} \left[[1 - \mathfrak{E}(t)] V_{n+1}(q - u + A', S') + \mathfrak{E}(t) V_{n+1}(q + A', S') \right] \Big\}.$$

To derive structural properties of the optimal value function and optimal policy, we use the standard form of value function.

We derive the transition probability between states:

$$\Pr(Q_{n+1} = j, S_{n+1} = z | Q_n = q, U_n = u, S_n = s, T_n = t, A_n = a).$$

Since $\{A_n\}$ and $\{S_n\}$ processes are independent processes, and state of the channel is an uncontrolled process, we can deduce:

$$\begin{aligned} &\Pr(Q_{n+1} = j, S_{n+1} = z | Q_n = q, U_n = u, S_n = s, T_n = t, A_n = a) \\ =&\Pr(S_{n+1} = z | S_n = s). \Pr(Q_{n+1} = j | Q_n = q, U_n = u, T_n = t, A_n = a) \end{aligned}$$

We find an explicit expression for the term $p(Q_{n+1} = j|Q_n = q, U_n = u, T_n = t, A_n = a)$.

Based on the dynamics of the queue there are two possible scenarios for the state of the queue. With probability $\mathfrak{E}(t)$, next state of the queue is

$$j = A_{n+1} + q,$$

and with probability $1 - \mathfrak{E}(t)$, next state of the queue is

$$j = q - u + A_{n+1}.$$

As a result we can find the probability of these two events from the arrival process as following:

$$\begin{aligned}
&\Pr(Q_{n+1} = j|Q_n = q, U_n = u, S_n = s, T_n = t, A_n = a) \\
=&\Pr(Q_{n+1} = j|Q_n = q, U_n = u, T_n = t, A_n = a) \\
=&\Pr(A_{n+1} = j - q).\mathfrak{E}(t) + \Pr(A_{n+1} = j - q + u).(1 - \mathfrak{E}(t)).
\end{aligned}$$

As a result, we can derive the state transition probability as following:

$$\begin{aligned}
&\Pr(Q_{n+1} = j, S_{n+1} = z|Q_n = q, U_n = u, S_n = s, T_n = t, A_n = a) \\
=&\Pr(S_{n+1} = z|S_n = s).\Pr(Q_{n+1} = j|Q_n = q, U_n = u, T_n = t, A_n = a) \\
=&\Pr(S_{n+1} = z|S_n = s).\Big[\Pr(A_{n+1} = j - q).\mathfrak{E}(t) + p(A_{n+1} = j - q + u).(1 - \mathfrak{E}(t))\Big].
\end{aligned}$$

Using this expression, we can define the reverse CDF function. For $q \in F(q)$, and $k \in \mathbb{N}$, we define reverse CDF function for the state q as following:

$$q_1(k \mid q, u, s, t) = \sum_{q' \geq k} \Pr(q' \mid q, u, s, t).$$

We can further simplify this expression:

$$\begin{aligned}
q_1(k \mid q, u, s, t) &= \sum_{j=k}^{\infty} \Big[\Pr(A_{n+1} = j - q).\mathfrak{E}(t) + \Pr(A_{n+1} = j - q + u).(1 - \mathfrak{E}(t))\Big] \\
=&\mathfrak{E}(t)\Big[Pr[A_{n+1} \in [k - q, \infty)]\Big] + (1 - \mathfrak{E}(t))\Big[Pr[A_{n+1} \in [k - q + u, \infty)]\Big].
\end{aligned}$$

In the following section, we utilize these expressions to prove properties of cost and value function terms. Similarly we define the reverse CDF function for the state of the channel as following:

$$q_2(k \mid q, u, s, t) = \sum_{s' \geq k} \Pr(s' \mid q, u, s, t).$$

Since state of the channel is an uncontrolled independent process we can deduce:

$$q_2(k \mid q, u, s, t) = \sum_{s' \geq k} \Pr(s' \mid q, u, s, t) = \sum_{s' \geq k} \Pr(s' \mid s) = \Pr(s' \in [k, \infty)).$$

5.4 Properties of the Cost and Reverse CDF Function

In this section, we try to summarize properties of the cost term and expectation term of the value function. Following properties hold for the cost and reverse CDF function:

1. $d_N(q, s)$ is increasing function of q for fixed s.

   Proof. It follows directly from the assumptions on the model. □

2. $d_N(q, s)$ is increasing in s for fixed q.

   Proof. $d_N(q, s)$ is a constant function of s, and hence weakly increasing. □

3. $c(q, u, t, s)$ is increasing in q for fixed u, t, s.

   Proof. Recall that we have :

   $$c(q, u, s, t) = \lambda_1 \pi(u, t) h(s) + \lambda_2 \mathfrak{E}(t) d(q) + \lambda_2 (1 - \mathfrak{E}(t)) d(q - u).$$

   As a result, since $d(q-u)$ and $d(q)$ are increasing functions of q, $c(q, u, s, t)$, is increasing in q. □

4. $c(q, u, t, s)$ is increasing in s for fixed q, u, t.

   Proof. By assumptions, we know $h(s)$ is increasing in s, we infer for fixed u, t and q, $c(q, u, t, s)$ is increasing in s. □

5. $q_1(k \mid q, u, s, t)$ is increasing in q for fixed u, s, t and all $k \in \mathbb{N}$.

   Proof. Recall that we have:

   $$q_1(k \mid q, u, s, t) = \mathfrak{E}(t)(Pr[A_{n+1} \in [k-q, \infty)]) + (1 - \mathfrak{E}(t))(Pr[A_{n+1} \in [k-q+u, \infty)]).$$

   It is immediate that this expression is an increasing function of variable q, when u, s, t are fixed. □

6. $q_2(k \mid q, u, s, t)$ is increasing in s for fixed u, q, t and all $k \in S$.

   Proof. Recall:

   $$q_2(k \mid q, u, s, t) = \Pr(s' \in [k, \infty)).$$

   Since $\{S_n\}$ is an $i.i.d$ process, $q_2(k \mid q, u, s, t)$ is a constant function of s and hence increasing. □

   From this property onwards, we show joint behavior of cost function on pair of states and actions.

7. $c(q, u, t, s)$ is a restricted submodular function on $Q \times U$.

   Proof. We have:

   $$c(q, u, s, t) = \lambda_1 \pi(u, t) h(s) + \lambda_2 \mathfrak{E}(t) d(q) + \lambda_2 (1 - \mathfrak{E}(t)) d(q - u).$$

   Notice that first and second terms do not have joint dependence on (q, u), as a result, it is sufficient to show the property for the term:

   $$\lambda_2 (1 - \mathfrak{E}(t)) d(q - u).$$

   Which is equivalent to showing $d(q - u)$ is submodular on (q, u) which follows from (4.3.6) since $d(.)$ is a convex increasing function. □

8. $c(q, u, t, s)$ is supermodular on $S \times U$ for fixed q, t.

Proof. In order to prove $c(q,u,t,s)$ is supermodular, for arbitrary states s^+, s^- such that $s^+ \geq s^-$, and arbitrary actions u^+, u^- such that $u^+ \geq u^-$, we have to show the following:

$$c(q,u^+,t,s^+) + c(q,u^-,t,s^-) \geq c(q,u^-,t,s^+) + c(q,u^+,t,s^-).$$

Consider $c(q,u,s,t)$, since variables s, u do not jointly exist in second and third term, we infer it is sufficient to prove suppermodularity only for the first term. For the first term we have:

$$\begin{aligned}
\lambda_1\pi(u^+,t)h(s^+) + \lambda_1\pi(u^-,t)h(s^-) &\geq \lambda_1\pi(u^+,t)h(s^-) + \lambda_1\pi(u^-,t)h(s^+) \\
\Leftrightarrow \lambda_1\pi(u^+,t)(h(s^+) - h(s^-)) &\geq \lambda_1\pi(u^-,t)(h(s^+) - h(s^-)) \\
\overset{a}{\Leftrightarrow} \lambda_1\pi(u^+,t) &\overset{b}{\geq} \lambda_1\pi(u^-,t),
\end{aligned}$$

where (a) follows from the fact $h(s^+) \geq h(s^-)$. If $h(s^+) = h(s^-)$ the inequality holds by equality, otherwise, we can cancel the term $\Big(h(s^+) - h(s^-)\Big)$ from both sides. Moreover, (b) follows from the assumption that $h(s)$ is an increasing function of s. □

9. $c(q,u,t,s)$ is supermodular on $Q \times T$ for fixed u, s.

Proof.

$$\begin{aligned}
c(q,u,s,t) =& \lambda_1\pi(u,t)h(s) + \lambda_2\mathfrak{E}(t)d(q) + \lambda_2(1-\mathfrak{E}(t))d(q-u) && (5.1)\\
=& \lambda_1\pi(u,t)h(s) + \lambda_2 d(q-u) + \lambda_2\mathfrak{E}(t)(d(q) - d(q-u)) && (5.2)
\end{aligned}$$

The first term does not depend on q and the second term does not depend on t. Thus both are trivially submodular in (q,t). We now check the submodularity property for the 3^{rd} term.

It is obvious that for the first and second term the statement is true. Hence, in order to

prove the supermodularity property it is sufficient to prove it for the third term.

$$\begin{aligned}
&\lambda_2\mathfrak{E}(t^+)(d(q^+)-d(q^+-u))+\lambda_2\mathfrak{E}(t^-)(d(q^-)-d(q^--u))\\
\geq&\lambda_2\mathfrak{E}(t^+)(d(q^-)-d(q^--u))+\lambda_2\mathfrak{E}(t^-)(d(q^+)-d(q^+-u))\\
\Leftrightarrow&\lambda_2\mathfrak{E}(t^+)\Big[d(q^+)-d(q^-)-(d(q^+-u)-d(q^--u))\Big]\\
\geq&\lambda_2\mathfrak{E}(t^-)\Big[d(q^+)-d(q^-)-(d(q^+-u)-d(q^--u))\Big]\overset{a}{\Leftrightarrow}\lambda_2\mathfrak{E}(t^+)\overset{b}{\geq}\lambda_2\mathfrak{E}(t^-),
\end{aligned}$$

where (a) follows from the fact that $d(.)$ is an increasing convex function and as a result $d(q^+)-d(q^-)-(d(q^+-u)-d(q^--u))$ is non-negative and (b) follows from the fact that $\mathfrak{E}(t)$ is an increasing function of t. □

10. $c(q,u,t,s)$ is a submodular function on $S\times T$ for fixed q,u.

Proof. Consider the expression for $c(q,u,t,s)$ given in 5.1, since joint expression of (s,t) exists only in first term, trivially it is sufficient to show the submodularity in the first term. By the definition of submodularity, we have:

$$\begin{aligned}
&\lambda_1\pi(u,t^+)h(s^+)+\lambda_1\pi(u,t^-)h(s^-)\leq\lambda_1\pi(u,t^+)h(s^-)+\lambda_1\pi(u,t^-)h(s^+)\\
\Leftrightarrow\ &h(s^-)(\lambda_1\pi(u,t^-)-\lambda_1\pi(u,t^+))\leq h(s^+)(\lambda_1\pi(u,t^-)-\lambda_1\pi(u,t^+))\\
\overset{a}{\Leftrightarrow}\ &h(s^-)\overset{b}{\leq}h(s^+),
\end{aligned}$$

where (a) follows from the fact that $\pi(u,t)$ is decreasing in t and (b) follows from the fact that $h(\cdot)$ is an increasing function. □

Following properties try to show the joint behavior of $q(k|q,u,s,t)$ as a function of control and state variables.

11. $q_1(k\mid q,u,s,t)$ is supermodular function on $Q\times T$ for fixed u,s and for all $k\in Q$.

Proof. we showed before that

$$q_1(k \mid q,u,s,t,a) = \sum_{j=k}^{\infty}\Big[\Pr(A_{n+1}=j-q).\mathfrak{E}(t) + \Pr(A_{n+1}=j-q+u).(1-\mathfrak{E}(t))\Big]$$
$$=\mathfrak{E}(t)\Big[\sum_{j=k}^{\infty}\Pr(A_{n+1}=j-q) - \sum_{j=k}^{\infty}\Pr(A_{n+1}=j-q+u)\Big] + \sum_{j=k}^{\infty}\Pr(A_{n+1}=j-q+u)$$
$$=\mathfrak{E}(t)\Big[\sum_{j=k}^{k+u}\Pr(A_{n+1}=j-q)\Big] + \sum_{j=k}^{\infty}\Pr(A_{n+1}=j-q+u).$$

Since variable t is not in the second term, it is sufficient to prove the supermodularity condition for the first term. Now we write the supermodularity definition for the first term.

$$\mathfrak{E}(t^+)\Big[\sum_{j=k}^{k+u}\Pr(A_{n+1}=j-q^+)\Big] + \mathfrak{E}(t^-)\Big[\sum_{j=k}^{k+u}\Pr(A_{n+1}=j-q^-)\Big]$$
$$\geq \mathfrak{E}(t^+)\Big[\sum_{j=k}^{k+u}\Pr(A_{n+1}=j-q^-)\Big] + \mathfrak{E}(t^-)\Big[\sum_{j=k}^{k+u}\mathfrak{E}_t(A_{n+1}=j-q^+)\Big]$$
$$\Leftrightarrow [\mathfrak{E}(t^+)-\mathfrak{E}(t^-)]\Big[\sum_{j=k}^{k+u}\Pr(A_{n+1}=j-q^+)\Big] \geq [\mathfrak{E}(t^+)-\mathfrak{E}(t^-)]\Big[\sum_{j=k}^{k+u}\Pr(A_{n+1}=j-q^-)\Big]$$
$$\overset{a}{\Leftrightarrow} \sum_{j=k}^{k+u}p_t(A_{n+1}=j-q^+) \overset{b}{\geq} \sum_{j=k}^{k+u}p_t(A_{n+1}=j-q^-),$$

where (a) follows from the fact that $\mathfrak{E}(t)$ is increasing in t and (b) follows from the assumption that arriving pdf $P(A_{n+1})$ is a non-increasing function. □

12. $q_2(k \mid q,u,s,t)$ is submodular on $S \times T$ for all $k \in S$ for fixed q,u.

Proof. We showed $q_2(k \mid q,u,s,t) = \Pr(S_{n+1} \in [k,\infty) \mid S_n = s)$ since there is no t in the aforementioned statement, it is obvious that definition of submodularity inequality holds with equality. □

13. For the model presented we have the following property:
Property: For arbitrary queue states q^+ and q^-, any control action $u^+ \in \{U_{q^+} \setminus U_{q^-}\}$ there exists at least an action $u^- \in U_{q^-}$ for which we have:

(i) $c(q^+, u^+, t, s) \geq c(q^-, u^-, t, s)$ for fixed s, t.

(ii) $q_1(k \mid q^+, u^+, s, t) \geq q_1(k \mid q^-, u^-, s, t)$ for fixed s, t and for all $k \in Q$.

Proof. In order to prove this, we give a method to construct u^- from known variables q^+,u^+ and q^-. First we define $\delta^+ = q^+ - u^+$. Based on definition of δ^+, we consider two cases:

(a) if $q^- \geq \delta^+$, then we choose $u^- = q^- - \delta^+$.

(b) if $q^- < \delta^+$, then we choose arbitrary $u^- \in U_{q^-}$.

Recall:

$$c(q, u, s, t) = \lambda_1 \pi(u, t) h(s) + \lambda_2 \mathfrak{E}(t) d(q) + \lambda_2 (1 - \mathfrak{E}(t)) d(q - u).$$

Since $\pi(\cdot, \cdot)$ is an increasing function in u for fixed t and $d(.)$ is increasing function we get:

$$\lambda_1 \pi(u^+, t) h(s) + \lambda_2 \mathfrak{E}(t) d(q^+) \geq \lambda_1 \pi(u^-, t) h(s) + \lambda_2 \mathfrak{E}(t) d(q^-).$$

As a result, the first and second terms clearly satisfy the theorem. For the case (a), third terms are equal and for the case (b) it is obvious that $d(q^+ - u^+) \geq d(q^- - u^-)$. As a result, we showed for any action $u^+ \in \{U_{q^+} \setminus U_{q^-}\}$, there exists an action $u^- \in U_{s^-}$ such that $c(q^+, u^+, t, s) \geq c(q^-, u^-, t, s)$.
In order to show the theorem for $q(k \mid q, u, s, t)$, recall :

$$q_1(k \mid q, u, s, t) = \mathfrak{E}(t)[\Pr(A_{n+1} \in [k - q, \infty)))] + (1 - \mathfrak{E}(t))[\Pr(A_{n+1} \in [k - q + u, \infty))].$$

Here again the first term is trivially satisfying the theorem and we should restrict our attention to the second term. Consider the term $Pr(A_{n+1} \in [k - \delta, \infty))$, in case (a) the second terms are equal. In (b), we have :

$$Pr(A_{n+1} \in [k - \delta^-, \infty)) < Pr(A_{n+1} \in [k - \delta^+, \infty)).$$

So we showed for any action $u^+ \in \{U_{q^+} \setminus U_{q^-}\}$ there exists at least an action $u^- \in U_{q^-}$ for which we have $q(k \mid q^+, u^+, s, t) \geq q(k \mid q^-, u^-, s, t)$. □

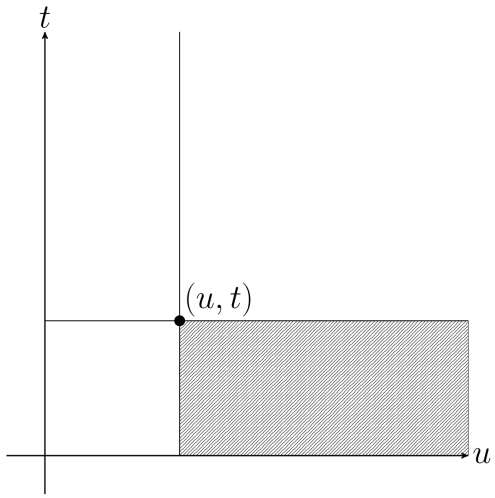

Fig. 5.2 The shaded are represent all the action variables $(u,t) \lesssim (u',t')$.

5.5 Main Results

In this section, we derive structure of optimal policy. Since the action space is multi-dimensional, we have to define a partial order on the action space and then use the results for multi-dimensional action spaces in chapter (2).

***Definition* 5.5.1.** We define partial order on $U \times T$ as follows $z_1 = (u_1, t_1) \preceq z_2 = (u_2, t_2)$ if we have two following properties :

1. $u_1 \leq u_2$
2. $t_1 \geq t_2$.y

Notice that this partial order induces a lattice on action space since two elements of $z_1 \vee z_2$ and $z_1 \wedge z_2$ exists in the set. Using this definition of partial order, we can establish following result:

***Theorem* 5.5.1.** *Cost function $c(q,u,s,t)$ has decreasing difference on $Q \times \{T \times U\}$. In other words, for $q^- \leq q^+$, and $z_1 \precsim z_2$ with defined partial order, we have:*

$$c(q^+, z_1, s) - c(q^-, z_1, s) \geq c(q^+, z_2, s) - c(q^-, z_2, s).$$

Proof. We can rewrite $c(q,u,s,t)$ as following:

$$\begin{aligned} c(q,u,s,t) &= \lambda_1\pi(u,t)h(s) + \lambda_2\mathfrak{E}(t)d(q) + \lambda_2(1-\mathfrak{E}(t))d(q-u) \\ &= \lambda_1\pi(u,t)h(s) + \lambda_2 d(q-u) + \lambda_2\mathfrak{E}(t)(d(q)-d(q-u)). \end{aligned}$$

Now we want to prove the following:

$$c(q^+,z_1,s) - c(q^-,z_1,s) \geq c(q^+,z_2,s) - c(q^-,z_2,s).$$

We derive the left hand side of the equation as following:

$$\begin{aligned} &h(s)(\pi(u_1,t_1) - \pi(u_1,t_1)) \\ &\quad + \mathfrak{E}(t_1)\Big[d(q^+) - d(q^+-u_1) - d(q^-) + d(q^--u_1)\Big] + d(q^+-u_1) - d(q^--u_1) \\ &= \mathfrak{E}(t_1)\Big[d(q^+) - d(q^+-u_1) - d(q^-) + d(q^--u_1)\Big] + d(q^+-u_1) - d(q^--u_1). \end{aligned}$$

Similarly for the right hand side we have the following statement:

$$\begin{aligned} &h(s)(\pi(u_2,t_2) - \pi(u_2,t_2)) \\ &\quad + \mathfrak{E}(t_2)\Big[d(q^+) - d(q^+-u_2) - d(q^-) + d(q^--u_2)\Big] + d(q^+-u_2) - d(q^--u_2) \\ &= \mathfrak{E}(t_2)\Big[d(q^+) - d(q^+-u_2) - d(q^-) + d(q^--u_2)\Big] + d(q^+-u_2) - d(q^--u_2). \end{aligned}$$

Now we define Δ and Γ as follows:

$$\begin{aligned} \Delta_1 &= d(q^+-u_1) - d(q^--u_1), \\ \Delta_2 &= d(q^+-u_2) - d(q^--u_2), \\ \Gamma &= d(q^+) - d(q^-). \end{aligned}$$

Hence for the left hand side we can write:

$$\mathfrak{E}(t_1)\Big[\Gamma - \Delta_1\Big] + \Delta_1 = \Delta_1(1-\mathfrak{E}(t_1)) + \mathfrak{E}(t_1)\Gamma.$$

And for the right hand side we can write:

$$\mathfrak{E}(t_2)\Big[\Gamma - \Delta_2\Big] + \Delta_2 = \Delta_2(1 - \mathfrak{E}(t_2)) + p(t_2)\Gamma.$$

Now we want to prove:

$$\Delta_1(1 - \mathfrak{E}(t_1)) + \mathfrak{E}(t_1)\Gamma - \Big[\Delta_2(1 - \mathfrak{E}(t_2)) + \mathfrak{E}(t_2)\Gamma\Big] \geq 0,$$

with knowing these facts:

$$\Delta_2 \leq \Delta_1 \leq \Gamma,$$
$$\mathfrak{E}(t_2) \leq \mathfrak{E}(t_1).$$

Where the first statement is directly comes from the fact that $d(.)$ is convex increasing function, and second one is from the fact that $t_1 \geq t_2$ and $\mathfrak{E}(\cdot)$ is an increasing function. First, we prove the following lemma:

$$\Big[\Delta_2(1 - \mathfrak{E}(t_2)) + \mathfrak{E}(t_2)\Gamma\Big] \leq \Big[\Delta_2(1 - \mathfrak{E}(t_1)) + \mathfrak{E}(t_1)\Gamma\Big]. \tag{5.3}$$

proof:

$$\begin{aligned}
&\Delta_2(1 - \mathfrak{E}(t_2)) + \mathfrak{E}(t_2)\Gamma \leq \Delta_2(1 - \mathfrak{E}(t_1)) + \mathfrak{E}(t_1)\Gamma \\
\Leftrightarrow\ &\Delta_2 - \Delta_2\mathfrak{E}(t_2) + \mathfrak{E}(t_2)\Gamma \leq \Delta_2 - \Delta_2\mathfrak{E}(t_1) + \mathfrak{E}(t_1)\Gamma \\
\Leftrightarrow\ &\mathfrak{E}(t_2)(\Gamma - \Delta_2) \leq \mathfrak{E}(t_1)(\Gamma - \Delta_2)
\end{aligned}$$

where since $\Gamma - \Delta_2$ is positive and $\mathfrak{E}(t_2) \leq \mathfrak{E}(t_1)$ last inequality becomes trivial.
Now by lemma we infer the following statement:

$$\begin{aligned}
&\Delta_1(1 - \mathfrak{E}(t_1)) + \mathfrak{E}(t_1)\Gamma - \{\Delta_2(1 - \mathfrak{E}(t_1)) + \mathfrak{E}(t_1)\Gamma\} \\
\leq&\Delta_1(1 - \mathfrak{E}(t_1)) + \mathfrak{E}(t_1)\Gamma - \{\Delta_2(1 - \mathfrak{E}(t_2)) + \mathfrak{E}(t_2)\Gamma\}.
\end{aligned}$$

now we show:

$$\begin{aligned}\Delta_1(1-\mathfrak{E}(t_1))+\mathfrak{E}(t_1)\Gamma-\{\Delta_2(1-\mathfrak{E}(t_1))+\mathfrak{E}(t_1)\Gamma\}\geq 0\\ \Leftrightarrow(\Delta_1-\Delta_2)+\Gamma(\mathfrak{E}(t_1)-\mathfrak{E}(t_1))+(\Delta_2-\Delta_1)\mathfrak{E}(t_1)\geq 0\\ \Leftrightarrow(\Delta_1-\Delta_2)(1-\mathfrak{E}(t_1))\geq 0.\end{aligned}$$

where the last inequality is obvious from the fact that $\mathfrak{E}(t_1)\leq 1$ and $\Delta_1\geq\Delta_2$.

□

***Theorem* 5.5.2.** *For the defined partial order on $U\times T$, reverse CDF function $q_1(k|q,s,z)$ has increasing difference on $Q\times(U\times T)$. In other words, for $q^-\leq q^+$, and $z_1\precsim z_2$ with defined partial order, we have:*

$$q_1(k\mid q^+,z_2,s)-q_1(k\mid q^-,z_2,s)\leq q_1(k\mid q^+,z_1,s)-q_1(k\mid q^-,z_1,s). \tag{5.4}$$

Proof. Recall:

$$q_1(k\mid q,u,s,t)=\mathfrak{E}(t)[\Pr(A_{n+1}\in[k-q,\infty)])+[1-\mathfrak{E}(t)][\Pr(A_{n+1}\in[k-q+u,\infty)]).$$

we can write the left hand side of equation 5.4 as following:

$$\begin{aligned}&q_1(k\mid q^+,z_1,s)-q_1(k\mid q^-,z_1,s)\\ &=\mathfrak{E}(t_1)\Big[\Pr(A_{n+1}\in[k-q^+,\infty))-\Pr(A_{n+1}\in[k-q^-,\infty))\Big]\\ &\quad+(1-\mathfrak{E}(t_1))\Big[(\Pr(A_{n+1}\in[k-q^++u_1,\infty))-(\Pr(A_{n+1}\in[k-q^-+u_1,\infty))\Big]\\ &=\mathfrak{E}(t_1)\Big[\Pr(A_{n+1}\in[k-q^+,k-q^-])\Big]+(1-\mathfrak{E}(t_1))\Big[\Pr(A_{n+1}\in[k-q^++u_1,k-q^-+u_1])\Big].\end{aligned}$$

And similarly we obtain for the right hand side that:

$$\mathfrak{E}(t_2)\Big[\Pr(A_{n+1}\in[k-q^+,k-q^-])\Big]+(1-\mathfrak{E}(t_2))\Big[\Pr(A_{n+1}\in[k-q^++u_2,k-q^-+u_2])\Big].$$

Now in order to prove the theorem we define following symbols:

$$\begin{aligned}\Gamma &= \Pr(A_{n+1} \in [k-q^+, k-q^-])\\ \Delta_1 &= \Pr(A_{n+1} \in [k-q^+ + u_1, k-q^- + u_1])\\ \Delta_2 &= \Pr(A_{n+1} \in [k-q^+ + u_2, k-q^- + u_2]).\end{aligned}$$

Due to the fact that PMF of arrival process $\{A_n\}$ is assumed to be decreasing and convex, and properties of $\mathfrak{E}(t)$ we infer:

$$\begin{aligned}\Delta_2 \leq \Delta_1 \leq \Gamma\\ \mathfrak{E}(t_2) \leq \mathfrak{E}(t_1).\end{aligned}$$

So we write the theorem with new symbols:

$$\mathfrak{E}(t_2)\Gamma + (1-\mathfrak{E}(t_2))\Delta_2 \leq \mathfrak{E}(t_1)\Gamma + (1-\mathfrak{E}(t_1))\Delta_1,$$

exactly like the above theorem we can show the last inequality. □

We check the conditions of theorems 3.3.6, 3.3.7 to prove the monotonicity of value function and optimal policy for the proposed model.

***Theorem* 5.5.3.** *Value function $V(q,s)$ is an increasing function of state q, for fixed s.*

Proof. To prove this theorem we check the conditions of theorem 3.3.6.

- Due to property (3), $c(q,u,s,t)$ is increasing in q for fixed s, for all action pairs (u,t).
- Due to property (5), $q(k \mid q,u,s,t)$ is increasing in q, for all action pairs (u,t), for all s,k.
- Due to property (1), final cost $d_N(q,s)$ is increasing in q for fixed s.
- If $q^+ \geq q^-$, then obviously we have:
  1. $\mathcal{U}_{q^-} \subseteq \mathcal{U}_{q^+}$.
  2. $\forall (u',t') \in \mathcal{U}_{q^+} \setminus \mathcal{U}_{q^-}$, we have:

$$(u,t) \lesssim (u',t') \quad \forall (u,t) \in \mathcal{U}_{q^-}.$$

3. For any action $(u,t) \in \mathcal{U}_{q^+} \setminus \mathcal{U}_{q^-}$, there exists an action (u',t') for which we have the following properties:
   i. $c(q^+,(u,t),s) \geq c(q^-,(u',t'),s)$.
   ii. $q(k \mid q^+,(u,t),s) \geq q(k \mid q^-,(u',t'),s)$.

Item (1) is obvious since action space T is fixed and $\mathcal{U}_{q^-} \subseteq \mathcal{U}_{q^+}$. With similar logic, item (2) is also trivial. In order to construct items (3-*i*), (3-*ii*), we fix control action t, and use property (14). These conclude the proof.

□

***Theorem* 5.5.4.** *Imposing the assumption that power function $\pi(u,t)$ satisfies:*

$$\pi(u^-,t^+) + \pi(u^+,t^-) \leq \pi(u^-,t^-) + \pi(u^+,t^+),$$

Then, there exists an optimal transmission/feedback policy (f,g) which is increasing in q for the defined partial order $\overset{a}{\lesssim}$, on $\mathcal{U} \times \mathcal{T}$.

Proof. To prove this theorem we check the conditions of theorem 3.3.7. Conditions 1-4, are shown in previous theorem.

- Condition (7) of theorem3.3.7, is proved in theorem 5.5.1.
- Condition (8) of theorem 3.3.7, is proved in theorem 5.5.2.
- Conditions (5), and (6) are submodularity in $c(\cdot,\cdot)$ and $q(\cdot,\cdot)$ in (u,t) for the defined partial order.

Consider control actions (u_1,t_1) and (u_2,t_2). We use following notation for simplicity:

$$t^- = \min\{t_1,t_2\}, \quad t^+ = \max\{t_1,t_2\},$$
$$u^- = \min\{u_1,u_2\}, \quad u^+ = \max\{u_1,u_2\}.$$

Notice that for the defined partial order, we can derive the lowest upper bound and highest lower bound as following:

$$(u_1,t_1) \wedge (u_2,t_2) = (\min\{u_1,u_2\}, \max\{t_1,t_2\}) = (u^-,t^+)$$
$$(u_1,t_1) \vee (u_2,t_2) = (\max\{u_1,u_2\}, \min\{t_1,t_2\}) = (u^+,t^-).$$

Now our goal is to show following expression for the control pairs (u_1, t_1) and (u_2, t_2).

$$c\Big((u_1,t_1)\wedge(u_2,t_2),q,s\Big)+c\Big((u_1,t_1)\vee(u_2,t_2),q,s\Big)\leq c(u_1,t_1,q,s)+c(u_2,t_2,q,s)$$
$$\Leftrightarrow c(u^-,t^+,q,s)+c(u^+,t^-,q,s)\leq c(u_1,t_1,q,s)+c(u_2,t_2,q,s)$$

If $\Big[u_1 \leq u_2 \text{ and } t_1 \geq t_2\Big]$ or $\Big[u_1 \geq u_2 \text{ and } t_1 \leq t_2\Big]$, then:

$$\begin{aligned} &c(u^-,t^+,q,s)+c(u^+,t^-,q,s)\leq c(u_1,t_1,q,s)+c(u_2,t_2,q,s)\\ \Leftrightarrow\ &c(u^-,t^+,q,s)+c(u^+,t^-,q,s)\leq c(u^-,t^+,q,s)+c(u^+,t^-,q,s). \end{aligned} \tag{5.5}$$

Where the last inequality holds by equality . If $\Big[u_1 \leq u_2 \text{ and } t_1 \leq t_2\Big]$ or $\Big[u_1 \geq u_2 \text{ and } t_1 \geq t_2\Big]$, then:

$$\begin{aligned} &c(u^-,t^+,q,s)+c(u^+,t^-,q,s)\leq c(u_1,t_1,q,s)+c(u_2,t_2,q,s)\\ \Leftrightarrow\ &c(u^-,t^+,q,s)+c(u^+,t^-,q,s)\leq c(u^+,t^+,q,s)+c(u^-,t^-,q,s). \end{aligned} \tag{5.6}$$

To show this inequality note that cost term $\mathfrak{E}(t)d(q)$, cancels out from both sides and we have to show this property for $(1-\mathfrak{E}(t))d(q)$. We have following chain of inequalities:

$$\begin{aligned} &\Big[1-\mathfrak{E}(t^+)\Big]d(q-u^-)+\Big[1-\mathfrak{E}(t^-)\Big]d(q-u^+)\\ \leq&\Big[1-\mathfrak{E}(t^+)\Big]d(q-u^+)+\Big[1-\mathfrak{E}(t^-)\Big]d(q-u^-)\\ \Leftrightarrow&\mathfrak{E}(t^+)d(q-u^+)+\mathfrak{E}(t^-)d(q-u^-)\leq\mathfrak{E}(t^+)d(q-u^-)+\mathfrak{E}(t^-)d(q-u^+)\\ \Leftrightarrow&\mathfrak{E}(t^-)\Big[d(q-u^-)-d(q-u^+)\Big]\leq\mathfrak{E}(t^+)\Big[d(q-u^-)-d(q-u^+)\Big]\\ \Leftrightarrow&\mathfrak{E}(t^-)\leq\mathfrak{E}(t^+). \end{aligned}$$

Which is trivial due to the assumption that function $\mathfrak{E}(\cdot)$ is an increasing function of t. Also by assumption of the theorem we know $\pi(u,t)$ satisfies following inequality:

$$\pi(u^-,t^+)+\pi(u^-,t^+)\leq\pi(u^-,t^-)+\pi(u^+,t^+). \tag{5.7}$$

As a result, we showed cost function $c(u,t,q,s)$ is submodular based on the partial order

$\overset{a}{\lesssim}$. At last, we check condition (6). Similar to the equations 5.5 and 5.6, we can show $q_1(k \mid q, u, t, s)$ should satisfy following inequality for all k.

$$q_1(k \mid u^-, t^+, q, s) + q_1(k \mid u^+, t^-, q, s) \leq q_1(k \mid u^+, t^+, q, s) + q_1(k \mid u^-, t^-, q, s).$$

We have following chain of inequalities:

$$\begin{aligned}
&(1 - \mathfrak{E}(t^+))[\Pr[A_{n+1} \in [k - q + u^-, \infty)]] + (1 - \mathfrak{E}(t^-))[\Pr[A_{n+1} \in [k - q + u^+, \infty)]] \\
\leq &(1 - \mathfrak{E}(t^+))[\Pr[A_{n+1} \in [k - q + u^+, \infty)]] + (1 - \mathfrak{E}(t^-))[\Pr[A_{n+1} \in [k - q + u^-, \infty)]] \\
\Leftrightarrow &\mathfrak{E}(t^+)[\Pr[A_{n+1} \in [k - q + u^+, \infty)]] + \mathfrak{E}(t^-)[\Pr[A_{n+1} \in [k - q + u^-, \infty)]] \\
\leq &\mathfrak{E}(t^+)[\Pr[A_{n+1} \in [k - q + u^-, \infty)]] + \mathfrak{E}(t^-)[\Pr[A_{n+1} \in [k - q + u^+, \infty)]] \\
\Leftrightarrow &\mathfrak{E}(t^-)\Big[\Pr[A_{n+1} \in [k - q + u^-, \infty)] - \Pr[A_{n+1} \in [k - q + u^+, \infty)]\Big] \\
\leq &\mathfrak{E}(t^+)\Big[\Pr[A_{n+1} \in [k - q + u^-, \infty)] - \Pr[A_{n+1} \in [k - q + u^+, \infty)]\Big].
\end{aligned}$$

Where the last inequality follows from the fact that function $p(\cdot)$ is an increasing function in t. We have showed $q_1(k \mid q, u, s, t)$ is submodular, recall that $q_2(k \mid q, u, s, t)$ is not a function of (u, t), as a result, reverse CDF function is submodular in (u, t) based on the partial order $\overset{a}{\lesssim}$, and hence, condition (6) of theorem 3.3.7, is satisfied and this closes the proof. □

Remark 5.5.1. As it is proved in the previous theorem, function $q_1(k \mid q, u, s, t)$, and the cost part related to delay $\mathfrak{E}(t)d(q) + (1 - \mathfrak{E}(t))d(q - u)$ satisfies the submodularity condition for the partial order $\overset{a}{\lesssim}$; however, The structure of power function is a function of channel type. As a result, the $\pi(.,.)$ might not be submodular. In this case, this condition can be verified by checking following sufficient condition:

$$\frac{\partial^2 c(u, t, q, s)}{\partial u \partial t} \geq 0.$$

Designer of the system can verify bounds for which this equation is satisfied and as a result, he can extract intervals for the parameters of the system for which optimal policy is monotone with the order $\overset{a}{\lesssim}$.

Remark 5.5.2. The definition of submodularity for the defined partial order, as it is proved

in previous theorem, results in the definition of supermodularity on normal order on (u, t). As a result, Designer need to check the condition:

$$\frac{\partial^2 c(u,t,q,s)}{\partial u \partial t} \geq 0.$$

Next, we try to establish similar motonicity results for the the state of the channel s. We prove two results related to joint behavior of action and space.

***Theorem* 5.5.5.** *Cost function $c(q, u, t, s)$ has an increasing difference on $(\mathcal{U} \times \mathcal{T}) \times \mathcal{S}$. In other words, for action pairs $(u_1, t_1) \overset{a}{\lesssim} (u_2, t_2)$, and $s^- \leq s^+$ we have:*

$$c(q, s^+, (u_1, t_1)) - c(q, s^-, (u_1, t_1)) \leq c(q, s^+, (u_2, t_2)) - c(q, s^-, (u_2, t_2)).$$

Proof. For fixed control pairs (u_1, t_1), we have:

$$c(q, s^+, (u_1, t_1)) - c(q, s^-, (u_1, t_1)) = h(s^+)\pi(u_1, t_1) - h(s^-)\pi(u_1, t_1).$$

Since terms

$$\mathfrak{E}(t)d(q) + (1 - \mathfrak{E}(t))d(q - u)$$

get canceled from both sides. Then, we can write following chain of inequalities:

$$\begin{aligned} h(s^+)\pi(u_1, t_1) - h(s^-)\pi(u_1, t_1) &\leq h(s^+)\pi(u_2, t_2) - h(s^-)\pi(u_2, t_2) \\ \Leftrightarrow \pi(u_1, t_1)(h(s^+) - h(s^-)) &\leq \pi(u_2, t_2)(h(s^+) - h(s^-)) \overset{a}{\Leftrightarrow} \\ \pi(u_1, t_1) &\overset{b}{\leq} \pi(u_2, t_2), \end{aligned}$$

where (a) follows form the fact that $h(s^+) - h(s^-) \geq 0$ and (b) follows from the fact that $\pi(\cdot, \cdot)$ is increasing in variable u, and decreasing in variable t. □

***Theorem* 5.5.6.** *The function $q_2(k \mid q, u, t, s)$ has an increasing difference on $(\mathcal{U} \times \mathcal{T}) \times \mathcal{S}$ for fixed q. In other words, for action pairs $(u_1, t_1) \overset{a}{\lesssim} (u_2, t_2)$, and $s^- \leq s^+$ we have:*

$$q_2(k \mid q, s^+, (u_1, t_1)) - q_2(k \mid q, s^-, (u_1, t_1)) \leq q_2(k \mid q, s^+, (u_2, t_2)) - q_2(k \mid q, s^-, (u_2, t_2)).$$

The proof is trivial since we have shown:

$$q_2(k \mid q, u, s, t, a) = \Pr(S_{n+1} \in [k, \infty) \mid S_n = s),$$

as a result, last inequality is satisfied by equality.

***Theorem* 5.5.7.** *Value function $V(q, s)$ is an increasing function of state s, for fixed q.*

Proof. To prove this theorem we check the conditions of theorem 3.2.1.

- Due to property (4), $c(q, u, s, t)$ is increasing in s for fixed q, for all action pairs (u, t).
- Due to property (6), $q_2(k \mid q, u, s, t)$ is increasing in s, for all action pairs (u, t), for all q, k.
- Due to property (2), final cost $d_N(q, s)$ is increasing in q for fixed s.

As a result, value function $V(q, s)$ is increasing in s for fixed q. □

In the following theorem, we prove that there exists an optimal transmission/feedback policy (f, g) which is decreasing for the defined partial order $\overset{a}{\lesssim}$, on $\mathcal{U} \times \mathcal{T}$.

***Theorem* 5.5.8.** *Imposing the assumption that power function $\pi(u, t)$ satisfies eq.(5.7), then there exists an optimal transmission/feedback policy (f, g) which is decreasing for the defined partial order $\overset{a}{\lesssim}$, on $\mathcal{U} \times \mathcal{T}$.*

Proof. To prove this theorem we check the conditions of theorem 3.3.8. Conditions 1-3, are shown in previous theorem.

- Condition (6) of theorem3.3.8, is proved in theorem 5.5.5.
- Condition (7) of theorem 3.3.7, is proved in theorem 5.5.6.
- Finally, notice that conditions (4), and (5) are proved in theorem 5.5.4.

As a result, all conditions of theorem 3.3.8, are satisfied and proof is completed.

□

At last, we use theorems 3.3.10, and 3.3.9, to derive weaker results for the optimal policy when submodularity of $\pi(.,.)$ is hard to verify for the designer. These results constraint the structure of optimal policy using the defined partial order; however, they cannot clearly depict the structure as previous theorems.

***Theorem* 5.5.9.** *The optimal transmission/feedback policy (f, g) is non-decreasing in q for the defined partial order $\overset{a}{\lesssim}$, on $\mathcal{U} \times \mathcal{T}$, when $\pi(u, t)$ does not necessarily satisfy equation* (5.7).

Proof. In the proof of theorem 5.5.4, it is established that $Q(s, a)$ has decreasing difference in $Q \times \{T \times U\}$, as a result, proof of the claim follows from 3.3.10. □

***Theorem* 5.5.10.** *The optimal transmission/feedback policy (f, g) is non-increasing in s for the defined partial order $\overset{a}{\lesssim}$, on $\mathcal{U} \times \mathcal{T}$, when $\pi(u, t)$ does not necessarily satisfy equation* (5.7).

Proof. In the proof of theorem 5.5.8, it is established that $Q(s, a)$ has increasing difference in $Q \times \{T \times U\}$, as a result, proof of the claim follows from 3.3.9. □

5.6 Conclusion

In this chapter, we investigate the problem of transmission of bursty traffic over a fading channel when two control parameter of transmission rate and decoding power threshold are available. Structural properties of optimal value function and optimal policies are investigated in the model. With realistic assumptions on the model we show optimal transmission policy satisfies certain monotonicity properties as a function of queue state and channel state.

Chapter 6

Monotonicity in Energy Harvesting Communication Systems

6.1 Introduction

With emerging novel applications of wireless networks in sensing and monitoring, it is becoming increasingly important to design sustainable networks which can meet real time latency constraints. A promising approach to increase the lifetime of wireless networks is to use nodes that harvest energy from the environment. However, for a network with energy harvesting nodes to be efficient, one needs to design transmission strategies that can adapt to the unreliability of available energy.

The optimal design of wireless networks with energy harvesting sensors has received considerable attention in the literature. Broadly speaking, the literature may be classified into papers that take a physical layer view and characterize channel capacity and papers that take a cross layer view and determine throughput or delay optimal packet scheduling policies.

In this chapter, we take the latter approach and refer the readers to [6–9] and references therein for an overview of channel capacity with an energy harvesting transmitter. The problem of cross layer design of energy harvesting communication systems has been investigated in [10–20] under different modeling assumptions. The key criteria are: deterministic or stochastic data arrivals, deterministic or stochastic energy arrivals, finite or infinite data buffer and battery capacity, and throughput or delay optimal scheduling policies. Most

papers use Markov decision theory to investigate the model and establish the existence of optimal strategies, but typically resort to numerically solving the dynamic program to identify optimal policies.

The motivation for this work is to characterize qualitative properties of the optimal policies. For example, in queuing theory, it is often possible to establish that the optimal policy is monotone increasing in the queue length [21, 22]. Such a property, in addition to being intuitively satisfying, simplifies the search and implementation of the optimal strategies. Such monotonicity properties are also known to hold for cross layer design of communication systems when a constant source of energy is available at the transmitter [3]. So it is natural to ask if such qualitative properties hold for energy harvesting transmitters.

Partial answers to this question for throughput optimal policies are provided in [10–13]. Under the assumption of backlogged traffic or the assumption of a deterministic data arrival process, these papers show that the optimal policy is weakly increasing in the queue state and/or weakly increasing in the battery state.

In this chapter, we consider delay optimal policies for system with energy harvesting transmitter where *both* the data and energy arrival processes are stochastic. We show that the value function is weakly increasing in the queue state and weakly decreasing in the battery state. However, quite surprisingly, optimal policy is not necessarily monotone in battery or queue state. We present counterexamples to show that the optimal policy need not be weakly increasing in queue state nor in the battery state. Furthermore, the performance of the optimal policy is about 8-17% better than the performance of the best monotone policy. These counterexamples continue to hold for i.i.d. fading channels as well.

6.1.1 Notation

Uppercase letters (e.g., E, N, etc.) represent random variables; the corresponding lowercase letters (e.g., e, n, etc.) represent their realizations. Cursive letters (e.g., $\mathcal{L}$, $\mathcal{B}$, etc.) represent sets. The sets of real, positive integers, and non-negative integers are denoted by $\mathbb{R}$, $\mathbb{Z}_{>0}$, and $\mathbb{Z}_{\geq 0}$ respectively. The notation $[a]_L$ is a short hand for $\min\{a, L\}$.

6.2 Model And Problem Formulation

Consider a communication system shown in Fig 6.1. A source generates bursty data packets that have to be transmitted to a receiver by an energy-harvesting transmitter. The trans-

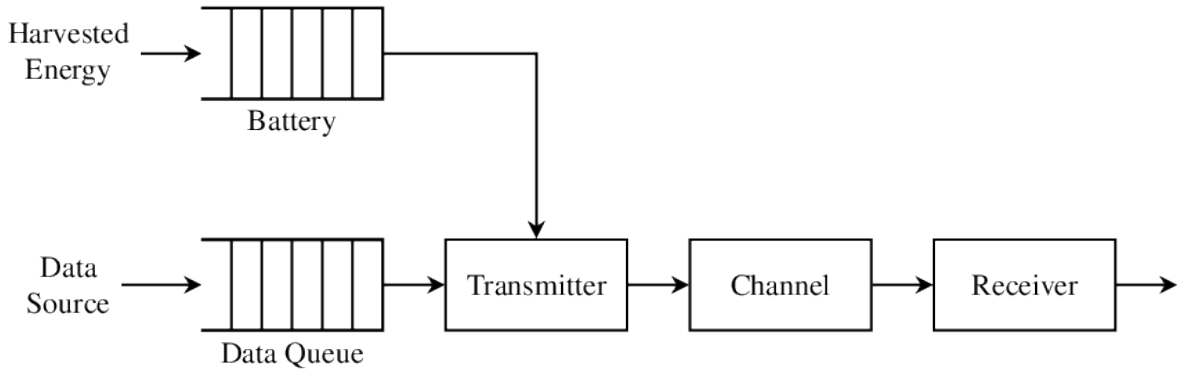

Fig. 6.1 Model of a transmitter with energy-harvester

mitter has finite buffer where the data packets are queued and a finite capacity battery where the harvested energy is stored. The system operates in discrete time slots. The data packets and the energy that arrive during a time slot are available only at the beginning of the next time slot.

At the beginning of a time slot, the transmitter picks some data packets from the queue, encodes them, and transmits the encoded symbol. Transmitting a symbol requires energy that depends on the number of encoded packets in the symbol. At the end of the time slot, the system incurs a delay penalty that depends on the number of packets remaining in the queue.

Time slots are indexed by $k \in \mathbb{Z}_{\geq 0}$. The length of the buffer is denoted by L and the size of the battery by B; $\mathcal{L}$ and $\mathcal{B}$ denote the sets $\{0, 1, \ldots, L\}$ and $\{0, 1, \ldots, B\}$, respectively. Other variables are as follows:

- $N_k \in \mathcal{L}$ denotes the number of data packets in the queue at the beginning of the time slot k.
- $A_k \in \mathcal{L}$ denotes the number of packets that arrive during time slot k.
- $S_k \in \mathcal{B}$ denotes the energy stored in the battery at the beginning of time slot k.
- $E_k \in \mathcal{B}$ denotes the energy that is harvested during time slot k.
- U_k denotes the number of packets transmitted during time slot k. The feasible choices

of U_k are denoted by $\mathcal{U}(N_k, S_k)$ where

$$\mathcal{U}(n, s) \coloneqq \{u \in \mathcal{L} : u \leq n \text{ and } p(u) \leq s\},$$

where $p(u)$ denotes the amount of power needed to transmit u packets. In our examples, we model the channel as a band-limited AWGN channel with bandwidth W and noise level N_0. The capacity of such a channel when transmitting at power level P is $W \log_2(1 + P/(N_0 W))$. Therefore, for such channels we assume $p(u) = \lfloor N_0 W(2^{u/W} - 1) \rfloor$. In general, we assume that $p : \mathcal{L} \to \mathbb{R}_{\geq 0}$ is a strictly convex and increasing function with $p(0) = 0$.

The dynamics of the data queue and the battery are

$$N_{k+1} = [N_k - U_k + A_k]_L \quad \text{and} \quad S_{k+1} = [S_k - p(U_k) + E_k]_B.$$

Packets that are not transmitted during time slot k incur a delay penalty $d(N_k - U_k)$, where $d : \mathcal{L} \to \mathbb{R}_{\geq 0}$ is a convex and increasing function with $d(0) = 0$.

The data arrival process $\{A_k\}_{k \geq 0}$ is assumed to be an independent and identically distributed process with pmf (probability mass function) P_A. The energy arrival process $\{E_k\}_{k \geq 0}$ is an independent process that is also independent of $\{A_k\}_{k \geq 0}$ with pmf P_E.

The number U_k of packets to transmit are chosen according to a scheduling policy $f \coloneqq \{f_k\}_{k \geq 0}$, where

$$U_k = f_k(N_k, S_k), \quad U_k \in \mathcal{U}(N_k, S_k).$$

The performance of a scheduling policy f is given by

$$J(f) \coloneqq \mathbb{E}^f \Big[\sum_{k=0}^{\infty} \beta^k d(N_k - U_k) \;\Big|\; N_0 = 0, S_0 = 0 \Big], \tag{6.1}$$

where $\beta \in (0, 1)$ denotes the discount factor and the expectation is taken with respect to the joint measure on the system variables induced by the choice of f.

We are interested in the following optimization problem.

Problem 3. *Given the buffer length L, battery size B, power cost $p(\cdot)$, delay cost $d(\cdot)$, pmf P_A of the arrival process, pmf P_E of the energy arrival process, and the discount factor β, choose a feasible scheduling policy f to minimize the performance $J(f)$ given by* (6.1).

6.3 Dynamic Programming Decomposition

The system described above can be modeled as an infinite horizon time homogeneous Markov decision process (MDP) [23]. Since the state and action spaces are finite, standard results from Markov decision theory imply that there exists an optimal policy which is time homogeneous and is given by the solution of a dynamic program. To succinctly write the dynamic program, we define the following Bellman operator: Define the operator $\mathscr{B} : [\mathcal{L} \times \mathcal{B} \to \mathbb{R}] \to [\mathcal{L} \times \mathcal{B} \to \mathbb{R}]$ that maps any $V : \mathcal{L} \times \mathcal{B} \to \mathbb{R}$ to

$$[\mathscr{B}V](n,s) \;=\; \min_{u \in \mathcal{U}(n,s)} \Big\{ d(n \,-\, u) \,+\, \beta \mathbb{E}[V([n \,-\, u \,+\, A]_L, [s \,-\, p(u) \,+\, E]_B)] \Big\}, \quad (6.2)$$

where A and E are independent random variables with pmfs P_A and P_E. Then, an optimal policy for the infinite horizon MDP is given as follows [23].

***Theorem* 6.3.1.** *Let $V^* : \mathcal{L} \times \mathcal{B} \to \mathbb{R}$ denote the unique fixed point of the following equation:*

$$V(n,s) = [\mathscr{B}V](n,s), \quad \forall (n,s) \in \mathcal{L} \times \mathcal{B}. \quad (6.3)$$

Furthermore, let f^* be such that $f^*(n,s)$ attains the minimum in the right hand side of (6.3). Then, the time homogeneous policy $f^{*,\infty} = (f^*, f^*, \dots)$ is optimal for Problem 3.

The dynamic program described in (6.3) can be solved using value iteration, policy iteration, or linear programming algorithms [23].

6.3.1 Properties of the value function

Let $\mathcal{M}$ denote the family of the functions $V : \mathcal{L} \times \mathcal{B} \to \mathbb{R}$ such that

1. for any $s \in \mathcal{B}$, $V(n,s)$ is weakly increasing in n.
2. for any $n \in \mathcal{L}$, $V(n,s)$ is weakly decreasing in s.

Furthermore, let $\mathcal{F}_s$ denote the family of functions $f : \mathcal{L} \times \mathcal{B} \to \mathcal{U}$ such that for any $n \in \mathcal{L}$, $f(n,s)$ is weakly increasing in s. Similarly, let $\mathcal{F}_n$ be family of functions $f : \mathcal{L} \times \mathcal{B} \to \mathcal{U}$, such that for any $s \in \mathcal{B}$, $f(n,s)$ is weakly increasing in n.

***Proposition* 6.3.1.** *The optimal value function $V^* \in \mathcal{M}$.*

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	1	1	1	1
2	0	1	1	1	2	2
3	0	1	1	1	1	2
4	0	1	1	1	1	2
5	0	1	1	1	1	2

(a) The optimal policy

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	1	1	1	1
2	0	1	1	1	1	2
3	0	1	1	1	1	2
4	0	1	1	1	1	2
5	0	1	1	1	1	2

(b) The best monotone policy

Fig. 6.2 The optimal and the best monotone policies for the example of Sec. 6.4.1.

The proof is presented in the Appendix.

Proposition 6.3.1 says that if we follow an optimal policy, the optimal cost when starting from a smaller queue state is lower than that starting from a larger queue state. Similarly, the optimal cost when starting from a larger battery state is lower than that starting from a smaller battery state. Such a result appears to be intuitively obvious.

One might argue that it should be the case that the optimal policy should be weakly increasing in state of the queue, and weakly increasing in the available energy in the battery. In particular, if it is optimal to transmit u packets when the queue state is n, then (for the same battery state) the optimal number of packets to transmit at any queue state larger than n should be at least u. Similarly, if it is optimal to transmit u packets when the battery state is s, then (for the same queue state) the optimal number of packets to transmit at any battery state larger than s should be at least u. In the next section, we present counterexamples that show both of these properties do not hold. The code for all the results is available at [24].

6.4 Counterexamples on the monotonicity of optimal policies

6.4.1 On the monotonicity in queue state

Consider the communication system with a band-limited AWGN channel where $\mathcal{L} = 5$, $\mathcal{B} = 5$, $\beta = 0.99$, $N_0 = 2.0$, $W = 1.75$ (thus, $p(u) = \lfloor 3.5 \cdot (2^{(u/1.75)} - 1) \rfloor$), $d(q) = q$, data arrival distribution $P_A = [0.59, 0.41, 0, 0, 0]$, and energy arrival distribution $P_E = [0.385, 0.23, 0.385, 0, 0]$.

The optimal policy for this system (obtained by policy iteration [23]) is shown in Fig. 6.2a, where the rows correspond to the current queue length and the columns correspond to the current energy level. Note that the policy is not weakly increasing in queue state (i.e, $f^* \notin \mathcal{F}_n$). In particular, $f(3,4) < f(2,4)$.

Given that the optimal policy is not monotone, one might wonder how much do we lose if we follow a monotone policy instead of the optimal policy. To characterize this, we define the best queue-monotone policy as:

$$f_n^\circ = \arg\min_{f \in \mathcal{F}_n} \left\{ \max_{(n,s)\in\mathcal{L}\times\mathcal{B}} |V(n,s) - V^*(n,s)| \right\}$$

and let V_n° denote the corresponding value function.

The best monotone policy cannot be obtained using dynamic programming and one has to resort to a brute force search over all monotone policies. For the model described above, there are 86400 monotone policies.[1] The best monotone policy obtained by searching over these is shown is Fig. 6.2b. The worst case difference between the two value functions is given by

$$\alpha_n = \max_{(n,s)\in\mathcal{L}\times\mathcal{B}} \left\{ \frac{V^\circ(n,s) - V^*n(n,s)}{V^*(n,s)} \right\} = 0.1764.$$

Thus, for this counterexample, the best queue-monotone policy performs 17.46% worse than the optimal policy.

6.4.2 On the monotonicity in battery state

Consider the communication system described in Sec. 6.4.1 but with $P_A = [0.33, 0.67, 0, 0, 0]$ as the data arrival distribution and $P_E = [0.033, 0.934, 0.033, 0, 0]$ as the energy arrival distribution .

The optimal policy (obtained using policy iteration [23]) is shown in Fig. 6.3a. Note that the policy is not weakly increasing in the battery state (i.e $f^* \notin \mathcal{F}_s$). In particular, we have that $f^*(5,2) > f^*(5,3)$.

Given that optimal policy is not monotone, the previous question arises again that how much do we lose if we follow a monotone policy instead of the optimal policy. To

[1] Due to the power constraint $\mathcal{U}(n,s)$, it is not possible to count the number of monotone functions using combinatorics. The number above is obtained by explicit enumeration.

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \left[\begin{array}{cccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 2 & 2 \\ 0 & 1 & 1 & 1 & 2 & 2 \\ 0 & 1 & 1 & 1 & 2 & 2 \\ 0 & 1 & 1 & 0 & 2 & 2 \end{array}\right] \end{array}$$

(a) The optimal policy

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \left[\begin{array}{cccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 2 & 2 \\ 0 & 1 & 1 & 1 & 2 & 2 \\ 0 & 1 & 1 & 1 & 2 & 2 \\ 0 & 0 & 0 & 0 & 2 & 2 \end{array}\right] \end{array}$$

(b) The best monotone policy

Fig. 6.3 The optimal and the best monotone policies for the example of Sec. 6.4.1.

characterize this, we define the best battery-monotone policy as:

$$f_s^\circ = \arg\min_{f \in \mathcal{F}_s} \left\{ \max_{(n,s) \in \mathcal{L} \times \mathcal{B}} |V(n,s) - V^*(n,s)| \right\}$$

and let V_s° denote the corresponding value function.

As before, we find the best monotone policy by a a brute force search over all 303750 monotone battery-policies. The resultant policy is shown in Fig. 6.3b.

The worst case difference between the two value functions is given by

$$\alpha_s = \max_{(n,s) \in \mathcal{L} \times \mathcal{B}} \left\{ \frac{V^\circ(n,s) - V^* s(n,s)}{V^*(n,s)} \right\} = 0.0860.$$

Thus, for this counterexample, the best battery-monotone policy performs 8.60% worse than the optimal policy.

6.5 Counterexamples for fading channels

6.5.1 Channel model with i.i.d. fading

Consider the model in Sec. 6.2 where the channel has i.i.d. fading. In particular, let $H_k \in \mathcal{H}$ denote the channel state at time k and $g(H_k)$, where $g : \mathcal{H} \to \mathbb{R}_{>0}$, denote the attenuation at state H_k. Thus, the power needed to transmit u packets when the channel is in state h is given by $p(u)/g(h)$. We assume that $\{H_k\}_{k \geq 0}$ is an i.i.d. process with pmf P_H that is independent of the data and energy arrival processes $\{A_k\}_{k \geq 0}$ and $\{E_k\}_{k \geq 0}$.

6.5.2 On the monotonicity in queue state

Consider the model in Sec. 6.4.1 with $N_0 = 1$, $W = 1.75$, and an i.i.d. fading channel where $\mathcal{H} = \{1, 2, 3\}$, $g(\cdot) = \{0.4, 0.7, 0.8\}$ and $P_H = [0.15, 0.25, 0.6]$. The optimal policy for this model (obtained using policy iteration) is shown in Fig. 6.4a–6.4c. Note that for all h, the optimal policy in not monotone in the queue length.

In this case, there are $(4320) \times (1296) \times (362) \approx 10^8$ monotone policies. Therefore, a brute force search to find the best monotone policy is not possible. We choose a heuristic monotone policy f_n° which differs from f^* only at the following points: $f_n^\circ(1, 2, 1) = 0$, $f_n^\circ(4, 3, 1) = 1$, $f_n^\circ(5, s, 1) = 1$, for $s \in \{3, 4\}$, $f_n^\circ(2, 3, 2) = 1$, $f_n^\circ(5, 1, 2) = 1$, $f_n^\circ(3, 5, 2) = 2$, $f_n^\circ(3, 5, 2) = 2$, $f_n^\circ(5, 4, 2) = 2$, $f_n^\circ(5, 1, 3) = 1$ and $f_n^\circ(3, 4, 3) = 2$. The policy f_n° may be thought of as the queue-monotone policy that is closest to f^*. Let V_n° denote the corresponding value function. The worst case difference between the two value functions is given by

$$\alpha_n = \max_{(n,s,h)\in\mathcal{L}\times\mathcal{B}\times\mathcal{H}} \frac{|V^*(n, s, h) - V_n^\circ(n, s, h)|}{|V^*(n, s, h)|} = 0.2052.$$

Thus, the heuristically chosen queue-monotone policy performs 20.52% worse than the optimal policy.

6.5.3 On the monotonicity in the battery state

Consider the model in Sec. 6.4.2 with $N_0 = 1.55$, $W = 1.75$, and an i.i.d. fading channel where $\mathcal{H} = \{1, 2\}$, $g(\cdot) = \{0.75, 0.80\}$, and $P_H = [0.3, 0.7]$. The optimal policy for this model (obtained using policy iteration) is shown in Fig. 6.4d–6.4e. Note that for $h \in \{1, 2\}$, the optimal policy is not monotone in the battery state.

In this case, there are $(629856) \times (30375019) \approx 10^{10}$ monotone policies. Therefore, a brute force search is not possible. As before, we choose a heuristic policy f_s° which is the battery-monotone policy that is closest to f^*. In particular, f_s° differs from f^* only at two points: $f_s^\circ(5, 3, 1) = 1$ and $f_s^\circ(5, 3, 2) = 1$. Let V_s° denote the corresponding value function. The worst case difference between the two value functions is given by

$$\alpha_s = \max_{(n,s,h)\in\mathcal{L}\times\mathcal{B}\times\mathcal{H}} \frac{|V^*(n, s, h) - V_s^\circ(n, s, h)|}{|V^*(n, s, h)|} = 0.1114.$$

Thus, the heuristically chosen battery-monotone policy performs 11.14% worse than the

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \left[\begin{array}{cccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \end{array}$$

(a) $f^*(\cdot,\cdot,h=1)$

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \left[\begin{array}{cccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 2 & 2 & 2 \\ 0 & 1 & 1 & 1 & 2 & 3 \\ 0 & 1 & 1 & 1 & 2 & 2 \\ 0 & 0 & 1 & 1 & 1 & 2 \end{array}\right] \end{array}$$

(b) $f^*(\cdot,\cdot,h=2)$

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \left[\begin{array}{cccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 2 & 2 & 2 \\ 0 & 1 & 2 & 2 & 3 & 3 \\ 0 & 1 & 2 & 2 & 2 & 3 \\ 0 & 0 & 2 & 2 & 2 & 3 \end{array}\right] \end{array}$$

(c) $f^*(\cdot,\cdot,h=3)$

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \left[\begin{array}{cccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 2 & 2 \\ 0 & 1 & 1 & 1 & 2 & 2 \\ 0 & 1 & 1 & 1 & 2 & 2 \\ 0 & 1 & 1 & 0 & 2 & 2 \end{array}\right] \end{array}$$

(d) $f^*(\cdot,\cdot,h=1)$

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \left[\begin{array}{cccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 2 & 2 \\ 0 & 1 & 1 & 1 & 2 & 2 \\ 0 & 1 & 1 & 1 & 2 & 2 \\ 0 & 1 & 1 & 0 & 2 & 2 \end{array}\right] \end{array}$$

(e) $f^*(\cdot,\cdot,h=2)$

Fig. 6.4 The optimal policy for the examples of Sec. 6.5.2 shown in subfigures (a)–(c) and Sec. 6.5.3 shown in subfigures (d)–(e).

optimal policy.

6.6 Conclusion

In this chapter, we consider delay optimal strategies in cross layer design with energy harvesting transmitter. We show that the value function is weakly increasing in the queue state and weakly decreasing in the battery state. We show via counterexamples that the optimal policy is not monotone in queue length nor in the available energy in the battery.

6.6.1 Discussion about the counterexamples

One might ask why the optimal policy is not monotone in the above model. The standard argument in MDPs to establish monotonicity of the optimal policies is to show that the value-action function is submodular in the state and action. The value-action function is given by

$$H(n,s,u) = d(n-u) + \beta \mathbb{E}[V([n-u+A]_L, [s-p(u)+E]_B)]$$

A sufficient condition for the optimal policy to be weakly increasing in the queue length is:

(S1) for every $s \in \mathcal{B}$, $H(n,s,u)$ is submodular in (n,u).

Note that since $d(\cdot)$ is convex, $d(n-u)$ is submodular in (n,u). Thus, a sufficient condition for (S1) to hold is:

(S2) for all $s \in \mathcal{B}$, $\mathbb{E}[V([n-u+A]_L, [s-p(u)+E]_B)]$ is submodular in (n,u).

Since submodularity is preserved under addition, a sufficient condition for (S2) to hold is:

(S3) for all $s \in \mathcal{B}$, $V(n-u, s-p(u))$ is submodular in (n,u).

By a similar argument, it can be shown that a sufficient condition for the optimal policy to be weakly increasing in battery state is:

(S4) for all $n \in \mathcal{L}$, $V(n-u, s-p(u))$ is submodular in (s,u).

We have not been able to identify sufficient conditions under which (S3) or (S4) hold. Note that if the data were backlogged, then we do not need to keep track of the queue state; thus, the value function is just a function of the battery state. In such a scenario, (S4) simplifies to $V(s-p(u))$ is submodular in (s,u). Since $p(\cdot)$ is convex, it can be shown that convexity of $V(s)$ is sufficient to establish submodularity of $V(s-p(u))$. This is the essence of the argument given in [10, 12].

Similarly, if the transmitter had a steady supply of energy, then we do not need to keep track of the battery state; thus, the value function is just a function of the queue state. In such a scenario, (S3) simplifies to $V(n-u)$ is submodular in (n,u). It can be shown that convexity of the $V(n)$ is sufficient to establish submodularity of $V(n-u)$. This is the essence of the argument given in [3].

In our model, data is not backlogged and energy is intermittent. As a result, we have two queues—the data queue and the energy queue—which have coupled dynamics. This coupling makes it difficult to identify conditions under which $V(n-u, s-p(u))$ will be submodular in (n,u) or (s,u).

6.6.2 Implication of the results

In general, there are two benefits if one can establish that the optimal policy is monotone. The first advantage is that monotone policies are easier to implement. In particular, one needs a $(L+1)\times(B+1)$-dimensional look-up table to implement a general transmission policy (similar to the matrices shown in Figs. 6.2 and 6.3). In contrast, one only needs to store the thresholds boundaries of the decision regions (which can be stored in a sparse matrix) to implement a queue- or battery-monotone policy. Our counterexamples show that such a simpler implementation will result in a loss of optimality in energy-harvesting systems.

The second advantage is that if we know that the optimal policy is monotone, we can search for them efficiently using monotone value iteration and monotone policy iteration [1]. Our counterexamples show that these more efficient algorithms cannot be used in energy-harvesting systems.

One might want to restrict to monotone policies for the sake of implementation simplicity. However, if the system does not satisfy properties (S3) and (S4) mentioned in the previous section, then dynamic programming cannot be used to find the best monotone policy. Thus, one has to resort to a brute force search, which suffers from the curse of dimensionality.

Finally, one might simply choose a heuristic monotone policy rather than the best monotone policy. In that case, Markov decision theory can be used to bound the degree of suboptimality of the optimal policy. In particular, suppose f° is a heuristic policy and V° is the corresponding value function. Let V' denote $\mathscr{B}V^\circ$. Then, [23, Vol II, Proposition 3.1] implies that the optimal value function V^* is bounded by

$$\underline{\delta} \leq V' - V^* \leq \bar{\delta}, \tag{6.4}$$

where

$$\underline{\delta} = \frac{\beta}{1-\beta} \min_{(n,s)\in\mathcal{L}\times\mathcal{B}} \Big\{ V'(n,s) - V^{\circ}(n,s) \Big\},$$
$$\bar{\delta} = \frac{\beta}{1-\beta} \max_{(n,s)\in\mathcal{L}\times\mathcal{B}} \Big\{ V'(n,s) - V^{\circ}(n,s) \Big\}.$$

Based on the degree of suboptimality $\underline{\delta}$ and $\bar{\delta}$, the system designer may decide whether the benefit due to the simpler implementation outweighs the performance loss.

6.6.3 Monotonicity of Bellman operator

***Lemma* 6.6.1.** *Given $V \in \mathcal{M}$, let*

$$H(n,s,u) = d(n-u) + \beta \mathbb{E}[V([n-u+A]_L, [s-p(u)+E]_B)] \tag{6.5}$$

for all $n \in \mathcal{L}$, $s \in \mathcal{B}$, and $u \in \mathcal{U}(n,s)$. Furthermore, let $W = \mathscr{B}V$, i.e.,

$$W(n,s) = \min_{u\in\mathcal{U}(n,s)} H(n,s,u), \quad \forall n \in \mathcal{L}, s \in \mathcal{B}. \tag{6.6}$$

Then, for all $n \in \mathcal{L}$, $s \in \mathcal{B}$, and $u \in \mathcal{U}(n,s)$, we have:

1. $H(n,s,u) \leq H([n+1]_L, s, u)$.
2. $H(n,s,n) \leq H([n+1]_L, s, [n+1]_L)$.
3. $H(n,[s+1]_B,u) \leq H(n,s,u)$.

As a consequence of the above, $W \in \mathcal{M}$.

Proof. We first prove properties of H.

1. We assume that $n+1 \in \mathcal{L}$ (otherwise, the result is trivially true). Since $u \in \mathcal{U}(n,s)$, $u \leq n$ and therefore $u < n+1$. Thus, from monotonicity of $d(\cdot)$, we have

$$d(n+1-u) \geq d(n-u).$$

In addition, since $V \in \mathcal{M}$, we infer that for any $u \in \mathcal{U}(n,s)$,

$$\mathbb{E}[V([n+1-u+A]_L, [s-p(u)+E]_B)] \geq \mathbb{E}[V([n-u+A]_L, [s-p(u)+E]_B)].$$

Combining the above two equations, we get

$$H(n, s, u) \leq H(n + 1, s, u).$$

2. We assume that $n + 1 \in \mathcal{L}$ (otherwise, the result is trivially true). Note that

$$d(n + 1 - n - 1) = d(n - n) = d(0).$$

In addition, since $V \in \mathcal{M}$, we infer that for any $u \in \mathcal{U}(n, s)$,

$$\mathbb{E}[V([n + 1 - n - 1 + A]_L, [s - p(u + 1) + E]_B)] \\ \geq \mathbb{E}[V([n - n + A]_L, [s - p(u) + E]_B)].$$

Combining the above two equations, we get

$$H(n, s, n) \leq H(n + 1, s, n + 1).$$

3. We assume that $s+1 \in \mathcal{B}$ (otherwise, the result is trivially true). Note that $\mathcal{U}(n, s) \subseteq \mathcal{U}(n, s + 1)$. In addition, since $V \in \mathcal{M}$, we infer that for any $u \in \mathcal{U}(n, s)$,

$$\mathbb{E}[V([n - u + A]_L, [s - p(u) + E]_B)] \geq \mathbb{E}[V([n - u + A]_L, [s + 1 - p(u) + E]_B)].$$

Hence, we get that

$$H(n, s + 1, u) \leq H(n, s, u).$$

We now prove that $W \in \mathcal{M}$. For any $n \in \mathcal{L}$ and $s \in \mathcal{B}$, let $f(n, s)$ denote an arg min of the right hand side of (6.6). Now we consider two cases: $f(n + 1, s) \neq n + 1$ and $f(n + 1, s) = n + 1$.

1. Suppose $u^* = f(n+1, s) \neq n+1$. Then, it must be the case that $u^* \in \mathcal{U}(n, s)$. Thus,

$$W(n + 1, s) = H(n + 1, s, u^*) \overset{(a)}{\geq} H(n, s, u^*) \\ \geq \min_{u \in \mathcal{U}(n,s)} H(n, s, u) = W(n, s),$$

where (a) follows from Property 1.

2. Suppose $u^* = f(n+1, s) = n+1$. Then, it must be the case that $p(n+1) \leq s$ and, therefore, $p(n) \leq s$. Hence $n \in \mathcal{U}(n, s)$. Thus,

$$\begin{aligned} W(n+1, s) = H(n+1, s, n+1) &\overset{(b)}{\geq} H(n, s, n) \\ &\geq \min_{u \in \mathcal{U}(n,s)} H(n, s, u) = W(n, s), \end{aligned}$$

where (b) follows from Property 2.

As a result of both of these cases, we get that

$$W(n, s) \leq W(n+1, s). \tag{6.7}$$

Now let $u^* = f(n, s)$, recall that $\mathcal{U}(n, s) \subseteq \mathcal{U}(n, s+1)$ then $u^* \in \mathcal{U}(n, s+1)$ thus

$$\begin{aligned} W(n, s) = H(n, s, u^*) &\overset{(c)}{\geq} H(n, s+1, u^*) \\ &\overset{(d)}{\geq} \min_{u \in \mathcal{U}(n,s+1)} H(n, s+1, u) = W(n, s+1), \end{aligned} \tag{6.8}$$

Where (c) follows from Property 3 and (d) follows from the fact that $u^* \in \mathcal{U}(n, s+1)$.

As a result of previous cases, From (6.7) and (6.8) we infer $W \in \mathcal{M}$. □

6.6.4 Proof of Proposition 6.3.1

Arbitrarily initialize $V^{(0)} \in \mathcal{M}$ and for $n \in \mathbb{Z}_{>0}$, recursively define $V^{(n+1)} = \mathscr{B}V^{(n)}$. Since $V^{(0)} \in \mathcal{M}$, Lemma 6.6.1 implies that $V^{(n)} \in \mathcal{M}$, for all $n \in \mathbb{Z}_{>0}$. Since monotonicity is preserved under the limit, we have that $\lim_{n\to\infty} V_0^{(n)} \in \mathcal{M}$. By [23], $\lim_{n\to\infty} V_0^{(n)} = V$. Hence, $V \in \mathcal{M}$.

6.7 Bounds on the suboptimality of monotone policies

In this section we investigate the importance of results and counterexamples we found in this paper. In previous sections, we brought counterexamples for monotonicity property of optimal policy in energy harvesting systems. In this section we try to illustrate the

importance of this result. In practice, designer has the choice between best monotone policy and optimal policy. The common interest in monotone policies is because of their simple implementation. In other words, a monotone strategy can be stored in a memory only by indices of increases and their values. On the other hand, in order to implement optimal policy, designer has to store the policy look up table for all the states. The other difference between optimal and best monotone policy is the difficulty in finding these strategies. To the best of authors' knowledge, there is no low computation algorithm to extract the best monotone policy when there is no guarantee on monotonicity of optimal policy. The monotone policy and value iteration discussed in [1], only work under sufficient conditions which shows optimal policy is monotone. As a result, in order to extract best monotone strategy, one should conduct a brute force search over the space of all monotone policies while optimal policy can be extracted with a simple policy iteration algorithm. If designer extracts the best monotone policy f_n° or f_s°, applying the results in approximate dynamic programming [25], he can bound the performance from the optimal policy. Suppose V_s° is the value function corresponds to f_s°, and $V_s^1 = \mathscr{B}V_s^\circ$ then we have:

$$\begin{aligned}&\frac{\beta}{1-\beta}\min_{(n,s)\in\mathcal{L}\times\mathcal{B}}\{V_s^1(n,s)-V_s^\circ(n,s)\}\leq V^*(n,s)\\ \leq&\frac{\beta}{1-\beta}\min_{(n,s)\in\mathcal{L}\times\mathcal{B}}\{V_s^1(n,s)-V_s^\circ(n,s)\}\end{aligned}$$

This lower and higher bound on the value function based on the value function of best monotone policy, let the designer understand the degree of suboptimality of the best monotone policy. Base on degree of suboptimality and quality of service constraints, designer can decide between simplicity in the implementation and optimality of the strategy.

Authors think these counterexamples and the degree of suboptimality in their performance can show that energy harvesting system designers should consider the trade off between simplicity in the implementation and degree of suboptimality more carefully. In addition, in online implementation setups, monotone policies are not easily achievable nor close to optimal in the performance. This, we think bring new insights about the existing trade-offs in designing energy harvesting communication systems.

References

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

[2] D. M. Topkis, *Supermodularity and complementarity*. Princeton university press, 1998.

[3] R. A. Berry, *Power and delay trade-offs in fading channels*. PhD thesis, Massachusetts Institute of Technology, 2000.

[4] Edmund Yeh, "Personal communication," 2009.

[5] E. Altman, *Constrained Markov decision processes*, vol. 7. CRC Press, 1999.

[6] W. Mao and B. Hassibi, "Capacity analysis of discrete energy harvesting channels," *IEEE Trans. Inf. Theory*, vol. 63, pp. 5850–5885, Sept 2017.

[7] R. Rajesh, V. Sharma, and P. Viswanath, "Capacity of gaussian channels with energy harvesting and processing cost," *IEEE Trans. Inf. Theory*, vol. 60, pp. 2563–2575, May 2014.

[8] O. Ozel and S. Ulukus, "Achieving AWGN capacity under stochastic energy harvesting," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6471–6483, 2012.

[9] D. Shaviv, P. M. Nguyen, and A. Özgür, "Capacity of the energy-harvesting channel with a finite battery," *IEEE Trans. Inf. Theory*, vol. 62, pp. 6436–6458, Nov 2016.

[10] A. Sinha, "Optimal power allocation for a renewable energy source," in *Communications (NCC), 2012 National Conference on*, pp. 1–5, IEEE, 2012.

[11] I. Ahmed, K. T. Phan, and T. Le-Ngoc, "Optimal stochastic power control for energy harvesting systems with delay constraints," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3512–3527, 2016.

[12] S. Mao, M. H. Cheung, and V. W. Wong, "Joint energy allocation for sensing and transmission in rechargeable wireless sensor networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 6, pp. 2862–2875, 2014.

[13] M. Kashef and A. Ephremides, "Optimal packet scheduling for energy harvesting sources on time varying wireless channels," *Journal of Communications and Networks*, vol. 14, no. 2, pp. 121–129, 2012.

[14] V. Sharma, U. Mukherji, V. Joseph, and S. Gupta, "Optimal energy management policies for energy harvesting sensor nodes," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 1326–1336, April 2010.

[15] C. K. Ho and R. Zhang, "Optimal energy allocation for wireless communications with energy harvesting constraints," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4808–4818, 2012.

[16] O. Ozel, K. Tutuncuoglu, J. Yang, S. Ulukus, and A. Yener, "Transmission with energy harvesting nodes in fading wireless channels: Optimal policies," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1732–1743, 2011.

[17] K. Tutuncuoglu and A. Yener, "Optimum transmission policies for battery limited energy harvesting nodes," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 1180–1189, 2012.

[18] K. Wu, C. Tellambura, and H. Jiang, "Optimal transmission policy in energy harvesting wireless communications: A learning approach," in *Commun. (ICC), 2017 IEEE Int. Conf. on*, pp. 1–6, IEEE, 2017.

[19] J. Yang and S. Ulukus, "Optimal packet scheduling in an energy harvesting communication system," *IEEE Trans. Commun.*, vol. 60, no. 1, pp. 220–230, 2012.

[20] B. Varan and A. Yener, "Delay constrained energy harvesting networks with limited energy and data storage," *IEEE J. Sel. Areas Commun.*, vol. 34, pp. 1550–1564, May 2016.

[21] S. Stidham Jr and R. R. Weber, "Monotonic and insensitive optimal policies for control of queues with undiscounted costs," *Operations research*, vol. 37, no. 4, pp. 611–625, 1989.

[22] E. Gallisch, "On monotone optimal policies in a queueing model of M/G/1 type with controllable service time distribution," *Advances in Applied Probability*, vol. 11, no. 4, pp. 870–887, 1979.

[23] D. P. Bertsekas, *Dynamic programming and optimal control*, vol. 1. Athena scientific Belmont, MA, 2005.

[24] B. Sayedana and A. Mahajan, "Counterexamples on the monotonicity of delay optimal transmission policies in energy harvesting communication systems," May 2019. Available at `https://codeocean.com/capsule/0148099/tree/v2`.

[25] D. P. Bertsekas, *Dynamic programming and optimal control*, vol. 1. Athena scientific Belmont, MA, 1995.