CALL ADMISSION CONTROL IN WIRELESS DS-CDMA SYSTEMS USING REINFORCEMENT LEARNING

Pitipong Chanloha

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Telecommunication Engineering

Suranaree University of Technology

Academic Year 2006

การควบคุมการเรียกเข้าในระบบ ดีเอส-ซีดีเอ็มเอไร้สาย
โดยใช้การเรียนรู้แบบรีอินฟอร์สเมนท์

นายปิติพงศ์ ชาญโลหะ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมโทรคมนาคม
มหาวิทยาลัยเทคโนโลยีสุรนารี
ปีการศึกษา 2549

CALL ADMISSION CONTROL IN WIRELESS DS-CDMA SYSTEMS USING REINFORCEMENT LEARNING

Suranaree University of Technology has approved this thesis submitted in partial fulfillment of the requirements for a Master's Degree.

Thesis Examining Committee

_____

(Asst. Prof. Dr. Rangsan Tongta)

Chairperson

_____

(Asst. Prof. Dr. Wipawee Hattagam)

Member (Thesis Advisor)

_____

(Assoc. Prof. Dr. Kitti Attakitmongcol)

Member

_____

(Assoc. Prof. Dr. Saowanee Rattanaphani)

Vice Rector for Academic Affairs

_____

(Assoc. Prof. Dr. Vorapot Khompis)

Dean of Institute of Engineering

ปิติพงศ์ ชาญโลหะ : การควบคุมการเรียกเข้าในระบบ ดีเอส-ซีดีเอ็มเอ ไร้สายโดยใช้การเรียนรู้แบบรีอินฟอร์สเมนท์ (CALL ADMISSION CONTROL IN WIRELESS DS-CDMA SYSTEMS USING REINFORCEMENT LEARNING). อาจารย์ที่ปรึกษา : ผศ. ดร. วิภาวี หัตถกรรม, 95 หน้า.

วัตถุประสงค์ของงานวิจัยคือ การหานโยบายที่ดีที่สุดที่เป็นไปได้ในการควบคุมการเรียกเข้าสำหรับผู้ใช้งานเสียงหลายระดับในระบบดีเอส-ซีดีเอ็มเอ ไร้สาย ซึ่งทำให้ผลรางวัลตอบแทนระยะยาวของระบบมีค่าสูงสุด โดยที่ยังสามารถคงเงื่อนไขบังคับของคุณภาพการให้บริการได้

การควบคุมการเรียกเข้านี้ได้กำหนดปัญหาเป็นแบบ การตัดสินใจแบบกึ่งมาคอฟ (semi-Markov decision process) โดยที่สองเงื่อนไขบังคับธรรมชาติวิสัยในระบบดีเอส-ซีดีเอ็มเอที่พิจารณาคือ เงื่อนไขบังคับของระดับของอัตราส่วนของสัญญาณต่อสิ่งแทรกสอด (signal-to-interference ratio) และเงื่อนไขบังคับของความน่าจะเป็นในการติดขัด (blocking probability) เพื่อหลีกเลี่ยงภาระการคำนวณของวิธีการไดนามิคโปรแกรมมิง (dynamic programming) เราได้นำวิธีการเรียนรู้แบบรีอินฟอร์สเมนท์แบบแอคเตอร์-คริติค (actor-critic reinforcement learning) เพื่อแก้ปัญหาการควบคุมการเรียกเข้า นอกจากนี้เราได้ทำการรวมเอาฟังก์ชันทำโทษเข้าไปในสัญญาณผลรางวัลตอบแทนเพื่อควบคุมเงื่อนไขบังคับของความน่าจะเป็นในการติดขัด ส่วนเงื่อนไขบังคับของอัตราส่วนสัญญาณต่อสิ่งแทรกสอดนั้นควบคุมโดยการกำหนดค่าในปริภูมิสเตต (state space) ของระบบ

จากผลการทดลองพบว่าอัลกอริธึมที่นำเสนอสามารถให้ผลที่ดีกว่าเทคนิคที่มีอยู่เดิมและสามารถเข้าถึง 91-95% ของผลจากวิธีการไดนามิคโปรแกรมมิงซึ่งเหมาะที่สุดและยังสามารถรักษาเงื่อนไขบังคับของคุณภาพการให้บริการโดยใช้ปริมาณการคำนวณและความต้องการในการเก็บข้อมูลในระดับที่พอเหมาะ

สาขาวิชาวิศวกรรมโทรคมนาคม ลายมือชื่อนักศึกษา_____

ปีการศึกษา 2549 ลายมือชื่ออาจารย์ที่ปรึกษา_____

PITIPONG CHANLOHA : CALL ADMISSION CONTROL IN WIRELESS DS-CDMA SYSTEMS USING REINFORCEMENT LEARNING. THESIS ADVISOR : ASST. PROF. WIPAWEE HATTAGAM, Ph.D. 95 PP.

DIRECT-SEQUENTIAL CODE DIVISION MULTIPLE ACCESS (DS-CDMA)/ CALL ADMISSION CONTROL/ REINFORCEMENT LEARNING/ ACTOR-CRITIC REINFORCEMENT LEARNING (ACSMDP)/ SEMI-MARKOV DECISION PROCESS (SMDP)

The underlying aim of this research is to find the best possible call admission control policy for multiclass voice services in wireless direct-sequential code division multiple access (DS-CDMA) systems that maximize the long-term reward of the system while satisfying multiple quality-of-service (QoS) constraints.

The call admission control problem is formulated as a semi-Markov decision process. Two important constraints inherent in CDMA systems are considered which are signal-to-interference ratio (SIR) constraints and blocking probability constraints. To circumvent the computational burden of conventional dynamic programming (DP) methods, we employ an actor-critic reinforcement learning method to solve the call admission control problem. Furthermore, we incorporate a penalty function into the reward signal in order to account for the blocking probability constraints. The SIR constraints are accounted for by embedding them into the system state space.

The numerical results show that the proposed algorithm can outperform existing techniques where it can achieve up to 91-95% of the optimal DP solution while maintaining the QoS requirement constraints with reasonable computational and storage requirements.

School of Telecommunication Engineering Student's Signature_____

Academic Year 2006 Advisor's Signature_____

ACKNOWLEDGEMENT

I am grateful to all those, who by their direct or indirect involvement have helped in the completion of this thesis.

First and foremost, I wish to express my sincere thanks to my thesis advisor, Asst. Prof. Dr. Wipawee Hattagam for her invaluable help and constant encouragement throughout the course of this research. She provided many insightful ideas and comments throughout my masters degree.

I would like to thank the lecturers in the School of Telecommunication Engineering, Asst. Prof. Dr. Rangsan Wongsan, Asst. Prof. Dr. Rangsan Tongta, Dr. Chutima Prommak, Dr. Chanchai Tongsopa and many others for suggestions and all their help.

I am grateful to the engineers, technicians and personnel at the Telecommunication Laboratory for their suggestions and help. My special thanks goes to Mr. Prapol Jaratakul for his full support.

Finally, I am most grateful to my parents and my friends both in both masters and doctoral degree courses for all their support throughout the period of this research.

Pitipong Chanloha

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

SYMBOLS AND ABBREVIATIONS

CAC	Call admission control
DS-CDMA	Direct-sequential code division multiple access
CDMA	Code division multiple access
RL	Reinforcement learning
ACSMDP	Actor-Critic reinforcement learning
SIR	Signal-to-interference ratio
DP	Dynamic programming
LP	Linear programming
QoS	Quality-of-service
SMDP	Semi-Markov decision process
$\{X_t\}$	A stochastic process at time t
X_t	State of the process at time t
$E\{.\}$	Expectation operator
x	State
a	Action
r	Reward
X	State space
A	Action space
π	Policy
Π	Policy space

SYMBOLS AND ABBREVIATIONS (Continued)

$v(\pi)$	Average reward for policy π
MDP	Markov decision process
$c(x,a)$	Expected cost incurred until the next decision epoch if action a is chosen in state x
TD	Temporal-difference
ω	Event
Ω	Event Space
$\Psi(i)$	SIR value of class i voice user
$\beta(i)$	Minimum SIR requirement for class i voice user
μ_θ	Randomize stationary policy parameterized by θ
∇	Gradient operator
τ	Transition time, expected time for each state
ψ^θ	Actor feature with parametric θ
ϕ^θ	Critic feature with parametric θ
LMMSE	Linear minimum mean square error
BEP	Bit error probability
λ	Arrival rate
μ	Service rate

SYMBOLS AND ABBREVIATIONS (Continued)

R_+	Set of positive real numbers
$I_+^{K_v}$	Set of vector of positive integer numbers of size K_v
K_v	Number of classes of voice users
$e(i)$	Vector with all zero components except for the i-th component which is unity
$p_{xy}(a)$	Transition probability that the state at the next decision epoch is y, given that an action a is selected at current state x
$\nu(i)$	Weight factor for blocking probability constraints of class i voice user
$B(i)$	Maximum allowable blocking probability for class i voice user
z_{xa}^*	Optimal solution for LP
$\bar{h}(i)$	Channel gain for voice class i user
$P(i)$	Transmission power for voice class i user
$\xi^2(i)$	Channel variance for voice class i user
N	Processing gain of the channel
$T(i)$	Threshold limit for class i voice user
χ	Lagrange multiplier

CHAPTER I

INTRODUCTION

This chapter provides a background on code division multiple access (CDMA) mobile communication systems and highlights the significance of call admission control in wireless CDMA systems and the motivation for applying reinforcement learning which is the main focus of this thesis.

1.1 Significance of the Problem

Since its introduction, cellular telephones have fascinated millions of subscribers in the United States, Asia and Europe (Chen, 1998). The first generation (1G) of cellular systems has been designed to mainly support voice users. Channel access is provided by frequency division multiple access (FDMA). In FDMA, a user is assigned a particular channel which is not shared by other users in the vicinity (Rappaport, 2002). However, as the number of subscribers increased, the frequency spectrum became exhausted which resulted in capacity limitation. The spectrum exhaustion of 1G cellular systems led to the development of digitized transmission in the second generation (2G) cellular systems which provided increased capacity and supported low rate data transfer. To access the channel, time division multiple access (TDMA) was employed. TDMA allows users to share the same frequency channel by dividing the signal into different time slots. Each user then takes turn transmitting and receiving over the channel in a round robin fashion. The 2G cellular systems support

voice, data and facsimile services—all of which require low data rate transfer. However, apart from voice and low rate data transfer services, mobile users in modern communication systems demand for support of more sophisticated services, such as, multimedia and video-on-demand, which require high bit rate transfer and therefore increased system capacity.

An alternative multiple access namely the code division multiple access (CDMA) scheme, has been recently developed to increase the capacity of cellular systems (LEE, 1991) in order to achieve high bit rate data transfer for supporting multimedia which consumes more bandwidth. CDMA uses unique spreading codes to spread the baseband data before transmission. The receiver then dispreads the wanted signal which is then passed through a narrow band pass filter. The unwanted signals are not dispreaded nor passed through the filter. The main advantages of CDMA are:

1. Soft capacity limit – Unlike FDMA and TDMA, an increase of the number of users in CDMA system raises the noise level in the system. Hence, the CDMA system capacity can be increased however at the cost of degraded signal quality. This is referred to as the *soft* capacity limit. Therefore, unlike FDMA and TDMA systems, there is no *hard* limit on the number of users in CDMA systems. Note that the users' signal-to-interference ratio gradually degrades as the CDMA capacity increases and improves as the capacity of the system decreases.
2. RAKE receiver – One of the main advantages of CDMA systems is the capability of using signals that arrive at the receivers with different time delays. This phenomenon is called multipath. More specifically, multipath is a scenario where radio signals reach the receiving antenna via different

paths which merely provide multiple versions of the transmitted signal at the receiver. Multipath causes interference and phase shifting of the signal. To correct multipath effects, a RAKE receiver can be used by collecting time delayed versions of the required signal. RAKE receivers employ a set of sub-receivers to search for different multipaths and feed the information to the other sub-receivers. Each sub-receiver then finds a strong multipath signal. The results are then combined together to create a signal with higher signal-to-noise ratio.

3. Power control – One of the major problems in mobile communication is the near-far problem. The near-far problem occurs when many mobile users share the same channel. Consider a receiver at a base station which serves many mobile users. Assuming that all users transmit at the same power level, the receiver at the base station will receive stronger signals from nearer users than users further away. In fact, a user's transmitted signal appears as other users' interference. If the near users transmit at several orders of magnitude higher than users further away, then the signal-to-interference ratio (SIR) of further away users may become undetectable. To reduce the interference from mobile users transmitting at different power levels, CDMA employs a power control scheme. Under such scheme, the base station controller (BSC) which controls all radio transmission will dynamically adjust power the nearer users so that the SIR levels of all users are roughly the same.
4. Soft handoff – In non-CDMA cellular systems which employ traditional hard handoff, each mobile user makes a connection to one base station at a

time. It is only when the connection to the current cell is broken, a connection to the new cell is then established. However, connection re-establishment to the new cell may not always be successful due to multipath fading and insufficient capacity at subsequent cells. Hence, ongoing calls may be interrupted or even forced terminated. Alternatively, CDMA systems employ *soft handoff* where a mobile user is simultaneously connected to two or more cells. The soft handoff is performed by the Mobile Switching Center (MSC) which chooses the best version of the signal at anytime without switching frequencies. Therefore, soft handoffs can help reduce the number of forced terminated handoffs.

5. Security broadcasting – In FDMA and TDMA cellular systems, data transmission security is usually not considered. CDMA cellular systems, on the other hand, can improve transmission security owing to its use of spreaded codes such as pseudo noise (PN) long code, PN short code, etc. For example, the PN long code has more than 4.4 billions codes.

The aforementioned main advantages allow CDMA to support the various classes of multimedia traffic, such as, voice, video streaming, images, web documents, data or a combination thereof (Bartolini and Chlamtac, 2002). Such traffic requires high bit rate which FDMA and TDMA cannot satisfy. Furthermore, CDMA networks can also support multiple classes of traffic with different quality-of-service (QoS). Suppose a mobile user requests to join the CDMA network. The base station then makes a decision whether to admit or reject the call request. This process is called call admission control (CAC) which is employed to administrate the required level of QoS. The basic of CAC mechanism is to admit a new user into the system

only when the QoS constraints of all existing users in the system and that of the new user are satisfied. Therefore, CAC requires quantification of QoS constraints.

One significant QoS constraint in CDMA systems is in the *physical* layer in terms of *signal-to-interference ratio* (SIR). It is the minimum SIR requirement that governs the system capacity in CDMA systems. The call admission controller rejects calls to maintain the SIR requirement. This results in an increase in *blocking probability*, which is in many cases, an important QoS constraint in the *network* layer. Hence, unlike CAC in other cellular systems, there exists a significant interplay of the physical layer (SIR) and the network layer (blocking probability) QoS requirements in CDMA CAC mechanisms.

It should be noted that there are many existing approaches which investigated CAC in mechanisms in communication networks. These CAC methods can be classified as complete sharing, threshold policy and the semi-Markov decision process (SMDP) approach.

1.1.1 Complete Sharing

Complete sharing is the simplest CAC policy which always accepts new users as long as there is sufficient capacity, or in the case of CDMA systems, as long as the SIR requirements of all users in the system are satisfied. Comaniciu and Narayan (2000), Lee and Wang (1998) and Liu and Zarki (1994) investigated the CAC problem in CDMA systems by considering the SIR level in the physical layer alone. They do not consider the problem of finding an admission policy that considers the blocking probability of the system which is a functionality in the network layer. Complete sharing CAC policies are generally suboptimal. This policy can be easily

implemented but it cannot satisfy any QoS constraints—apart from the SIR level requirements.

1.1.2 Threshold Policy

Threshold policy is a call admission policy which accepts a user of some class k if the number of users in such class is less than a threshold T_k (Liu and Silvester, 1998) and (Soroushnejad and Geraniotis, 1995). However, similar to the complete sharing policy, it is generally known that threshold policies cannot satisfy blocking probability constraints (Ross, 1995) either—unless it is found through brute force empirical search. Therefore, threshold policies are impractical for maintaining QoS requirements in actual networks.

1.1.3 Semi-Markov Decision Process (SMDP)

From the two aforementioned approaches, the blocking probability QoS constraints cannot be easily satisfied in any approach. Alternatively, the semi-Markov decision process approach, on the other hand, can deal with multiple QoS constraints and can also guarantee the optimal call admission control policy. Several existing works have been proposed to find the optimal CAC policy using SMDP framework. These works can be classified as follows.

1.1.3.1 CAC in Cellular Network

Choi J., Kwon, Choi, Y. and Naghshineh (2000), Singh and Bertsekas (1997) and El-Alfy, Yao and Heffers (2001) have investigated the CAC problem in cellular networks using the SMDP framework. To solve the CAC problem, Choi et al. (2000) employed the linear programming method which can deal with QoS constraints on call blocking probability and call dropping probability. On the other hand, Singh and Bertsekas (1997) and El-Alfy et al. (2001) have applied an

alternative tool called reinforcement learning to solve the CAC problem in cellular networks with capacity constraints and call dropping probability constraints. These two works have dealt with CAC problems in generic cellular networks which are not specifically CDMA networks.

1.1.3.2 CAC in CDMA with Blocking Probability Constraints Only

Yang and Geraniotis (1994) dealt with the CAC problem which considered the blocking probability constraints alone. To solve the CAC problem, the authors used value iteration technique which is a dynamic programming tool. Makarevitch (2000) also dealt with the CAC problem which considered the power control affect and QoS constraints in terms of blocking probability constraints only. This work employed reinforcement learning to solve for the near-optimal CAC policy.

1.1.3.3 CAC in CDMA with Dual Constraints

Comaniciu and Poor (2003) investigated the joint optimization call admission control problem in multiple data services and dealt with dual constraints which are SIR levels and blocking probability constraints. To solve for the solution, linear programming which is a dynamic programming technique is employed.

Singh, Krishnamurthy and Poor (2002) employed a conventional dynamic programming (DP) method to solve for an optimal CAC policy. Given an explicit model of the system, i.e., the transition probability matrix and the expected rewards, DP method is guaranteed to deliver an optimal CAC policy. However, DP has limited applicability due to the curse of dimensionality and curse of modeling (see **Chapter 2** for details). Therefore, the approach of Singh et al. (2002) inevitably becomes too complex to solve when the scale of the CAC problem increases.

To circumvent the computational burden of DP, this thesis proposes an alternative approach based on reinforcement learning (RL) (Sutton and Barto, 1998) to determine near-optimal CAC policies instead. RL methods can provide near-optimal solutions to complex DP problems through experience learned from simulations and direct interaction with the environment. Consequently, they do not require an explicit model of the system. The scalability of RL is therefore better than classical DP methods. RL has already been successfully applied to solve CAC problems in many communication networks, such as, in ATM networks (Tong and Brown, 2000), non-CDMA cellular networks (Singh and Bertsekas, 1997 and El-Alfy et al., 2001) and CDMA networks (Makarevitch, 2000 and Vazquez-Abad and Krishnamurthy, 2002).

More specifically in CDMA networks, Makarevitch (2002) investigated the CAC problem with blocking probability constraint alone. They employed RL to solve for a near-optimal CAC policy. Vazquez-Abad and Krishnamurthy (2002) handled the dual constraints, i.e., SIR and blocking probability constraints. Their work also employed RL to solve for a near-optimal CAC policy. However, their proposed work is table-based meaning that a parameterized component is needed for every possible system configuration. Hence, as the scale of the CAC problem increases, the scalability of their approach will become a concern—particularly in terms of computational and storage requirements.

In response to these outstanding issues, the aim of this thesis is to develop an online call admission decision-making algorithm for multiple voice services in a wireless DS-CDMA system which has low computational and storage requirements and maximizes the long-term performance criterion while satisfying dual QoS

constraints on the SIR level and the required blocking probability. The contribution of this thesis is placed on the development of an actor-critic RL method (The details of actor-critic can be found in **section 2.5**) which deals with multiple QoS constraints. The proposed method is an extension of Usaha and Barria (2007) which developed an actor-critic RL method and applied it to solve a call admission control and routing problem in low-earth orbit satellite networks. Whereas Usaha and Barria (2007) considered an *unconstrained* SMDP problem, the proposed method in this thesis differs from Usaha and Barria (2007) work where we consider a *constrained* SMDP problem instead. The actor-critic method proposed here differs from Usaha and Barria (2007) where the reward is modulated to account for the constraints (Chanloha and Usaha, 2007). The results in this thesis show that the proposed actor-critic algorithm can satisfy the dual constraints on the SIR level and blocking probability. Furthermore, our proposed method employs function approximation which has the advantage of scalability when compared to Vazquez-Abad and Krishnamurthy (2002) which also employed RL to deal with dual constraints for CAC in CDMA networks.

1.2 Research Objectives

The objectives of this research are as follows:

1.2.1 To apply reinforcement learning (RL) to solve the CAC problem in wireless CDMA systems supporting multiple class voice users subject to the dual QoS constraints on blocking probability and SIR requirements.

1.2.2 To apply reinforcement learning (RL) to alleviate the curse of dimensionality and the curse of modeling of dynamic programming (DP).

1.2.3 To compare the reinforcement learning solution with dynamic programming solution in terms of average reward, computational complexity and memory storage.

1.3 Assumptions

1.3.1 CAC in CDMA systems can be formulated as a SMDP problem.

1.3.2 Reinforcement learning can reduce the computational complexity and memory storage of the solution when compared to dynamic programming.

1.3.3 Reinforcement learning can achieve a near-optimal CAC policy and can satisfy multiple QoS constraints.

1.4 Scope of the Thesis

This thesis consists of two main parts. Firstly, the call admission control problem for a small multiclass voice service CDMA system is formulated as a SMDP and is solved with a conventional dynamic programming method. A classical dynamic programming method, i.e. linear programming, will be compared to the complete sharing and threshold policy CAC methods. Among these three methods, the dynamic programming method can give the optimal CAC policy. The dynamic programming method in this part is obtained from Singh et al. (2002) and deals with multiclass voice services only. To quantify the performance, we compared two metrics, namely, the blocking probability and average long-term reward. We use the blocking probability metric to demonstrate the ability to maintain the QoS requirements, whereas the average long-term reward metric is used to demonstrate the optimality of the policy. The SIR constraint (Details of the SIR computation can be found in

appendix I) which is the other QoS constraint considered is embedded in the state space of the system.

In the second part, we extended the small scale CAC problem to a more realistic scenario by significantly increasing the state space. An actor-critic RL technique is proposed to solve the call admission control problem in the large scale network. In this part, we deal with QoS constraints by proposing a modification of the reward to account for the constraints on blocking probability. The SIR constraint is, however, embedded in the state space of the system. The storage and computational requirements and numerical results for the DP method, threshold policy and the proposed actor-critic RL method are compared and analyzed.

1.5 Expected Usefulness

1.5.1 To obtain a call admission control algorithm by using RL to control the desired blocking probability while the SIR requirements are satisfied in DS-CDMA systems.

1.5.2 To obtain a call admission control algorithm by using RL which can reduce the high computational complexity and memory storage of the conventional dynamic programming in DS-CDMA systems.

1.5.3 To obtain a conclusion about the application of reinforcement learning in DS-CDMA networks and suggest its possible applications to other call admission control problems, for example, to deal with multimedia traffic such as multiclass voice and data traffic.

1.6 Organization of Thesis

The remainder of this thesis is organized as follows. **Chapter 2** presents the theoretical background which underlies the contribution of this thesis. Firstly, the concept of dynamic programming (DP) method is reviewed. This is followed by an introduction of reinforcement learning (RL) which deals with the dual curses (The curses of DP can be found in **section 2.3**) of DP which are the curse of dimensionality and the curse of modeling. RL can avoid the curses of DP method by directly learning through experience from simulation or interaction with the environment. We then classify three categories of RL which are actor-only methods, critic-only methods and actor-critic methods and justify the reasons for selecting the latter method in this research.

In **Chapter 3**, we study the call admission control problem (CAC) with multiple voice services in wireless DS-CDMA network. Already addressed and solved in Singh et al (2002), the CAC problem is formulated as semi-Markov decision process (SMDP) problem and solved analytically by using a DP method called linear programming (LP). The objective of this chapter is to demonstrate that the CAC policy obtained by the DP method is the optimal policy. Two metrics are considered which are the blocking probability and average long-term reward. The SIR constraint is inherently embedded in the system capacity of the CDMA system. Numerical results show that the DP CAC policy outperforms the complete sharing and threshold policy methods. However, we point out that the storage and computational requirements of the DP method becomes prohibitive and can grow intractably as the scale of the system increases.

To avoid the computational burden caused by DP, an alternative method, namely, the actor-critic RL method is proposed in **Chapter 4** to solve the CAC problem. By using function approximation, the proposed method needs only a small number of parameters for making each CAC decision. The performance metrics are blocking probability and the average long-term reward. The SIR level constraint is embedded in the state space of the system. Numerical results are obtained from a significantly larger scale system. Results show that our proposed actor-critic RL method can achieve near-optimal CAC policy, satisfy QoS constraints and demand less memory storage and low computational complexity when compared to DP.

Chapter 5 summarizes all the findings and original contribution in this thesis and points out possible future research direction.

CHAPTER II

BACKGROUND THEORY

2.1 Introduction

In this thesis, we study the call admission control (CAC) problem in wireless DS-CDMA networks. A common approach to solve call admission control problems is to formulate the problem as a semi-Markov decision process. This is due to the fact that the call admission decisions occur at instances whereby the system exhibits Markov property. Analytical tools such as dynamic programming (DP) can then be applied to solve for the optimal policy. It is generally known that the computation of this technique needs an explicit model of the SMDP process. The explicit model refers to the state transition probability and expected reward which grow exponentially as the size of the state space increases. DP therefore becomes impractical or even intractable to solve as the dimension of the problem increases.

An alternative method to avoid the computational burden of dynamic programming method is reinforcement learning (RL). RL evaluates the optimal policy by experience sampled from simulation or direct interaction with the environment. Such sampling allows RL to *learn* the state dynamics model instead of computing the exact analytical model as in dynamic programming. In addition, RL can also employ function approximation to deal with large state space. Function approximation is extremely important as it allows experience learned from a limited subset of the state

space to be usefully generalized to produce a good approximation of decision variables over a much larger subset.

This chapter therefore provides an essential foundation of Markov decision theory framework and presents algorithms which are used to solve it in this thesis. This chapter is organized as follows. The next section gives a theoretical background on Markov decision processes. Section 2.3 describes the dynamic programming concept which is used to solve the SMDP formulation. An introduction of reinforcement learning is given in section 2.4. In section 2.5, the actor-critic reinforcement learning which the contribution of this thesis is based on is presented. Finally, the conclusion of chapter is in the last section.

2.2 Markov Decision Theory Background

As mentioned in the previous section, CAC problems in communication networks can be viewed as a SMDP process (Singh, Krishnamurthy and Poor, 2002). The basic idea of SMDP lies on Markov property which states that the probability of visiting the next state depends only on the present state of the system.

2.2.1 Markov Property

Let $\{X_t\}$ be a stochastic process where X_t refers to the state of the process at any time t. If the future of the process, given that the process is presently in state X_{t_k}, is independent of the past, then $\{X_t\}$ is called a Markov process. That is, $\{X_t\}$ is a Markov process if

$$\begin{aligned} P[X_{t_{k+1}} = x_{k+1} \mid X_{t_k} = x_k, ..., X_{t_1} = x_1] \\ = P[X_{t_{k+1}} = x_{k+1} \mid X_{t_k} = x_k] \end{aligned} \tag{2.1}$$

where $t_1 < t_2 < \cdots < t_k < t_{k+1}$, t_k is the present time and t_{k+1} is the time instant in the future. Equation (2.1) is referred to as the Markov property. In other words, a stochastic process has Markov property if the probability distribution of future states of the process, given the present state and all past states, depends only upon the present state and not on any past states.

2.2.2 Markov Decision Process (MDP)

A Markov decision process (MDP) is a discrete-time stochastic process characterized by a set of states, actions and rewards. Let $\{X_n\}$ be a discrete value Markov chain[1] for all n, where n is the present and $n+1$ is the future (discrete) time index. At each state, the decision-maker can select an action from a set of permissible actions at the given state. For a state x and an action a, a state transition function $p_{xx'}(a)$ defines the transition probability to the next state x',

$$p_{xx'}(a) = P[X_{n+1} = x' \mid X_n = x, a_n = a] \tag{2.2}$$

As a result of taking action a at state x and transiting into state x', the decision-maker earns an expected reward $r_{xx'}(a)$ given by

$$r_{xx'}(a) = E\{r_{n+1} \mid X_n = x, a_n = a, X_{n+1} = x'\} \tag{2.3}$$

where $E\{.\}$ is the expectation operator and r_{n+1} is the reward earned where $n+1$ is the future discrete time index. Note that equations (2.2) and (2.3) completely specify the dynamics of the Markov decision process. Dynamic programming requires the

[1] A Markov chain is a series of states of a system that has Markov property.

exact knowledge of these two functions in order to determine the optimal policy. Reinforcement learning, on the other hand, does not as RL learns these functions from interacting with the environment. However, before we proceed to determine an optimal policy for the MDP, some measure of how good a policy is must be established. The next subsection describes such policy measure.

2.2.2.1 Value Function

Value functions are estimations of *how good* it is to be in a given state x. A value function is defined as the amount of future reward that can be expected. Since whatever actions are taken affect the rewards received in the future, policies which govern what actions to take at a given state will characterize the value function accordingly. Suppose that π is a policy that maps a state $x \in X$ to some action $a \in A(x)$, where $A(x)$ is the set of available actions in state x, to the probability $\pi(x,a)$ of taking action a when in state x. Denote $h(x,\pi)$ as the expected return when starting in x and following policy π thereafter, given by

$$h(x,\pi) = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^{k} r_{n+k+1} \middle| x_n = x\right\} \tag{2.4}$$

where $E_{\pi}\{.\}$ is the expectation operator given that the actions taken follow policy π, and n is any time step. The function $h(x,\pi)$ is referred to as the *state value function* for policy π at state x. The fundamental property of value functions used throughout reinforcement learning and dynamic programming is that they satisfy a particular recursive relationship. In particular, the relationship between the value function of state x and the value function of its possible successor states, is given by

$$h(x,\pi)+v(\pi)=r_x(a)+\sum_{x'}p_{xx'}(a)h(x',\pi) \tag{2.5}$$

where x' is the successor state, $v(\pi)$ is the average reward following some policy π, $r_x(a)=\sum_{x'}p_{xx'}(a)r_{xx'}(a)$, $p_{xx'}(a)$ is the probability of selection action a from current state x to state x'. Equation (2.5) is called the *Bellman equation* for $h(x,\pi)$.

2.2.2.2 Optimal Value Function

Now that the notion of value function has been established, the notion of optimal value function can now be introduced. The policy π^* is said to be average reward optimal if and only if $v(\pi^*)\geq v(\pi)$ for every other policy π, where $v(\pi)$ is the average reward given by

$$v(\pi)=\lim_{N\to\infty}\frac{1}{N}E_\pi\left\{\sum_{k=0}^{\infty}r_{k+1}\right\} \tag{2.6}$$

Furthermore, the optimal value function $h(x,\pi^*)$, $\forall x\in X$ can then be defined as

$$h(x,\pi^*)+v(\pi^*)=\max_{a\in A(x)}\left\{r_x(a)+\sum_{\forall x'\in X}p_{xx'}(a)h(x',\pi^*)\right\} \tag{2.7}$$

$$h(\tilde{x},\pi^*)=0$$

where $\tilde{x}$ is the recurrent state. The function $h(x,\pi^*)$ can be interpreted as follows. Under policy π^*, $h(x,\pi^*)-h(x',\pi^*)$ is the difference in the optimal expected reward over an infinitely long-time when starting in state x rather than state x', where x, $x'\in X$. Note that the optimal value function is the unique solution of

the Bellman equation in (2.5). The optimal value function in (2.7) is called the *Bellman optimality equation* for the average reward MDP criterion.

The objective of MDP formulated problems is to find the best possible policy that maximizes the long-term average reward for *discrete-time* stochastic processes. However, the CAC problem investigated in this thesis is viewed as a *continuous-time* stochastic process. Even so, some researches have formulated the call admission control problem in CDMA networks with the discrete-time MDP model before. For example, Sung, Hwang, Chen and Hsu (2004) has formulated CAC problem with channel reservation in CDMA systems and solved it by policy iteration which is a dynamic programming method. Makarevitch (2002) studied the CAC problem in CDMA systems and solved it by reinforcement learning.

As aforementioned, CAC problems are generally viewed as continuous-time stochastic processes. A continuous-time version of MDP called the semi-Markov decision process (SMDP) framework is often used.

2.2.3 Semi-Markov Decision Process (SMDP)

In the previous section, we have introduced a discrete-time Markov decision model where the decisions can be made only at fixed epochs $n = 0,1,\ldots$. However, in many sequential decision-making problems, the time between each consecutive decision epoch is not identical but random. In fact, the random duration between epochs is drawn from a general probability distribution which may or may not be independent of the past history. The *semi-Markov decision process* is generally used to model such problems.

Consider a Markov process $\left\{X_{t_k}\right\}$ whose state transition follows the transition probability matrix controlled by policy π denoted by P^{π}. Let the set of

possible states be denoted by X. Suppose that at time t_k, the system is in state $X_{t_k} = x \in X$, action $a \in A(x)$ is chosen where $A(x)$ is the set of all available actions in state x, and the system transits into a new state $x' \in X$ with probability $p_{xx'}(a) \in P^{\pi}$. Associated with the transition is the expected reward of $r_{k+1} = r_{xx'}(a)$. In this thesis, we consider optimizing the average reward criterion. Under policy π, the average reward SMDP criterion is given by

$$v(\pi) = \lim_{N \to \infty} \frac{E_{\pi} \left\{ \sum_{k=0}^{\infty} r_{k+1} \right\}}{t_N}, \tag{2.8}$$

where π is a *stationary policy*[2] that maps a particular state into a particular action, t_N is the duration of the sequence of *N* state transitions. Under the unichain assumption which states that under any stationary policy π, a state can be reached by any other state under π, the limit in the above equation exists and is independent of the initial state. The objective of the SMDP formulation is to find an optimal stationary policy π^* that maximizes the average reward criterion such that $v(\pi^*) \geq v(\pi)$ for every other policy π.

2.3 Dynamic Programming (DP)

Dynamic programming (DP) is an analytical method used to solve for the optimal policy which is the solution for the objective function in (2.8). DP is tailored

[2] Let π_k be the policy at time t_k. A set of polices $\{\pi_0, \pi_1, ..., \pi_{N-1}\}$ is said to be the *stationary policy* if and only if $\pi_0 = \pi_1 = ... = \pi_{N-1} = \pi$.

to solve sequential decision-making problems under uncertainty (Uncertainty in real-world problems are modeled by transition probabilities which are often inaccurate). Efficient numerical computation techniques which have been developed to solve the DP optimization problems include value iteration (Tijms, 1986), policy iteration (Sutton and Barto, 1998) and linear programming (Bertsekas, 1995). Note that linear programming has the advantage of having widely available codes, whereas value and policy iteration usually involves the writing of its own code. The number of iterations required by linear programming depends heavily on the specific problem considered, whereas the policy iteration algorithm and value iteration requires typically only a very small number of iteration regardless of the problem size (Tijms, 1986).

Consider a Markov chain in section 2.2.3 whereby the average reward criterion in (2.8) under some stationary policy π holds. To solve for the optimal policy π^*, one must solve for all the $|X|+1$ unknowns (which are $v(\pi^*)$ and $h(x,\pi^*), \forall x \in X$) to the following Bellman optimality equation for the average reward *SMDP* criterion,

$$h(x,\pi^*)+v(\pi^*)\tau(x,\pi^*)=\max_{a\in A(x)}\left\{r_x(a)+\sum_{\forall x'\in X}p_{xx'}(a)h(x',\pi^*)\right\}, x\in X, \quad (2.9)$$

$$h^*(\tilde{x})=0,$$

where $\tilde{x}$ is the recurrent state, $\tau(x,\pi^*)$ is the expected time that the system remains in state x under policy π^* and $|X|$ is the size of the system state space.

Note that DP is able to attain an optimal solution to SMDP formulated problems. However, it is generally known that for complex systems with large state

spaces, dynamic programming methods demand enormous amount of computation for developing transition probabilities and expected reward expressions. This problem is called *Bellman's curse of dimensionality*. Moreover, DP also requires an analytical model of state transitions which is usually hard to identify in real applications. This problem is called the *curse of modeling*. DP methods are therefore difficult to implement in actual networks. In light of the limitations of DP methods, an alternative method which is based on experience learned from interacting directly with the environment or simulation is presented in the next section.

2.4 Reinforcement Learning

Reinforcement learning (RL) is a computational approach for automated goal-directed learning and decision-making (Sutton and Barto, 1998), in order to maximize a numerical reward signal. Reinforcement learning is a type of unsupervised learning system. RL provides new methods to deal with the curses of DP problem by finding and achieving a near-optimal set of actions through experience instead. Such experience is learned through interaction between the agent or decision-maker and the environment. Figure 2.1 shows the agent-environment interaction in reinforcement learning. Let x_t, a_t and r_t be the state, action and reward incurred at time t, respectively. At each time step t, the agent receives some representation of the environment's state x_t and selects an action a_t. One time step later, the agent receives a numerical reward r_{t+1} and finds itself in a new state x_{t+1}. The immediate reward is returned to evaluate the action taken by the agent from the environment. These events interact continually and the agent's goal is to maximize (minimize) the total amount

of reward (cost) it receives over the long run (Sutton and Barto, 1998). RL methods can be broadly classified into three main categories as follows.

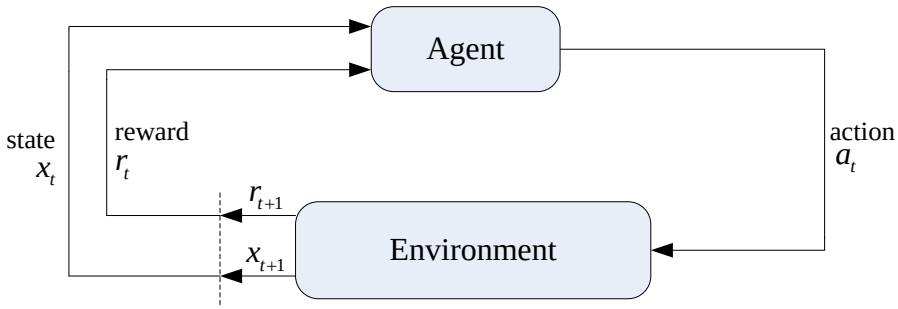

Figure 2.1 Diagram of agent-environment interaction in reinforcement learning

2.4.1 Actor-Only Methods

Actor-only methods operate under a parameterized family of probabilistic policy to select an action in each state (Marbach and Tsitsiklis, 1999). The parameterized components are updated in the direction that improves the gradient of some performance measure with respect to the parameter. These methods therefore update the policy *directly*. However, the general weakness of the actor-only methods is the large variance of the gradient estimators obtained directly from simulations which results in a slow learning rate.

2.4.1.1 Actor-only methods for CAC in CDMA networks

Vazquez-Abad and Krishnamurthy (2002) proposed a CAC policy for multiple voice and data services in CDMA systems based on an actor-only method. Their method has the ability to deal with both SIR requirements and blocking probability constraints. However, the shortcoming of their proposed method is that it is table-based—meaning that a parameterized component is required for every state and action pair in the system. Consequently, the scalability of the approach in

Vazquez-Abad and Krishnamurthy (2002) will become a concern when the size of the system increases.

Lui, Zhang Y. and Zhang H. (2005) has investigated the multiple voice and data services CAC problem in CDMA systems. They employed an adaptive self-learning control which determines the grade-of-services (GoS) in blocking probability terms of new call and handoff calls. The probability of selecting an action for whether admitting or rejecting a call in this work is fixed at 0.5. The shortcoming for this work is the time spent in learning the network before it can be used.

2.4.2 Critic-Only Methods

Critic-only methods rely on approximations of value functions and aim to solve for an approximation to the Bellman equation in (2.5) (Sutton and Barto, 1998), (Bertsekas and Tsitsiklis, 1996). Then, a greedy policy based on the approximated value function is applied aiming at improving the policy currently being followed. Critic-only methods update the policy *indirectly* through learning the value function approximations. Convergence to a near-optimal policy can be achieved in a timely manner. However, for these methods, policy improvement is not always guaranteed, even if good approximations of the value functions are obtained. Policy improvement is only guaranteed in limited settings (Bertsekas and Tsitsiklis, 1996). Successful applications of critic-only RL methods to solve for CAC policies in cellular networks include El-Alfy, Yao and Heffers (2001) and Lilith and Dogancay (2005).

2.4.2.1 Critic-only methods for CAC in CDMA networks

Makarevitch (2000) used a critic-only method to obtain a call admission control policy for a multiple class voice services in CDMA system which is formulated as a MDP. This work investigated the power control effects and network layer constraints in terms of blocking probability constraints only.

Liao, Yu, Leung and Chang (2006) has investigated the CAC problem in CDMA systems under dynamic cell configuration. The problem is formulated as a MDP with no blocking probability constraints. This work differs from Makarevitch (2000) where Liao, et al (2006) additionally considered the maximum link power constraints in their problem.

2.4.3 Actor-Critic Methods

Actor-critic methods combine the strong feature of the two previous methods together. The critic part attempts to learn value functions from simulation and uses them to update the actor's policy parameters in the direction of improvement of the performance measure gradient. Policy improvement is guaranteed as long as the policy is gradient-based which is a strong feature of actor-only methods. Faster convergence is achieved by using approximated value functions which is a strong feature of critic-only methods (Bertsekas and Tsitsiklis, 1996). An actor-critic reinforcement learning has been employed to solve the CAC and routing problem in low earth orbit satellite networks (Usaha and Barrier, 2007). However, this work formulated the CAC problem as an unconstrained SMDP. To the best of our knowledge, there has yet been any work which applied actor-critic RL methods to SMDPs with multiple constraints.

2.4.3.2 Actor-Critic methods for CAC in CDMA networks

Pandana and Liu (2004) has proposed an actor-critic RL approach in a mobile transmission problem for one base station which aims at maximizing an average reward criterion. This work considered the SIR level constraints which determines the capacity of the CDMA system. However, this work did not consider any blocking probability constraints.

From the aforementioned works which employed RL, none of these works have employed the actor-critic RL algorithm to solve the CAC problem in CDMA systems with the dual constraints on the SIR and call blocking probability before. This led us to consider the CAC problem with dual constraints and develop an actor-critic approach to solve the problem. A brief introduction of our proposed method is presented in the next section.

2.5 Actor-Critic Method in this Thesis

The actor-critic methods combine two strong features of the critic-only and the actor-only methods together which are fast convergence and guaranteed performance improvement. Figure 2.2 shows the diagram of the actor-critic architecture. The architecture is comprised of two parts, namely, the critic and actor part. Upon an action selection, the critic part estimates the value function which predicts future reward from the temporal-difference (TD) error. The TD error is used to evaluate how well the selected action was. A positive TD error means that the tendency to select that action in the future should be encouraged. On the contrary, if the TD error is negative, the tendency in selecting that action should be discouraged. The TD error is calculated from the difference between the immediate reward and predicted reward.

Afterwards, the TD error value is then fed back to the critic part which then drives the probability of selecting actions in the actor part (Sutton and Barto, 1998).

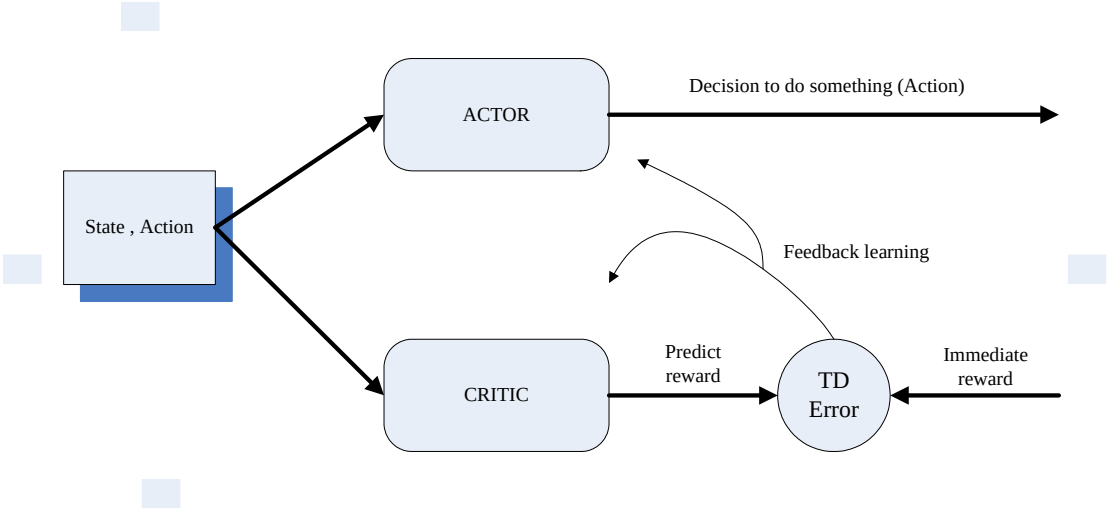

Figure 2.2 Diagram of actor-critic architecture

As stated earlier in **section 2.4.1,** the natural weakness of the actor-only methods is the large variance of the gradient estimator obtained directly from simulation which results in a slow learning rate. To improve the learning rate, the estimated value function should be involved. The estimated value function gives a fast convergence rate which helps expedite the learning rate which is a strong characteristic of the critic-only methods (Usaha, 2004).

However, as mentioned in **section 2.4.2,** policy improvement in critic-only methods is not always guaranteed. The reason is that the improvement of policy experiences an *abrupt change*. This has been identified as the key reason why policy improvement in critic-only methods is guaranteed in some limited cases only (Bertsekas and Tsitsiklis, 1996). To avoid abrupt policy changes, a probabilistic policy should be used. The probabilistic policy is then gradually updated in the direction that improves the performance measure. Under the condition that the

probabilistic policy is gradient-based, policy improvement is guaranteed which is a strong characteristic of actor-only methods.

To avoid the drawbacks caused by actor-only and critic-only methods, the actor-critic reinforcement learning is therefore selected in this thesis. Hence, by employing the actor-critic method in the policy search, a near-optimal solution can be achieved. In the following subsections, we proceed to present the actor-critic method, the performance criterion of which we wish to improve in this thesis and its gradient.

2.5.1 Average Reward Criterion

Consider a Markov decision process with finite state space X and finite action space A. Let μ_θ be a probabilistic policy which belongs to a family of policies parameterized by vector $\theta \in R^M$. That is, $\mu_\theta\left(a|x\right)$ is a mapping from the current state of the process to a distribution of actions. Suppose that the transition probability transition matrix of this Markov decision process is P_θ whose elements are the probability of transiting from state x into state x' when the actions taken are controlled policy by μ_θ. Upon selecting each action $a \in A(x)$ at state $x \in X$, a reward $g\left(x,a\right)$ is generated. The objective is to find the average reward criterion for a randomize stationary policy that $v\left(\theta^*\right) \geq v\left(\theta\right)$ for every θ which is given by

$$v\left(\theta\right) = \lim_{N\to\infty} \frac{1}{N} E_{\mu_\theta}\left\{\sum_{k=0}^{N-1} g\left(x_k, a_k\right)\right\} \tag{2.10}$$

where $E_{\mu_\theta}\left\{.\right\}$ is expectation operator and N is the number of transitions.

2.5.2 Gradient Estimation

Let ∇ denote the gradient with respect to the parameterized vector θ. The parameterized policy, μ_θ, can be tuned by improving the gradient of the average reward which is given by (Usaha, Barria 2007).

$$\nabla v(\theta) = \frac{\sum_{\forall x \in X} \sum_{\forall a \in A(x)} p_\theta(x) \mu_\theta(a|x) \psi^\theta(x,a) Q^\theta(x,a)}{\sum_{\forall x \in X} p_\theta(x) \sum_{\forall a \in A(x)} \mu_\theta(a|x) \bar{\tau}(x,a)} \tag{2.11}$$

where $Q^\theta(x,a)$ is an action-value function of starting in state-action pair (x,a) and following policy μ_θ thereafter, $\bar{\tau}(x,a)$ is the average transition time corresponding to state-action pair (x,a) and

$$\psi^\theta(x,a) = \frac{\nabla \mu_\theta(a|x)}{\mu_\theta(a|x)} \tag{2.12}$$

2.5.3 The Actor

The parameter $\psi^\theta(x,a)$ in (2.12) is the actor feature of the actor-critic architecture. Note that at any time step, the parametric vector θ controls μ_θ (i.e. the policy of the actor) and $\psi^\theta(x,a)$ (i.e. the estimated gradient $\nabla v(\theta)$). Hence, by tuning the parametric vector θ, the estimated gradient can be improved. In particular, at the k-th transition occurring at time t_k, the parametric vector θ_k is updated by

$$\theta_{k+1} = \theta_k + \eta_k \Gamma(r_k) \tilde{Q}_{r_k}^{\theta_k}(x_{k+1}, a_{k+1}) \psi^{\theta_k}(x_{k+1}, a_{k+1}) \tag{2.13}$$

where $\tilde{Q}_{r_k}^{\theta_k}$ is the estimated action-value function at time t_k, η_k and $\Gamma(r_k)$ are the stepsize parameters, r_k is the parameterized vector of the critic at time t_k.

2.5.4 The Critic

The job of the critic part is to critique the actor. In other words, the critic quantifies how well the action of the actor is. This is done by maintaining action-value functions $Q^{\theta}(x,a)$ which is a measure of how good it is to perform a given action a in a given state x and following policy μ_θ thereafter. In the case where the transition probability matrix P_θ is unknown (this is generally the case when the system dynamics is generally too complex to extract that one must resort to simulation or direct interaction with the system), the critic must approximate $Q^{\theta}(x,a)$ by an estimated action-value function. Such approximation at any time step t_k is denoted by the parameter $\tilde{Q}_{r_k}^{\theta_k}(x_k,a_k)$ in the below equation

$$\tilde{Q}_{r_k}^{\theta_k}(x_k,a_k) = r_k^T \phi^{\theta_k} = \sum_{j=0}^{K-1} r(j)\phi_j^{\theta_k}(x_k,a_k) \tag{2.14}$$

where $\phi^{\theta_k}(x_k,a_k) = \left[\phi_0^{\theta_k}(x_k,a_k),...,\phi_{K-1}^{\theta_k}(x_k,a_k)\right]^T$ is the feature vector for state-action pair (x_k,a_k) which is dependent on parameter vector θ_k, $r_k = \left[r(1),...,r(K)\right]^T$ is the critic parameter vector at time t_k. The critic parametric vector can be updated as follows

$$r_{k+1} = r_k + \gamma_k d_k z_k \tag{2.15}$$

where γ_k is the stepsize, z_k is the eligibility trace and d_k is the temporal difference (TD) error.

In the above treatment, we have assumed that the problem state space is small enough that we can use a *look-up table representation*. A look-up table representation means that a separate action-value function $\tilde{Q}_r^{\theta}(x,a)$ is kept for every state-action pair (x,a). That is, the number of entries required for look-up table representation in (2.15) is $|r| = K = |X| \times |A|$. However, as the problem size increases and the number of state-action pair becomes large, look-up table representation becomes infeasible. Alternatively, we can use *compact representation* whereby action-value functions can be represented by a smaller set of parameters using a function approximator. That is by using function approximation in (2.14), the critic approximates $Q^{\theta}(x,a)$ with $\tilde{Q}_{r_k}^{\theta_k}(x_k,a_k)$ by using a linear function approximation architecture. The number of entries in (2.15) required for compact representation can be chosen such that $|r| = K << |X| \times |A|$. By using such representation, the number of parameters required for actor-critic method is greatly reduced and the scalability of the method is enhanced.

2.6 Conclusion

The SMDP framework has been used to formulate call admission control problems in CDMA networks. In this chapter, we have briefly reviewed the SMDP concept, and introduced an analytical tool called dynamic programming (DP) to solve the SMDP formulated problem. The advantage of DP is that the optimal solution can be determined. However, DP has two main drawbacks which are Bellman's curse of

dimensionality and the curse of modeling. These curses are caused by DP's requirement of complete mathematical formulization of the system dynamics.

To solve such drawbacks, reinforcement learning (RL) has been introduced. RL methods provide an approximate solution to SMDP formulated problems. RL circumvents the curses of dimensionality by employing function approximation. Such approach demands less computation and parameter storage. Furthermore, the curse of modeling is avoided by simulation or direct interaction with the system which do not require an explicit model of the system dynamics.

In the next chapter, an SMDP formulation of the CAC problem for multiple voice services in DS-CDMA systems is presented. The purpose is to demonstrate the optimality of the solution obtained by means of a dynamic programming method. The SMDP formulation is proposed and solved by Singh et al. (2002).

CHAPTER III

CALL ADMISSION CONTROL IN WIRELESS DS-CDMA SYSTEMS: A DP APPROACH

3.1 Introduction

In this chapter, a conventional dynamic programming (DP) method is employed to solve the CAC problem in wireless direct-sequential code division multiple access (DS-CDMA) systems which support multiple voice services subject to multiple quality-of-service (QoS) requirements. Existing approaches have been proposed to find for the CAC policy under QoS constraints in CDMA systems and have employed DP methods to solve for the optimal CAC policy.

In **section 1.1.3.2**, Yang and Geraniotis (1994) investigated the call blocking probability which is the network layer constraint alone. In this work, a dynamic programming method called value iteration is employed to solve for the CAC policy. Singh et al. (2002) investigated a call admission control problem which deals with dual constraints, namely, the SIR and blocking probability constraints. The solution from Singh et al. (2002) is solved by another DP method called linear programming (LP) which gave the optimal CAC policy under multiple QoS constraints. This chapter is therefore dedicated to present the underlying concepts, advantages and drawbacks of the method proposed by Singh et al. (2002).

The emphasis of this chapter is focused on the following issues:

1. The introduction of the network model and basic assumptions of CAC in CDMA systems.
2. The semi-Markov decision process (SMDP) formulation for the CAC problem in CDMA systems.
3. Performance quantification of DP compared to other CAC policies.

The structure of this chapter is organized as follows. The network model for DS-CDMA systems which support multiple voice services will be described in section 3.2. The following session is dedicated to describing the semi-Markov decision process (SMDP) formulation of the CAC problem. Section 3.4 presents the construction of optimal CAC policy by means of a DP method called LP. In section 3.5, the numerical results will be presented. Finally, section 3.6 summarizes the entire chapter.

3.2 Network Model

In this chapter, we study the work in Singh et al. (2002) for call admission control problem as a SMDP for the *uplink*, which is the mobile connection to base station (BS) of a synchronous DS-CDMA cellular system in multiple voice services only.

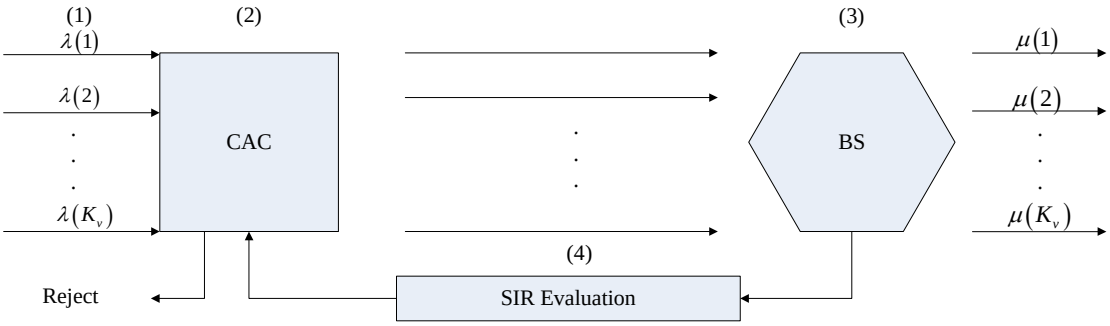

Figure 3.1 CAC diagram in CDMA network

Consider the incoming calls which require access to the CDMA network which can be classified into K_v classes. Assume that class i voice call is generated according to a homogenous Poisson distribution with intensity $\lambda(i)$ where $i = 1,...,K_v$. The call duration of class-i call is randomly generated according to an exponential distribution with mean rate $\frac{1}{\mu(i)}$. Figure 3.1 depicts the CAC model in CDMA systems which can be described as follows.

1. A voice user of class i where $i = 1,...,K_v$ requests to access the network.
2. The call admission controller receives the SIR level computed from the SIR evaluation. Based on the SIR level and the required SIR constraints, the admission controller will make a CAC decision for the base station (BS).
3. If the call admission controller decides to accept the call request, the request is sent to BS and connected to the public network i.e., public switched telephone network (PSTN).
4. The SIR evaluation is passed back to the call admission controller in (2) to decide whether to admit or reject the next incoming user request. Thus, the performance in the physical layer affects the admission of the new users.

3.3 Semi-Markov Decision Process Formulation

The aim of this section is to formulate the call admission control in CDMA network problem for the uplink process in figure 3.1 as a SMDP. The general assumptions of the SMDP framework in this chapter are as follows.

1. A discrete-valued the state space specifies the profile of the users, i.e., the number of users in the network. The SIR requirement constraints are incorporated into the framework by truncating the state space to include only the user profiles which satisfy the SIR constraints.

2. The state dynamics is defined by the arrival process which is a homogeneous Poisson distribution and the duration time for voice users which is assumed to be exponentially distributed.

3. The SMDP must be unichain which is the property that defines the average reward (or cost) performance criterion to be optimized.

Under these assumptions, the optimal CAC algorithm, which is designed to optimize a certain performance criterion and satisfy the QoS requirements in both the physical and network layers, can be attained by formulating the CAC problem as a SMDP as follows.

3.3.1 State Space

Consider a continuous-time stochastic process $\{X_t\}$ where X_t is a random variable representing the state of the system at time t where $t \in R_+$. Let X denote the state space which specifies all possible profiles of the number of users in the network. Suppose that at a given time $t \in R_+$, $X_t = x$ where $x \in X \subset I_+^{K_v}$ represents a state vector of the BS. Let $x(i)$ be the number of class i voice users

where $i = 1, 2, ..., K_v$. The state vector, x, which depicts the profile of current users in the system can then be defined by

$$x = \left[x(1), ..., x(K_v) \right]^T \tag{3.1}$$

To define the state space of the system, recall that the profile of the number of users in equation (3.1) in the system at any time is controlled by the level of SIR requirement in the CDMA system. In particular, the profile of the number of users in the system must be such that the actual SIR level for any class i must not fall below the minimum SIR requirement for class i where $i = 1, 2, ..., K_v$. Let $\Psi_x(i)$ be the SIR value for class i voice users when the system is in state x (See **Appendix I** for more details of the calculation of $\Psi_x(i)$). Let the minimum SIR requirement for class i users be denoted by $\beta(i)$. The minimum SIR requirement vector β can be defined as

$$\beta = \left[\beta(1), ..., \beta(K_v) \right]^T \tag{3.2}$$

Therefore, the state space of all the possible profiles of users in the CDMA system is given by

$$X = \left\{ x = \left[x(1), ..., x(K_v) \right]^T : \Psi_x(i) \geq \beta(i),\ i = 1, ..., K_v \right\} \tag{3.3}$$

Note that the state space X in (3.3) inherently embeds the SIR constraints.

3.3.2 Decision Epochs

Consider a CDMA system which has a user requesting to join the network. At this particular instant, the BS must decide whether to admit or reject that

call request. Such instant is generally referred to as a *decision epoch*. The definition of decision epoch, according to Ross and Tsang (1989), is the instance when the stochastic process $\{X_t\}$, $t \in R_+$ changes state. The transition time between the changes of state in the stochastic process $\{X_t\}$ is usually defined as t_k where $k = 0,1,2,...$ is the index of the state transition sequence.

3.3.3 Actions

For each decision epoch, the decision from the call admission controller at the BS is made whether it is to admit or reject the call arrival. The decisions are referred to *actions*. The action space, denoted by A, is the set of all possible actions which is defined as follows

$$A = \left\{ a = \left[a(1), ..., a(K_v) \right]^T : a \in \{0,1\}^{K_v} \right\} \tag{3.4}$$

where $a(i)$ refers to the action for class i voice user such that

$$a(i) = \begin{cases} 1 & \text{, if class } i \text{ voice user is admitted} \\ 0 & \text{, otherwise} \end{cases} \tag{3.5}$$

In other words, action $a(i) = 1$ is the action that accepts the request of class i voice user; otherwise, the user is blocked by the BS.

Suppose at decision epoch t_k, the system is in state $X_{t_k} = x$, where $k = 0,1,2,...$. At state x, an action must be selected from a state-dependent subset of A denoted by $A(x)$. More specifically, $A(x)$ is the set of all possible actions at state x which can be defined as

$$A(x) = \{a \in A : a(i) = 0, \quad \text{if } x + e(i) \notin X \} \tag{3.6}$$

where $e(i)$ denotes a vector with all zero components except for the i-th component which is unity. Equation (3.6) ensures that any action a taken in state x, does not result in a transition to a state outside the state space.

3.3.4 State Dynamics

In this chapter, the CAC problem is formulated as a SMDP. As stated in **section 3.1**, the purpose of this chapter is to employ a dynamic programming method which has been proposed by Singh et al. (2002) to solve the CAC problem. However, as explained in **section 2.3**, the DP method requires a complete knowledge of the system dynamics which is described by the transition probabilities and expected reward of the system. Denote the transition probabilities for each state action pairs by $p_{xy}(a)$ which is defined as the probability that the state at the next decision epoch is y, given that an action a is selected at current state x. That is, the transition probability can be defined as

$$p_{xy}(a) \triangleq P\left(x_{t_{k+1}} = y \middle| x_{t_k} = x, a_{t_k} = a\right) \tag{3.7}$$

where $x_{t_{k+1}}$, x_{t_k}, a_{t_k} refer to the states at time t_{k+1} and t_k, and the action taken at time t_k, respectively. Let $\tau_x(a)$ be the expected sojourn time of the system. The expected sojourn time is the expected time at which the system remains in the state x when action a is taken. The expected sojourn time influences a change of state in terms of a specified period of time. Its mathematical term is given by

$$\tau_x(a) \triangleq E\left\{t_{k+1} - t_k \,\middle|\, x_{t_k} = x, a_{t_k} = a\right\} \tag{3.8}$$

For a call admission control problem formulated as a SMDP, equations (3.7) and (3.8) can be expressed as follows (Bertsekas, 1995 and Singh et al., 2002)

$$p_{xy}(a) = \begin{cases} \lambda(i)a(i)\tau_x(a) & \text{,if } y = x + e(i) \in X \\ \mu(i)x(i)\tau_x(a) & \text{,if } y = x - e(i) \in X \\ 0 & \text{,otherwise,} \end{cases} \tag{3.9}$$

and

$$\tau_x(a) = \left[\sum_{i=1}^{K_v}\left(a(i)\lambda(i) + x(i)\mu(i)\right)\right]^{-1}, \tag{3.10}$$

where $e(i)$ denotes the vector with all elements equal to zero except the i-th component which is unity, $\lambda(i)$ and $\mu(i)$ are the arrival rate and call duration time class i voice users which follow a homogeneous Poisson distribution and an exponential distribution, respectively, where $i = 1,...,K_v$ is the class index of voice services.

Note that $\tau_x(a)$ in equation (3.10) is the reciprocal of the sum of the rates which exit state x when action a is taken. In other words, $\tau_x(a)$ is the expected time the system remains in state x when action a is taken, i.e., the expected time until the system transits into a new state. In equation (3.9), $p_{xy}(a)$ given by $\lambda(i)a(i)\tau_x(a)$ is the probability of transiting from state x to a new state $y = x + e(i) \in X$. In other words, this is the probability of transiting into a state which has one more class i

voice user than state x. Note that this probability is non-zero only when a class i call is admitted (i.e., $a(i)=1$). On the other hand, equation (3.9) states that the probability of transiting from state x to a new state $y=x-e(i)\in X$ is given by $p_{xy}(a)=\mu(i)x(i)\tau_x(a)$. In other words, this is the probability of transiting into a state which has one less class i voice user than state x. This refers to the scenario when a class i voice user terminates a call and leaves the system. Note that this probability is non-zero so long as there exists class i voice users in the system (i.e., $x(i)>0$).

3.3.5 Policy

A call admission control policy is a rule that maps each state x into an action a. Suppose that at any given state x, an action a which decides whether to accept or reject the call, is selected according to a specified policy π. A stationary policy (See **section 2.2.3**) π is a function that maps the state space X into the action space A and is independent of time. An admissible policy is a stationary policy that satisfies

$$\Pi=\left\{\pi: X\to A \,\middle|\, \pi(x)\in A(x), \forall x\in X\right\} \tag{3.11}$$

where $\pi(x)$ refers to the action taken at state x under policy π. Hence, we are focused on finding a policy that is optimal over all other admissible policies subject to some performance criterion.

3.3.6 Performance Criterion

The aim of this chapter is to solve the CAC problem in a CDMA system by *minimizing* the average *cost* performance criterion and satisfy multiple QoS requirement constraints. The cost criterion $c(x,a)$ in this chapter, which is proposed by Singh et al. (2002), is defined as follows

$$c(x,a)=\sum_{i=1}^{K_v}\nu(i)\left(1-a(i)\right) \tag{3.12}$$

where $\nu(i)\in R_+$ where $\nu(i)$, $i=1,2,...,K_v$ is some weight factor. As stated earlier that the CAC problem in CDMA systems must deal with two constraints. The first constraint is the SIR level constraint which determines the capacity of the system and is embedded into the state space as shown in equation (3.3). The second constraint is the blocking probability constraint which can be embedded in the cost function in terms of $\nu(i)\in R_+$ where $\nu(i)$, $i=1,2,...,K_v$ is the weight factor for the blocking probability constraints. In this chapter, the average cost criterion is used as the performance criterion to be optimized. Suppose the average cost for a given admissible policy $\pi\in\Pi$ is denoted by $v(\pi)$ is defined as

$$v(\pi)=\lim_{N\to\infty}\frac{1}{t_N}E_\pi\left\{\sum_{k=0}^{N-1}c(x_k,a_k)\right\} \tag{3.13}$$

where $E_\pi\{.\}$ is the expectation under policy π. The optimal policy π^* can then be defined as

$$v(\pi^*) = \min_{\forall \pi \in \Pi} v(\pi) \tag{3.14}$$

3.4 Constructing the Optimal CAC Policy with Constraints

This section explains the construction of the optimal policy. Here, the proposed method by Singh et al. (2002) is investigated, where the dynamic programming technique called linear programming (LP) method is employed to solve for the optimal CAC for the average cost criterion, in (3.14). The optimal policy π^* is obtained by solving the following LP

$$\min_{z_{xa} \geq 0, x \in X, a \in A(x)} \sum_{x \in X} \sum_{a \in A(x)} c(x,a) \tau_x(a) z_{xa},$$

$$\text{Subject to } \sum_{a \in A(y)} z_{ya} - \sum_{x \in X} \sum_{a \in A(x)} p_{xy}(a) z_{xa} = 0, \; y \in X,$$

$$\sum_{x \in X} \sum_{a \in A(x)} \tau_x(a) z_{xa} = 1 \tag{3.15}$$

and the blocking probability constraints

$$\sum_{x \in X} \sum_{a \in A(x)} c^i(x,a) \tau_x(a) z_{xa} \;\leq\; B(i), \qquad i = 1, ..., K_v \tag{3.16}$$

where $c^i(x,a) = 1 - a(i)$, $B(i)$ is the maximum allowable blocking probability for class i voice user. The optimal solution of the above LP problem is denoted by z^*_{xa}. To obtain a CAC policy from z^*_{xa}, the optimal CAC policy can be constructed as a *randomized stationary policy* as follows. Denote the optimal randomized stationary policy as $\pi^*(x,a)$ where

$$\pi^*(x,a) = \frac{\tau_x(a) z^*_{xa}}{\sum_{\forall a \in A(x)} \tau_x(a) z^*_{xa}} \tag{3.17}$$

However, if the optimal solution for some state $x \in X$ is such that $z^*_{xa} = 0$, $\forall a \in A(x)$, choose an arbitrary $a \in A(x)$ and set $\pi^*(x) = a$.

3.5 Numerical Study

In this section, a numerical study is carried out to demonstrate the optimality of CAC policies obtained in the previous section which was proposed in Singh et al. (2002). For performance comparison, three approaches have been investigated which include the complete sharing approach (CS), the threshold policy and the SMDP approach solved by DP. We consider a system with two classes of voice services, i.e., $K_v = 2$. Each arrival of class i users is generated according to a homogenous Poisson distribution with rate $\lambda(i)$, where $i = 1,2$. The mean call holding time for class i call is exponentially distributed with parameter $\mu(i)$ where $i = 1,2$. The cost function $c(x,a)$ used in (3.12) and (3.13) is modified into a reward function in order to consider an average reward performance criterion (as oppose to the average cost performance criterion). The physical interpretation of the reward function ($r(i)$ in tables 3.1-3.3) can be interpreted as the income of the system earned by admitting class i voice users. Thus, the optimization objective is to maximize the long-term average reward instead. Performance metrics are measured in terms of the long-term average reward, the blocking probability and the SIR levels. We assume that there is

no fading in the channel. The parameter $\beta(i)$ denotes the minimum SIR requirement for class i which determines the upper limit on the number of users in the network as shown in equations (3.2) and (3.3) which is 20, 21 dB, respectively. The parameter $P(i)$ denotes the transmission power for class i user which is 1.2 and 1.7 watts, respectively. The channel gain for class i is $\bar{h}(i)=1$, where $i=1,2$. Finally, the channel variance of class i user is given by $\xi^2(i)=0$, where $i=1,2$. N is the processing gain of the channel which is 10. Note that $\bar{h}(i)$, $P(i)$ and $\xi^2(i)$ are used to calculate, $\Psi_x(i)$, which is the actual SIR value of class i users when the system is in state x (See **Appendix I)**. The value of $\Psi_x(i)$ is then employed to determine the state space X in equation (3.3). Simulation is run for 10^7 time steps in each CAC method.

We study 6 cases of parameters settings as shown in Tables 3.1-3.3. Table 3.1 tests the ability to satisfy the blocking probability constraints of class 2 users when the constraint on blocking probability requirement is reduced. Table 3.2 shows the ability to maintain the blocking probability constraints of class 1 users while the traffic arrival rates of each class are increased from 1.5 to 2.0 calls/min. Table 3.3 shows the ability to maintain the constraint on blocking probability requirement in a non-trivial scenario.

The parameter $T(i)$ in tables 3.1-3.3 is the threshold for class i voice user under the threshold CAC policy. The threshold policy shown in the tables are obtained by empirically reducing the number of calls in each class, thereby decreasing the blocking probability for the desired class until its blocking probability constraint is

satisfied. We select the threshold levels that give the maximum long-term average reward in order to obtain the best possible threshold policy that satisfies the desired blocking probability constraints.

Table 3.1 Multiservice Parameters: cases 1-2

	Class1	**Class2**
Case 1, Blocking probability constraints, $B(i)$	-	0.0001
Case 2, Blocking probability constraints, $B(i)$	-	0.00005
Mean arrival rate (call/min), $\lambda(i)$	1.5	1.5
Threshold limited, $T(i)$	2	8
Rewards ($), $r(i)$	8	4
Mean call holding time(min/call), $1/\mu(i)$	1.1	1.1

Table 3.2 Multiservice Parameters: cases 3-4

	Class1	**Class2**
Case 3, Blocking probability constraints, $B(i)$	0.0001	-
Case 4, Blocking probability constraints, $B(i)$	0.00005	-
Mean arrival rate (call/min), $\lambda(i)$	2.0	2.0
Threshold limited, $T(i)$	9	1
Rewards ($), $r(i)$	8	4
Mean call holding time(min/call), $1/\mu(i)$	1.1	1.1

Table 3.3 Multiservice Parameters: cases 5-6

	Class1	**Class2**
Case 5, Blocking probability constraints, $B(i)$	-	0.00005
Case 6, Blocking probability constraints, $B(i)$	-	0.00001
Mean arrival rate (call/min), $\lambda(i)$	4	1
Threshold limited, $T(i)$	3	7
Rewards ($), $r(i)$	2	8
Mean call holding time(min/call), $1/\mu(i)$	2	1

Tables 3.4-3.9 reveal the optimality of the DP method over all other policies. The DP method can achieve the optimal long-term average reward while maintaining constraints on the call blocking probability requirement. The CS policy can also give the high long-term average reward. However, CS *cannot* satisfy the blocking probability constraints required. The empirical threshold policy can maintain the constraint on the call blocking probability requirement in all cases. However, the long-term average reward is the lowest of all.

As stated earlier, the objective of this chapter is to deal with the CAC problem with dual constraints, namely, the call blocking probability and SIR level. Results in tables 3.5-3.10 depict the ability of the Singh et al (2002)'s method to maintain the call blocking probability constraints.

Table 3.4 Blocking probability measured for cases 1-2

	DP		CS		Threshold Policy	
	Class1	**Class2**	**Class1**	**Class2**	**Class1**	**Class2**
Unconstraint	0.000343	0.000445	0.000412	0.000380	0.000412	0.000380
Case1	0.006027	0.000108	0.000380	0.000440	0.281156	0.000074
Case2	0.025400	0.000060	0.000408	0.000399	0.280863	0.000093

Table 3.5 Average reward measured for cases 1-2

	DP	CS	Threshold Policy
	Avg. Reward	**Avg. Reward**	**Avg. Reward**
Unconstraint	17.9625	17.9924	17.9924
Case1	17.9167	17.9761	14.6206
Case2	17.6844	17.9894	14.6288

Table 3.6 Blocking probability measured for cases 3-4

	DP		CS		Threshold Policy	
	Class1	**Class2**	**Class1**	**Class2**	**Class1**	**Class2**
Unconstraint	0.002715	0.042713	0.002797	0.002942	0.002797	0.002942
Case3	0.000136	0.224125	0.002893	0.003167	0.000124	0.645501
Case4	0.000054	0.413666	0.003232	0.003189	0.000060	0.645284

Table 3.7 Average reward measured for cases 3-4

	DP	CS	Threshold Policy
	Avg. Reward	**Avg. Reward**	**Avg. Reward**
Unconstraint	23.6002	23.9509	23.9509
Case3	22.18	23.9858	18.7727
Case4	20.7061	23.9858	18.8338

Table 3.8 Blocking probability measured for cases 5-6

	DP		CS		Threshold Policy	
	Class1	**Class2**	**Class1**	**Class2**	**Class1**	**Class2**
Unconstraint	0.000877	0.000110	0.000867	0.000868	0.000802	0.000719
Case5	0.002256	0.000050	0.000828	0.000836	0.210803	0.000036
Case6	0.003930	0.000010	0.000802	0.000719	0.400232	0.000008

Table 3.9 Average reward measured for cases 5-6

	DP	**CS**	**Threshold Policy**
	Avg. Reward	**Avg. Reward**	**Avg. Reward**
Unconstraint	15.9972	15.9854	15.9454
Case5	15.9703	15.9931	14.3628
Case6	15.9596	15.9454	12.8427

In terms of the constraints on the SIR level, figures 3.2 and 3.3 show the moving average of the SIR level under case 4 setting. Results show that the SIR level of both classes can indeed be satisfied.

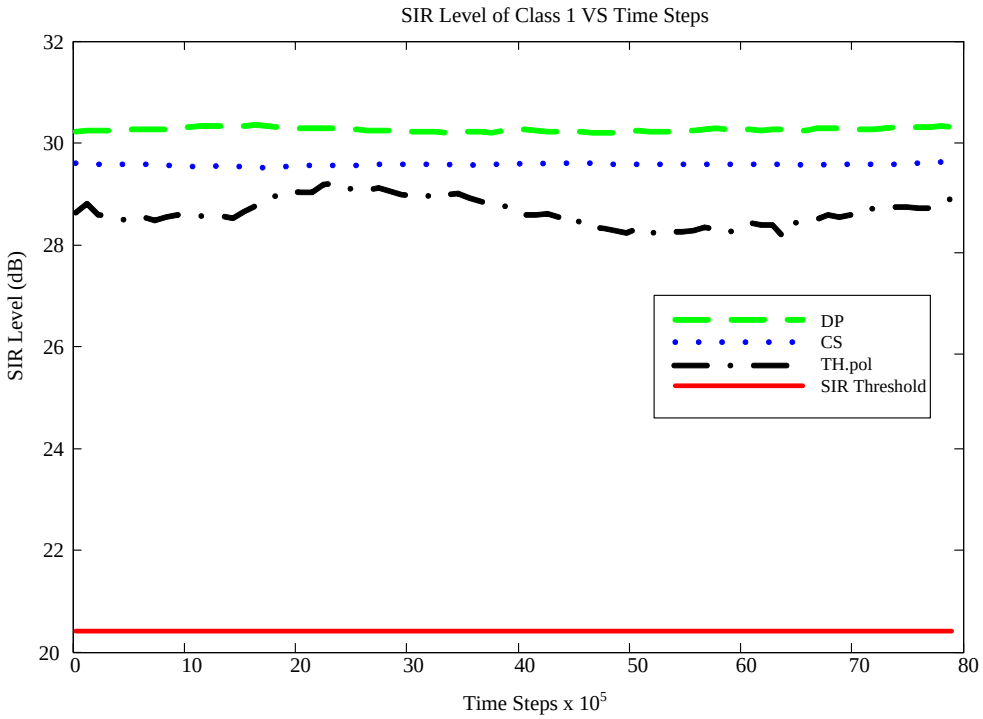

Figure 3.2 The SIR level for class 1 voice user

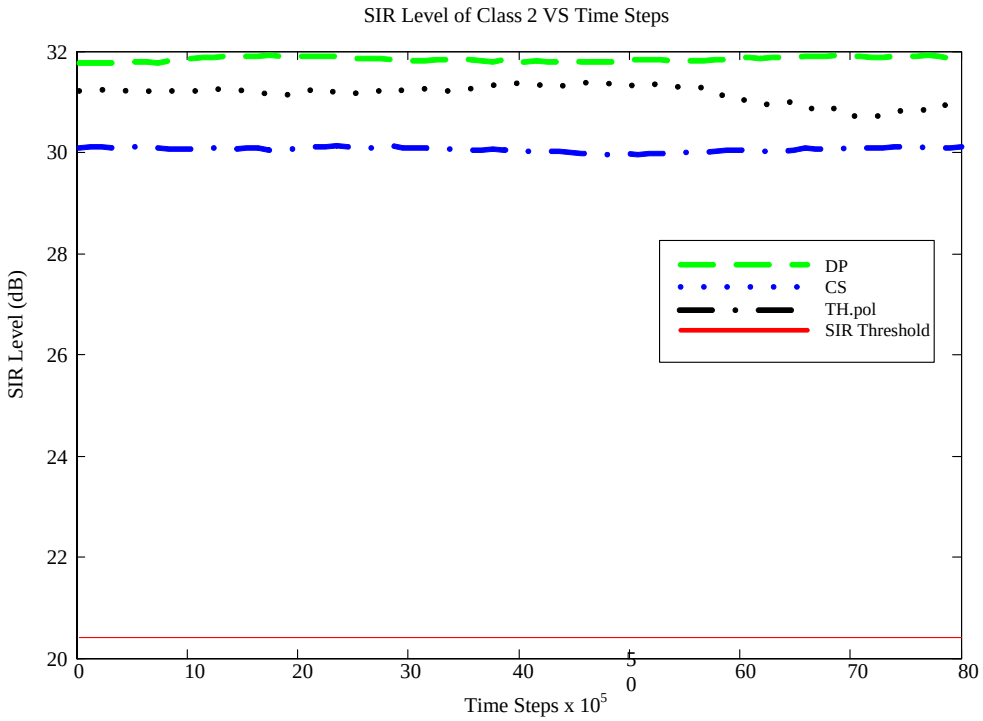

Figure 3.3 The SIR level for class 2 voice user

3.6 Conclusion

In this chapter, we studied the CAC problem for multiclass voice services in CDMA networks subject to two types of QoS constraints, namely, the SIR level constraint which is a physical layer constraint, and the call blocking probability constraint which is a network layer constraint. The focus of this chapter is on the approach of Singh et al. (2002) which formulates the CAC problem as a semi-Markov decision process (SMDP) by embedding the SIR level constraints into the state space. The call blocking probability constraint is taken into account by virtue of a dynamic programming (DP) method called linear programming (LP). This particular DP method was selected by Singh et al. (2002) due to its ability to solve optimization problems subject to multiple constraints.

The optimality of the DP-based CAC policy is demonstrated in the numerical study. The performance metrics measured are the long-term average reward, the call blocking probability and the SIR requirements. The numerical results show that the policy obtained from the DP method outperforms the complete sharing and empirical threshold policies by attaining the highest long-term average reward while still satisfying the dual constraints on the SIR level and call blocking probability.

In the next chapter, we extend the CAC problem to a more realistic scenario by increasing the state space which results in increased computational and storage complexity of the DP approach. To circumvent computational burden of DP, we propose to use an alternative method called actor-critic reinforcement learning (RL) to solve for a near-optimal CAC policy in the CDMA network instead.

CHAPTER IV

CALL ADMISSON CONTROL IN WIRELESS DS-CDMA SYSTEMS: A RL APPROACH

4.1 Introduction

In the previous chapter, we obtained an optimal call admission control (CAC) policy in wireless DS-CDMA systems with multiple voice services constructed from dynamic programming (DP). Although an optimal solution is obtained, the scalability of such method becomes a major concern as the system size is increased to a more realistic scale. This problematic issue is referred to as the curse of dimensionality. Furthermore, DP methods must also cope with the curse of modeling which requires a complete knowledge of state transition probabilities and expected reward which are difficult to determine exactly in many scenarios.

To overcome the computational difficulty of DP, we propose an alternative approach based on a reinforcement learning (RL) technique (Sutton and Barto, 1998) to determine near-optimal CAC policies instead of the optimal solution as obtained by DP. The RL method can provide near-optimal solutions to complex DP problems through experience learned from simulations or direct interaction with the environment. Consequently, RL does not require knowledge of the explicit model of system dynamics. Furthermore, RL methods also permit the use of function approximation which allows approximation of decision variables by a small set of parameters. Therefore, the scalability of RL is greater than classical DP methods.

The aim of this chapter is to propose an actor-critic RL method to solve the CAC problem in a CDMA network with multiple voice services, which incorporates both SIR levels and blocking probability constraints. The method builds on an earlier version of an actor-critic RL method in Usaha and Barria (2007) which has been successfully applied to solve a CAC and routing problem in LEO satellites networks. Whereas Usaha and Barria (2007) considered an *unconstrained* SMDP problem, the method proposed in this chapter considers a *constrained* SMDP problem. It should be noted that Vazquez-Abad and Krishnamurthy (2002) proposed a RL method to solve the CAC problem in CDMA networks under the SIR and blocking probability constraints. However, their method is based on a look-up table representation which will be infeasible as the problem dimension becomes large. On the contrary, our proposed method employs function approximation which has the advantage of less memory storage and computational requirements than Vazquez-Abad and Krishnamurthy (2002).

The emphasis of this chapter is on the following issues:

1. The semi-Markov decision process (SMDP) formulation which differs from the formulation in **chapter 3**.

2. A modified reward to deal with the blocking probability constraints which differs from that of Usaha and Barria (2007).

3. The performance analysis of the proposed RL method compared to the DP and empirical threshold policy methods.

4. The analysis of memory storage and computational requirements.

The structure of this chapter is organized as follows. Section 4.2 describes the CDMA network model. Section 4.3 presents the SMDP formulation for the RL

framework. Section 4.4 describes the RL technique called the actor-critic RL method which combines the strong points of the actor-only and critic-only methods together. Section 4.5 presents the numerical study of RL approach and section 4.6 concludes this chapter.

4.2 DS-CDMA Network Model

The network model remains the same as in **section 3.2**. For convenience, the network model for the CDMA system is summarized as follows.

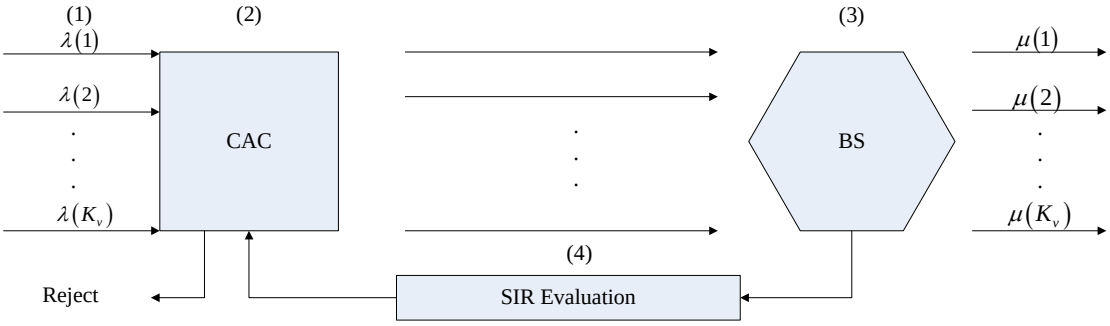

Figure 4.1 Network model for DS-CDMA systems

Consider a CDMA system which supports multiclass voice users of K_v classes in Figure 4.1. In phase (1) in the figure, we denote the arrival rate for voice class i users by $\lambda(i)$ where $i = 1, 2, ..., K_v$. The arrival of incoming call requests of class i users follows a homogeneous Poisson distribution. In phase (2) of Figure 4.1, the call admission controller (CAC) receives the SIR level from the SIR evaluation in (4). In phase (3) of the figure, the call admission controller will make a CAC decision for the base station (BS) based on the SIR level and the required SIR constraints. If the decision is to admit the call request, the BS connects the call request to the

network. Finally in phase (4) of the figure, the current SIR level must be reevaluated as the current profile of the users has changed. The SIR value is then fed back to the call admission controller for the next CAC decision. Once admitted to the network, the duration of a class i call in the system is exponentially distributed with mean rate $\frac{1}{\mu(i)}$.

4.3 SMDP Formulation

The CAC problem in CDMA networks can be formulated as an SMDP in a similar manner as **chapter 3**. However, unlike **chapter 3**, we modify the reward signal by including a penalty function which allows control over the blocking probability so that it meets the blocking probability requirements. For convenience, a complete formulation is provided in this section.

4.3.1 State Space

Consider a continuous-time stochastic process $\{X_t\}$ where X_t is a random variable representing the number of users in the system at time $t \in R_+$. Let X denote the state space which represents the number of users in the system. Let $x \in X \subset I_+^{K_v}$ be the state vector of the system and K_v be the number of voice classes where

$$x = \left[x(1), ..., x(K_v) \right]^T \tag{4.1}$$

where $x(i)$, $i = 1,...,K_v$ denotes the number of class i voice users. Suppose that the current state is given by $X_t = x$. In CDMA systems, the maximum number of allowable users depends on the SIR levels of the users present in the system. Such SIR levels must not violate the minimum SIR level requirements.

Let $\Psi_x(i)$ denote the SIR value for all class i calls when the system is in state x. Let vector β represent the minimum SIR requirements for all classes specified as follows

$$\beta = \left[\beta(1),...,\beta(K_v)\right]^T \tag{4.2}$$

Since equation (4.2) must be satisfied for each admission of voice user into the CDMA system, the number of users in the system must give a SIR value $\Psi_x(i)$ which strictly satisfies the minimum SIR requirement $\beta(i)$ for every class i. Hence, the state space X truncated by the required SIR levels can then be defined as follows

$$X = \left\{x = \left[x(1),...,x(K_v)\right]^T : \Psi_x(i) \geq \beta(i),\ i = 1,2,...,K_v\right\} \tag{4.3}$$

Note that the minimum SIR requirements in vector β determine the size of the state space. In other words, vector β determines the maximum allowable number of users in the system or the capacity of the system. Details of the calculation of the SIR level $\Psi_x(i)$ is given in **appendix I.**

4.3.2 Decision Epochs

A decision epoch refers to the time instant where an event occurs such that a decision must be made. Let the event space Ω be the finite set of all possible events defined by

$$\Omega = \left\{ \omega = \left[\omega(1), ..., \omega(K_v) \right]^T : \omega(i) \in \{1, 0, -1\} \right\} \tag{4.4}$$

The event vector ω is defined as

$$\omega = \left[\omega(1), ..., \omega(K_v) \right]^T \tag{4.5}$$

The event $\omega \in \Omega$ indicates whether a call arrival or call departure occurs. In particular, $\omega(i)$ refers to an event associated to class i calls and has the following meaning,

$$\omega(i) = \begin{cases} 1 & \text{, if a call arrival of class } i \text{ occurs} \\ -1 & \text{, if a call depature of class } i \text{ occurs} \\ 0 & \text{, otherwise} \end{cases} \tag{4.6}$$

Furthermore, at any decision epoch, we assume that only one event can incur, i.e.,

$$\sum_{i=1}^{K_v} \left| \omega(i) \right| = 1 \tag{4.7}$$

4.3.3 Action Sets

Let t_k where $k \in I_+$ be the k-th decision epoch at which the event $\omega_k \in \Omega$ occurs. Let x_k be the state of the system in time interval $[t_{k-1}, t_k)$. Suppose

that $x_k = x$ and $\omega_k = \omega$. If the event refers to an arrival of a new call request of class i, the call admission controller in the BS must decide whether to admit or reject the call request. Such action depends on the current state of the system and the type of event incurred. Let A be the set of all possible actions of the call admission controller. The action $a(x,\omega) \in A$ refers to the action taken at state x and event ω which is given by

$$a(x,\omega) = \begin{cases} 1 & \text{, accept the call} \\ 0 & \text{, reject the call} \end{cases} \tag{4.8}$$

Note that all call termination events must be allowed, therefore no action is required.

4.3.4 Immediate Reward

Suppose that the state and event at the k-th decision epoch be x_k and event ω_k, respectively. Suppose that action $a_k = a(x_k, \omega_k)$ is taken and the system then transits into the next state x_{k+1}. Depending on how good the selected action was at a given state, an immediate reward is generated. The immediate reward function $g(x_k, \omega_k, a_k)$ is given by

$$g(x_k, \omega_k, a_k) = \begin{cases} r(i) & \text{, if } x_k \in X \text{ and } \omega_k \text{ is such that } \omega(i)=1 \\ 0 & \text{, otherwise} \end{cases} \tag{4.9}$$

where $r(i) \in R$ can be interpreted as the revenue earned by the service provider for each decision to admit a class i user into the system.

4.3.5 Policy

The goal of the SMDP formulation is to find a state dependent rule or a stationary policy that maximizes the long-term average reward. Let Π be the set of stationary policies defined as

$$\Pi = \{\pi : X \times \Omega \to A\} \tag{4.10}$$

To solve for the optimal stationary CAC policy, the performance criterion of the SMDP formulation must be defined.

4.3.6 Performance criterion

The aim of this chapter is to find the CAC policy that maximizes the long-term average reward for CDMA systems. The long-term average reward for policy π can be defined as

$$v(\pi) = \lim_{N\to\infty} \frac{1}{t_N} E_\pi \left\{ \sum_{k=0}^{N-1} g(x_k, \omega_k, a_k) \right\} \tag{4.11}$$

where t_N is the completion time of the N-th decision epoch, and the actions a_k are selected according to policy π, i.e., $a_k = \pi(x_k, \omega_k)$. Let the optimal policy be denoted by π^* where the associating long-term average reward under such policy is denoted by $v(\pi^*)$. The objective is to find an optimal policy π^* such that

$$v(\pi^*) = \max_{\forall \pi \in \Pi} v(\pi) \tag{4.12}$$

4.3.7 Modified Reward for Blocking Probability Constraints

In this section, the blocking probability constraints are accounted for by modulating the immediate reward. To incorporate the constraints on blocking probability into the SMDP framework, the immediate reward function is changed from equation (4.9) by including a penalty term to the original reward. The penalty term is a function of the current blocking probability multiplied by a Lagrange constant and the time elapsed since the last decision epoch which is given by

$$g_{\chi}\left(x_k, \omega_k, a_k\right) = g\left(x_k, \omega_k, a_k\right) - \chi B(i)\tau \tag{4.13}$$

where $B(i)$ is the measured blocking probability of class i users over the past history up to the instant when action a_k is made in state x_k, τ is the mean sojourn time since the last decision epoch and χ is the Lagrange multiplier (Tong and Brown, 1999). The Lagrange multiplier can be physically interpreted as the penalty incurred whenever a blocking probability constraint is violated. The modified reward function in equation (4.13) is then fed back to the call admission controller to evaluate how good the selected action was.

In addition, to maximize the long-term average reward, the performance criterion must be changed by replacing the reward signal in equation (4.11) by equation (4.13). The new performance criterion is thus given by

$$v(\pi) = \lim_{N\to\infty} \frac{1}{t_N} \sum_{k=0}^{N-1} g_{\chi}\left(x_k, \omega_k, a_k\right) \tag{4.14}$$

where t_N is the completion time of the N-th decision epoch, and the actions a_k are selected according to policy π. Note that policy associated to the long-term average reward must also satisfy the blocking probability constraints. The blocking probability constraint for class i users can be written as follows

$$\lim_{N\to\infty} \frac{1}{t_N} \sum_{k=0}^{N-1} B_{k+1}(i)\tau_{k+1} \le B(i) \tag{4.15}$$

The purpose of equation (4.15) is to guarantee the long-term blocking probability requirement of class i users in the system.

4.4 Actor-Critic Reinforcement Learning

Actor-critic reinforcement learning combines the two strong characteristics of the actor-only and critic-only RL methods. The critic part of the actor-critic algorithm estimates the value functions based on some approximation architecture and simulation. The estimated value function is used to update the parameterized policy in the actor part in the direction which improves the performance gradient. Moreover, the estimation of value functions in the actor-critic algorithm may help reduce the variance which might deliver faster convergence speed when compared to actor-only methods (Usaha, 2004). Hence, faster convergence may be achieved (strong feature of critic-only methods) and policy improvement is guaranteed as long as the policy is gradient-based (strong feature of actor-only methods).

A gradient-based policy can be described as follows. Consider a SMDP with state space X and action space A. Let μ_θ be a randomized stationary policy parameterized by some vector θ where $\theta \in R^M$ and M is the number of tunable

parameters. The actor-critic algorithm selects an action according to some probabilistic distribution over the set of allowable actions parameterized by vector $\theta \in R^M$. In particular, a randomized stationary policy $\mu_\theta(a|x,\omega)$ maps a probability distribution over the action space A to each state and event pair $x \in X$ and $\omega \in \Omega$. The policy $\mu_\theta(a|x,\omega)$ is controlled by the parameter vector θ. We are interested in finding a policy $\mu_\theta(a|x,\omega)$ such that $v(\mu_{\theta^*}) = \max_{\forall \theta \in R^M} v(\mu_\theta)$ for every other policy μ_θ. This can be achieved by estimating the gradient of the average reward from simulation and updating θ in the direction that improves the gradient direction.

In the proposed actor-critic RL algorithm, the actor selects an action according to the following randomized stationary policy architecture

$$\mu_\theta(a|x,\omega) = \begin{cases} p_\theta(x,a,\omega) & ,\text{if } \omega(i) = 1 \\ 1 & ,\text{if } \omega(i) = -1 \end{cases} \tag{4.16}$$

where $p_\theta(x,a,\omega)$ is the probability of selecting action a at state x and event ω associated to parameter vector θ. The probability distribution of $p_\theta(x,a,\omega)$ can be written in the form as follows

$$p_\theta(x,\omega,a) = \frac{\exp(s_\theta(a))}{\sum_{\forall u \in A} \exp(s_\theta(u))} \tag{4.17}$$

where the function $s_\theta(a)$ is defined by

$$s_\theta(a) = S(x,\omega,a) + \theta(\omega,a) \tag{4.18}$$

Note that the probability distribution function in equation (4.17) and (4.18) is used here as it defines the probability of selecting an action which is a continuously differentiable with respect to the parametric vector θ. This is a necessary condition for the existence of the gradient of the performance criterion (Usaha, 2004) of which we wish to improve. The scalar function $S(x,\omega,a)$ is the state representation which should be chosen in such a way that characterizes the features of the event ω, state x and action a. In this work, following scalar function is used

$$S(x,\omega,a)=\sum_{i=1}^{K_v} x(i) \tag{4.19}$$

It should be noted that equation (4.19), captures the characteristic of current users in the CDMA system. The state representation function in (4.19) is selected from Table I in Usaha and Barria (2007) because it gave the best results for CAC in their work.

The actor-critic algorithm consists of the critic and actor feature of the following forms

Actor Feature

$$\psi_i^{\theta}(x,\omega,a)=\frac{\frac{\partial}{\partial\theta(i)}\mu_{\theta}(a|x,\omega)}{\mu_{\theta}(a|x,\omega)}, \quad i=0,...,M-1 \tag{4.20}$$

Critic Feature

$$\phi_i^{\theta}(x,\omega,a)=\begin{cases}\psi_i^{\theta}(x,\omega,a) & ,\text{for } i=0,...,M-1\\ \phi_i^{\theta}(x,\omega,a) & ,\text{for } i=M\end{cases} \tag{4.21}$$

Note that the actor-critic feature in equation (4.20) and (4.21) are *updated only at decision epochs associated to the incoming calls* of the system. The critic feature in (4.21) can be expressed to be mathematical function as follows

$$\phi_i^\theta(x,\omega,a) = x(i),\ \forall\omega(i) = 1 \tag{4.22}$$

4.4.1 Actor-Critic Algorithm

The proposed method is an extension of Usaha and Barria (2007) which developed an actor-critic method and applied it to solve an unconstrained CAC (and routing) problem. More specifically, the actor-critic method proposed here deals with a constrained CAC problem by employing the immediate reward function in equation (4.13) to account for the constraints.

The actor-critic algorithm is used to train two parameter vectors, namely, the critic parameter vector $r \in R^{M+1}$, and the actor parameter vector $\theta \in R^M$. In this algorithm, let γ_k and η_k be small stepsize parameters, $\tilde{Q}_r^\theta(x,\omega,a) = r^T\phi^\theta(x,\omega,a)$ and $\nabla_r\tilde{Q}_r^\theta(x,\omega,a) = \phi^\theta(x,\omega,a)$. $\Gamma(r_k)$ is a normalizing scalar parameter that controls the learning rate of the actor. $\Gamma(r_k)$ must satisfy $\frac{R_1}{1+|r_k|} \le \Gamma(r_k) \le \frac{R_2}{1+|r_k|}$ for some constant $0 < R_1 < R_2$, where $R_1, R_2 \in R_+$. The update of d_k, z_k, r_k, $\tilde{v}_k$ and θ_k are perform at every instant an event occurs (i.e., at user arrival and departure instants). The proposed algorithm called the actor-critic for SMDP algorithm (ACSMDP) is presented as follows. The convergence results of the algorithm is provided in Usaha and Barria (2007).

The ACSMDP algorithm

1) Initialize r_0, $z_0 \in R^{M+1}$, $\theta_0 \in R^M$, $x_0 \in X$ and $\tilde{v}_0$ arbitrarily.

2) **for** $k = 1$ **to** N **do**

3) At t_k, an event ω_k is generated at state x_k.

4) $\tau_k = t_k - t_{k-1}$

5) Generate $a_k \in A$ from $\mu_{\theta_{k-1}}\left(a \middle| x_k, \omega_k\right)$.

6) Get reward $g_\chi\left(x_k, \omega_k, a_k\right)$ and system transit state to next state x_{k+1}.

7) Perform updates.

a) Temporal difference:

$$d_k = g\left(x_{k-1}, \omega_{k-1}, a_{k-1}\right) - \tilde{v}_{k-1}\tau_k + \tilde{Q}_{r_{k-1}}^{\theta_{k-1}}\left(x_{k-1}, \omega_{k-1}, a_{k-1}\right) - \tilde{Q}_{r_{k-1}}^{\theta_{k-1}}\left(x_{k-1}, \omega_{k-1}, a_{k-1}\right)$$

b) Critic Parameter:

$$z_{k-1} = \lambda z_{k-1} + \nabla_r \tilde{Q}_{r_{k-1}}^{\theta_{k-1}}\left(x_k, \omega_k, a_k\right)$$

c) Tunable Parameter r **Update:**

$$r_k = r_{k-1} + \gamma_k d_k z_k$$

d) Average Reward:

$$\tilde{v}_k = \tilde{v}_{k-1} + \eta_k \left(g_\chi \left(x_k, \omega_k, a_k \right) - \tilde{v}_{k-1} \tau_k \right)$$

e) Actor Parameter:

$$\theta_k = \theta_{k-1} + \beta_k \Gamma \left(r_{k-1} \right) \tilde{Q}_{r_{k-1}}^{\theta_{k-1}} \left(x_k, \omega_k, a_k \right) \psi^{\theta_{k-1}} \left(x_k, \omega_k, a_k \right)$$

8) **end for** k

Note that in line 6) of the algorithm, the modified immediate reward function in equation (4.13) is employed. The proposed algorithm is used to estimate the gradient of the average reward and the randomized policy μ_θ parameterized by $\theta \in R^M$. The parameter vector θ is gradually updated in the direction which improves the gradient. Under conditions that ensure the existence of the gradient, and condition which guarantee that the actor parameter vector θ is learned at a slower rate than the critic parameter r_k, it can be shown that the above algorithm will eventually approximate the optimal randomized policy μ_{θ^*} (Usaha and Barria, 2007).

4.5 Numerical Results

4.5.1 General Settings and results

In the numerical study, we assume a discrete event simulator which generates traffic streams for new call requests according to mutually independent Poisson processes. The mean call holding time is exponentially distributed. We compare performance metrics in terms of the long-term average reward and blocking

probability for each class. The proposed ACSMDP method will be compared with the CAC policy obtained from a conventional DP method in **chapter 3** proposed by Singh et al. (2002) and an empirical threshold policy. The DP method attains an optimal CAC policy whereas the threshold policy empirically attains a policy that satisfies the blocking probability constraint for each class.

The CDMA network under consideration has K_v=2 classes of voice users. Each class has a SIR level threshold given by $\beta = \left[\beta(1), \beta(2)\right]^T$, respectively and the minimum SIR level threshold is given by 20, 21 dB, respectively. We also assume that there is no fading to avoid the unpredictable noise in the channel. We study 6 cases with parameter settings as shown in Table 4.1-4.3. In Table 4.1, we test the ability to satisfy the blocking probability constraints of class 1 voice users by changing the blocking probability constraints. In Table 4.2, we test the ability to satisfy the blocking probability of class 2 voice users. This is done by increasing that arrival rate of class 2 users. Table 4.3 investigates a nontrivial scenario where the immediate rewards, arrival rates and the mean call holding time are all varied. The purpose is to test the performance of the proposed ACSMDP method whether it can guarantee the blocking probability requirements in such scenario. The parameter $P(i)$ denotes the transmission power for class i user which is 1.2 and 1.7 watts, respectively. The channel gain for class i is $\bar{h}(i)=1$, where $i=1,2$. Finally, the channel variance of class i user is given by $\xi^2(i)=0$, where $i=1,2$. N is the processing gain of the channel which is 32. Note that $\bar{h}(i)$, $P(i)$ and $\xi^2(i)$ are used to calculate, $\Psi_x(i)$, which is the actual SIR value of class i users when the system is in state x (See **Appendix I)**. It should be emphasized that the SIR level requirements

are inherently incorporated into the framework by the truncation of the state space to points that satisfy the SIR requirements as shown in equation (4.3). The SMDP formulation is then solved over the truncated state space.

Each CAC method is evaluated in a simulation with run length of 2×10^7 time steps. The stepsizes used in the ACSMDP method are

$$\gamma_k = \eta_k = \frac{0.0001}{1+\frac{k}{20000}} \tag{4.23}$$

and $\beta_k = \frac{\eta_k}{2}$. The normalizing parameter in the ACSMDP method is given by

$$\Gamma\left(r_k\right) = 0.5\left(\frac{1}{1+\left|r_k\right|}+\frac{2}{\left|r_k\right|}\right) \tag{4.24}$$

Table 4.1 Multiservice Parameters: cases 1-2

	Class1	**Class2**
Case 1, Blocking probability constraints, $B(i)$	0.001	-
Case 2, Blocking probability constraints, $B(i)$	0.0005	-
Mean arrival rate (call/min), $\lambda(i)$	12	12
Rewards ($), $r(i)$	8	4
Mean call holding time(min/call), $1/\mu(i)$	1.1	1.1

Table 4.2 Multiservice Parameters: cases 3-4

	Class1	**Class2**
Case 3, Blocking probability constraints, $B(i)$	-	0.01
Case 4, Blocking probability constraints, $B(i)$	-	0.005
Mean arrival rate (call/min), $\lambda(i)$	12	15
Rewards ($), $r(i)$	8	4
Mean call holding time(min/call), $1/\mu(i)$	1.1	1.1

Table 4.3 Multiservice Parameters: cases 5-6

	Class1	**Class2**
Case 5, Blocking probability constraints, $B(i)$	-	0.01
Case 6, Blocking probability constraints, $B(i)$	-	0.001
Mean arrival rate (call/min), $\lambda(i)$	16	7
Rewards ($), $r(i)$	2.5	10
Mean call holding time(min/call), $1/\mu(i)$	3	1

From the numerical study, the obtained results in Table 4.4-4.9 reveal that the proposed approach can achieve up to 91-95% of the average reward achievable by the DP method, while the blocking probability constraint is still satisfied. From Table 4.4-4.5, the results of case 1 and 2 show that our algorithm can achieve a near-optimal solution compared to dynamic programming solution whereas the empirical threshold policy cannot satisfy the required blocking probability constraint.

Table 4.4 Blocking probability measured for cases 1-2

	DP		**ACSMDP**		**Threshold Policy**	
	Class1	**Class2**	**Class1**	**Class2**	**Class1**	**Class2**
Unconstraint	0.0047	0.0181	0.0181	0.0177	0.0130	0.0130
Case1	0.0009	0.0455	0.0010	0.0460	0.0010	0.2537
Case2	0.0004	0.0651	0.0004	0.0831	0.0006	0.3140

Table 4.5 Average reward measured for cases 1-2

	DP	**ACSMDP**	**Threshold Policy**
	Avg. Reward	**Avg. Reward**	**Avg. Reward**
Unconstraint	142.810	142.759	142.135
Case1	141.964	141.600	131.659
Case2	140.986	138.224	128.945

In case 3 and 4, the blocking probability requirement of the ACSMDP method can be satisfied. Note that in these cases, the empirical threshold policy can also

satisfy the blocking probability constraint. However, its long-term average reward is less than the ACSMDP method.

In case 5 and 6, results also show that the ACSMDP method can consistently perform well under nontrivial scenarios.

Table 4.6 Blocking probability measured for cases 3-4

	DP		**ACSMDP**		**Threshold Policy**	
	Class1	**Class2**	**Class1**	**Class2**	**Class1**	**Class2**
Unconstraint	0.0090	0.062	0.0390	0.0270	0.0350	0.0350
Case3	0.0840	0.010	0.1130	0.0090	0.2540	0.0010
Case4	0.1130	0.005	0.1484	0.0050	0.3110	0.0060

Table 4.7 Average reward measured for cases 3-4

	DP	**ACSMDP**	**Threshold Policy**
	Avg. Reward	**Avg. Reward**	**Avg. Reward**
Unconstraint	151.690	151.283	150.097
Case3	147.647	144.371	131.012
Case4	145.187	142.007	125.631

Table 4.8 Blocking probability measured for cases 5-6

	DP		**ACSMDP**		**Threshold Policy**	
	Class1	**Class2**	**Class1**	**Class2**	**Class1**	**Class2**
Unconstraint	0.0960	0.033	0.1080	0.0380	0.0870	0.0420
Case5	0.0590	0.007	0.0890	0.0083	0.0940	0.0090
Case6	0.0460	0.001	0.0630	0.0010	0.0850	0.0010

Table 4.9 Average reward measured for cases 5-6

	DP	**ACSMDP**	**Threshold Policy**
	Avg. Reward	**Avg. Reward**	**Avg. Reward**
Unconstraint	104.085	103.438	102.147
Case5	107.643	106.719	104.772
Case6	107.993	107.281	106.034

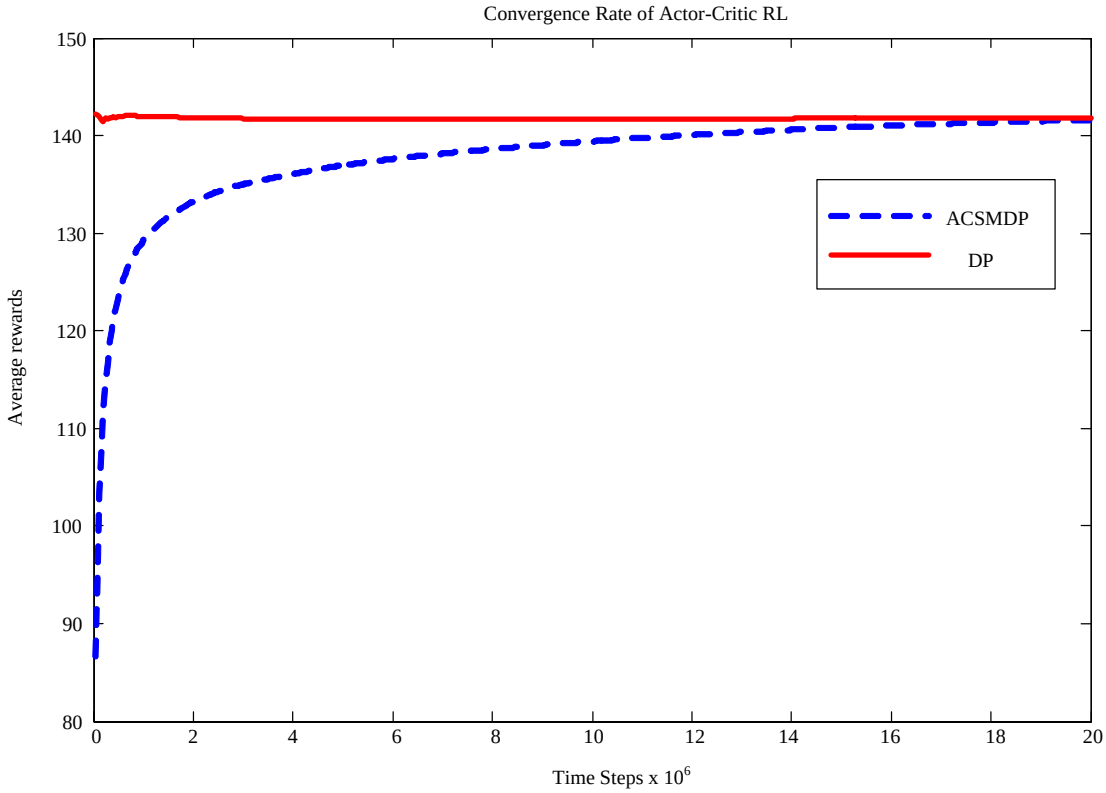

Figure 4.2 Learning curve of ACSMDP method

Figure 4.2 illustrates that the learning curve of the ACSMDP method required in the training process before it approaches the long-term average reward of the dynamic programming method. Due to the long training time (i.e. up to about 10^7 events), the graph suggests that the ACSMDP method should be trained offline prior to actual online implementation in order to achieve a near-optimal CAC policy.

Figure 4.3-4.4 illustrates the SIR level under case 2 settings of the ACSMDP method in the training process for each class. During the training of the ACSMDP, the SIR does not fall below the SIR threshold limit.

Figure 4.5-4.6 illustrate that all policies can maintain the SIR levels for both classes. Note that after training, the ACSMDP can still satisfy the SIR requirements

throughout the simulation process. The tests in Figure 4.5-4.6 used a runlength of 10^7 time steps.

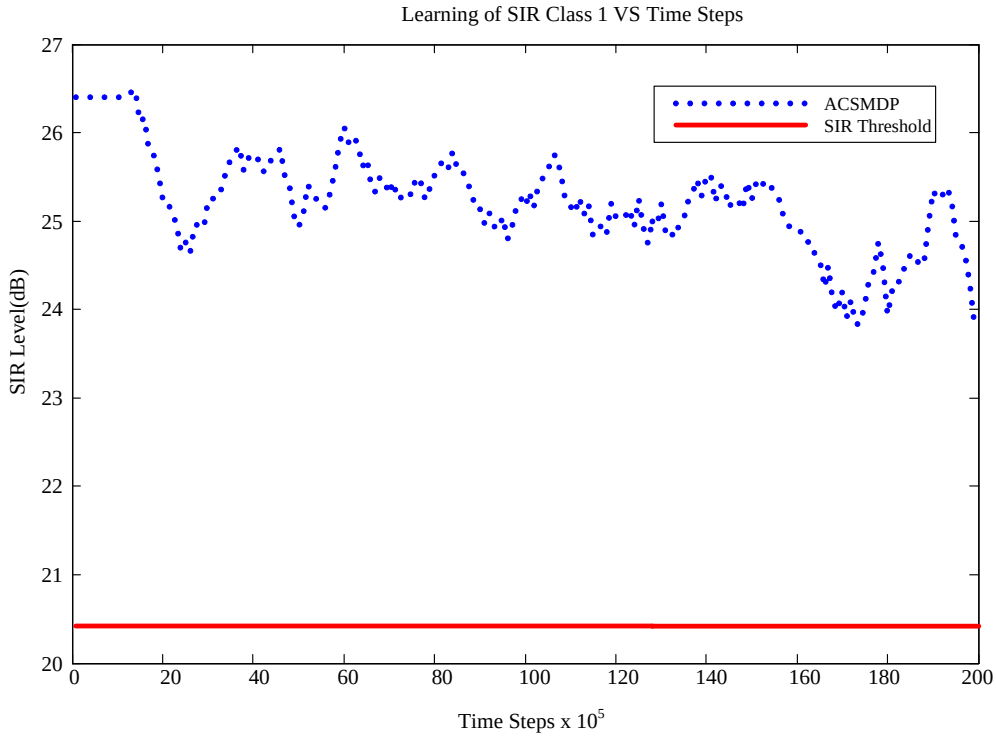

Figure 4.3 SIR level of class 1 user in training mode of ACSMDP

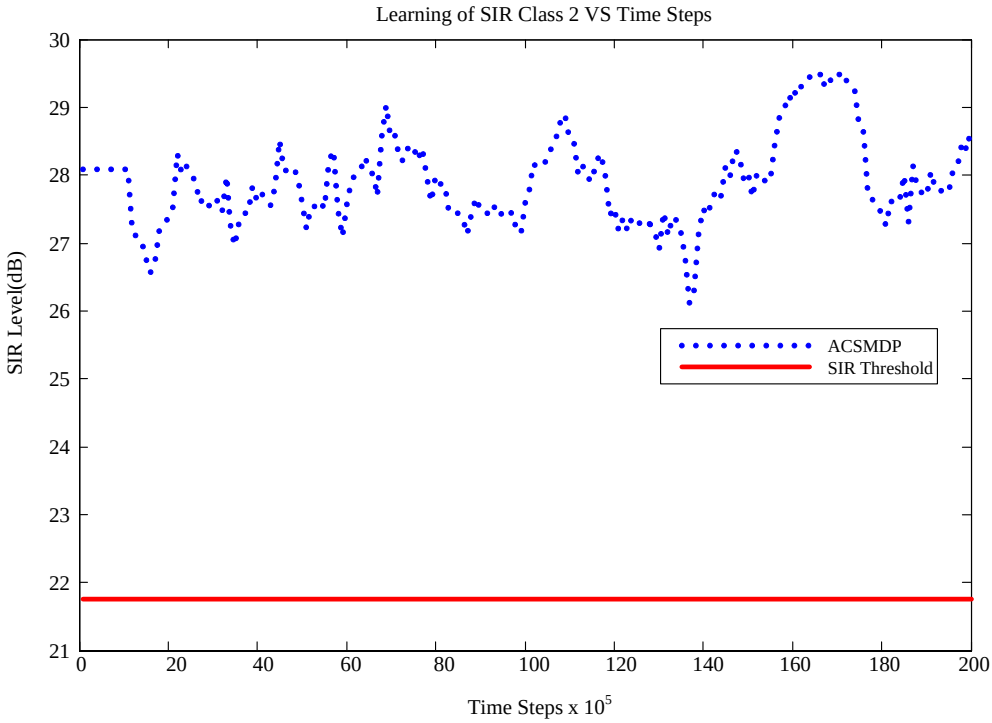

Figure 4.4 SIR level of class 2 user in training mode of ACSMDP

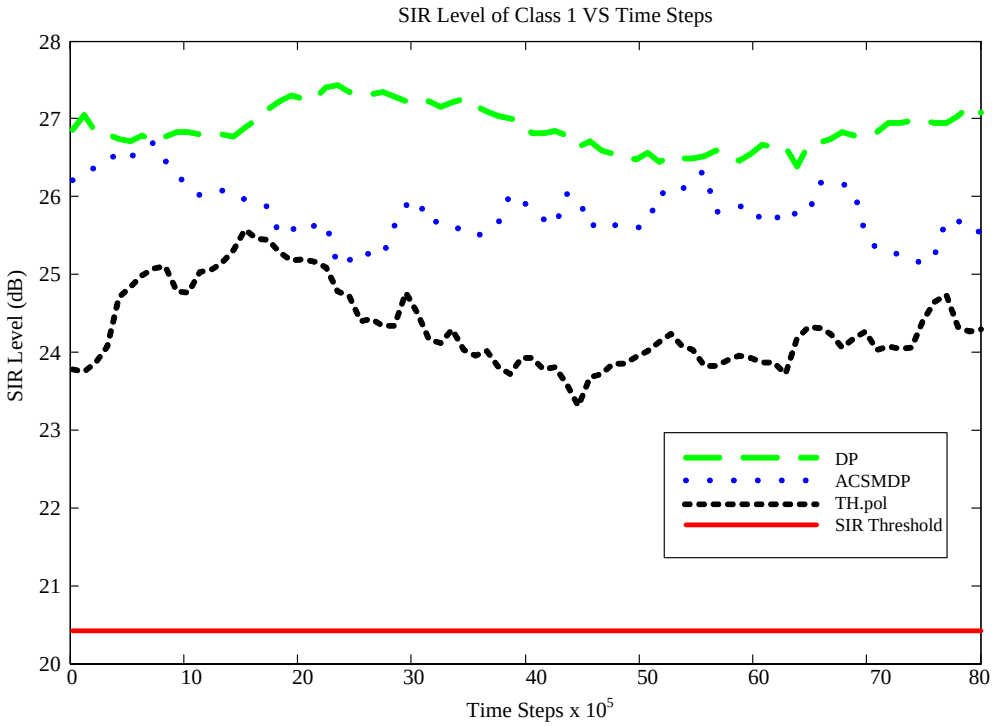

Figure 4.5 SIR of class 1 users for each policy

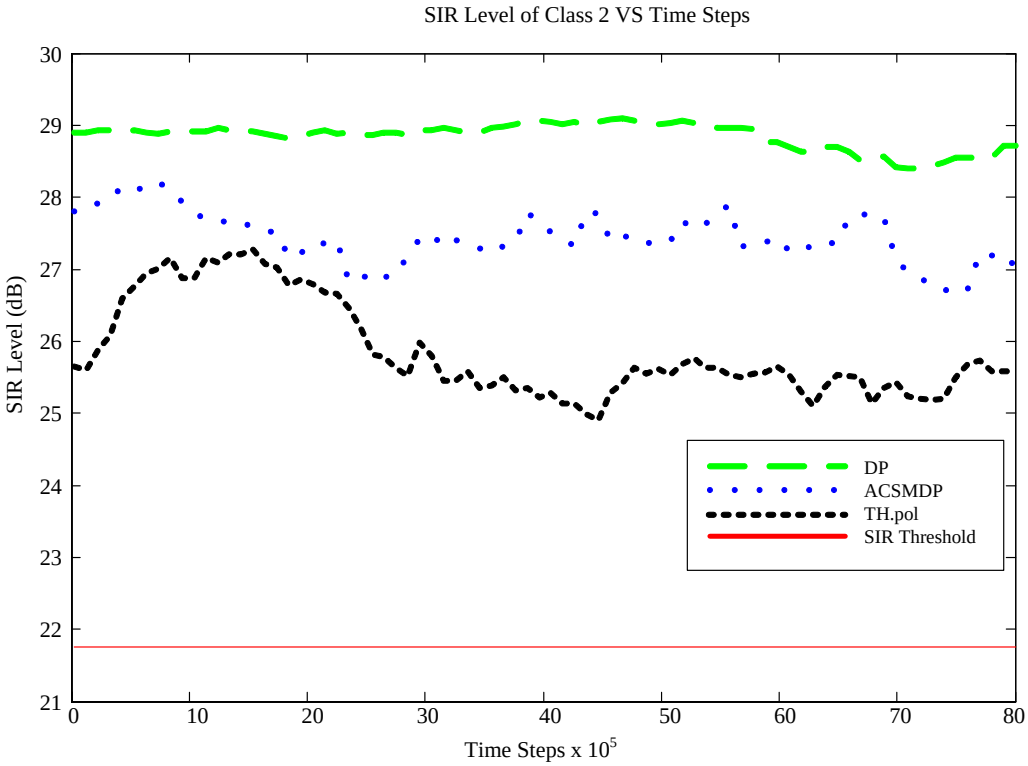

Figure 4.6 SIR of class 2 users for each policy

4.5.2 Memory Storage Analysis

In terms of storage requirements, it is stated in **chapter 3** that the DP method requires a complete knowledge of the system dynamics in terms of state transition probability matrix. In particular, the transition probability matrix requires a number of parameters of $|X|\times|A|$ parameters, where $|X|=32^2$ and $|A|=4$ are the size of state spaces and action spaces, respectively. Therefore, in this numerical study, the total number of parameters needed for the transition probability matrix is $32^2\times 4=4096$ parameters. On the contrary, the ACSMDP method does not require a complete knowledge of the system dynamics. It can improve CAC policies by learning from direct interaction with the environment. However, the ACSMDP method still needs storage for i) the tunable parameters, which are the actor parameter

vector $\theta \in R^M$ and the critic parameter vector $r \in R^{M+1}$, ii) the feature structures which are $\psi^\theta \in R^M$ for the actor part and $\phi^\theta \in R^{M+1}$ for the critic part; iii) the eligibility trace for the critic part, $z \in R^{M+1}$, where $M = K_v \times |A|$ and K_v is the number of classes of voice users in the system. Consequently, the total number of parameter usage for the ACSMDP method is

$$|\theta| + |r| + |\psi^\theta| + |\phi^\theta| + |z| = 2M + 3(M+1) \tag{4.25}$$

Hence, the ACSMDP method requires $2 \times (2 \times 2) + 3 \times ((2 \times 2) + 1) = 23$ parameters. In this numerical study, the ACSMDP method demands significantly lower memory storage requirement than the DP counterpart. Thus, as the scale of the CDMA network under consideration increases, either by increasing the number of users in the system, the number of user classes, reducing the SIR level requirement thresholds, increasing the service area, etc., the state space of the system grows exponentially. This is referred to as the curse of dimensionality. Under these circumstances, the ACSMDP method can still scale well and provide near-optimal CAC decisions to the optimal CAC policy obtained by DP. The savings of the memory storage requirement of the ACSMDP method over the DP method becomes even more significant as the scale of the state space increases.

4.5.3 Complexity Analysis

In this subsection, we analyze the computational complexity of the DP and the ACSMDP methods, in terms of the amount of computation required to compute one online decision. Let the size of the state space of the considered system

be denoted by $|X|$ and the size of action space of the call admission controller be denoted by $|A|$. Recall that K_v denotes the number of classes in the system.

For the DP method, the amount of computation required for computing the optimal action, which is constructed from the linear programming solution z_{xa} in equations (3.15) and (3.16), is $O(|X||A|)$. On the other hand, the amount of computation required for each action selection for the ACSMDP method in equation (4.16) is $O(|\theta|) = O(K_v|A|)$. Note that $|X|$ grows exponentially as a function of K_v and the SIR level requirements. Hence, it can be seen that the amount of computation required in computing one online decision in the ACSMDP method is significantly less than that of the DP method. Furthermore, the savings in the computational requirement becomes even more apparent as the size of the state space increases.

4.6 Conclusion

In this chapter, a reinforcement learning approach called the actor-critic for semi-Markov decision process (ACSMDP) method is employed to learn a near-optimal call admission control decision policy for a multiple voice service DS-CDMA cellular system with QoS constraints at the physical layer (SIR level) and network layer (blocking probability). The approach circumvents the curse of modeling and dimensionality of the conventional DP method. The SMDP formulation has been reformulated and differs from that of **chapter 3** where a modified reward signal has been employed to deal with the blocking probability constraints.

Numerical study shows that the proposed approach can achieve an average reward of 91-95% of that obtained from the DP method and still satisfy the QoS

constraints. Furthermore, the storage requirements and the amount of computational complexity to compute an online decision of the proposed approach is generally less demanding than the DP approach which allows easy implementation. However, the tradeoff of using the ACSMDP method is the requirement of tuning and training the parameters involved in the algorithm.

It should be noted that the parameter training in the ACSMDP method can be performed offline. However, in the event of unpredictable changes in real CDMA networks or drastic traffic variations, the parameters can be recomputed offline and uploaded to the call admission controller in the base station in a timely manner.

CHAPTER V

CONCLUSIONS

5.1 Conclusion

In this thesis, we proposed a framework that enables multiple QoS constraints for adaptive call admission control in wireless DS-CDMA systems with multiclass voice users based on reinforcement learning. The work carried out in this thesis can be divided into two parts which are dynamic programming and reinforcement learning. The dynamic programming approach for determining an optimal call admission control policy is presented in **chapter 3**. An actor-critic reinforcement learning approach which is employed to solve for near-optimal call admission control policies is presented in **chapter 4**. The findings of this thesis can be summarized as follows.

5.1.1 Chapter 3. Call Admission Control in Wireless DS-CDMA Systems: A DP Approach

The purpose of this chapter is to study the call admission control problem by using the conventional dynamic programming method which has been proposed by Singh, Krishnamurthy and Poor (2002). In this chapter, the formulation of the problem is formulated as the semi-Markov decision process (SMDP) and the solution is exactly solved using linear programming (LP). In the numerical study, the performance in terms of the blocking probability controlled and the long-term average reward is compared between dynamic programming, complete sharing and threshold

policies under a small network scenario. The results show that the dynamic programming method can achieve the maximum long-term average reward while the desired blocking probability can be satisfied.

5.1.2 Chapter 4. Call Admission Control in Wireless DS-CDMA Systems: A RL Approach

The purpose of this chapter is to extend the call admission control (CAC) scheme to more realistically larger network. A type of reinforcement learning (RL) technique namely the “actor-critic” reinforcement learning has been successfully employed to solve the CAC problem in a large scale wireless DS-CDMA network. The large scale of the problem is obtained by increasing the capacity of the system in **chapter 3.** The CAC problem is cast as a semi-Markov decision process (SMDP) by using a modified reward signal to deal with the blocking probability constraints. The obtained numerical results in this chapter reveal that the actor-critic method can achieve an average reward between 91-95% of the optimal average reward achievable by the DP method while constraints can still be satisfied. Furthermore, in the terms of storage and computational requirements needed to compute an online decision of the proposed approach is generally demanding than the DP approach. We make note that the savings of the memory storage and computational requirements of the proposed method over the DP method becomes even more significant as the scale of the CDMA network under consideration increases.

5.2 Recommendation for Future Work

5.2.1 Prioritized in Handover for Adaptive Call Admission Control

In this thesis, we focus on the call admission control problem in the uplink where only new calls arrivals are considered. We can extend the framework to consider handoff calls whereby priority should be given over the new call arrivals as a forced termination of an existing call is likely to cause more dissatisfaction than blocking of new calls. Using this framework, we can construct a near-optimal CAC policy that minimizes the probability of dropping handoff calls or include handoff call dropping probability as additional constraints.

5.2.2 Multiclass Data and Voice Services for Wireless DS-CDMA

The actor-critic method can be extended to deal with multiservice voice and data call admission control. The arriving data user can be admitted into a buffer (all arrivals of data users must be buffered by default). The CAC problem can be reformulated to decide whether to admit queued data users or voice users subject to QoS constraints for both data and voice services as suggested in Singh et al (2002).

5.2.3 Optimization of Reward Signals Design

In **chapter 4**, the main focus of the chapter is to modify the reward signals to deal with the QoS constraints in call admission control problem in CDMA systems. However, the performance of the proposed algorithm depends on the modified reward signals. Investigations on how to find the best possible modified reward signal remains a subject for further study.

5.2.4 Optimization in Parametric Tuning

The actor-critic algorithm’s performance criterion, learning rate and function approximation in the actor and critic parts are affected by the tunable

parameters. The optimization of these tunable parameters in the actor-critic algorithm therefore warrants further investigation.

5.2.5 Comparison with other Actor-Critic Approaches

In **chapter 4**, the actor-critic algorithm is chosen as it combines two strong points of actor-only and critic-only RL methods. However, comparisons between the actor-critic method proposed in this thesis and other types of actor-critic reinforcement learning approaches in terms of learning rate, complexity, memory storage and performance criterion has not yet been covered in this thesis and is a matter worthwhile to investigate.

REFERENCES

Bartolini, N. and Chlamtac, I. (2002). Call admission control in wireless multimedia networks. **The 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2002**, pp: 285-289.

Bertsekas, D.P. (1995). **Dynamic Programming and Optimal Control Volume II**. Massachusetts Institute of Technology.

Bertsekas, D.P. and Tsitsiklis, J.N. (1996). **Neuro-Dynamic Programming.** Belmont, MA. Athena Scientific.

Chanloha, P. and Usaha, W. (2007). Call Admission Control in Wireless DS-CDMA systems using Actor-Critic Reinforcement Learning. **The 2nd IEEE International Symposium on Wireless Pervasive Computing.**

Chen, A.C. (1998). Overview of Code Division Multiple Access Technology for Wireless Communications. **Industrial Electronics Society, 1998. IECON '98. Proceedings of the 24th Annual Conference of the IEEE**, pp: T15-T24.

Choi, J., Kwon, T., Choi, Y. and Naghshineh, M. (2000). Call Admission control for multimedia services in mobile cellular networks: a Markov decision approach. **IEEE International. Symposium Computers. Communnications**, pp: 594-599.

Comaniciu, C. and Narayan, M.B. (2000). QoS guarantees for third generation (3G) CDMA systems via admission and flow control. **Vehicular Technology Conference**, pp: 249-256.

Comaniciu, C. and Poor, H. V. Jointly Optimal Power and Admission Control for Delay Sensitive Traffic in CDMA Networks With MMSE Receivers. **IEEE Transactions on Signal Processing**, pp: 2031-2042

El-Alfy, E., Yao, Y. and Heffers, H. (2001). Autonomous Call Admission Control with Prioritized Handoff in Cellular Networks. **IEEE International Conference on Communications**, pp: 1386-1390.

Hoshino, Y. and Kamei, K. (2003). A Proposal of Reinforcement Learning System to Use Knowledge Effectivity**.** SICE2003 **Annual Conference**, pp: 1582-1585.

Lee, T.H. and Wang, J.T. (1998). Admission control for VSG-CDMA systems supporting integrated services. **IEEE Global Telecommunications Conference**, pp: 2050–2055.

Lee, W.C.Y. (1991). Overview of cellular CDMA. **IEEE Transactions on Vehicular Technology**, pp: 291-302.

Liao, Y.C., Yu, F. Leung, V.C.M. and Chang, J.C. (2006). A novel dynamic cell configuration scheme in next-generation situation-aware CDMA networks. **IEEE Journal on Selected Areas in Communications**, pp: 16-25.

Lilith, N. and Dogancay, K. (2005). Distributed Dynamic Call Admission Control and Channel Allocation Using SARSA. **IEEE TENCON 2005,** pp: 1-6.

Liu; D., Zhang; Y. and Zhang, H. (2005). A Self-Learning Call Admission Control Scheme for CDMA Cellular Networks. **IEEE Transactions on Neural Networks**, pp: 1219–1228.

Liu, T. and Silvester, J.A. (1998). Joint admission/congestion control for wireless CDMA systems supporting integrated services. **IEEE Journal on Selected Areas in Communications**, pp: 845–857.

Liu, Z. and Zarki, M.E. (1994). SIR-based call admission control for DS-CDMA cellular systems. **IEEE Journal on Selected Areas in Communications**, pp: 638–644.

Makarevitch, B. (2000). Application of reinforcement learning to admission control in CDMA network. **Proc. of Personal, Indoor and Mobile Radio Communication**, pp: 1353-1357.

Marbach, P. and Tsitsiklis, J.N. (1999). Simulation-Based Optimization of Markov Reward Processes: Implementation Issues. **Proceedings of the 38th IEEE Conference on Decision and Control**, pp: 1769-1774.

Pandana, C. and Liu, K.J.R (2004). Throughput maximization for energy efficient multi-node communications using actor-critic approach. **IEEE Global Telecommunications Conference,** pp: 3578-3582.

Prasad, R. (1998). A Survey on CDMA: Evolution Towards Wideband CDMA. **IEEE 5th International Symposium on Spread Spectrum Techniques and Applications,** pp: 323-331.

Rao, R.M., Comaniciu, C., Lakshman, T.V. and Poor V.H. (2004). Call Admission Control in Wireless Multimedia Network. **IEEE Signal Processing Magazine**, pp: 51-58.

Rappaport, T.S. (2002) **Wireless Communication.** Prentice Hall, Inc.

Ross, K.W. (1995). **Multiservice Loss Models for Broadband TelecommunicationNetworks**. London, U.K.: Springer-Verlag.

Ross, K.W. and Tsang, D. (1989) Optimal circuit access policies in an ISDN environment: A Markov decision approach. **IEEE Transactions on Communnications,** pp: 934-939.

Soroushnejad, M. and Geraniotis, E. (1995). Multi-access strategies for an integrated voice/data CDMA packet radio network. **IEEE Transactions on Communnications**, pp: 934–945.

Singh S., Krishnamurthy V. and Poor V. H. (2002). Integrated Voice/Data Call Admission Control for Wireless DS-CDMA Systems. **IEEE Transactions on Signal Processing**, pp: 1481-1495.

Singh, S. and Bertsekas, D.P. (1997). Reinforcement learning for dynamic channel allocation in cellular telephone systems. **Proc. of Advances in Neural Information Processing Systems 10**, pp: 974-980.

Sung; K. Y., Hwang; H. R., Chen; M.X and Hsu, J. M. (2004). Adaptive Call Admission Control Mechanism for DS-CDMA Cellular System. **The IEEE 6th Circuits and Systems Symposium on Emerging Technologies: Frontiers of Mobile and Wireless Communication**, pp: 549-522.

Sutton, R.S. and Barto, A.G. (1998). **Reinforcement Learning: An Introduction**. Cambridge, MA: MIT Press.

Tijms, H.K. (1986). **Stochastic Modeling and Analysis: A Computational Approach**. Wiley&Son.

Tong, H. (1999). **Adaptive Admission Control and Routing under Quality of Service Constraints in Broadband Communications.** Doctor of Philosophy Thesis, University of Colorado, Boulder, United State of America.

Tong H. and Brown, X. T. (1999) Adaptive Call Admission Control Under Quality of Service Constraints: A Reinforcement Learning Solution. **IEEE Journal on Selected Areas in Communications,** pp: 209-221.

Usaha, W. (2004). **Resource Allocation in Networks with Dynamic Topology**. PhD Thesis,University of London, London, U.K.

Usaha, W. and Barria, J. (2007) Reinforcement Learning for Resource Allocation in LEO satellite networks. **IEEE Transactions on Systems, Man, and Cybernetics** Part B. Volume 37, Issue 3, pp: 515-527.

Vazquez-Abad, F.J. and Krishnamurthy, V. (2002). Self learning Call Admission Control for multimedia wireless DS-CDMA Systems. **Proc. of International Workshop on Discrete Event Systems**, pp: 399-404.

Yang, W. and Geraniotis E. (1994). Admission policies for integrated voice and data traffic in CDMA packet radio networks. **Journal on Selected Areas in Communications**, pp: 654–664.

APPENDIX I

SIR Computation

SIR Computation

The call admission control mechanism relies on the soft capacity of the CDMA network which is characterized by the SIR level. The SIR measurement in CDMA call admission control problems presented in this appendix is summarized from (Liu and Zarki, 1994). Suppose that there are M cells is in DS-CDMA network of interest with n_k calls in progress in cell k. Let $R(s)$ denote the received field strength. We model the reception at the receiving antenna of a particular cell's base station (BS) by taking into account of the path loss, log-normal shadowing, and multipath fading through the following expression

$$R(s) = 10^{\zeta/10} s^{-\alpha} \tag{A.1}$$

where α is a constant typically ranging from two to four, s is the distance between the receiver and transmitter, and ζ is the transmit field strength in decibels (dB) which is normally distributed. The total power received by the BS in cell k is the sum of the power from all the mobiles in the system and is given by

$$I(k) = \sum_{l=1}^{M} \sum_{i=1}^{n_h} I_i(l,k) \tag{A.2}$$

where $I_i(l,k)$ is the power received by cell k's BS from mobile i of cell l. Suppose that, with ideal power control i.e., each mobile's signal is perceived with the

same strength at its BS, S is the power level of a mobile's signal at its home cell BS and $s_{ik}^{(l)}$ is the distance between mobile i of cell l and the BS of the cell k. Then

$$I(k) = Sn_k + S\sum_{l\neq k}\sum_{i=1}^{n_l}\left(\frac{s_{il}}{s_{ik}^{(l)}}\right)^{-\alpha} 10^{(\zeta_{ik}-\zeta_{il})/10} \tag{A.3}$$

Equation (A.3) assumes the controlled power per mobile at each home BS is the same. Thus, the SIR at BS k is given by

$$\mathrm{SIR_k} = \frac{S}{I(k) - S} \tag{A.4}$$

The variable quantity in the expression for SIR in (A.4) is $I(k)$, the total power received at the k-th BS. Equation (A.3) shows that this is a random variable because it depends on several other random variables, namely, the number of callers, their positions, and the transmitted power of interfering calls in neighboring cells. The capacity of CDMA systems is limited by the level of multiaccess interference in the system, which is measured by the SIR. In general, because the SIR drops as the number of users increases, it appears reasonable to maintain the SIR level above the set thresholds by limiting the number of incoming users.

Although equation (A.4) looks at SIR purely as it is perceived at the BS antenna, it is obviously affected by the processing at the receiver. The standard signal model for DS-CDMA systems in fading environment has a received signal $\mathbb{R}_{s_m}$ for the m-th transmitted bit which is given by (Singh et al., 2002)

$$\mathbb{R}_{s_m} = \sqrt{P_1} b_{1,m} h_{1,m} \upsilon_1 + \sum_{n=2}^{N} \sqrt{P_n} b_{n,m} h_{n,m} \upsilon_n + \sigma w_m \tag{A.5}$$

where P_n, $b_{n,m} = \pm 1$ and $h_{n,m}$ denote the transmit power, the m-th transmitted bit, and the channel gain respectively for the n-th user and $\upsilon_n \in R^N$ is the n-th users' signature sequence. The final additive term denotes the white noise with variance σ^2 with w_m being a zero mean unit variance, circularly symmetric, complex Guassian random variable. Suppose that the sequence $h_{n,m}$ is also complex valued and random with mean and variance denoted by $\bar{h}_n$ and ξ_n^2, respectively. The SIR expression can then be approximated by (Rao, Comaniciu, Lakshman, and Poor, 2004)

$$\mathrm{SIR}(1) \approx \widehat{\mathrm{SIR}}(1) \triangleq \frac{P_1 \left|\bar{h}(1)\right|^2 \beta}{1 + P(1) \xi^2(1) \beta} \tag{A.6}$$

Equation (A.6) is the signal-to-interference ratio (SIR) of the CDMA system. The parameter $\bar{h}(i)$ is the channel gain of the system and $\xi^2(i)$ is the variance of the system where i denotes the class of users where $i = 1, 2, ..., K_v$, K_v denotes the number of class. Parameter β is the unique fixed point in interval $(0, \infty)$. This fixed point equation can be expressed as follows

$$\beta = \left[\sigma^2 + \frac{1}{N} \sum_{i=1}^{K_v} I\left(P_i \left(\xi^2(i) + \left|\bar{h}(i)\right|^2 \right), \beta \right) \right]^{-1} \tag{A.7}$$

where

$$I(p,\beta) \triangleq \frac{p}{1+p\beta} \tag{A.8}$$

From the fixed point equation in (A.7) and (A.8), we can express and construct the SIR function $f_x^{\beta}(i)$ for the capacity constraint as follows

$$f_x^{\beta}(i) = \beta_x(i) = \left[\sigma^2 + \frac{x(i)-1}{N} I\left(P(i)\xi^2(i) + P(i)\left|\overline{h}(i)\right|^2, \beta\right) + \sum_{j=1, j\neq i}^{K_v} \frac{x(j)}{N} I\left(P(j)\xi^2(j) + P(j)\left|\overline{h}(j)\right|^2, \beta\right)\right]^{-1} \tag{A.9}$$

where N is the channel gain of the system, I is defined in (A.8), $x(i)$ is the number of class i voice users currently in the system where $i = 1, 2, ..., K_v$, $P(i)$ is the power transmission for each class i voice user. In this thesis, the power transmission is a deterministic value for each class i voice user. Finally, the SIR function $\Psi_x(i)$ can be defined as follows

$$\Psi_x(i) = \begin{cases} \infty & \text{,if } x(i) = 0 \\ \dfrac{P(i)\left|\overline{h}(i)\right|^2}{\sigma^2 + P(i)\xi^2(i)} & \text{,if } x = e(i) \\ \dfrac{P(i)\left|\overline{h}(i)\right|^2 \beta}{1 + P(i)\xi^2(i)\beta} & \text{,else } \beta > 0 \text{ and } \beta = f_x^{\beta}(i) \end{cases} \tag{A.10}$$

where $e(i)$ denotes the vector with all elements equal to zero except the i-th component which is unity. Equations (A.9) and (A.10) evaluate the classwise SIR level which are used to determine the CDMA soft capacity constraints.

APPENDIX II

List of Publications

List of Publication

Chanloha, P. and Usaha, W. (2007). Call Admission Control in Wireless DS-CDMA systems using Actor-Critic Reinforcement Learning. **The 2nd IEEE International Symposium on Wireless Pervasive Computing.**

BIOGRAPHY

Mr. Pitipong Chanloha was born on June 16, 1983 in Bangkae District, Bangkok Province. In 2000, he began studying for his Bachelors degree at School of Telecommunication Engineering, Institute of Engineering at Suranaree University of Technology, Nakhon Ratchasima Province. After graduating, he continued to study for a Masters degree at the School of Telecommunication Engineering, Institute of Engineering, Suranaree University of Technology.