creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

조선대학교
CHOSUN UNIVERSITY

February 2024
Master's Degree Thesis

Multi-Objective Reinforcement Learning-based Power Allocation for Massive MIMO Networks

- A Study on Spectral and Energy Trade-Off Optimization -

Graduate School of Chosun University

Department of Computer Engineering

Youngwoo Oh

조선대학교
CHOSUN UNIVERSITY

Multi-Objective Reinforcement Learning-based Power Allocation for Massive MIMO Networks

- A Study on Spectral and Energy Trade-Off Optimization -

Massive MIMO 네트워크를 위한 다중목표 강화학습 기반 전력
할당: 스펙트럼 및 에너지 Trade-Off 최적화에 대한 연구

February 23, 2024

Graduate School of Chosun University

Department of Computer Engineering

Youngwoo Oh

조선대학교
CHOSUN UNIVERSITY

Multi-Objective Reinforcement Learning-based Power Allocation for Massive MIMO Networks

- A Study on Spectral and Energy Trade-Off Optimization -

Advisor: Wooyeol Choi

A thesis submitted in partial fulfillment of the
requirements for a Master's degree

October 2023

Graduate School of Chosun University

Department of Computer Engineering

Youngwoo Oh

조선대학교
CHOSUN UNIVERSITY

오영우의 석사학위논문을 인준함

위원장 신석주 (인)

위 원 강문수 (인)

위 원 최우열 (인)

2023년 12월

조선대학교 대학원

조선대학교 CHOSUN UNIVERSITY

Contents

List of Contents i

List of Tables iii

List of Figures iv

Abstract (English) vii

Abstract (Korean) ix

Acknowledgments xi

Ⅰ. Introduction 1

A. Research Background 1

B. Contributions 5

C. Thesis Organization 6

Ⅱ. Related Work 7

A. Deep Reinforcement Learning Approach 7

B. Multi-Agent Reinforcement Learning Approach 8

C. Multi-Objective Reinforcement Learning Approach 9

Ⅲ. Downlink Multi-Cell Massive MIMO Network 11

A. Channel Estimation and Spectral Efficiency 11

B. Power Consumption Model and Energy Efficiency 15

C. Definition of Joint Optimization Problem 17

조선대학교
CHOSUN UNIVERSITY

Ⅳ. MORL Algorithms and MOO Problem Conversion 18
A. MORL Technique for MOO Problem ······ 18
B. MOMDP-Based Conversion of MOO to SOO ······ 22

Ⅴ. Proposed MO-A3Cs Model for Power Allocation 25
A. Bayesian Rule-Based Preference Weights Updates ······ 28
B. Multi-Objective Advantage Function ······ 30
C. Optimization of MO-A3Cs Model Updates ······ 33

Ⅵ. Simulation Results and Analyses 37
A. Simulation Parameters ······ 37
B. MO-A3Cs Architecture ······ 38
C. Benckmark Methods ······ 39
1. Algorithmic Approaches ······ 40
2. Reinforcement Learning Approaches ······ 41
D. Performance Comparison and Analysis ······ 41
1. Training Performance Evaluation ······ 42
2. Simulation Results ······ 61

Ⅶ. Conclusion 69
A. Summary ······ 69
B. Future Work ······ 70

Abbreviations 71

References 73

조선대학교
CHOSUN UNIVERSITY

List of Tables

6.1 System parameters of the downlink multi-cell massive MIMO network setup. 38

6.2 Hyperparameters of the utilized DRL and MORL models. 39

6.3 Impact of quantization on training efficiency of the MO-A3Cs model. 48

6.4 Comparison of trade-offs optimization in downlink transmit PA methods across varying numbers of transmit antenna M from 20 to 100. 64

6.5 Comparison of trade-offs optimization in downlink transmit PA methods across varying maximum transmit power $P_{\max}$ from 20 to 60. 64

조선대학교
CHOSUN UNIVERSITY

List of Figures

1.1 Illustrative example of the conventional and massive MIMO systems. 1

1.2 Illustration of increased energy consumption according to the evolving wireless networks. ········· 2

3.1 Illustration of the downlink multi-cell massive MIMO networks. ····· 11

4.1 Comparison of DRL and MORL algorithms: Interaction between the agent and environment. ········· 18

4.2 Illustration of policy convergence in the MORL algorithms driven by preference weights for solving the MOO problem; (a) Initial steps of the MORL model training process, (b) Adjusting step of the policy by using preference weights, (c) Policy convergence step. ········· 20

5.1 The proposed MO-A3Cs based transmit PA framework for SE-EE joint optimization in the downlink multi-cell massive MIMO networks. ·· 26

5.2 Structure of the neural network of the single actor and multi-critics network on the proposed MO-A3Cs model. ········· 26

6.1 Training results of MO-A3Cs model at different learning rates: Average SE reward. ········· 42

6.2 Training results of MO-A3Cs model at different learning rates: Average EE reward. ········· 42

6.3 Training results of MO-A3Cs model at various discount factors: Average SE reward. ········· 44

조선대학교
CHOSUN UNIVERSITY

6.4 Training results of MO-A3Cs model at various discount factors: Average EE reward. ········· 44

6.5 Training results of MO-A3Cs model according to the different number of deployed local agents: Average SE reward. ········· 46

6.6 Training results of MO-A3Cs model according to the different number of deployed local agents: Average EE reward. ········· 46

6.7 Training results of MO-A3Cs model at different quantization levels: Comparison of training efficiency. ········· 48

6.8 Training results of MO-A3Cs model for each different preference weights initialization technique: Average SE reward. ········· 50

6.9 Training results of MO-A3Cs model for each different preference weights initialization technique: Average EE reward. ········· 50

6.10 Training results of MO-A3Cs model for each different preference weights update methods: Average SE reward. ········· 52

6.11 Training results of MO-A3Cs model for each different preference weights update methods: Average EE reward. ········· 52

6.12 Training results of MO-A3Cs model at different reward-aggregation methods: Average SE reward. ········· 54

6.13 Training results of MO-A3Cs model at different reward-aggregation methods: Average EE reward. ········· 54

6.14 Validation of the proposed MO-A3Cs model: Comparison and analysis of Average SE reward with PQN model. ········· 56

6.15 Validation of the proposed MO-A3Cs model: Comparison and analysis of average EE reward with PQN model. ········· 56

조선대학교 CHOSUN UNIVERSITY

6.16 Validation of the proposed MO-A3Cs model: Comparison and analysis of convergence results with various benchmark models. ········ 58

6.17 Validation of the proposed MO-A3Cs model: Comparison of multi-objective reward with various benchmark models. ········ 60

6.18 An evaluation of the total power consumption for downlink transmit PA methods across different numbers of transmit antenna M from 20 to 100. ········ 61

6.19 Comparison of the CDF for average SE across different downlink transmit PA methods at varying transmit antenna M from 20 to 100. ········ 62

6.20 Comparison of the CDF for average EE across different downlink transmit PA methods at varying transmit antennas M from 20 to 100. ········ 62

6.21 Performance evaluation and analysis of average SE and EE for each downlink PA method in densely deployed UEs per each cell in massive MIMO networks including pre-trained models; (a) Average SE with varying numbers of UEs, (b) Average EE with varying numbers of UEs. ········ 66

6.22 Execution time comparison for each downlink PA method for different numbers of UEs and $L = 16$ and $M = 40$ including the pre-trained models. ········ 68

조선대학교
CHOSUN UNIVERSITY

Abstract

Multi-Objective Reinforcement Learning for Power Allocation in Massive MIMO Networks: A Study on Spectral and Energy Trade-Off Optimization

Oh, Youngwoo
Advisor : Prof. Choi, Wooyeol, Ph.D.
Department of Computer Engineering,
Graduate School of Chosun University

The joint optimization of spectral and energy efficiency through power allocation techniques is a critical requirement for emerging fifth-generation and beyond networks. While various algorithmic approaches, such as genetic algorithms and convex optimization, have been considered for optimizing the trade-offs between spectral and energy efficiency in cellular networks, these methods suffer from high computational costs. Deep reinforcement learning-based methods have shown promise in addressing the computational challenges of single-objective optimization problems in wireless networks. Despite the potential of deep reinforcement learning approaches, utilizing them for the joint optimization of spectral and energy efficiency has yet to be noticed in the existing literature.

In this thesis, we propose a downlink transmit power allocation method based on a multi-objective asynchronous advantage single actor–multiple critics model. This

조선대학교 CHOSUN UNIVERSITY

method aims to optimize spectral and energy efficiency trade-offs in massive multiple-input-multiple-output assisted multi-cell networks. Furthermore, we also propose a Bayesian rule-based preference weight updating mechanism, multi-objective advantage function, and balanced-reward aggregation method. These proposed methods ensure effective training and control biases toward any specific objective reward during the training process of our model.

Based on extensive simulations, we demonstrate that the proposed model-based power allocation method outperforms the other techniques, especially Pareto front approximation policy-driven multi-objective reinforcement learning-based power allocation strategies.

조선대학교
CHOSUN UNIVERSITY

국문요약

Massive MIMO 네트워크를 위한 다중목표 강화학습 기반 전력 할당: 스펙트럼 및 에너지 Trade-Off 최적화에 대한 연구

오영우
지도교수 : 최우열
컴퓨터공학과,
조선대학교 대학원

5G 네트워크와 차세대 이동통신 기술의 발전에 따라 스펙트럼 및 에너지 효율성의 공동 최적화 기술은 핵심적인 연구 주제로 자리 잡아 왔다. 그러나, 스펙트럼과 에너지 효율성 사이의 복잡한 trade-off 관계로 인해 genetic 알고리즘 및 convex optimization solver 기반의 전통적인 최적화 솔루션은 높은 계산 복잡도가 요구된다. 이러한 계산 복잡도 문제를 해결하기 위한 방안으로, 다양한 강화학습 알고리즘 기반의 솔루션 연구가 활발히 수행되고 있으나, 이를 활용한 무선 네트워크의 다중목표를 공동으로 최적화하기 위한 연구는 크게 활성화되지 않았다.

따라서, 본 학위 논문에서는 downlink multi-cell massive MIMO 네트워크 시나리오에서의 기존 솔루션에서 야기되는 계산 복잡도를 해결함과 동시에 스펙트럼 효율성과 에너지 효율성 사이에 발생하는 trade-off를 효과적으로 최적화하기 위한 multi-objective asynchronous advantage single actor – multiple critics (MO-A3Cs) 모델 기반 하향링크 전송 전력 할당 기법을 제안한다. 제안하는 모델은 multi-objective Markov decision process에 의해 모델링 되며, 기존의 단일 스칼라 형태의 보상은

조선대학교 CHOSUN UNIVERSITY

보상 벡터로 확장된다. 더불어, 본 논문에서는 학습 과정에서의 스펙트럼 및 에너지 효율성 사이의 특정 목표로 수렴되는 것을 방지함과 동시에 스펙트럼 및 에너지 효율성을 공동으로 최적화하기 위한 행동 정책을 학습시키기 위해 Bayesian rule 기반 선호도 가중치 갱신 기법, multi-objective advantage function 및 balanced-reward aggregation 기법을 소개한다.

실험 결과를 통해, 제안하는 MO-A3Cs 모델은 학습 과정에서의 기존의 강화학습 모델에서 야기되는 특정 목표에 대한 편향 없이 trade-off 최적화를 위한 효율적인 학습이 가능한 것을 확인할 수 있으며, Pareto front approximation policy 기반의 다중목표 강화학습의 대표적인 알고리즘을 포함한 여러 하향링크 전송 전력 할당 기법을 능가하는 공동 최적화 성능을 달성할 수 있음을 확인하였다.

조선대학교 CHOSUN UNIVERSITY

Acknowledgments

19년 9월에 시작된 저의 학부 연구생 활동이 22년 3월에 시작된 석사 과정을 거쳐, 어느덧 석사 학위 논문을 준비하는 단계에 이르렀습니다. 대학원에 입학했을 당시 2년이라는 시간이 멀게만 느껴져 때로는 안일하게 생각했던 제 모습이 후회됩니다. 스마트 네트워킹 연구실의 문을 처음 열었을 때, 주말과 낮·밤의 구분 없이 최선을 다하기로 다짐했지만, 항상 그렇게 했다고 말할 수 없어 아쉬움이 남습니다. 이렇게 부족한 저에게 많은 도움과 격려를 해주셨던 많은 분께 서면으로나마 감사의 마음을 전하고자 합니다.

우선 학부 연구생과 석사 과정을 포함한 4년가량의 여정 간 저를 지도해주신 최우열 지도 교수님께 진심 어린 존경과 감사를 표합니다. 교수님께서 베풀어주신 아낌없는 믿음과 배려가 있었기에, 인격적으로 한층 더 성장할 수 있었습니다. 또한, 교수님께서 바쁘신 일정 중에도 새벽과 주말을 마다하지 않고 저를 지도해주신 덕분에 힘든 상황에서도 버틸 수 있는 큰 원동력이 될 수 있었습니다. 졸업 앞둔 지금까지도 교수님의 품을 벗어나 사회로 진출하는 것이 가끔은 무섭게 느껴지지만, 교수님과 함께하며 '나 스스로 더 단단해지는 법'을 배웠기에 교수님과 함께한 시간을 생각하며, 다시 한번 큰 도약을 준비할 수 있을 것 같습니다. 다시 한번, 진심 어린 감사의 말씀 드립니다. 그리고 바쁘신 일정 속에서도 저의 석사 학위 논문을 심사해주시고, 조언을 아끼지 않으신 신석주 교수님, 강문수 교수님께 머리 숙여 감사 인사 올립니다.

조선대학교
CHOSUN UNIVERSITY

제 삶의 유일한 쉼터이자, 활력인 우리 가족들에게도 감사의 마음을 전하고 싶습니다. 하늘의 별이 되어 우리 가족을 항상 지켜주시는 할아버지께서 지어주신 '빛날 영(煐)', '집 우(宇)'의 뜻대로 우리 집을 빛내는 사람이 되도록 항상 최선을 다하여, 자랑스러운 손자가 되겠습니다. 표현이 서툴러 항상 가벼운 인사만 올렸던 과거의 모습이 너무 후회되고, 지금이라도 감사하고, 사랑한다는 말씀을 드리고 싶습니다. "내가 배운 공부는 도둑질 못 한다."라는 말씀을 항상 해주신 할머니, 지금은 그 말씀이 무엇을 의미하는지 조금이나마 이해할 수 있을 것 같습니다. 가까운 곳에 계시지만 자주 찾아뵙지 못한 것 같습니다. 항상 건강 주의하시고, 진심으로 사랑합니다. 저를 위해 궂은일을 마다하지 않으시며, 지금까지 보살펴주신 아버지와 어머니, 표현이 서툴러 사랑한다는 말도 제대로 못 드렸던 것 같습니다. 앞으로 더 훌륭한 아들이 되어 대가 없이 주신 아낌없는 사랑에 보답하겠습니다. 사랑합니다. 그리고 가장 믿고 의지할 수 있는 형에게도 감사의 말을 전하고 싶습니다. 어렸을 때부터, 바쁘신 부모님을 대신하여 항상 저를 보살펴주고, 지금까지도 진심 어린 조언과 격려를 해준 덕에 매 힘든 순간을 잘 이겨낼 수 있었습니다. 감사합니다.

제가 4년가량 보내왔던, 우리 스마트 네트워킹 연구실의 모든 구성원에게도 고맙다는 말을 전하고 싶습니다. 저에게 영어는 자신감이라며, 하고 싶은 말이 있다면 뭐든 뱉고 보라고 독려해준 Sifat, Rodoshi에게 고마움을 전합니다. 지금은 각자 독일과 미국에서 박사 과정에 진학 중이지만, 그 누구보다 훌륭한 연구원이 될 것이라 믿습니다. 함께 강화학습과 무선 네트워크를 연구하면서, 서로 부족한 부분에 대해 의논하며 웃기도 많이 웃었던 Faisal과 Pulok에게도 고맙고, 그립다는 말을 전합니다. 그리고 항상 힘든 순간 진심 어린 조언과 도움을 주며, 저와 함께 연구실에 가장

조선대학교
CHOSUN UNIVERSITY

늦게까지 남아줬던 Islam에게도 감사를 드립니다. 해외 저널 작업과 관련하여, 질문하면 항상 적극적으로 알려주셨던 Arif, Ganiga 박사님들께도 감사드립니다. 언어와 문화적 장벽이 있었음에도, 저의 부족한 부분을 채워주기 위해 노력하고, 배려하는 우리 연구실 구성원들의 마음을 항상 감사히 하며, 평생 잊지 않겠습니다.

대학원 생활까지 함께할지 상상도 못 했지만, 고교 시절부터 석사 과정까지 11년을 같이 보내온 강민이, 힘들 때면 복도에서 함께 이야기 나눴던 유일한 입학 동기 영서, 과거의 저를 보는 것 같아 항상 걱정되는 연구실 막내 정태, 그리고, 항상 먼저 밝게 인사해주는 멀티미디어컴퓨팅 연구실 동생들에게도 고맙다는 말을 전하고 싶습니다. 각자가 원하는 목표를 위해, 묵묵히 걸어가는, 걸어왔던 우리들의 노력이 밝은 미래가 될 수 있도록 언제 어디서든 응원하겠습니다.

끝으로, 대학원 생활 동안 아낌없는 조언과 도움을 주신 모든 분께 다시 한번 고개 숙여 감사 인사 올립니다.

조선대학교
CHOSUN UNIVERSITY

Chapter Ⅰ

Introduction

A. Research Background

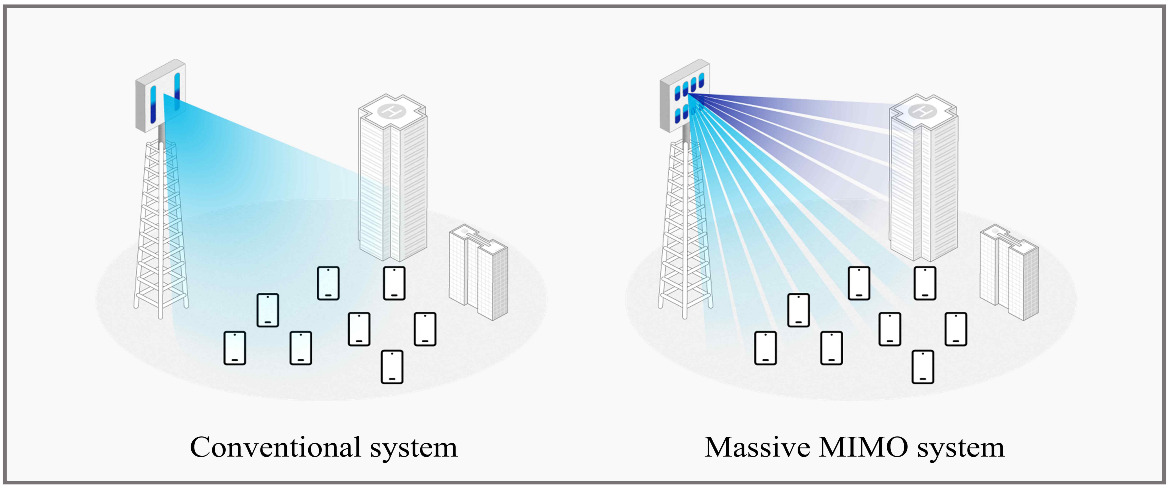


Figure 1.1: Illustrative example of the conventional and massive MIMO systems.

Massive multiple-input multiple-output (MIMO) is one of the key technologies of fifth-generation (5G) and beyond networks and is capable of enhancing the spectral efficiency (SE) and cell coverage by utilizing multi-antenna transmissions at the base station (BS) to simultaneously serve multiple user equipment (UE) [1], [2], as shown in Fig. 1.1. The impact of fading and interference in massive MIMO can be reduced through spatial diversity and multiplexing gain. Moreover, the link reliability and transmission rate are improved by leveraging the spatial domain to precisely focus energy toward the intended UE.

조선대학교
CHOSUN UNIVERSITY

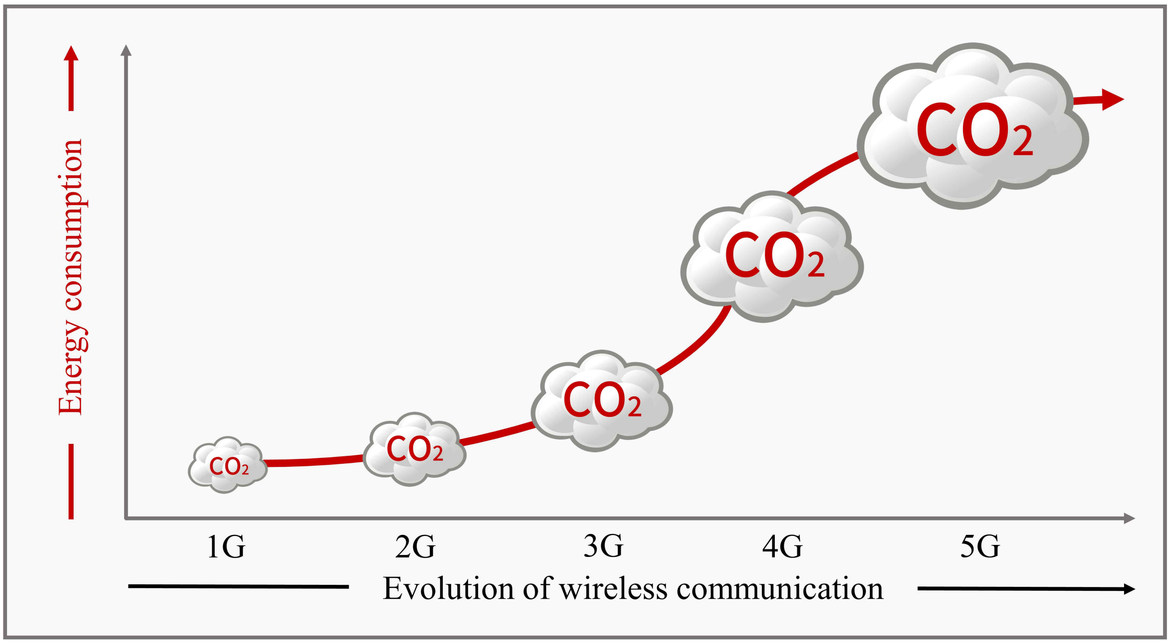


Figure 1.2: Illustration of increased energy consumption according to the evolving wireless networks.

However, deploying a large number of antennas requires high transmit power, leads to high interference, degrades the overall network performance, and significantly increases overall network energy consumption. For this reason, energy consumption increases as network technology evolves to fulfill the required wireless traffic, as shown in Fig. 1.2. As a key technology, multi-user MIMO (MU-MIMO) systems utilize the same frequency resources to serve multiple users at the same time, which leads to more efficient use of scarce spectrum resources and provides more tolerance to propagation losses compared to single-user MIMO (SU-MIMO) systems. A remarkable advancement has been made to improve the performance of downlink MU-MIMO systems, primarily focusing on tackling the high energy consumption

조선대학교
CHOSUN UNIVERSITY

of the cellular network by achieving a reasonable trade-off between SE and energy efficiency (EE) [3], [4]. However, resource allocation techniques seriously influence the overall performance of MU-MIMO systems and play a vital role in harnessing the full potential of massive MIMO systems. To ensure that the available power resources are efficiently utilized, and the quality-of-service (QoS) requirements of the UEs are fulfilled, an efficient power allocation (PA) is challenging.

Furthermore, in varying dynamic environments with the dense deployment of network nodes and UEs, the PA becomes more crucial to managing inter-cell and intra-cell interference and ensuring equitable service to all UEs in the network. Given PA complexities in massive MIMO, innovative solutions are essential. Traditional techniques have limitations, and there is a growing need for more advanced methods. As conventional methods face scalability issues, alternative approaches now regard multiple objective optimization (MOO) problems.

The MOO has recently attracted the interest of researchers to simultaneously optimize different objective functions in 5G and beyond networks. Conventional genetic algorithms and convex optimization techniques are primarily used to solve MOO-based PA problems. However, the computational complexity of these conventional PA techniques increases exponentially with the number of antennas in massive MIMO systems [5], [6]. In this regard, deep learning (DL) based PA schemes are proposed that can achieve near-optimal performance while addressing the computational complexity issues inherited by the iterative algorithm-based PA

조선대학교
CHOSUN UNIVERSITY

techniques [7]-[9]. However, DL-based approaches face challenges, especially in dynamically changing wireless network scenarios, requiring additional training datasets.

Deep reinforcement learning (DRL) is an emerging technique that employs the Markov decision process (MDP) framework to optimize network objectives. Through a trial-and-error strategy, DRL algorithms utilize interactions between agents and wireless network environments to determine optimal policies for solving optimization problems. DRL has the potential to effectively deal with computationally complex optimization problems in dynamic wireless networks [10]-[16]. Despite the escalating importance of addressing the joint optimization of SE and EE in the 5G and next-generation networks, the DRL-based transmit PA techniques are overlooked in the context of multi-cell massive MIMO systems. In contrast to the conventional DRL algorithm, an efficient MORL framework is needed to effectively train and solve multi-objective resource allocation problems in massive MIMO networks.

Therefore, This work addresses the PA technique and joint optimization of SE and EE in the downlink multi-cell massive MIMO networks by proposing multi-objective asynchronous advantage single actor-multiple critics (MO-A3Cs) architecture.

조선대학교
CHOSUN UNIVERSITY

B. Contributions

The key contributions are summarized as follows:

- We propose a PA technique based on the novel MORL algorithm for the downlink multi-cell massive MIMO networks. The proposed MO-A3Cs algorithm utilizes MORL to optimize a trade-off between SE and EE in a massive MIMO network. the MO-A3Cs model follow Bayesian rule-based preference weights updating, the multi-objective advantage function, and the balanced-reward aggregation methods to solve the trade-off problem effectively, The proposed PA technique optimally allocates the transmission power in a massive MIMO network while ensuring an overall SE and EE balanced increase.

- We define a multi-objective MDP (MOMDP) for the proposed MO-A3Cs model comprising the state space, action space, and the extended reward vector. In addition, We provide the proposed MO-A3Cs model-based downlink transmit PA strategies in multi-cell massive MIMO networks. This procedure offers insights into the MORL algorithm for optimizing trade-offs, a critical aspect of 5G networks and next-generation wireless communications.

조선대학교
CHOSUN UNIVERSITY

- Extensive simulations are conducted to analyze the performance of the proposed MO-A3Cs for downlink PA in multi-cell massive MIMO networks. Compared with other benchmark schemes, the proposed MO-A3Cs provide better performance regarding average SE and power consumption in the massive MIMO networks. Furthermore, the simulation results depict the effectiveness of the proposed MO-A3Cs in achieving a joint-optimized SE and EE.

C. Thesis Organization

The rest of the thesis organized as follows. Chapter Ⅱ presents the system model for the downlink multi-cell massive MIMO networks. Chapter Ⅲ presents the background and problem formulation, while Chapter Ⅳ presents the proposed MO-A3Cs model for downlink PA in multi-cell massive MIMO networks. The simulation setup and the detailed discussion related to simulation results are presented in Chapter Ⅴ. Finally, the paper is summarized and concluded in Chapter Ⅵ.

조선대학교 CHOSUN UNIVERSITY

Chapter Ⅱ

Related Work

In this chapter, we introduce and analyze the various DRL, MARL, and MORL approach-based PA studies in cellular networks.

A. Deep Reinforcement Learning Approach

The authors in [10] proposed a deep Q-network (DQN)-based PA method to enhance the sum rate in multi-cell networks. This approach maximizes the sum rate and is used as the reward. The states considered for the actions selected by the DQN agents include normalized interference, downlink rate, and transmit power. Similarly, the authors in [11] defined the MDP for sum rate maximization considering the previous transmission power and channel gain as a state. However, this approach results in a high dimensional problem [17]. To deal with this, the actor-critic (A2C) algorithm is utilized in [12] to reduce the complexity of the action space in the DQN-based PA methods. Similarly, the authors in [13] consider continuous action space for the downlink max-min power control problem in cell-free (CF) massive MIMO systems and propose a deep deterministic policy gradient (DDPG) method.

Furthermore, the objective function is maximized considering max-min fairness

조선대학교
CHOSUN UNIVERSITY

[18] and the maximum product signal-to-interference-plus-noise ratio (SINR) [19] methods. However, single agent-based PA strategies in DRL algorithms require extensive training to determine optimal policies in case of optimization in complex environments.

B. Multi-Agent Reinforcement Learning Approach

To deal with training overhead in single agents-based DRL technique for PA in dynamic wireless networks, the multi-agent reinforcement learning (MARL) approach was adopted with enhanced training strategy, scalable distributed learning, and execution in [14], [15]. The authors in [14] introduce a multi-agent DQN-based PA technique to maximize the sum rate in multi-cell networks. The sum rate is maximized using local agents with uniform target parameters while the global network updates the replay buffer gathered by these local agents. Furthermore, the authors demonstrate that the multi-agent DQN outperforms the single DQN in model training efficiency. Similarly, a multi-agent double DQN (DDQN)-based PA framework is proposed in [15] to maximize the capacity in multi-cell massive MIMO networks. The multi-agent DDQN model is split into sub-networks, i.e., the target Q-network and the evaluation Q-network, to avoid overestimating the Q-value in the DQN model. It is concluded that the proposed multi-agent DDQN provides improved convergence stability compared to the conventional DQN approach.

조선대학교 CHOSUN UNIVERSITY

C. Multi-Objective Reinforcement Learning Approach

Recently, the emerging MORL algorithm has been used to solve MOO problems in the CF massive MIMO networks. The authors in [16] use a reward vector, defined as the sum rate and user fairness. In addition, to solve the MOO problem by transforming the problem into a single objective optimization (SOO). Moreover, the twin-delayed DDPG (TD3) algorithm with a replay buffer effectively maximizes the sum rate and fairness. These replay buffer-based training strategies can enhance sampling diversity and efficiency in massive MIMO networks.

However, the model does not undergo training through real-time interactions between the network and the agent in buffer memory-based training strategies. Instead, it relies on old data saved in the buffer with limited memory size and uses it for future training. Furthermore, instead of using weight adjustment among multiple objectives such as sum rate and fairness, interpolation preference weights are considered, which are scenario-limited. These training strategies can lead to sub-optimal policies in the case of massive MIMO networks.

To solve this problem, it is crucial to develop an advanced MORL algorithm designed to optimize transmit PA, thereby enhancing the overall SE and EE in massive MIMO networks. This paper introduces a novel MORL algorithm that leverages a MARL training strategy for efficient MORL model training. This approach enables interactions between each local agent and independent environments, leading

조선대학교
CHOSUN UNIVERSITY

to the acquisition of diverse and immediate experience-data to training jointly optimization policy. Furthermore, we also propose and implement a Bayesian rule-based preference weight updating mechanism that dynamically adjusts the weightings of multi-objectives, including SE and EE, informed by the trajectories collected from each local agent. These innovations ensure that our proposed MORL algorithm not only trains from a diversity of experience-data but also improves both SE and EE in downlink multi-cell massive MIMO networks.

조선대학교
CHOSUN UNIVERSITY

Chapter Ⅲ

Downlink Multi-Cell Massive MIMO Network

In this chapter, we present the network layout followed by the main system assumptions, SINR and SE, the network power consumption model, and an overview of the joint spectral-energy optimization problem.

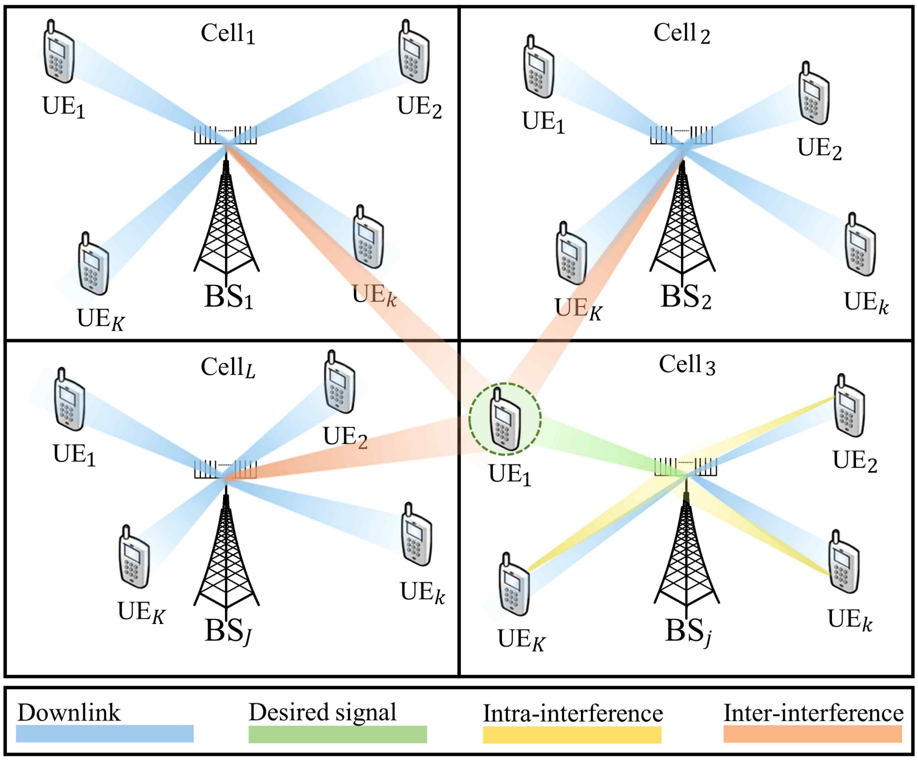


Figure 3.1: Illustration of the downlink multi-cell massive MIMO networks.

조선대학교
CHOSUN UNIVERSITY

A. Channel Estimation and Spectral Efficiency

A downlink multi-cell massive MIMO network is considered with L number of cells as shown in Fig. 3.1. The BS is deployed at the center of each cell l where j-th BS in the cellular network is equipped with M number of antennas. The UEs are assumed to be located randomly in the l-th cell. Furthermore, we assume that each BS simultaneously serves a K number of UEs by sharing the same frequency band.

The channel matrices between the j-th BS and k-th UE located in l-th cell is denoted by $h_{j,k}^{l} \in C^{M}$ and can be expressed as

$$h_{j,k}^{l} \sim N_{C}\left(0, R_{j,k}^{l}\right), \tag{3.1}$$

where C^{M} and $R_{j,k}^{l} \in C^{M \times M}$ denote the complex-valued vector space of dimension M and the spatial correlation matrix, respectively. Furthermore, we assume the BSs and UEs are perfectly synchronized and operate under the time division duplex (TDD) protocol. Before performing downlink transmission BS, each user transmits the pilot signal in the uplink to estimate the channel at the BS. The UEs reuse the pilot signal in the cell, and the reuse factor $\tau_{p} = K$ is employed to reduce interference in the adjacent cells [20].

Based on this assumption, we utilize the minimum mean-square error (MMSE) estimation method at the BS to effectively estimate the imperfect channel condition

조선대학교 CHOSUN UNIVERSITY

corrupted by the interference and noise in the network [21]. The estimated channel between the j-th BS and k-th UE. computed from the uplink pilot signal ρ^{UL}, is denoted by $\hat{h}_{j,k}^{l}$. The MMSE-based estimated channel is given by

$$\hat{h}_{j,k}^{l}=R_{j,k}^{l}\,Q_{j,k}^{-1}\left(\sum_{l\neq j}^{L}h_{j,k}^{l}+\frac{1}{\tau_p}\frac{\sigma^2}{\rho^{UL}}n_{j,k}\right), \tag{3.2}$$

where $Q_{j,k}=\sum_{l\neq j}^{L}R_{j,k}^{l}+\frac{1}{\rho^{UL}}I_M$, I_M denotes the identity matrix, and σ^2 is the noise variance. The noise added by the system is represented as $\frac{1}{\tau_p}\frac{\sigma^2}{\rho^{UL}}n_{j,k}$. Based on the MMSE technique, the channel estimation is performed by minimizing the estimation error between the actual and estimated channels and is expressed as $e_{j,k}^{l}=h_{j,k}^{l}-\hat{h}_{j,k}^{l}$.

The downlink signal received at k-th UE contains the desired signal transmitted from the j-th BS, inter-cell and intra-cell interference, and the system-added noise. The downlink signal received at the k-th UE from the j-th BS located in the l-th cell can be expressed as

$$y_{j,k}=z_{j,k}s_{j,k}+\sum_{l\neq j}^{L}\sum_{i=1}^{K}z_{l,i}s_{l,i}+\sum_{l=j}^{L}\sum_{i\neq k}^{K}z_{j,i}s_{j,i}+n_{j,k}, \tag{3.3}$$

where $s_{j,k}$ denote the transmitted signal from the j-th BS to each k-th UE, $z_{j,k}$ denote the regularized zero-forcing (RZF) precoding vector [22]. and $s_{j,k}z_{j,k}$ represents the actual transmitted downlink signal to k-th UE.

조선대학교
CHOSUN UNIVERSITY

The received SINR at the k-th UE from the j-th BS is written as

$$\lambda_{j,k}=\frac{p_{j,k}\alpha_{j,k}}{\sum_{l=1}^{L}\sum_{i=1}^{K}p_{l,i}\beta_{l,i}+\sigma^2}\ , \tag{3.4}$$

where $p_{j,k}$, $\alpha_{j,k}$, and $\beta_{j,k}$ denote the downlink transmit power, the channel gain between the j-th BS and the k-th UE, and the interference signal power received at the k-th user from the l-th BS [10].

According to Shannon's theorem, the channel capacity is defined as the maximum amount of information that can be transferred over a channel [7]. The achievable channel capacity of the established link between the k-th UE and the j-th BS is expressed as

$$C_{j,k}=\frac{\tau_d}{\tau_c}\log_2\left(1+\lambda_{j,k}\right), \tag{3.5}$$

where τ_d and τ_c represent the number of samples used for downlink data transmission and per coherence block, respectively. The downlink SE is defined as the total achievable data rate over the available bandwidth in massive MIMO networks and is measured in bits per second per Hertz (b/sec/Hertz). Based on the received SINR in (3.4) and the achievable channel capacity in (3.5), the total achievable SE in multi-cell massive MIMO networks can be formulated as [8]

$$SE_{DL}=\sum_{j=1}^{L}\sum_{k=1}^{K}C_{j,k}. \tag{3.6}$$

조선대학교
CHOSUN UNIVERSITY

B. Power Consumption Model and Energy Efficiency

The total power consumption in the downlink multi-cell massive MIMO networks is the sum of the effective transmit power $p_{j,k}$ allocated based on the PA technique and the circuit power consumption P_{CR}. The total consumed power can be mathematically expressed as

$$P_{total} = \sum_{j=1}^{L} \sum_{k=1}^{K} p_{j,k} + \sum_{j=1}^{L} P_{CR}. \tag{3.7}$$

The circuit power consumption of each j-th BS in the massive MIMO network comprises the constant power consumed at BS denoted by P_{FIX} and the constant power incurred during the signal processing denoted by P_{SP}.

Therefore, the total circuit power consumption of a BS can be expressed as

$$P_{CR} = P_{CH} + P_{CE} + P_{BH} + P_{ED} + P_{FIX} + P_{SP}. \tag{3.8}$$

A large fraction of the power consumed in the network comprises the power consumed at the BS [23]. The power consumption of the BS comprises circuit powers required in operations such as the number of transmit antennas, channel estimation, and encoding and decoding [24].

In particular, the circuit power due to the transceiver chain, which accounts for the most power consumption, includes components such as filters, mixers, digital-to-analog converters (DAC), and analog-to-digital converters (ADC). The power consumption of the transceiver chain component can be written as

조선대학교
CHOSUN UNIVERSITY

$$P_{CH} = Mp_{BS} + P_{LO} + Kp_{UE}, \tag{3.9}$$

where p_{BS}, p_{LO}, and p_{UE} denote the transmission power of a single BS antenna, the local oscillator (LO), and the circuit power coefficient of the UE, respectively. From (3.9), the power consumption of the BS is proportional to the number of antennas. Furthermore, the power consumed during the channel estimation at the BS for each coherent block is also taken into consideration [2]. The power consumption in terms of the channel estimation can be calculated as

$$P_{CE} = \frac{3B}{\tau_c L_{BS}} K \times M\tau_p + M^2, \tag{3.10}$$

where B and L_{BS} denote the bandwidth and the computational efficiency of the BS, respectively [25]. The circuit power consumed in the backhaul during the uplink and downlink data transmission can be expressed as

$$P_{BH} = p_{BT} TP, \tag{3.11}$$

where p_{BT} denotes the backhaul traffic power and TP represents the achievable throughput within a cell. The value of TP is calculated as $B\sum_{k=1}^{K} C_{j,k}$. Similarly, the circuit power consumed in channel encoding and decoding is denoted by P_{ED} and is given by

$$P_{ED} = (p_{ENC} + p_{DEC}) TP, \tag{3.12}$$

where p_{ENC} and p_{DEC} represent the power consumption coefficients incurred during

조선대학교 CHOSUN UNIVERSITY

the encoding and decoding processes, respectively. Therefore, The EE of the downlink massive MIMO network is the ratio of SE to the total power consumption and can be formulated as

$$EE_{DL} = \frac{SE_{DL}}{P_{total}}. \tag{3.13}$$

C. Definition of Joint Optimization Problem

To evaluate the joint optimization in the multi-cell massive MIMO networks, simultaneously SE and EE must be optimized. Let us define a joint objective function of SE and EE by $f(SE_{DL}, EE_{DL})$. Thus, the joint optimization problem can be formulated as [4]

$$\begin{aligned} &\max_{p_{j,k}} f\left(SE_{DL}, EE_{DL}\right) \\ &s.t.\ \ 0 \leq p_{j,k} \leq P_{\max}, \forall j,k, \end{aligned} \tag{3.14}$$

where $P_{\max}$ denote the maximum transmit power. The transmit power $p_{j,k}$ that affects both SE and EE is defined as a constraint and is required in the joint optimization problem [26]. The joint optimization problem in (3.14) is classified as multi-objective non-convex and NP-hard and requires high computations [10], [15]. Therefore, in this thesis, we convert the SOO problem to the MOO problem through MOMDP to solve the optimization problem. In addition, we propose a PA technique based on the MO-A3Cs model to effectively solve the converted SOO problem.

조선대학교
CHOSUN UNIVERSITY

Chapter Ⅳ

MORL Algorithms and Problem Conversion

In this chapter, we first briefly present the background of the MORL techniques for MOO problems. Then, the detailed description of the MOMDP is presented to transform the MOO problem into the SOO problem, which is later utilized in allocating power using the proposed MO-A3Cs model.

A. MORL Technique for MOO Problem

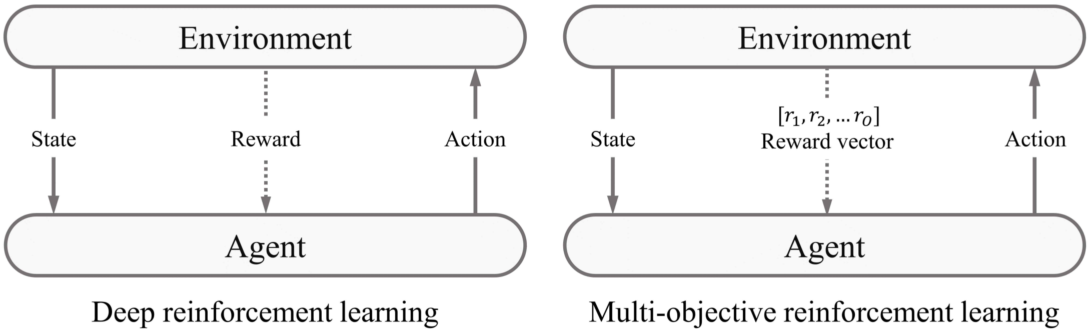


Figure 4.1: Comparison of DRL and MORL algorithms: Interaction between the agent and environment.

Numerous studies have adopted and validated the DRL algorithms for PA to achieve enhanced performance in wireless networks. However, it is challenging to achieve better performance complexity trade-offs in the emerging MORL algorithm [27]. Compared to DRL, the MORL algorithms utilize multiple rewards in the form of

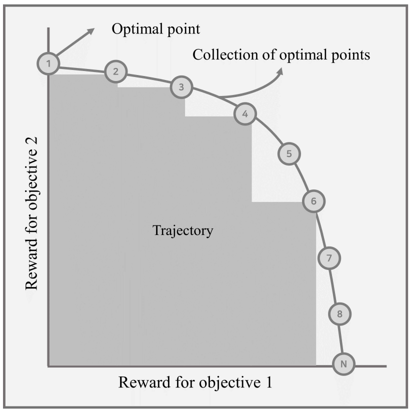
조선대학교 CHOSUN UNIVERSITY

reward vectors to maximize multiple objectives, as shown in Fig. 4.1. To effectively tackle these reward vectors, the MORL algorithm either uses Pareto front approximation (PFA) [28], [29] or uses MOO transformation to SOO problems [27], [30].

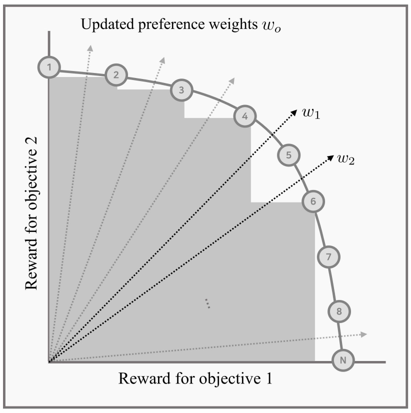
The PFA approach utilizes the reward vectors of the selected actions to determine the optimal point among multiple objectives based on Pareto dominance and Pareto fronts. The collected data from agent-environment interactions are used to construct a Pareto set and derive the required solution for the MOO problem. This approach requires a considerable buffer memory used to generate the Pareto set. Moreover, the Pareto fronts [31] used to determine the optimal solution require significant training time in large-scale environments such as massive MIMO systems [32].

On the other hand, the transformation approaches that transform MOO problems into SOO problems employ strategies such as linear weighted sums [27] and constraints [30] and objective-preference concepts. Fig. 4.2 illustrates the process of determining the optimal policy for solving MOO problems using a MORL algorithm that uses preference weights to determine the optimal solution. In the initial stage of the MORL algorithm, a hypervolume is generated between multiple objectives through interactions between the agent and the environment, as shown in Fig. 4.2(a) In addition, Fig. 4.2(b) and Fig. 4.2(c) present the graphical illustration of weight assignment and the selection of optimal point for multiple objectives by using

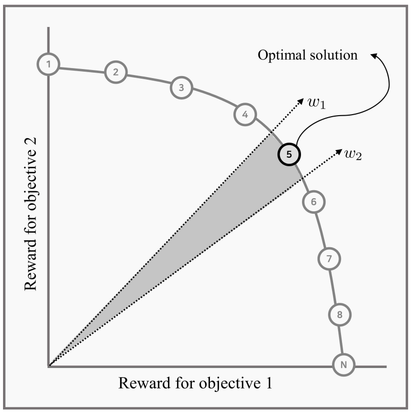
조선대학교 CHOSUN UNIVERSITY



(a) Initial steps of the MORL model training process



(b) Adjusting step of the policy by using preference weights



(c) Policy convergence step

Figure 4.2: Illustration of policy convergence in the MORL algorithms driven by preference weights for solving the MOO problem.

조선대학교
CHOSUN UNIVERSITY

the relative priorities ω_1 and ω_2. This transformation strategy may achieve faster convergence compared to the PFA approach. However, it can lead to limited training efficiency due to a bias towards specific objectives and the potential for converging to sub-optimal solutions depending on the preference weight settings [33].

To this end, various methods have been suggested to effectively determine preference weights for MORL. These methods include the utilization of uniform weights [34], random weights [35], and dynamic weights [36]. The uniform and random preference weight approaches have limitations in that the convergence of the MORL model must be verified through various experiments to train the optimal points for multiple objectives. On the other hand, the dynamic weight approach allows for dynamically determined weights to optimize the policy designed to solve the MOO problem. However, this method requires additional buffer memory to update the weights for each objective.

Therefore, we propose a novel MORL algorithm to solve multi-objective functions and trade-off problems between SE and EE. The proposed model employs a MOMDP framework and integrates a Bayesian rule-based technique for updating preference weights, a multi-objective advantage function, and a balanced reward aggregation method. The multi-objective advantage function allows for the individual evaluation of action value for each objective. Moreover, the balanced-reward aggregation method aggregates rewards considering each preference weight, ensuring a more efficient approach to action selection by the agents.

조선대학교
CHOSUN UNIVERSITY

B. MOMDP-Based Conversion of MOO to SOO

The MOMDP is an extension of the MDP and deals with multiple rewards in the form of a reward vector. In addition, the MOMDP can be defined as a tuple $\langle S, A, P, R, \gamma \rangle$, where S, A, $P(s'|s,a)$ denotes the state space, action space, and the transition probability of taking action a for state transitions from s to s', respectively. The reward vector R consists of the respective objective rewards for SE and EE. Thus, R can be expressed as $\{R_o \mid\sim \forall o \in \{1, 2 \cdots, O\}\}$, where O represents the total number of objectives. Similarly, the discount factor γ, which determines how much the agent considers long-term rewards, is defined as $\gamma \in (0, 1)$. Furthermore, we employ preference weights $\{\omega_o \mid\sim \forall o \in \{1, 2 \cdots, O\}\}$ indicate the relative priority of each objective [37], [38].

Therefore, the scalarized function for processing the reward vector in scalar form is defined as $f\omega_o(R_o) = \omega_o \times R_o$. The joint optimization problem in (3.14) is transformed and can be rewritten as

$$\begin{aligned} &\max_{\omega_1, \omega_2} f\omega_1(SE_{DL}) + f\omega_2(EE_{DL}) \\ &s.t.\ \omega_1 + \omega_2 = 1,\ \omega_1, \omega_2 \geq 0, \end{aligned} \tag{4.1}$$

where ω_1 and ω_2 denote the preference weights that indicate the relative priorities of SE and EE, respectively. The states s_t, which facilitate the observation of various features related to the problem in (4.1), can be defined as

조선대학교
CHOSUN UNIVERSITY

$$s_t = \left\{ \alpha_{j,k}^t, C_{j,k}^t, p_{j,k}^t \right\}, \quad \forall j,k, \tag{4.2}$$

where $\alpha_{j,k}^t$, $C_{j,k}^t$, and $p_{j,k}^t$ denote the channel gain, achievable channel capacity, and the transmit power between j-th BS and k-th UE at the time step t, respectively. These state s_t are utilized by an agent to efficiently observe the SE and EE while interacting with the downlink multi-cell massive MIMO network.

The action space A consists of feasible downlink transmission powers between all BSs and UEs. However, defining the action space as the set of all possible transmission powers in a multi-cell massive MIMO network, the dimensional issue arises [17]. To this end, we utilize a discretization strategy using quantization [39] of transmit power between $P_{\min}$ and $P_{\max}$ with a specific quantization level to select action a_t. The discretized action space based on quantization can be expressed as

$$A = \left\{ 0, P_{\min}, P_{\min} \left(\frac{P_{\max}}{P_{\min}} \right)^{\frac{1}{|Q|-2}}, \cdots, P_{\max} \right\}, \tag{4.3}$$

where $|Q|$ denotes the quantization level, which indicates the degree at which the transmission power range between $P_{\min}$ and $P_{\max}$ be divided into discrete values. This discretization approach allows an increase in the power at each $|Q|$ level and effectively generates a variety of power action space between $P_{\min}$ and $P_{\max}$.

Finally, the immediate reward vector obtained through the interaction between the agent and environment in a massive MIMO network at a certain time instant can be expressed as

조선대학교
CHOSUN UNIVERSITY

$$r_t = \left[\overline{SE}_{DL}(t), \overline{EE}_{DL}(t) \right], \tag{4.4}$$

where $\overline{SE}_{DL}(t)$, $\overline{EE}_{DL}(t)$ denote average SE and EE, respectively. In general, the total value of SE is higher than the achievable total EE, which can lead to convergence instability and extend to training duration, Thus, we used the average SE and EE to reduce the variability of rewards and ensure more smoother convergence during the model training process.

조선대학교
CHOSUN UNIVERSITY

Chapter Ⅴ

Proposed MO-A3Cs Technique for Power Allocation

In this chapter, we introduce the proposed MO-A3Cs model-based PA framework followed by a Bayesian rule-based preference update mechanism, multi-objective function with reward aggregation method, and optimization of each actor and critic network. The proposed MO-A3Cs model uses the multi-critic model to simultaneously consider multi-objectives and estimate the expected value for each objective. The single actor determines the optimal power value for different objectives by aggregating the predicted values from each critic model. The proposed downlink PA framework based on MO-A3Cs model, the utilized single-actor, and multi-critic neural networks is illustrated in Fig. 5.1 and Fig. 5.2, respectively.

Furthermore, the MO-A3Cs model integrates the extension of the asynchronous advantage actor-critic (A3C) [40] along with the proposed Bayesian rule-based preference weight update, multi-objective advantage function, and balanced-reward aggregation method. Our proposed approach is inspired by the A3C model, which consists of an actor-network that selects actions and a critic network that evaluates the chosen actions. The actor and critic networks interact with each other and decide whether to take a specific action a_t from the available action space A at a particular s_t. The critic evaluates the selected a_t by the actor using the value function $V_{\phi}(s_t)$.

조선대학교
CHOSUN UNIVERSITY

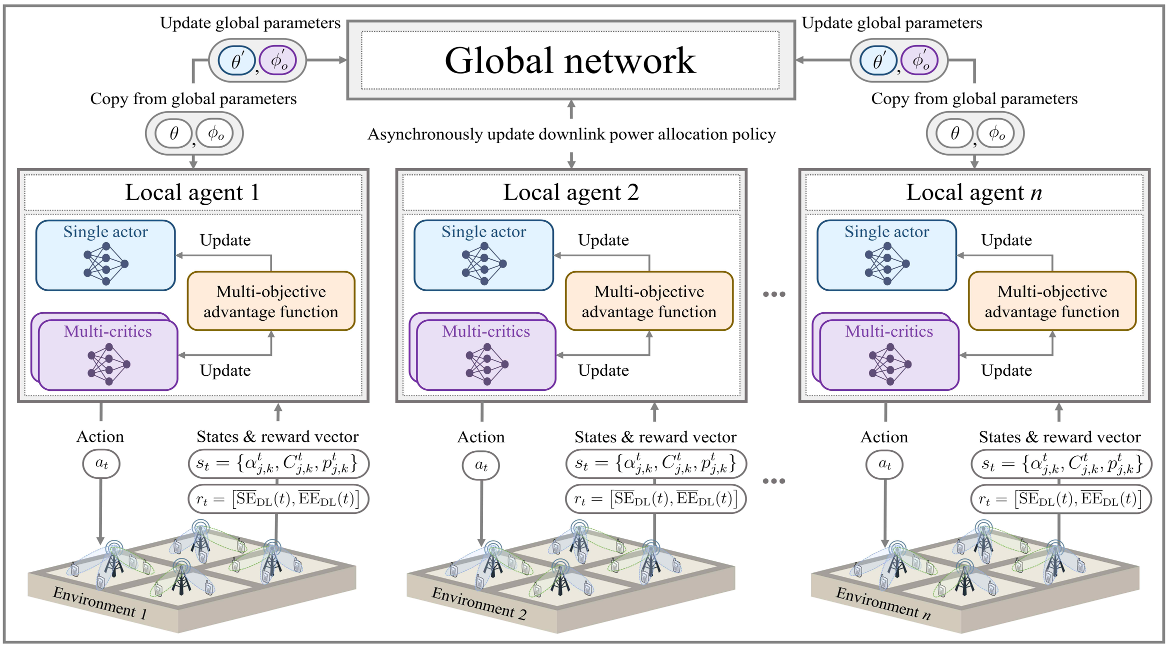


Figure 5.1: The proposed MO-A3Cs based transmit PA framework for SE-EE joint optimization in the downlink multi-cell massive MIMO networks.

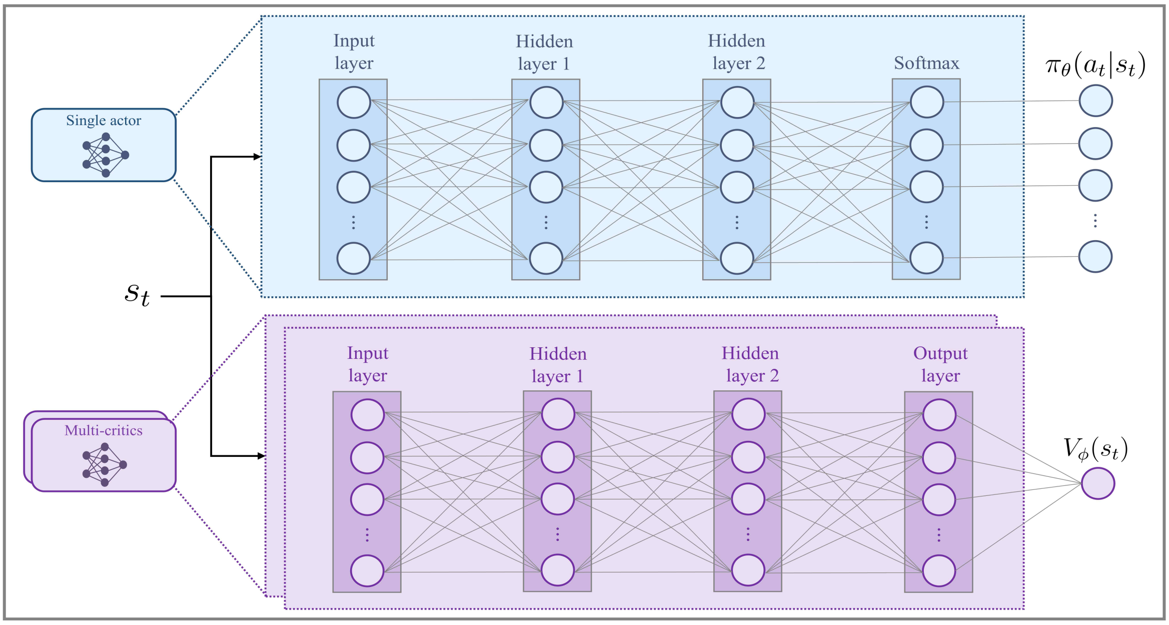


Figure 5.2: Structure of the neural network of the single actor and multi-critics network on the proposed MO-A3Cs model.

조선대학교 CHOSUN UNIVERSITY

The update process of the actor network is given as

$$\theta \leftarrow \theta + \eta \sum_{t=1}^{T} \left(R_t - V_\phi(s_t)\right) \times \nabla \log \pi_\theta(a_t|s_t), \tag{5.1}$$

where θ represents the policy parameters of the actor, η is the learning rate, t denotes the time step, and T indicates the maximum number of episode, while $R_t = \sum_{i=0}^{\infty} \gamma^i r_{t+i}$ represents the accumulated reward computed by the discount factor γ, $V_\phi(s_t)$ is the value function for state s_t, and $\nabla \log \pi_\theta(a_t|s_t)$ denotes the gradient of the actor network. The network aims to maximize the expected reward by utilizing the advantage function, which is based on the difference between R_t and $V_\phi(s_t)$ for a given state s_t.

For the critic, it aims to minimize the error between the accumulated reward R_t obtained from the actor and its predicted value $V_\phi(s_t)$. The critic evaluates the value of the selected action a_t based on

$$\phi \leftarrow \phi + \eta \frac{\partial \left(R_t - V_\phi(s_t)\right)^2}{\partial \phi}, \tag{5.2}$$

where ϕ denotes the parameters of the critic network and $\frac{\partial \left(R_t - V_\phi(s_t)\right)^2}{\partial \phi}$ represents the gradient of the squared error, respectively.

조선대학교
CHOSUN UNIVERSITY

A. Bayesian Rule-Based Preference Updates

The proposed preference weight update method utilizes trajectories collected through multiple local agents to update the preference weights ω_o for observed objective, i.e., SE and EE in a multi-cell massive MIMO network. The update strategy ensures diverse experiences for global network training without relying on potentially biased experiences from initially collected data and helps in the joint optimization of SE and EE. In the MORL algorithm, the preference weights follow uniform and random initialization [36]. In the case of uniform initialization, each objective is initialized with an equal weight given by

$$\omega_o = \frac{1}{O}, \forall o = 1, 2, \cdots, O. \tag{5.3}$$

Using this uniform method, SE and EE are initialized with the same preference weights despite the local agents interacting with distinct and independent multi-cell massive MIMO environments. Such an initialization implies that the agents must update it more frequently, and more training is required to find the joint optimization policy. To address this issue, we propose an adaptive update mechanism using Bayesian rules with random preference initialization. The preference weight is initialized from uniform distribution as $\omega_o \sim U(0,1)$. To quantify the relative priority of the objectives, the weights are normalized as follows.

조선대학교
CHOSUN UNIVERSITY

$$p(x_o|\psi_0) = \frac{w_o}{\sum_{o=1}^{O} w_o}, \tag{5.4}$$

where x_o represents the objective function for each SE and EE, and ψ_0 denotes the initial trajectory capturing the sequence of interactions between the agent and the environment. For the given state, action, and reward vector at time step t, $\psi_t = (s_1, a_1, r_1, \cdots, s_{t-1}, a_{t-1}, r_{t-1}, s_t, a_t, r_t)$. The normalization process facilitates the comparison of the relative importance of each objective. Later, each local agent utilizes the ω_o as prior probabilities for calculating the likelihood of objectives. The likelihood function for each objective can be expressed as

$$p(\psi_t|x_o) = \prod_{t=1}^{T} p(s_t, a_t|x_o), \tag{5.5}$$

where the likelihood $p(\psi_t|x_o)$ represents the probability of observing the trajectory ψ_t, which consists of states s_t, actions a_t, given the objective function x_o. This metric indicates how well the observed ψ_t aligns with x_o, and the preference weights update is made based on the posterior probability of the x_o using Bayesian rules. Therefore, the agent assesses whether the trajectory ψ_t of the current state s_t sufficiently aligns which objective function x_o. As the samples in the trajectory ψ_t change at step t, the agent adaptively adjusts the weight ω_o using the Bayesian rule and is given by [41]

조선대학교
CHOSUN UNIVERSITY

$$w_o = p(x_o|\psi_t) = \frac{p(\psi_t|x_o)p(x_o)}{\sum_{o=1}^{O} p(\psi_t|x_o)p(x_o)}, \tag{5.6}$$

where the prior probability and the likelihood function are given in (5.4) and (5.5), respectively.

The process described by (5.6) represents updating preference weights for the objective function based on trajectory ψ_t, and prior probabilities of previous preference weights. Therefore, the proposed preference weights updating techniques utilize the posterior probability of x_o, updated from the Bayesian rule, as its preference weight. By adopting this proposed updates strategy, the DRL agents can more efficiently observe the trade-off between SE and EE during the model training process. Moreover, the proposed preference weight update mechanism overcomes shortcomings of the conventional interpolation weights-based [16], [34] and buffer memory-based dynamic weights update approaches [36]. Next, we present the multi-objective advantage function for the proposed MO-A3Cs method.

B. Multi-Objective Advantage Function

The A3C model predicts the $V_\phi(s_t)$ from the current state s_t in the critic network and employs the advantage function to evaluate and update the actor-critic network. This function measures the difference between the R_t and $V_\phi(s_t)$ for a specific action

조선대학교 CHOSUN UNIVERSITY

a_t taken on the current state s_t. It is beneficial to evaluate the value of the chosen action a_t [40].

However, such an advantage function is optimized for training a single objective and ineffective for solving MOO problems. To solve the MOO problem, this work extends the single objective advantage function to a multi-objective advantage function. Let O be the number of objectives, then the multi-objective advantage function can be expressed as

$$G_o = R_t - V_\phi(s_t) = \sum_{i=t}^{\infty} \gamma^{i-t} \delta_i, \;\; o \in \{1, 2, \cdots, O\}, \tag{5.7}$$

where $\delta_t = r_{t+1} + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$ denotes the temporal difference error (TD error) [42]. Using this advantage function, the value of actions for SE and EE is independently assessed, simultaneously considering each objective. The actor update is performed using the proposed balanced-reward aggregation method utilizing this multi-objective advantage function.

In the MORL algorithm, reward aggregation represents the summation of multi-objective rewards based on the scalarized function $f\omega_o(R_o)$, considering their relative priorities [43]. Generally, in the MORL algorithm, where trade-offs are not considered, all the objective rewards in the reward vector are summed and employed as a single reward. This can be mathematically expressed as

$$\theta \leftarrow \theta + \eta \sum_{t=1}^{T} \left(\sum_{o=1}^{O} G_o \right) \times \nabla \log \pi_\theta(a_t | s_t). \tag{5.8}$$

조선대학교 CHOSUN UNIVERSITY

For SE and EE as objectives, the absence of preference weights ω_o makes it challenging to be used to solve the trade-off between them. To apply the preference weights ω_o, most MORL algorithms utilized the combined-reward aggregation [44] that reflects preference weights for each objective and can be expressed as

$$\theta \leftarrow \theta + \eta \sum_{t=1}^{T} \left(\sum_{o=1}^{O} \widetilde{G_o} \right) \times \nabla \log \pi_\theta(a_t | s_t), \tag{5.9}$$

where $\widetilde{G_o} = f\omega_o(R_t) - V_\phi(s_t)$ denotes the aggregation of the accumulated reward R_t considering ω_o for each objective using the scalarized function $f\omega_o$ which is applied in the multi-objective advantage function. Unlike (5.8), this approach allows for considering priorities for each objective, providing a more effective way to address SE-EE trade-off problems. Nevertheless, an issue with this method is that the ω_o for each objective is not reflected in the value function $V_\phi(s_t)$. This suggests that the agent might select actions biased towards a specific objective, not fully considering the preferences for both SE and EE, due to asymmetric updates to the value function $V_\phi(s_t)$ in each critic network. For instance, if the value function for SE is considered more important than that for the value function of EE, and without preference weights ω_o, the agent focuses on prioritizing actions that maximize SE, which may potentially lead to EE degradation.

To deal with this, this work considers a balanced-reward aggregation method that can be defined as

조선대학교 CHOSUN UNIVERSITY

$$\theta \leftarrow \theta + \eta \sum_{t=1}^{T} \left(\sum_{o=1}^{O} f_{\omega_o}(G_o) \right) \times \nabla \log \pi_\theta(a_t | s_t). \tag{5.10}$$

This method applies preference weights ω_o to both the cumulative reward R_t and the value function $V_\phi(s_t)$ for each objective. This approach leads to more efficient training to determine and improve the optimal PA policy for jointly optimizing the SE and EE.

C. Optimization of MO-A3Cs Updates

In this section, we introduce the update methods for the single-actor and multi-critic networks for the proposed model. The update for the single actor network, which determines the optimal action a_t, in the MO-A3Cs model, can be expressed as

$$L(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \left(\sum_{o=1}^{O} f \omega_o(G_o) \right) \times \log \pi_\theta(a_t | s_t), \tag{5.11}$$

where N denotes the number of trajectories ψ, i denotes the trajectory index ψ_t, and $\pi_\theta(a_t|s_t)$ is the probability of selected a_t for state s_t according to policy π_θ. The loss function, $L(\theta)$, for the single actor, computes the difference between the expected value function $V_\phi(s_t)$ and the actual reward R_t using multi-objective advantage function and the balanced-reward aggregation methods. Hence, the actor updates the action policy by considering the preference weights ω_o for each objective.

The structure of the proposed multi-critic networks is depicted in Fig. 5.1 and

조선대학교 CHOSUN UNIVERSITY

have an independent critic network for each SE and EE, which leads to a more accurate estimation of the value function $V_{\phi}(s_t)$ for each objective. The update and optimized multi-critics can be defined as

$$L(\phi_o)=\frac{1}{N}\sum_{i=1}^{N}\left(V_{\phi_o}(s_t)-R_t\right)^2, \tag{5.12}$$

where ϕ_o represents the parameters of the critic network for the o-th objective. Since each critic network is updated independently for each objective, updates can be made without influencing the value function estimation for other objectives. This extended multi-critic enables the estimation of the optimal value function $V_{\phi}(s_t)$ for each SE and EE and facilitates a more appropriate balance between SE and EE trade-offs.

Furthermore, this thesis paper introduces action distribution entropy to encourage agents to select and explore various actions in a multi-cell massive MIMO network environment. This entropy prevents premature convergence to sub-optimal solutions and enhances long-term convergence performance [45]. Thus, the action distribution entropy utilized in this paper can be expressed as

$$H(\pi_\theta)=-\sum_{A}\pi_\theta(a_t|s_t)\log\pi_\theta(a_t|s_t). \tag{5.13}$$

The larger entropy value $H(\pi_\theta)$ enables the agent to explore the environment and search in the expanded action space to collect diverse trajectories, leading to effective training of MO-A3Cs. Finally, the total loss function utilized for the MO-A3Cs model is given by

조선대학교 CHOSUN UNIVERSITY

$$L_{total} = L(\theta) + \sum_{o=1}^{O} L(\phi_o) + \mu H(\pi_\theta), \tag{5.14}$$

where $L(\theta)$ represents the loss function of the single actor, and $L(\phi_o)$ denotes the loss function of the critic network. Additionally, μ is a weight used for regularizing the action distribution entropy, and its value ranges between 0 and 1. The value of μ is set to 0.001. The total loss function is minimized for single actor and multi-critic networks.

The training procedure of the MO-A3Cs model for each thread is given in Algorithm 1. The algorithm initializes by synchronizing the key parameters between the global network and local threads followed by the preference weight initialization. For each time instant of the local thread, random preference weights are assigned to each local agent based on a distinct downlink multi-cell massive MIMO environment. The collected trajectories from each agent are leveraged and the global network is updated asynchronously. This training strategy benefits from the independent evaluation of SE and EE by the multi-critic network, distinct from existing MORL algorithms. This evaluation directs the joint optimization policy updates through the proposed balanced-reward aggregation function. Ultimately, by integrating a MARL-based training strategy and the proposed innovative MORL algorithm.

조선대학교
CHOSUN UNIVERSITY

Algorithm 1 The proposed MO-A3Cs model training procedure for each single-actor multi-critics thread

```
Initialize global parameters $\theta$ and $\phi_o$, $o \in \{1, 2, \dots, O\}$.
Initialize global shared counter $\mathrm{T} = 0$.
Initialize local thread step counter $t \leftarrow 1$.
Initialize random preference weights $\omega_o$ using (5.4).
while $\mathrm{T} < \mathrm{T}_{\max}$ do
    Reset gradients: $d\theta \leftarrow 0$ and $d\phi_o \leftarrow 0$.
    Synchronize specific parameters: $\theta^{'} = \theta$ and $\phi^{'}_o = \phi_o$.
    $t_{\mathrm{start}} = t$.
    Get state $s_t$ extracted from massive MIMO networks.
    repeat
        Perform action $a_t$ according to policy $\pi(a_t|s_t;\theta')$.
        Receive reward vector $r_t$ and next state $s_{t+1}$.
        $t \leftarrow t + 1$.
        $\mathrm{T} \leftarrow \mathrm{T} + 1$.
    until terminal $s_t$ or $t - t_{\mathrm{start}} == t_{\max}$;
    for $i \in \{t - 1, \dots, t_{\mathrm{start}}\}$ do
        Update cumulative rewards $\mathcal{R}_i \leftarrow \mathcal{R}_i + \gamma \times \mathcal{R}_{i+1}$.
        Update the weights $\omega_o$ with the Bayesian rule (5.6).
        Compute multi-objective advantage function by (5.7).
        Apply the balanced-reward aggregation with (5.10).
        for $o \in \{1, 2, \dots, O\}$ do
            Update multi-critic networks for $\phi^{'}_o$ with (5.12).
        Update single actor network for $\theta^{'}$ with (5.11).
    Asynchronous global update of $\theta$ and $\phi_o$ with $\theta^{'}$ and $\phi^{'}_o$.
```

조선대학교
CHOSUN UNIVERSITY

Chapter Ⅵ

Simulation Results and Analyses

In this chapter, we present the simulation setup, the employed benchmarks, and simulation results to evaluate the performance of the proposed MO-A3Cs-based PA in downlink multi-cell massive MIMO networks.

A. Simulation Parameters

In the simulation setup, we consider 16 square cells with one BS per cell, and each cell has an area of 250 × 250 m. The UEs in the network are equipped with a single antenna and are randomly and uniformly distributed in each cell. The minimum distance between the BS and the UE is set to 25 m. The channel gain at a distance of 1 km is -148.1 dB, and the path loss exponent is set to 3.76. The noise power of the receiver and the noise figure of each BS are set to -94 dBm and 7 dBm, respectively. The parameters considered in the massive MIMO system and the power consumption model utilized in simulations are listed in Table 6.1.

조선대학교 CHOSUN UNIVERSITY

Table 6.1: System parameters of the downlink multi-cell massive MIMO network setup.

System parameters	Network setup
Number of cells (L)	16
Number of UEs per cell (K)	[5, 10]
Number of transmit antennas (M)	[20, 100]
Bandwidth (B)	20 MHz
Pilot reuse factor (τ_p)	4
Coherence block length (τ_c)	200
Power for BS antennas (p_{BS})	0.4 W
Power for BS local oscillator (p_{LO})	0.2 W
Power per UE (p_{UE})	0.2 W
Power for backhaul traffic (p_{BT})	0.25 W/(Gbit/s)
Power for data encoding (p_{ENC})	0.1 W/(Gbit/s)
Power for data decoding (p_{DEC})	0.8 W/(Gbit/s)
BS computation efficiency (L_{BS})	75 Gflops/W
UL transmit power (p^{UL})	0.1 W
Fixed BS power (P_{FIX})	10 W
Fixed power for signal process (P_{SP})	0.1 W
Minimum transmission power ($P_{\min}$)	5 dBm
Maximum transmission power ($P_{\max}$)	38 dBm

B. MO-A3Cs Architecture

The proposed MO-A3Cs model comprises four fully connected layers, including two hidden layers and an input and output layer. The state space size is utilized as an input to the first layer, and the size of both the first and second hidden layers is the size of 128 and uses a ReLU activation function. The single actor network outputs the probability of possible action a_t given the state s_t through the softmax function, which results in a probability distribution between 0 and 1. On the other hand, the multi-critic networks comprise several critic networks based on the number of objectives O.

조선대학교
CHOSUN UNIVERSITY

To effectively train the proposed model and various benchmark models, the hyper-parameters of each DRL and MORL model utilized are set as described in Table 6.2.

Table 6.2: Hyperparameters of the utilized DRL and MORL models.

Hyper-parameters	SE-DQN	EE-DQN	PQN	MO-A3Cs
Learning rate	0.001	0.001	0.001	0.001
Discount factor	0.98	0.98	0.98	0.98
Hidden layers	2	2	2	2
Hidden size	[128, 128]	[128, 128]	[128, 128]	[128, 128]
Batch size	64	64	64	64
Update interval	10	10	10	N/A
Number of agents	N/A	N/A	N/A	16
Optimizer	Adam	Adam	Adam	Adam
Activation functions	ReLU	ReLU	ReLU	ReLU
Maximum steps	100,000	100,000	100,000	100,000
Warm-up steps	10,000	10,000	30,000	N/A
Replay buffer size	50,000	50,000	100,000	N/A
Initial ϵ	1.0	1.0	1.0	N/A
Final ϵ	0.01	0.01	0.01	N/A
ϵ-decay	0.995	0.995	0.995	N/A
Loss function	Huber	Huber	MSE	(5.14)

C. Benchmark Methods

The performance of the proposed PA scheme is compared with the existing benchmarks, including iterative algorithm-based PA methods, conventional DRL models, and MORL model-based PA techniques.

조선대학교
CHOSUN UNIVERSITY

1. Algorithmic Approaches

The considered benchmark algorithms comprise an equal PA method which allocates equal transmission power, and a PA technique based on the Dinkelbach algorithm [46]. Typically, the Dinkelbach algorithm addresses the fractional programming problem [47]. For Dinkelbach algorithm, we transformed the problem in (3.14) into $\max_{x}\frac{f(x)}{g(x)}$. Here, $f(x)$ and $g(x)$ represent the objective function of SE and EE, respectively, and x indicates the downlink transmission power. This fractional programming problem exhibits non-linear and non-convex characteristics. In the Dinkelbach algorithm, the problem is transformed into a sub-problem of the form $\max_{x}[f(x)-\kappa g(x)]$ and then addressed through an iterative process based on an arbitrary scalar value κ updated using the ratio of $f(x)$ to $g(x)$ in each iteration. This value is updated to $\frac{f(x^*)}{g(x^*)}$ using the optimal solution x^* obtained at each stage. In addition, this iterative process is conducted by gradually adjusting the transmission power up to the $P_{\max}$.

Therefore, the Dinkelbach-based downlink PA method optimizes the transmit power until the ratio of SE to EE in the transformed sub-problem becomes smaller than the parameter ζ. Considering the computational complexity and accuracy of the Dinkelbach algorithm, the value ζ is set to 0.001.

조선대학교 CHOSUN UNIVERSITY

2. Reinforcement Learning Approaches

The considered DRL and MORL benchmarks include SE-DQN, EE-DQN, and the PFA-based DQN (PQN), where the SE-DQN and EE-DQN aim to maximize SE and EE, respectively. The PQN model jointly optimizes SE and EE. During training, these models use the ϵ-greedy algorithm, a strategy for balancing exploration and exploitation [48]. The value of $\epsilon \in (0, 1)$ while the remaining hyperparameters for each model are listed in Table 6.2. In the DRL models, the state space of SE-DQN includes channel gain and downlink user rate, while the EE-DQN model state space consists of power consumption and computed EE. PQN adopts the same state space as the proposed MO-A3Cs model.

D. Performance Comparison and Analyses

In this section, we evaluate the performance of the proposed MO-A3Cs-based PA in the downlink multi-cell massive MIMO network.

조선대학교 CHOSUN UNIVERSITY

1. Training Performance Evaluation

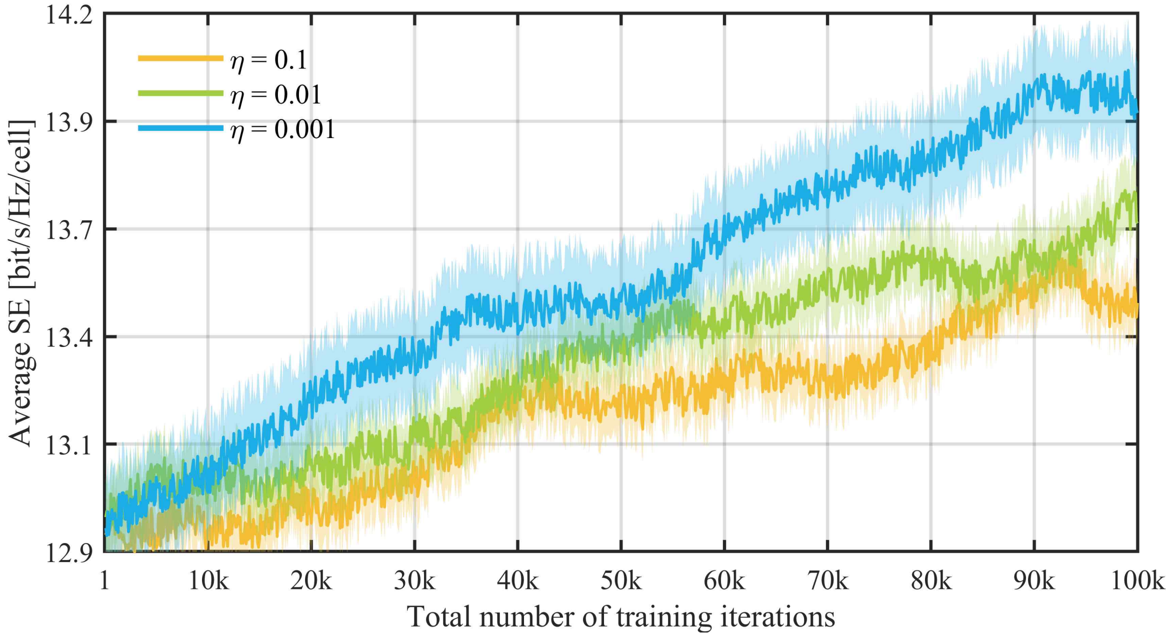


Figure 6.1: Training results of MO-A3Cs model at different learning rates: Average SE reward.

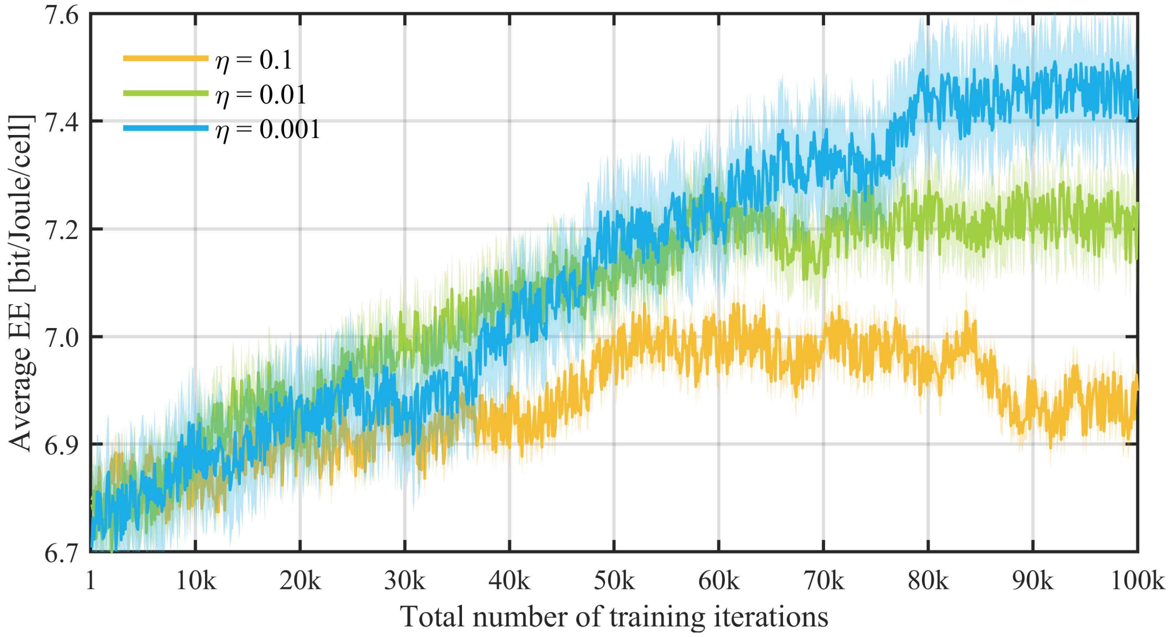


Figure 6.2: Training results of MO-A3Cs model at different learning rates: Average EE reward.

조선대학교 CHOSUN UNIVERSITY

Fig. 6.1 and 6.2 depict the comparison of reward for average SE and average EE based on the learning rate in the proposed MO-A3Cs model. The learning rate η belongs to the set $\{0.1, 0.01, 0.001\}$, which determines the speed at which the model trains from the collected data in the environment.

For the setting $\eta = 0.1$, significant instabilities and fluctuations in the average objective rewards were observed during the training process. On the other hand, the $\eta = 0.01$ showed improvements in the instability and fluctuations at $\eta = 0.1$ from the perspective of average SE and EE rewards.

Futeremore, based on the median of each objective reward, the $\eta = 0.001$ demonstrated improvements of 1.90% and 0.85% in average SE compared to the $\eta = 0.1$ and $\eta = 0.01$, and the average EE enhanced by 3.36% and 0.52%, respectively.

조선대학교
CHOSUN UNIVERSITY

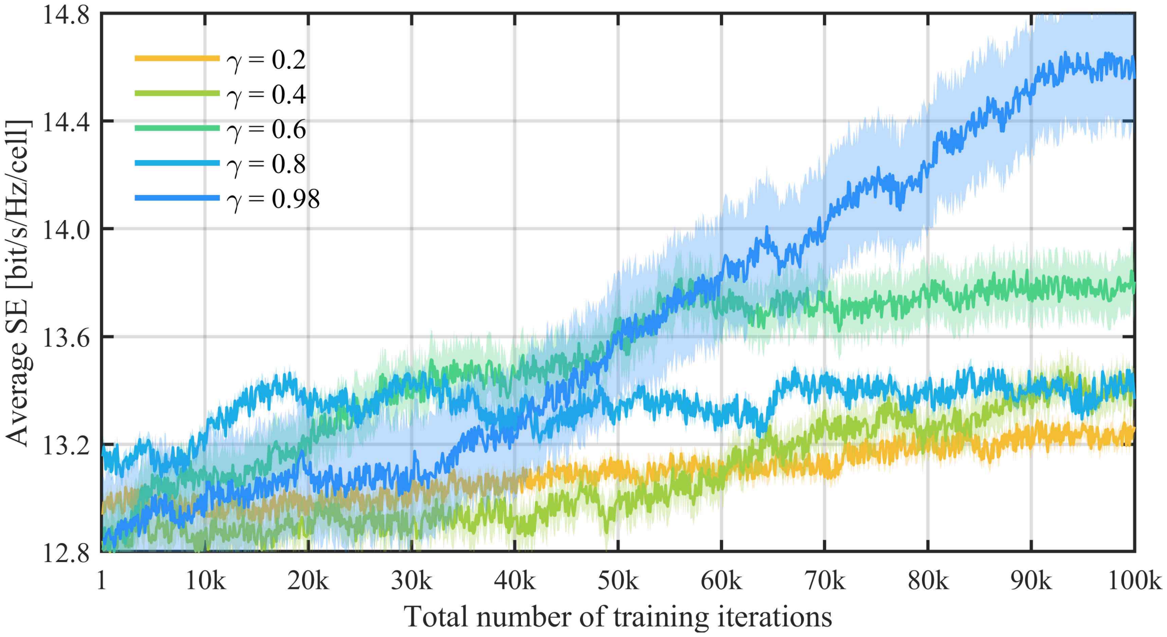


Figure 6.3: Training results of MO-A3Cs model at various discount factors: Average SE reward.

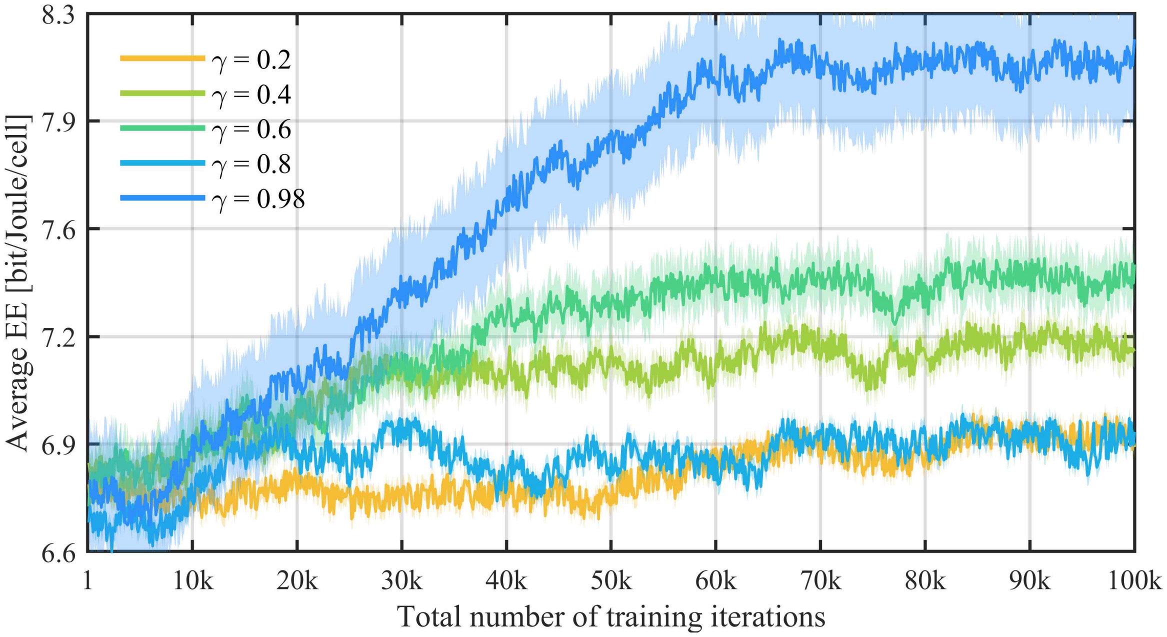


Figure 6.4: Training results of MO-A3Cs model at various discount factors: Average EE reward.

조선대학교 CHOSUN UNIVERSITY

Fig. 6.3 and 6.4 illustrate the training performance of the MO-A3Cs model for each different discount discount factor. In $\gamma = 0.98$, the improved median reward for SE was 3.96%, 4.77%, 0.18%, 1.81%, and EE reward was 15.93%, 9.96%, 6.92%, 14.28% compared to the $\gamma = \{0.2, 0.4, 0.6, 0.8\}$, respectively.

These training results confirm that the proposed MO-A3Cs model can enhance its overall training performance by focusing more on the future rewards. However, it was observed that in the setting of $\gamma = 0.2$ and $\gamma = 0.8$, despite exhausting all of the training iterations, the rewards for SE and EE did not increase. This suggests that the discount factor γ setting can significantly influence the training performance of the MO-A3Cs model.

조선대학교
CHOSUN UNIVERSITY

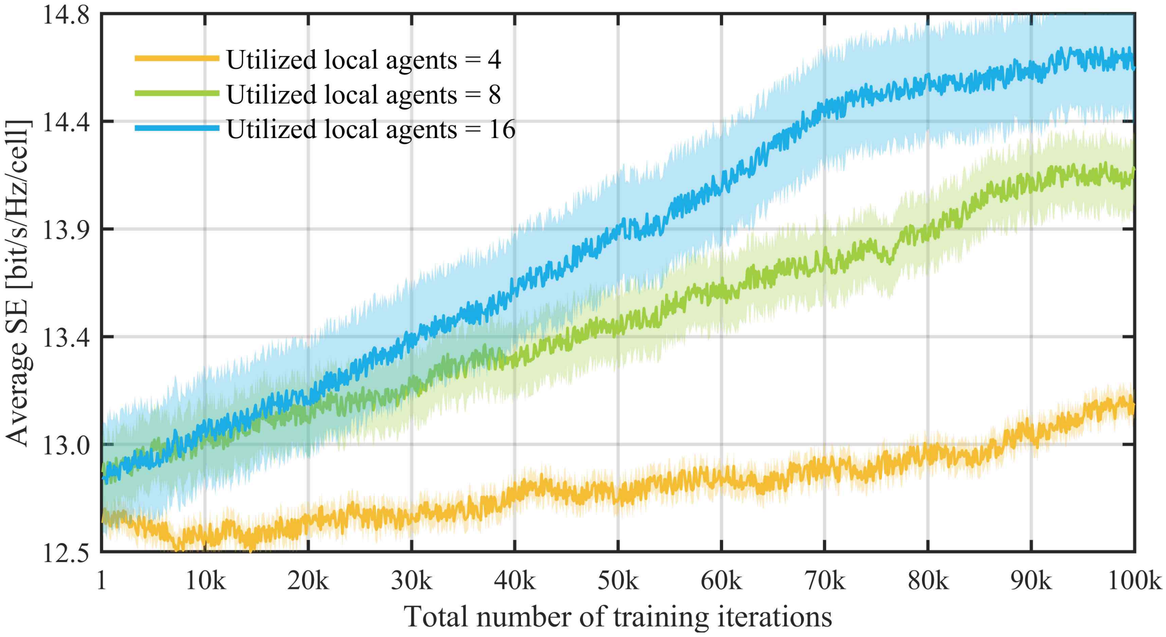


Figure 6.5: Training results of MO-A3Cs model according to the different number of deployed local agents: Average SE reward.

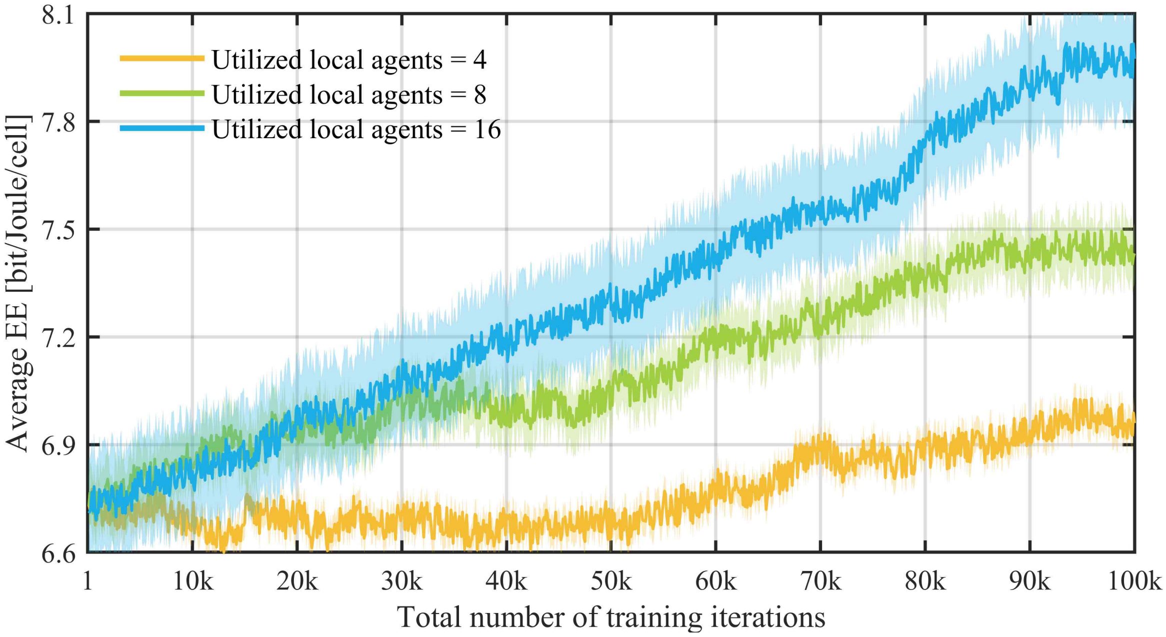


Figure 6.6: Training results of MO-A3Cs model according to the different number of deployed local agents: Average EE reward.

조선대학교
CHOSUN UNIVERSITY

In this thesis, we propose the MO-A3Cs model designed based on the fundamental structure of the A3C model, utilizing multiple local agents to address the training speed degradation issue associated with the single-agent DRL and buffer memory usage. The proposed MORL architecture allows each local agent to interact independently within a downlink multi-cell massive MIMO network environment, collecting diverse trajectories and training the global network. Consequently, it is required to analyze the influence of the number of local agents on the training performance of the proposed MO-A3Cs model.

Fig. 6.5 and 6.6 demonstrate the analysis results of each objective reward of MO-A3Cs according to the number of local agents. The results of 4 local agents have limited diversity in the collected samples in the environments, leading to relatively lower training performance. In contrast, employing 16 local agents results demonstrated an improvement of 9.48% and 3.37% in the rewards for average SE, and 7.97% and 3.83% in the average EE reward, compared to the number of 4 and 8 agents, respectively.

조선대학교
CHOSUN UNIVERSITY

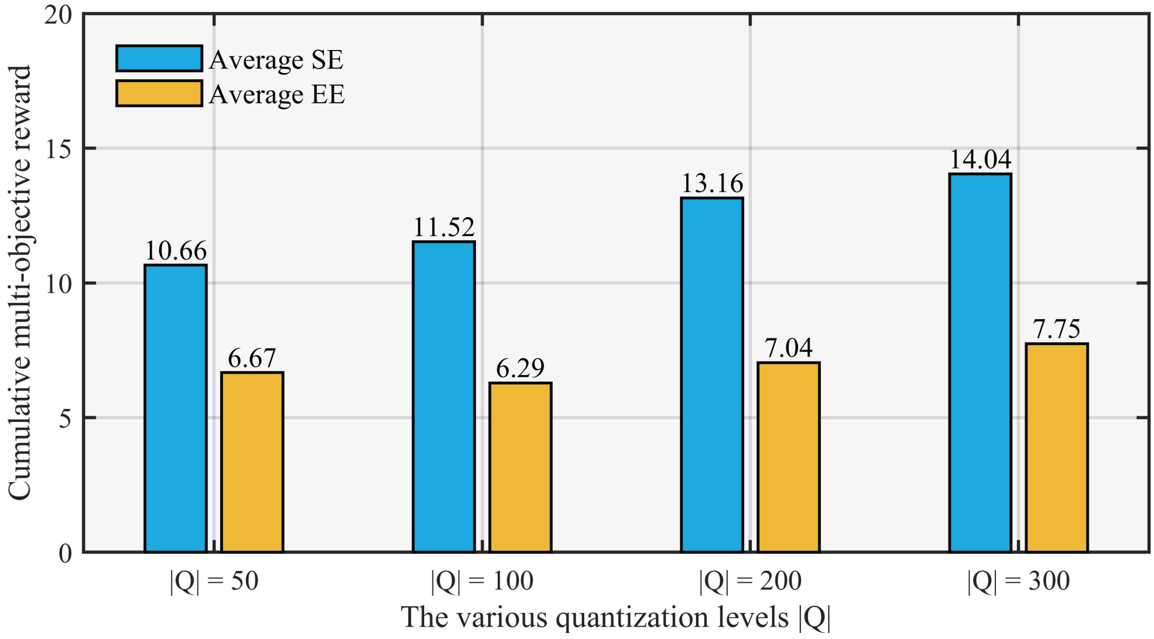


Figure 6.7: Training results of MO-A3Cs model at different quantization levels: Comparison of training efficiency.

Table 6.3: Impact of quantization on training efficiency of the MO-A3Cs model.

Metrics	$\|Q\| = 50$	$\|Q\| = 100$	$\|Q\| = 200$	$\|Q\| = 300$	$\|Q\| = 500$
Convergence training steps	19,144	28,421	42,345	**47,451**	66,753
Convergence time (min.)	11.21	34.19	58.43	**71.16**	203.54
Total training time (min.)	60.30	94.24	141.44	**191.02**	282.42

조선대학교
CHOSUN UNIVERSITY

As discussed in Chapter Ⅲ-B, the quantization level $|Q|$ is a key parameter affecting the training performance of the proposed model. Fig. 6.7 demonstrates the training performance of the MO-A3Cs model at various $|Q|$ levels. In $|Q| = 30$, the average cumulative rewards for SE and EE were the lowest at 10.66 and 6.67, respectively. With the setting of $|Q| = 100$, the SE increased by 0.86, but the EE reward decreased by 0.38 compared to $|Q| = 30$, while with $|Q| = 200$, improvements were recorded in both SE and EE, achieving respective values of 13.16 and 7.04. Furthermore, $|Q| = 300$ showed the highest SE and EE rewards at 14.04 and 7.75.

Table 6.3 shows the training complexity of the MO-A3Cs model for different values of $|Q|$. In $|Q| = 50$, the fastest convergence was achieved in 11.21 minutes. In contrast, the $|Q| = 500$ settings took 203.54 minutes to converge, indicating a significant increase in training time and duration. In addition, changing from 200 to 300 raised 5,106 steps, while changing from 300 to 500 greatly increased by 19,302 steps for convergence. The experiment demonstrates that the most balanced setting between performance and training complexity for the proposed MO-A3Cs model is $|Q| = 300$.

조선대학교 CHOSUN UNIVERSITY

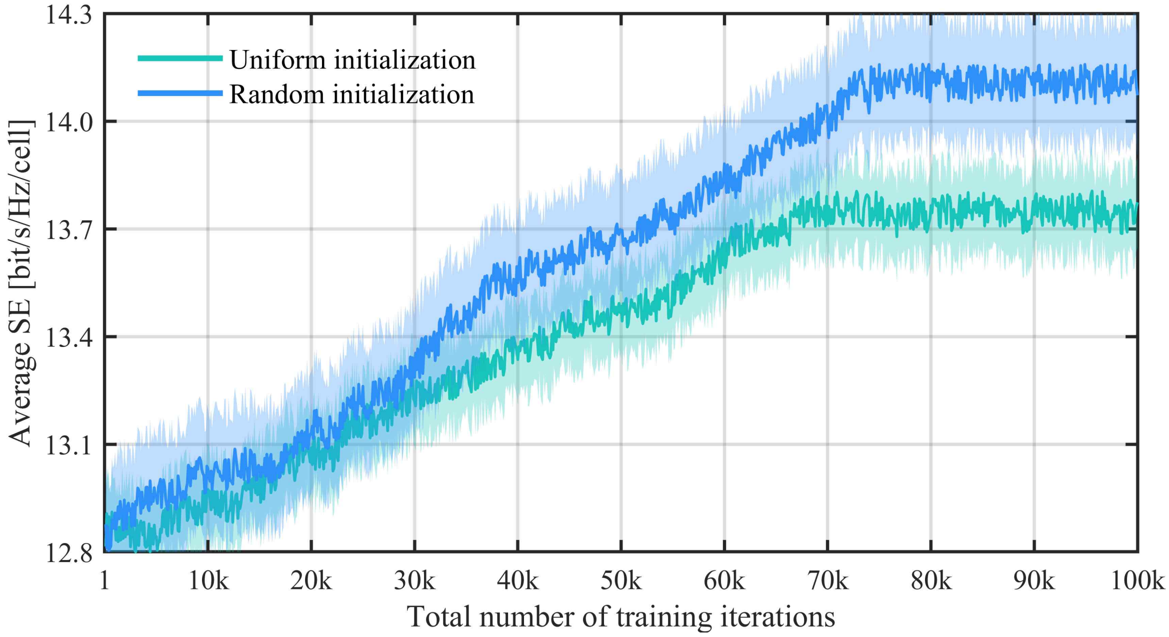


Figure 6.8: Training results of MO-A3Cs model for each different preference weights initialization technique: Average SE reward.

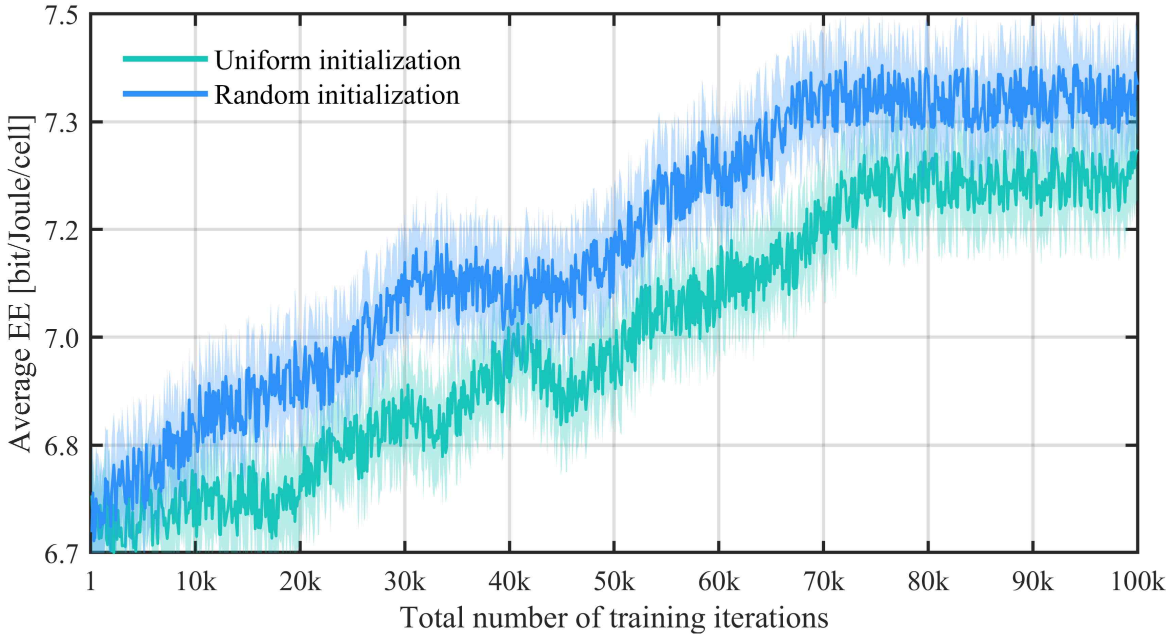


Figure 6.9: Training results of MO-A3Cs model for each different preference weights initialization technique: Average EE reward.

조선대학교
CHOSUN UNIVERSITY

Fig. 6.8 and 6.9 show the average SE and EE rewards for uniform and random initialization methods in the MO-A3Cs model training, respectively. The simulation results demonstrate that utilizing the random method in the proposed MO-A3Cs model achieves the median of SE and EE rewards that are 1.57% and 1.55% higher than the uniform initialization. This result suggests that random initialization is more effective in the proposed Bayesian rule-based preference update mechanism, as discussed in Chapter V-A.

조선대학교 CHOSUN UNIVERSITY

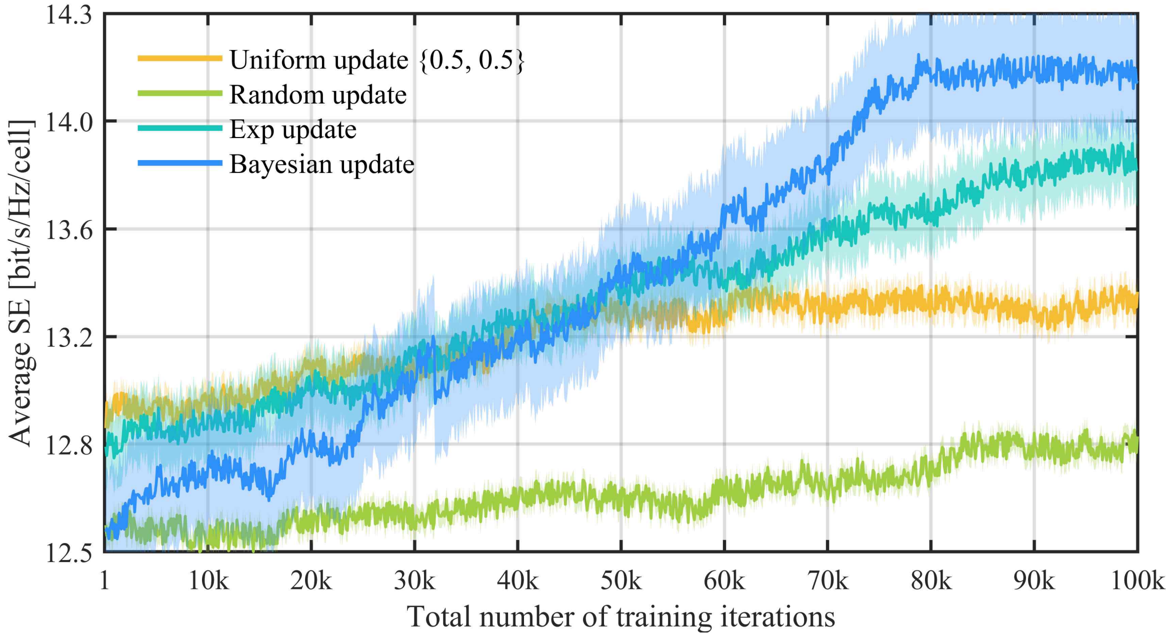


Figure 6.10: Training results of MO-A3Cs model for each different preference weights update methods: Average SE reward.

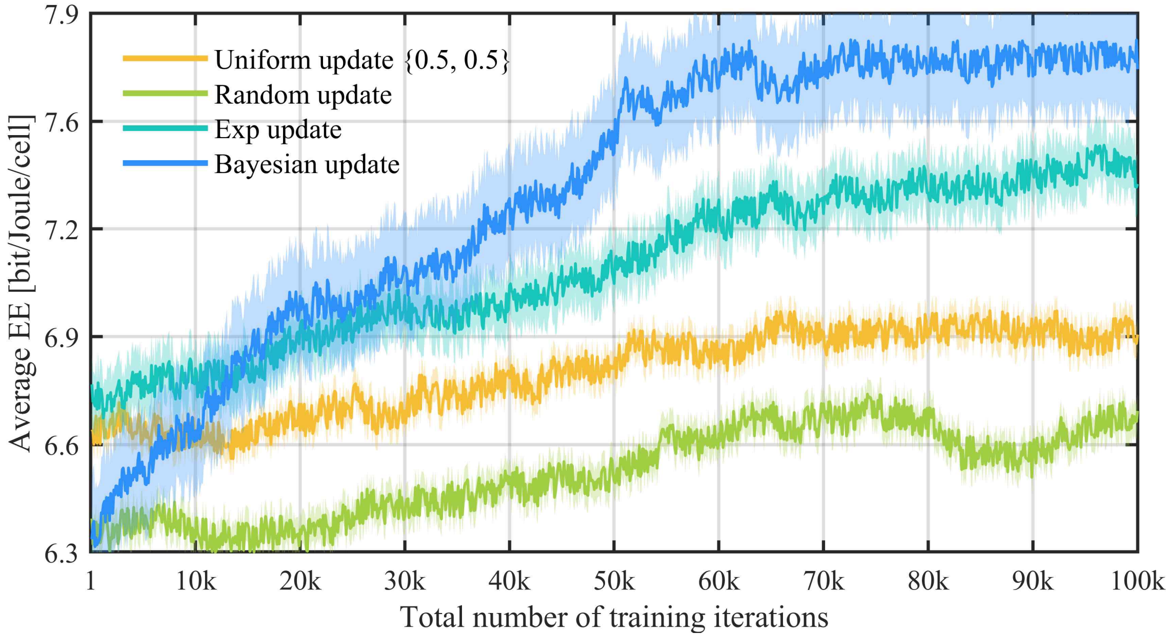


Figure 6.11: Training results of MO-A3Cs model for each different preference weights update methods: Average EE reward.

조선대학교 CHOSUN UNIVERSITY

Fig. 6.10 and 6.11 depict the average objective rewards of the MO-A3Cs model for various preference weight update methods. The uniform update method sets both preference weights to 0.5. The random update method adjusts preference weights with random values between 0 and 1. The exp update method designed to adjust preference weights employs a strategy that exponentially decreases calculated as $\omega_o(t) = \exp(-\upsilon_o \times t)$. Here, υ_o represents the parameter determining the weight decrease rate. This method encourages agents to reduce exploration towards objectives with high rewards and intensively explore objectives with lower rewards. In this simulation, to minimize bias towards SE, the υ_o values for SE and EE were set to 0.08 and 0.06, respectively.

The simulation results show that the proposed Bayesian rule-based preference weight update technique outperforms other methods in achieving the highest objective rewards. Specifically, compared to the exp, uniform, and random methods, the median rewards for SE improved by 0.55%, 1.58%, and 5.95%, while those for EE improved by 3.16%, 7.88%, and 12.75%, respectively. Conventional update methods in the MORL algorithm tend to prioritize SE over EE due to relatively higher values, leading to a decreased priority for EE. On the other hand, our proposed Bayesian rule-based update method adaptively adjusts weights based on obtained trajectories from an interaction between the agents and the environment.

조선대학교 CHOSUN UNIVERSITY

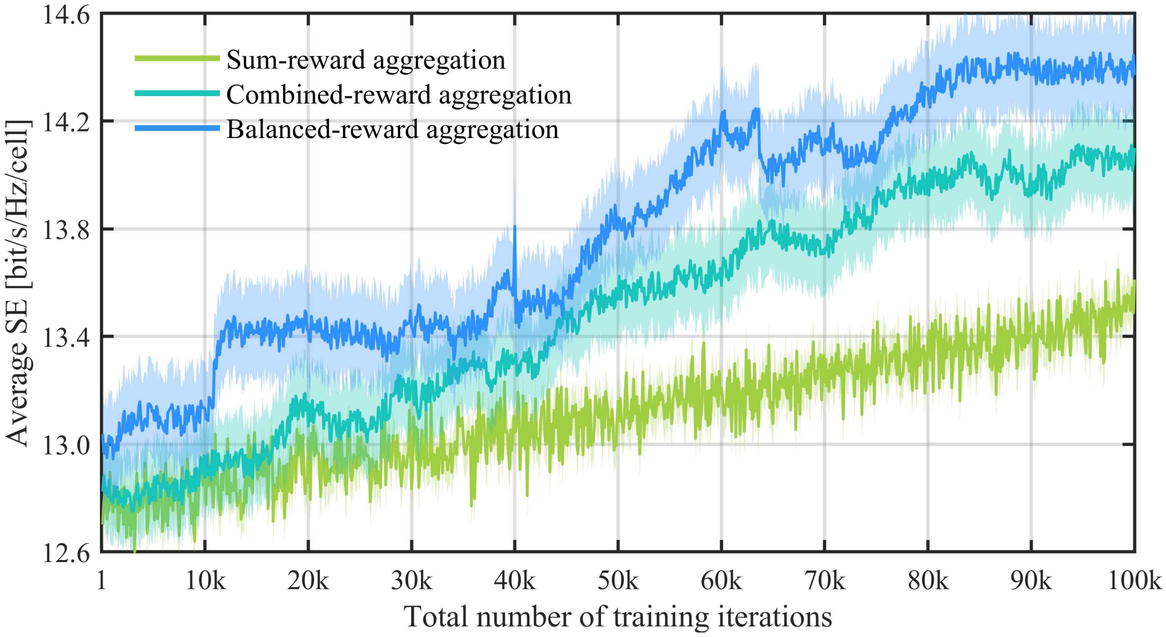


Figure 6.12: Training results of MO-A3Cs model at different reward-aggregation methods: Average SE reward.

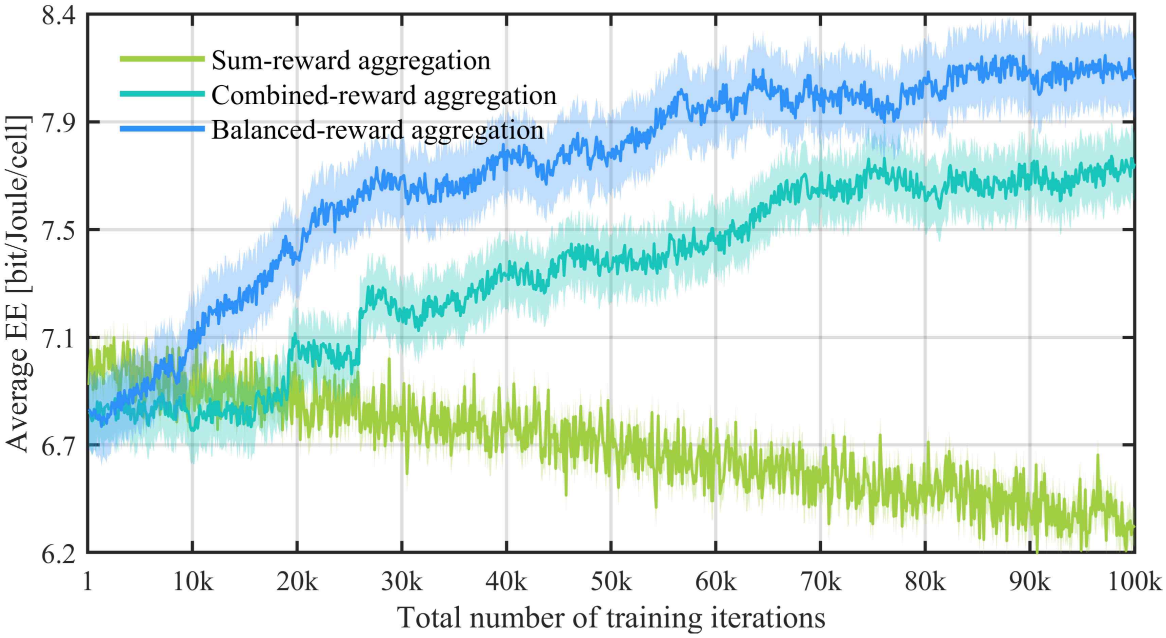


Figure 6.13: Training results of MO-A3Cs model at different reward-aggregation methods: Average EE reward.

조선대학교
CHOSUN UNIVERSITY

Fig. 6.12 and 6.13 show the evaluation of various aggregation methods such as sum-reward, combined-reward, and the proposed balanced-reward as addressed in Section V-B. The simulation results reveal that the sum-reward method mainly focuses on increasing SE during model training, overlooking the trade-off between SE and EE. In contrast, the combined-reward method, which reflects preference weights, showed enhanced performance of 3.24% for SE and 11.10% for EE compared to the sum-reward method.

Furthermore, our proposed balanced-reward aggregation method outperformed other aggregation methods, with enhancements of 5.23% and 17.92% over the sum-reward method and 1.93% and 6.13% compared to the combined-reward method.

조선대학교 CHOSUN UNIVERSITY

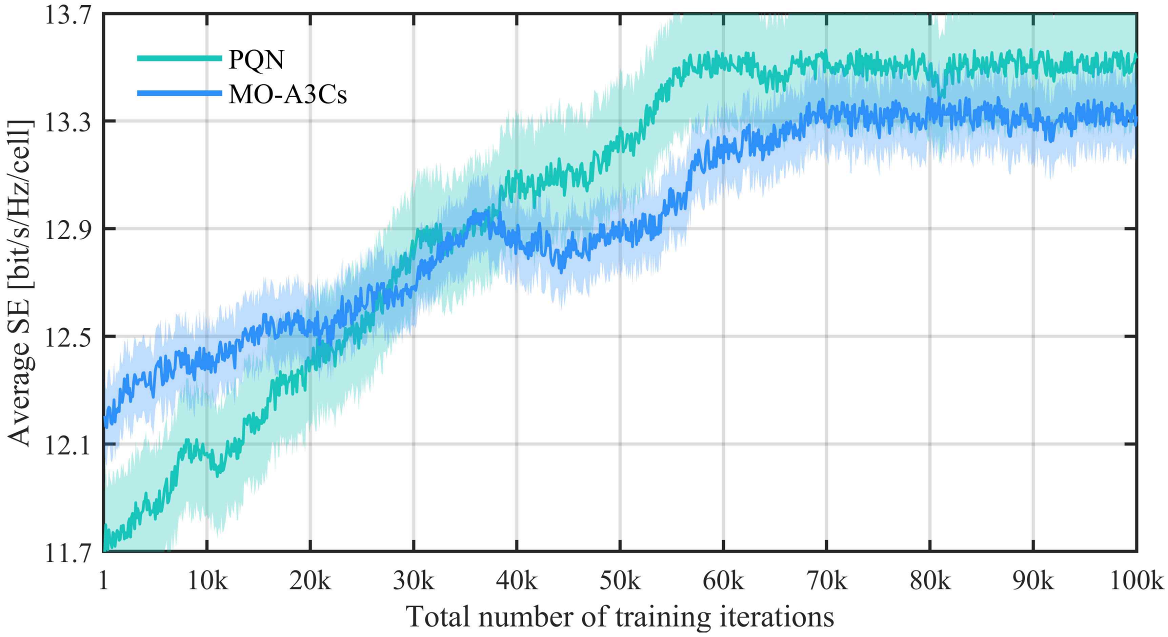


Figure 6.14: Validation of the proposed MO-A3Cs model: Comparison and analysis of Average SE reward with PQN model.

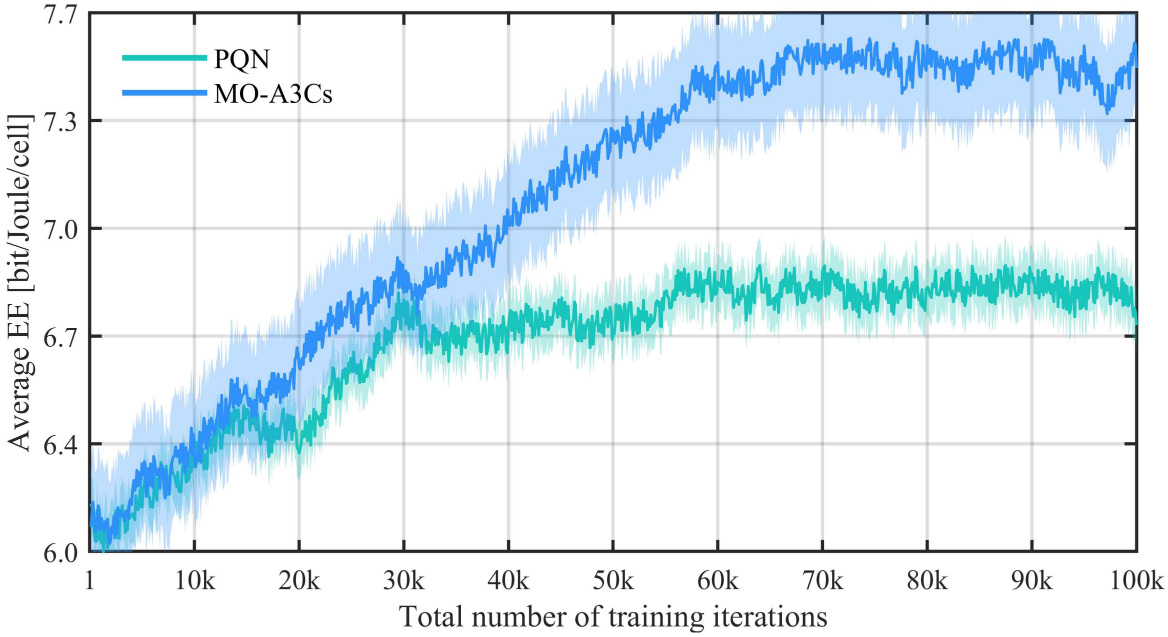


Figure 6.15: Validation of the proposed MO-A3Cs model: Comparison and analysis of average EE reward with PQN model.

조선대학교
CHOSUN UNIVERSITY

Fig. 6.14 and 6.15 present the training performance of the proposed MO-A3Cs model in comparison to the PQN model, a representative model of the MORL algorithm. The simulation results indicate that the average SE reward of MO-A3Cs was approximately 2.26% lower than PQN, while its EE reward was about 7.56% higher. This suggests that the proposed MO-A3Cs model achieves more effective joint optimization of the average SE and EE rewards compared to the PQN model.

Furthermore, the PQN model converges rapidly to a sub-optimal solution for the average EE reward due to its reliance on obtained samples from replay buffers. In contrast, the MO-A3Cs model employs a multi-agent training strategy, enhancing sampling efficiency without utilizing buffer memory. This approach enables the MO-A3Cs model to train more efficiently to achieve joint optimization of SE and EE.

조선대학교
CHOSUN UNIVERSITY

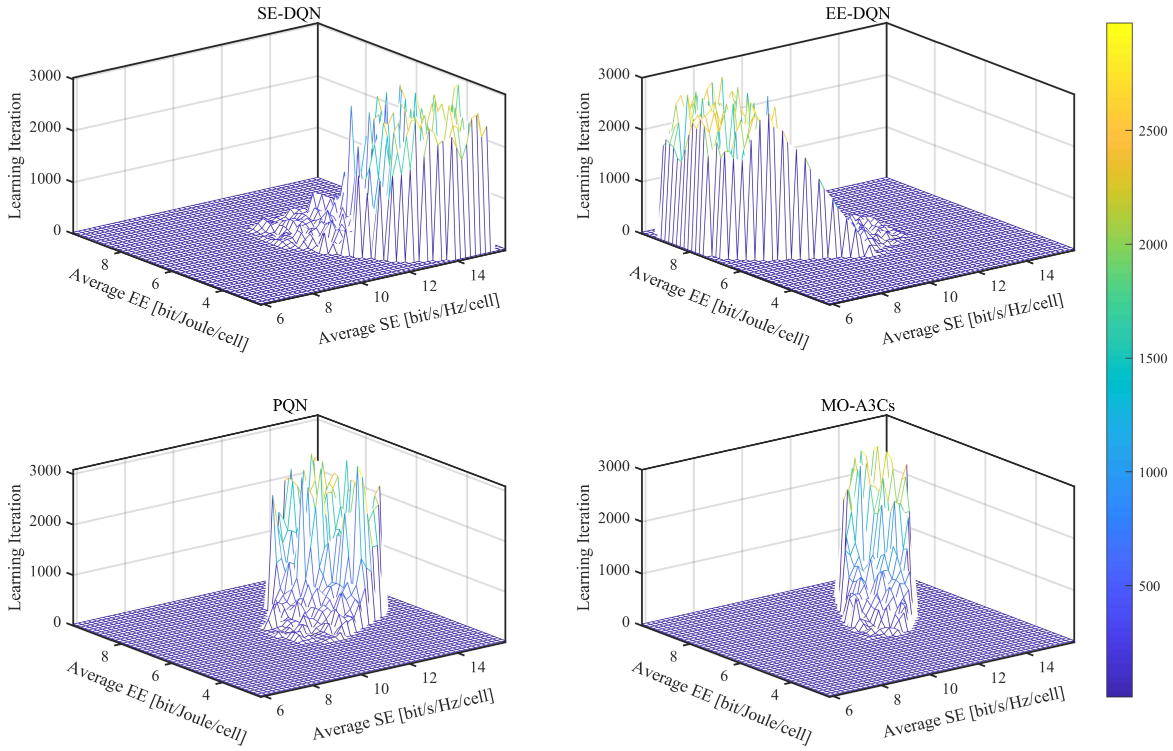


Figure 6.16: Validation of the proposed MO-A3Cs model: Comparison and analysis of convergence results with various benchmark models.

Fig. 6.16 presents a comparative analysis of the convergence direction of multi-objective rewards during the training processes of the SE-DQN, EE-DQN, PQN, and the proposed MO-A3Cs model. For the SE-DQN and EE-DQN models, which were designed to maximize average SE and EE respectively, it can be observed that as the training progresses, the models converge towards maximizing the target objective reward with other objective rewards being ignored.

In contrast, the PQN model is a representative MORL algorithm utilized to validate the convergence direction of the MO-A3Cs model. The PQN converges by considering the multi-objectives, jointly optimizing SE and EE. However, as the number of training iterations increases, the model exhibits a direction similar to that of the SE-DQN

조선대학교 CHOSUN UNIVERSITY

model. This result is due to the limited sample data in the buffer memory used during the Pareto set generation and the approximate Pareto Front process in the PQN model. This suggests that while the PQN can learn a convergence policy that jointly optimizes SE and EE, it may lead to a biased training direction toward a specific objective reward in the model training process.

On the other hand, the proposed model, through independent exploration by multiple local agents in various downlink multi-cell massive MIMO networks, trains the global network based on diverse trajectories. This model training strategy, unlike buffer memory-based model training, can reflect various real-time data and observe states s_t collected by each local agent, enabling more effective model training. Furthermore, by employing the proposed multi-objective advantage function and balanced-reward aggregation method for joint optimization of SE and EE during the model training process, it was experimentally verified that the proposed MO-A3Cs model achieves the most joint optimization performance compared to the utilized DRL benchmarks.

조선대학교
CHOSUN UNIVERSITY

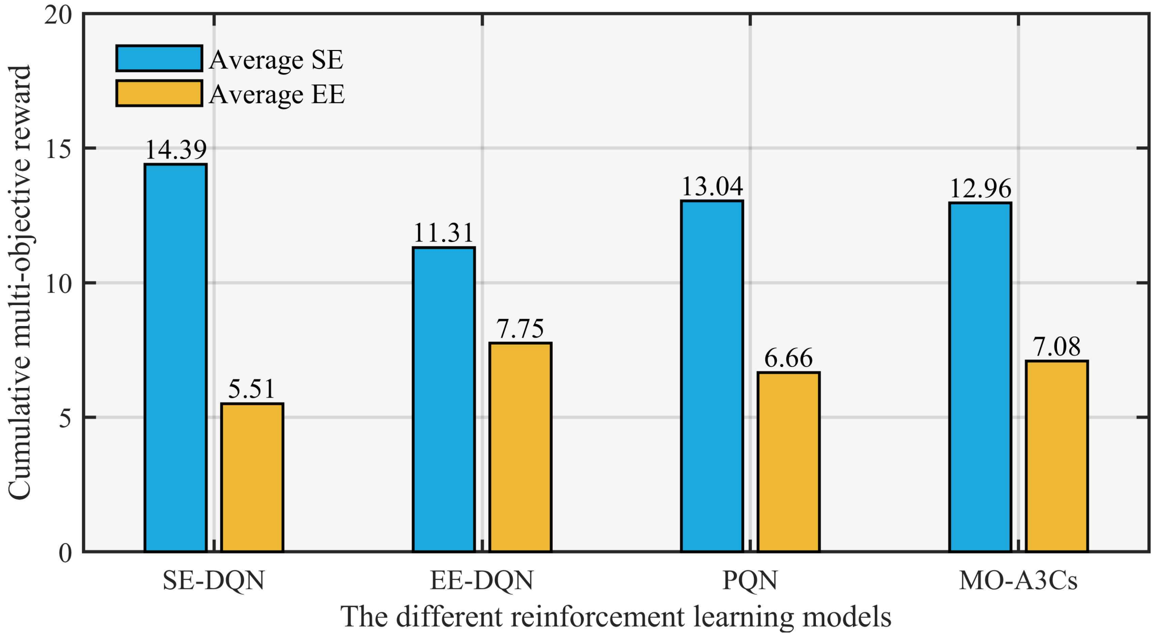


Figure 6.17: Validation of the proposed MO-A3Cs model: Comparison of multi-objective reward with various benchmark models.

Fig. 6.17 depicts the average cumulative rewards achieved for each SE and EE during the training process of the utilized DRL and MORL models. The SE-DQN, aiming to maximize SE, reached the highest average cumulative reward of 14.39 for SE. Similarly, the EE-DQN, focusing on EE maximization, recorded a reward of 7.75 for EE. However, the single-objective models tend to maximize one target reward at the expense of other objectives. In contrast, the proposed MO-A3Cs achieved average cumulative rewards of 12.96 and 7.08 for SE and EE, respectively.

In addition, The difference between the average cumulative SE and EE rewards is 5.88 for MO-A3Cs, 8.88 for SE-DQN, 3.56 for EE-DQN, and 6.38 for the PQN model. These results demonstrate that our proposed MO-A3Cs model achieves the most efficient joint optimization compared to the benchmark models.

조선대학교 CHOSUN UNIVERSITY

2. Simulation Results

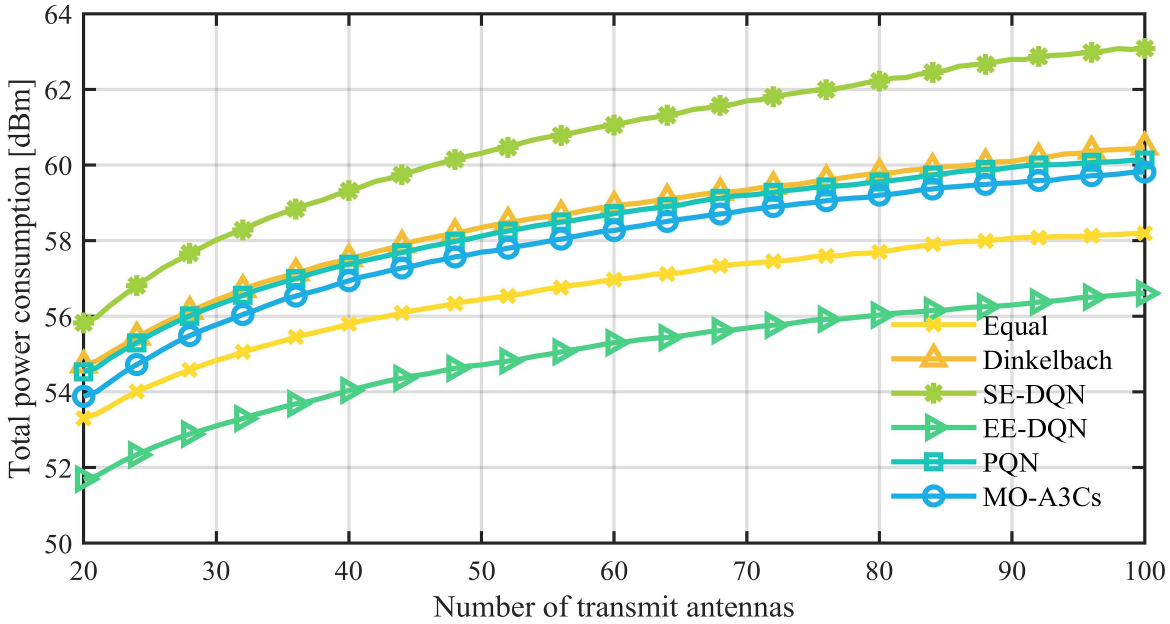


Figure 6.18: An evaluation of the total power consumption for downlink transmit PA methods across different numbers of transmit antenna M from 20 to 100.

Fig. 6.18 illustrates the total power consumption of different PA techniques at settings $L = 16$, $K = 10$, and varying transmit antenna from $M = 20$ to $M = 100$. The SE-DQN method achieved the highest average power consumption at 60.37 dBm, while the EE-DQN method recorded the lowest at 54.76 dBm. The Dinkelbach and PQN methods consumed 58.32 dBm and 58.13 dBm, respectively. Moreover, the proposed MO-A3Cs-based PA showed an average consumption of 57.68 dBm, which is about 4.66% lower than the SE-DQN and 5.07% higher than the EE-DQN. It also consumed 0.78% and 1.11% less power than the PQN and Dinkelbach, respectively. In summary, the simulation results demonstrate that the proposed MO-A3Cs-based PA method can effectively power control at the downlink transmission.

조선대학교 CHOSUN UNIVERSITY

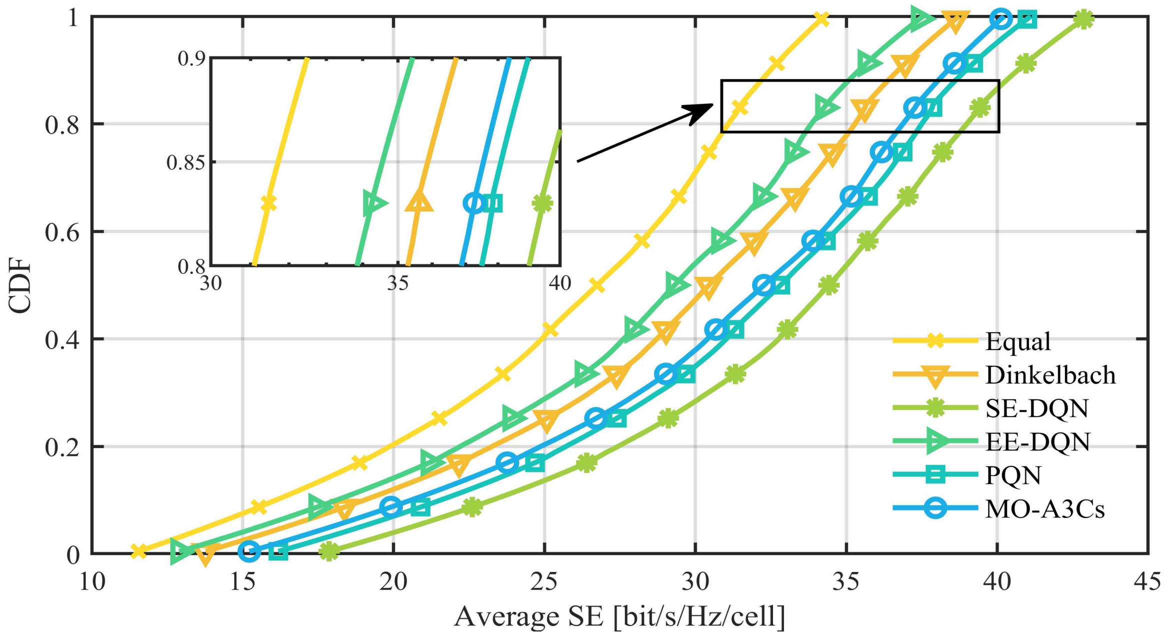


Figure 6.19: Comparison of the CDF for average SE across different downlink transmit PA methods at varying transmit antenna M from 20 to 100.

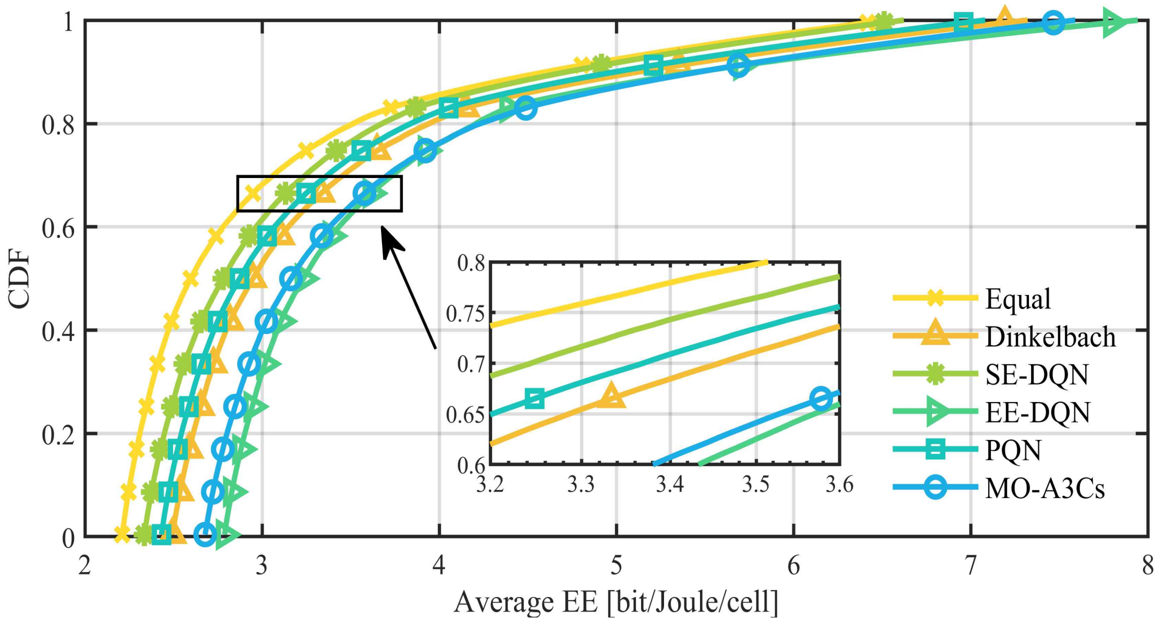


Figure 6.20: Comparison of the CDF for average EE across different downlink transmit PA methods at varying transmit antenna M from 20 to 100.

조선대학교
CHOSUN UNIVERSITY

Fig. 6.19 presents the cumulative distribution function (CDF) analysis results for the SE and EE of various PA techniques under the settings of $L=16$, $K=5$, and varying numbers of downlink transmit antenna from $M=20$ to $M=100$.

In Fig. 6.19, the MO-A3Cs-based PA technique demonstrates a performance that is approximately 2.19% lower than the PQN method. However, compared to EE-DQN and the Dinkelbach techniques, it achieves higher performances by approximately 9.07% and 5.39%, respectively. The enhanced SE performance of the PQN compared to the MO-A3Cs can be attributed to differences in their training policies, leading to different optimal transmission powers.

Moreover, Fig. 6.20 indicates that the EE performance of the MO-A3Cs is close to the EE-DQN method. In contrast to Fig. 6.19, the CDF of EE shows that both PQN and Dinkelbach perform 8.79% and 6.32% lower than the MO-A3Cs, respectively. This analysis confirms that the proposed MO-A3Cs-based PA technique jointly optimizes the overall SE and EE.

조선대학교
CHOSUN UNIVERSITY

Table 6.4: Comparison of trade-offs optimization in downlink transmit PA methods across varying numbers of transmit antenna M from 20 to 100.

PA methods	Optimal point	Avg.SE	Avg.EE
Equal	(21.43, 4.61)	24.90	3.96
Dinkelbach	(25.08, 5.16)	29.16	4.42
SE-DQN	(28.16, 4.38)	29.75	3.91
EE-DQN	(19.87, 6.72)	28.78	4.55
PQN	(25.84, 4.89)	29.70	4.15
MO-A3Cs	**(25.69, 5.26)**	**29.60**	**4.37**

Table 6.5: Comparison of trade-offs optimization in downlink transmit PA methods across varying maximum transmit power $P_{\max}$ from 20 to 60.

PA methods	Optimal point	Avg.SE	Avg.EE
Equal	(19.77, 4.01)	22.87	3.37
Dinkelbach	(23.55, 4.56)	27.20	3.80
SE-DQN	(24.49, 4.18)	28.01	3.38
EE-DQN	(23.53, 4.69)	26.82	3.90
PQN	(24.36, 4.30)	27.80	3.59
MO-A3Cs	**(24.22, 4.63)**	**27.74**	**3.71**

Table 6.4 and 6.5 show the trade-off optimization performance with varying antennas, and maximum transmit powers, respectively. The optimal point is where SE and EE are jointly optimized.

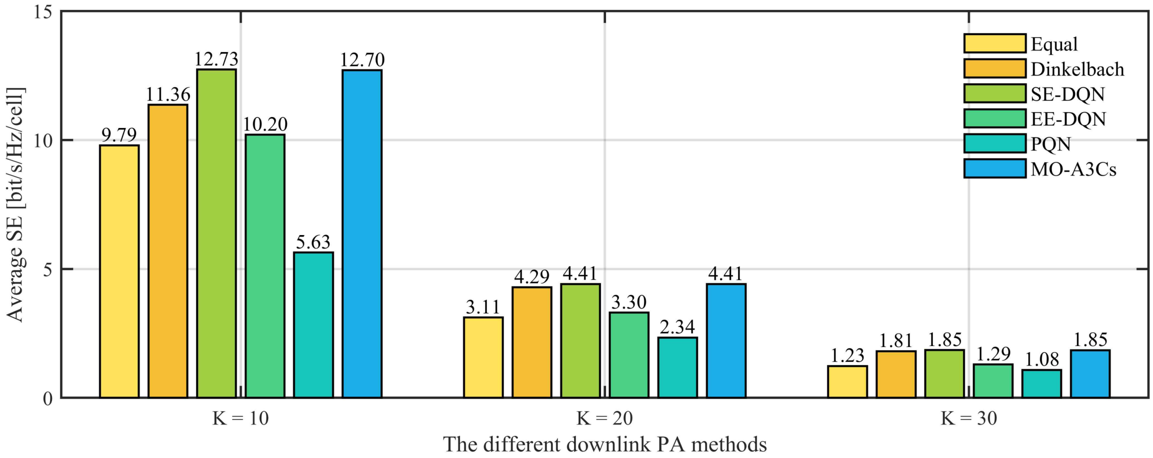
Table 6.4 demonstrates the trade-off optimization performance for each PA method with varying numbers of transmit antennas from 20 to 100. The equal PA technique shows the lowest performance across all metrics, as it does not undertake efficient power control. In addition, the optimal points for the SE-DQN and EE-DQN methods were recorded as (28.16, 4.38) and (19.87, 6.72), respectively. These methods can maximize specific objectives while sacrificing other objectives. In contrast, the

조선대학교 CHOSUN UNIVERSITY

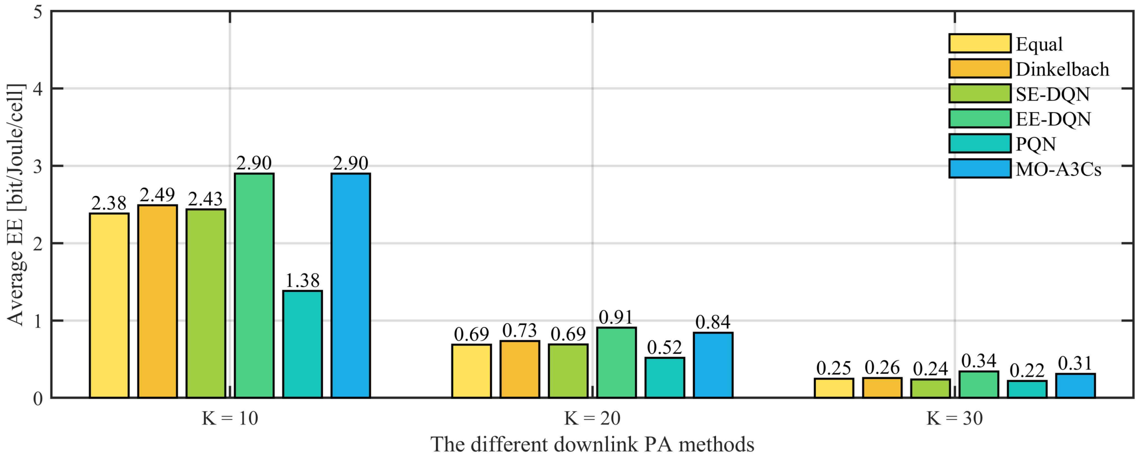
Dinkelbach achieved points of (25.08, 5.16). Moreover, the proposed MO-A3Cs recorded the most balanced point (25.69, 5.26), while the PQN showed a slightly higher SE value of (25.84, 4.89) than the MO-A3Cs. However, a notable difference is observed in EE values. Regarding average SE, the PQN method approximates the performance of SE-DQN with a value of 29.70, while the MO-A3Cs record a slightly lower at 29.60. For average EE, the MO-A3Cs achieve an improved value of 4.37, representing a 0.22 enhancement over the PQN-based PA approach.

Table 6.5 presents the SE-EE trade-off optimization performance with the number of antennas fixed at 40, while the maximum transmission power ranges from 20 to 60 dBm. These changes in transmission power constraints directly impact the action space and the estimation accuracy of the utilized models. The simulation results demonstrate that the proposed MO-A3Cs method achieves the most efficient optimal points at (24.22, 4.63). In contrast, the PQN appears to closely approximate the performance of the SE-DQN. Moreover, with its adopted PFA approach, the PQN method requires diverse samples to generate the Pareto set and approximate the Pareto front, especially with changes in key parameters such as $P_{\max}$. This implies a need for expanded buffer memory and training duration, in contrast to our proposed MO-A3Cs method.

조선대학교
CHOSUN UNIVERSITY



(a) Average SE with varying numbers of UEs



(b) Average EE with varying numbers of UEs

Figure 6.21: Performance evaluation and analysis of average SE and EE for each downlink PA method in densely deployed UEs per each cell in massive MIMO networks including pre-trained models; (a) Average SE with varying numbers of UEs, (b) Average EE with varying numbers of UEs.

Fig. 6.21 (a) and (b) demonstrate a decrease in average SE and EE as the number of UEs increases in the downlink multi-cell massive MIMO network with $L = 16$ and $M = 40$. This decline is attributed to increased power consumption at BSs due to a higher number of deployed UEs, coupled with the growth of network density

조선대학교 CHOSUN UNIVERSITY

and interference, leading to a decrease in SE and negatively impacting both SE and EE metrics.

Unlike the results in Fig. 6.19 and 6.20, there is a considerable performance difference between the proposed MO-A3Cs model and the PQN model. In particular, the PQN model, despite its capability to handle changing the key network parameters such as the number of antennas, shows limitations in the case of an increasing number of UEs. An increase in UEs seriously influences the overall network density, leading to more complex changes in network environments compared to variations in the number of transmit antennas. As a result, when the PQN was applied in dense scenarios not matching the number of UE sets from the training environment, it performed the lowest. In contrast, the proposed MO-A3Cs model-based downlink PA technique approximately achieved the performance of SE-DQN in Fig. 6.21(a) and the EE-DQN in Fig. 6.21(b), respectively.

The simulation results reveal that the MO-A3Cs model-based downlink PA method achieves adaptive and robust performance, even in environments different from the training setup, outperforming iterative algorithm-based, DRL, and MORL models.

조선대학교 CHOSUN UNIVERSITY

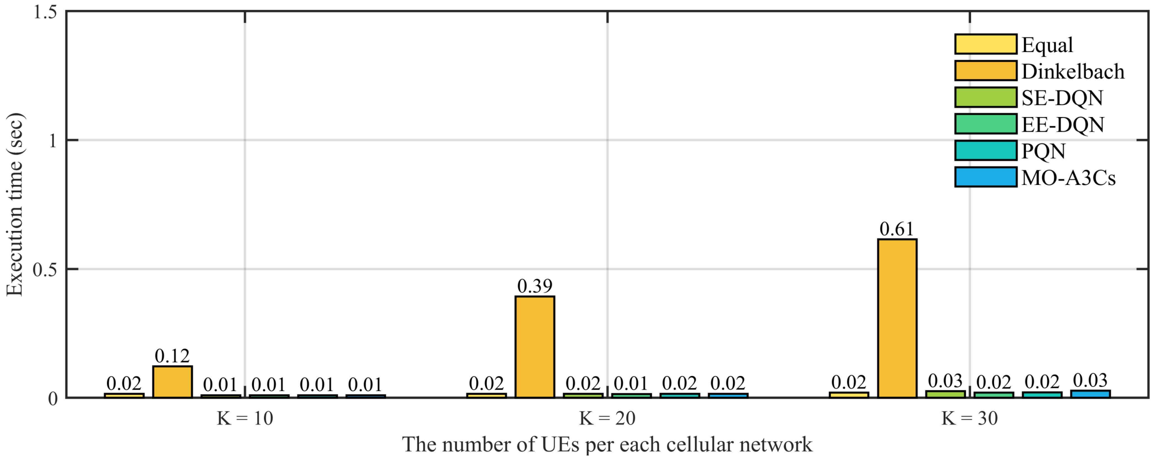


Figure 6.22: Execution time comparison for each downlink PA method for different numbers of UEs and $L = 16$ and $M = 40$ including the pre-trained models.

Fig. 6.22 presents the execution time of the proposed pre-trained MO-A3Cs model as a function of K in comparison with the other learning and model-based algorithms. Fig. 6.22 depicts that the execution time for the Dinkelbach-based downlink PA method increases exponentially with the number of UEs whereas the DRL, MORL-based, and equal PA techniques have less computational complexity even for a higher number of UEs.

These comprehensive simulation results demonstrate that the proposed MO-A3Cs model-based downlink PA framework provides reduced computational costs compared to the interactive algorithms while ensuring robust and joint optimization of SE and EE in dynamically changing downlink multi-cell massive MIMO networks.

조선대학교
CHOSUN UNIVERSITY

Chapter Ⅶ

Conclusion

A. Summary

In this thesis, we propose a transmit PA technique based on the MO-A3Cs model to achieve the SE-EE trade-off in multi-cell massive MIMO networks. The proposed model learns the optimal joint policies to optimize the SE and EE by integrating the MARL-based training strategy with the proposed MORL algorithm. Unlike deep learning and iterative algorithms, trial-and-error-based reinforcement learning maximizes the rewards, takes optimal action through real-time interactions with the environment, and ensure adaptability and robustness in various network scenarios. Comprehensive simulation results demonstrate that our proposed MO-A3Cs model-based downlink PA method optimizes the SE-EE trade-off more effectively and outperforms the conventional MORL algorithm with the PFA approach in terms of joint SE-EE optimization in a dynamic environment. In particular, our proposed PA technique shows robust and flexible performance when varying key network parameters, such as $P_{\max}$ and M in the multi-cell massive MIMO networks. Lastly, we demonstrated that our MO-A3Cs model-based PA method has the possibility of an innovative MORL-based solution for PA techniques in massive MIMO networks.

조선대학교
CHOSUN UNIVERSITY

B. Future Work

The future work for two issues are summarized as follows:

Issue 1: Expansion of the action space to solve the real-world problems

In this thesis, we utilized a quantization level-based discretized action space to address the downlink power control in multi-cell massive MIMO networks. This approach reduces the overhead associated with training models in a continuous action space and provides an efficiently generated action space. However, for real-world problems such as adaptive decision-making with multiple objectives, there is a need to further refine and expand the action space. To this end, our future work aims to extend the proposed MO-A3Cs architecture and the MOMDP to incorporate a continuous action space tailored for such real-world challenges.

Issue 2: Scalability of the proposed MO-A3Cs model

The MO-A3Cs model introduced in this work is designed to optimize the trade-off between SE and EE. However, to address the challenges of next-generation networks, we must consider the characteristics of heterogeneous networks. Therefore, our future work will focus on verifying and expanding the MO-A3Cs model in heterogeneous and next-generation wireless networks.

조선대학교
CHOSUN UNIVERSITY

Abbreviations

5G	Fifth-Generation
A2C	Advantage Actor-Critic
A3C	Asynchronous A2C
ADC	Analog-to-Digital Converter
AWGN	Additive White Gaussian Noise
BS	Base Station
CDF	Cumulative Distribution Function
CF	Cell-Free
DAC	Digital-to-Analog Converter
DDPG	Deep Deterministic Policy Gradient
DDQN	Double DQN
DL	Deep Learning
DNN	Deep Neural Network
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
EE	Energy Efficiency
EE-DQN	Energy Efficiency-DQN
LO	Local Oscillator
MARL	Multi-Agent Reinforcement Learning
MDP	Markov Decision Process
MIMO	Multiple-Input Multiple-Output
MMSE	Minimum Mean-Square Error
MO-A3Cs	Multi-Objective Asynchronous Advantage Actor-multiple Critics
MOMDP	Multi-Objective MDP
MOO	Multiple Objective Optimization
MSE	Mean-Squared Error
MU-MIMO	Multi-User MIMO
PA	Power Allocation
PFA	Pareto front approximation
PQN	PFA-based DQN

조선대학교 CHOSUN UNIVERSITY

RZF	Regularized Zero-Forcing
SE	Spectral Efficiency
SE-DQN	Spectral Efficiency-DQN
SINR	Signal-to-Interference-plus-Noise Ratio
SOO	Single Objective Optimization
SU-MIMO	Single-User MIMO
TD error	Temporal Difference error
TD3	Twin Delayed DDPG
TDD	Time Division Duplex
UE	User Equipment

조선대학교
CHOSUN UNIVERSITY

References

1. F. A. Pereira de Figueiredo, "An overview of massive MIMO for 5G and 6G," *IEEE Latin America Transactions*, vol. 20, no. 6, pp. 931–940, 2022.
2. E. Bjornson, J. Hoydis, and L. Sanguinetti, "Massive MIMO has unlimited capacity," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 574–590, 2018.
3. P. Gandotra, R. K. Jha, and S. Jain, "Green communication in next generation cellular networks: A survey," *IEEE Access*, vol. 5, pp. 11 727–11 758, 2017.
4. Z. Liu, W. Du, and D. Sun, "Energy and spectral efficiency tradeoff for massive MIMO systems with transmit antenna selection," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4453–4457, 2017.
5. W.-Y. Chen, P.-Y. Hsieh, and B.-S. Chen, "Multi-objective power mini-mization design for energy efficiency in multicell multiuser MIMO beam-forming system," *IEEE Transactions on Green Communications and Networking*, vol. 4, no. 1, pp. 31–45, 2020.
6. X. Wang, Y. Wang, W. Ni, R. Sun, and S. Meng, "Sum rate analysis and power allocation for massive MIMO systems with mismatch channel," *IEEE Access*, vol. 6, pp. 16 997–17 009, 2018.

조선대학교
CHOSUN UNIVERSITY

7. L. Sanguinetti, A. Zappone, and M. Debbah, “Deep learning power allocation in massive MIMO,” in *2018 52nd Asilomar Conference on Signals, Systems, and Computers (ACSSC)*, 2018, pp. 1257–1261.
8. R. H. Y. Perdana, T.-V. Nguyen, and B. An, “Deep learning-based power allocation in massive MIMO systems with SLNR and SINR criterions,” in *2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN)*, 2021, pp. 87–92.
9. M. Zhang and M. Chen, “Power allocation in multi-cell system using distributed deep neural network algorithm,” in *2019 International Confer-ence on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 2019, pp. 1–4.
10. F. Meng, P. Chen, and L. Wu, “Power allocation in multi-user cellular networks with deep q learning approach,” in *2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–6.
11. A. Anzaldo and G. Andrade, “Training effect on ai-based resource allocation in small-cell networks,” in *2021 IEEE Latin-American Conference on Communications (LATINCOM)*, 2021, pp. 1–6.
12. S. Zhang, L. Li, J. Yin, W. Liang, X. Li, W. Chen, and Z. Han, “A dynamic power allocation scheme in power-domain NOMA using actor-critic reinforcement learning,” in *2018 IEEE/CIC International Conference on Communications in China (ICCC)*, 2018, pp. 719–723.

조선대학교
CHOSUN UNIVERSITY

13. L. Luo, J. Zhang, S. Chen, X. Zhang, B. Ai, and D. W. K. Ng, “Downlink power control for cell-free massive MIMO with deep reinforcement learning,” *IEEE Transactions on Vehicular Technology*, vol. 71, no. 6, pp. 6772–6777, 2022.
14. Y. S. Nasir and D. Guo, “Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2239–2250, 2019.
15. Y. Yang, F. Li, X. Zhang, Z. Liu, and K. Y. Chan, “Dynamic power allo-cation in cellular network based on multi-agent double deep reinforcement learning,” *Computer Networks*, vol. 217, p. 109342, 2022.
16. M. Rahmani, M. Bashar, M. J. Dehghani, P. Xiao, R. Tafazolli, and M. Debbah, “Deep reinforcement learning-based power allocation in uplink cell-free massive MIMO,” in *2022 IEEE Wireless Communications and Net-working Conference (WCNC)*, 2022, pp. 459–464.
17. A. G. Barto and S. Mahadevan, “Recent advances in hierarchical reinforcement learning,” *Discrete event dynamic systems*, vol. 13, no. 1-2, pp. 41–77, 2003.
18. T. Van Chien, E. Bjomson, and E. G. Larsson, “Joint pilot sequence design and power control for max-min fairness in uplink massive MIMO,” in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–6.

조선대학교
CHOSUN UNIVERSITY

19. H. T. Dao and S. Kim, “Disjoint pilot power and data power allocation in multi-cell multi-user massive MIMO systems,” *IEEE Access*, vol. 6, pp. 66 513–66 521, 2018.

20. A. S. Al-hubaishi, N. K. Noordin, A. Sali, S. Subramaniam, and A. Mohammed Mansoor, “An efficient pilot assignment scheme for addressing pilot contamination in multicell massive MIMO systems,” *Electronics*, vol. 8, no. 4, p. 372, 2019.

21. S. M. Kay, *Fundamentals of statistical signal processing: Estimation theory.* Prentice-Hall, Inc., 1993.

22. B. Saleeb, M. Shehata, H. Mostafa, and Y. Fahmy, “Performance evalua-tion of RZF precoding in multi-user MIMO systems,” in *2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2019, pp. 1207–1210.

23. H. Liu, H. Deng, Y. Yi, Z. Zhu, G. Liu, and J. Zhang, “Energy efficiency optimization based on power allocation in massive MIMO downlink systems,” *Symmetry*, vol. 14, no. 6, pp. 1145–1160, 2022.

24. J. Zhang, H. Deng, Y. Li, Z. Zhu, G. Liu, and H. Liu, “Energy efficiency optimization of massive MIMO system with uplink multi-cell based on imperfect CSI with power control,” *Symmetry*, vol. 14, no. 4, pp. 780–795, 2022.

조선대학교
CHOSUN UNIVERSITY

25. H. Yang and T. L. Marzetta, “Total energy efficiency of cellular large scale antenna system multiple access mobile networks,” in *2013 IEEE Online Conference on Green Communications (OnlineGreenComm)*, 2013, pp. 27–32.

26. O. Amin, E. Bedeer, M. H. Ahmed, and O. A. Dobre, “Energy efficiency–spectral efficiency tradeoff: A multiobjective optimization approach,” *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 1975–1981, 2016.

27. T. T. Nguyen, N. D. Nguyen, P. Vamplew, S. Nahavandi, R. Dazeley, and C. P. Lim, “A multi-objective deep reinforcement learning framework,” *Engineering Applications of Artificial Intelligence*, vol. 96, p. 103915, 2020.

28. P. Vamplew, R. Dazeley, A. Berry, R. Issabekov, and E. Dekker, “Empirical evaluation methods for multi-objective reinforcement learning algorithms,” *Machine learning*, vol. 84, pp. 51–80, 2011.

29. R. Yang, X. Sun, and K. Narasimhan, “A generalized algorithm for multi-objective reinforcement learning and policy adaptation,” *arXiv preprint arXiv:1908.08342*, 2019.

30. S. Huang, A. Abdolmaleki, G. Vezzani, P. Brakel, D. J. Mankowitz, M.Neunert, S. Bohez, Y. Tassa, N. Heess, M. Riedmiller et al., “A constrained multi-objective reinforcement learning framework,” in *Conference on Robot Learning.* PMLR, 2022, pp. 883–893.

조선대학교
CHOSUN UNIVERSITY

31. P. Vamplew, J. Yearwood, R. Dazeley, and A. Berry, "On the limitations of scalarisation for multi-objective reinforcement learning of pareto fronts," in *21st Australasian Joint Conference on Artificial Intelligence (AJCAI)*. Springer, 2008, pp. 372–378.

32. J. Xu, Y. Tian, P. Ma, D. Rus, S. Sueda, and W. Matusik, "Prediction-guided multi-objective reinforcement learning for continuous robot control," in *International conference on machine learning*. PMLR, 2020, pp. 10 607–10 616.

33. C. Liu, X. Xu, and D. Hu, "Multiobjective reinforcement learning: A comprehensive overview," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 3, pp. 385–398, 2015.

34. E. Friedman and F. Fontaine, "Generalizing across multi-objective reward functions in deep reinforcement learning," *arXiv preprint arXiv:1809.06364*, 2018.

35. T. Basaklar, S. Gumussoy, and U. Y. Ogras, "Pd-morl: Preference-driven multi-objective reinforcement learning algorithm," *arXiv preprint arXiv:2208.07914*, 2022.

36. A. Abels, D. Roijers, T. Lenaerts, A. Nowe, and D. Steckelmacher, "Dynamic weights in multi-objective deep reinforcement learning," in *International conference on machine learning*. PMLR, 2019, pp. 11–20.

조선대학교
CHOSUN UNIVERSITY

37. C. F. Hayes, R. Radulescu, E. Bargiacchi, J. Kallström, M. Macfarlane, M.Reymond, T. Verstraeten, L. M. Zintgraf, R. Dazeley, F. Heintz et al., “A practical guide to multi-objective reinforcement learning and planning,” *Autonomous Agents and Multi-Agent Systems*, vol. 36, no. 1, p. 26, 2022.

38. K. Van Moffaert, M. M. Drugan, and A. Nowe, “Scalarized multi-objective reinforcement learning: Novel design techniques,” in *2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, 2013, pp. 191–199.

39. K. I. Ahmed and E. Hossain, “A deep q-learning method for downlink power allocation in multi-cell networks,” *arXiv preprint arXiv:1904.13032*, 2019.

40. V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, “Asynchronous methods for deep reinforcement learning,” in *International conference on machine learning*. PMLR, 2016, pp. 1928–1937.

41. J. V. Stone, *Bayes’rule: A tutorial introduction to bayesian analysis*. Sebtel Press, 2013.

42. R. S. Sutton, “Learning to predict by the methods of temporal differences,” *Machine learning*, vol. 3, pp. 9–44, 1988.

43. H. Lu, D. Herman, and Y. Yu, “Multi-objective reinforcement learning: Convexity, stationarity and pareto optimality,” in *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

조선대학교
CHOSUN UNIVERSITY

44. R. Pasunuru and M. Bansal, “Multi-reward reinforced summarization with saliency and entailment,” *arXiv preprint arXiv:1804.06451*, 2018.

45. J. Schulman, X. Chen, and P. Abbeel, “Equivalence between policy gradients and soft q-learning,” *arXiv preprint arXiv:1704.06440*, 2017.

46. W. Dinkelbach, “On nonlinear fractional programming,” *Management science*, vol. 13, no. 7, pp. 492–498, 1967.

47. S. Schaible, “Fractional programming: Applications and algorithms,” *European Journal of Operational Research*, vol. 7, no. 2, pp. 111–120, 1981.

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

조선대학교
CHOSUN UNIVERSITY

Curriculum Vitae

Name : Youngwoo Oh
Birth Date : Oct. 22, 1997
Birth Place : Gwangju, Republic of Korea
Permanent Address : Gwangju, Republic of Korea

Education

2022.03 - 2024.02	M.S. in Computer Engineering, Chosun University, Gwangju, Republic of Korea
2016.03 - 2022.02	B.S. in Computer Engineering (*cum laude*), Chosun University, Gwangju, Republic of Korea

Publications

- International Journal Papers

1. **Youngwoo Oh**, Arif Ullah, and Wooyeol Choi, "Multi-objective reinforcement learning for power allocation in massive MIMO networks: A solution to spectral and energy trade-offs," *Under Review in IEEE Access*, 2023.
2. Yonggang Kim, Gyungmin Kim, **Youngwoo Oh** and Wooyeol Choi, "Transmission delay-based uplink multi-user scheduling in IEEE 802.11ax networks," *Applied Sciences*, vol. 11, no. 19, article no. 9196, October 2021.

조선대학교
CHOSUN UNIVERSITY

• International Conference Papers

1. Raghavendra Ganiga, **Youngwoo Oh** and Wooyeol Choi, “Design and implementation of middleware platform for real-time monitoring of vital body signs in mobile aplication,” in Proc. of *International Symposium on Advanced Intelligent Systems (ISIS)*, Jeju, Republic of Korea, December 6-9, 2023.
2. **Youngwoo Oh** and Wooyeol Choi, “Deep reinforcement learning-based power allocation in multi-cell massive MIMO,” in Proc. of *International Conference on Maritime IT Convergence (ICMIC)*, Jeju, Republic of Korea, September 22-23, 2022.

• Domestic Journal Papers

1. **Youngwoo Oh** and Wooyeol Choi, “Design and implementation of TDMA-based multi-hop relay network with adaptive equalizer for inter-symbol interference compensation,” *The Journal of Korea Information and Communications Society*, vol. 46, no. 6, June 2021.

• Domestic Conference Papers

1. **Youngwoo Oh** and Wooyeol Choi, “A study on deep reinforcement learnings-based resource allocation for 5G networks,” in Proc. of *Summer Conference on Korea Information and Communications Society (KICS)*, Jeju, Republic of Korea, June 21-24, 2023.

조선대학교
CHOSUN UNIVERSITY

2. Raghavendra Ganiga, **Youngwoo Oh** and Wooyeol Choi, “Design and implementation of lightweight, scalable, and secure REST APIs for seamless integration with hospital ICT infrastructure,” in Proc. of *Summer Conference on Korea Information and Communications Society (KICS)*, Jeju, Republic of Korea, June 21-24, 2023.
3. Raghavendra Ganiga, **Youngwoo Oh** and Wooyeol Choi, “Streamlining healthcare with NLP and AI: Extracting medical information from unstructured text and linking of medical codes,” in Proc. of *Summer Conference on Korea Information and Communications Society (KICS)*, Jeju, Republic of Korea, June 21-24, 2023.
4. **Youngwoo Oh** and Wooyeol Choi, “Multi-objective reinforcement learning-based power allocation for joint optimization of spectral efficiency and user fairness in massive MIMO systems,” in Proc. of *The 33rd Joint Conference on Communications and Information (JCCI)*, Yeosu, Republic of Korea, April 26-28, 2023.
5. Hyeju Han, **Youngwoo Oh**, Minsu Park, Kwang Myung Jeon, Chaejun Leem and Wooyeol Choi, “Design and implementation of Internet of Things (IoT)-based realtime indoor localization system,” in Proc. of *Winter Conference on Korean Institute of Electromagnetic Engineering and Science (KIEES)*, Jeju, Republic of Korea, February 15-17, 2023.

조선대학교 CHOSUN UNIVERSITY

6. Bumsu Kim, **Youngwoo Oh**, Yong Suk Oh and Wooyeol Choi, “Design and implementation of medical-ICT convergence healthcare application,” in Proc. of *Winter Conference on Korean Institute of Electromagnetic Engineering and Science (KIEES)*, Jeju, Republic of Korea, February 15-17, 2023.
7. **Youngwoo Oh** and Wooyeol Choi, “An actor-critic deep reinforcement learning-based antenna selection scheme for MIMO systems,” in Proc. of *Summer Conference on Korea Information and Communications Society (KICS)*, Jeju, Republic of Korea, June 22-24, 2022.
8. **Youngwoo Oh** and Wooyeol Choi, “Design and performance analysis of automatic modulation classification using convolutional neural network,” in Proc. of *Korea Artificial Intelligence Conference (KoreaAI)*, Jeju, Republic of Korea, September 29 - October 1, 2021.
9. Kwang Myung Jeon, Inchul Ryu, Nuri Kim, Chaejun Leem, **Youngwoo Oh**, Chanjun Chun and Wooyeol Choi, “Livestock eartag recognition system based on smart glasses and lightweight OCR,” in Proc. of *Korea Artificial Intelligence Conference (KoreaAI)*, Jeju, Republic of Korea, September 29 - October 1, 2021.
10. **Youngwoo Oh**, Dongmin Kim and Wooyeol Choi, “Performance analysis of M-QAM/OFDM system using LMS-based adaptive equalizer in Rayleigh fading,” in Proc. of *Summer Conference on Korean Institute of Electromagnetic Engineering and Science (KIEES)*, Jeju, Republic of Korea, August 18-21, 2021.

조선대학교
CHOSUN UNIVERSITY

11. **Youngwoo Oh** and Wooyeol Choi, "Implementation and performance analysis of multi-hop relay network based on time division multiple access," in Proc. of *Fall Conference on Korea Information and Communications Society (KICS)*, November 13, 2020.
12. **Youngwoo Oh**, Junsu Kim, Siwoong Park and Wooyeol Choi, "Design and implementation of multi-hop relay network based on software-defined radio testbed," in Proc. of *Summer Annual Conference on The Institute of Electronics and Information Engineers (IEIE)*, Jeju, Republic of Korea, August 19-21, 2020.