

Cooperative Multi-task Semantic Communication: Semantic-Aware NOMA Adaptation

Master Thesis

submitted by

Sakib Absar

Matr.-No.: 6091109

Bremen, July 01, 2025

Cooperative Multi-task Semantic Communication: Semantic-Aware NOMA Adaptation



Fachbereich 1 - Physics and Electronical Engineering
Institute for Telecommunications and High-Frequency Techniques (ITH)
Department of Communications Engineering
P.O. Box 33 04 40
D-28334 Bremen

Supervisor: A. Halimi Razlighi, M.Sc.

First Examiner: Prof. Dr.-Ing. A. Dekorsy

Second Examiner: Dr.-Ing. C. Bockelmann

I ensure the fact that this thesis has been independently written and no other sources or aids, other than mentioned, have been used.

Bremen, July 01, 2025

Sakib absar

.....

Hinweise zu den offiziellen Erklärungen

1. Die folgende Seite mit den offiziellen Erklärungen

- A)** Eigenständigkeitserklärung
- B)** Erklärung zur Veröffentlichung von Bachelor- oder Masterarbeiten
- C)** Einverständniserklärung über die Bereitstellung und Nutzung der Bachelorarbeit / Masterarbeit in elektronischer Form zur Überprüfung durch eine Plagiatsoftware

ist entweder direkt in jedes Exemplar der Bachelor- oder Masterarbeit fest mit einzubinden oder unverändert im Wortlaut in jedes Exemplar der Bachelor- oder Masterarbeit zu übernehmen.

Bitte achten Sie darauf, jede Erklärung in allen drei Exemplaren der Arbeit zu unterschreiben.

2. In der digitalen Fassung kann auf die Unterschrift verzichtet werden. Die Angaben und Entscheidungen müssen jedoch enthalten sein.

Zu B)

Die Einwilligung kann jederzeit durch Erklärung gegenüber der Universität Bremen, mit Wirkung für die Zukunft, widerrufen werden.

Zu C)

Das Einverständnis der dauerhaften Speicherung des Textes ist freiwillig.

Die Einwilligung kann jederzeit durch Erklärung gegenüber der Universität Bremen, mit Wirkung für die Zukunft, widerrufen werden.

Weitere Informationen zur Überprüfung von schriftlichen Arbeiten durch die Plagiatsoftware sind im Nutzungs- und Datenschutzkonzept enthalten. Diese finden Sie auf der Internetseite der Universität Bremen.

A) Eigenständigkeitserklärung

Ich versichere, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Alle Teile meiner Arbeit, die wortwörtlich oder dem Sinn nach anderen Werken entnommen sind, wurden unter Angabe der Quelle kenntlich gemacht. Gleiches gilt auch für Zeichnungen, Skizzen, bildliche Darstellungen sowie für Quellen aus dem Internet, dazu zählen auch KI-basierte Anwendungen oder Werkzeuge. Die Arbeit wurde in gleicher oder ähnlicher Form noch nicht als Prüfungsleistung eingereicht. Die elektronische Fassung der Arbeit stimmt mit der gedruckten Version überein. Mir ist bewusst, dass wahrheitswidrige Angaben als Täuschung behandelt werden.

Ich habe KI-basierte Anwendungen und/oder Werkzeuge genutzt und diese im Anhang "Nutzung KI-basierte Anwendungen" dokumentiert.

B) Erklärung zur Veröffentlichung von Bachelor- oder Masterarbeiten

Die Abschlussarbeit wird zwei Jahre nach Studienabschluss dem Archiv der Universität Bremen zur dauerhaften Archivierung angeboten. Archiviert werden:

- 1) Masterarbeiten mit lokalem oder regionalem Bezug sowie pro Studienfach und Studienjahr 10 % aller Abschlussarbeiten
- 2) Bachelorarbeiten des jeweils ersten und letzten Bachelorabschlusses pro Studienfach und Jahr.

Ich bin damit einverstanden, dass meine Abschlussarbeit im Universitätsarchiv für wissenschaftliche Zwecke von Dritten eingesehen werden darf.

Ich bin damit einverstanden, dass meine Abschlussarbeit nach 30 Jahren (gem. §7 Abs. 2 BremArchivG) im Universitätsarchiv für wissenschaftliche Zwecke von Dritten eingesehen werden darf.

Ich bin nicht damit einverstanden, dass meine Abschlussarbeit im Universitätsarchiv für wissenschaftliche Zwecke von Dritten eingesehen werden darf.

C) Einverständniserklärung zur Überprüfung der elektronischen Fassung der Bachelorarbeit / Masterarbeit durch Plagiatsoftware

Eingereichte Arbeiten können nach § 18 des Allgemeinen Teil der Bachelor- bzw. der Masterprüfungsordnungen der Universität Bremen mit qualifizierter Software auf Plagiatvorwürfe untersucht werden.

Zum Zweck der Überprüfung auf Plagiate erfolgt das Hochladen auf den Server der von der Universität Bremen aktuell genutzten Plagiatsoftware.

Ich bin damit einverstanden, dass die von mir vorgelegte und verfasste Arbeit zum oben genannten Zweck dauerhaft auf dem externen Server der aktuell von der Universität Bremen genutzten Plagiatsoftware, in einer institutionseigenen Bibliothek (Zugriff nur durch die Universität Bremen), gespeichert wird.

Ich bin nicht damit einverstanden, dass die von mir vorgelegte und verfasste Arbeit zum o.g. Zweck dauerhaft auf dem externen Server der aktuell von der Universität Bremen genutzten Plagiatsoftware, in einer institutionseigenen Bibliothek (Zugriff nur durch die Universität Bremen), gespeichert wird.

01.07.2025

Datum

Sakib absar

Unterschrift

Contents

1	Introduction	2
1.1	Research Questions and Challenges	3
1.2	Methodology Overview	4
1.3	Structure of the Thesis Report	5
2	Semantic Communication	6
2.1	Approaches to Semantic Communication	7
2.1.1	Classical Semantic Information Approach	7
2.1.2	Knowledge Graph Approach	8
2.1.3	Machine Learning (ML) Approach	8
2.1.4	Significance-Based Approach	9
2.1.5	Information Theory Approach	10
2.2	Existing Research Directions	10
2.3	Cooperative Multi-Task Semantic Communication (CMT-SemCom) System .	11
3	Non-Orthogonal Multiple Access (NOMA)	14
3.1	Working Principle of Power-domain NOMA	14
3.2	Classification	15
3.2.1	Based on the multiplexing technique	15
3.2.2	Based on architecture	16
3.3	Advantages	18
4	CMT-SemCom Enabled by NOMA	19
4.1	System Model	20
4.1.1	System Probabilistic Modeling	20
4.1.2	Optimization Problem	22
4.2	Simulation Results of CMT-SemCom Enabled by NOMA	24
4.2.1	Simulation Setup	24
4.2.2	Case Study	26
4.3	Summary of CMT-SemCom Enabled by NOMA	31
5	Power Allocation for Wireless Networks	32
5.1	Power Allocation in NOMA	33
5.1.1	Traditional Power Allocation Methods in NOMA	33
5.2	Semantic-Aware Power Allocation	35
6	Semantic-Aware Power Allocation	37
6.1	Full-Observation Setting	38
6.1.1	Model 1: Digital Communication Design	38
6.1.2	Model 2: Analog E2E Design	42
6.2	Distributed Partial-Observation Setting	47
6.2.1	Model 3: Digital Communication Design with Distributed Partial Ob- servation Setting	48

6.2.2	Model 4: Analog E2E Design with Distributed Partial Observation Setting	50
7	Simulation Results	55
7.1	Simulation Setup	55
7.1.1	Dataset	55
7.1.2	Neural Network Architecture	55
7.1.3	Semantic Tasks	56
7.1.4	Training and Evaluation	56
7.1.5	Summary of Simulation Cases	57
7.2	Case Study	57
7.2.1	Digital Communication Design (Full-Observation)	57
7.2.2	Analog E2E Design (Full-Observation)	62
7.2.3	Digital vs. Analog Design Comparison (Full-Observation)	67
7.2.4	Digital Communication Design (Partial-Observation)	68
7.2.5	Analog E2E Design (Partial-Observation)	73
7.2.6	Digital vs. Analog Design Comparison (Partial-Observation)	77
7.2.7	Full vs. Partial Observation Comparison	78
8	Conclusion	81
	Acronyms	83
	List of Symbols	84
	Bibliography	86

List of Figures

2.1	Comparison between traditional bit level and goal-oriented semantic communication [KSM07].	7
2.2	An example knowledge graph representing entities (nodes) and relationships (edges) [GQA ⁺ 23].	9
2.3	Probabilistic graphical modeling of the semantic source [RBD24].	11
2.4	Cooperative multi-task semantic communication system model [RBD24]. . . .	12
3.1	Working principle of a 3-user uplink power-domain NOMA[Pea22].	16
3.2	Signal transmission and decoding technique in Downlink NOMA [ATG ⁺ 18]. .	17
3.3	Signal transmission and decoding technique in Uplink NOMA [ATG ⁺ 18]. . . .	17
4.1	Block diagram of the proposed CMT-SemCom Enabled by DNN-NOMA framework.	20
4.2	Block diagram of the proposed non-cooperative task oriented semantic communication system (without CU and without DNN-NOMA).	25
4.3	Block diagram of the proposed Tx-side cooperative multi-tasking semantic communication system (with CU and without DNN-NOMA).	26
4.4	Block diagram of the proposed Rx-side cooperative multi-tasking semantic communication system (without CU and with DNN-NOMA).	26
4.5	Block diagram of both Tx-side and Rx-side cooperative multi-tasking semantic communication system (with CU and DNN-NOMA).	27
4.6	Performance comparison for “With CU” Vs. “Without CU” for Non-NOMA. .	27
4.7	Performance comparison for “With CU” Vs. “Without CU” for NOMA. . . .	28
4.8	Performance comparison of NOMA Vs. Non-NOMA for “With CU” and “Without CU”.	29
4.9	Performance comparison of NOMA Vs. Non-NOMA for “With CU”.	29
4.10	Performance comparison of NOMA Vs. Non-NOMA for “Without CU”.	30
4.11	Performance comparison of SIC-Base Decoder Vs. DNN-Based Decoder. . . .	31
6.1	Block diagram of the proposed digital communication design of CMT-SemCom enabled by NOMA framework with semantic-aware power allocation.	38
6.2	The semantic value - BER functionsfor task1 and task2.	42
6.3	Block diagram of the proposed analog E2E design of CMT-SemCom enabled by NOMA framework with semantic-aware power allocation.	43
6.4	The MSE - SNR functionsfor task1 and task2	46
6.5	The Semantic value - MSE functionsfor task1 and task2	47
6.6	Block diagram of the proposed digital communication design of CMT-SemCom enabled by NOMA framework with semantic-aware power allocation and distributed partial observation setting.	48
6.7	The semantic value - BER functionsfor task1 and task2.	50

6.8	Block diagram of the proposed analog E2E design of CMT-SemCom enabled by NOMA framework with semantic-aware power allocation and distributed partial observation setting.	51
6.9	The MSE - SNR functions for task1 and task2	53
6.10	The Semantic value - MSE functionsfor task1 and task2	54
7.1	Task error rate comparison of semantic-aware power allocation vs. equally allocated power allocation for digital communication design.	58
7.2	Task error rate comparison of semantic-aware power allocation vs. channel-based power allocation for digital communication design.	59
7.3	Allocated power comparison of semantic-aware power allocation vs. channel-based power allocation for digital communication design	60
7.4	Power scaling effect of semantic-aware power allocation for digital communication design.	62
7.5	Task error rate comparison of semantic-aware power allocation vs. equally allocated power allocation for analog E2E design.	63
7.6	Task error rate comparison of semantic-aware power allocation vs. channel-based power allocation for analog E2E design	64
7.7	Allocated power comparison of semantic-aware power allocation vs. channel-based power allocation for analog E2E design	65
7.8	Power Scaling Effect of semantic-aware power allocation for analog E2E design.	66
7.9	Task error rate of analog E2E and digital communication enabled semantic-aware power allocation structure comparison.	67
7.10	Task error rate comparison of semantic-aware power allocation vs. equally allocated power allocation for digital communication design with distributed partial observation setting	69
7.11	Task error rate comparison of semantic-aware power allocation vs. channel-based power allocation for digital communication design with distributed partial observation setting	70
7.12	Allocated power comparison of semantic-aware power allocation vs. channel-based power allocation for digital communication design with distributed partial observation setting	71
7.13	Power Scaling Effect of semantic-aware power allocation for digital communication design with distributed partial observation setting	72
7.14	Task error rate comparison of semantic-aware power allocation vs. equally allocated power allocation for analog communication enabled structure with distributed partial observation setting	73
7.15	Task error rate comparison of semantic-aware power allocation vs. channel-based power allocation for analog E2E design with distributed partial observation setting	74
7.16	Allocated power comparison of semantic-aware power allocation vs. channel-based power allocation for analog E2E design with distributed partial observation setting	75
7.17	Power Scaling Effect of semantic-aware power allocation for analog E2E design with distributed partial observation setting	76
7.18	Task error rate of analog E2E and digital communication enabled semantic-aware power allocation structure comparison with distributed partial observation setting	77
7.19	Task error rate of digital communication design with and without distributed partial observation setting semantic-aware power allocation structure comparison.	79

7.20 Task error rate of analog E2E with and without distributed partial observation setting semantic-aware power allocation structure comparison	80
---	----

List of Tables

4.1	Simulation line styles for Task 1	25
4.2	Simulation line styles for Task 2	25
7.1	Neural network architecture specifications	56
7.2	Simulation line styles digital communication design	57
7.3	Simulation line styles analog E2E design	57

Chapter 1

Introduction

The rapid development of wireless communication systems has enabled smarter networks, transforming how devices, people, and applications interact with each other. Traditional communication systems primarily aim at reliably communicating bit sequences. They prioritize accurate transmission of raw bits across channels [GQA⁺23]. Although this paradigm has served well for many decades, the emerging demands of intelligent systems need a shift toward semantic and task-oriented communication [Lea21]. Specifically, these systems require an understanding of the semantics or the meaning behind the transmitted data, rather than focusing only on transmitting the bits accurately.

Semantic communication introduces a revolutionary approach by aligning data transmission with the task-specific goals of end-users. Semantic communication optimizes resource utilization and improves system efficiency using task-relevant information. This paradigm is particularly relevant in applications such as autonomous vehicles, Internet of Things (IoT) devices, and intelligent decision-making systems, where only task-relevant information is essential to achieve optimal performance. A semantic communication structure is proposed in [RBD24] for cooperative multi-task semantic communication (CMT-SemCom). This system leverages a cooperative processing unit at the transmitter side to perform the cooperative processing and several specific units to extract task-specific features for individual tasks.

In case of small edge devices, cooperative processing at the transmitter side proposed by the authors of [RBD24] is often difficult due to resource constraints. To address this, cooperative processing can be shifted from the transmitter side to the receiver side with the help of a deep neural network-based decoder within an uplink Non-Orthogonal Multiple Access (NOMA) architecture. In addition, the impact of using cooperative processing both at the transmitter and receiver side should also be investigated. The impact of using a unified decoder to handle all tasks instead of individual decoders can also be explored.

Moreover, NOMA has emerged as a groundbreaking physical layer technique that can enable massive connectivity and enhance spectral efficiency. Unlike traditional orthogonal access schemes, NOMA facilitates simultaneous transmission for multiple users through power-domain multiplexing [Pea22]. This can address the growing demand for high-capacity networks.

Integrating the principles of semantic communication, which represents the higher layers of communication over a network, with a physical layer like NOMA can build an end-to-end (E2E) communication network.

Allocating power is a key part of power-domain NOMA. By allocating higher power to users with weaker channels and lower power to users with stronger channels, the superimposed signal can be separated at each receiver. This power allocation technique based on channel condition ensures user fairness [DWD⁺18]. Existing resource allocation schemes often distribute

resources without considering the semantic importance. The power of each task is allocated according to the channel condition of each user. The user with better channel condition gets less power and the user with worse channel condition gets more power. However, semantics communication brings a new challenge. Semantic information is heterogeneous and not all features have equal semantic values [XMM⁺25]. Treating every semantic feature equally will result in allocating less power for a more important semantic feature and more power for a less important semantic feature. This is why the semantic value should also be considered along with channel conditions to allocate appropriate power.

However, there are some challenges in performing this integration and allocating proper power. Balancing the unique requirements of semantic communication, such as task-specific encoding, with NOMA's principles of efficient resource utilization, requires a deep understanding of both paradigms.

This thesis investigates the integration of CMT-SemCom with DNN-NOMA, including a semantic-aware power-allocation strategy in real-world scenarios.

1.1 Research Questions and Challenges

The focus of this thesis is to investigate the impact of cooperative processing on the transmitter and receiver side by integrating Deep Neural Network (DNN) based NOMA with semantic communication principles. Then we will try to investigate how transmission power can be allocated in a semantic-aware manner to improve task performance. Finally, we will evaluate the proposed model's performance in a real-world scenario in which each user observes only a part of the source image. The study is driven by the following research questions:

RQ1: How can a CMT-SemCom enabled by DNN-NOMA framework be structured for shifting cooperative processing from the transmitter to the receiver?

The first research question investigates the feasibility of enabling cooperative multitask semantic communication over an uplink DNN NOMA-based framework. The key focus lies in relocating cooperative processing from the transmitter side to the receiver side. This is especially advantageous in scenarios where small edge devices are limited in computation and energy resources. The goal is to evaluate whether such a structure can match the performance of traditional orthogonal transmission or transmitter side cooperative systems in maintaining task-level accuracy.

RQ2: How can transmission power be optimally allocated across tasks based on their semantic requirements and channel conditions?

The second research question focuses on the integration of semantic-aware power allocation into the CMT-SemCom enabled by DNN-NOMA framework. Traditional power allocation methods prioritize channel quality or fairness without regard to semantic value. This thesis instead considers an approach where each task is assigned a semantic value target and aims to compute the transmit power that satisfies this constraint. The research explores how semantic performance metrics can be mapped to physical-layer parameters like BER or MSE, and how this mapping can be used to establish power allocation strategies.

RQ3: What are the performance trade-offs between partial and full observation settings?

The third research question examines the impact of distributed partial observation settings on the system model. In the full-observation case, every encoder sees the entire image before feature extraction. This is an ideal scenario with no hidden information. In the distributed partial-observation case, each encoder only sees a part of the image. This models a real-world scenarios where each sensor is observing a part of the world. The goal is to evaluate whether the structure with partial observation settings can match the performance of the structure that has full observation for all users.

1.2 Methodology Overview

This thesis adopts a step-by-step approach that combines system design, mathematical modeling, simulation and performance evaluation to address the proposed research questions.

The work begins with the proposed framework for CMT-SemCom enabled by DNN-NOMA including semantic source modeling, encoder architecture, superimposition and DNN NOMA decoder. The impact of cooperative processing are shown with the help of this structure.

The systems is built on the principle of NOMA, which enables simultaneous transmission of multiple tasks using power-domain multiplexing. Each encoder is assigned with a distinct semantic task such as binary identification or categorical classification. Each specific unit encoder encodes its input using neural network. These features are then superimposed and transmitted to a DNN based decoder.

Another focus of the thesis is on optimizing power allocation in a semantic-aware manner. Instead of simply distributing power equally or based on channel quality, the goal is to assign power according to the semantic importance of each task. This is achieved by modeling the relationship between semantic accuracy and error (such as BER or MSE), and then relating these error values to required transmit power.

Two system architectures are proposed for this:

1. A digital communication design, where semantic features are quantized, digitally modulated, and transmitted through the wireless channel.
2. An analog E2E design, where real-valued features are transmitted directly over the channel.

In the digital communication system design, the training is divided into two phases. First, the encoders and decoder are trained and then, these trained components are used in a digital transmission model. In contrast, the analog structure is trained E2E, jointly optimizing the encoder, channel, and decoder components as a single pipeline. Finally, both the digital and analog E2E design architectures are extended to a distributed partial observation setting where each encoder sees only a portion of the source to meat the real-world scenarios where each sensor is observing a part of the world.

The final part of the methodology involves detailed simulations under varying wireless channel conditions. The outcomes are analyzed in terms of task error rates and total power consumption.

1.3 Structure of the Thesis Report

This thesis is organized into eight chapters to gradually develop and validate a semantic-aware CMT-SemCom enabled by DNN-NOMA framework.

- **Chapter 1:** This chapter presents the motivation behind the research, identifies the core research questions, and provides an overview of the methodology.
- **Chapter 2:** This chapter reviews foundational concepts of semantic and goal-oriented communication including different approaches and existing research directions. It discusses prior work in these areas and identifies limitations in existing approaches that motivate the proposed system.
- **Chapter 3:** This chapter reviews foundational concepts of NOMA such as working principle, classification and benefits of using it.
- **Chapter 4:** This chapter presents the initial system architecture that integrates CMT-SemCom with uplink NOMA. Different variations of the structure, with and without Common Units (CU), and with and without DNN-based joint decoding are explored through simulation to evaluate the performance of different transmitter and receiver side cooperative processing structures.
- **Chapter 5:** This chapter investigates conventional and semantic-aware power allocation strategies relevant to NOMA systems. It establishes the limitations of existing methods when applied to a multi-task semantic setup like ours.
- **Chapter 6:** This chapter introduces four system models that incorporate semantic-aware power allocation into the existing CMT-SemCom enabled by DNN-NOMA framework. The digital communication design and the analog E2E design are both presented along with their respective system model and optimization problems. Then both the digital and analog E2E design architectures are extended to a distributed partial observation setting.
- **Chapter 7:** This chapter presents simulation results that compare the performance of the proposed semantic-aware models against traditional baselines. Metrics such as task error rate, total power consumption, and channel adaptability are analyzed to demonstrate the performance of incorporating semantic-aware power allocation into the CMT-SemCom enabled by DNN-NOMA framework.
- **Chapter 8:** The final chapter summarizes the main findings of the research and discusses potential directions for future work in semantic-aware wireless communication.

Chapter 2

Semantic Communication

Modern wireless communication systems have transformed society by enabling advanced IoT applications, Machine-to-Machine (M2M), and Human-to-Machine (H2M) interactions. Many advancements in wireless technologies are based on Claude Shannon's foundational work on information theory. He defined communication as the problem of "reproducing at one point either exactly or approximately a message selected at another point [Sha48]."

Shannon emphasized that the semantic aspects of communication are irrelevant to the engineering problem. Consequently, existing communication systems have been designed around rate-centric metrics, such as:

- Throughput,
- Spectrum/Energy efficiency,
- Latency.

While Shannon's framework has been foundational, it increasingly shows limitations in addressing the goals of the new generation of wireless systems. Specific shortcomings include:

- A narrow focus on communication reliability without considering the meaning of transmitted data
- Inefficient coupling between the relevance of data content and transmission strategies
- Redundancy in transmission, such as sending information that lacks relevance or freshness [ADI⁺12, 3GP21, Bro17].

These issues cause challenges in transmitting information effectively for exponential data traffic growth [Lea21, SB21].

Warren Weaver, Shannon's collaborator, extended Shannon's framework to define communication across three levels [WS49]:

1. **Technical Problem:** Addressed by Shannon, focusing on "How accurately can the symbols of communication be transmitted?"
2. **Semantic Problem:** Concerned with "How precisely do the transmitted symbols convey the desired meaning?"
3. **Effectiveness Problem:** Focused on "How effectively does the received meaning influence conduct in the desired way?"

The need for more efficient communication systems has led to a shift from semantic neutrality to semantic communication. This paradigm aims to integrate the meaning and effectiveness of communication with transmission strategies [Lea21, SB21].

Semantic communication deals with the meaning and goal of communication systems. It focuses on transmitting only task-relevant semantic information rather than just ensuring bit-level reliability. This approach integrates the contextual understanding of messages and the goals of communication [Lea21, GQA⁺23].

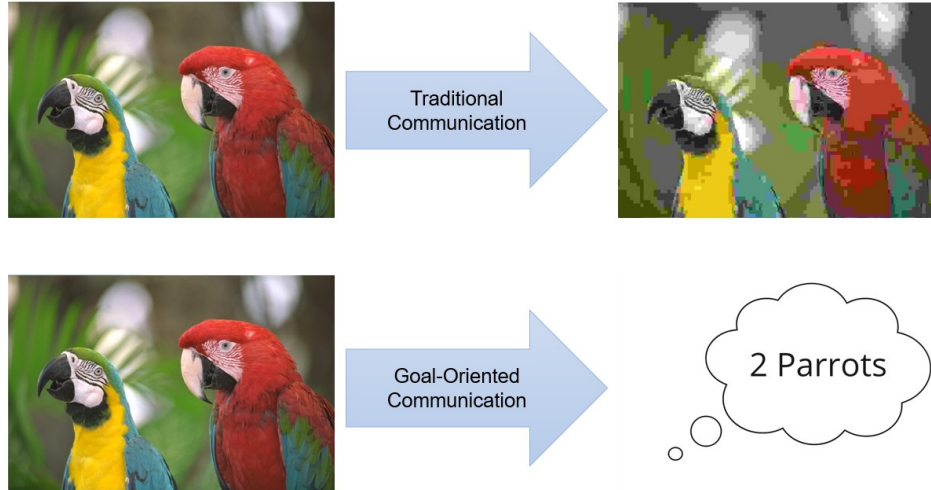


Figure 2.1: Comparison between traditional bit level and goal-oriented semantic communication [KSM07].

Figure 2.1 illustrates the distinction between traditional bit level and goal-oriented communication. The top row represents a traditional bit level approach. Here, the transmitted image is reconstructed with added noise but retains its visual content. The bottom row shows a goal-oriented approach, where the transmitted information focuses on task-specific features relevant to the goal. Thereby, it minimizes redundancy and improves task efficiency. This highlights the potential of goal-oriented semantic communication in focusing on task-relevant content rather than data fidelity.

2.1 Approaches to Semantic Communication

Four approaches to semantic communication are mentioned in [WN23]. They are the Classical Semantic Information Approach, Knowledge Graph Approach, Machine Learning (ML) Approach, and Significance-Based Approach. The Information Theory Approach, inspired by Weaver, can be recognized as a fifth approach to semantic communication [XQLJ21].

2.1.1 Classical Semantic Information Approach

In traditional systems, Shannon’s model quantifies the amount of information based on probabilities. However, this approach focuses on logical probabilities to include semantics. Researchers like Carnap and Bar-Hillel proposed a way to calculate the semantic entropy of a message by measuring its consistency with a logical system. This approach evaluates how true or meaningful a message is within a specific context [WN23].

This method is foundational because it provides theoretical measures such as semantic entropy and semantic mutual information. However, it faces challenges when dealing with contradictions. For instance, a paradox or inconsistency might carry maximum semantic information

but could be meaningless in practical terms. Moreover, this approach depends on predefined logical structures. This makes it less flexible in dynamic and real-world scenarios where these structures might not exist [WN23].

2.1.2 Knowledge Graph Approach

The knowledge graph approach represents semantics by structuring knowledge into nodes and edges. Nodes represent concepts or entities and edges represent relationships between them. For instance, in a KG about animals, “dog” might be a node connected to another node, “mammal,” by the relationship “is a type of.” This structured representation enables machines to easily understand and utilize semantic information.

In this approach, shared knowledge bases between the sender and receiver ensure effective communication. Metrics like semantic similarity can measure how closely two concepts in the KG are related. This helps in tasks like reasoning or making inferences. For example, the system can determine that “cat” and “lion” are semantically closer than “cat” and “tree.” The KG approach is particularly useful because it provides a clear and interpretable representation of knowledge. However, it has its challenges. Both the sender and receiver need to have access to the same knowledge base. This can be difficult to synchronize in real-time communication. Additionally, creating and using large-scale KGs becomes computationally expensive and harder to manage as the amount of data grows [WN23].

Figure 2.2 illustrates a knowledge graph, which represents entities (nodes) and their relationships (edges). This graph-based representation encodes contextual and relational information about objects, concepts, or events. This makes it useful for interpreting and processing semantic data.

2.1.3 Machine Learning (ML) Approach

The ML approach uses artificial intelligence (AI) to learn semantic information directly from any data. In this approach, models like neural networks are trained on text, images, or speech to create representations of meaning. For example, natural language processing models like BERT or GPT analyze text to extract semantic embeddings, which capture the context and meaning of words or phrases. These embeddings can then be used for tasks like translating languages or summarizing documents [WN23].

This method is highly adaptable and can handle different types of data in dynamic environments. It does not require predefined rules, as the models learn relationships directly from different datasets. For example, ML models can process images to understand objects or analyze speech to extract the context behind spoken words [WN23].

Although the ML approach is powerful and flexible, it also has limitations. Training these models requires large datasets and computational resources. They often operate in a black-box manner. They rely on DNNs without explicit modeling. Additionally, the results are often hard to interpret. This makes it difficult to understand why a model made a certain decision. This can be a drawback in applications where explainability is essential [WN23].

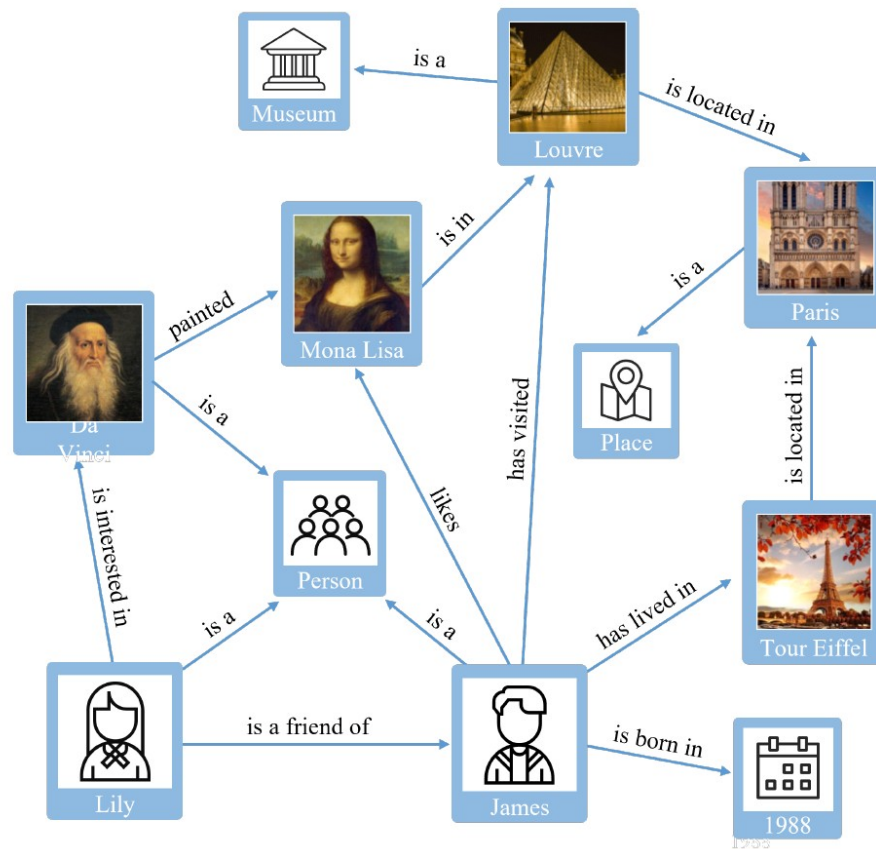


Figure 2.2: An example knowledge graph representing entities (nodes) and relationships (edges) [GQA⁺23].

2.1.4 Significance-Based Approach

The significance-based approach focuses on the importance or relevance of the transmitted information rather than its detailed meaning. In this method, communication systems prioritize data that is critical for achieving specific goals. For example, in a system designed for autonomous vehicles, only data relevant to avoiding obstacles or navigating roads would be transmitted and irrelevant information would be discarded [WN23].

This approach uses metrics such as "Age of Information," which measures how timely the data is. It ensures that only the most relevant and up-to-date information is communicated. This approach is especially effective in real-time systems where delays can have significant consequences [WN23].

The significance-based approach is efficient in reducing unnecessary data transmission. This is particularly useful in resource-limited environments. However, it has its own challenges. Determining what information is most relevant can vary widely depending on the application. For example, what is critical for a machine-learning model may not align with what a human user considers important. As a result, this approach often requires careful design to balance relevance and comprehensiveness [WN23].

2.1.5 Information Theory Approach

This approach extends Shannon's traditional information theory by integrating semantic and effectiveness layers into the communication framework, as proposed by Weaver [XQLJ21]. It moves beyond the technical level, which focuses on the accurate transmission of symbols. It includes two additional levels: the Semantic Layer (Level 2), which emphasizes the meaning of messages, and the Effectiveness Layer (Level 3), which focuses on how well the received meaning achieves the intended outcomes. By adapting Shannon's statistical probability models, this approach quantifies meaning and task effectiveness using measures such as semantic entropy and semantic mutual information. It aims to unify the technical, semantic, and effectiveness aspects of communication within a single theoretical framework [XQLJ21].

The Information Theory Approach provides a strong foundation for bridging the gap between traditional bit level and goal-oriented communication. It enables task-specific optimization. This makes it highly suitable for applications requiring both accurate meaning transmission and actionable outcomes, such as H2M or M2M interactions. However, implementing semantic and effectiveness measures in practice remains challenging due to the complexity of real-time semantic interpretation and the computational resources required. Despite these challenges, the approach offers significant potential for systems where understanding meaning and achieving specific goals are critical. By building directly on the theories of Shannon and Weaver, this approach adds a valuable dimension to the field of semantic communication [XQLJ21].

2.2 Existing Research Directions

Research in semantic communication is broadly categorized into two main areas: data reconstruction and task execution. Data reconstruction focuses on extracting semantic information at the transmitter and recovering it at the receiver using received semantic data [XQLJ21]. Initial investigations used ML approaches to reconstruct text, speech, and image sources [XQLJ21]. Subsequent studies extended the focus to explore communication concepts like efficiency and resource allocation [XBMMT24, TYW⁺21]. In contrast, task execution, also known as task-oriented communication, prioritizes encoding information relevant to a specific task. This ensures that the transmitted data contributes directly to the desired task outcome. For task-oriented communication, [SMZ22] proposed a communication method using the information bottleneck (IB) framework. This approach allows encoding information for a specific task while adapting to changing wireless channel conditions. In [SMZ23], the same authors explored a distributed approach to encode relevant information for collaborative feature extraction to accomplish a single task.

Despite significant advancements, several gaps persist in the field of semantic communication. Existing research has mainly focused on single-task-oriented communication. This leaves multi-task processing largely unexplored, especially from an information-theoretic perspective. Current multi-task approaches often rely heavily on ML models. This treats them as "black boxes" without sufficient theoretical analysis. These limitations hinder the development of efficient multi-task semantic communication systems [RBD24].

To address these gaps, the authors of [RBD24] proposed a probabilistic framework for modeling semantic sources. This model allows multiple semantic interpretations to be extracted simultaneously from a single observation. The paper also presented an innovative semantic

encoding structure that divides the encoder into a Common Unit (CU) for shared information and Specific Units (SUs) for task-specific processing. This structure allows for efficient cooperative task execution when tasks share statistical relationships, while also supporting independent processing when tasks are unrelated.

2.3 Cooperative Multi-Task Semantic Communication (CMT-SemCom) System

The system model presented in [RBD24] focuses on enabling cooperative multi-task processing for semantic communication. It incorporates a probabilistic semantic source, semantic encoding and decoding mechanisms. The model is designed to optimize task-specific performance. The semantic source is modeled using a probabilistic approach, where an observation serves as the input, semantic variables represent task-specific features, and tasks correspond to specific objectives. This structure allows the system to infer multiple semantic representations from a single observation.

The encoding process is divided into two components: a CU and SUs. The CU captures shared semantic information relevant to multiple tasks by performing cooperative processing. The SUs extract task-specific features tailored to individual tasks. This structure allows flexibility in handling both shared and task-specific semantic requirements. Encoded semantic information is transmitted over a noisy channel and the decoders at the receiver reconstruct task-relevant information.

The system supports both cooperative and independent task processing. Cooperative processing leverages the CU to share semantic information across tasks. This improves performance when tasks have statistical relationships. Conversely, independent processing is supported by the SUs. Overall, the system model provides a flexible framework to extract and transmit semantic features efficiently.

The semantic source is modeled to enable the simultaneous extraction of multiple semantic variables from a single observation. N independent tasks are considered, each associated with a unique semantic variable, denoted by $\mathbf{z} = [z_1, z_2, \dots, z_N]$. The observation \mathbf{S} is entailed with these semantic variables in a stochastic manner, forming a semantic source (\mathbf{z}, \mathbf{S}) , fully described by the probability distribution of $p(\mathbf{z}, \mathbf{S})$. The semantic source can be more precisely described by $p(\mathbf{z})p(\mathbf{S}|\mathbf{z})$. Here, $p(\mathbf{z})$ represents the prior distribution of semantic variables and $p(\mathbf{S}|\mathbf{z})$ reflects the semantic channel to link the semantic variables to the observation. Figure 2.3 represents a probabilistic graphical modeling of the semantic source.

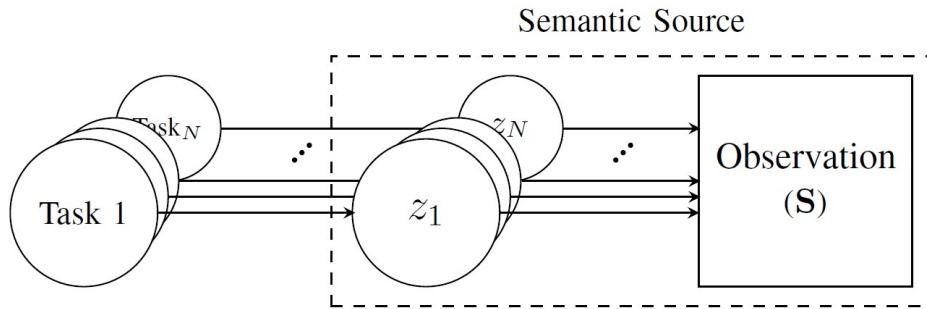


Figure 2.3: Probabilistic graphical modeling of the semantic source [RBD24].

This formulation allows the system to address multiple tasks simultaneously. For example, in an image containing both a tree and a number, one task may identify the presence of a tree (binary variable), while another may detect the number (multinomial variable).

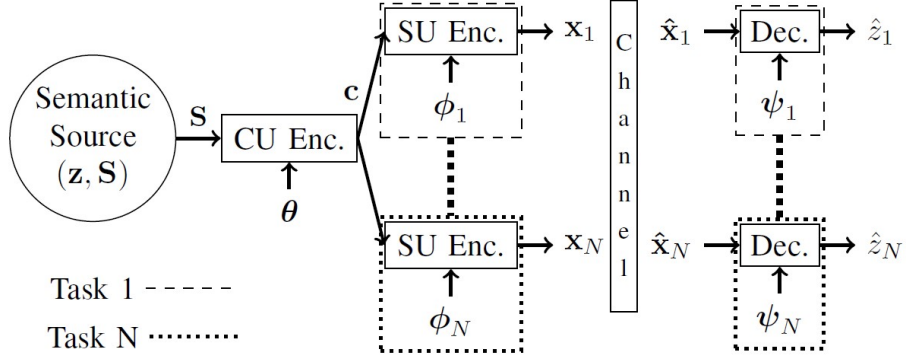


Figure 2.4: Cooperative multi-task semantic communication system model [RBD24].

Figure 2.4 depicts the system architecture for CMT-SemCom, showing the division of the encoder into a CU and SUs. The CU extracts shared semantic features from the input observation S , which are then passed to the task-specific SUs for further processing. The encoded task-specific outputs are transmitted over a noisy channel to the decoder.

The system leverages statistical relationships among semantic variables to enable cooperative multi-task processing. The semantic encoder is divided into:

1. **Common Unit:**

- Extracts information shared across tasks from the observation S .
- Modeled as $p^{\text{CU}}(\mathbf{c}|\mathbf{S})$, where \mathbf{c} represents the common encoded information.

2. **Specific Units:**

- Extract task-specific information from \mathbf{c} for each task.
- Modeled as $p^{\text{SU}_i}(\mathbf{x}_i|\mathbf{c})$, where \mathbf{x}_i is the encoded task-specific information for the i -th task.

The transmission is modeled using an Additive White Gaussian Noise (AWGN) channel. The received signal is:

$$\hat{\mathbf{x}}_i = \mathbf{x}_i + \mathbf{n}, \quad \mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I}), \quad (2.1)$$

where \mathbf{n} is the noise.

The probabilistic representation for the i -th semantic variable is:

$$p(\hat{z}_i, \hat{\mathbf{x}}_i, \mathbf{x}_i, \mathbf{c} | \mathbf{S}) = p^{\text{Dec}_i}(\hat{z}_i | \hat{\mathbf{x}}_i) p^{\text{Channel}}(\hat{\mathbf{x}}_i | \mathbf{x}_i) p^{\text{SU}_i}(\mathbf{x}_i | \mathbf{c}) p^{\text{CU}}(\mathbf{c} | \mathbf{S}). \quad (2.2)$$

This structure ensures the CU captures shared information, while the SUs focus on task-specific needs.

The optimization goal is to design the split semantic encoder architecture using the information maximization principle in an E2E learning framework. This framework has proven effective for task-oriented communication. The optimization problem is defined as:

$$[p^{\text{CU}}(\mathbf{c}|\mathbf{S})^*, p^{\text{SU}}(\mathbf{x}|\mathbf{c})^*] = \arg \max_{p^{\text{CU}}(\mathbf{c}|\mathbf{S}), p^{\text{SU}}(\mathbf{x}|\mathbf{c})} \sum_{i=1}^N b_i I(\hat{\mathbf{x}}_i; z_i), \quad (2.3)$$

where b_i is a constant coefficient fixed at 1, as the relationship or prioritization among semantic variables is not explored in this work. The objective is to maximize the mutual information between the channel output $\hat{\mathbf{x}}_i$ and the semantic variables z_i .

Expanding the mutual information, the approximated objective function can be expressed as:

$$L(\boldsymbol{\theta}, \boldsymbol{\Phi}) = \sum_{i=1}^N I(\hat{\mathbf{x}}_i; z_i) \approx \mathbb{E}_{p^{\text{CU}}_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{S})} \left[\sum_{i=1}^N \mathbb{E}_{p(\mathbf{S}, z_i)} \left[\mathbb{E}_{p^{\text{SU}}_{\boldsymbol{\Phi}_i}(\hat{\mathbf{x}}_i|\mathbf{c})} [\log p(z_i | \hat{\mathbf{x}}_i)] \right] \right], \quad (2.4)$$

where the semantic encoding is split into a CU and multiple SUs. The term $p^{\text{SU}}_{\boldsymbol{\Phi}_i}(\hat{\mathbf{x}}_i | \mathbf{c})$ accounts for joint semantic and channel coding. $\boldsymbol{\theta}$ represents the neural network (NN) parameters approximating the CU, and $\boldsymbol{\Phi}$ represents the NN parameters approximating the SU encoders. The outer expectation represents the cooperative multi-task processing enabled by sharing the CU.

The decoder for the i -th task is expressed as:

$$p^{\text{Dec}_i}(\hat{z}_i | \hat{\mathbf{x}}_i) = \frac{\int p^{\text{SU}}_{\boldsymbol{\Phi}_i}(\hat{\mathbf{x}}_i | \mathbf{c}) p^{\text{CU}}_{\boldsymbol{\theta}}(\mathbf{c} | \mathbf{S}) p(\mathbf{S}, z_i) ds dc}{p(\hat{\mathbf{x}}_i)}. \quad (2.5)$$

Due to the intractability of the high-dimensional integrals, a variational approximation technique is applied. This leads to the adjusted objective function:

$$L(\boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\Psi}) \approx \mathbb{E}_{p^{\text{CU}}_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{S})} \left[\sum_{i=1}^N \mathbb{E}_{p(\mathbf{S}, z_i)} \left[\mathbb{E}_{p^{\text{SU}}_{\boldsymbol{\Phi}_i}(\hat{\mathbf{x}}_i|\mathbf{c})} [\log q^{\text{Dec}_i, \boldsymbol{\Psi}_i}(z_i | \hat{\mathbf{x}}_i)] \right] \right] \quad (2.6)$$

where $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_N]$ represents the decoder's NN parameters, and $q^{\text{Dec}_i, \boldsymbol{\Psi}_i}(z_i | \hat{\mathbf{x}}_i)$ approximates the true distribution of the decoders.

The proposed architecture for semantic communication offers several key benefits, as highlighted in [RBD24]. The model efficiently leverages shared semantic information for tasks with statistical dependencies while maintaining task-specific encoding for independent tasks by dividing the encoder into a CU and SUs. This leads to optimized multi-task processing. This design reduces computational complexity by minimizing redundant processing, as the CU eliminates the need for duplicative feature extraction across tasks. The model provides lower task execution error rate compared to structures which do not have a CU. Also, faster improvement in task accuracy can be seen in case of using this structure.

In summary, Chapter 2 has explored the key paradigms in semantic communication from classical logical-probability measures and knowledge-graph representations to modern machine-learning embeddings, significance-based freshness metrics, and information-theoretic frameworks. It also highlighted both their individual merits and limitations. While most of these approaches have successfully addressed single-task scenarios, they fall short when multiple, interdependent tasks must be performed. The CMT-SemCom architecture addresses this gap by enabling cooperative, multi-task semantic processing in a single E2E framework.

Chapter 3

Non-Orthogonal Multiple Access (NOMA)

Wireless mobile communication has become an essential part of modern life. The development of modern wireless communication systems has led to an increased number of users. Moreover, the demand for faster data rates and better connectivity have increased. Traditional orthogonal multiple access (OMA) methods like frequency division multiple access (FDMA), time division multiple access (TDMA), code division multiple access (CDMA), and orthogonal frequency division multiple access (OFDMA) cannot meet these needs because they share resources like time slots, bandwidth, or codes among users in an orthogonal way. These techniques limit the number of users that can connect to the network. NOMA has emerged as a promising solution to achieve massive connectivity. It employs superposition coding at the transmitter to assign varying power levels to different users. At the receiver, the transmitted superimposed signal is separated using different techniques [Pea22].

Over time, multiple access techniques have evolved significantly. For 1G, FDMA assigned resources to users in the frequency domain. This ensured that no signal would overlap. For 2G, TDMA allocated resources in the time domain. Here, users communicated one after another in assigned time slots. For 3G, CDMA distributed resources in the code domain. For 4G, OFDMA introduced many modulated subcarriers placed orthogonally. These methods, called OMA, assign resources to users without overlap [DWY⁺15].

The fast development of wireless communication in 5G has created new demands, such as massive connectivity, low latency, high spectral efficiency, and diverse services. However, OMA techniques cannot handle a large number of users effectively because they rely on dividing limited resources among users. OMA struggles to provide the required connectivity and efficiency as the number of users grows [Pea22].

NOMA has been identified as a solution to meet these challenges regarding OMA [ATG⁺18]. NOMA allows more users to share the full spectrum, increasing spectral efficiency and supporting massive connectivity. Unlike OMA, NOMA's non-orthogonal resource allocation significantly improves network capacity [MCBA20]. However, separating user signals at the receiver requires more complex processing and introduces interference. NOMA also achieves a higher sum rate and capacity compared to OMA. By allowing non-orthogonal resource allocation, NOMA supports a larger number of users effectively [Pea22].

3.1 Working Principle of Power-domain NOMA

The basic techniques used in power-domain NOMA are superposition coding and decoding of the superimposed signals.

Superposition coding (SC) technique processes the data by assigning different power levels to each user, ensuring the system's requirements are met.

For example, two signals x_1 and x_2 are superimposed with two different power levels P_1 and P_2 respectively.

$$x = \sqrt{P_1} x_1 + \sqrt{P_2} x_2, \quad (3.1)$$

In the downlink, the base station (BS) combines the signals for various users linearly using superposition coding. In uplink NOMA, users transmit their signals simultaneously to the BS and the BS receives the superimposed signal.

The superimposed signal can be detected using 2 techniques. They are Successive Interference Cancellation (SIC) and DNN-Based Decoding Approach [Lea19].

In SIC, the receiver captures a composite signal containing all user transmissions.

Each user- i observes

$$y_i = h_i x + w_i, \quad (3.2)$$

where h_i is the complex channel gain to user- i , and w_i is gaussian noise plus inter-cell interference with power density $N_{0,i}$ [SKB⁺13].

Users are ordered by their normalized channel gains $\frac{|h_1|^2}{N_{0,1}} \geq \frac{|h_2|^2}{N_{0,2}} \geq \dots$. The weaker user (with the smaller normalized gain) decodes its own signal directly, treating the other user's signal as noise. Then the stronger user first decodes and subtracts the weaker user's transmission, then decodes its own [SKB⁺13].

In DNN-Based Decoder, the decoder is trained using a large dataset of labeled examples, including composite received signals and their corresponding user messages. The loss function is minimized to improve decoding accuracy [Lea19, LCC⁺21].

The traditional SIC receiver is built as a pipeline of distinct modules. First, a channel estimator obtains channel state information (CSI). Then users are ordered by descending received power. After that, an iterative loop decodes the strongest user, reconstructs its signal, and subtracts it from the composite. On the other hand, in a DNN-based Decoder, DNN replaces all separate blocks. It takes the received superposed signal as input and each users' signal separately. The network layers are trained to cancel interference without any explicit subtraction step [Lea19].

Figure3.1 shows how signals are sent and received via an uplink NOMA system.

3.2 Classification

3.2.1 Based on the multiplexing technique

NOMA schemes are generally divided into two main categories: power-domain multiplexing and code-domain multiplexing.

In power-domain multiplexing, users are assigned different power levels based on their channel conditions. At the transmitter, the information signals from multiple users are superimposed into a single signal. At the receiver, the superimposed signal is decoded for each user. This approach strikes a good balance between system throughput and user fairness [DWD⁺18].

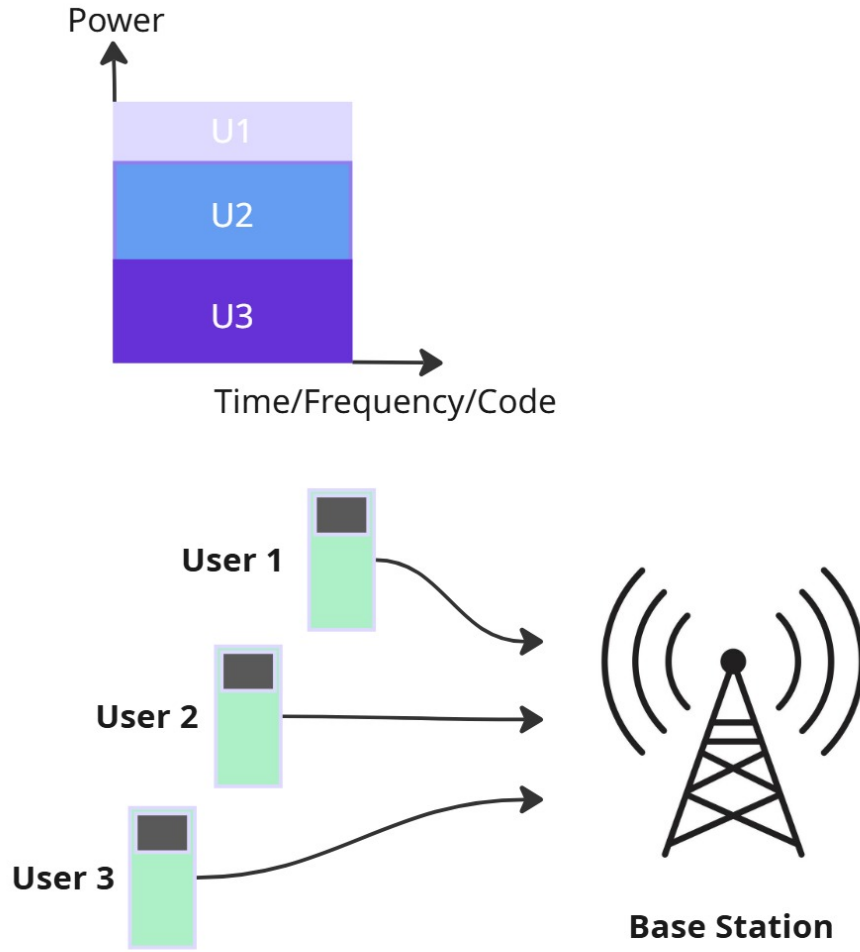


Figure 3.1: Working principle of a 3-user uplink power-domain NOMA[Pea22].

Code-domain multiplexing embeds each user's data in a higher-dimensional sparse code or signature space. Users transmit over the same time–frequency resources but with distinct sparse codewords. While code-domain multiplexing can improve spectral efficiency, it typically requires higher transmission bandwidth. Moreover, it is not easily compatible with existing networks [DWD⁺18].

On the other hand, power-domain multiplexing is simpler to implement and requires minimal changes to current systems. Moreover, it enhances spectral efficiency without the need for additional bandwidth like code-domain multiplexing[DWD⁺18].

3.2.2 Based on architecture

The NOMA architecture can be two types: Uplink NOMA and Downlink NOMA [Pea22, ATG⁺18].

In the downlink NOMA system, the base station (BS) employs superposition coding to combine messages intended for multiple users with different power level into a single transmit signal. Users with poorer channel conditions are allocated higher power coefficients

[Pea22, ATG⁺18].

Each user decodes its message using a decoding technique, such as successive interference cancellation (SIC). In SIC, the far user directly decodes its signal by treating the other users' signals as interference. Next user applies SIC to remove far user's signal before decoding its own. Similarly other users remove previous users' signals before decoding its message [Pea22, ATG⁺18].

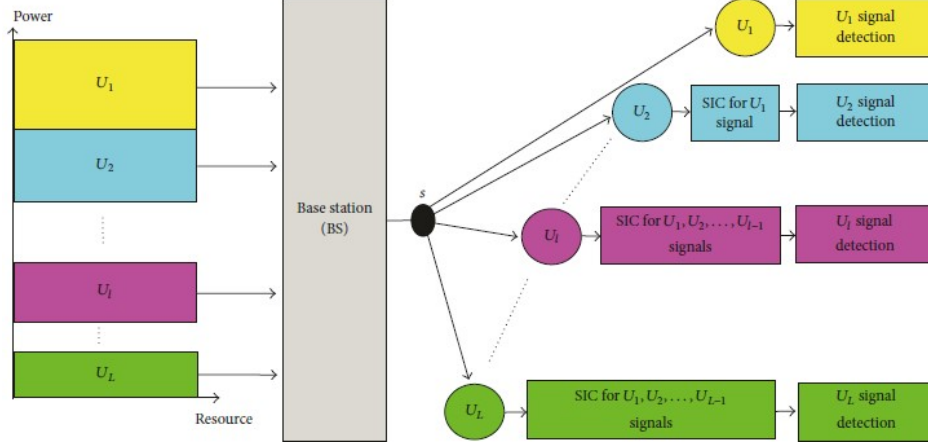


Figure 3.2: Signal transmission and decoding technique in Downlink NOMA [ATG⁺18].

Figure 3.2 shows the signal transmission and decoding technique in Downlink NOMA.

In the uplink NOMA system, users transmit their signals simultaneously to the BS with different power levels. Powers are usually predefined and set by the base station and the information is sent to each user. The BS which performs SIC to decode each user's message.

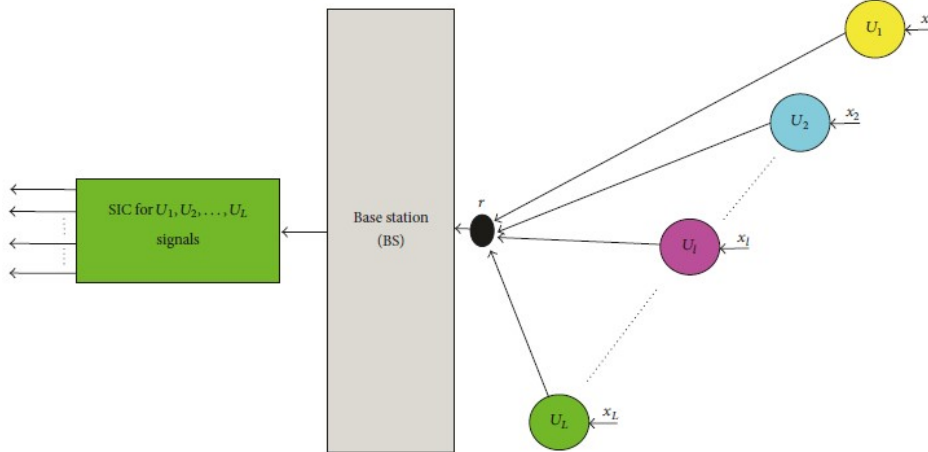


Figure 3.3: Signal transmission and decoding technique in Uplink NOMA [ATG⁺18].

Figure 3.3 shows the signal transmission and decoding technique in Uplink NOMA.

Uplink NOMA maps perfectly to task-oriented communications and edge-inference setups. This thesis considers edge devices which transmit their task-relevant information and a unified DNN-based Decoder that performs cooperative processing on the receiver side. This scenario can be perfectly implemented with the help of Uplink NOMA architecture.

3.3 Advantages

NOMA offers several advantages over traditional OMA techniques, making it a strong candidate for next-generation wireless communication systems. These benefits can be summarized as follows [ATG⁺18]:

1. **Higher Spectral Efficiency and Throughput:** In OMA, such as OFDMA, each user is assigned a specific frequency resource, regardless of their channel condition. This results in inefficient resource usage and lower overall system performance.

NOMA allows multiple users with varying channel conditions to share the same frequency resource simultaneously. By assigning weaker users higher power and mitigating interference in the decoding process, NOMA improves spectral efficiency and overall throughput.

2. **Enhanced User Fairness, Reduced Latency, and Massive Connectivity:** OMA often prioritizes users with better channel conditions. This causes users with poorer conditions to wait, resulting in fairness issues and higher latency.

NOMA addresses this by enabling simultaneous service for multiple users with different channel conditions. This approach ensures better user fairness, reduces latency, and supports massive connectivity.

3. **Compatibility with Existing and Future Systems:** NOMA integrates seamlessly with current and upcoming communication architectures, requiring minimal changes to the existing infrastructure.

4. **Flexibility in Applications:** NOMA is designed to meet the diverse requirements of 5G use cases, including enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable low-latency communication (URLLC).

In summary, Chapter 3 has established NOMA as a powerful alternative to traditional orthogonal schemes. It enables simultaneous transmission of multiple user signals through power-domain superposition and decoding techniques. We reviewed how superposition coding and successive interference cancellation or modern DNN-based decoders allow a base station to separate overlapped signals. We classified NOMA both by its multiplexing principles (power- vs. code-domain) and by deployment scenarios (uplink vs. downlink). NOMA delivers higher spectral efficiency, improved user fairness, and enhanced support for massive connectivity. With these advantages in mind, we are now well positioned to integrate semantic communication concepts into an uplink NOMA framework.

Chapter 4

CMT-SemCom Enabled by NOMA

The rapid advancements in semantic communication have created opportunities for data transmission with task-specific goals, optimizing resource utilization and system efficiency. NOMA enhances physical-layer connectivity by allowing simultaneous transmissions for multiple users through power-domain multiplexing. However, several research challenges remain unresolved, particularly in the context of integrating semantic communication, which represents the higher layers of communication over a network with advanced physical layer techniques such as NOMA.

For edge devices, the resource-intensive cooperative processing at the transmitter side is often impractical. These devices are constrained by limited resources. For example, in an edge inference scenario where the cameras or sensors send only distilled feature vectors may have limited power and computational resources. To solve this, the cooperative processing can be shifted to the receiver side using a DNN-based decoder in an uplink NOMA architecture. To fully assess the trade-offs and gains of this new architecture, we need to build a complete design and study the physical-layer framework. By employing a single, E2E trained DNN-based NOMA decoder, we can perform cooperative processing and all PHY-layer functions can be described by this Uplink NOMA structure.

Key research questions that drive this work are:

1. Can the CMT-SemCom structure maintain efficient multi-tasking when integrated with the multi-access physical layer technique?
 - The CMT-SemCom structure is a method for multi-task semantic communication. This research will evaluate whether this structure can maintain its efficiency when combined with NOMA, which introduces additional constraints like power allocation and superimposition.
2. What is the impact of replacing the cooperative processing from the transmitter side to the receiver side with the help of DNN-based NOMA?
 - Shifting cooperative processing from the transmitter side to the receiver side can alleviate the resource burden on edge devices. The study will assess whether this change helps edge devices without compromising the overall system efficiency.
3. What is the impact of using cooperative processing on both the transmitter and receiver sides?
 - This research will explore whether cooperative processing at both ends can offer performance improvements.

To address these challenges, an architecture is proposed that combines the CMT-SemCom

structure of [RBD24] with DNN-NOMA to build an E2E communication network. The proposed model leverages:

- A shared CU) for extracting semantic information,
- Task-specific SUs for encoding, and
- NOMA-based signal transmission to enable efficient multi-tasking over a shared communication channel.
- A DNN NOMA-based base station decoder for reconstructing the task-specific semantic variables.

4.1 System Model

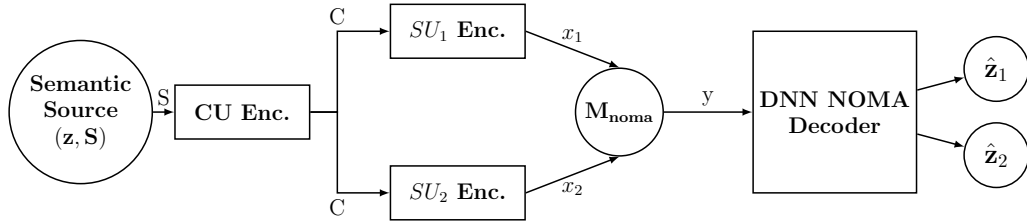


Figure 4.1: Block diagram of the proposed CMT-SemCom Enabled by DNN-NOMA framework.

Figure 4.1 illustrates the architecture of the proposed semantic communication system combining CMT-SemCom structure with NOMA. The system has a CU Encoder, which extracts shared semantic features from the input source consisting of semantic variables $[z_1, z_2, \dots, z_N]$. These features are then processed by task-specific SU Encoders to generate task-specific signals $[x_1, x_2, \dots, x_N]$. The outputs of the SU encoders are superimposed into a single signal (M_{noma}) using NOMA-based power-domain multiplexing. The superimposed signal is transmitted through a noisy communication channel, received as y at the base station. A DNN-based Base Station Decoder processes y to reconstruct the semantic variables $[z_1, z_2, \dots, z_N]$ corresponding to their respective tasks. This architecture efficiently integrates semantic communication principles with NOMA for multi-task processing.

Similar to the CMT-SemCom system model in 2.3, the semantic source is modeled as a tuple (z, S) .

4.1.1 System Probabilistic Modeling

The proposed architecture leverages a probabilistic framework that integrates shared and task-specific encoding processes with NOMA to efficiently transmit and reconstruct semantic information for multiple tasks. This section provides a detailed description of the probabilistic modeling used in the system, which encompasses the CU Encoder, SU Encoders, NOMA-based superimposition, and DNN-based decoding.

The CU encoder extracts shared semantic features (c) from the input observation (S). These shared features process the common information of all tasks and are modeled probabilistically as:

$$p^{\text{CU}}(c|S),$$

where \mathbf{c} denotes the common relevant information. The CU encoder maximizes the mutual information between the input observation (\mathbf{S}) and the shared features (\mathbf{c}), ensuring that \mathbf{c} retains the maximum relevant information for the task-specific encoders.

The shared features (\mathbf{c}) are further refined by task-specific encoders, referred to as SUs. Each SU processes the shared features to produce task-specific encoded outputs ($\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$). For the i -th task, the SU encoder is modeled as:

$$p^{\text{SU}_i}(\mathbf{x}_i|\mathbf{c}), \quad i \in \{1, 2, \dots, N\}.$$

The task-specific encoded signals ($\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$) are superimposed using power-domain NOMA. The superimposed signal, \mathbf{M}_{noma} , is a linear combination of the encoded signals with different power levels:

$$\mathbf{M}_{\text{noma}} = \sqrt{q_1}\mathbf{x}_1 + \sqrt{q_2}\mathbf{x}_2 + \dots + \sqrt{q_N}\mathbf{x}_N, \quad (4.1)$$

Here, q_i is the power allocation coefficients for i -th task.

This superimposition technique allows simultaneous transmission of task-specific signals over the same channel, leveraging NOMA to efficiently utilize available resources. The superimposed signal (\mathbf{M}_{noma}) is transmitted over an Additive White Gaussian Noise (AWGN) channel. The received signal (\mathbf{y}) at the base station is modeled as:

$$\mathbf{y} = \mathbf{M}_{\text{noma}} + \mathbf{n}, \quad (4.2)$$

where $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2)$ represents Gaussian noise with zero mean and variance σ^2 . The channel introduces noise to the transmitted signal, which the decoder must account for during reconstruction.

The received signal (\mathbf{y}) is processed by a DNN-based decoder at the base station. The decoder's goal is to extract semantic vector $\hat{\mathbf{z}} = [\hat{z}_1, \hat{z}_2, \dots, \hat{z}_N]$ from \mathbf{y} . Here, \hat{z}_i represents the reconstructed semantic variable for Task i . The decoding process is modeled probabilistically as:

$$p^{\text{Dec}}(\hat{\mathbf{z}}|\mathbf{y}), \quad i \in \{1, 2, \dots, N\},$$

The DNN architecture includes shared hidden layers to process the received signal (\mathbf{y}) and extract features shared across tasks. It also has task-specific output layers for each task to produce the task-specific outputs.

The decoder is trained to minimize the reconstruction error, ensuring accurate recovery of the semantic variables.

The entire system, from input to output, is modeled probabilistically with the help of Markov representation. The joint probability distribution for Task i is:

$$p(\hat{\mathbf{z}}, \mathbf{y}, \mathbf{M}_{\text{noma}}, \mathbf{x}_i, \mathbf{c}|\mathbf{S}) = p^{\text{Dec}}(\hat{\mathbf{z}}|\mathbf{y})p^{\text{Channel}}(\mathbf{y}|\mathbf{M}_{\text{noma}})p^{\text{Sup}}(\mathbf{M}_{\text{noma}}|\mathbf{x})p^{\text{SU}_i}(\mathbf{x}_i|\mathbf{c})p^{\text{CU}}(\mathbf{c}|\mathbf{S}), \quad (4.3)$$

where:

- $p^{\text{Dec}}(\hat{\mathbf{z}}|\mathbf{y})$: Decoder to reconstruct $\hat{\mathbf{z}}$ where \mathbf{y} is the received information passed through the channel,
- $p^{\text{Channel}}(\mathbf{y}|\mathbf{M}_{\text{noma}})$: AWGN Channel model incorporating noise,
- $p^{\text{Sup}}(\mathbf{M}_{\text{noma}}|\mathbf{x})$: Superimposition of all the task specific encoder outputs denoted as \mathbf{x} ,

- $p^{\text{SU}_i}(\mathbf{x}_i|\mathbf{c})$: Task-specific encoder distribution, extracting task-specific information and providing the channel input,
- $p^{\text{CU}}(\mathbf{c}|\mathbf{S})$: CU that extracts the common relevant information amongst all tasks, from the observation.

The probabilistic modeling framework integrates shared and task-specific processing with NOMA for efficient multi-task semantic communication. The CU and SU encoders enable hierarchical representation of semantic information, while the NOMA-based superimposition optimizes resource utilization. The DNN NOMA-based decoder ensures accurate reconstruction of semantic variables. With this integration, cooperative processing is also employed on the receiver side through the DNN-NOMA architecture. Furthermore, the simulation results examine scenarios where employing cooperative processing on the transmitter side is not feasible and cooperative processing is performed only on the receiver side.

4.1.2 Optimization Problem

The information maximization principle together with the E2E learning manner has been adopted to formulate an optimization problem for our semantic architecture. The optimization problem aims to maximize the mutual information between the received signal (\mathbf{y}) and the semantic variables (z_i) for all tasks. It is defined as:

$$[p^{\text{CU}}(\mathbf{c}|\mathbf{S})^*, p^{\text{SU}}(\mathbf{x}|\mathbf{c})^*] = \arg \max_{p^{\text{CU}}(\mathbf{c}|\mathbf{S}), p^{\text{SU}}(\mathbf{x}|\mathbf{c})} \sum_{i=1}^N I(\mathbf{y}; z_i), \quad (4.4)$$

where $I(\mathbf{y}; z_i)$ is the mutual information between the received signal (\mathbf{y}) and the semantic variable (z_i).

This formulation ensures that the CU and SU encoders maximize information transfer efficiency for all tasks while minimizing redundancy. The approximated objective function is derived expanding the mutual information in our optimization problem.

Using the objective function:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\Phi}) = \sum_{i=1}^N I(\mathbf{y}; z_i) \quad (4.5)$$

$$= \sum_{i=1}^N \iint p(\mathbf{y}, z_i) \log \frac{p(z_i|\mathbf{y})}{p(z_i)} dz_i dy \quad (4.6)$$

$$= \sum_{i=1}^N \left[\iint p(\mathbf{y}, z_i) \log p(z_i|\mathbf{y}) dz_i dy + H(z_i) \right] \quad (4.7)$$

Further ignoring the constant term $H(z_i)$ and leveraging the Markov chain relationship:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\Phi}) \approx \sum_{i=1}^N \int \int \int \int \int p(z_i, \mathbf{S}) p_{\boldsymbol{\theta}}^{\text{CU}}(\mathbf{c}|\mathbf{S}) p_{\phi_i}^{\text{SU}}(\mathbf{x}_i|\mathbf{c}) p^{\text{Sup}}(\mathbf{M}_{\text{noma}}|\mathbf{x}) \\ p^{\text{Channel}}(\mathbf{y}|\mathbf{M}_{\text{noma}}) \log p(z_i|\mathbf{y}) dz_i d\mathbf{S} d\mathbf{c} d\mathbf{x}_i d\mathbf{M}_{\text{noma}} dy \end{aligned} \quad (4.8)$$

$$\approx \sum_{i=1}^N \int \int \int \int p(z_i, \mathbf{S}) p_{\theta}^{CU}(\mathbf{c}|\mathbf{S}) p_{\phi_i}^{SU}(\mathbf{y}|\mathbf{c}) \log p(z_i|\mathbf{y}) dz_i d\mathbf{S} d\mathbf{c} d\mathbf{y} \quad (4.9)$$

$$\approx \mathbb{E}_{p_{\theta}^{CU}(\mathbf{c}|\mathbf{S})} \left[\sum_{i=1}^N \left\{ \mathbb{E}_{p(\mathbf{S}, z_i)} \left[\mathbb{E}_{p_{\phi_i}^{SU}(\mathbf{y}|\mathbf{c})} [\log p(z_i|\mathbf{y})] \right] \right\} \right] \quad (4.10)$$

We employ the variational method, which is a common approach to simplify complex computations by introducing adjustable parameters, such as weights in neural networks (NNs). This technique is widely applied in machine learning and task-oriented communication research [RBD24].

Our posterior distributions, $p^{CU}(\mathbf{c}|\mathbf{S})$ and $p^{SU}(\mathbf{x}|\mathbf{c}) = [p^{SU_1}(\mathbf{x}_1|\mathbf{c}), \dots, p^{SU_N}(\mathbf{x}_N|\mathbf{c})]$, are approximated using NNs. This results in $p_{\theta}^{CU}(\mathbf{c}|\mathbf{s})$ and $p_{\phi_i}^{SU}(\mathbf{x}_i|\mathbf{c})$, where θ represents the NN parameters for the CU, and $\Phi = [\phi_1, \dots, \phi_N]$ represents the parameters for the SU encoders.

The channel outputs are used to emphasize the role of joint semantic and channel coding handled by the SUs. Using

$$p_{\phi_i}^{SU}(\mathbf{y}|\mathbf{c}) = \int p_{\phi_i}^{SU}(\mathbf{x}_i|\mathbf{c}) p^{Channel}(\mathbf{y}|\mathbf{x}_i) d\mathbf{x}_i, \quad (4.11)$$

we aim to optimize $p_{\phi_i}^{SU}(\mathbf{y}|\mathbf{c})$.

The objective function demonstrates the E2E design approach, where both encoders and decoders are optimized together. The AWGN channel is included directly in this design because its transfer function is differentiable. Additionally, the outer expectation distinguishes this approach from single-task processing by incorporating cooperation among the SU blocks, achieved by sharing the CU.

Regarding the decoder in the objective function, the $p^{Dec}(\hat{\mathbf{z}}|\mathbf{y})$ can be fully determined using the known distributions and underlying probabilistic relationship as:

$$p^{Dec}(\hat{\mathbf{z}}|\mathbf{y}) = \frac{\int p_{\phi_i}^{SU}(\mathbf{y}|\mathbf{c}) p_{\theta}^{CU}(\mathbf{c}|\mathbf{S}) p(\mathbf{S}, z_i) d\mathbf{s} d\mathbf{c}}{p(\mathbf{y})} \quad (4.12)$$

However, due to the high-dimensional integrals, this equation becomes intractable, and we need to follow the variational approximation technique, resulting in the following:

$$\mathcal{L}(\theta, \Phi, \Psi) \approx \mathbb{E}_{p_{\theta}^{CU}(\mathbf{c}|\mathbf{S})} \left[\sum_{i=1}^N \left\{ \mathbb{E}_{p(\mathbf{S}, z_i)} \left[\mathbb{E}_{p_{\phi_i}^{SU}(\mathbf{y}|\mathbf{c})} [\log q_{\psi_i}^{Dec}(\hat{\mathbf{z}}|\mathbf{y})] \right] \right\} \right] \quad (4.13)$$

Here, $\Psi = [\psi_1, \dots, \psi_N]$ represents NNs approximating the true distribution of decoders.

The system's training framework involves the following steps:

1. **Sampling:** Observations (\mathbf{S}) and semantic variables (z_1, z_2, \dots, z_N) are sampled from the dataset.
2. **Encoding:** The CU encoder generates shared features (\mathbf{c}), which are refined into task-specific signals ($\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$) by the SU encoders.

3. **Transmission:** Task-specific signals are superimposed via NOMA and transmitted through the channel, resulting in the received signal (\mathbf{y}).
4. **Decoding:** The DNN-based decoder reconstructs the semantic variables ($\hat{z}_1, \hat{z}_2, \dots, \hat{z}_N$).

This optimization framework ensures efficient encoding, transmission, and decoding of semantic information while leveraging NOMA for resource optimization.

4.2 Simulation Results of CMT-SemCom Enabled by NOMA

4.2.1 Simulation Setup

The simulation setup for the proposed architecture is designed to validate its performance in multi-task semantic communication. For this purpose, the MNIST dataset was selected. It provides a well-defined benchmark for image-based classification tasks. The dataset contains 60,000 training samples and 10,000 test samples of handwritten digits ranging from 0 to 9, each in grayscale and with a resolution of 28×28 pixels.

Two semantic tasks are defined: Task 1 is a binary identification problem that detects whether a given digit is a “2”, Task 2 is a categorical classification task that identifies the actual digit class from 0 to 9. These tasks are designed to reflect both simple and complex tasks. This enables the evaluation of the system’s multi-task capabilities.

The neural network architecture employed for this simulation comprises three primary components: a CU encoder, task-specific SU encoders, and a DNN-NOMA decoder. The CU encoder is responsible for extracting features that are commonly relevant across all tasks. It consists of a fully connected layer with 64 units and uses a Tanh activation function. This shared representation is then passed to the two SU encoders. Each SU encoder is dedicated to a specific task and includes a fully connected layer with 16 units and Tanh activation.

At the receiver side, the system uses a DNN-based decoder that reconstructs the semantic variables from the received signal. The decoder includes two fully connected layers with 64 ReLU-activated units. It produces two outputs: a binary output for Task 1 using a Sigmoid activation function and a categorical output for Task 2 using a Softmax activation function.

The CU and SU encoders, along with the decoder, are trained jointly in an E2E fashion. A composite loss function is used, consisting of binary cross-entropy loss for the binary task and categorical cross-entropy loss for the categorical task. The optimization is carried out using the Adam optimizer. Throughout the training process, the system is evaluated based on task execution error rate.

To provide a clear understanding of the cases studied and their comparisons, we have designed two summary tables for Task 1 and Task 2. These tables visually represent the corresponding line styles used in the simulation results.

Table 4.1 highlights the configurations for Task 1, where the line styles are represented as dotted lines with distinct colors for each case.

Table 4.2 presents the configurations for Task 2, where the line styles are solid lines with distinct colors for each case.

The results are analyzed by examining the task error rates over multiple training epochs. These results help quantify the benefits of cooperative processing through the CU at the

Table 4.1: Simulation line styles for Task 1

Category	Line Style
WithoutCU Non-NOMA
WithCU Non-NOMA
WithoutCU DNN-NOMA
WithCU DNN-NOMA
WithCU SIC-NOMA

Table 4.2: Simulation line styles for Task 2

Category	Line Style
WithoutCU Non-NOMA	———
WithCU Non-NOMA	———
WithoutCU DNN-NOMA	———
WithCU DNN-NOMA	———
WithCU SIC-NOMA	———

transmitter side and DNN-NOMA Decoder at the receiver side.

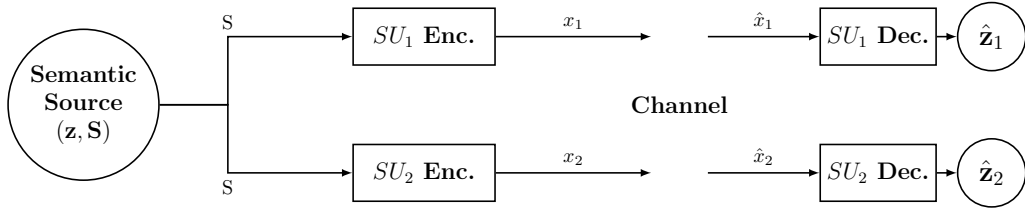


Figure 4.2: Block diagram of the proposed non-cooperative task oriented semantic communication system (without CU and without DNN-NOMA).

Figures 4.2, 4.3, 4.4 and 4.5 illustrate the four transmitter–receiver topologies evaluated in the simulation campaign. All share the same semantic-source model and task definitions described earlier.

Non-Cooperative Task Oriented Semantic Communication System (Without-CU, Non-NOMA)

In Figure 4.2, the semantic source directly goes to each SU encoder. No common semantic representation is extracted, so there is no cooperative processing at the transmitter side. The two encoded vectors are sent through channel. At the base station, two independent decoders reconstruct z_1 and z_2 . So, there is no cooperative processing on the receiver side as well.

Tx-Side Cooperative Multi-Tasking Semantic Communication System (With-CU, Non-NOMA)

In Figure 4.3, the observation first passes through a CU that extracts features shared by both tasks. Task-specific SUs then refine these shared features. Transmission is still orthogonal

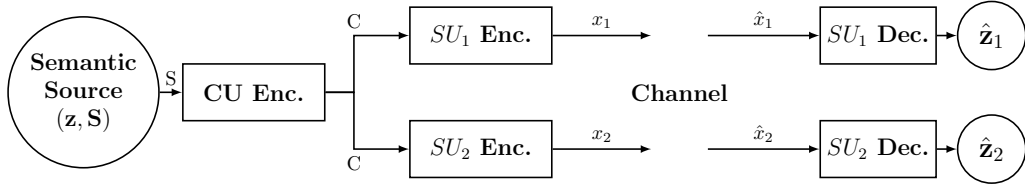


Figure 4.3: Block diagram of the proposed Tx-side cooperative multi-tasking semantic communication system (with CU and without DNN-NOMA).

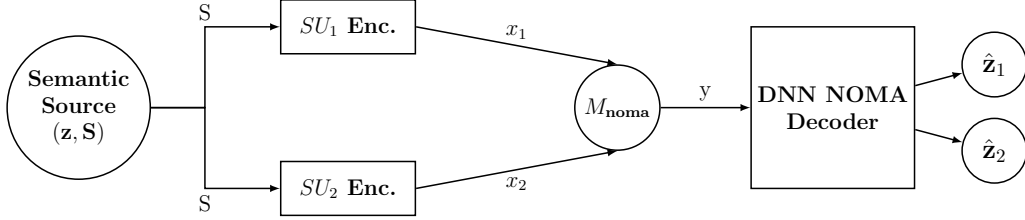


Figure 4.4: Block diagram of the proposed Rx-side cooperative multi-tasking semantic communication system (without CU and with DNN-NOMA).

(Non-NOMA), so the receiver applies two separate decoders. This setting has cooperative processing on the transmitter side only.

Rx-Side Cooperative Multi-Tasking Semantic Communication System (Without-CU, DNN-NOMA)

In Figure 4.4, the semantic source directly goes to each SU encoder. No common semantic representation is extracted, so there is no cooperative processing at the transmitter side. At the receiver side, it has the DNN-NOMA Decoder to decode the superimposed signal. As a result, this setting has cooperative processing on the receiver side.

Both Tx-Side and Rx-Side Cooperative Multi-Tasking Semantic Communication System (With-CU, DNN-NOMA)

Figure 4.5, the observation first passes through a CU that extracts features shared by both tasks. Task-specific SUs then refine these shared features. At the receiver side, it has the DNN-NOMA Decoder to decode the superimposed signal. As a result, this setting has cooperative processing both on the transmitter and receiver side.

4.2.2 Case Study

To evaluate the effectiveness of the designed architecture, the system was tested under various scenarios and configurations to observe its performance. These cases examined the architecture's ability to handle different tasks and environmental conditions, such as varying Signal-to-Noise Ratios (SNR), with and without the CU, and the influence of DNN NOMA

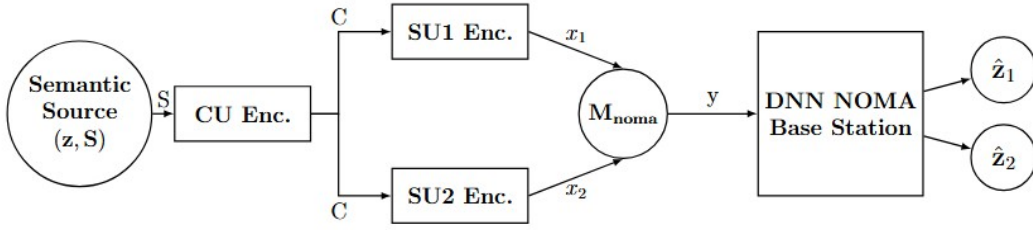


Figure 4.5: Block diagram of both Tx-side and Rx-side cooperative multi-tasking semantic communication system (with CU and DNN-NOMA).

integration. The observations will focus on Task Error Rate across binary and categorical tasks.

Case 1: “With CU” Vs. “Without CU” for Non-NOMA

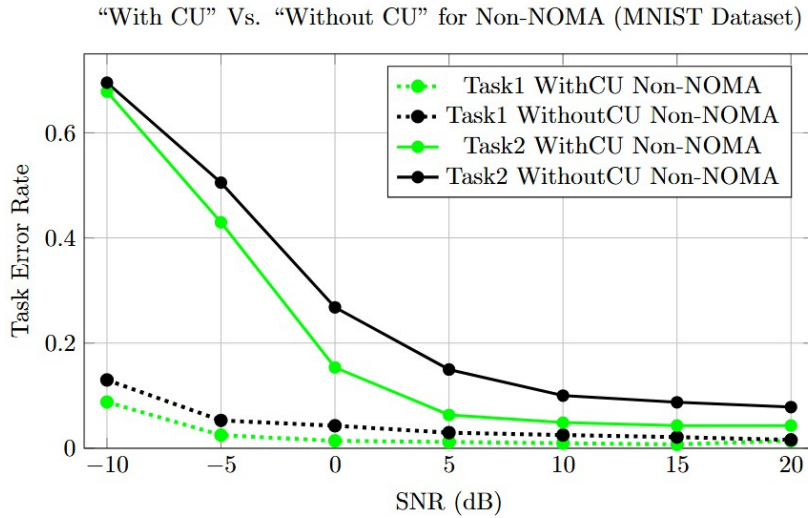


Figure 4.6: Performance comparison for “With CU” Vs. “Without CU” for Non-NOMA.

The "With CU" setup consistently exhibits lower task error rate values compared to "Without CU" for both Task1 and Task2. The difference between the two configurations is more pronounced for Task2 which is the Categorical Identification of numbers, where cooperative processing plays a critical role in reducing errors. For Task2, the Task error rate for "With CU" decreases rapidly with increasing SNR.

Figure 4.6 shows the comparison between the “With CU” and “Without CU” setup for Non-NOMA architecture.

The "With CU" architecture showcases clear advantages in Task error rate. For binary tasks, the performance gap between "With CU" and "Without CU" is smaller, due to the simpler nature of the task. For categorical tasks, the cooperative processing enabled by the CU significantly improves performance, as it benefits from shared semantic representations. These results underscore the effectiveness of incorporating a CU for cooperative multi-task processing, making the system more resilient to channel noise and better suited for complex tasks.

Case 2: “With CU” Vs. “Without CU” for NOMA

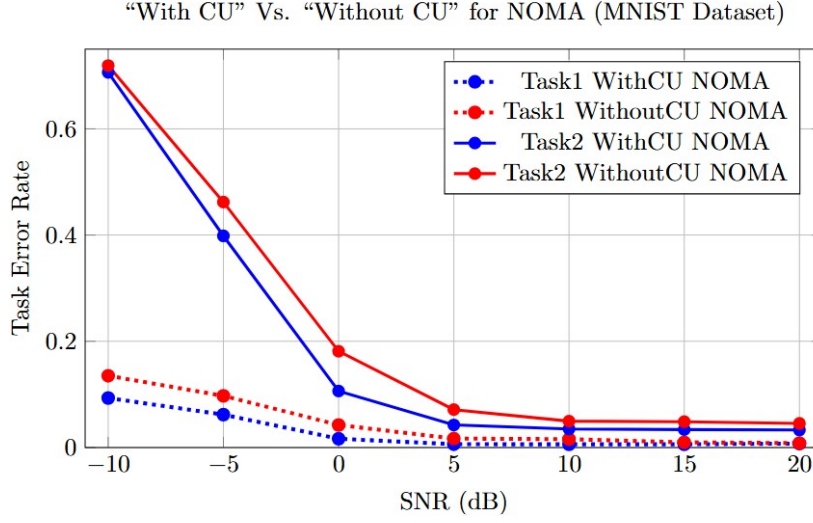


Figure 4.7: Performance comparison for “With CU” Vs. “Without CU” for NOMA.

The "With CU" setup consistently exhibits lower Task error rate across all SNR values compared to "Without CU." for both Task1 and Task2. The difference between the two configurations is more pronounced at low SNRs, where cooperative processing plays a critical role in reducing errors. For Task2, the Task error rate for "With CU" decreases rapidly with increasing SNR.

Figure 4.7 shows the comparison between the “With CU” and “Without CU” setup for NOMA-based architecture.

The "With CU" architecture showcases advantages in Task error rate. These results underscore the effectiveness of incorporating a CU for cooperative processing, making the system better suited for complex tasks.

Case 3: NOMA Vs. Non-NOMA for “With CU” Vs. “Without CU”

The gap between "With CU" and "Without CU" for complex Task2 is larger for Non-NOMA than for NOMA.

Figure 4.8 shows the comparison between the gap between “With CU” and “Without CU” setup for NOMA-based Vs. Non-NOMA architectures.

DNN-NOMA reduces the dependency on CU, narrowing the performance gap between "With CU" and "Without CU" setups. This highlights DNN-NOMA’s ability to handle tasks independently while still benefiting from CU when available. CU provides a substantial improvement for both setups, but its impact is more significant in Non-NOMA scenarios, where task performance without CU suffers more.

Case 4: NOMA Vs. Non-NOMA for “With CU”

In the very low SNR region (-10 dB), the Task error rate of the NOMA-based system is higher than that of the Non-NOMA system, due to the inherent complexity of NOMA in handling

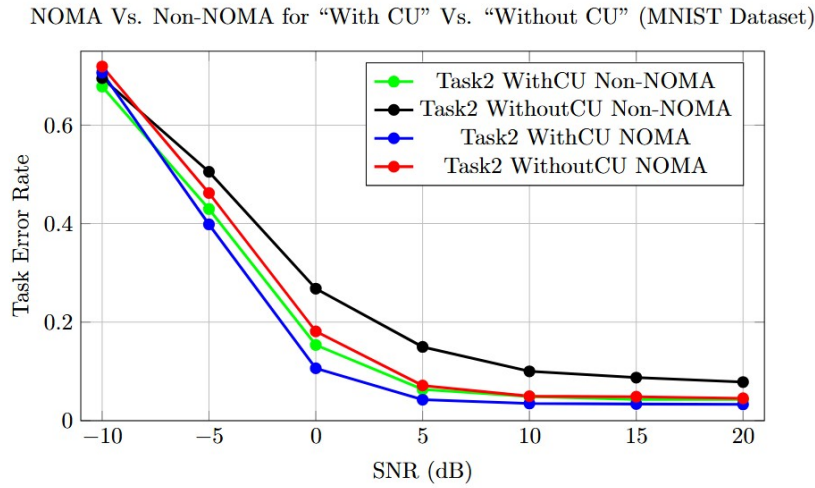


Figure 4.8: Performance comparison of NOMA Vs. Non-NOMA for “With CU” and “Without CU”.

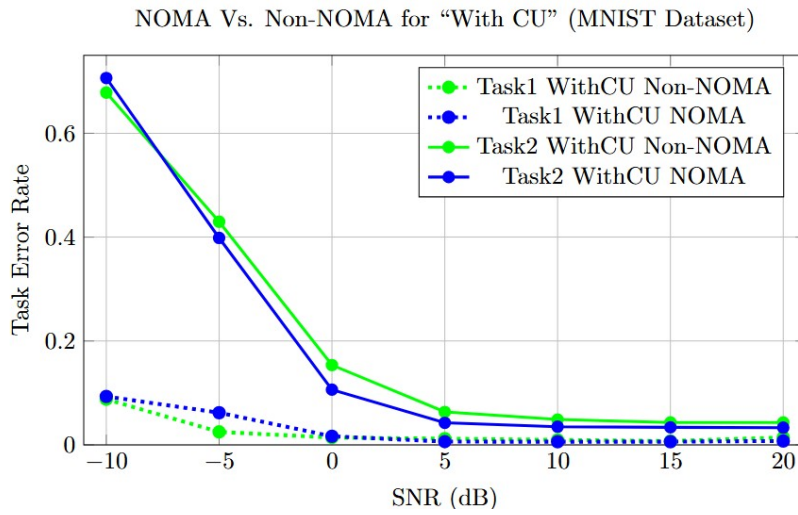


Figure 4.9: Performance comparison of NOMA Vs. Non-NOMA for “With CU”.

low SNR conditions. As the SNR improves, the Task error rate performance of both setups improves. The Task error rate of NOMA based system improves more than Non-NOMA system and remains lower for higher SNR values.

Figure 4.9 shows the comparison between the NOMA and Non-NOMA architecture for “With CU” setup.

For less complex dataset, in the low SNR region, the Task error rate of the NOMA system is higher than that of the Non-NOMA system but as SNR increases, the performance gap between NOMA and Non-NOMA diminishes. For more complex dataset the Task error rate performance of the NOMA based architecture is always better than the Non-NOMA architecture. This suggests that, the DNN-NOMA architecture is beneficial for complex task processing due to its DNN based receiver.

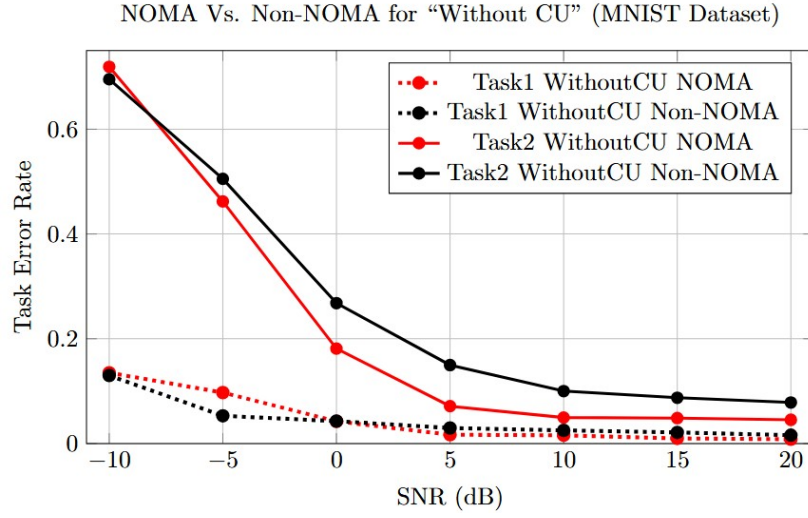


Figure 4.10: Performance comparison of NOMA Vs. Non-NOMA for “Without CU”.

Case 5: NOMA Vs. Non-NOMA for “Without CU”

For very low SNR (-10 dB), the Task error rate of the NOMA-based system is higher than that of the Non-NOMA system, due to the inherent complexity of NOMA in handling low SNR conditions. As the SNR improves, the Task error rate performance of both setups improves. The NOMA based architecture shows lower Task error rate than Non-NOMA at high SNR range.

Figure 4.10 shows the comparison between the NOMA and Non-NOMA architecture for “Without CU” setup.

For less complex dataset, in the low SNR region, the Task error rate of the NOMA system is slightly higher than that of the Non-NOMA system but as SNR increases, the performance gap between NOMA and Non-NOMA diminishes. For more complex dataset the Task error rate performance of the NOMA based architecture is better than the Non-NOMA architecture. This suggests that, the DNN-NOMA architecture is beneficial for complex task processing due to its DNN based receiver.

Case 6: SIC-Base Decoder Vs. DNN-Based Decoder

The DNN-based decoder achieves consistently higher accuracy across all SNR ranges compared to the SIC-based decoder. The accuracy difference is most noticeable in the low SNR region (-10 dB to 0 dB), where the DNN decoder effectively manages the challenging channel conditions better than the SIC decoder. As SNR increases, the accuracy gap between the two decoders reduces. However, the DNN-based decoder consistently retains higher accuracy across all SNR values.

Figure 4.11 shows the comparison between the DNN-based decoder and SIC-based decoder setups.

The DNN-based decoder consistently outperforms the SIC-based decoder, particularly in the low SNR region. For binary tasks, the performance gap between the two decoders is less pronounced, due to the simpler nature of the task. However, for categorical tasks, the DNN-

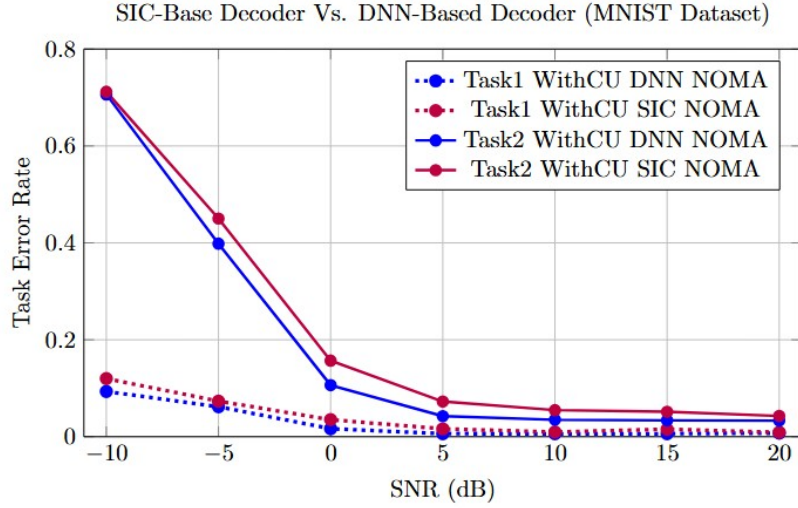


Figure 4.11: Performance comparison of SIC-Base Decoder Vs. DNN-Based Decoder.

based decoder shows a significant advantage, demonstrating its strength in handling complex multi-class classification problems.

4.3 Summary of CMT-SemCom Enabled by NOMA

CMT-SemCom can be well integrated with uplink NOMA to support efficient and intelligent multi-task communication. The integration of DNN-NOMA minimizes the reliance on CU in the transmitter side. This makes the system particularly suitable for edge devices with limited transmission-side cooperation capability.

While the proposed system offers significant advancements, power allocation in the NOMA framework is fixed in the current system. The present system relies on fixed power levels in the NOMA configuration. This may not adequately capture the varying channel conditions and semantic importance of transmitted features. Introducing semantic-aware power control could enable the system to allocate greater power to tasks that carry higher semantic value.

To address this, the next chapter explores different power allocation methods to understand their impact on performance. Building on those findings, Chapter 6 presents enhanced system models that incorporate semantic-aware power allocation into the CMT-SemCom with DMM-NOMA architecture. Our DNN-NOMA structure not only supports cooperative multi-task semantic processing but also enables semantic-aware PHY-layer management, which will be explored in the upcoming chapters.

Chapter 5

Power Allocation for Wireless Networks

Power allocation plays an important role in wireless systems. Optimized power allocation enhances system efficiency by ensuring that source data is transmitted effectively to the destination. As wireless data networks continue to evolve as integral components of next-generation infrastructure, they aim to provide seamless data access anytime, anywhere. The devices within these networks exhibit diverse computing and storage capabilities. This makes efficient power usage a critical design concern. As a result, power management remains one of the key challenges in wireless communication. Recent studies have explored various power allocation strategies to address this issue effectively [MRJ12].

An adaptive closed-loop power allocation scheme was introduced in [JQC⁺04] for non-regenerative relaying systems. This method adjusts the power based on feedback from the channel state information (CSI). However, its reliance on continuous CSI feedback poses practical challenges in fast-changing environments. To address this, open-loop schemes like those in [HA03] avoid reliance on instantaneous CSI but instead use statistical CSI. These schemes are advantageous in reducing complexity and feedback overhead while still providing performance gains.

In [SKAAS10], a power allocation framework was proposed for OFDM-based relay-assisted cognitive radio networks. It combined diversity gains from relaying with spectrum-sharing techniques. The scheme aims to maximize spectral efficiency under both peak and average interference power constraints.

A method focused on maximizing the average SNR at the destination in a single-relay setup was detailed in [DH05]. The authors demonstrated performance improvements over equal-power allocation by optimizing how power was distributed between the source and the relay. In contrast, the scheme in [FU08] focused on bit-error rate (BER) minimization. Here, the authors formulated a power allocation method to minimize a union bound on BER in fading relay channels using the amplify-and-forward protocol.

Chunhui Liu et al. [LSM10] proposed a water-filling-based power and rate allocation scheme where power was distributed optimally across subcarriers. This method significantly improved system performance by matching transmission rates to subchannel conditions. However, the strategy had substantial signaling overhead due to the complexity of maintaining current mapping schemes. To mitigate this, the authors also introduced a constant-rate scheme where power was still allocated adaptively across subcarriers but with simplified rate constraints. This reduced overhead while maintaining performance under BER and rate constraints.

In [Wan09], a power allocation framework was proposed that jointly considered long-term path loss and short-term Rayleigh fading. The scheme aimed to minimize the system BER by deriving expressions that effectively distribute power in variable environments. It demonstrated strong results in fading scenarios.

These power allocation techniques have played an important role in improving performance in traditional wireless systems. But as wireless technologies continue to evolve, new challenges and opportunities are emerging. NOMA introduces a different kind of power-sharing mechanism that allows multiple users to transmit simultaneously over the same resources. It requires specialized power allocation methods to manage user interference and maximize system throughput. Furthermore, with the rise of semantic communication, the focus shifts from just delivering bits reliably to prioritizing the meaning of the content. The following sections explore these two emerging directions in more detail.

5.1 Power Allocation in NOMA

In power-domain NOMA, the users' signals are superimposed by assigning corresponding power coefficients. The system allocates more power to the user with poorer channel conditions to balance throughput and fairness [ZWKL16]. This fundamental trade-off between maximizing the cell's sum-rate and ensuring cell-edge users achieve acceptable service depends on proper power allocation. This is why power allocation is one of the pivotal design factors in NOMA systems.

In NOMA, all M users' signals are superimposed [DYFP14]:

$$x = \sum_{m=1}^M \sqrt{P \alpha_m} s_m, \quad (5.1)$$

where s_m is the unit-power information symbol for user m , P is the total transmit power, α_m is the power-allocation coefficient, with

$$\sum_{m=1}^M \alpha_m = 1.$$

At the receiver side, user m observes

$$y_m = h_m x + n_m = h_m \sum_{m=1}^M \sqrt{P \alpha_m} s_m + n_m, \quad (5.2)$$

where $n_m \sim \mathcal{CN}(0, 1)$ is additive white Gaussian noise at user m .

The base station can flexibly control the throughput of each user by adjusting the power allocation ratio. The overall cell throughput, cell-edge throughput, and user fairness depend on the power allocation scheme [SKB⁺13]. Moreover, the outage performance of NOMA depends largely on the choices of allocated power for the users. If the power split is not correctly aligned with users' rate targets, certain users may be permanently in outage [DYFP14]. So, the benefit of NOMA depends on the allocated power among users.

5.1.1 Traditional Power Allocation Methods in NOMA

Over the past decade, a lot of research has emerged on NOMA power allocation. Vast majority of these researches focus on allocating power based on the channel conditions of the users.

In [DFP16], a Fixed Power-Allocation NOMA (F-NOMA) is introduced. It considers the simplest non-orthogonal scenario involving two users who share the same resource. Rather than dynamically adjusting power in response to instantaneous channel fading, the authors fix the transmit power for both users. They introduce two coefficients, a_m for the weaker user and a_n for the stronger user. They are chosen so that

$$a_m^2 + a_n^2 = 1 \quad \text{and} \quad a_m \geq a_n.$$

This choice reflects the core NOMA principle of allocating more power to the user who faces worse channel conditions.

A more adaptive NOMA strategy called Cognitive-Radio-Inspired NOMA (CR-NOMA) is introduced in [DFP16]. CR-NOMA dynamically adjusts power allocation based on the immediate quality-of-service (QoS) needs of the users. In CR-NOMA systems, the user experiencing poorer channel conditions is referred to as user m and it is treated similarly to a "primary user" in cognitive radio frameworks. This user is given strict priority to meet its QoS requirements. It is often expressed in terms of a minimum required data rate. Meanwhile, the user with a stronger channel (user n) functions as the "secondary user". It is permitted to utilize the remaining resources only if this does not compromise the QoS commitments for the primary user. The authors explain this concept by directly embedding the QoS constraint into the power allocation strategy. The transmit power allocated to the stronger user n is limited to ensure that weaker user m achieves its target data rate. As a result, the communication reliability of the more vulnerable user remains protected.

Saito et al. [SKB⁺13] observed that finding the absolute optimal power split among multiple users in NOMA is a combinatorial problem. The complexity grows prohibitively with the number of user groupings. To mitigate this, they introduced a method known as Fractional Transmit Power Allocation (FTPA). Instead of searching for every possible power split, FTPA applies a fractional rule to allocate more power to users with weaker channel conditions. Here, each user's allocated power is set inversely proportional to a power of its measured channel-to-noise ratio. The network can smoothly control how strongly it favors weak channels by tuning a single decay factor.

Saito et al. [SKB⁺13] also proposed Tree-Search Power Allocation (TTPA) to recover much of the lost performance without full exhaustive search. TTPA sees the set of all possible power splits as the leaves of a conceptual search tree. The algorithm branches progressively rather than examining each option. It starts with broad splits, evaluates partial assignments, and prunes entire subtrees when it determines that no further split can outperform the current best.

In [YPNP⁺23] the authors introduce a deep-learning framework whose primary role is to predict the optimal power-allocation coefficients. The authors found that jointly finding the best power split among users and the optimal phase settings involves a complex, non-convex search. It must be repeated every time the channel or user positions change. To overcome this, they proposed training a neural network to act as a real-time oracle. The network directly predicts near-optimal power allocation ratios without relying on iterative solvers. It used high-level inputs like users' relative positions and recent channel data. This deep-learning model captures the underlying relationship between network geometry, channel conditions, and optimal power distribution. As a result, it allows the system to adapt power allocations dynamically as users move or the environment changes.

5.2 Semantic-Aware Power Allocation

Conventional power-allocation methods focus only on users' channel conditions and QoS requirements. However, this approach is insufficient in semantic communications. In semantic communication, different data streams contribute unequally to perceived quality. Therefore, transmit power should be assigned based on each stream's impact on the final semantic performance.

In [XBMMT24], Semantic-Aware Proportional Method was introduced which decouples the semantic-performance constraint into per-stream requirements. Each stream is assigned just enough power so that its individual semantic value meets a predetermined threshold semantic value. Streams with higher semantic importance receive more power. A closed-form allocation is established which directly links required bit-error rates to power via the inverse Q-function.

The goal was to minimize total transmit power while ensuring that each semantic stream achieves a minimum perceptual-quality threshold \bar{P} . The transmit power for stream i is denoted by q_i , its quasi-static channel gain by h_i , and the noise variance by σ_i^2 . Under these definitions, the received SNR and resulting bit-error rate ψ_i for stream i are given by:

$$\text{SNR}_i = \frac{q_i |h_i|^2}{\sigma_i^2}. \quad (5.3)$$

$$\psi_i = \frac{a_i}{\log_2 M_i} Q\left(\sqrt{b_i \text{SNR}_i}\right), \quad (5.4)$$

where M_i is the modulation order and a_i, b_i are modulation-specific constants.

Instead of enforcing a single overall perceptual-quality constraint

$$P(\{\psi_i\}) \leq \bar{P} \quad (5.5)$$

they decouple this into independent requirements

$$\hat{L}_i(\psi_i) \geq \bar{L}_i, \quad \forall i,$$

where $\hat{L}_i(\psi_i)$ denotes the semantic value of stream i as a function of its bit-error rate ψ_i , and \bar{L}_i is the minimum semantic threshold for that stream. Under the insight that semantic value decreases monotonically with BER, the target BER ψ_i^* can be found by solving $\hat{L}_i(\psi_i^*) = \bar{L}_i$. This BER is then mapped to the required transmit power through the Q-function relationship:

$$q_i^* = \frac{\sigma_i^2}{b_i |h_i|^2} \left(Q^{-1}\left(\frac{\log_2 M_i}{a_i} \psi_i^*\right) \right)^2. \quad (5.6)$$

Because each stream is handled independently, the semantic-aware proportional method yields closed-form power allocations. This prioritizes streams with higher semantic importance. It can achieve significant energy savings when strict quality guarantees are needed.

Another method called Semantic-Aware Bisection Method was introduced in [XBMMT24] for cases where streams interact. The power allocation problem is reformulated over the joint error-performance surface. A bisection search along the contour identifies the point where further increasing one stream's power at the cost of the other no longer reduces total power.

In the Semantic-Aware Bisection Method, the allocation problem for two semantic streams is tackled by transforming the joint power-minimization with a perceptual-quality constraint. Concretely, rather than decoupling streams, they enforced

$$P(\psi_1, \psi_2) = \bar{P}, \quad (5.7)$$

so any feasible pair (ψ_1, ψ_2) lies on this “perception-error” curve. The algorithm repeatedly bisects the interval in the first stream’s error rate starting from two end-points on this curve. At each step it picks the midpoint ψ_1 , solves for the corresponding ψ_2 that keeps the total quality exactly \bar{P} , and then evaluates the objective’s directional derivative. It discards the half-interval that cannot contain the minimum and continues bisecting depending on the sign of this derivative. This process yields a locally optimal split without exhaustive two-dimensional search. The bisection method achieves tighter power savings than the proportional approach, especially when the two streams’ impacts interact in complex ways.

In this chapter, we reviewed two broad families of power-allocation policies.

Traditional NOMA strategies, such as fixed split, channel-gain ordering, and fairness-oriented water-filling distribute power according to physical-layer metrics. A user experiencing a poor channel is given more power and a strong-channel user receives less. Although this rule keeps signal-to-interference-plus-noise ratios (SINRs) balanced, it is agnostic to what the transmitted bits actually mean.

More recent semantic-aware methods go a step further by linking power to the information value of feature vectors. Yet none of them are suitable for the CMT-SemCom enabled by DNN-NOMA architecture, where several tasks are superimposed using different powers. To bridge this gap, Chapter 6 introduces a semantic-aware power-allocation framework tailored to the CMT-SemCom enabled by DNN-NOMA model. For every task we first express the desired semantic value as a target BER or MSE by fitting empirical curves. Those error targets are then inverted to SNR and finally to transmit power. It also takes the channel gain into account. The methodology is implemented in both a digital communication design in which quantized semantic symbols are modulated and an analog E2E design that sends continuous latent features directly.

Chapter 6

Semantic-Aware Power Allocation

In this chapter, we extend the CMT-SemCom enabled by DNN-NOMA framework by introducing semantic-aware power allocation mechanisms. Our work is driven by two central research questions:

1. **How can transmit power be tailored to the semantic requirements of each task?**

This question asks how we can allocate powers accordingly to the semantic importance of each task instead of treating every semantic task equally or allocating power solely on the basis of channel gain. This entails developing models that map semantic-level metrics into physical-layer parameters, and then solving for allocated power for each task.

2. **How does a semantic-aware power allocation strategy differ from conventional, channel-centric strategies?**

Traditional power control in NOMA focuses on maximizing channel throughput or ensuring user-fairness based on channel conditions. By contrast, a semantic-aware approach must balance semantic value and channel conditions requirements. We need to investigate how this semantic-aware power allocation model performs against the conventional channel-based and fixed-power models in terms of task error rate and total allocated power.

To answer these questions, we propose two system architectures:

- **Model 1: Digital-Communication Design**

Here, task-specific semantic encoders and decoders are first trained and optimized. Then their outputs are used and quantized into bitstreams for transmission over a power-domain NOMA based digital communication channel. Each bitstream's transmit power calculation is adjusted according to its task's semantic-value versus BER curve. This ensures that we allocate just enough energy to satisfy each task's semantic value requirement.

- **Model 2: Analog E2E Design**

In this design, quantization is omitted entirely. Task-specific feature vectors are superimposed directly in analog form and passed through the wireless channel. Power allocation calculated is based on semantic-value versus MSE relationships.

The remainder of this chapter presents the mathematical modeling and optimization formulations for each architecture.

First, we begin by analyzing the full-observation setting. Here, the receiver has access to the complete source signal. Then we will move to the partial-observation setting to reflect a real-world scenario. Here, each encoder only observes a subset of the source prior to semantic encoding.

6.1 Full-Observation Setting

In the full-observation setting, each encoder has access to the complete source signal, which in our case is the entire 28×28 MNIST image before semantic feature extraction. This represents the idealized scenario in which no information is hidden from any user.

6.1.1 Model 1: Digital Communication Design

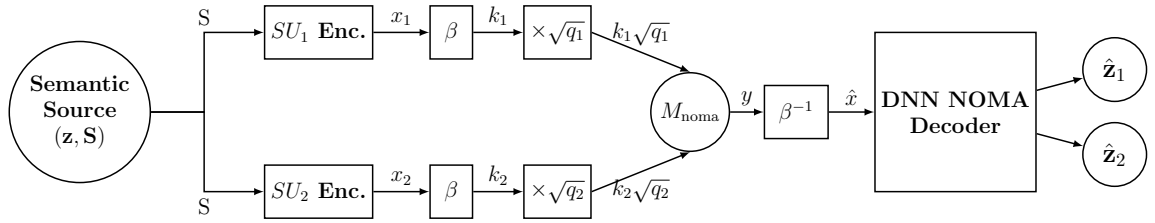


Figure 6.1: Block diagram of the proposed digital communication design of CMT-SemCom enabled by NOMA framework with semantic-aware power allocation.

Figure 6.1 illustrates the overall architecture of the digital communication design. Here, task-specific semantic encoders and decoders are first trained and optimized, then their outputs are used and quantized into bitstreams for transmission over a power-domain NOMA based digital communication channel.

Similar to the CMT-SemCom enabled by NOMA structure in 4, the semantic source is modeled as a tuple (\mathbf{z}, \mathbf{S}) . Each SU processes the shared features to produce task-specific encoded outputs $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$. For the i -th task, the SU encoder is modeled as:

$$p^{\text{SU}_i}(\mathbf{x}_i|\mathbf{S}), \quad i \in \{1, 2, \dots, N\}.$$

To ensure compatibility with existing digital communication systems, the semantic feature \mathbf{x}_i is converted into the bit sequence denoted as \mathbf{k}_i . We have $\mathbf{k}_i = \mathcal{B}(\mathbf{x}_i)$, where $\mathcal{B}(\cdot)$ is a binary mapping function such as ASCII, Unicode encoding, and quantization.

The task-specific binary signals $(\mathbf{k}_1, \mathbf{k}_2, \dots, \mathbf{k}_N)$ are superimposed using power-domain NOMA. The superimposed signal, \mathbf{M}_{noma} , is a linear combination of the encoded signals with different power levels. The received signal (\mathbf{y}) at the base station is modeled as:

$$\mathbf{y} = \sqrt{q_1} \mathbf{h}_1 \mathbf{k}_1 + \sqrt{q_2} \mathbf{h}_2 \mathbf{k}_2 + \dots + \sqrt{q_N} \mathbf{h}_N \mathbf{k}_N + \mathbf{n}, \quad (6.1)$$

$q_1 + q_2 + \dots + q_N = q$, the total transmission power

$$\mathbf{y} = \mathbf{M}_{\text{noma}} + \mathbf{n}, \quad (6.2)$$

This superimposition technique allows simultaneous transmission of task-specific signals over the same channel.

The received semantic data \mathbf{y} is first reconverted into the semantic features $\hat{\mathbf{x}} = \mathcal{B}^{-1}(\mathbf{y})$. Here, $\mathcal{B}^{-1}(\cdot)$ is the inverse operation of $\mathcal{B}(\cdot)$.

The signal ($\hat{\mathbf{x}}$) is processed by a DNN-based decoder at the base station. The decoding process is modeled probabilistically as:

$$p^{\text{Dec}}(\hat{\mathbf{z}}|\hat{\mathbf{x}}), \quad i \in \{1, 2, \dots, N\},$$

Probabilistic Representation: The entire system, from input to output, is modeled probabilistically with the help of Markov representation. The joint probability distribution for Task i is:

$$p(\hat{\mathbf{z}}, \hat{\mathbf{x}}, \mathbf{y}, \mathbf{x}_i | \mathbf{S}) = p^{\text{Dec}}(\hat{\mathbf{z}}|\hat{\mathbf{x}}) p^{\text{SemFet}}(\hat{\mathbf{x}}|\mathbf{y}) p^{\text{Channel}}(\mathbf{y}|\mathbf{k}_i) p^{\text{SU}_i}(\mathbf{x}_i | \mathbf{S}), \quad (6.3)$$

where:

- $p^{\text{Dec}}(\hat{\mathbf{z}}|\hat{\mathbf{x}})$: Decoder to reconstruct $\hat{\mathbf{z}}$ where $\hat{\mathbf{x}}$ is the extracted semantic features from the binary signal,
- $p^{\text{SemFet}}(\hat{\mathbf{x}}|\mathbf{y})$: Semantic feature extractor where \mathbf{y} is the received information passed through the channel,
- $p^{\text{Channel}}(\mathbf{y}|\mathbf{k}_i)$: Rayleigh fading Channel model incorporating noise,
- $p^{\text{Bin}}(\mathbf{k}_i|\mathbf{x}_i)$: Binary mapper to convert extracted semantic features into binary signals,
- $p^{\text{SU}_i}(\mathbf{x}_i | \mathbf{S})$: Task-specific encoder distribution, extracting task-specific information from the observation and providing the channel input,
- $\mathbf{k}_i = \mathcal{B}(\mathbf{x}_i)$ is the binary representation of semantic feature \mathbf{x}_i .

Optimization Problem

Stage 1 (Encoder and Decoder Optimization - Multi-Task Learning Problem): The optimization problem aims to maximize the mutual information between the received signal (\mathbf{y}) and the semantic variables (z_i) for all tasks. It is defined as:

$$\mathbf{P1} : \arg \max_{p^{\text{SU}}} \sum_{i=1}^N I(\mathbf{y}; z_i), \quad (6.4)$$

where $I(\mathbf{y}; z_i)$ is the mutual information between the received signal (y) and the semantic variable (z_i).

Using the objective function:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\Phi}) = \sum_{i=1}^N I(\mathbf{y}; z_i) \quad (6.5)$$

$$= \sum_{i=1}^N \iint p(\mathbf{y}, z_i) \log \frac{p(z_i | \mathbf{y})}{p(z_i)} dz_i dy \quad (6.6)$$

$$= \sum_{i=1}^N \left[\iint p(\mathbf{y}, z_i) \log p(z_i|\mathbf{y}) dz_i dy + H(z_i) \right] \quad (6.7)$$

Further ignoring the constant term $H(z_i)$ and leveraging the Markov chain relationship:

$$\mathcal{L}(\Phi) \approx \sum_{i=1}^N \int \int \int \int \int p(z_i, S) p_{\phi_i}^{SU}(\mathbf{x}_i|\mathbf{S}) p^{Channel}(\mathbf{y}|\mathbf{x}_i) \log p(z_i|\mathbf{y}) dz_i dS dx_i dy \quad (6.8)$$

$$\approx \sum_{i=1}^N \int \int \int \int p(z_i, \mathbf{S}) p_{\phi_i}^{SU}(\mathbf{y}|\mathbf{S}) \log p(z_i|\mathbf{y}) dz_i dS dy \quad (6.9)$$

$$\approx \sum_{i=1}^N \left\{ \mathbb{E}_{p(\mathbf{S}, z_i)} \left[\mathbb{E}_{p_{\phi_i}^{SU}(\mathbf{y}|\mathbf{S})} [\log p(z_i|\mathbf{y})] \right] \right\} \quad (6.10)$$

The channel outputs are used to emphasize the role of joint semantic and channel coding handled by the SUs. Using

$$p_{\phi_i}^{SU}(\mathbf{y}|\mathbf{S}) = \int p_{\phi_i}^{SU}(x_i|S) p^{Channel}(\mathbf{y}|\mathbf{x}_i) dx_i, \quad (6.11)$$

we aim to optimize $p_{\phi_i}^{SU}(\mathbf{y}|\mathbf{S})$.

The objective function demonstrates the E2E design approach, where both encoders and decoders are optimized together.

Regarding the decoder in the objective function, the $p^{Dec}(\hat{\mathbf{z}}|\mathbf{y})$ can be fully determined using the known distributions and underlying probabilistic relationship as:

$$p^{Dec}(\hat{\mathbf{z}}|\mathbf{y}) = \frac{\int p_{\phi_i}^{SU}(\mathbf{y}|\mathbf{S}) p(\mathbf{S}, z_i) ds}{p(y)} \quad (6.12)$$

However, due to the high-dimensional integrals, this equation becomes intractable, and we need to follow the variational approximation technique, resulting in the following:

$$\mathcal{L}(\Phi, \Psi) \approx \sum_{i=1}^N \left\{ \mathbb{E}_{p(\mathbf{S}, z_i)} \left[\mathbb{E}_{p_{\phi_i}^{SU}(\mathbf{y}|\mathbf{S})} [\log q_{\psi_i}^{Dec}(\hat{\mathbf{z}}|\mathbf{y})] \right] \right\} \quad (6.13)$$

Here, $\Psi = [\psi_1, \dots, \psi_N]$ represents NNs approximating the true distribution of decoders.

Stage 2 (Semantic-Aware Power Allocation): The objective is to minimize total power consumption while guaranteeing semantic performance.

$$\mathbf{P2} : \min_{q_i} \sum_{i=1}^N q_i \quad (6.14)$$

$$\text{s.t. } \hat{L}_i(\xi_i) \geq \bar{L}_i \quad (6.15)$$

Here, $\hat{L}_i(\xi_i)$ is the semantic value as a function of the BER (ξ_i), and \bar{L}_i is the target minimum semantic value requirement for semantic task i .

The received signal at the decoder can be written as

$$\mathbf{y} = \sqrt{q_1}\mathbf{h}_1\mathbf{k}_1 + \sqrt{q_2}\mathbf{h}_2\mathbf{k}_2 + \dots + \sqrt{q_N}\mathbf{h}_N\mathbf{k}_N + \mathbf{n}, \quad (6.16)$$

where h_i is channel assumed to be quasi-static and modelled as [XBMMT24]

$$h_i \triangleq \sqrt{h_0 \left(\frac{d}{d_0}\right)^{-\alpha}} \tilde{h}_i$$

Here, $h_0 \left(\frac{d}{d_0}\right)^{-\alpha}$ is the path loss at distance d with h_0 being the path loss at reference distance d_0 . \tilde{h}_i and n_i are the Rayleigh fading channel with a covariance of 1 and the Gaussian noise following the distributions of $n_i \sim \mathcal{CN}(0, \sigma_i^2)$. q_i is the allocated power for each symbol of the i -th semantic data stream.

Under the quasi-static channel, the received SNR of each symbol is equal, which is given by

$$\text{SNR}_i = \frac{q_i |h_i|^2}{\sigma_i^2}. \quad (6.17)$$

The BER of each bit of the i -th semantic data is given by

$$\xi_i = \frac{a_i}{\log_2 M_i} Q\left(\sqrt{b_i \text{SNR}_i}\right), \quad (6.18)$$

where

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} du$$

is the Q-function. Parameters a_i and b_i depend on the adopted modulation type with an order of M_i .

Figure 6.2 shows the semantic value - BER functions for both task1 and task2. The functions are derived through curve fitting of numerical simulation data.

The functions in the figures are obtained using nonlinear curve fitting methods, the Levenberg–Marquardt (L-M) algorithm. This algorithm is designed to minimize the sum of squared errors between the simulated data points and the chosen nonlinear model. It combines the strengths of the gradient descent method and the Gauss–Newton method. Here, `scipy.optimize.curve_fit()` was used to perform the fitting. This internally defaults to the L-M algorithm for unconstrained problems. This algorithm iteratively adjusts model parameters to best approximate the empirical data trends [Gav13]. Curve fitting is used because some of our unique simulated performance data do not follow simple closed-form laws. We obtain our required expression by fitting a nonlinear model with the L-M algorithm.

The semantic value - BER function for task1:

$$L(\xi) = -5.510 \times 10^1 \xi^3 + 8.500 \times 10^0 \xi^2 - 7.937 \times 10^{-1} \xi + 9.957 \times 10^{-1} \quad (6.19)$$

The semantic value - BER function for task2:

$$L(\xi) = 1.258 \times 10^2 \xi^3 + 1.235 \times 10^1 \xi^2 - 6.930 \times 10^0 \xi + 9.551 \times 10^{-1} \quad (6.20)$$

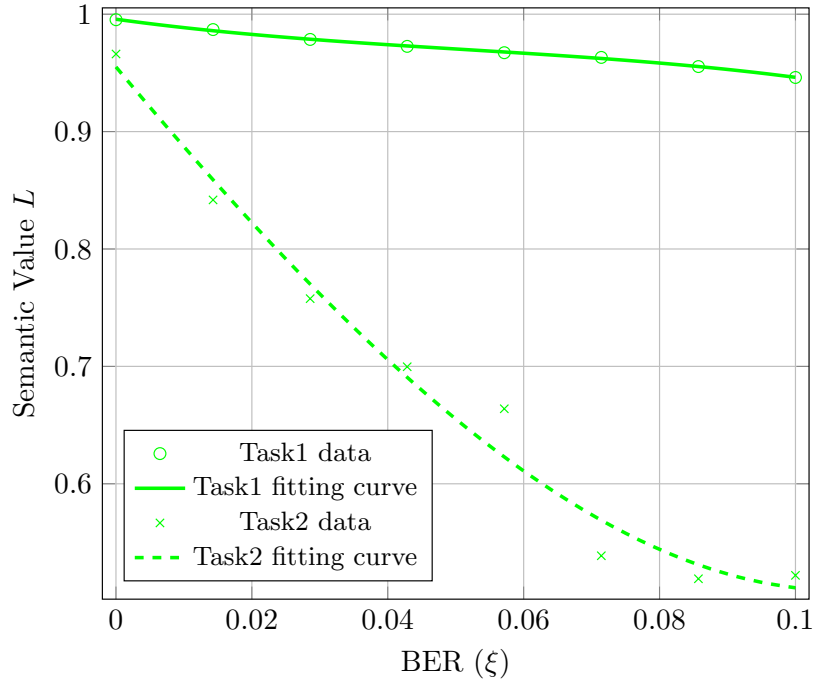


Figure 6.2: The semantic value - BER functions for task1 and task2.

The problem minimizing the total power consumption while ensuring semantic value remains equal or above the semantic value requirement of the i -th received semantic data can be formulated as

$$\min_{q_i} \sum_{i=1}^N q_i \quad (6.21)$$

$$\text{s.t. } \hat{L}_i(\xi_i) \geq \bar{L}_i, \quad \forall i \in I, \quad (6.22)$$

where $\hat{L}_i(\xi_i)$ is a monotonically non-increasing function of BER and \bar{L}_i is the semantic value requirement of the i -th received semantic data. Denoting ξ_i^* as the solution obtained by solving the equation $\hat{L}_i(\psi_i) = \bar{L}_i$, the optimal solutions can be obtained by substituting ξ_i^* , which is given by

$$q_i^* = \frac{\sigma_i^2}{b_i |h_i|^2} \left(Q^{-1} \left(\frac{\log_2 M_i}{a_i} \xi_i^* \right) \right)^2, \quad (6.23)$$

6.1.2 Model 2: Analog E2E Design

Unlike the digital-enabled design of Model 1, Model 2 maintains semantic features in continuous form throughout the entire transmission chain. Figure 6.3 depicts the complete analog processing pipeline. The observation \mathbf{S} is transformed by task-specific encoders (SUs) into continuous feature vectors \mathbf{x}_i . Instead of converting these vectors into bits, each \mathbf{x}_i is scaled by q_i and superimposed. This waveform traverses a Rayleigh-fading channel with additive noise and producing \mathbf{y} . Finally, a DNN ingests \mathbf{y} and disentangles the superimposed streams and gives per-task semantic estimates \hat{z}_i .

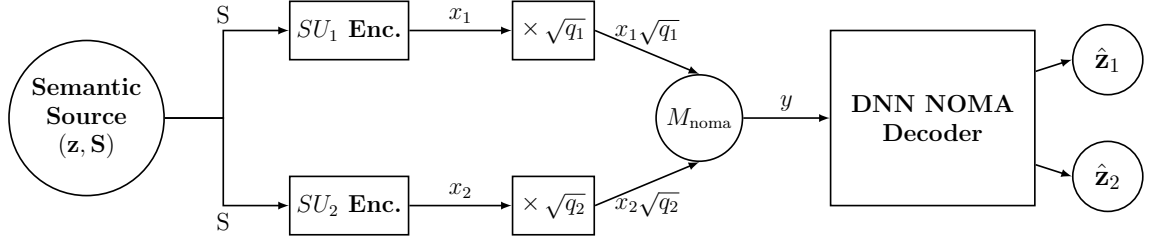


Figure 6.3: Block diagram of the proposed analog E2E design of CMT-SemCom enabled by NOMA framework with semantic-aware power allocation.

By retaining the analog nature of semantic information, Model 2 avoids the information loss. This enables higher semantic fidelity. In the following sections that, we formalize the probabilistic model and derive its joint optimization framework.

Similar to the CMT-SemCom enabled by DNN-NOMA structure in 4, the semantic source is modeled as a tuple (\mathbf{z}, \mathbf{S}) . Each SU processes the shared features to produce task-specific encoded outputs $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$. For the i -th task, the SU encoder is modeled as:

$$p^{\text{SU}_i}(\mathbf{x}_i|\mathbf{S}), \quad i \in \{1, 2, \dots, N\}.$$

The task-specific semantic features $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$ are superimposed using power-domain NOMA. This can be represented as:

$$\mathbf{y} = \sqrt{q_1}\mathbf{h}_1\mathbf{x}_1 + \sqrt{q_2}\mathbf{h}_2\mathbf{x}_2 + \dots + \sqrt{q_N}\mathbf{h}_N\mathbf{x}_N + \mathbf{n}, \quad (6.24)$$

Here,

$$\mathbf{y} = \mathbf{M}_{\text{noma}} + \mathbf{n}, \quad (6.25)$$

The decoding process is modeled probabilistically as:

$$p^{\text{Dec}}(\hat{\mathbf{z}}|\mathbf{y}), \quad i \in \{1, 2, \dots, N\},$$

Probabilistic Representation: The entire system, from input to output, is modeled probabilistically with the help of Markov representation. The joint probability distribution for Task i is:

$$p(\hat{\mathbf{z}}, \mathbf{y}, \mathbf{x}_i|\mathbf{S}) = p^{\text{Dec}}(\hat{\mathbf{z}}|\mathbf{y})p^{\text{Channel}}(\mathbf{y}|\mathbf{x}_i)p^{\text{SU}_i}(\mathbf{x}_i|\mathbf{S}), \quad (6.26)$$

where:

- $p^{\text{Dec}}(\hat{\mathbf{z}}|\mathbf{y})$: Decoder to reconstruct $\hat{\mathbf{z}}$ where \mathbf{y} is the received information passed through the channel,
- $p^{\text{Channel}}(\mathbf{y}|\mathbf{x}_i)$: Rayleigh fading Channel model incorporating noise,
- $p^{\text{SU}_i}(\mathbf{x}_i|\mathbf{S})$: Task-specific encoder distribution, extracting task-specific information from the observation and providing the channel input.

Optimization Problem

The primary goal of this model is to jointly optimize the semantic communication system. The optimization objective aims to maximize the total mutual information between transmitted semantic signals and corresponding semantic tasks, while minimizing the total power consumption. Formally, the joint optimization problem is expressed as:

$$\begin{aligned} \mathbf{P1} : \quad & \max_{p^{\text{SU}}, \{q_i\}} \sum_{i=1}^N I(\mathbf{y}; z_i) - \lambda \sum_{i=1}^N q_i \\ & \text{s.t.} \quad \hat{L}_i(E_i) \geq \bar{L}_i, \quad i = 1, \dots, N \end{aligned} \quad (6.27)$$

where:

- λ is a weighting parameter balancing semantic fidelity and power efficiency.
- $\hat{L}_i(E_i)$ is the semantic value as a function of the mean square error (MSE) E_i , and \bar{L}_i is the target minimum semantic value requirement for semantic task i .

Although Model 2 is an E2E pipeline from observation to reconstructed semantics, the optimization task intertwines two different sets of variables. On one side are the millions of encoder and decoder network weights that shape how semantic features are extracted, superimposed, and decoded. On the other side are just a few scalar power coefficients that scale each task's signal on the analogue NOMA channel. Solving for both groups simultaneously in one gradient-descent loop is numerically fragile and slow. It is because the network weights live on a highly non-convex surface while the power terms enter only as scalar gains. Solving the joint optimization problem directly is impractical due to these significant challenges. We therefore decompose the original joint objective into two subproblems: (i) mutual information maximization, and (ii) a semantic aware power allocation.

Subproblem 1 (Mutual Information Maximization):

The optimization problem aims to maximize the mutual information between the received signal (\mathbf{y}) and the semantic variables (z_i) for all tasks. It is defined as:

$$\mathbf{P2} : \arg \max_{p^{\text{SU}}} \sum_{i=1}^N I(\mathbf{y}; z_i), \quad (6.28)$$

where, $I(y; z_i)$ is the mutual information between the received signal (\mathbf{y}) and the semantic variable (z_i).

Using the objective function:

$$\mathcal{L}(\Phi) = \sum_{i=1}^N I(\mathbf{y}; z_i) \quad (6.29)$$

$$= \sum_{i=1}^N \iint p(\mathbf{y}, z_i) \log \frac{p(z_i | \mathbf{y})}{p(z_i)} dz_i dy \quad (6.30)$$

$$= \sum_{i=1}^N \left[\iint p(\mathbf{y}, z_i) \log p(z_i | \mathbf{y}) dz_i dy + H(z_i) \right] \quad (6.31)$$

Further ignoring the constant term $H(z_i)$ and leveraging the Markov chain relationship:

$$\mathcal{L}(\Phi) \approx \sum_{i=1}^N \int \int \int \int p(z_i, \mathbf{S}) p_{\phi_i}^{SU}(\mathbf{x}_i | \mathbf{S}) p^{Channel}(\mathbf{y} | \mathbf{x}_i) \log p(z_i | \mathbf{y}) dz_i d\mathbf{S} dx_i dy \quad (6.32)$$

$$\approx \sum_{i=1}^N \int \int \int p(z_i, \mathbf{S}) p_{\phi_i}^{SU}(\mathbf{y} | \mathbf{S}) \log p(z_i | \mathbf{y}) dz_i d\mathbf{S} dy \quad (6.33)$$

$$\approx \sum_{i=1}^N \left\{ \mathbb{E}_{p(\mathbf{S}, z_i)} \left[\mathbb{E}_{p_{\phi_i}^{SU}(\mathbf{y} | \mathbf{S})} [\log p(z_i | \mathbf{y})] \right] \right\} \quad (6.34)$$

The channel outputs are used to emphasize the role of joint semantic and channel coding handled by the SUs. Using

$$p_{\phi_i}^{SU}(\mathbf{y} | \mathbf{S}) = \int p_{\phi_i}^{SU}(\mathbf{x}_i | \mathbf{S}) p^{Channel}(\mathbf{y} | \mathbf{x}_i) dx_i, \quad (6.35)$$

we aim to optimize $p_{\phi_i}^{SU}(\mathbf{y} | \mathbf{S})$.

Regarding the decoder in the objective function, the $p^{Dec}(\hat{\mathbf{z}} | \mathbf{y})$ can be fully determined using the known distributions and underlying probabilistic relationship as:

$$p^{Dec}(\hat{\mathbf{z}} | \mathbf{y}) = \frac{\int p_{\phi_i}^{SU}(\mathbf{y} | \mathbf{S}) p(\mathbf{S}, z_i) ds}{p(\mathbf{y})} \quad (6.36)$$

However, due to the high-dimensional integrals, this equation becomes intractable, and we need to follow the variational approximation technique, resulting in the following:

$$\mathcal{L}(\Phi, \Psi) \approx \sum_{i=1}^N \left\{ \mathbb{E}_{p(\mathbf{S}, z_i)} \left[\mathbb{E}_{p_{\phi_i}^{SU}(\mathbf{y} | \mathbf{S})} [\log q_{\psi_i}^{Dec}(\hat{\mathbf{z}} | \mathbf{y})] \right] \right\} \quad (6.37)$$

Here, $\Psi = [\psi_1, \dots, \psi_N]$ represents NNs approximating the true distribution of decoders.

Subproblem 2 (Minimum Power Allocation):

The objective is to minimize total power consumption while guaranteeing semantic performance.

$$\begin{aligned} \mathbf{P3}: \quad & \min_{\{q_i\}} \sum_{i=1}^N q_i \\ \text{s.t.} \quad & \hat{L}_i(\xi_i) \geq \bar{L}_i, \quad \forall i \in I \end{aligned} \quad (6.38)$$

The received signal at the decoder can be written as

$$\mathbf{y} = \sqrt{q_1} \mathbf{h}_1 \mathbf{x}_1 + \sqrt{q_2} \mathbf{h}_2 \mathbf{x}_2 + \dots + \sqrt{q_N} \mathbf{h}_N \mathbf{x}_N + \mathbf{n}, \quad (6.39)$$

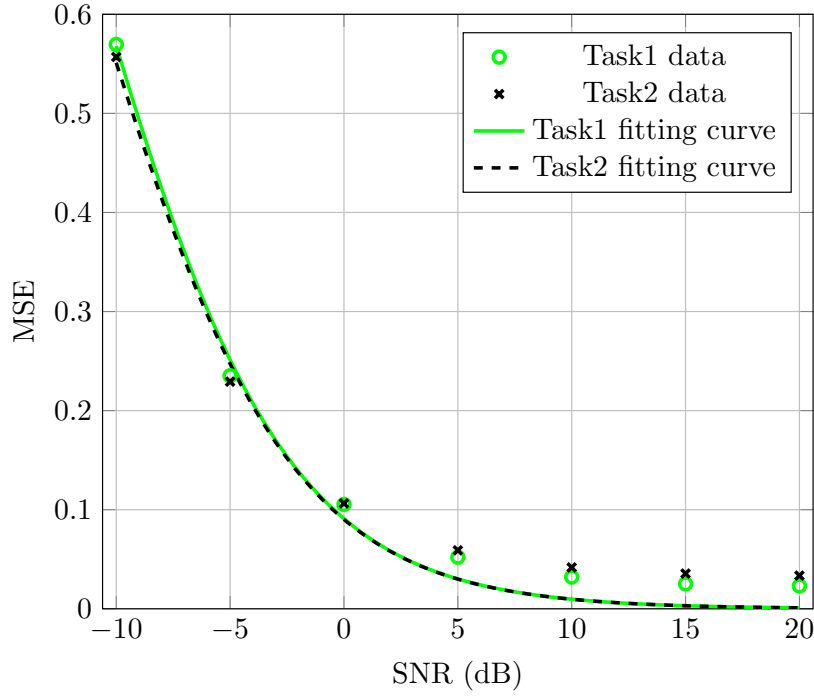


Figure 6.4: The MSE - SNR functions for task1 and task2

Under the quasi-static channel, the received SNR of each symbol is equal, which is given by

$$\text{SNR}_i = \frac{q_i |h_i|^2}{\sigma_i^2}. \quad (6.40)$$

Figure 6.4 shows the MSE - SNR functions for both task1 and task2. The functions are derived through curve fitting of numerical simulation data. The functions in the figures are also obtained using nonlinear curve fitting methods, the Levenberg-Marquardt (L-M) algorithm [Gav13].

General MSE - SNR function:

$$E = \frac{\alpha}{1 + \beta \text{SNR}}, \quad (6.41)$$

MSE - SNR function for task1:

$$E = \frac{1.3610}{1 + 13.9598 \text{SNR}}. \quad (6.42)$$

MSE - SNR function for task2:

$$E = \frac{1.2757}{1 + 13.1469 \text{SNR}}. \quad (6.43)$$

Figure 6.5 shows the semantic value - MSE functions for both task1 and task2. The functions are derived through curve fitting of numerical simulation data. The functions in the figures are also obtained using nonlinear curve fitting methods, the Levenberg-Marquardt (L-M) algorithm [Gav13].

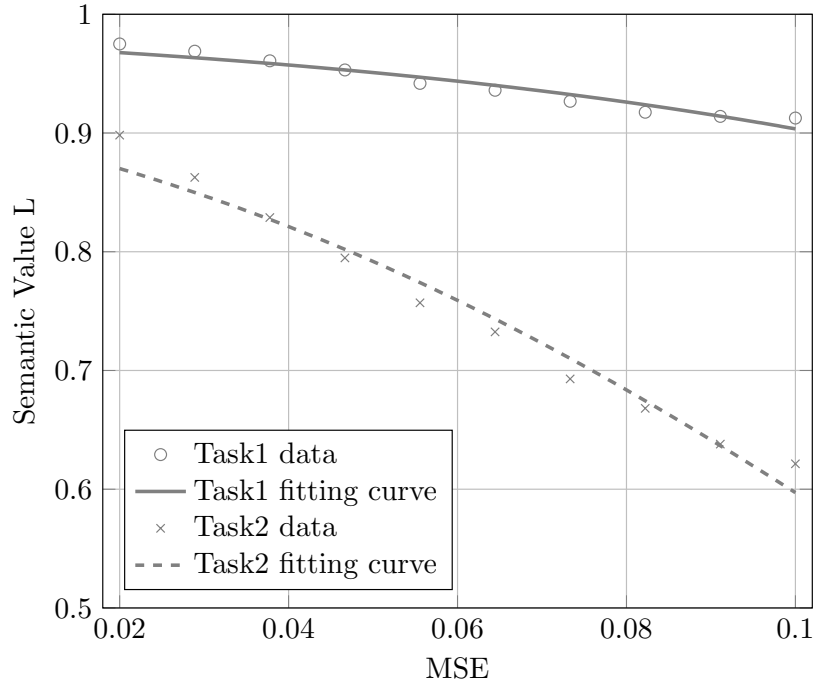


Figure 6.5: The Semantic value - MSE functions for task1 and task2

Semantic value - MSE function for task1:

$$L = \frac{1}{1 + \exp(14.5523E - 3.6914)} \quad (6.44)$$

Semantic value - MSE function for task2:

$$L = \frac{1}{1 + \exp(18.8438E - 2.2781)} \quad (6.45)$$

Using $E = \alpha / (1 + \beta \text{SNR})$:

$$\frac{\alpha}{1 + \beta \text{SNR}_i} = \bar{E}_i \implies \text{SNR}_i^* = \frac{1}{\beta} \left(\frac{\alpha}{\bar{E}_i} - 1 \right).$$

Since $\text{SNR}_i^* = \frac{q_i^* |h_i|^2}{\sigma^2}$, we get

$$q_i^* = \frac{\sigma^2}{|h_i|^2} \frac{1}{\beta} \left(\frac{\alpha}{\bar{E}_i} - 1 \right) \quad (6.46)$$

6.2 Distributed Partial-Observation Setting

In the distributed partial-observation setting, each encoder only observes a subset of the source prior to semantic encoding. For instance, when the source is a 28×28 MNIST image, one encoder processes the top 14×28 half while a second encoder processes the bottom 14×28 half. This models practical multi-agent or edge-sensor deployments in which no single node has full access to the source. Similar distributed partial-view scenarios have been studied in recent

semantic recovery frameworks (e.g. [BBD23]). This study demonstrated that cooperative, E2E architectures can nearly close the gap to full-view performance. This can be useful in realistic scenarios where only distributed partial observations are available, such as in a production line monitored by multiple sensing nodes [RH24].

6.2.1 Model 3: Digital Communication Design with Distributed Partial Observation Setting

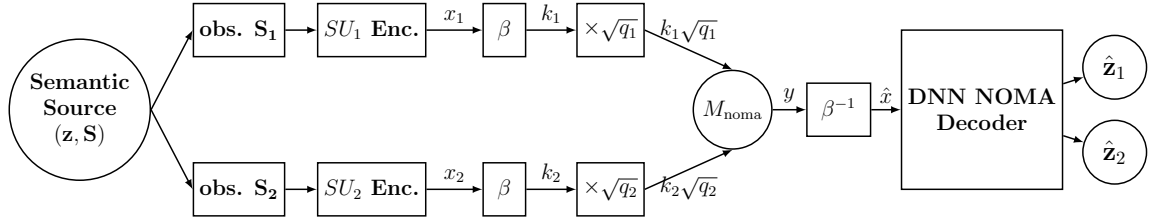


Figure 6.6: Block diagram of the proposed digital communication design of CMT-SemCom enabled by NOMA framework with semantic-aware power allocation and distributed partial observation setting.

Figure 6.6 illustrates the overall architecture of the digital communication design with distributed partial observation setting. Here, each task-specific encoders observe only a fraction of the scene. The task-specific semantic encoders and decoders are first trained and optimized, then their outputs are used and quantized into bitstreams for transmission over a power-domain NOMA based digital communication channel.

The semantic source is modeled as a tuple $(\mathbf{z}, \mathbf{S}_j)$, where \mathbf{S}_j represents j -th partially observed data.

For the i -th task, the SU encoder is modeled as:

$$p^{\text{SU}_i}(\mathbf{x}_i|\mathbf{S}_j), \quad i \in \{1, 2, \dots, N\}, \quad j \in \{1, 2, \dots, N\}.$$

In Model 3, the binary mapping, superposition, semantic feature extraction, and DNN-based decoding are performed exactly as in Model 1 (6.1.1)

Probabilistic Representation: The joint probability distribution for Task i is:

$$p(\hat{\mathbf{z}}, \hat{\mathbf{x}}, \mathbf{y}, \mathbf{x}_i|\mathbf{S}_j) = p^{\text{Dec}}(\hat{\mathbf{z}}|\hat{\mathbf{x}})p^{\text{SemFet}}(\hat{\mathbf{x}}|\mathbf{y})p^{\text{Channel}}(\mathbf{y}|\mathbf{k}_i)p^{\text{SU}_i}(\mathbf{x}_i|\mathbf{S}_j), \quad (6.47)$$

where:

- $p^{\text{SU}_i}(\mathbf{x}_i|\mathbf{S}_j)$: Task-specific encoder distribution, extracting task-specific information from the partial-observation and providing the channel input.

Optimization Problem

Stage 1 (Encoder and Decoder Optimization - Multi-Task Learning Problem): The optimization problem aims to maximize the mutual information between the received signal (\mathbf{y})

and the semantic variables (z_i) for all tasks. It is defined as:

$$\mathbf{P1} : \arg \max_{p^{\text{SU}}} \sum_{i=1}^N I(\mathbf{y}; z_i), \quad (6.48)$$

The subsequent steps are carried out exactly as in Model 1 (6.1.1), leading to:

$$\mathcal{L}(\Phi, \Psi) \approx \sum_{i=1}^N \left\{ \mathbb{E}_{p(\mathbf{s}_j, z_i)} \left[\mathbb{E}_{p_{\phi_i}^{\text{SU}}(\mathbf{y}|\mathbf{s}_j)} \left[\log q_{\psi_i}^{\text{Dec}}(\hat{\mathbf{z}}|\mathbf{y}) \right] \right] \right\} \quad (6.49)$$

Stage 2 (Semantic-Aware Power Allocation): The objective is to minimize total power consumption while guaranteeing semantic performance.

$$\mathbf{P2} : \min_{q_i} \sum_{i=1}^N q_i \quad (6.50)$$

$$\text{s.t. } \hat{L}_i(\xi_i) \geq \bar{L}_i \quad (6.51)$$

Here, $\hat{L}_i(\xi_i)$ is the semantic value as a function of the bit error rate (BER) ξ_i , and \bar{L}_i is the target minimum semantic value requirement for semantic task i .

The received signal at the decoder can be written as

$$\mathbf{y} = \sqrt{q_1} \mathbf{h}_1 \mathbf{k}_1 + \sqrt{q_2} \mathbf{h}_2 \mathbf{k}_2 + \dots + \sqrt{q_N} \mathbf{h}_N \mathbf{k}_N + \mathbf{n}, \quad (6.52)$$

Under the quasi-static channel, the received SNR of each symbol is equal, which is given by

$$\text{SNR}_i = \frac{q_i |h_i|^2}{\sigma_i^2}. \quad (6.53)$$

The BER of each bit of the i -th semantic data is given by

$$\xi_i = \frac{a_i}{\log_2 M_i} Q \left(\sqrt{b_i \text{SNR}_i} \right), \quad (6.54)$$

where

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} du$$

is the Q-function. Parameters a_i and b_i depend on the adopted modulation type with an order of M_i .

Figure 6.7 shows the semantic value - BER functions for both task1 and task2. The functions are derived through curve fitting of numerical simulation data. The functions in the figures are obtained using nonlinear curve fitting methods, the Levenberg-Marquardt (L-M) algorithm [Gav13].

The semantic value - BER function for task1:

$$L(\xi) = -1.776 \times 10^1 \xi^3 + 3.718 \times 10^0 \xi^2 - 5.637 \times 10^{-1} \xi + 9.712 \times 10^{-1} \quad (6.55)$$

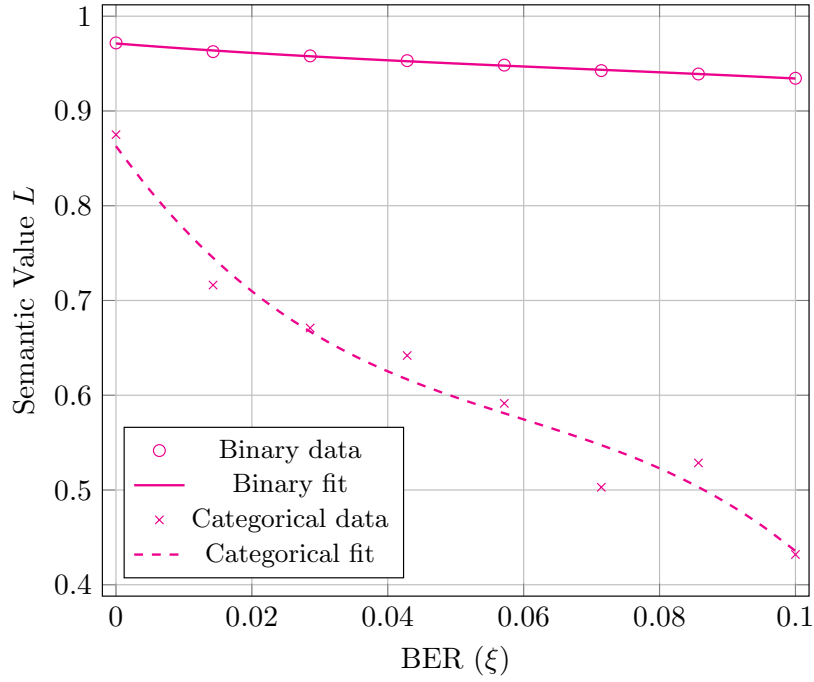


Figure 6.7: The semantic value - BER functions for task1 and task2.

The semantic value - BER function for task2:

$$L(\xi) = -7.235 \times 10^2 \xi^3 + 1.291 \times 10^2 \xi^2 - 9.948 \times 10^0 \xi + 8.629 \times 10^{-1} \quad (6.56)$$

The problem minimizing the total power consumption while ensuring semantic value remains equal or above the semantic value requirement of the i th received semantic data can be formulated as

$$\min_{q_i} \sum_{i=1}^N q_i \quad (6.57)$$

$$\text{s.t. } \hat{L}_i(\xi_i) \geq \bar{L}_i, \quad \forall i \in I, \quad (6.58)$$

where $\hat{L}_i(\xi_i)$ is a monotonically non-increasing function of BER and \bar{L}_i is the semantic value requirement of the i -th received semantic data. Denoting ξ_i^* as the solution obtained by solving the equation $\hat{L}_i(\xi_i) = \bar{L}_i$, the optimal solutions can be readily obtained by substituting ξ_i^* , which is given by

$$q_i^* = \frac{\sigma_i^2}{b_i |h_i|^2} \left(Q^{-1} \left(\frac{\log_2 M_i}{a_i} \xi_i^* \right) \right)^2, \quad (6.59)$$

6.2.2 Model 4: Analog E2E Design with Distributed Partial Observation Setting

Unlike the digital-enabled design of Model 3, Model 4 maintains semantic features in continuous form throughout the entire transmission chain by removing quantization and bit-mapping. Figure 6.8 depicts the complete analog processing pipeline.

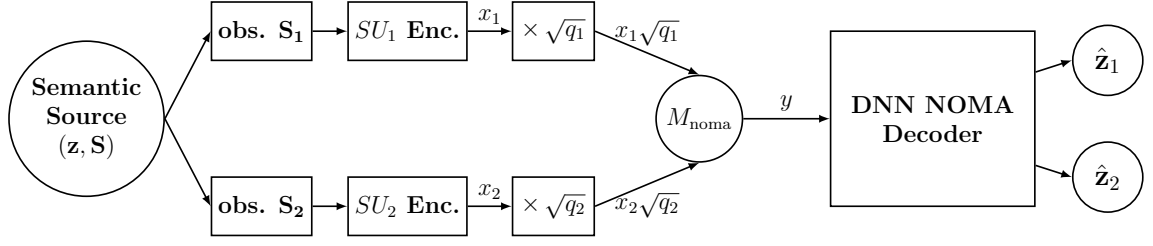


Figure 6.8: Block diagram of the proposed analog E2E design of CMT-SemCom enabled by NOMA framework with semantic-aware power allocation and distributed partial observation setting.

The semantic source is modeled as a tuple $(\mathbf{z}, \mathbf{S}_j)$, where \mathbf{S}_j represents j -th partially observed the observed data.

For the i -th task, the SU encoder is modeled as:

$$p^{\text{SU}_i}(\mathbf{x}_i | \mathbf{S}_j), \quad i \in \{1, 2, \dots, N\}, \quad j \in \{1, 2, \dots, N\}.$$

In Model 4, superposition and DNN-based decoding are performed exactly as in Model 2 (6.1.2)

Probabilistic Representation: The entire system, from input to output, is modeled probabilistically with the help of Markov representation. The joint probability distribution for Task i is:

$$p(\hat{\mathbf{z}}, \mathbf{y}, \mathbf{x}_i | \mathbf{S}_j) = p^{\text{Dec}}(\hat{\mathbf{z}} | \mathbf{y}) p^{\text{Channel}}(\mathbf{y} | \mathbf{x}_i) p^{\text{SU}_i}(\mathbf{x}_i | \mathbf{S}_j), \quad (6.60)$$

Optimization Problem

The primary goal of this model is to jointly optimize the semantic communication system, comprising semantic encoders, power allocation for NOMA, and a DNN-based decoder. The optimization objective aims to maximize the total mutual information between transmitted semantic signals and corresponding semantic tasks, while simultaneously minimizing the total power consumption. Formally, the joint optimization problem is expressed as:

$$\begin{aligned} \mathbf{P1} : \quad & \max_{p^{\text{SU}}, \{q_i\}} \sum_{i=1}^N I(\mathbf{y}; z_i) - \lambda \sum_{i=1}^N q_i \\ & \text{s.t.} \quad \hat{L}_i(E_i) \geq \bar{L}_i, \quad i = 1, \dots, N \end{aligned} \quad (6.61)$$

Similar to Model 2 (6.1.2), we decompose the original joint objective into two subproblems: (i) mutual information maximization, and (ii) a semantic aware power allocation. The next two subsections detail these subproblems and their solutions.

Subproblem 1 (Mutual Information Maximization):

The optimization problem aims to maximize the mutual information between the received

signal (\mathbf{y}) and the semantic variables (z_i) for all tasks. It is defined as:

$$\mathbf{P2} : \arg \max_{p^{\text{SU}}} \sum_{i=1}^N I(\mathbf{y}; z_i), \quad (6.62)$$

The subsequent steps are carried out exactly as in Model 2 (6.1.2), leading to:

$$\mathcal{L}(\Phi, \Psi) \approx \sum_{i=1}^N \left\{ \mathbb{E}_{p(\mathbf{S}_j, z_i)} \left[\mathbb{E}_{p_{\phi_i}^{\text{SU}}(\mathbf{y}|\mathbf{S}_j)} \left[\log q_{\psi_i}^{\text{Dec}}(\hat{\mathbf{z}}|\mathbf{y}) \right] \right] \right\} \quad (6.63)$$

Subproblem 2 (Minimum Power Allocation):

The objective is to minimize total power consumption while guaranteeing semantic performance.

$$\begin{aligned} \mathbf{P3} : \quad & \min_{\{q_i\}} \sum_{i=1}^N q_i \\ & \text{s.t.} \quad \hat{L}_i(\xi_i) \geq \bar{L}_i, \quad \forall i \in I \end{aligned} \quad (6.64)$$

The received signal at the decoder can be written as

$$\mathbf{y} = \sqrt{q_1} \mathbf{h}_1 \mathbf{x}_1 + \sqrt{q_2} \mathbf{h}_2 \mathbf{x}_2 + \dots + \sqrt{q_N} \mathbf{h}_N \mathbf{x}_N + \mathbf{n}, \quad (6.65)$$

Under the quasi-static channel, the received SNR of each symbol is equal, which is given by

$$\text{SNR}_i = \frac{q_i |h_i|^2}{\sigma_i^2}. \quad (6.66)$$

Figure 6.9 shows the MSE - SNR functions for both task1 and task2. The functions are derived through curve fitting of numerical simulation data. The functions in the figures are also obtained using nonlinear curve fitting methods, the Levenberg–Marquardt (L-M) algorithm [Gav13].

General MSE - SNR function:

$$E = \frac{\alpha}{1 + \beta \text{SNR}}, \quad (6.67)$$

MSE - SNR function for task1:

$$E = \frac{1.2859}{1 + 13.8334 \text{SNR}}. \quad (6.68)$$

MSE - SNR function for task2:

$$E = \frac{1.4770}{1 + 14.0419 \text{SNR}}. \quad (6.69)$$

Figure 6.10 shows the semantic value - MSE functions for both task1 and task2. The functions are derived through curve fitting of numerical simulation data. The functions in these figures are also obtained using nonlinear curve fitting methods, the Levenberg–Marquardt (L-M) algorithm [Gav13].

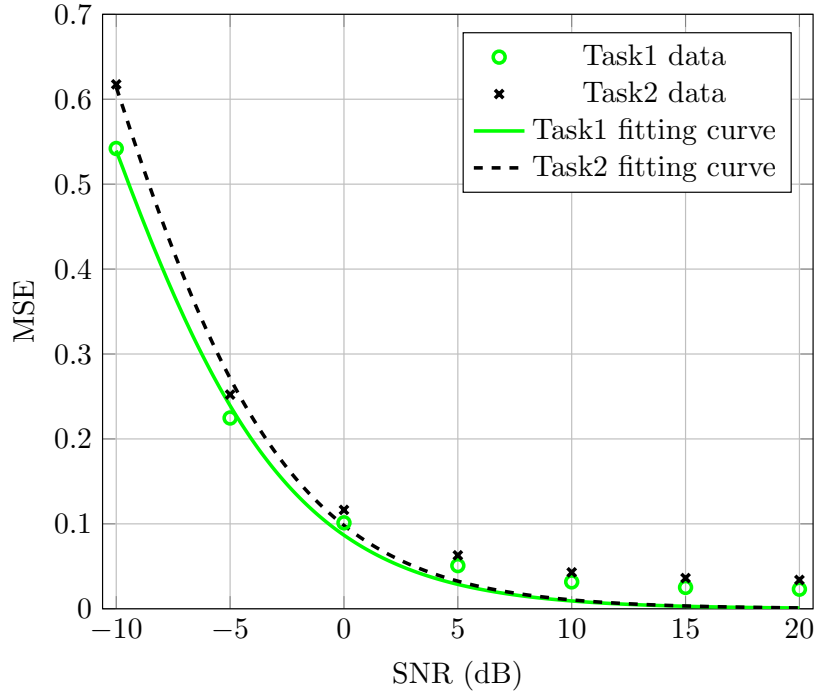


Figure 6.9: The MSE - SNR functions for task1 and task2

Semantic value - MSE function for task1:

$$L = \frac{1}{1 + \exp(14.6460E - 3.2232)} \quad (6.70)$$

Semantic value - MSE function for task2:

$$L = \frac{1}{1 + \exp(20.3716E - 2.1683)} \quad (6.71)$$

Using $E = \alpha / (1 + \beta \text{SNR})$:

$$\frac{\alpha}{1 + \beta \text{SNR}_i} = \bar{E}_i \implies \text{SNR}_i^* = \frac{1}{\beta} \left(\frac{\alpha}{\bar{E}_i} - 1 \right).$$

Since $\text{SNR}_i^* = \frac{q_i^* |h_i|^2}{\sigma^2}$, we get

$$q_i^* = \frac{\sigma^2}{|h_i|^2} \frac{1}{\beta} \left(\frac{\alpha}{\bar{E}_i} - 1 \right) \quad (6.72)$$

In this chapter, we introduced four distinct ways of bringing semantic awareness into power allocation for our CMT-SemCom enabled by DNN-NOMA system. The first and third models are structured for digital transmission while the second and fourth models are structured in an E2E analog form. For each approach, we laid out the system model, developed a probabilistic framework, and solved optimisation problem. We will evaluate each model's performance in Chapter 7 and compare the models to evaluate the performance of them.

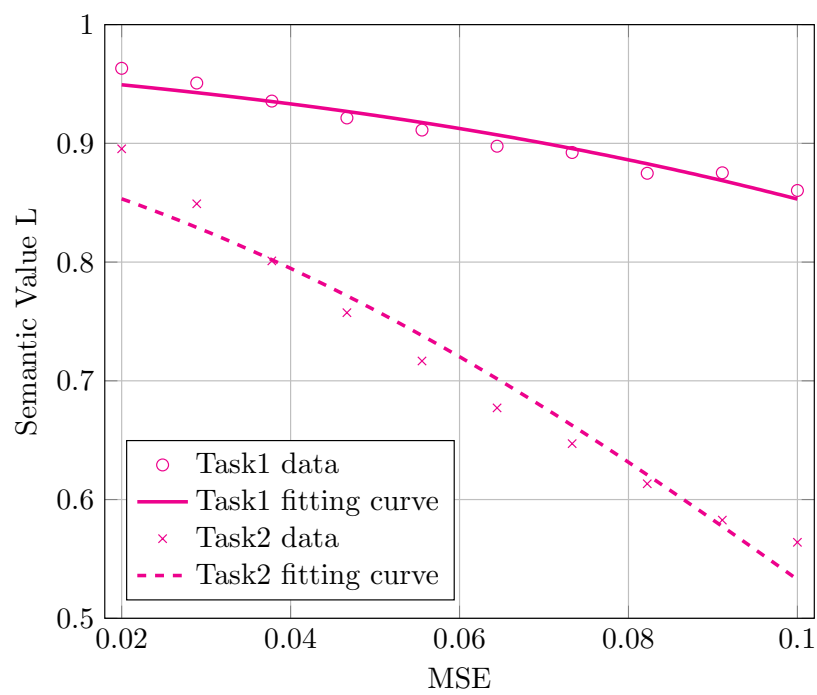


Figure 6.10: The Semantic value - MSE functions for task1 and task2

Chapter 7

Simulation Results

This chapter presents the simulation results of the proposed semantic-aware CMT-SemCom enabled by DNN-NOMA framework. The main focus is to evaluate how different power allocation strategies affect the system's performance across various configurations. Throughout the experiments, we compare equal power allocation with semantic-aware power allocation, where power is distributed based on the semantic importance and task complexity. The goal is to understand whether adapting power according to semantic requirements can improve task accuracy and overall system efficiency. These simulations help highlight the impact of power allocation in multitask semantic communication. Moreover, we will see the performance comparison of partial vs. full observation settings to understand their performance trade-offs.

7.1 Simulation Setup

The simulation setup for the proposed architecture is designed to validate its performance in multi-task semantic communication using a benchmark dataset and specific neural network configurations. The details of the setup are provided below.

7.1.1 Dataset

- **MNIST Dataset:**

- The MNIST dataset consists of grayscale images of handwritten digits ranging from 0 to 9.
- It includes 60,000 training samples and 10,000 test samples, each with a resolution of 28×28 pixels.
- The dataset is preprocessed to associate each sample with semantic variables for two tasks:
 - * **Task 1 (Binary Classification):** Determines the presence or absence of a specific digit (digit "2").
 - * **Task 2 (Categorical Classification):** Identifies the digit class (0–9).

7.1.2 Neural Network Architecture

The architecture comprises two main components: SU encoders, and the DNN-based base station. The specifications for each module are as follows:

- **SU Encoders:**

- Refine the shared features into task-specific representations (x_1, x_2):
 - * **SU Encoder 1:** Dedicated to Task 1, consists of an FC layer with 16 units and Tanh activation.
 - * **SU Encoder 2:** Dedicated to Task 2, consists of an FC layer with 16 units and Tanh activation.

- **DNN-Based Base Station:**

- Acts as a decoder to reconstruct the semantic variables for both tasks from the received signal.
- Composed of two FC layers with 64 units and ReLU activation.
- Outputs:
 - * **Binary Classification (Task 1):** Sigmoid activation for binary output.
 - * **Categorical Classification (Task 2):** Softmax activation for multi-class output.

Table 7.1: Neural network architecture specifications

Neural Network	Layers
SU Encoder 1	FC (16 units, Tanh activation)
SU Encoder 2	FC (16 units, Tanh activation)
DNN-Based Base Station	Two FC (64 units, ReLU) + Binary output (Sigmoid) + Categorical output (Softmax)

7.1.3 Semantic Tasks

- **Task 1 (Binary Classification):**

- The goal is to determine the presence or absence of a specific feature in the input image.
- For MNIST, this corresponds to identifying whether a specific digit ("2") is present.

- **Task 2 (Categorical Classification):**

- The objective is to classify the input into one of multiple categories.
- For MNIST, this involves digit classification (0–9).

7.1.4 Training and Evaluation

- **Preprocessing:**

- Each dataset sample is preprocessed to include labels for both tasks.
- Data normalization is applied to ensure consistency during training.

- **Training Framework:**

- The SU encoders, along with the DNN-based decoder, are trained to minimize the multi-task loss function.
- **Loss Functions:**
 - * Binary Cross-Entropy for Task 1.
 - * Categorical Cross-Entropy for Task 2.
- Optimization is performed using the Adam optimizer with a suitable learning rate.

7.1.5 Summary of Simulation Cases

To provide a clear understanding of the cases studied and their comparisons, we have designed two summary tables for Task 1 and Task 2. These tables visually represent the corresponding line styles used in the simulation results.

Table 7.2: Simulation line styles digital communication design









Category	Task1 Line Style	Task2 Line Style
Equally allocated power		
Channel-based power		
Semantic-aware power		
Semantic-aware power with partial observation		

Table 7.3: Simulation line styles analog E2E design









Category	Task1 Line Style	Task2 Line Style
Equally allocated power		
Channel-based power		
Semantic-aware power		
Semantic-aware power with partial observation		

Table 7.2 table highlights the configurations for digital communication design, where the line styles are represented with distinct colors for each case.

Table 7.3 presents the configurations for analog E2E design, where the line styles are represented with distinct colors for each case.

7.2 Case Study

7.2.1 Digital Communication Design (Full-Observation)

Case 1: Task error rate comparison of semantic-aware power allocation vs. equally allocated power allocation for digital communication design

In this experiment, the performance of semantic-aware power allocation of Model 1 in 6.1.1 is compared against a baseline where equal transmit power is allocated to both tasks, across varying channel gain pairs (h_1, h_2) .

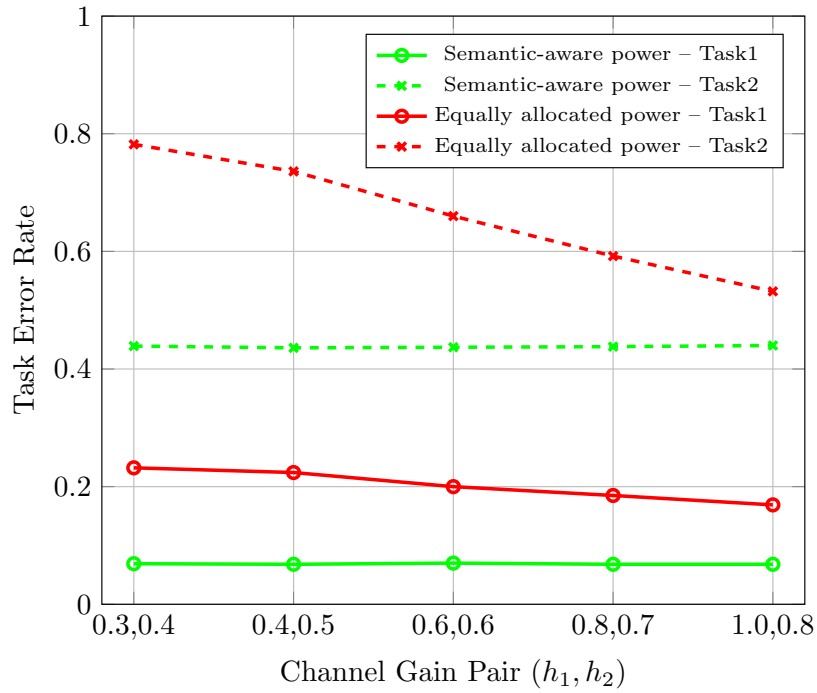


Figure 7.1: Task error rate comparison of semantic-aware power allocation vs. equally allocated power allocation for digital communication design.

In the semantic-aware setup, power is assigned to each task based on its semantic value requirement and the channel gain.

In contrast, the equally allocated power allocation setup uses fixed, identical power levels for both tasks.

Figure 7.1 shows the simulation result for the task error rate comparison of semantic-aware power allocation vs. equally allocated power allocation for digital communication design.

Simulation Results

- Task error rates for both Task 1 (binary identification) and Task 2 (multi-class classification) are plotted against increasing channel gain pairs.
- Under semantic-aware allocation, both tasks maintain stable and low error rates across all channel conditions.
- Under equal power allocation:
 - Task 1 performs reasonably well and improves slightly with better channel gains.
 - Task 2 shows significantly higher error rates at lower channel gains.
- The performance gap between semantic-aware and equal-power setups is especially visible for Task 2 in low-to-mid channel gain regimes.

General Observations

- The semantic-aware allocation meets each task's semantic accuracy requirement by dynamically adjusting transmit power based on semantic value and channel quality. This ensures consistent task performance.
- As the Task 2 requires higher semantic precision due to its multi-class nature, it benefits significantly from semantic-aware power allocation. In contrast, the equal-power setup under-allocates power to Task 2. This leads to high error rates.
- Task 1 is simpler and requires less power to achieve its semantic goal. The equal-power system often over-allocates power to it. This wastes power resources.
- These results highlight that semantic-aware power allocation leads to more efficient use of transmit power by allocating power with both semantic value and channel conditions.

Case 2: Task error rate comparison of semantic-aware power allocation vs. channel-based power allocation for digital communication design

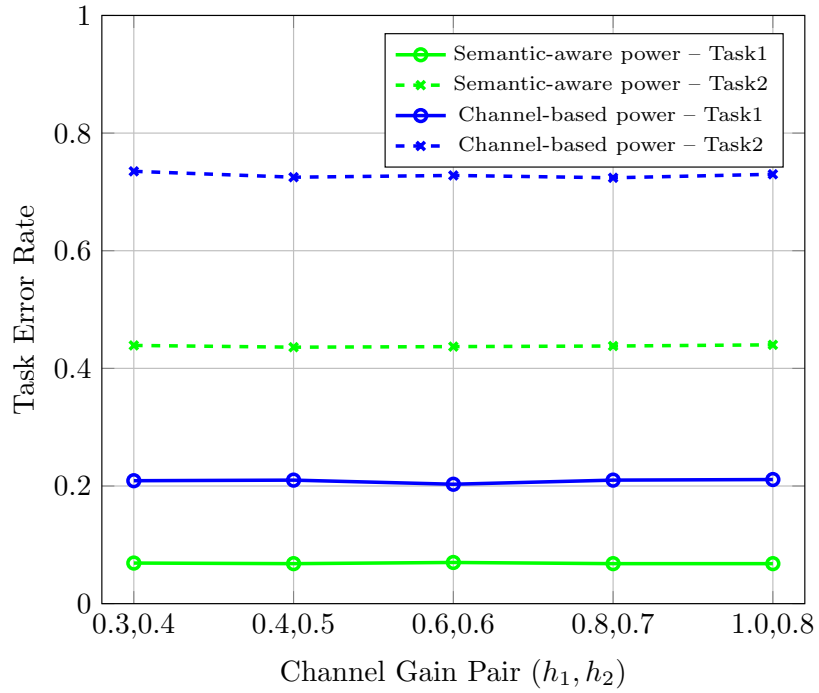


Figure 7.2: Task error rate comparison of semantic-aware power allocation vs. channel-based power allocation for digital communication design.

In this experiment, semantic-aware power allocation of Model 1 in 6.1.1 is compared against a channel-based power allocation strategy.

In the semantic-aware setup, power is individually assigned to each task based on its required semantic value. In the channel-based setup, power is allocated solely based on channel condition.

Figure 7.2 shows the simulation result for the task error rate comparison of semantic-aware power allocation vs. channel-based power allocation for the digital communication design.

Simulation Results

- Under semantic-aware allocation, both tasks maintain consistent error rates throughout all channel conditions.
- Channel-based allocation results in higher error rates, especially for Task 2.
- Task 1 also shows slightly higher errors under the channel-based approach.

General Observations

- The results highlight a limitation of channel-based allocation. It assumes that allocating power only based on channel conditions is sufficient, which is not the case in semantic communications.
- Semantic-aware allocation ensures higher accuracy by translating semantic requirements into physical-layer parameters.
- The higher error rate in the channel-based approach is due to its inability to account for the semantic complexity of task 2.
- These findings confirm that semantic-aware power allocation achieves higher accuracy as it assigns power not only based on channel conditions but also based on semantic value.

Case 3: Allocated power comparison of semantic-aware power allocation vs. channel-based power allocation for digital communication design

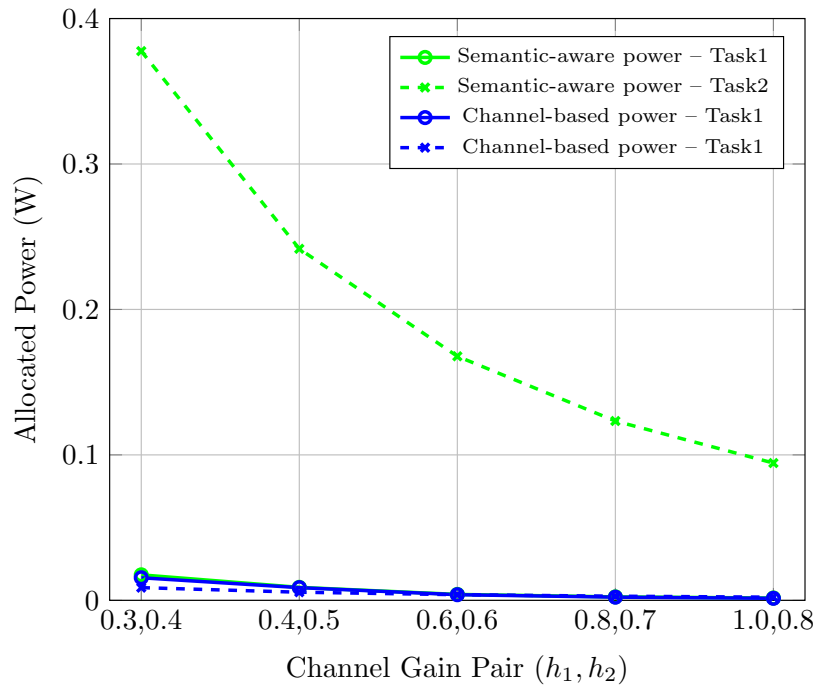


Figure 7.3: Allocated power comparison of semantic-aware power allocation vs. channel-based power allocation for digital communication design

In this experiment, we compare the transmit power allocation under semantic-aware structure of Model 1 in 6.1.1 and channel-based power strategy.

Figure 7.3 shows the simulation result for the allocated power comparison of semantic-aware power allocation vs. channel-based power allocation for the digital communication design.

Simulation Results

- As shown in the figure, the power allocated to Task 1 is identical in both strategies across all channel gain pairs. This is because the same semantic accuracy and BER target were used for Task 1 in both setups.
- The key difference lies in Task 2's power allocation. The semantic-aware method allocates substantially more power to Task 2 than the channel-based method, especially under low channel gain conditions.
- The power allocated to Task 2 under the channel-based strategy remains low and consistent across all channel conditions.

General Observations

- This result highlights a trade-off between power consumption and semantic performance.
- In order to meet the higher semantic accuracy required for Task 2, the semantic-aware strategy allocates more power.
- The channel-based strategy under-allocates to Task 2 because it ignores the semantic complexity.
- The consistently higher power assigned to Task 2 in the semantic-aware case ensures that task-level performance is preserved.

Case 4: Power Scaling Effect of semantic-aware power allocation for digital communication design

This experiment investigates the relationship between total transmit power and average task error rate under the semantic-aware power allocation scheme of Model 1 in 6.1.1. The aim is to determine whether increasing total transmit power beyond the computed semantic-optimal levels leads to improved performance.

The system uses fixed semantic targets for Task 1 and Task 2. Then computes their baseline powers accordingly. These base powers are then scaled by fixed multiplicative factors to simulate increasing total transmit power.

Figure 7.4 shows the simulation result for the power scaling Effect of semantic-aware power allocation for digital communication design.

Simulation Results

- The figure plots the average task error rate (mean of Task 1 and Task 2 errors) against increasing total transmit power.
- The baseline (scale = 1) corresponds to the minimum power required to satisfy the semantic targets for both tasks.

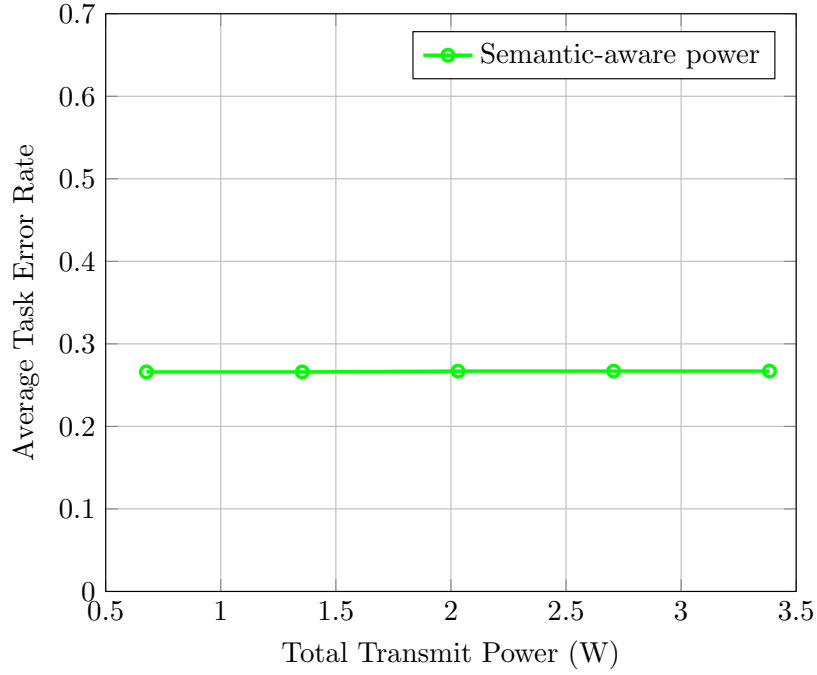


Figure 7.4: Power scaling effect of semantic-aware power allocation for digital communication design.

- As the power is scaled beyond this point ($2\times$, $3\times$, etc.), the average task error rate remains nearly constant and flat.
- No further reduction in error is observed despite significantly increasing the total transmit power.

General Observations

- This result confirms that the semantic-aware allocation strategy finds a power-optimal operating point. It allocates just enough power to meet semantic requirements without waste.
- This behavior demonstrates that semantic-aware communication systems not only improve task performance but also operate with power efficiency.

7.2.2 Analog E2E Design (Full-Observation)

Case 5: Task error rate comparison of semantic-aware power allocation vs. equally allocated power allocation for analog E2E design

In this experiment, we compare the performance of semantic-aware power allocation against a fixed equal power allocation strategy under the analog E2E design (Model 2 in 6.1.2), across different channel gain pairs (h_1, h_2) .

Figure 7.5 shows the simulation result for the task error rate comparison of semantic-aware power allocation vs. equally allocated power allocation for analog E2E design.

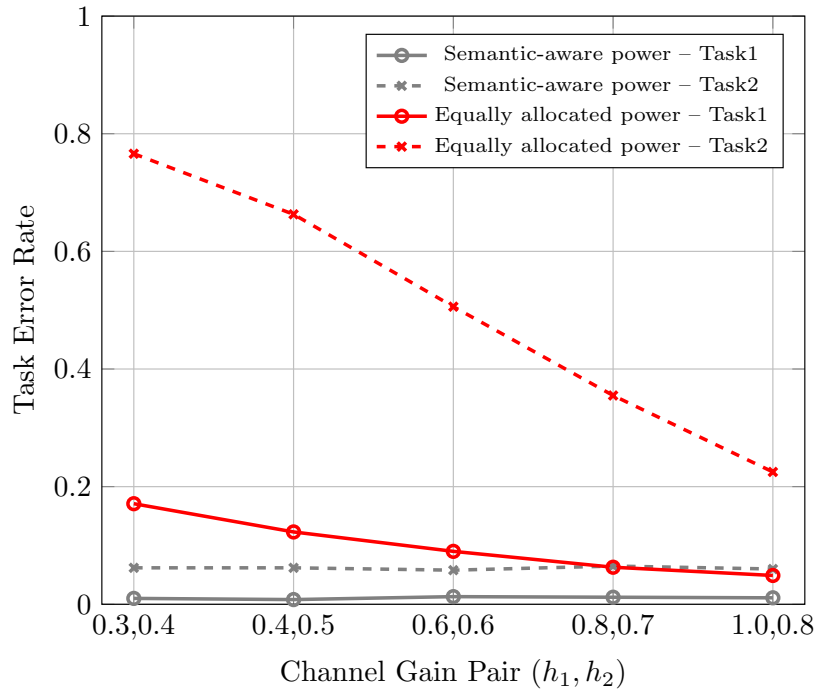


Figure 7.5: Task error rate comparison of semantic-aware power allocation vs. equally allocated power allocation for analog E2E design.

Simulation Results

- Task 1 and Task 2 error rates are plotted for five different channel gain scenarios.
- Under semantic-aware allocation, both Task 1 and Task 2 maintain consistently low error rates.
- In contrast, the equal power setup shows significantly higher error rates, especially for Task 2.
- Task 1 also performs worse under equal allocation.

General Observations

- Semantic-aware power allocation clearly outperforms equal allocation by adapting power to both semantic value and channel quality.
- Task 2 requires more power to maintain higher semantic performance. Semantic-aware allocation correctly allocates power for this but equal allocation underpowers it.
- The simpler Task 1 also benefits from tailored power.
- This case confirms that semantic-aware allocation performs better by allocating power according to semantic value and channel condition.

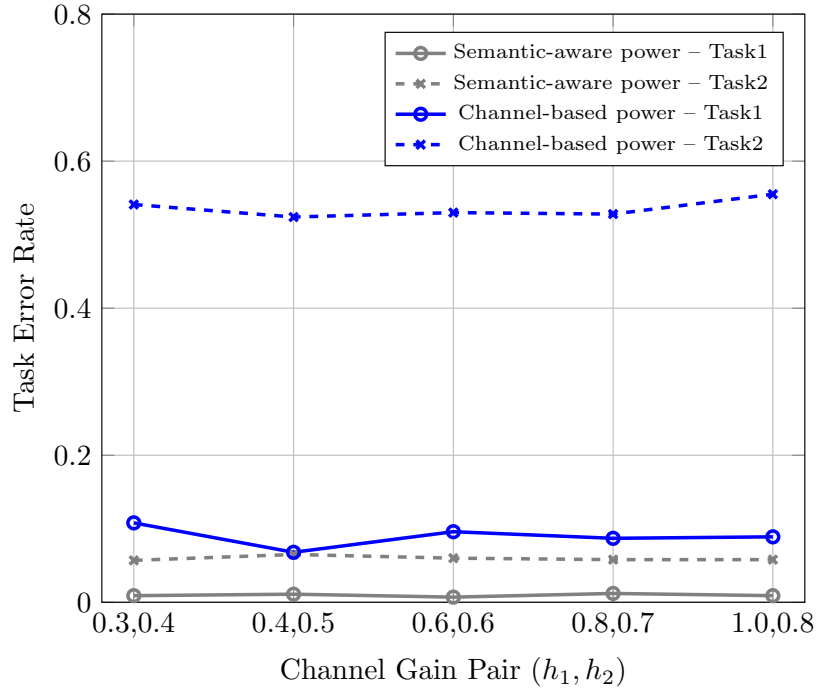


Figure 7.6: Task error rate comparison of semantic-aware power allocation vs. channel-based power allocation for analog E2E design

Case 6: Task error rate comparison of semantic-aware power allocation vs. channel-based power allocation for analog E2E design

In this experiment, we compare semantic-aware power allocation and channel-based power allocation for the analog E2E design (Model 2 in 6.1.2) across varying channel gain pairs (h_1, h_2).

In the semantic-aware setup, each task's power is determined based on a task-specific semantic value and channel condition. In contrast, the channel-based strategy each task's power is determined based on only and channel condition.

Figure 7.6 shows the simulation result for the task error rate comparison of semantic-aware power allocation vs. channel-based power allocation for the analog E2E design.

Simulation Results

- The figure presents task error rates for Task 1 and Task 2 under both allocation strategies.
- The semantic-aware scheme yields low and stable error rates for both tasks across all channel conditions.
- In the channel-based setup, both Task 1 and Task 2 show.

General Observations

- This result shows that task performance in semantic communication is not solely a function of BER or channel SNR, but instead depends on how well the transmission

supports the semantic value of each task.

- The channel-based strategy's relies on a fixed channel condition for semantically demanding tasks. This leads to high error rates.
- Semantic-aware allocation dynamically adjusts transmit power to meet semantic goals directly. This ensures that each task receives only as much power as needed to meet its purpose. This results in better accuracy and more efficient energy usage.

Case 7: Allocated power comparison of semantic-aware power allocation vs. channel-based power allocation for analog E2E design

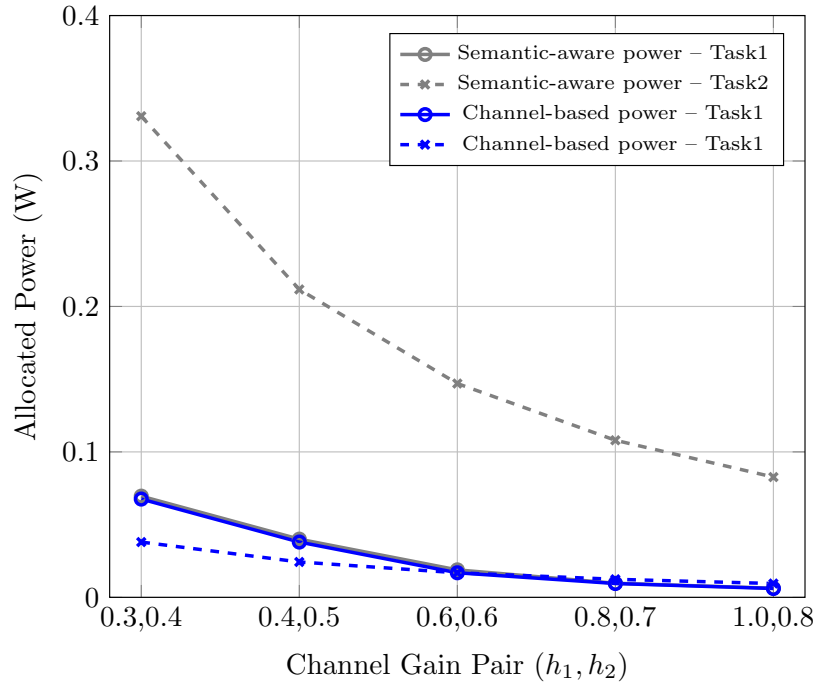


Figure 7.7: Allocated power comparison of semantic-aware power allocation vs. channel-based power allocation for analog E2E design

In this experiment, we analyze the allocated transmit power per task under semantic-aware and channel-based power allocation strategies for the analog E2E system (Model 2 in 6.1.2).

Figure 7.7 shows the simulation result for the allocated power comparison of semantic-aware power allocation vs. channel-based power allocation for the analog E2E design.

Simulation Results

- The plot displays the power allocated to Task 1 and Task 2 across five different channel gain pairs.
- Under semantic-aware allocation, Task 2 consistently receives significantly more power than Task 1.
- For Task 1, both allocation schemes yield similar values because the semantic requirement and MSE mapping used in both approaches for this task are nearly aligned.

- The channel-based strategy allocates lower and more uniform power to both tasks.

General Observations

- This result highlights the power-performance trade-off inherent in semantic-aware allocation. To meet stricter semantic goals, the system allocates more power.
- The channel-based strategy appears more energy-efficient but results in semantic underperformance.
- Overall, this experiment confirms that semantic-aware systems prioritize meaningful task completion, even at the cost of higher power.

Case 8: Power Scaling Effect of semantic-aware power allocation for analog E2E design

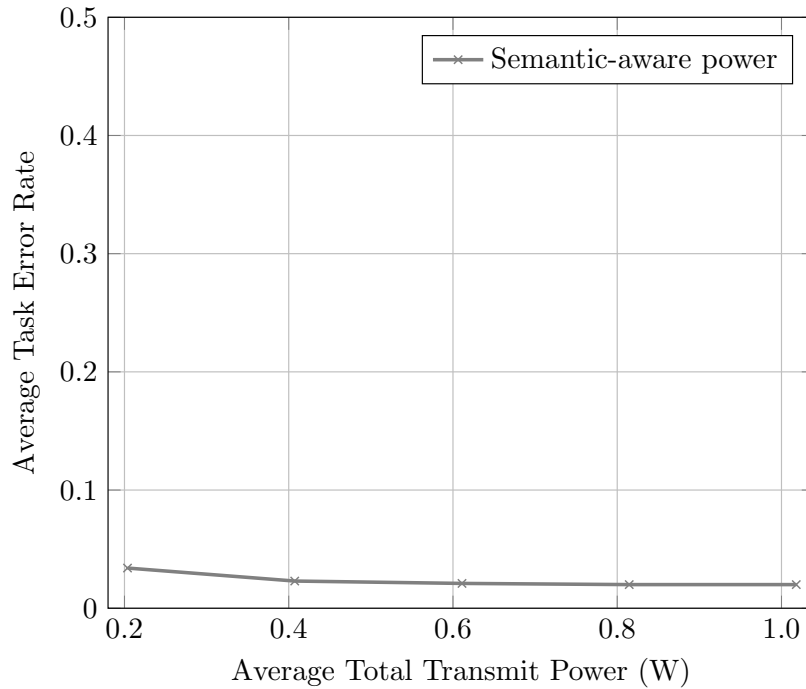


Figure 7.8: Power Scaling Effect of semantic-aware power allocation for analog E2E design.

This experiment investigates the effect of scaling total transmit power under semantic-aware allocation in the analog E2E design (Model 2 in 6.1.2).

Figure 7.8 shows the simulation result for the power scaling Effect of semantic-aware power allocation for analog E2E design.

Simulation Results

- The figure plots the average task error rate (mean of Task 1 and Task 2 errors) against increasing total transmit power.
- The baseline (scale = 1) corresponds to the minimum power required to satisfy the semantic targets for both tasks.

- As the power is scaled beyond this point ($2\times$, $3\times$, etc.), the average task error rate remains nearly constant and flat.
- No further reduction in error is observed despite significantly increasing the total transmit power.

General Observations

- This result confirms that the semantic-aware allocation strategy finds a power-optimal operating point. It allocates just enough power to meet semantic requirements without waste.
- This behavior demonstrates that semantic-aware communication systems not only improve task performance but also operate with power efficiency.

7.2.3 Digital vs. Analog Design Comparison (Full-Observation)

Case 9: Task error rate of analog E2E and digital communication enabled semantic-aware power allocation structure comparison

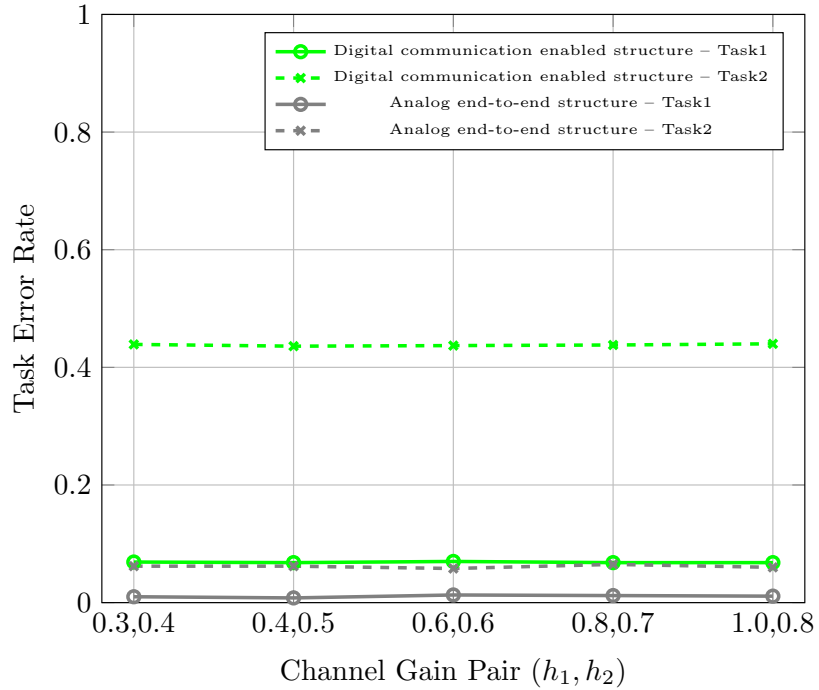


Figure 7.9: Task error rate of analog E2E and digital communication enabled semantic-aware power allocation structure comparison.

This experiment compares the performance of two semantic-aware communication structures: the Digital Communication Design (Model 1 in 6.1.1) and the Analog E2E design (Model 2 in 6.1.2).

In both models, power is allocated semantically based on task-specific semantic values. However, the digital model quantizes and maps latent features to discrete symbols for digital

channel transmission. The analog model directly transmits continuous-valued semantic representations through the physical channel, without quantization or digital modulation.

Figure 7.9 shows the simulation result for task error rate of analog E2E and digital communication enabled semantic-aware power allocation structure comparison.

Simulation Results

- The figure presents task error rates for both Task 1 and Task 2 across five different channel gain conditions.
- The analog structure consistently outperforms the digital structure for both tasks.
- For Task 2, the analog system maintains low and consistent error rates, while the digital system shows significantly higher errors.

General Observations

- These results highlight the superiority of the analog E2E design in preserving semantic fidelity and achieving lower task error rates.
- One key factor is the training methodology. The analog structure is trained E2E as a unified system, allowing the encoder, channel interaction, and decoder to co-adapt jointly for optimal semantic reconstruction.
- In contrast, the digital structure is trained in two stages. The encoder and decoder are first trained separately, and then reused in a digital communication design. This staged approach introduces a mismatch between training and deployment conditions.
- The analog model avoids losses by directly transmitting continuous semantic features.

7.2.4 Digital Communication Design (Partial-Observation)

Case 10: Task error rate comparison of semantic-aware power allocation vs. equally allocated power allocation for digital communication design with distributed partial observation setting

In this experiment, the performance of semantic-aware power allocation structure with distributed partial observation setting (Model 3 in 6.2.1) is compared against a baseline where equal transmit power is allocated to both tasks, across varying channel gain pairs (h_1, h_2) .

Figure 7.10 shows the simulation result of the task error rate comparison of semantic-aware power allocation vs. equally allocated power allocation for digital communication design with distributed partial observation setting.

Simulation Results

- Task error rates for both Task 1 and Task 2 are plotted against increasing channel gain pairs.
- Under semantic-aware allocation, both tasks maintain stable and low error rates.

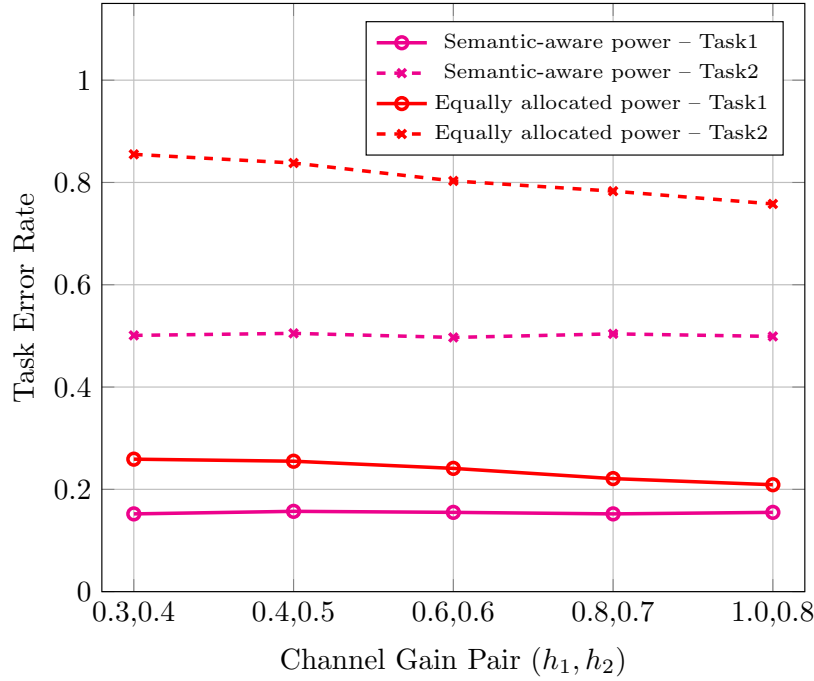


Figure 7.10: Task error rate comparison of semantic-aware power allocation vs. equally allocated power allocation for digital communication design with distributed partial observation setting

- Under equal power allocation:
 - Task 1 performs reasonably well and improves slightly with better channel gains.
 - Task 2 shows significantly higher error rates at lower channel gains.
- The performance gap between semantic-aware and equal-power setups is especially visible for Task 2.

General Observations

- The semantic-aware allocation dynamically adjusts transmit power based on task importance and channel quality. This ensures consistent task performance across all scenarios.
- Task 2, which requires higher semantic precision due to its multi-class nature, benefits significantly from the semantic-aware power allocation.
- Task 1 requires less power to achieve its semantic goal. The equal-power system often over-allocates power to it. This wastes power resources.

Case 11: Task error rate comparison of semantic-aware power allocation vs. channel-based power allocation for digital communication design with distributed partial observation setting

In this experiment, semantic-aware power allocation of Model 3 in 6.2.1 is compared against a channel-based power allocation strategy.

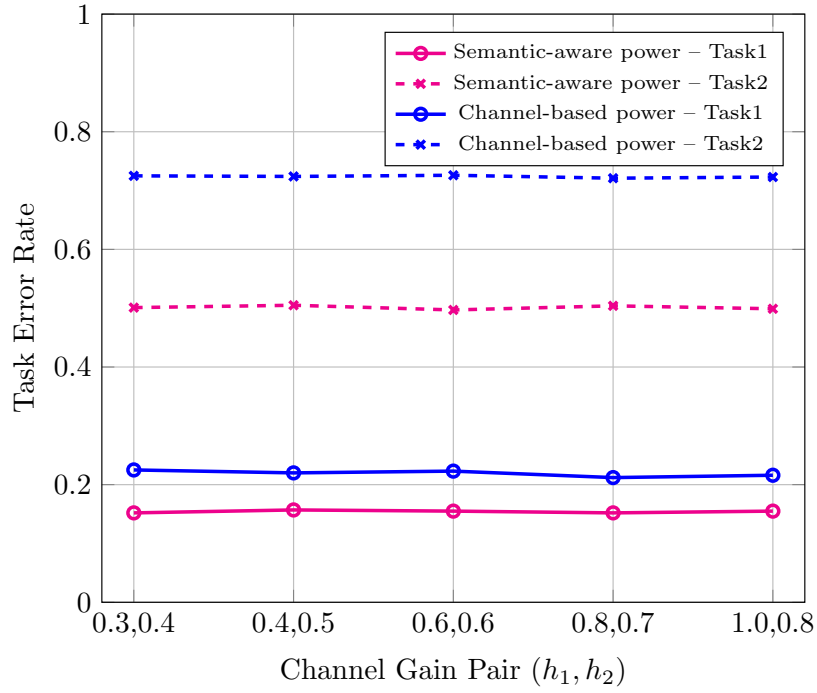


Figure 7.11: Task error rate comparison of semantic-aware power allocation vs. channel-based power allocation for digital communication design with distributed partial observation setting

Figure 7.11 shows the simulation result for the task error rate comparison of semantic-aware power allocation vs. channel-based power allocation for the digital communication design.

Simulation Results

- Task error rates for Task 1 and Task 2 are shown for both allocation strategies across five different channel gain combinations.
- Under semantic-aware allocation, both tasks maintain consistent error rates throughout all channel conditions.
- Channel-based allocation results in noticeably higher error rates, especially for Task 2.

General Observations

- The results highlight a limitation of channel-based allocation. It assumes that allocating power only based on channel conditions is sufficient.
- Semantic-aware allocation translates semantic requirements into physical-layer parameters which ensures higher accuracy.
- These findings confirm that semantic-aware power allocation achieves higher accuracy by allocating power based on both semantic value and channel condition.

Case 12: Allocated power comparison of semantic-aware power allocation vs. channel-based power allocation for digital communication design with distributed partial observation setting

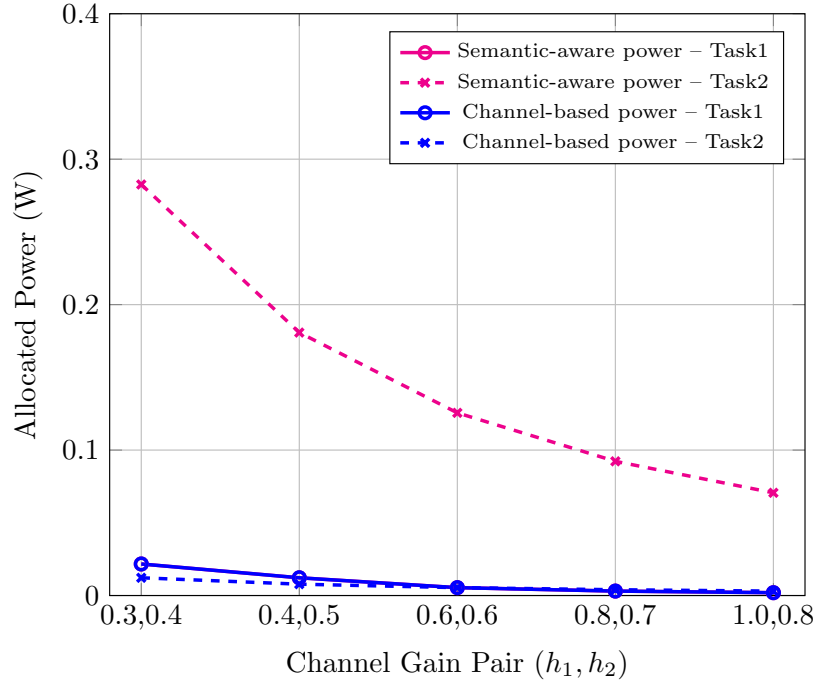


Figure 7.12: Allocated power comparison of semantic-aware power allocation vs. channel-based power allocation for digital communication design with distributed partial observation setting

In this experiment, we compare the transmit power allocation under semantic-aware structure of Model 3 in 6.2.1 and channel-based strategies.

Figure 7.12 shows the simulation result for the allocated power comparison of semantic-aware power allocation vs. channel-based power allocation for the digital communication design.

Simulation Results

- The power allocated to Task 1 is identical in both strategies across all channel gain pairs. This is because the same semantic accuracy and BER target were used in both setups.
- The key difference lies in Task 2's power allocation. The semantic-aware method allocates substantially more power to Task 2, especially under low channel gain conditions.
- Meanwhile, the power allocated to Task 2 under the channel-based strategy remains low and consistent across all channel conditions.

General Observations

- This result highlights a critical trade-off between power consumption and semantic performance.

- In order to meet the higher semantic accuracy required for Task 2, the semantic-aware strategy must allocate more power, particularly under weaker channel conditions.
- The channel-based strategy under-allocates to Task 2 because it ignores the semantic complexity.
- This experiment clearly illustrates that semantic-aware communication prioritizes semantic value over required power.

Case 13: Power Scaling Effect of semantic-aware power allocation for digital communication design with distributed partial observation setting

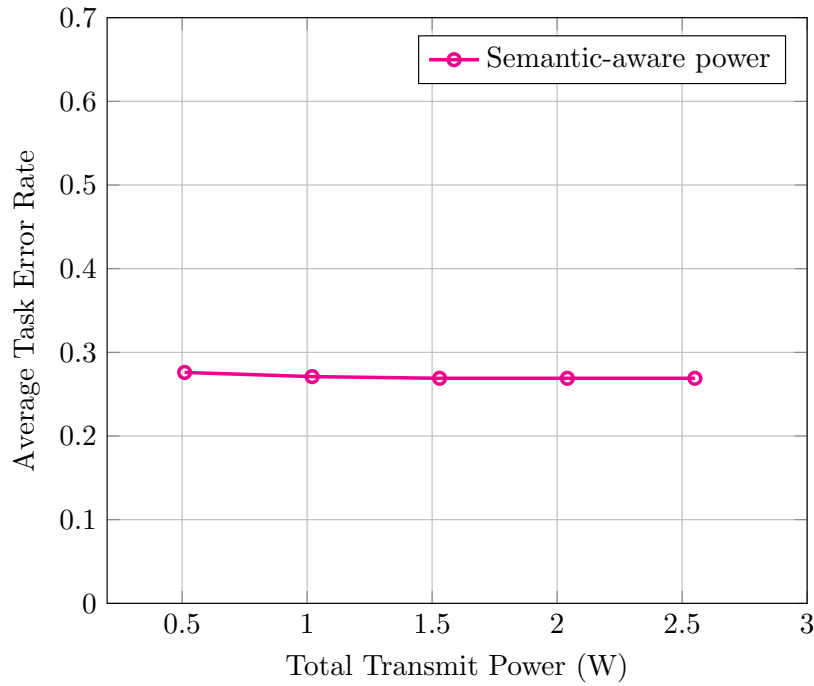


Figure 7.13: Power Scaling Effect of semantic-aware power allocation for digital communication design with distributed partial observation setting

This experiment investigates the relationship between total transmit power and average task error rate under the semantic-aware power allocation scheme of Model 3 in 6.2.1.

Figure 7.13 shows the simulation result for the power scaling Effect of semantic-aware power allocation for digital communication design.

Simulation Results

- The figure plots the average task error rate (mean of Task 1 and Task 2 errors) against increasing total transmit power.
- The baseline (scale = 1) corresponds to the minimum power required.
- As the power is scaled beyond this point (2×, 3×, etc.), the average task error rate remains nearly constant and flat.

General Observations

- This result confirms that the semantic-aware allocation strategy finds a power-optimal operating point. It allocates just enough power to meet semantic requirements without waste.

7.2.5 Analog E2E Design (Partial-Observation)

Case 14: Task error rate comparison of semantic-aware power allocation vs. equally allocated power allocation for analog E2E design with distributed partial observation setting

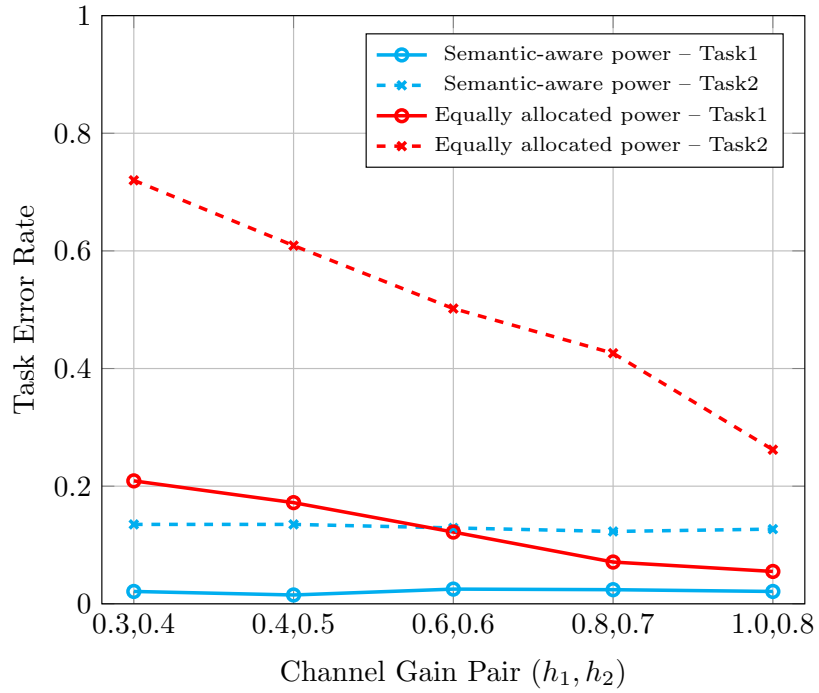


Figure 7.14: Task error rate comparison of semantic-aware power allocation vs. equally allocated power allocation for analog communication enabled structure with distributed partial observation setting

In this experiment, we compare the performance of semantic-aware power allocation against a fixed equal power allocation strategy under the analog E2E design (Model 4 in 6.2.2).

Figure 7.14 shows the simulation result for the task error rate comparison of semantic-aware power allocation vs. equally allocated power allocation for analog E2E design.

Simulation Results

- Task 1 and Task 2 error rates are plotted for five different channel gain scenarios.
- Under semantic-aware allocation, both Task 1 and Task 2 maintain consistently low error rates.
- In contrast, the equal power setup shows significantly higher error rates.

General Observations

- Semantic-aware power allocation clearly outperforms equal allocation by adapting power to both semantic value and channel quality.

Case 15: Task error rate comparison of semantic-aware power allocation vs. channel-based power allocation for analog E2E design with distributed partial observation setting

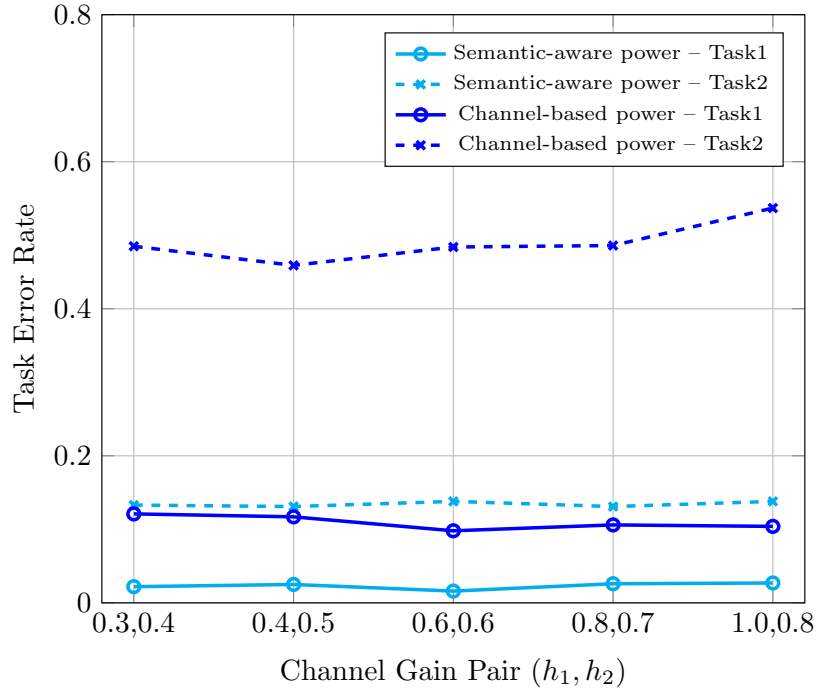


Figure 7.15: Task error rate comparison of semantic-aware power allocation vs. channel-based power allocation for analog E2E design with distributed partial observation setting

In this experiment, we compare semantic-aware power allocation and channel-based power allocation for the analog E2E design (Model 4 in 6.2.2).

Figure 7.15 shows the simulation result for the task error rate comparison of semantic-aware power allocation vs. channel-based power allocation for the analog E2E design.

Simulation Results

- The figure presents task error rates for Task 1 and Task 2 under both allocation strategies.
- The semantic-aware scheme yields low and stable error rates for both tasks.
- In the channel-based setup, Task 1 and Task 2 shows moderately higher error rates.

General Observations

- This result reinforces that task performance in semantic communication is not solely a function of BER or channel SNR.
- Semantic-aware allocation dynamically adjusts transmit power to meet semantic goals directly. It ensures that each task receives only as much power as needed to meet its semantic purpose.

Case 16: Allocated power comparison of semantic-aware power allocation vs. channel-based power allocation for analog E2E design with distributed partial observation setting

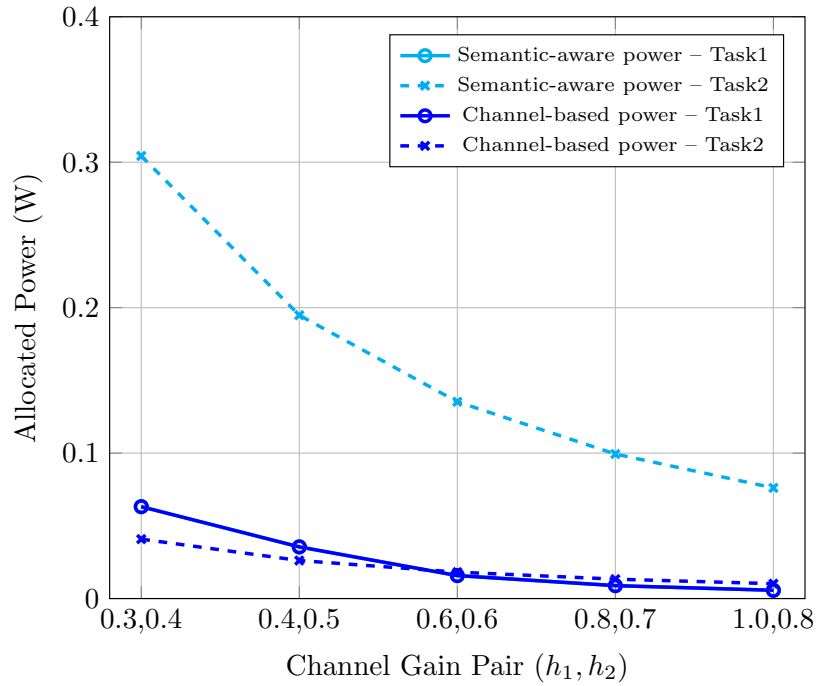


Figure 7.16: Allocated power comparison of semantic-aware power allocation vs. channel-based power allocation for analog E2E design with distributed partial observation setting

In this experiment, we analyze the allocated transmit power per task under semantic-aware and channel-based power allocation strategies for the analog E2E system (Model 4 in 6.2.2).

Figure 7.16 shows the simulation result for the allocated power comparison of semantic-aware power allocation vs. channel-based power allocation for the analog E2E design.

Simulation Results

- The plot displays the power allocated to Task 1 and Task 2 across five different channel gain pairs.
- Under semantic-aware allocation, Task 2 consistently receives significantly more power than Task 1.
- For Task 1, both allocation schemes yield similar values, as the semantic requirement and MSE mapping used in both approaches for this task are nearly aligned.

- The channel-based strategy allocates lower and more uniform power to both tasks.

General Observations

- This result highlights the power-performance trade-off inherent in semantic-aware allocation.
- The channel-based strategy appears more energy-efficient on the surface but results in semantic underperformance, as seen in prior experiments.
- This experiment confirms that semantic-aware systems prioritize meaningful task completion, even at the cost of higher power.

Case 17: Power Scaling Effect of semantic-aware power allocation for analog E2E design with distributed partial observation setting

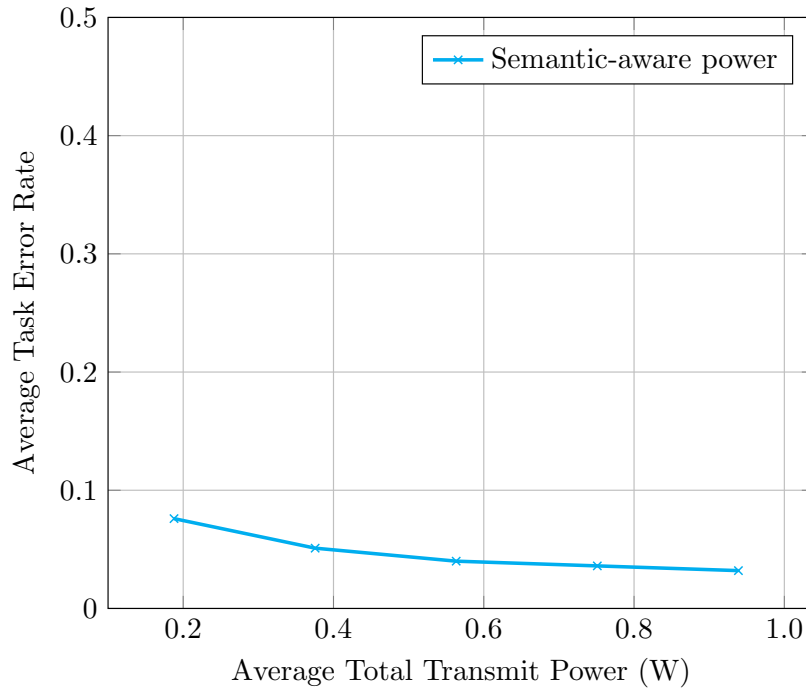


Figure 7.17: Power Scaling Effect of semantic-aware power allocation for analog E2E design with distributed partial observation setting

This experiment investigates the effect of scaling total transmit power under semantic-aware allocation in the analog eE2E design (Model 4 in 6.2.2).

Figure 7.17 shows the simulation result for the power scaling Effect of semantic-aware power allocation for analog E2E design.

Simulation Results

- The figure plots the average task error rate (mean of Task 1 and Task 2 errors) against increasing total transmit power.

- The baseline (scale = 1) corresponds to the minimum power required to satisfy the semantic targets.
- As the power is scaled beyond this point ($2\times$, $3\times$, etc.), the average task error rate remains nearly constant and flat.

General Observations

- This result confirms that the semantic-aware allocation strategy finds a power-optimal operating point.
- This behavior demonstrates that semantic-aware communication systems not only improve task performance but also operate with power efficiency.

7.2.6 Digital vs. Analog Design Comparison (Partial-Observation)

Case 18: Task error rate of analog E2E and digital communication enabled semantic-aware power allocation structure comparison with distributed partial observation setting

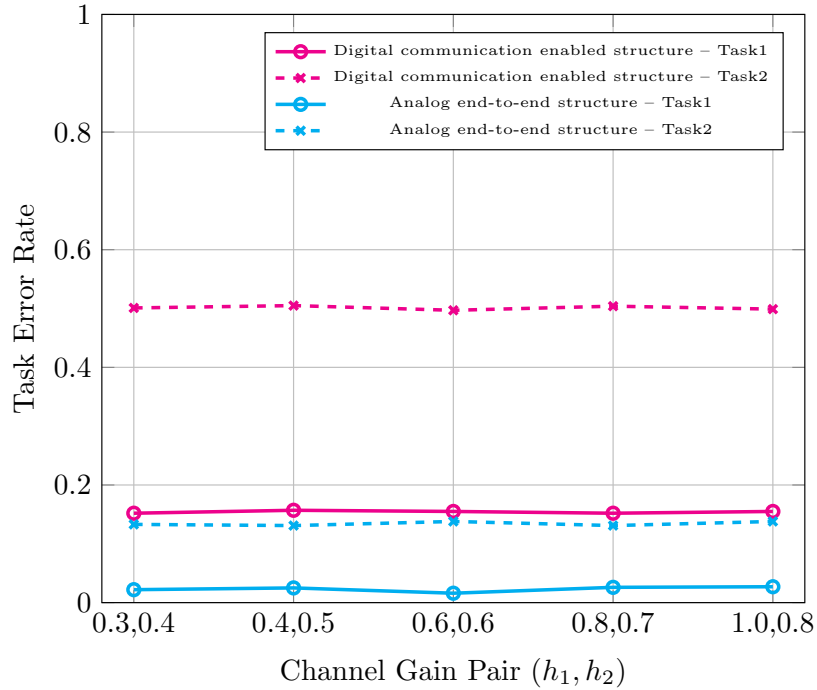


Figure 7.18: Task error rate of analog E2E and digital communication enabled semantic-aware power allocation structure comparison with distributed partial observation setting

This experiment compares the performance of two semantic-aware communication structures: the Digital Communication Design (Model 3 in 6.2.1) and the Analog E2E Design (Model 4 in 6.2.2).

Figure 7.18 shows the simulation result for task error rate of analog E2E and digital communication enabled semantic-aware power allocation structure comparison.

Simulation Results

- The figure presents task error rates for both Task 1 and Task 2 across five different channel gain conditions.
- The analog structure consistently outperforms the digital structure for both tasks.
- For Task 2, the analog system maintains low and consistent error rates, while the digital system shows significantly higher errors.

General Observations

- These results highlight the superiority of the analog E2E design in achieving lower task error rates.
- One key factor is the training methodology. The analog structure is trained E2E as a unified system. This allows the encoder, channel interaction, and decoder to co-adapt jointly for optimal semantic reconstruction.
- In contrast, the digital structure is trained in two stages. The encoder and decoder are first trained separately, and then reused in a digital communication design. This staged approach introduces a mismatch between training and deployment conditions.
- This training difference coupled with structural simplicity and continuous representation explains why the analog model achieves superior semantic outcomes.

7.2.7 Full vs. Partial Observation Comparison

Case 19: Task error rate comparison of digital communication enabled semantic-aware power allocation structure with and without distributed partial observation setting

We compare two digital semantic communication designs under identical conditions:

- **Full-observation (Model 1 in 6.1.1):** each encoder processes the entire MNIST image before quantization and digital transmission.
- **Partial-observation (Model 3 in 6.2.1):** two encoders each handle only a half-image (14×28 top or bottom patch) prior to encoding.

Both systems reconstruct Task 1 and Task 2 at a common receiver using a unified DNN decoder.

Figure 7.19 shows the simulation result for the task error rate comparison of digital communication enabled semantic-aware power allocation structure with and without distributed partial observation setting.

Simulation Results

- For Task 1, both full- and partial-observation curves show low task error rates. The full observation setting performs slightly better in terms of task error rate.
- The full observation setting performs slightly better in terms of task error rate also for Task 2.

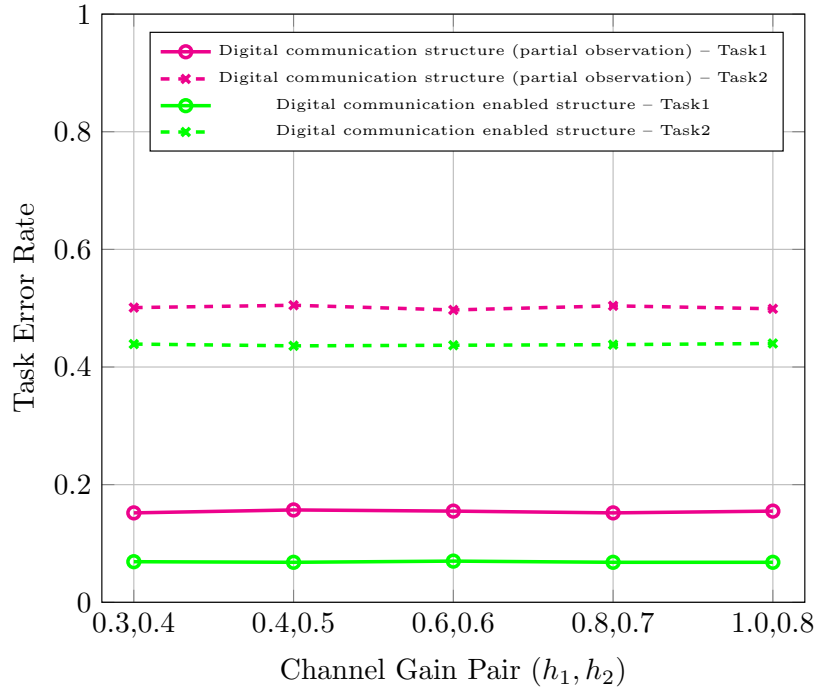


Figure 7.19: Task error rate of digital communication design with and without distributed partial observation setting semantic-aware power allocation structure comparison.

General Observation

- By splitting the image, each encoder conveys fewer features. So, the decoder needs somewhat stronger channel conditions to match full-view accuracy.
- Even with this modest degradation, the partial-observation setting still reconstructs the semantic variables nearly as effectively as the full-view case.

Case 20: Task error rate of analog E2E with and without distributed partial observation setting semantic-aware power allocation structure comparison

General Description This case evaluates two analog, E2E semantic communication pipelines under identical channel conditions:

- **Full-observation (Model 2 in 6.1.2):** each encoder processes the 28×28 MNIST image into continuous-valued features before transmission.
- **Partial-observation (Model 4 in 6.2.2):** two encoders each observe only a 14×28 half-image (top or bottom patch).

Both setups jointly reconstruct Task 1 and Task 2 at a shared receiver via a DNN NOMA decoder.

Simulation Results

- For Task 1, both full- and partial-observation curves show low task error rates. The full observation setting performs slightly better in terms of task error rate.

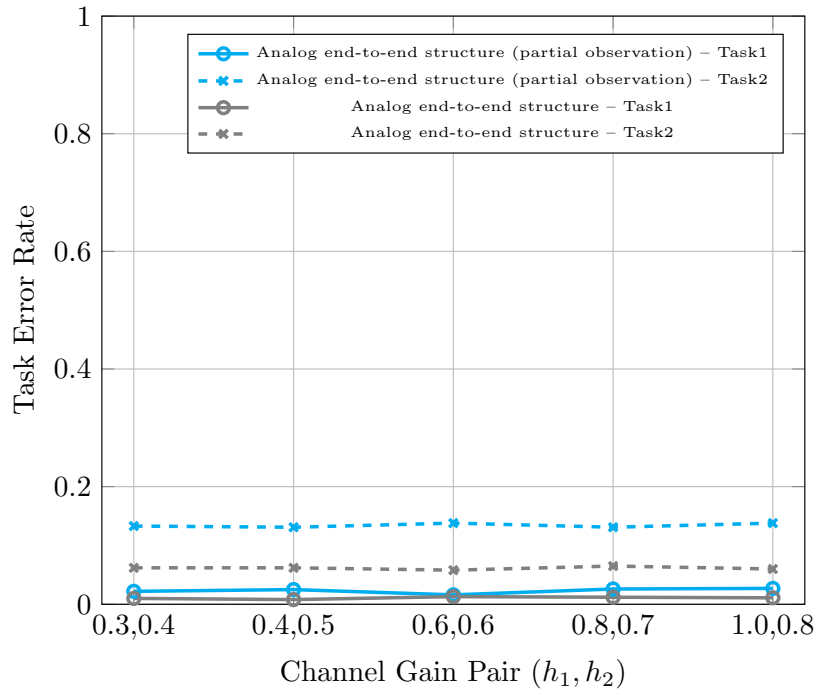


Figure 7.20: Task error rate of analog E2E with and without distributed partial observation setting semantic-aware power allocation structure comparison

- The full observation setting performs slightly better in terms of task error rate also for Task 2.

General Observation

- With only half the pixels per encoder, the decoder receives fewer analog features.
- Despite this modest performance gap, partial observation still collectively reconstructs the semantic variables with performance close to the full-observation setting.

Chapter 8

Conclusion

This thesis investigated receiver side cooperative processing by integrating DNN-NOMA with CMT-SemCom framework. It developed the CMT-SemCom enabled by DNN-NOMA architecture with a particular emphasis on semantic-aware power allocation. The motivation behind this work stemmed from the increasing demands for intelligent decision-making in modern wireless communication systems, emphasizing task-specific communication rather than mere accurate transmission of bits. Traditional bit-centric communication models are insufficient in meeting the requirements of emerging applications such as autonomous vehicles, IoT networks, and intelligent edge devices. These emerging applications require efficient semantic and task-oriented communication strategies.

In addressing these needs, three central research questions guided the research presented in this thesis. The first question focused on the integration of DNN-NOMA to support CMT-SemCom for shifting cooperative processing from transmitter to receiver. The second examined how transmission power can be optimally allocated based on both semantic requirements and channel conditions of multiple tasks. The third question examined performance trade-offs of this model for a practical scenario in which each user observes only a portion of the source image against an ideal scenario in which all users have access to the complete image.

To systematically address these questions, the thesis first reviewed the foundational concepts and recent advancements in semantic communication and NOMA. Semantic communication was explored through various existing approaches, including classical semantic information theory, knowledge graph methods, machine learning techniques, significance-based approaches, and information-theoretic frameworks. This literature survey highlighted key limitations in existing methodologies in case of treating cooperative multi-task scenarios. A novel CMT-SemCom architecture comprising a CU and SUs by [RBD24] was described as a promising solution, allowing both cooperative processing and task-specific semantic communication.

Subsequently, the thesis reviewed NOMA with its advantages over traditional OMA techniques. NOMA facilitates simultaneous transmission of multiple users' signals through power-domain multiplexing. It offers significant improvements in spectral efficiency, connectivity, latency, and compatibility with existing and future wireless systems. The review highlighted key NOMA decoding techniques and surveyed power-allocation strategies for multi-task scenarios.

Following the literature review, a novel integration framework was presented that combined CMT-SemCom with DNN-NOMA. This architecture utilized a DNN-based decoder at the receiver to shift cooperative processing from transmitter to the receiver in case of resource-constrained edge devices.

Another core contribution of this thesis was the introduction of semantic-aware power allocation methods that integrated semantic importance into power allocation decisions. Unlike

conventional power allocation strategies that mainly consider channel conditions, the proposed semantic-aware methods allocated resources based on task-specific semantic metrics. Semantic accuracy requirements were mapped to physical-layer parameters, resulting in formulations that guided optimal power distribution.

The architecture was analyzed in digital and analog transmission contexts, first under idealized full-observation conditions, and then extended to more practical distributed partial-observation settings, simulating realistic constraints of edge-based distributed sensing environments.

The detailed simulation studies provided comprehensive validations and performance comparisons under various configurations and channel conditions. Key results demonstrated substantial improvements in task performance when using semantic-aware power allocation compared to traditional channel-only or equally allocated schemes. In particular, semantic-aware methods significantly enhanced resource efficiency, improved semantic fidelity, and maintained performance even under stringent power constraints or distributed partial-observation settings.

The detailed simulation results provided comprehensive performance comparisons and addressed our three core research questions:

- **Receiver-side cooperative processing (RQ1):** By integrating DNN-NOMA with the CMT-SemCom framework, cooperative processing can be shifted to the receiver side.
- **Semantic-aware power allocation (RQ2):** Semantic-aware power allocation consistently outperformed conventional power allocation strategies in terms of task error rates.
- **Partial vs. full observation (RQ3):** In distributed partial-observation scenarios, our semantic-aware CMT-SemCom enabled by DNN-NOMA framework still reconstructs semantic variables with performance nearly matching the full-observation case.

Future research directions could include further optimization of neural architectures, exploration of more sophisticated semantic metrics, and development of adaptive schemes for dynamically evaluating semantic value. Additionally, incorporating more real-world datasets and testbeds for validation could further demonstrate the practical applicability of the proposed system model.

In conclusion, this thesis has demonstrated that by integrating DNN-NOMA with a CMT-SemCom framework, it is possible to shift the cooperative processing to the receiver side for small edge devices. In addition, using semantic-aware power allocation can decrease the task error rate for this CMT-SemCom enabled by DNN-NOMA structure. Moreover, even in distributed partial-observation scenarios the proposed framework recovers semantic variables nearly as effectively as the ideal full-observation case. These results validate the viability of this semantic-aware CMT-SemCom enabled by DNN-NOMA framework for next-generation wireless systems.

Acronyms

IoT	Internet of Things
NOMA	Non-Orthogonal Multiple Access
CU	Common Unit
SU	Specific Unit
DNN	Deep Neural Network
SNR	Signal-to-Noise Ratio
AI	Artificial Intelligence
5G	Fifth Generation (wireless technology)
6G	Sixth Generation (wireless technology)
SemCom	Semantic Communication
QoS	Quality of Service
BER	Bit Error Rate
MSE	Mean Squared Error
QoS	Quality of Service
M2M	Machine-to-Machine
H2M	Human-to-Machine
IB	Information Bottleneck
CMT-SemCom	Cooperative Multi-Task Semantic Communication
KG	Knowledge Graph
ML	Machine Learning
DL	Deep Learning
NN	Neural Network
E2E	End-to-End
SC	Superposition Coding
BS	Base Station
SIC	Successive Interference Cancellation
eMBB	Enhanced Mobile Broadband

mMTC	Massive Machine-Type Communications
URLLC	Ultra-Reliable Low-Latency Communication
OMA	Orthogonal Multiple Access
FDMA	Frequency Division Multiple Access
TDMA	Time Division Multiple Access
CDMA	Code Division Multiple Access
OFDM	Orthogonal Frequency Division Multiplexing
AWGN	Additive White Gaussian Noise
MNIST	Modified National Institute of Standards and Technology
FC	Fully Connected
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
Tx	Transmitter
Rx	Receiver
CSI	Channel State Information

List of Symbols

ϕ	NN parameters for the SU
ψ	NN parameters for the DNN NOMA Decoder
θ	NN parameters for the CU
$\hat{\mathbf{z}}$	Reconstructed Semantic Variable
\hat{z}_i	Reconstructed Semantic Variable for i-th Task
\mathbf{c}	Common Encoded Information
\mathbf{M}_{noma}	Superimposed Signal
\mathbf{n}	Noise
\mathbf{S}	Input Observation
\mathbf{x}_i	Encoded Task-Specific Information for i-th Task
\mathbf{y}	Received Signal at Base Station
\mathbf{z}	Semantic Variable
h_i	Channel gain i-th user
q_i	Power Allocation Coefficients for i-th Task
z_i	Semantic Variable for i-th Task

Bibliography

- [3GP21] 3GPP. 5g system: Technical realization of service based architecture, 2021. [Online] https://www.3gpp.org/ftp/Specs/archive/29_series/29.500/. 6
- [ADI⁺12] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman. A survey of information-centric networking. *IEEE Communications Magazine*, 50(7):26–36, 2012. 6
- [ATG⁺18] M. Aldababsa, M. Toka, S. Gökçeli, G. G. K. Kurt, and O. L. Kucur. A tutorial on non-orthogonal multiple access for 5g and beyond, 2018. VII, 14, 16, 17, 18
- [BBD23] Edgar Beck, Carsten Bockelmann, and Armin Dekorsy. Semantic information recovery in wireless networks. *Sensors*, 23(14), 2023. 48
- [Bro17] G. Brown. Serviced-based architecture for 5g core network, 2017. [Online] https://www.3gpp.org/ftp/Specs/archive/29_series/29.500/. 6
- [DFP16] Zhiguo Ding, Pingzhi Fan, and H. Vincent Poor. Impact of user pairing on 5g nonorthogonal multiple-access downlink transmissions. *IEEE Transactions on Vehicular Technology*, 65(8):6010–6023, 2016. 34
- [DH05] Xitirnin Deng and A.M. Haimovich. Power allocation for cooperative relaying in wireless networks. *IEEE Communications Letters*, 9(11):994–996, 2005. 32
- [DWD⁺18] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo. A survey of non-orthogonal multiple access for 5g. *IEEE Communications Surveys & Tutorials*, 20(3):2294–2323, 2018. 2, 15, 16
- [DWY⁺15] L. Dai, B. Wang, Y. Yuan, S. Han, C. I, and Z. Wang. Nonorthogonal multiple access for 5g: solutions, challenges, opportunities, and future research trends. *IEEE Communications Magazine*, 53(9):74–81, 2015. 14
- [DYFP14] Zhiguo Ding, Zheng Yang, Pingzhi Fan, and H. Vincent Poor. On the performance of non-orthogonal multiple access in 5g systems with randomly deployed users. *IEEE Signal Processing Letters*, 21(12):1501–1505, 2014. 33
- [FU08] Muhammad Mehboob Fareed and Murat Uysal. Ber-optimized power allocation for fading relay channels. *IEEE Transactions on Wireless Communications*, 7(6):2350–2359, 2008. 32
- [Gav13] Henri P. Gavin. The levenberg-marquardt method for nonlinear least squares curve-fitting problems c ©. 2013. 41, 46, 49, 52
- [GQA⁺23] Deniz Gündüz, Zhijin Qin, Inaki Estella Aguerri, Harpreet S. Dhillon, Zhaohui Yang, Aylin Yener, Kai Kit Wong, and Chan-Byoung Chae. Beyond transmitting bits: Context, semantics, and task-oriented communications. *IEEE Journal on Selected Areas in Communications*, 41(1):5–41, 2023. VII, 2, 7, 9

- [HA03] M.O. Hasna and M.-S. Alouini. Optimal power allocation for relayed transmissions over rayleigh fading channels. In *The 57th IEEE Semiannual Vehicular Technology Conference, 2003. VTC 2003-Spring.*, volume 4, pages 2461–2465 vol.4, 2003. 32
- [JQC⁺04] Zhang Jingmei, Zhang Qi, Shao Chunju, Wang Ying, Zhang Ping, and Zhang Zhang. Adaptive optimal transmit power allocation for two-hop non-regenerative wireless relaying system. In *2004 IEEE 59th Vehicular Technology Conference. VTC 2004-Spring (IEEE Cat. No.04CH37514)*, volume 2, pages 1213–1217 Vol.2, 2004. 32
- [KSM07] Hisakazu Kikuchi, Kazuma Shinoda, and Shogo Muramatsu. Scalable lossless color image compression by a lossless-by-lossy approach. 01 2007. VII, 7
- [LCC⁺21] Y. Lu, P. Cheng, Z. Chen, W. H. Mow, Y. Li, and B. Vucetic. Deep multi-task learning for cooperative noma: System design and principles. *IEEE Journal on Selected Areas in Communications*, 39(1):61–78, 2021. 15
- [Lea19] Chuan Lin and et al. A deep learning approach for mimo-noma downlink signal detection. *Sensors*, 19(11):2526, 2019. 15
- [Lea21] Q. Lan and et al. What is semantic communication? a view on conveying meaning in the era of machine intelligence. *Journal of Communications and Information Networks*, 6(4):336–371, 2021. 2, 6, 7
- [LSM10] C. Liu, A. Schmeink, and R. Mathar. Constant-rate power allocation under constraint on average ber in adaptive ofdm systems. In *2010 IEEE International Conference on Communications*, pages 1–5, 2010. 32
- [MCBA20] B. Makki, K. Chitti, A. Behravan, and M.-S. Alouini. A survey of noma: current status and open research challenges. *IEEE Open Journal of Communications Society*, 1:179–189, 2020. 14
- [MRJ12] P. Mangayarkarasi, M. Ramya, and S. Jayashri. Analysis of various power allocation algorithms for wireless networks. In *2012 International Conference on Communication and Signal Processing*, pages 133–136, 2012. 32
- [Pea22] Ramesh Chandra Poonia and et al. *Proceedings of Third International Conference on Sustainable Computing*. Springer Nature, 2022. VII, 2, 14, 16, 17
- [RBD24] Ahmad Razlighi, Carsten Bockelmann, and Armin Dekorsy. Semantic communication for cooperative multi-task processing over wireless networks, 04 2024. VII, 2, 10, 11, 12, 13, 20, 23, 81
- [RH24] Ahmad Razlighi and et al. Halimi. Cooperative and collaborative multi-task semantic communication for distributed sources. arXiv preprint arXiv:2411.02150, Nov 2024. Accessed: 2025-06-25. 48
- [SB21] E. C. Strinati and S. Barbarossa. 6g networks: Beyond shannon towards semantic and goal-oriented communications, 2021. [Online] <https://arxiv.org/pdf/2011.14844.pdf>. 6
- [Sha48] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. 6
- [SKAAS10] M. Shamim Kaiser, Kazi M. Ahmed, and Raza Ali Shah. Power allocation in

- ofdm-based cognitive relay networks. In *2010 IEEE International Conference on Wireless Communications, Networking and Information Security*, pages 202–206, 2010. 32
- [SKB⁺13] Yuya Saito, Yoshihisa Kishiyama, Anass Benjebbour, Takehiro Nakamura, Anxin Li, and Kenichi Higuchi. Non-orthogonal multiple access (noma) for cellular future radio access. In *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*, pages 1–5, 2013. 15, 33, 34
- [SMZ22] Jiawei Shao, Yuyi Mao, and Jun Zhang. Learning task-oriented communication for edge inference: An information bottleneck approach. *IEEE Journal on Selected Areas in Communications*, 40(1):197–211, 2022. 10
- [SMZ23] Jiawei Shao, Yuyi Mao, and Jun Zhang. Task-oriented communication for multidevice cooperative edge inference. *IEEE Transactions on Wireless Communications*, 22(1):73–87, 2023. 10
- [TYW⁺21] Haonan Tong, Zhaohui Yang, Sihua Wang, Ye Hu, Walid Saad, and Changchuan Yin. Federated learning based audio semantic communication over wireless networks. In *2021 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, 2021. 10
- [Wan09] Shiguo Wang. Ber-optimized power allocation for amplify-and-forward in single relay system. In *2009 IEEE International Conference on Network Infrastructure and Digital Content*, pages 629–632, 2009. 32
- [WN23] D. Wheeler and B. Natarajan. Engineering semantic communication: A survey. *IEEE Access*, 11:13965–13995, 2023. 7, 8, 9
- [WS49] W. Weaver and C. Shannon. *Recent contributions to the mathematical theory of communication*. University of Illinois Press, 1949. 6
- [XBMMT24] Chunmei Xu, Mahdi Boloursaz Mashhadi, Yi Ma, and Rahim Tafazolli. Semantic-aware power allocation for generative semantic communications with foundation models, 07 2024. 10, 35, 41
- [XMM⁺25] Chunmei Xu, Mahdi Boloursaz Mashhadi, Yi Ma, Rahim Tafazolli, and Jiangzhou Wang. Generative semantic communications with foundation models: Perception-error analysis and semantic-aware power allocation. *IEEE Journal on Selected Areas in Communications*, pages 1–1, 2025. 3
- [XQLJ21] H. Xie, Z. Qin, G. Y. Li, and B. H. Juang. Deep learning enabled semantic communication systems. *IEEE Transactions on Signal Processing*, 69:2663–2675, 2021. 7, 10
- [YPNP⁺23] Ridho Hendra Yoga Perdana, Toan-Van Nguyen, Yushintia Pramitarini, Kyusung Shim, and Beongku An. Deep learning-based spectral efficiency maximization in massive mimo-noma systems with star-ris. In *2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, pages 644–649, 2023. 34
- [ZWKL16] Ningbo Zhang, Jing Wang, Guixia Kang, and Yang Liu. Uplink nonorthogonal multiple access in 5g systems. *IEEE Communications Letters*, 20(3):458–461, 2016. 33