Hybrid NOMA Assisted Heterogeneous Semantic and Bit Users Communication

Ishtiaque Ahmed, and Leila Musavian

***Abstract*—In this paper, we utilize a downlink hybrid Non-Orthogonal Multiple Access (NOMA) framework to support multiple semantic and bit users within the communication network. The hybrid NOMA setup exploits both NOMA and Orthogonal Multiple Access (OMA) which has the benefit of enhancing Spectral Efficiency (SE) by allowing users to dynamically access the resources in multiple heterogeneous slots. This enables integrating semantic and bit users based on their channel gains, while adopting bit-to-semantic decoding order in slots including heterogeneous users. This allows bit user to first decode its signal by treating the semantic signal as interference and subsequently semantic signal is decoded without interference. Semantic users are modeled using a deep learning-based system equipped with pre-trained neural networks. An optimization problem for the power allocation is formulated with the aim of maximizing the equivalent ergodic semantic SE with a constraint on the total available power of the Access Point (AP). The proposed algorithm uses NOMA in shared slots and OMA in bit-user-only slots. Simulation results validate the benefits of heterogeneous users hybrid NOMA setup in comparison to OMA-only for heterogeneous users.**

Keywords*—Downlink hybrid non-orthogonal multiple access, fading channel, power allocation, semantic communication, successive interference cancellation.

I. Introduction

The exponential surge in wireless data traffic presents a significant challenge for next-generation communication networks in terms of ensuring massive connectivity while meeting transmission capacity requirements [1]. On the other hand, building upon the foundational work by Shannon and Weaver, semantic communication emerges as a novel paradigm that revolutionizes communication, focusing on transmitting the semantic content of information, and is particularly suited in low Signal-to-Noise Ratio (SNR) environments [2]. However, at higher SNR regimes, traditional bit-based communication can achieve near-lossless data recovery, making semantic communication less beneficial. This suggests that semantic communication should complement and coexist with the traditional Shannon communication to enhance the overall performance [3]. This coexistence of semantic and bit-based communications within the same network defines heterogeneity within the users set in the network. In fact, with the evolution of semantic communication, new challenges emerge in systems design, particularly in implementing Successive Interference Cancellation (SIC) for supporting heterogeneous users within the network [4].

To support massive connectivity, next-generation communication systems necessitate advanced multiple access techniques that push the boundaries of classical schemes such as Orthogonal Multiple Access [5]. In this respect, Non-Orthogonal Multiple Access (NOMA) is considered a promising technology for enhancing the Spectral Efficiency (SE) of wireless systems by allowing multiple users to share the same time-frequency resources through power or code-domain multiplexing. However, NOMA also presents challenges which need to be carefully addressed, such as increased interference management complexity and fairness concerns [6], [7]. The increase in achievable SE using NOMA depends on the specific system configuration and use case, as NOMA does not always outperform orthogonal schemes, particularly in heterogeneous users communications [8]. In traditional NOMA, SIC is performed based on channel conditions, where the stronger user decodes the signal of the weaker user to mitigate interference. However, this approach cannot be directly applied in joint bit and semantic users communication systems. This is primarily due to the requirement of pre-trained neural networks at the semantic encoder and decoder for successful transmission and reception [9]. Consequently, the bit user is unable to decode the semantic signal from the superimposed signal, as traditional bit-based communication does not require any prior training at the transmitter or receiver. In contrast, the semantic user can decode the bit-based signal, resulting in a bit-to-semantic decoding order for joint bit and semantic users communication [10].

Although conventional NOMA enables multiple access through SIC, its decoding strategy must be adapted to support joint bit and semantic users [11]. Given the differences in their signal processing requirements, a more flexible approach is needed to optimize users pairing and interference management. In this context, hybrid NOMA can exploit the advantages of both OMA and pure NOMA through seamless transitions from one scheme to the other [12]. The crux of hybrid NOMA is to enable multiple users occupy and share the same resource via multiple time slots, for enhancing the overall system capacity. In addition to addressing SIC challenges, optimal power allocation plays a crucial role in the hybrid NOMA framework. By efficiently distributing power among bit and semantic users, the total expected semantic rate can be maximized while ensuring the bit-based users minimum rate requirements, as in [10].

To fully leverage the benefits of semantic communication within hybrid NOMA, it is important to understand how the semantic communication is characterized. The rate at which semantic information is transmitted with a prescribed accuracy is referred to as the semantic rate, which depends on the semantic similarity function with value ranging between zero

The authors are with the School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom (e-mail: {ishtiaque.ahmed, leila.musavian}@essex.ac.uk).

and one [3]. This function quantifies the resemblance between original signal at the transmitter and reconstructed one at the receiver. When semantic similarity function tends to one, the signal will increasingly resemble the original bit-based signal rather than its reconstructed semantic counterpart [9].

Since the semantic similarity function plays a crucial role in determining the effectiveness of semantic communication, a learning-based framework is needed to extract and reconstruct semantic features. To this end, a deep learning based system known as DeepSC is developed [9] which consists of semantic encoder/decoder and channel encoder/decoder. The DeepSC is equipped with neural networks for semantic features extraction. Semantic similarity function depends on neural network structure in DeepSC and the channel condition, and is based on the match level between the transmitted and received sentences using the Biderctional Encoder Representations from Transformers (BERT) model [9]. The pre-trained BERT model [13] captures the context from preceding and succeeding words in a sentence.

The works in [10] and [14] have advanced the integration of semantic and bit-based communications and the deployment of hybrid NOMA. In [10], a semi-NOMA framework is proposed that investigates the trade-offs between semantic and bit-based communications in a shared transmission scheme. However, this approach does not exploit hybrid NOMA with user pairing based on channel conditions, limiting its adaptability in dynamic wireless environments. In [14], a multi-user hybrid NOMA scheme is introduced that optimizes power allocation and minimizes energy consumption under latency constraints. However, this work focuses solely on bit-based communication and does not incorporate semantic communication.

Inspired by the above works, we harness the benefits of heterogeneous users by adopting hybrid NOMA and a tailored decoding strategy that improves interference management. The key contributions of this work are: i) we propose a hybrid NOMA framework for heterogeneous users that dynamically assigns bit and semantic users to time slots based on their channel conditions. To efficiently implement this, we develop a user pairing and decoding strategy, outlined in Algorithm 1 which enhances SE while ensuring adaptive resource allocation; ii) unlike previous works that assume a fixed decoding order, we deploy an adaptive decoding strategy that optimally manages interference in hybrid NOMA. In particular, our approach ensures a bit-to-semantic decoding order in time slots where both bit and semantic users coexist, while applying stronger-to-weaker SIC decoding in time slots containing only semantic users; iii) we formulate and solve a power allocation optimization problem that maximizes the equivalent ergodic semantic SE with a constraint on the total available power; iv) we demonstrate through simulations that the optimization prioritizes bit user in a heterogeneous users network by allocating more power, particularly at higher available power values.

II. System Model

We consider a downlink hybrid NOMA communication system consisting of a bit user (B) and two semantic users (S_1 and S_2), sharing the entire time-frequency resources. While this setup is chosen for analytical clarity, the proposed framework applies to any number of semantic users. In our proposed approach, we allow at most two users to share any slot, as shown in Fig. 1. Communication between the AP and the users occurs over Rayleigh fading channels, where the channel gains remain constant within each time slot but may vary across time slots. At the receiver, additive white Gaussian noise (AWGN) with mean zero and variance σ^2 is assumed to impair the superimposed signal. We deploy the DeepSC model in our hybrid NOMA assisted system that can be generally applied to any semantic architecture. The considered DeepSC neglects potential errors arising from finite neural network capacity, limited training data, and variations in real-time inference accuracy.

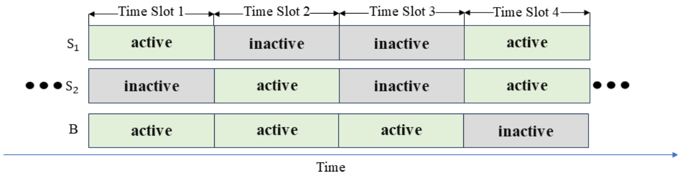

Fig. 1: Users allocation in hybrid NOMA.

Let $P_{\max}$ be the total available power budget at the AP. The channel gain coefficients from the AP to users B, S_1, and S_2 are denoted by h_{B}, $h_{\mathrm{S}1}$, and $h_{\mathrm{S}2}$, respectively, while the transmit powers allocated to them are P_{B}, $P_{\mathrm{S}1}$, and $P_{\mathrm{S}2}$. The semantic rate for any semantic user is given as [3]

$$R_{\mathrm{S}i}(K,L)=\frac{WI}{KL}\epsilon(K,\gamma), i\in\{1,2\} \tag{1}$$

where $\epsilon(K,\gamma)$ is the semantic similarity function, W is the channel bandwidth, I is the amount of semantic information in any transmitted message in units of semantic (suts), K represents the average number of semantic symbols transmitted for each word through DeepSC [9], γ represents the received SNR, and L denotes the number of words per sentence. With varying values of K and γ, $\epsilon(K,\gamma)$ was found to be monotonically non-decreasing with γ [3]. Moreover, its gradient change increases first with γ and then decreases, suggesting a sigmoid shape pattern for $\epsilon(K,\gamma)$. Authors in [8] deployed the data-regression method to tractably approximate the values of $\epsilon(K,\gamma)$ with a generalized logistic function (common form of sigmoid function), as

$$\epsilon(K,\gamma)\triangleq A_{K,1}+\frac{A_{K,2}-A_{K,1}}{1+e^{-(C_{K,1}\gamma+C_{K,2})}}, \tag{2}$$

where the lower (left) asymptote, upper (right) asymptote, growth rate, and the mid-point parameters of the logistic function are respectively denoted by $A_{K,1}$, $A_{K,2}$, $C_{K,1}$, and $C_{K,2}$ for different values of K. These parameters are determined by following the minimum mean square criterion for fitting the logistic function to $\epsilon(K,\gamma)$. The logistic function provides a differentiable and close approximation for tractable analysis and optimization. To ensure that semantic communication achieves desired accuracy at the receiver, $\epsilon(K,\gamma)$ must meet a minimum threshold ϵ_{th}, below which the transmission will be treated as invalid. This is represented as

$$\epsilon_K(\gamma)=\begin{cases}\epsilon(K,\gamma), & \text{if } \epsilon(K,\gamma)\geq\epsilon_{\mathrm{th}},\\ 0, & \text{otherwise.}\end{cases} \tag{3}$$

The approximated generalized logistic function depends explicitly on γ, while the effect of K is captured through the parameters $A_{K,1}$, $A_{K,2}$, $C_{K,1}$, and $C_{K,2}$.

We denote the rates of users S_1, S_2 and B by R_{S1}, R_{S2}, and R_{B}, respectively, where R_{S1} and R_{S2} are in suts/s/Hz, while R_{B} is in bits/s/Hz. The rate of User B is determined on the basis of Shannon's capacity formula. Perfect channel state information is assumed to be available at the AP for simplicity and tractability. However, achieving perfect CSI is difficult due to channel estimation error, feedback delay, or limited signaling overhead. To ensure efficient decoding of signals, the AP informs the receiver about the type of active users in each time slot via embedded control information in the headers of transmitted data packets.

Our goal is to maximize the expected equivalent semantic SE of the three users while ensuring that the system obeys the AP power constraint. We transform the rate of bit user R_{B} into its semantic counterpart R_{SB} in suts/s/Hz, as in [3], yielding

$$R_{\mathrm{SB}} = R_{\mathrm{B}} \frac{I}{\mu L} \epsilon_C, \tag{4}$$

where μ represents the average number of bits per word, the semantic similarity function for User B is denoted by ϵ_C and taken as one to reflect error-free transmission. This transformation allows a meaningful comparison between both types of users by expressing their efficiencies in the same unit.

For the hybrid NOMA structure, in each time slot, up to two users are scheduled to be active and share the resources. Users' activity is governed by the indicator variables I_{B}, I_{S1}, and I_{S2} for users B, S_1, and S_2, respectively. The selection of users for pairing is done dynamically based on real-time channel conditions, ensuring that the strongest user pairs with the weakest one across different time slots without a fixed ordering. When a semantic user has stronger channel gain than the bit user, the slot is allocated exclusively to the bit user, effectively operating as OMA. In contrast, when the channel conditions allow simultaneous transmission, NOMA is employed, where both users share the same time-frequency resources while following a bit-to-semantic decoding order. This adaptive hybrid NOMA strategy dynamically transitions between OMA and NOMA while maintaining the bit-to-semantic decoding strategy.

As semantic and bit users can simultaneously be active in each slot, a new decoding mechanism will be required. Let us denote the corresponding transmit signals for users B, S_1, and S_2 by x_{B}, x_{S1}, and x_{S2}, respectively. In slots where any two users share the channel via NOMA, the received signal will be represented as a superposition of their transmitted signals, scaled by their respective channel gains and corrupted by AWGN. In such slots, decoded signal of one user will always be affected by interference from the other. For the OMA slots, there will be no decoding interference experienced. Moreover, traditional channel-condition based SIC will no longer be beneficial in joint bit and semantic users communication networks as the bit user cannot decode a semantic signal due to the absence of prior training at transmitter and receiver. Consequently, we adopt the bit-to-semantic decoding order in slots involving both types of users, where semantic signals are decoded interference-free after SIC. However, the bit user being decoded first will treat the semantic user's signal as interference.

The pairing approach and decoding strategy for users are outlined in Algorithm 1.

Algorithm 1 User Pairing and Decoding Strategy for the Hybrid NOMA Framework

```
Initialization: Set $P_{\max}$ value, and $I_{\mathrm{B}} = 0$, $I_{\mathrm{S1}} = 0$, $I_{\mathrm{S2}} = 0$.
Users Pairing in Each Slot: Determine $h_{\mathrm{B}}$, $h_{\mathrm{S1}}$, $h_{\mathrm{S2}}$ and pair the strongest user with the weakest one. Set the indicators for the paired users to 1.
if slot contains B and either $\mathrm{S}_1$ or $\mathrm{S}_2$ then
    if involved semantic user has a stronger channel power gain than User B then
        Reset its indicator to 0 which allocates the slot to User B only (OMA).
    else
        follow Bit-to-semantic Decoding:
        B-receiver decodes its own signal treating semantic user's signal as noise.
        S-receiver firstly decodes B user's signal and applies SIC to subtract it, resulting in its own signal without interference.
    end if
else if slot contains $\mathrm{S}_1$ and $\mathrm{S}_2$ then
    Stronger Semantic User decodes its own signal treating the weaker user's signal as noise.
    Weaker Semantic User firstly decodes the stronger user's signal and applies SIC to subtract it, resulting in its own signal without interference.
end if
```

III. Optimal Power Allocation and Users Detection

After pairing users and establishing the decoding order in each time slot, we now formulate the power allocation optimization problem.

A. Power Allocation Optimization

Objective of our optimization problem is to maximize the expected value of the equivalent semantic SE, which can be formulated as

$$\max_{\{P_{\mathrm{B}}, P_{\mathrm{S1}}, P_{\mathrm{S2}}\}} \mathbb{E}\left[I_{\mathrm{S1}} R_{\mathrm{S1}} + I_{\mathrm{S2}} R_{\mathrm{S2}} + I_{\mathrm{B}} R_{\mathrm{SB}}\right], \tag{5}$$

$$\begin{aligned} \text{s.t.} \quad & P_{\mathrm{B}} + P_{\mathrm{S1}} + P_{\mathrm{S2}} \leq P_{\max}, \\ & P_{\mathrm{B}},\ P_{\mathrm{S1}},\ P_{\mathrm{S2}} \geq 0 \end{aligned} \tag{6}$$

where $\mathbb{E}[.]$ represents the expectation operator. Each of the considered rates in (5) can be generalized with a closed-form approximation of (2), exhibiting a logistic distribution model due to semantic similarity function [8], where $A_{K,1} = 0.37$, $A_{K,2} = 0.98$, $C_{K,1} = 0.25$, and $C_{K,2} = -0.7895$. The resulting rate curve is resilient to moderate parameter variations due to the robustness of the neural network-based DeepSC.

Unlike conventional rate functions, the approximated semantic similarity function does not exhibit concavity with respect to γ which renders the optimization problem as non-convex. However, the problem satisfies the "time-sharing" condition following the proof in [15] and [16]. If any problem satisfies the "time-sharing" condition [16], then strong duality holds between the primal and Lagrangian dual problems, regardless of convexity of the original problem. Moreover, when strong duality exists, the Karush-Kuhn-Tucker (KKT) conditions are both sufficient and necessary for optimality [17].

B. Solution Method

We solve the power allocation problem in (5) by developing its Lagrangian and then using the KKT conditions in each time slot. The complementary slackness and primal feasibility conditions ensure that the solution respects the power constraint and precludes any negative power allocation. The KKT conditions are developed based on the Lagrangian formulation given by

$$\begin{aligned} &\mathcal{L}(P_{\mathrm{B}}, P_{\mathrm{S1}}, P_{\mathrm{S2}}, \lambda) \\ &= \mathbb{E}\left[I_{\mathrm{S1}} R_{\mathrm{S1}} + I_{\mathrm{S2}} R_{\mathrm{S2}} + I_{\mathrm{B}} R_{\mathrm{SB}}\right] + \\ &\lambda \left(P_{\max} - P_{\mathrm{B}} - P_{\mathrm{S1}} - P_{\mathrm{S2}}\right), \end{aligned} \tag{7}$$

where λ is the corresponding Lagrange multiplier for the power constraint.

C. Decoding Strategy in Slots with Semantic and Bit Users

Bit-to-semantic decoding order is followed where the rate of User B will always be adversely impacted by interference from the semantic user. The roots of (8) provide the solution for optimal power of the bit user.

$$\frac{\partial \left(\mathbb{E}\left[I_{\mathrm{B}} R_{\mathrm{SB}} + I_{\mathrm{S}i} R_{\mathrm{S}i}^{\mathrm{SB}}\right]\right)}{\partial P_{\mathrm{B}}} = 0, \quad i \in \{1, 2\} \tag{8}$$

Specifically, the corresponding rate for bit user in such time slots is defined by

$$R_{\mathrm{B}i} = \log_2 \left(1 + \frac{P_{\mathrm{B}} \left|h_{\mathrm{B}}\right|^2}{\sigma^2 + P_{\mathrm{S}i} \left|h_{\mathrm{B}}\right|^2}\right), i \in \{1, 2\} \tag{9}$$

where $P_{\mathrm{S}i}$ denotes the power of semantic user paired with User B. Although we consider the equivalent semantic rate of bit user in our objective function, R_{SB} is merely a scalar multiple of conventional rate $R_{\mathrm{B}i}$. User B first decodes its own signal, while the semantic user extracts its desired signal without interference through SIC. The corresponding rate for semantic user involved in such slots is given by (10).

$$R_{\mathrm{S}i}^{\mathrm{SB}} = \frac{I}{KL} \epsilon_K \left(\frac{P_{\mathrm{S}i} \left|h_{\mathrm{S}i}\right|^2}{\sigma^2}\right), i \in \{1, 2\} \tag{10}$$

Say, we solve for P_{B} when S_1 is active by putting $I_{\mathrm{B}} = 1$, $I_{\mathrm{S1}} = 1$, and $I_{\mathrm{S2}} = 0$. In the considered time slot with B and S_1 active, the closed-form solution for optimum P_{B} can be found from

$$\begin{aligned} &\frac{I \left|h_{\mathrm{B}}\right|^2}{\mu L \ln(2)(\sigma^2 + \left|h_{\mathrm{B}}\right|^2 P_{\max} - \left|h_{\mathrm{B}}\right|^2 P_{\mathrm{B}})} - \\ &\frac{I \left|h_{\mathrm{S1}}\right|^2 (A_{K,2} - A_{K,1}) C_{K,1}\, e^{\left(-C_{K,1} \frac{(P_{\max} - P_{\mathrm{B}}) \left|h_{\mathrm{S1}}\right|^2}{\sigma^2} + C_{K,2}\right)}}{KL\sigma^2 (1 + e^{\left(-C_{K,1} \frac{(P_{\max} - P_{\mathrm{B}}) \left|h_{\mathrm{S1}}\right|^2}{\sigma^2} + C_{K,2}\right)})^2} = 0 \end{aligned} \tag{11}$$

Similarly, taking the partial derivative with respect to P_{S1} in such slots and solving for P_{B} leads to a function with transcendental value [18] that is solved numerically for the optimal power P_{B}^* of bit user.

$$\begin{aligned} f(P_{\mathrm{B}}) = &\frac{\left|h_{\mathrm{B}}\right|^2}{\mu \ln(2)(\sigma^2 + \left|h_{\mathrm{B}}\right|^2 P_{\max} - \left|h_{\mathrm{B}}\right|^2 P_{\mathrm{B}})} - \\ &\frac{\left|h_{\mathrm{S1}}\right|^2 (A_{K,2} - A_{K,1}) C_{K,1}\, e^{\left(-C_{K,1} \frac{(P_{\max} - P_{\mathrm{B}}) \left|h_{\mathrm{S1}}\right|^2}{\sigma^2} + C_{K,2}\right)}}{K\sigma^2 \left(1 + e^{\left(-C_{K,1} \frac{(P_{\max} - P_{\mathrm{B}}) \left|h_{\mathrm{S1}}\right|^2}{\sigma^2} + C_{K,2}\right)}\right)^2} \end{aligned} \tag{12}$$

In such time slots, if the involved semantic user does not satisfy the set minimum threshold ϵ_{th}, the entire available power is allocated to the bit user. The optimal power allocation for the bit is thus given as

$$P_{\mathrm{B}}^* = \begin{cases} f(P_{\mathrm{B}}), & \text{if } \epsilon_K(\gamma) \geq \epsilon_{\mathrm{th}}, \\ P_{\max}, & \text{otherwise.} \end{cases} \tag{13}$$

D. Decoding Strategy in Slots with Semantic Users Only

In these slots, the Lagrangian is accordingly adjusted as

$$\begin{aligned} &\mathcal{L}'(P_{\mathrm{S1}}, P_{\mathrm{S2}}, \lambda) \\ &= \mathbb{E}\left[I_{\mathrm{S1}} R_{\mathrm{S1}} + I_{\mathrm{S2}} R_{\mathrm{S2}}\right] + \lambda \left(P_{\max} - P_{\mathrm{S1}} - P_{\mathrm{S2}}\right), \end{aligned} \tag{14}$$

focusing solely on the rates of semantic users under the power constraint. Performing SIC becomes challenging in such slots due to the joint training of semantic encoders and decoders. To address this issue, we exploit an innovative channel reconstruction-based SIC method proposed in [4]. This method allows the semantic user with better channel condition to be decoded first while treating the signal from the weaker user as noise. The stronger user's signal is reconstructed by reversing the channel effects and mitigating interference until a target semantic similarity ϵ_{th} is attained. Once the stronger signal is successfully reconstructed, the weaker semantic user can subtract the decoded stronger user's signal from the superimposed signal at the receiver. This approach remains effective even when semantic users exhibit similar channel conditions and rate requirements, allowing either user to be decoded first.

Putting $I_{\mathrm{S1}} = 1$, $I_{\mathrm{S2}} = 1$, and assuming S_2 experiences higher channel gain than S_1, then the roots of (15) provide the optimal solution for P_{S2}.

$$\frac{\partial \left(\mathbb{E}\left[R_{\mathrm{S2}} + R_{\mathrm{S1}}\right]\right)}{\partial P_{\mathrm{S2}}} = 0 \tag{15}$$

More generally, the achievable semantic rate of the first decoded user in such time slots is given by

$$R_{\mathrm{S2}} = \frac{I}{KL} \epsilon_K \left(\frac{P_{\mathrm{S2}} \left|h_{\mathrm{S2}}\right|^2}{P_{\mathrm{S1}} \left|h_{\mathrm{S2}}\right|^2 + \sigma^2}\right), \tag{16}$$

Expressing $P_{\mathrm{S1}} = P_{\max} - P_{\mathrm{S2}}$, the closed-form solution for the optimal power of stronger semantic user can be found from

$$\frac{I}{KL}\left[\frac{\gamma_2^2\left(\sigma^2+P_{\max}\left|h_{\mathrm{S1}}\right|^2\right)\left(A_{K,2}-A_{K,1}\right)C_{K,1}e^{-(C_{K,1}\gamma_2+C_{K,2}}}{P_{\mathrm{S2}}^2\left|h_{\mathrm{S2}}\right|^2\left(1+e^{-(C_{K,1}\gamma_2+C_{K,2})}\right)^2}\right.$$
$$\left.-\frac{\gamma_1\left(A_{K,2}-A_{K,1}\right)C_{K,1}e^{-(C_{K,1}\gamma_1+C_{K,2})}}{\left(P_{\max}-P_{\mathrm{S2}}\right)\left(1+e^{-(C_{K,1}\gamma_1+C_{K,2})}\right)^2}\right]=0 \tag{17}$$

where $\gamma_1 = \frac{(P_{\max}-P_{\mathrm{S2}})|h_{\mathrm{S1}}|^2}{\sigma^2}$, and $\gamma_2 = \frac{P_{\mathrm{S2}}|h_{\mathrm{S2}}|^2}{(P_{\max}-P_{\mathrm{S2}})|h_{\mathrm{S2}}|^2+\sigma^2}$

denote the received SNR at S_1 and S_2, respectively. Solving for P_{S2} after taking the partial derivative with respect to P_{S1} results in a transcendental value function for the optimal power of stronger semantic user.

IV. Simulation Results

In this section, we present the simulation figures and analyze the results for the proposed heterogeneous users hybrid NOMA communication setup. Unless specified otherwise, we use the parametric values of $\epsilon_{\mathrm{th}} = 0.9$, $K = 5$, $\mu = 40$, and $I/L = 1$. Users are randomly distributed for each Monte Carlo simulation within a 100 meters cell radius. Users channel coefficients for the independently distributed Rayleigh fading are determined based on the distance-dependent path-loss model $\rho = \rho_0(1/d)^{\beta}$. We assume a reference path-loss of $\rho_0 = -30$ dB at 1 meter, and the path-loss exponent $\beta = 4$. All the simulation results are averaged over 10^5 Monte Carlo channel realizations.

In the first simulation figure, we compare the performance of the heterogeneous users hybrid NOMA communication with a heterogeneous users OMA setup, where the OMA time fractions are proportionally allocated to each user according to their participation ratio under hybrid NOMA. Fig. 2 shows the equivalent ergodic semantic SE versus $P_{\max}$ for both setups. The results demonstrate that the proposed setup consistently outperforms its OMA counterpart. This indicates the effectiveness of hybrid NOMA for integrating both semantic and traditional bit-based communications, leveraging NOMA and OMA resource allocation strategies.

In Fig. 3, the equivalent ergodic semantic SE is plotted versus ϵ_{th} for $P_{\max}$ values of 5 dB, 10 dB, and 20 dB. It is evident that the proposed setup achieves higher ergodic semantic SE across all values of ϵ_{th} for different $P_{\max}$. For both the setups as the value of ϵ_{th} approaches one, a drop in SE is recorded due to the inherent limitations in accurate semantic reconstruction at higher semantic similarity thresholds. However, the hybrid NOMA still maintains a higher SE due to the higher resource utilization by the bit user aiding in accurate signal reconstruction.

Fig. 4 shows the bit user ergodic SE versus $P_{\max}$ for hybrid NOMA and OMA setups. As $P_{\max}$ increases, the hybrid NOMA curve consistently lies above the OMA curve, demonstrating the benefits of its flexible resource allocation strategy. If the semantic user has stronger channel gain than the bit user in shared time slots, the entire power is redirected to the bit user, otherwise the total power splits between both the users.

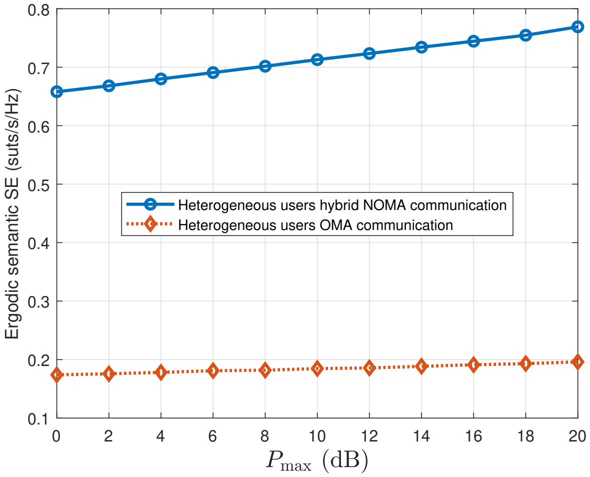

Fig. 2: Ergodic semantic SE versus $P_{\max}$ with $\epsilon_{\mathrm{th}} = 0.9$.

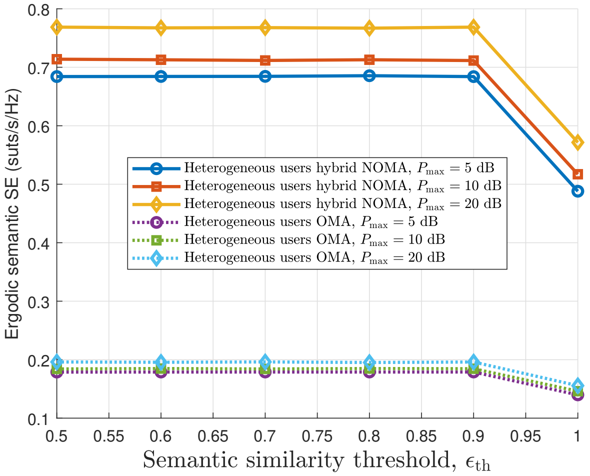

Fig. 3: Ergodic semantic SE versus ϵ_{th} for different $P_{\max}$.

Fig. 5 shows the allocated ergodic power for each user in the hybrid NOMA setup versus $P_{\max}$. The figure shows that with increasing $P_{\max}$ values, the power allocated to the bit user is significantly higher than that for the semantic users. The allocated power for each user increases with $P_{\max}$ but the growth rate for P_{S1} and P_{S2} is substantially lower than that for P_{B}. This indicates that the hybrid NOMA system prioritizes bit user which renders optimality to be unfair. This is due to the difference in capacity equations for bit and semantic users, respectively depending on log-rate and a saturating similarity function.

In Fig. 6, the equivalent ergodic semantic SE of bit user in the heterogeneous users hybrid NOMA setup is shown versus $P_{\max}$ for ϵ_{th} values of 0.85, 0.9, 0.95. The three curves corresponding to different values of ϵ_{th} show a steadily growing trend with $P_{\max}$. The figure shows that the equivalent semantic SE of bit user consistently increases with $P_{\max}$. This behavior can be attributed to the efficient resource management of the hybrid NOMA system which prioritizes bit user in a het-

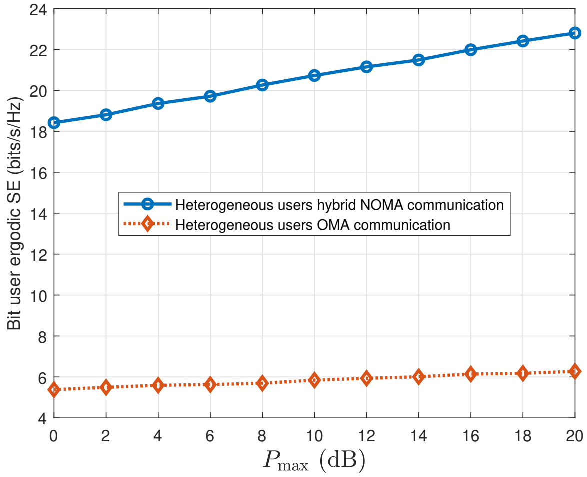

Fig. 4: Bit user ergodic SE versus $P_{\max}$ with $\epsilon_{\text{th}} = 0.9$.

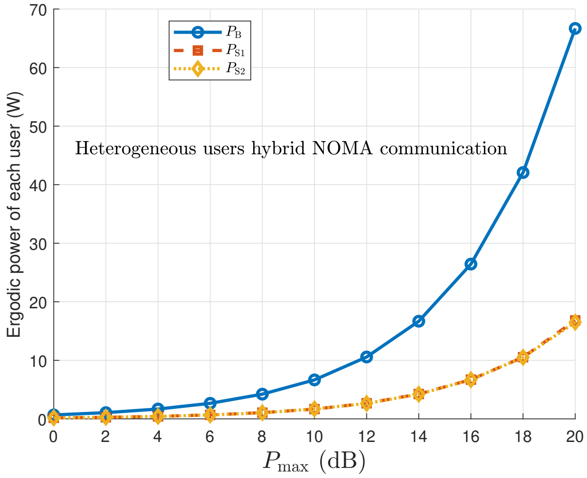

Fig. 5: Ergodic power of each user versus $P_{\max}$ with $\epsilon_{\text{th}} = 0.9$.

erogeneous users network for accurate signal reconstruction.

V. Conclusions

A novel hybrid NOMA framework is presented to enable the coexistence of semantic and bit users. Higher performance of the proposed heterogeneous users hybrid NOMA setup is demonstrated through analytical evaluation, with simulation results conforming to the related literature on hybrid NOMA and semantic communications. The proposed algorithm dynamically pairs users based on real-time channel conditions and implements a decoding strategy to ensure efficient interference management. The joint power allocation problem was solved under the power constraint to maximize the equivalent ergodic semantic SE. Bit-to-semantic decoding strategy is adopted in slots with both types of users, thereby enabling interference-free decoding of the semantic signal. The simulation results demonstrated the efficacy of the proposed communication setup in enhancing ergodic semantic SE. The proposed system integrates both NOMA and OMA strategies by dynamically

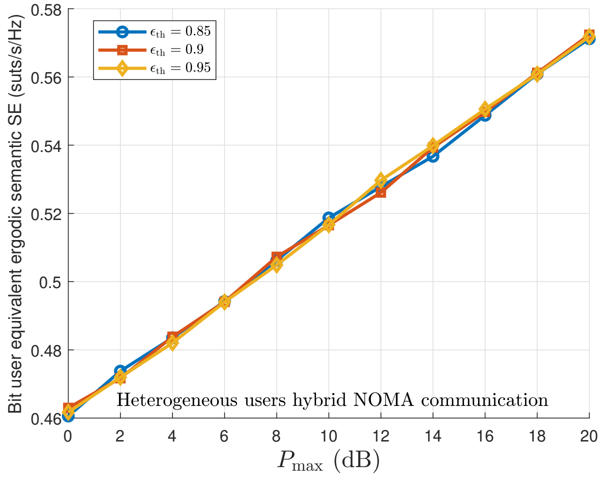

Fig. 6: Bit user equivalent ergodic semantic SE versus $P_{\max}$ for different ϵ_{th}.

assigning users to time slots, ensuring efficient resource utilization and making it a viable approach for heterogeneous users communication in future wireless networks.

Acknowledgment

This work was supported by the UK Research and Innovation under the UK government's Horizon Europe funding guarantee through MSCA-DN SCION Project Grant Agreement No.101072375 [grant number: EP/X027201/1].

References

[1] L. Wang, W. Wu, F. Zhou, Z. Yang, Z. Qin, and Q. Wu, "Adaptive resource allocation for semantic communication networks," *IEEE Trans. Commun.*, 2024.

[2] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen, and C. Miao, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 213–250, 2022.

[3] L. Yan, Z. Qin, R. Zhang, Y. Li, and G. Y. Li, "Resource allocation for text semantic communications," *IEEE Wireless Commun. Lett.*, vol. 11, no. 7, pp. 1394–1398, 2022.

[4] B. Chen, X. Wang, D. Li, R. Jiang, and Y. Xu, "Uplink NOMA semantic communications: Semantic reconstruction for SIC," in *IEEE/CIC Int. Conf. Commun. China (ICCC)*, Dalian, China, August 10–12, 2023, pp. 1–6.

[5] J. Wang, C. Jiang, H. Zhang, Y. Ren, K.-C. Chen, and L. Hanzo, "Thirty years of machine learning: The road to Pareto-optimal wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1472–1514, 2020.

[6] X. Pei, Y. Chen, M. Wen, H. Yu, E. Panayirci, and H. V. Poor, "Next-generation multiple access based on NOMA with power level modulation," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1072–1083, 2022.

[7] X. You, C.-X. Wang, J. Huang, X. Gao, Z. Zhang, M. Wang, Y. Huang, C. Zhang, Y. Jiang, J. Wang *et al.*, "Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts," *Sci. China Inf. Sci.*, vol. 64, pp. 1–74, 2021.

[8] X. Mu, Y. Liu, L. Guo, and N. Al-Dhahir, "Heterogeneous semantic and bit communications: A semi-NOMA scheme," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 155–169, 2022.

[9] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, “Deep learning enabled semantic communication systems,” *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021.

[10] X. Mu and Y. Liu, “Semantic communications in multi-user wireless networks,” *arXiv preprint arXiv:2211.08932*, 2022.

[11] I. Ahmed, Y. Sun, J. Fu, A. Köse, L. Musavian, M. Xiao, and B. Özbek, “Semantic communications in 6G: Coexistence, multiple access, and satellite networks,” *IEEE Commun. Stand. Mag.*, 2025.

[12] Z. Ding, R. Schober, P. Fan, and H. V. Poor, “Next generation multiple access for IMT towards 2030 and beyond,” *Sci. China Inf. Sci.*, vol. 67, no. 6, pp. 1–3, 2024.

[13] J. Devlin, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[14] Z. Ding, D. Xu, R. Schober, and H. V. Poor, “Hybrid NOMA offloading in multi-user MEC networks,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 5377–5391, 2022.

[15] X. Mu and Y. Liu, “Exploiting semantic communication for non-orthogonal multiple access,” *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2563–2576, 2023.

[16] W. Yu and R. Lui, “Dual methods for nonconvex spectrum optimization of multicarrier systems,” *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, 2006.

[17] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.

[18] S. Bernard, “Formalization of the Lindemann-Weierstrass theorem,” in *Interactive Theorem Proving: 8th International Conference, ITP 2017, Brasília, Brazil, September 26–29, 2017, Proceedings 8*. Springer, 2017, pp. 65–80.