Semantic Wireless Networks with Minimal Energy Consumption

Marija Poposka, *Graduate Student Member, IEEE*, Himal A. Suraweera, *Senior Member, IEEE*, George K. Karagiannidis, *Fellow, IEEE*, and Zoran Hadzi-Velkov, *Senior Member, IEEE*

***Abstract*—In this letter, we develop a novel resource allocation scheme for energy-efficient semantic-aware wireless networks. The wireless users (WUs) extract semantic information from their raw source data during the common feature extraction (FE) phase, and then transmit their semantic features to their paired semantic decoders at the BS over the uplink channels. In random fading channels, the semantic similarity between reconstructed and source data may occasionally fall below some desired threshold resulting in a communications outage, which leads to the notion of *average semantic rate*. The proposed schemes minimize total energy consumption of WUs by deriving closed-form global optimal solutions for the transmit powers, processor frequencies and transmission duration of each WU, and duration of common FE phase. Compared to conventional bit-based communications, energy gain of proposed schemes increases rapidly with increase of semantic rate and average SNR.**

***Index Terms*—Semantic communications, resource allocation, energy consumption, wireless networks.**

I. Introduction

As future sixth-generation wireless systems aim to provide significantly improved spectral and energy efficiency over current systems, semantic communications have recently emerged as a promising technology to achieve these goals by leveraging artificial intelligence (AI) [1]. In semantic communications, instead of transmitting every single bit of raw source data, the transmitter extracts and sends their essential features to enable the receiver to make the right inference, decision or action to perform a specific task, resulting in a significant reduction in data traffic. In the literature, different types of semantic systems have been studied for different types of source data, including text [2], speech [3], image [4], and video [5].

Motivated by the success of AI for natural language processing, the authors of [2] developed a joint source-channel coding method for text transmission, known as DeepSC, where a transformer-based encoder and decoder were developed for processing semantic information in text sentences. Based on the DeepSC framework, [6] introduces relevant performance metrics, *semantic rat*e and *semantic similarity*, which can be used for resource allocation of semantic-aware networks. Using these metrics, [6] maximizes the semantic rate for a multi-user system with optimal frequency allocation, while [7] characterizes the semantic versus bit rate region and power region of a two-user heterogeneous semantic and bit system. Although originally applied to textual semantic communications, these performance metrics are also applicable to semantic systems with source data types other than text; Specifically, [3] extended the DeepSC framework for speech transmission, while [8] shows that image and video transmission can be transformed into text transmission. Apart from schemes based on joint semantic-channel coding where the semantic encoders/decoders are trained jointly with the channel encoders/decoders with full-resolution constellations, there are also schemes that use variable semantic features-to-bit quantization but require continuous adjustment of the resource allocation parameters during the semantic communication session, thus increasing the computational load and signalling requirements of the system [9].

Semantic-aware wireless networks have also been studied in the context of energy efficiency in [10], however the proposed solution is suboptimal, the semantic information extraction subproblem is optimized independently from the resource allocation parameters, and the random channel fading is neglected. This paper arrives at the globally optimal solution for the resource allocation of energy-efficient semantic wireless networks in random fading channels. The main contributions of this paper specifically include: (1) We introduce a new performance metric for semantic systems in fading channels, termed *average semantic rate,* $\bar{R}_{th}$, which supports communication outage events between a semantic encoder and a semantic decoder in case of random channel fading, (2) We develop a resource allocation scheme that minimizes (computation and transmission) energy consumption for semantic-aware multi-user networks while guaranteeing a minimum (average) semantic rate for all users, and (3) Globally optimal closed-form solutions are derived for determining the transmit powers, computation and transmit durations, and central processing unit (CPU) clock frequencies of the wireless users (WUs).

II. System Model

A wireless network, consisting of a common base station (BS) and K WUs, is used for semantic communication. Each WU is equipped with a semantic encoder that is paired by a corresponding semantic decoder at the BS, and so the WU transmits to the BS independent semantic information to perform its own task. The semantic encoder of each WU extracts the semantic features from the text-based input by using, for example, transformer-based techniques of the DeepSC model [2], whereas the corresponding semantic decoder at the BS reconstructs the text from received semantic features. Actually, BS runs K semantic decoders, one for each WU encoder. The K semantic encoder-decoder pairs are trained for different channel conditions before the beginning of the actual semantic communication session, c.f. [2, Fig. 6]. We assume kth WU

Manuscript received 30 April 2024; accepted 24 May 2024. The associate editor coordinating the review of this letter and approving it for publication was Yingyang Chen. *(Corresponding author: Zoran Hadzi-Velkov.)*

M. Poposka and Z. Hadzi-Velkov are with the Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University, Skopje, North Macedonia (e-mail: {poposkam, zoranhv}@feit.ukim.edu.mk).

H. A. Suraweera is with Department of Electrical and Electronic Engineering, University of Peradeniya, Peradeniya, 20400, Sri Lanka (e-mail: himal@eng.pdn.ac.lk).

G. K. Karagiannidis is with with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece and also with the Artificial Intelligence & Cyber Systems Research Center, Lebanese American University (LAU), Lebanon (e-mail: geokarag@auth.gr).

($1 \leq k \leq K$) receives a sentence of L words at its input, and its encoder extracts needed semantic features from a sentence with an average length of s_k semantic symbols per word.

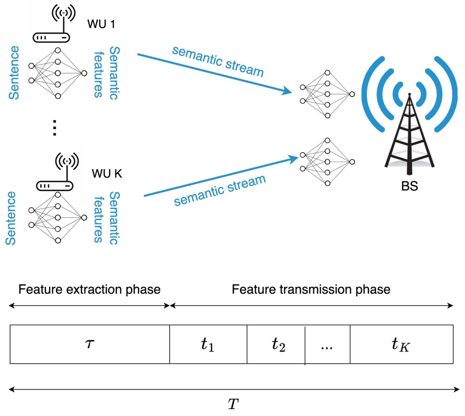

Fig. 1: System model of the semantic-aware network.

The system time is divided into epochs of duration T, which is subdivided into a *feature extraction (FE) phase* and a *feature transmission (FT) phase* (c.f. Fig. (1)). During the FE phase, each WU extracts its corresponding semantic feature vector. Let us denote by a_k the number of CPU cycles required by the kth WU to generate semantic feature vectors with an average length Ls_k per sentence during the FE phase, and the CPU frequency of the kth WU by f_k (in CPU cycles per second). Assuming $a_k = a_0, \forall k$, the duration of the FE phase of the kth WU is equal to a_0/f_k. For implementation simplicity, we assume the FE phase is completed at the same time instant by all WUs, and so its duration τ is determined as the maximum of durations of the individual FE phases of all WUs,

$$\tau = \max_{1 \leq k \leq K} \left\{ \frac{a_0}{f_k} \right\}. \tag{1}$$

The energy required by the kth WU to generate its semantic feature vector is $E_k^{FE} = \alpha a_0 f_k^2$, where α is the energy efficiency coefficient that depends on the WU's CPU architecture.

During the FT phase, the WUs employ time division multiple access (TDMA) to transmit their feature vectors to the BS in successive non-overlapping time periods t_1, t_2,..., t_K. Let us denote the gain of the uplink channel between the kth WU by g_k, the transmit power of the kth WU by p_k, the uplink channel bandwidth by B, and the power spectral density of the thermal noise at the BS receiver by N_0. Thus, the energy required by the kth WU to transmit its feature vector over the wireless channel is determined by $E_k^{FT} = p_k t_k$, whereas the signal-to-noise ratio (SNR) of the received signal from the kth WU at the BS is given by $\gamma_k = p_k g_k/(BN_0)$.

We adopt *semantic similarity*, ξ_k, as a measure for the "semantic accuracy" achieved between the kth WU encoder and the BS decoder over the available communications channel [6]. It is a function of s_k and γ_k, i.e., $\xi_k = \xi_k(s_k, \gamma_k)$, and can be approximated by [7, Eq. (6)]

$$\xi_k = \xi_k(s_k, \gamma_k) = A_{s_k,1} + \frac{A_{s_k,2} - A_{s_k,1}}{1 + e^{-\left(C_{s_k,1}\gamma_k + C_{s_k,2}\right)}}, \tag{2}$$

where $A_{s_k,1}$, $A_{s_k,2}$, $C_{s_k,1}$, and $C_{s_k,2}$ are the parameters defining the generalized logistic regression function. Although generally depend on s_k, we assume that these parameters are constant for all WUs, i.e., $\xi_k = \xi(\gamma_k)$, where $A_{s_k,1} = A_1$, $A_{s_k,2} = A_2$, $C_{s_k,1} = C_1$ and $C_{s_k,2} = C_2$.

Based on ξ_k, we define *semantic rate* of kth WU by [6]

$$R_k = \frac{IB}{s_k L}\xi_k = \frac{IB}{s_k L}\xi\left(\frac{p_k g_k}{BN_0}\right), \tag{3}$$

where I denotes the average amount of semantic information contained in a sentence, and is expressed in *semantic units* (*suts*). In (3), the ratio $I/(Ls_k)$ denotes the average amount of semantic information per symbol of the kth WU, and is expressed in *suts/symbol*, whereas R_k is expressed in *suts/second* since the number of transmitted symbols per second is numerically equal to the bandwidth B.

III. Resource Allocation over Static Channels

We propose a resource allocation scheme for semantic communication system that minimizes the total energy consumption of all WUs. Firstly, we assume that the uplink wireless channels for all WUs are static, which means that g_k is fixed during the entire communication session, i.e., $g_k = \Omega_k$. We thus set the following optimization problem:

$$\begin{aligned} &\underset{\tau, t_k, f_k, p_k}{\text{Minimize}} \sum_{k=1}^{K} \left(p_k t_k + \alpha a_0 f_k^2\right) \\ &\text{subject to: } C1: t_k \tfrac{IB}{s_k L}\xi\left(\tfrac{p_k \Omega_k}{BN_0}\right) \geq R_{th}T, \forall k \\ &\qquad C2: \xi\left(\tfrac{p_k \Omega_k}{BN_0}\right) \geq \xi_{th}, \forall k \\ &\qquad C3: \tau + \textstyle\sum_{k=1}^{K} t_k = T, \\ &\qquad C4: p_k \leq P_{max}, \forall k \\ &\qquad C5: f_k \leq f_{max}, \forall k \\ &\qquad C6: \tfrac{a_0}{f_k} \leq \tau, \forall k. \end{aligned} \tag{4}$$

In (4), the constraint $C1$ sets the minimal amount of received semantic information from the kth WU, $R_{th}T$, where R_{th} is the semantic rate threshold common for all WUs, and T is the duration of a single communication round (c.f. Fig. 1). The constraint $C2$ sets a minimal common threshold ξ_{th} for the semantic similarity between the kth WU input sentence and the corresponding decoded sentence at the BS. The constraint $C3$ sets the duration of a single communications round, $C4$ limits the maximum transmit power, and $C5$ limits the maximum CPU frequency. The constraint $C6$ imposes an upper bound on the duration of the FE phase of all WUs due to (1). Setting $r_k = R_{th}s_k L/(BI)$ and $\sigma_N^2 = BN_0$, the solution of (4) is given by the following theorem.

Theorem 1. *The feasibility conditions for the existence of an optimal solution of (4) is given by*

$$\frac{a_0}{Tf_{max}} + \frac{1}{\xi_{th}}\sum_{k=1}^{K} r_k \leq 1, \tag{5}$$

and

$$\xi\left(\frac{P_{max}\Omega_k}{\sigma_N^2}\right) \geq \xi_{th} \tag{6}$$

The optimal local computing time, τ^*, is determined by

$$\tau^* = T\left(1 - \frac{1}{\xi_{th}}\sum_{k=1}^{K} r_k\right), \tag{7}$$

the optimal transmit powers, p_k^*, are determined by

$$p_k^* = \frac{\sigma_N^2}{C_1\Omega_k}\left(\log\left(\frac{\xi_{th}-A_1}{A_2-\xi_{th}}\right)-C_2\right), \quad \forall k, \tag{8}$$

the optimal transmit durations, t_k, are determined by

$$t_k^* = \frac{r_k}{\xi_{th}}T, \quad \forall k, \tag{9}$$

and the optimal CPU frequencies, f_k^*, are determined by

$$f_k^* = \frac{a_0}{\tau^*}, \quad \forall k. \tag{10}$$

If (5) satisfies a strict inequality, then $f_k^* < f_{max}$, $\forall k$; Otherwise, if (5) satisfies a strict equality, then $f_k^* = f_{max}$, $\forall k$.

Proof: Please refer to Appendix A. ■

IV. Resource allocation over fading channels

If wireless channel is exposed to fading, the semantic communication system will occasionally experience outages when any channel g_k is in a "deep fading" state. Actually, when g_k is very low, $C2$ in (4) cannot be met even when the transmitter transmits at its peak power P_{max}. Thus, when exposed to fading, this communication system is better described by an averaged performance metric, i.e., the *average semantic rate*, $\bar{R}_{th}$. In this case, constraint $C1$ in (4) should be restated as

$$\tfrac{t_k BI}{s_k L}\, E\left[\xi\left(\tfrac{p_k g_k}{BN_0}\right)\, I\left(\xi\left(\tfrac{p_k g_k}{BN_0}\right) \geq \xi_{th}\right)\right] \geq \bar{R}_{th}T, \tag{11}$$

where $E[\cdot]$ denotes the expectation, and $I(\cdot)$ is an indicator function defined by

$$I\left(\xi \geq \xi_{th}\right) = \begin{cases} 1, & \xi \geq \xi_{th} \\ 0, & \xi < \xi_{th} \end{cases}. \tag{12}$$

The expectation in (11) is analytically determined by

$$E\left[\xi(x_k)\, I\left(\xi(x_k) \geq \xi_{th}\right)\right] = \textstyle\int_{\xi_{th}}^{\infty} y f_\xi(y)dy = \int_{x_{th}}^{\infty} \xi(x) f_{X_k}(x)dx, \tag{13}$$

where $f_\xi(\cdot)$ and $f_{X_k}(\cdot)$ are the probability density functions (PDFs) of random variables (RVs) f_ξ and X_k, respectively. Specifically, the RV X_k is defined by $x_k = p_k g_k/(BN_0)$, whereas x_{th} is its threshold value that satisfies $\xi(x_{th}) = \xi_{th}$.

To achieve high semantic similarity, the value of ξ_{th} is usually set above 0.7. In this case, the value of $\xi(x)$ is typically close to A_2, and (13) is tightly upper bounded by

$$E\left[\xi(x_k)\, I\left(\xi(x_k) \geq \xi_{th}\right)\right] \leq A_2 \textstyle\int_{x_{th}}^{\infty} f_{X_k}(x)dx = A_2 F_{X_k}(x_{th}), \tag{14}$$

where $F_{X_k}(\cdot)$ is the cumulative distribution function (CDF) of RV X_k. In the case of Rayleigh fading channel, the PDF of X_k is an exponential function with an average $p_k\Omega_k/(BN_0)$, where $\Omega_k = E[g_k]$. Thus, (13) is upper bounded by

$$E\left[\xi(x_k)\, I\left(\xi(x_k) \geq \xi_{th}\right)\right] \leq A_2 \exp\left(-\frac{x_{th}BN_0}{p_k\Omega_k}\right). \tag{15}$$

We therefore arrive at the following energy minimization problem in fading uplink channels

$$\underset{\tau, t_k, f_k, p_k}{\text{Minimize}} \sum_{k=1}^{K} p_k t_k + \alpha a_0 f_k^2$$

subject to:

$$\begin{aligned} &\bar{C}1: t_k \tfrac{BI}{Ls_k} A_2 \exp\left(-\tfrac{x_{th}BN_0}{p_k\Omega_k}\right) \geq \bar{R}_{th}T, \ \forall k \\ &C3: \tau + \textstyle\sum_{k=1}^{K} t_k = T, \\ &C4: p_k \leq P_{max}, \ \forall k \\ &C5: f_k \leq f_{max}, \ \forall k \\ &C6: \tfrac{a_0}{f_k} \leq \tau, \ \forall k \end{aligned} \tag{16}$$

Setting $\bar{r}_k = \bar{R}_{th}s_k L/(BI)$, $x_{th} = \xi^{-1}(\xi_{th})$ and $\sigma_N^2 = BN_0$, the solution of (16) is given by the following theorem.

Theorem 2. *The feasibility condition for the existence of an optimal solution of (16) is given by*

$$\frac{a_0}{Tf_{max}} + \frac{1}{A_2}\sum_{k=1}^{K} \bar{r}_k \exp\left(\frac{x_{th}\sigma_N^2}{P_{max}\Omega_k}\right) \leq 1. \tag{17}$$

In this case, the optimal local computing time, τ^*, is determined as the solution to the transcendental equation

$$\frac{1}{A_2}\sum_{k=1}^{K} \bar{r}_k T \exp\left(\frac{x_{th}\sigma_N^2}{p_k(\tau^*)\Omega_k}\right) = T - \tau^*, \tag{18}$$

where

$$p_k(\tau) = \frac{x_{th}\sigma_N^2}{2\Omega_k}\left(1 + \sqrt{1 + \frac{8K\alpha a_0^3\Omega_k}{\tau^3 x_{th}\sigma_N^2}}\right), \quad \forall k. \tag{19}$$

Given τ^*, the optimal transmit powers, p_k^*, are determined by

$$p_k^* = \min\left\{p_k(\tau^*), P_{max}\right\}, \quad \forall k, \tag{20}$$

the optimal transmit durations, t_k, are determined by

$$t_k^* = \frac{\bar{r}_k T}{A_2}\exp\left(\frac{x_{th}\sigma_N^2}{p_k^*\Omega_k}\right), \quad \forall k, \tag{21}$$

and the optimal CPU frequencies, f_k^*, are determined by

$$f_k^* = \frac{a_0}{\tau^*}, \quad \forall k. \tag{22}$$

If (17) satisfies a strict inequality, then $f_k^* < f_{max}$, $\forall k$; Otherwise, if (17) satisfies a strict equality, then $f_k^* = f_{max}$, $\forall k$.

Proof: Please refer to Appendix B. ■

V. Numerical Results

We illustrate the energy consumption of proposed schemes in a semantically aware wireless network with $K = 10$ WUs. The (average) gain of both static and fading channels is calculated according to $\Omega_k = \rho \cdot r_k^{-2.7}$, where $\rho = -30$dB is the referent path loss at one meter, the path loss exponent equals 2.7, and r_k is the distance between the kth WU and the BS. System parameters are set to $N_0 = -140$ dBm/Hz, $B = 10$ MHz, and $T = 1$ s; WUs' parameters are set to $\alpha = 10^{-28}$, $a_0 = 0.2$ CPU gigacycles, $P_{max} = 1$ Watts, $f_{max} = 3$ GHz, $\xi_{th} = 0.7$, and $s_k = s_0 = 4, \forall k$, semantic symbols per word. Similar to [7], the parameters of the generalized logistic function in (2) are set to $A_1 = 0.37, A_2 = 0.98, C_1 = 0.2525$, and $C_2 = -0.7985$. Both figures depict the average energy consumption per user, calculated as $E_{avg} = (1/K)\sum_{k=1}^{K}\left(p_k t_k + \alpha a_0 f_k^2\right)$.

Benchmark schemes: For comparative analysis, we use two resource allocation schemes for a conventional wireless

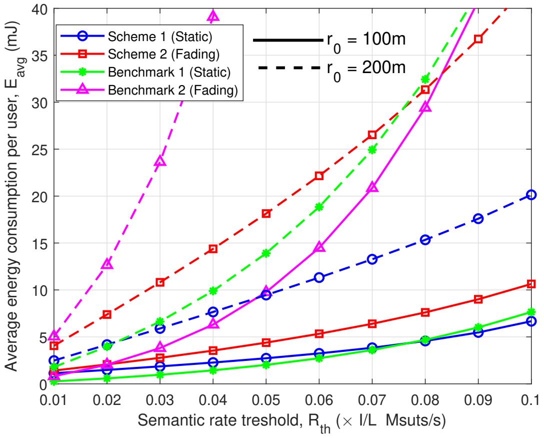

Fig. 2: Impact of semantic rate on energy consumption.

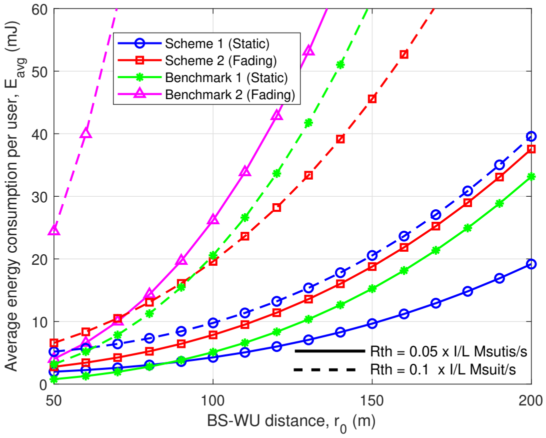

Fig. 3: Impact of channel strength on energy consumption.

network. Data processing in both benchmark schemes is neglected (i.e., $\tau = 0$), yielding a lower bound estimate on the energy consumption. The schemes minimize transmit energy consumption, $E_{c,avg} = (1/K)\sum_{k=1}^{K} p_k t_k$, subject to constraints $C3$ and $C4$, and scheme specific constraint $C1$. *Benchmark Scheme 1* applies to static channels and optimizes $(p_k, t_k), \forall k$, for transmit energy minimization with $C1$ given by $t_k B \log_2(1 + p_k \Omega_k/(BN_0)) \geq R_{c,th} T$; Here, $R_{c,th} = (\mu_0 L/I) R_{th}$ is the bit rate threshold equivalent to the semantic rate threshold R_{th}, where $\mu_k = \mu_0 = 32, \forall k$, is the average number of bits per word [6, Eq. (12)]. *Benchmark Scheme 2* applies to Rayleigh fading channels and optimizes $(p_k, t_k, R_k), \forall k$, for transmit energy minimization with $C1$ given by $t_k B R_k \cdot \Pr\{\log_2(1+p_k g_k/(BN_0)) > R_k\} \geq \bar{R}_{c,th} T$; Here, R_k is the fixed rate of kth (non-semantic) WU, the term $\Pr\{\cdot\}$ is the non-outage probability, and the equivalent average bit rate threshold is given by $\bar{R}_{c,th} = (\mu_0 L/I)\bar{R}_{th}$.

Fig. 2 depicts E_{avg} versus R_{th}, assuming $K = 10$ WUs are uniformly distributed in a circle with radius r_0 around the BS. E_{avg} increases with increasing R_{th} since WUs need to convey increasing amount of semantic data; E_{avg} also increases with r_0 due to the increasing transmit power needed to maintain the desired semantic similarity at the receiver. Relative to the static channel, random fading leads to higher energy consumption due to energy wastage when communication outages occur. For lower R_{th}, a conventional wireless network can consume less energy than a semantically aware network due to the additional energy needed for feature extraction local processing. However, when R_{th} and/or r_0 increase, the energy consumption of both benchmark schemes rapidly increases, because the transmit energy dominates over the energy needed for feature extraction and it takes much more energy to transmit non-semantic data rather than to transmit some of their features, especially in the case of fading channels.

Fig. 3 depicts E_{avg} versus r_0, assuming all WUs are placed along a circle with radius r_0 from the BS. Note, a low value of r_0 actually corresponds to a high value of received SNR, and vice versa. The benchmark schemes perform slightly better in the high SNR regime, because the SNR can support large data transfers while the semantic schemes spend some additional energy on feature extraction. As r_0 increases, the SNR decreases and the proposed semantic schemes significantly outperform the benchmark schemes, because the amount of semantic data is much lower than the non-semantic data. The energy gain of the proposed schemes rapidly increases with increasing r_0 and R_{th}. In our future work, we will apply the proposed energy minimization framework to develop resource allocation schemes for multiple semantic users aimed at realizing a common task.

Appendix A
Proof of Theorem 1

It can be shown that (4) has a unique optimal solution. First, let us observe that, to minimize the objective function, the optimal values of f_k must be as small as possible, but not smaller than the common value of a_0/τ that satisfies $C6$. Since f_k does not appear in the rest of the constraints, $C6$ is satisfied with strict equality, which yields (10). Let us further assume that, in addition to $C6$, the constraints $C1$, $C2$ and $C3$ are also satisfied by strict equalities, yielding (7), (8), and (9), whereas $C4$ and $C5$ are satisfied with strict inequalities. We obtain the feasibility condition (5) from (7) and $\tau \geq a_0/f_{max}$.

In the following, the Lagrange multiplier method is used to verify that the point $(\tau^*, p_k^*, t_k^*, f_k^*)$, determined by (7)-(10), is the desired global minimum solution of (4). Setting $b_k = \sigma_N^2/\Omega_k$, the Lagrangian of (4) is given by

$$\mathcal{L}_1 = \sum_{k=1}^{K}\left(p_k t_k + \frac{\alpha a_0^3}{\tau^2}\right) - \sum_{k=1}^{K}\lambda_k\left(t_k \xi\left(\frac{p_k}{b_k}\right) - r_k T\right) - \sum_{k=1}^{K}\mu_k\left(\xi\left(\frac{p_k}{b_k}\right) - \xi_{th}\right) - \lambda_0\left(\tau + \sum_{k=1}^{K} t_k - T\right), \quad (23)$$

where λ_k, μ_k and λ_0 in (23) are the Lagrange multipliers associated with the constraints $C1$, $C2$, and $C3$, respectively, which must be non-zero if these constraints are satisfied by equality. To show λ_k, μ_k and λ_0 have non-zero values, we set the first derivatives of (23) with respect to p_k, t_k and τ to zero, which yields

$$\frac{d\mathcal{L}_1}{dp_k} = t_k - \frac{\lambda_k t_k}{b_k}\xi'\left(\frac{p_k}{b_k}\right) - \frac{\mu_k}{b_k}\xi'\left(\frac{p_k}{b_k}\right) = 0, \quad \forall k. \quad (24)$$

$$\frac{d\mathcal{L}_1}{dt_k} = p_k - \lambda_0 - \lambda_k \xi\left(\frac{p_k}{b_k}\right) = 0, \quad \forall k. \quad (25)$$

$$\frac{d\mathcal{L}_1}{d\tau} = -\frac{2K\alpha a_0^3}{\tau^3} - \lambda_0 = 0, \quad (26)$$

where the first derivative of ξ at the proposed optimal point satisfies $\xi'(p_k/b_k) = C_1(A_1 - \xi_{th})(A_2 - \xi_{th})/(A1 - A2)$. From (26), we obtain $\lambda_0 = -2K\alpha a_0^3/\tau^3$, which is non-zero for any positive τ, thus validating the assumption that $C3$ is met by strict equality. From (25), we obtain $\lambda_k = \xi_{th}^{-1}(p_k + 2K\alpha a_0^3/\tau^3)$, which is non-zero for any positive τ, p_k, thus validating the assumption that $C1$ is met by strict equality.

From (24), we obtain $\mu_k = t_k\big[b_k/\xi'(p_k/b_k) - \xi_{th}^{-1}(p_k + 2K\alpha a_0^3/\tau^3)\big]$, which is non-zero for any positive τ, p_k, t_k, thus validating the assumption that $C2$ is met by strict equality. Actually, p_k must be as small as possible, but still not smaller than some value that guarantees the minimum semantic similarity of the decoded data, ξ_{th}, imposed by $C2$. Specifically, since the function $\xi(p_k/b_k)$ increases in p_k, the optimal transmit power should satisfy $C2$ by equality. Note, $C4$ is relevant for the feasibility of the solution, c.f. (6).

APPENDIX B
PROOF OF THEOREM 2

Although (16) is a non-convex optimization problem, it can still be shown to have a unique optimal solution. First, applying similar reasoning as in the proof of Theorem 1, we conclude $C6$ is satisfied with strict equality, which yields (22).

Case A) Let us now assume that, in addition to $C6$, the constraints $\bar{C}1$ and $C3$ are also satisfied with strict equalities, whereas $C4$ and $C5$ are satisfied with strict inequalities. In this case, the Lagrange multiplier method can be used to determine the local extrema of (16), which actually turns out to be a single (global) minimum. Setting $b_k = \sigma_N^2/\Omega_k$, the Lagrangian of (16) is given by

$$\mathcal{L}_2 = \sum_{k=1}^{K}\left(p_k t_k + \frac{\alpha a_0^3}{\tau^2}\right) - \sum_{k=1}^{K}\lambda_k \times \left(t_k \exp\left(-\frac{x_{th}b_k}{p_k}\right) - \frac{\bar{r}_k T}{A_2}\right) - \lambda_0\Big(\tau + \sum_{k=1}^{K} t_k - T\Big), \quad (27)$$

where λ_k and λ_0 in (27) are the Lagrange multipliers associated with the constraints $\bar{C}1$ and $C3$, respectively, which must be non-zero if these constraints are satisfied by equality. To determine the stationary points of (16), we set the first derivatives of (27) with respect to p_k, t_k and τ to zero, i.e.,

$$\tfrac{d\mathcal{L}_2}{dp_k} = t_k\left(1 - \lambda_k \exp\left(-\tfrac{x_{th}b_k}{p_k}\right)\tfrac{x_{th}b_k}{p_k^2}\right) = 0, \quad \forall k. \quad (28)$$

$$\tfrac{d\mathcal{L}_2}{dt_k} = p_k - \lambda_0 - \lambda_k \exp\left(-\tfrac{x_{th}b_k}{p_k}\right) = 0, \forall k. \quad (29)$$

$$\tfrac{d\mathcal{L}_2}{d\tau} = -\tfrac{2K\alpha a_0^3}{\tau^3} - \lambda_0 = 0. \quad (30)$$

From (28), we obtain $\lambda_k = (p_k^2/(x_{th}b_k))\exp(x_{th}b_k/p_k)$ which is non-zero for any p_k, thus validating the assumption that $\bar{C}1$ is met with strict equality. From (30), we obtain $\lambda_0 = -2K\alpha a_0^3/\tau^3$ which is non-zero for any $\tau > 0$, thus validating the assumption that $C3$ is met with strict equality. Inserting these expressions for λ_k and λ_0 into (29) yields a quadratic equation with respect to p_k,

$$p_k^2 - x_{th}b_k p_k - \frac{2K\alpha a_0^3 x_{th}b_k}{\tau^3} = 0, \quad (31)$$

which has a single positive solution given by (19). Combining equality constraints $\bar{C}1$ and $C3$ with (19) yields the transcendental expression given by (18). If $T > (1/A_2)\sum_{k=1}^{K}\bar{r}_k T$, then (18) has a unique solution τ^*, because, increasing τ from zero to T, the left hand side of (18) monotonically increases from $(1/A_2)\sum_{k=1}^{K}\bar{r}_k T$ to $(1/A_2)\sum_{k=1}^{K}\bar{r}_k T\exp(x_{th}b_k/p_k(T))$, whereas the right hand side decreases from T to zero. Inserting τ^* into (19) yields the optimal transmit power, $p_k^*(\tau^*)$, whereas the optimal transmit duration is given by $t_k^* = (1/A_2)\,\bar{r}_k T\exp(x_{th}b_k/p_k^*(\tau^*))$.

Case B) Let us assume that, in addition to $\bar{C}1$, $C3$ and $C6$, $C4$ is also satisfied with strict equality, whereas $C5$ is satisfied with strict inequality. In this case, the Lagrangian $\mathcal{L}_2$ is again given by (27) with p_k replaced by P_{max}, while t_k and τ remain as unknown variables. Setting the first derivatives of $\mathcal{L}_2$ with respect to t_k and τ to zero again yields (29) and (30), which implies non-zero values of λ_k and λ_0, thus validating the assumption that $\bar{C}1$ and $C3$ are satisfied by strict equalities.

Combining cases A and B gives (20) and (21). To show that the single stationary point $(\tau^*, f_k^*, p_k^*, t_k^*)$ is actually the (global) minimum of (16), we apply $\bar{C}1$ to the objective function to express t_k via p_k, which gives the function $(1/A_2)\sum_{k=1}^{K}\bar{r}_k T p_k \exp(x_{th}b_k/p_k) + \alpha a_0^3/\tau^2$. Since this function is convex in τ and $p_k, \forall k$, the point $(\tau^*, f_k^*, p_k^*, t_k^*)$ determined by Theorem 2 is the optimal solution of (16).

Case C) If all the constraints of (16) are satisfied by strict equalities, then its feasible solution set is a single point at the intersection of all equality constraints: $p_k^* = P_{max}$, $f_k^* = f_{max}$, $t_k^* = (1/A_2)\,\bar{r}_k T\exp(x_{th}b_k/P_{max})$, and $\tau^* = a_0/f_{max}$. In this case, the solution exists if T is exactly equal to $T_{min} = \tau^* + \sum_{k=1}^{K} t_k^*$. Thus, $T \geq T_{min}$, which implies that (17) is the feasibility condition for the existence of an optimal solution, because it is stricter than the condition $T > (1/A_2)\sum_{k=1}^{K}\bar{r}_k T$ mentioned in Case A.

REFERENCES

[1] D. Gündüz et al., ”Beyond transmitting bits: context, semantics, and task-oriented communications,” *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5-41, Jan. 2023.

[2] 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

[3] Z. Weng and Z. Qin, ”Semantic communication systems for speech transmission,” *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2434–2444, Aug. 2021.

[4] C.-H. Lee, J.-W. Lin, P.-H. Chen, and Y.-C. Chang, ”Deep learning constructed joint transmission-recognition for internet of things,” *IEEE Access*, vol. 7, pp. 76 547–76 561, Jun. 2019.

[5] T.-Y. Tung and D. Gündüz, ”DeepWiVe: Deep-learning-aided wireless video transmission,” *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2570-2583, Sept. 2022

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

[7] 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, Aug. 2023.

[8] J. Chen, J. Wang, C. Jiang and J. Wang, ”Age of incorrect information in semantic communications for NOMA aided XR applications,” *IEEE J. Sel. Topics Sig. Proc.*, vol. 17, no. 5, pp. 1093-1105, Sept. 2023

[9] L. Wang, W. Wu, F. Zhou, Z. Yang, and Z. Qin, ”Adaptive resource allocation for semantic communication networks,” Dec. 2023, *arXiv: 2312.01081*

[10] Z. Yang, M. Chen, Z. Zhang and C. Huang, ”Energy efficient semantic communication over wireless networks with rate splitting,” *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1484-1495, May 2023