Effective Capacity of URLLC over Parallel Fading Channels with Imperfect Channel State Information

Peng Hongsen, Tao Meixia*

Department of Electronic Engineering and the Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai 200240, China
* The corresponding author, email: mxtao@sjtu.edu.cn

Cite as: H. Peng, M. Tao, “Effective capacity of urllc over parallel fading channels with imperfect channel state information,” China Communications, 2024, vol. 21, no. 5, pp. 45-63. DOI: 10.23919/JCC.ea.2021-0795.202401

Abstract: This paper investigates the effective capacity of a point-to-point ultra-reliable low latency communication (URLLC) transmission over multiple parallel sub-channels at finite blocklength (FBL) with imperfect channel state information (CSI). Based on reasonable assumptions and approximations, we derive the effective capacity as a function of the pilot length, decoding error probability, transmit power and the sub-channel number. Then we reveal significant impact of the above parameters on the effective capacity. A closed-form lower bound of the effective capacity is derived and an alternating optimization based algorithm is proposed to find the optimal pilot length and decoding error probability. Simulation results validate our theoretical analysis and show that the closed-form lower bound is very tight. In addition, through the simulations of the optimized effective capacity, insights for pilot length and decoding error probability optimization are provided to evaluate the optimal parameters in realistic systems.

Keywords: effective capacity; finite blocklength regime; imperfect CSI; parallel fading channels; URLLC

I. INTRODUCTION

Ultra-reliable and low-latency communication (URLLC) is one of the main service scenarios supported by 5G wireless networks and beyond. It can enable many mission-critical applications such as autonomous driving, industrial automation, remote surgery and so on [1]. URLLC entails much stricter quality-of-service (QoS) requirements including a small packet error probability between $10^{-9} \sim 10^{-5}$ and a very low latency around 1 ms [1–4]. For URLLC, the latency is defined as end-to-end delay [1, 2], which not only contains the transmission delay in the physical layer but also includes the queueing delay in the buffer. The reliability is defined as the probability that a finite-size data packet is successfully delivered by the transmitter to the receiver within a target time duration [1]. Therefore, the reliability inherently includes the latency constraint and we can also call reliability as delay violation probability as it requires the latency constraint satisfied. Conventionally, due to the stochastic nature of wireless channels, it is quite challenging to support communications that have stringent requirement for delay and reliability.

To ensure low transmission latency, the packet size of URLLC is usually very small. Under the short-packet, or equivalently, finite blocklength (FBL), transmission constraint, the conventional Shannon formula derived from infinite blocklength model with error-free transmission is not applicable [5]. Specif-
ically, transmission at FBL will bring non-negligible loss in the achievable channel coding rate and non-zero probability of decoding error [6]. Another related and important problem is to study the impact of imperfect channel state information (CSI) under the FBL constraint [7]. Training based CSI acquisition will bring the tradeoff between the pilot length and the payload length. On the one hand, longer pilot length will improve the accuracy of the channel estimation but reduce the remaining blocklength for payload transmission. On the other hand, shorter pilot length will deteriorate the estimation of the channel while increase the blocklength for the payload transmission. The aim of this work is to investigate the throughput performance of URLLC from the link layer perspective in term of effective capacity, at FBL taking both statistical delay requirements and imperfect CSI into consideration.

1.1 Related Works

Information-theoretic study on the channel coding rate in the FBL regime was presented in [6] recently, where the authors proposed an accurate approximation (i.e., normal approximation) of the achievable rate over additive white Gaussian noise (AWGN) channels as a function of the signal-to-noise ratio (SNR), blocklength of the codeword, and decoding error probability. The extension to multiple-antenna fading channels was considered in [8, 9]. These new results have inspired many new research papers in different aspects.

Many of the existing works on URLLC have focused on how to achieve the stringent QoS requirement in the physical layer. Xu et al. investigated the energy-efficient packet scheduling problem over quasi-static block fading channels [10]. In [11], Sun et al. analyzed a nonorthogonal multiple access (NOMA) based downlink low latency transmission problem. Hu et al. [12] investigated the relay based cooperative URLLC transmission. Makki et al. [13] studied the hybrid automatic repeat request (ARQ) protocol in block fading channels. However, all of these works assumed perfect CSI at both the transmitter and the receiver. There are also many papers that consider both FBL and imperfect CSI. The works [14–16] investigated communications in block fading channels with pilot length optimizations under the assumption of the imperfect CSI estimation in the physical layer. Cheng et al. [17] investigated the effective capacity in the FBL regime was first analyzed by Gursoy in [23] with queueing constraints. It is proved that there exists a unique decoding error probability that maximizes the effective capacity. Hu et al. [24] investigated the optimal multiuser power allocation problem to maximize the normalized sum effective capacity with fixed decoding error probability. The authors in [25] obtained a closed-form approximation of the effective capacity under Rayleigh fading channel for machine type communications (MTC) through proper expansion and then investigated the power-delay tradeoff for fixed effective capacity. It is worthwhile to mention that, all the above works consider the channel coding performed only within one fading block. The following works consider the channel coding across multiple blocks. Choi [26] studied the effective capacity of parallel multi-channel for low latency communication for both infinite blocklength and FBL. Therein, FBL is only considered for the statistical CSI case where the channel coding rate remains unchanged for all fading blocks. Qiao et al. [27] investigated the effective capacity with FBL channel coding over multiple coherence blocks and re-
revealed the relationships between the decoding error probability, coherence block number and the effective capacity. Note that the tradeoffs among these system parameters are discussed via simulation results only. Nevertheless, none of these effective capacity related works considered the imperfect CSI scenario and finite blocklength at the same time.

The tool of SNC also has been applied for the link layer performance analysis for URLLC. Specifically, Xiao et al. [28, 29] investigated the power allocation problem and analyzed the delay performance in the link layer in downlink NOMA systems utilizing SNC. But these two papers did not take FBL channel coding and imperfect CSI into consideration. Schiessl et al. [30, 31] investigated the delay performance through a newly derived closed-form but approximate decoding error probability by taking both FBL and imperfect CSI into consideration with SNC. These two paper provided rate adaption strategies which lead to a minimum delay violation probability. However, none of these above SNC related works derived explicit closed-form relation between the delay violation probability and the considered parameters.

In this paper, we are primarily interested in the throughput performance with QoS guarantee instead of delay performance. Therefore, effective capacity is more suitable and we employ it as our analysis tool in this work.

1.2 Our Contributions

This work provides an analytical study on the performance in the link layer in terms of effective capacity at FBL with imperfect CSI over multiple parallel fading channels. Instantaneous CSI is assumed unavailable at the transmitter so that the transmitter must send pilot sequence to the receiver for channel estimation, then the receiver feeds back the estimated CSI. The main contributions and findings are as follows:

- Effective Capacity in Exponential Integral Expression and System Parameters’ Impact: We first derive an expression in the form of exponential integral for the effective capacity of the parallel channels following Rayleigh fading at FBL with imperfect CSI. This expression facilitates the evaluation the effective capacity with respect to key system parameters, including the pilot length, the decoding error probability, the transmit power as well as the sub-channel number. More specifically, it is proved that the effective capacity is concave with respect to the pilot length \( n_t \) and its inner term is also concave with respect to the decoding error probability \( \varepsilon \) respectively. This indicates that there exists a unique optimal pilot length \( n_t^* \) and unique optimal decoding error probability \( \varepsilon^* \) that maximize the effective capacity respectively.

- Closed-Form Lower Bound of Effective Capacity and Optimization Algorithm: With reasonable approximations, we derive a closed-form lower bound of the effective capacity possessing the same properties with respect to the aforementioned parameters. Based on the closed-form lower bound, an alternating optimization-based algorithm is proposed to find the optimal pilot length and decoding error probability for maximizing the effective capacity at given transmit SNR and sub-channel number.

- Numerical Validation and Key Observations: Numerical results validate that the lower bound of effective capacity is quite tight over a wide range of system parameters. Furthermore, through the numerical results of the optimized effective capacity, it is shown that the optimal decoding error probability \( \varepsilon^* \) decreases exponentially as the sub-channel number \( m \) or transmit SNR \( \gamma_0 \) (in dB) increases. The optimal pilot length \( n_t^* \) keeps constant as the sub-channel number increases. When the transmit SNR \( \gamma_0 \) increases, the optimal pilot length \( n_t^* \) decreases gradually and eventually keeps constant.

The remainder of this paper is organized as follows: Sec. II presents the system model and preliminaries of the FBL channel coding. Our main contributions are presented in Sec. III, identifying the impact of considered system parameters on the effective capacity and then providing a closed-form lower bound of the effective capacity. Numerical results are presented in Sec. IV. Finally, Sec. V concludes this paper.

II. SYSTEM MODEL AND EFFECTIVE CAPACITY

We consider data transmission from a source to a destination over multiple parallel fading channels using FBL coding as shown in Figure 1. The information...
2.1 Physical Layer Model

Let \( m \) denote the total number of parallel fading channels. Each sub-channel is subject to independent and identically distributed (i.i.d) Rayleigh block fading with additive white Gaussian noise. The fading coefficients, denoted as \( \{h_i\}_{i=1}^{m} \), remain constant within each block consists of \( n \in \mathbb{N} \) channel uses and change independently from one block to another. The channel gain is assumed to be normalized with \( E[|h_i|^2] = 1, \forall i \), and thus each channel coefficient follows the distribution \( h_i \sim \mathcal{CN}(0, 1) \). The transmission block structure of each sub-channel is illustrated in Figure 2, where the first \( n_t = \alpha n \in \mathbb{N} \) channel uses of each block are used for training and the remaining \( n_d = n - n_t \) channel uses are for data transmission.

We ignore the CSI feedback part due to the fact that very few bits will be used to feed back the estimated fading coefficient to the transmitter. Furthermore, the feedback is assumed to be error-free and delay-free for simplicity of analysis.

In the training phase, the transmitter sends a pilot sequence of length \( n_t \) on each sub-channel for channel estimation. The received pilot signal over the \( i \)-th sub-channel, for \( i = 1, ..., m \), can be written as

\[
y_i^p = \sqrt{n_t p} h_i^H q_i + n_i,
\]

where \( p \) is the pilot transmit power at the transmitter, \( q_i \) is the pilot sequence with \( q_i^H q_i = 1 \) and \( n_i \sim \mathcal{CN}(0, \sigma^2) \) is the additive white Gaussian noise.

Then the minimum mean square error (MMSE) estimate of the channel coefficient \( h_i \) is given by [32],[33]

\[
\hat{h}_i = \frac{n_t \gamma_0}{1 + n_t \rho} h_i + \frac{\sqrt{n_t \gamma_0}}{1 + n_t \gamma_0} n_i,
\]

where we define the transmit SNR as \( \gamma_0 = \frac{p}{\sigma^2} \). Hence, \( \hat{h}_i \) is also Gaussian distributed as \( \hat{h}_i \sim \mathcal{CN}(0, \frac{n_t \gamma_0}{1 + n_t \gamma_0}) \).

Figure 1. System diagram for multiple parallel fading channels URLLC transmission.

Figure 2. Illustration of the block structure on each sub-channel: the transmitter first sends pilot to the receiver, then the receiver feeds back the fading coefficient to the transmitter (omitted in the figure), finally the transmitter sends data to the receiver according to the estimated CSI.
The relationship between the MMSE estimate of the channel coefficient \( \hat{h}_i \) and its true value \( h_i \) can be given as [32]
\[
h_i = \hat{h}_i + z_i, \quad (3)
\]
where the channel estimation error \( z_i \sim \mathcal{CN}(0, \frac{1}{1+n_i \gamma_0}) \) is independent of \( \hat{h}_i \). After the training phase, the receiver feeds back the estimated channel coefficients to the transmitter using an ideal feedback link.

In the data transmission phase, both the receiver and the transmitter treat the estimated CSI as the realistic channel coefficients. Thus the estimation error \( z_i \) will be regarded as noise. The received signal on sub-channel \( i \) is given by:
\[
y_i = \sqrt{p}h_is_i + w_i = \sqrt{p}\hat{h}_is_i + \sqrt{p}z_is_i + w_i, \quad (4)
\]
where \( s_i \) is the transmitted signal with unit-power on the \( i \)th sub-channel, \( p \) is the data transmit power on each sub-channel and \( w_i \) is the complex Gaussian noise with zero mean and variance \( \sigma^2 \). Here we have assumed the training phase and data transmission phase have the same transmit power and the same noise power. According to the property of MMSE estimation, the estimation error and the estimated channel coefficient \( \hat{h}_i \) are independent. Thus the second noise term \( \{v_i\} \) in (4) can be viewed as independent to the first signal term.

The instantaneous received SNR of the \( i \)th sub-channel can be represented as
\[
\hat{\gamma}_i = p|\hat{h}_i|^2 \bigg/ \sigma^2 + \frac{p|\hat{h}_i|^2}{1+n_i \gamma_0} = \frac{n_i \gamma_0}{1 + \gamma_0 + n_i \gamma_0} x_i = Gx_i, \quad (5)
\]
where \( |\hat{h}_i|^2 = \frac{n_i \gamma_0}{1+n_i \gamma_0} x_i \), \( x_i \) is an exponential distributed random variable with probability density function \( f_{PDF}(x_i) = e^{-x_i} \), and \( G = \frac{n_i \gamma_0}{1+n_i \gamma_0} \) can be viewed as the average received SNR taking channel estimation error into account.

We assume that channel coding is performed across the \( m \) sub-channels, since these \( m \) parallel channels with estimated channel coefficients can be viewed as parallel AWGN channels with different noise variances. According to [34], the achievable coding rate in bits per channel use is shown in (6),
\[
R(\{\hat{\gamma}_i\}_1^m, \varepsilon) = \frac{1}{m} \sum_{i=1}^{m} \log_2 (1 + \hat{\gamma}_i) - \sqrt{\frac{V(\{\hat{\gamma}_i\}_1^m)}{n_d m}} Q^{-1}(\varepsilon) + o \left( \frac{\log_2 n_d}{mn_d} \right), \quad (6)
\]
where \( \varepsilon \) is the block decoding error probability, \( Q^{-1}(\cdot) \) is the inverse of the Gaussian Q function
\[
Q(x) = \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-t^2} dt, \quad (7)
\]
and \( V \) is the channel dispersion
\[
V(\{\hat{\gamma}_i\}_1^m) = \frac{1}{m} \sum_{i=1}^{m} \left( 1 - \frac{1}{(1 + \hat{\gamma}_i)^2} \right) \log_2 e. \quad (8)
\]
The achievable rate expression (6) is known as the normal approximation. For simplicity of analysis, we neglect the third term and approximate the channel dispersion term with an upper bound
\[
V(\{\hat{\gamma}_i\}_1^m) \approx \log_2 e. \quad (9)
\]
Then, an approximate lower bound of the achievable rate at FBL can be obtained as:
\[
R(\{\hat{\gamma}_i\}_1^m, \varepsilon) = \frac{1}{m} \sum_{i=1}^{m} \log_2 (1 + \hat{\gamma}_i) - \frac{\log_2 e}{\sqrt{n_d m}} Q^{-1}(\varepsilon). \quad (10)
\]
This approximation is shown to be accurate at high SNR (\( \geq 3 \)dB) in [35], especially when multiple parallel channels are considered.

### 2.2 Link Layer Model and Effective Capacity with Imperfect CSI at FBL

We assume that the transmitter applies a simple automatic repeat request (ARQ) mechanism as in [27]. Namely, when each transmission is over, the receiver can reliably detect the error transmission and then send a negative acknowledgement requesting for retransmission in case of transmission error. This feedback link is also assumed to be error-free and delay-free. When error occurs, the service rate of a specific transmission block is zero. Therefore, the service rate at the \( j \)th transmission block (in bits per block) can be
expressed as
\[
s(j) = \begin{cases} 
0, & \text{probability } \varepsilon \\
mn_d R(\{\hat{\gamma}_i\}_1^n, \varepsilon), & \text{probability } 1 - \varepsilon.
\end{cases}
\]

(11)

The sequence \( \{s(j), j = 1, 2, \ldots\} \) is a discrete-time stationary and ergodic stochastic service process.

Effective capacity is the maximum constant arrival rate that a given service process can support with statistical QoS guarantee specified by the QoS exponent \( \theta \) \[19\]. It can be calculated as \[19\]

\[
C_E(\theta) = - \lim_{t \to \infty} \frac{1}{\theta t} \ln \mathbb{E}\{e^{-\theta s(1,t)}\} \text{ bits/block},
\]

(12)

where \( S(1, t) = \sum_{i=1}^{t} s(i) \) is the accumulated amount of service bits up to the \( t \)-th transmission block.

Given the effective capacity \( C_E(\theta) = \mu \), the probability that the transmission delay of the arrived information bits in slot \( t \), denoted as \( D(t) \), exceeds a target delay bound \( D_{\text{max}} \), i.e., delay violation probability, can be expressed approximately as \[19\]

\[
P(D(t) > D_{\text{max}}) \approx \eta e^{-\theta \mu D_{\text{max}}},
\]

(13)

where \( \eta \) is the buffer non-empty probability and the QoS exponent \( \theta \) indicates the decaying speed. When the effective capacity and the delay bound are given, for larger \( \theta \), the delay violation probability decays faster. Therefore, throughout performance with different statistical QoS guarantee (i.e., delay violation probability) can be characterized appropriately by the effective capacity with different decaying exponent \( \theta \).

By definition (12), the effective capacity of \( m \) parallel channels at FBL can be given by

\[
C_E(\theta) = - \lim_{t \to \infty} \frac{1}{\theta t} \ln \mathbb{E}\{e^{-\theta s(1,t)}\}.
\]

(14)

\[
= - \lim_{t \to \infty} \frac{1}{\theta t} \ln \mathbb{E}\{e^{-\theta s(j)}\}^t.
\]

(15)

\[
= - \frac{1}{\theta} \ln \mathbb{E}\{e^{-\theta s(j)}\}.
\]

(16)

\[
= - \frac{1}{\theta} \ln \left\{ \varepsilon + (1 - \varepsilon)e^{-\theta mn_d R(\{\hat{\gamma}_i\}_1^n, \varepsilon)} \right\},
\]

(17)

where the expectation is with respect to \( \hat{\gamma} = [\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_m] \). Note that (15) follows from the fact that the services process \( s(j) \) changes independently from one block to another. By substituting (10) into (17), effective capacity is obtained as shown in (18)-(20).

\[
C_E(\theta) = - \frac{1}{\theta} \ln \mathbb{E}\{e^{-\theta mn_d R(\{\hat{\gamma}_i\}_1^n, \varepsilon)}\}
\]

(18)

\[
= - \frac{1}{\theta} \ln \left\{ \varepsilon + (1 - \varepsilon)e^{\theta \varepsilon Q^{-1}(\varepsilon)\ln mn_d e} \right\}.
\]

(19)

\[
= - \frac{1}{\theta} \ln \left\{ \varepsilon + (1 - \varepsilon) \left[ e^{-\theta mn_d (1 + \hat{\gamma})} + \theta mn_d (1 + \hat{\gamma}) e^{\theta \varepsilon Q^{-1}(\varepsilon)\ln mn_d e} \right] \right\}.
\]

(20)

2.3 Key System Parameters

The main focus of this paper is to investigate the impact of the key system parameters, including the pilot length \( n_t \), decoding error probability \( \varepsilon \), transmit power \( p \) and the sub-channel number \( m \) on the effective capacity. They are elaborated as follows:

**Impact of pilot length \( n_t \):** When using a long training sequence of \( n_t \) symbols, the channel estimation becomes more accurate, allowing transmissions with higher reliability but leaving fewer symbols \( n_d = n - n_t \) for the data transmission. While using a short training sequence, the channel estimation becomes inaccurate while the channel uses for the data transmission increases. Thus, the parameter \( n_t \) should be chosen carefully. This is particularly the case for FBL transmission since the blocklength reduction of
the data transmission part deteriorates the communication performance more rapidly due to the second order penalty term in (6).

**Impact of the decoding error probability $\varepsilon$:** When adopting a larger decoding error probability, the transmission rate is larger while more retransmissions will be needed; when adopting a smaller decoding error probability, the channel coding rate decreases while the number of retransmission will be small. Finding the optimal decoding error probability that maximizes the system throughput is quite important.

**Impact of the transmit power $p$ (i.e., transmit SNR):** The throughput increases as the transmit power increases by no means. But how does the throughput increase, especially in the FBL regime is still under investigated. Answering this question may help us design more practical power control scheme for URLLC systems.

**Impact of the number of sub-channels $m$:** Similar to the transmit power, how does the throughput increase as the number of sub-channels increases? This impact can provide valuable insights for subcarrier allocation in the OFDM system.

**III. PERFORMANCE ANALYSIS**

In this section, we first show the impact of the pilot length and the decoding error probability on the effective capacity and identify the optimal tradeoffs. Then the impact of the transmit power as well as the sub-channel number on the effective capacity are addressed. We then derive a closed-form lower bound of the effective capacity. Finally, we propose an alternating optimization based algorithm to maximize the effective capacity through optimizing the pilot length and the decoding error probability iteratively.

**3.1 Impact of the Considered Parameters on Effective Capacity**

In this subsection, we will investigate the impact of the pilot length $n_t$, transmit power $p$, sub-channel number $m$ and decoding error probability $\varepsilon$ on the effective capacity. Firstly, let us consider the impact of the decoding error probability $\varepsilon$. Denote the inner function in (18) with respect to $\varepsilon$ as shown in (21).

$$T(\varepsilon) = E \left\{ \varepsilon + (1 - \varepsilon) \varepsilon + \sum_{i=1}^{m} \log_2 \left( 1 + \hat{\gamma}_i \right) - \frac{\log_2 e}{mn_d Q^{-1}(\varepsilon)} \right\}. \tag{21}$$

**Theorem 1.** For given $\theta$, $m$, $n_t$ and $p$, $T(\varepsilon)$ is strictly convex in $\varepsilon$.

**Proof.** This theorem follows directly Theorem 1 in [27], with the only difference that our channel coding is performed over multiple sub-channels and while it is over multiple coherence blocks in [27].

Theorem 1 indicates that there exists an optimal decoding error probability that maximizes the effective capacity when the other parameters are given. This result reveals the optimal tradeoff between the decoding error probability (retransmissions) and the channel coding rate in terms of effective capacity.

Then we will consider the impact of the pilot length. Note that we shall remove the integer constraint of $n_t$ in the following analysis.

**Theorem 2.** For given $\theta$, $m$, $\varepsilon$ and $\gamma_0$, when $\alpha = \frac{n_t}{n} \in (0, 0.2)$, the effective capacity $C_E(\theta)$ is concave in $n_t$.

**Proof.** Please see appendix A.

Theorem 2 indicates that if the ratio of the pilot length over the entire transmission blocklength is upper bounded by a certain value, there exists a unique pilot length $n_t^*$ that maximizes the effective capacity $C_E(\theta)$. This result identifies the optimal tradeoff between the pilot length and the payload length in terms of the effective capacity.

**Theorem 3.** Assume that the received SNR $\hat{\gamma}_i > -3dB$, $\forall i \in \{1, 2, ..., m\}$, for given $\theta$, $m$, $\varepsilon$ and $n_t$, the effective capacity $C_E(\theta)$ is concave and monotonically increasing with respect to $p$. 
Proof. Please see appendix B.

This result indicates that under the given minimum received SNR assumption, as the transmit power increases, the effective capacity increases but in a diminishing manner. When the transmit power is large enough, the inner term of (21) is bounded by $\varepsilon$, hence the effective capacity is bounded by $-\frac{1}{\theta} \ln \varepsilon$. In this case and the decoding error probability becomes the main factor which affects the effective capacity.

**Theorem 4.** For given $\theta$, $p$, $n_t$ and $\varepsilon$, the effective capacity $C_E(\theta)$ is monotonically increasing with respect to $m$.

**Proof.** Please see appendix C.

Similarly, there is also an obvious upper bound of the effective capacity $-\frac{1}{\theta} \ln \varepsilon$. This result indicates, as the sub-channel number increases, the effective capacity also increases. When the sub-channel number is large enough, the effective capacity remains constant and the decoding error probability becomes the main factor which affects the effective capacity.

### 3.2 Closed-Form Approximation of the Effective Capacity

The aforementioned analysis indicates the existence of the impact of the parameters on the effective capacity. How to obtain the optimal parameters is a non-trivial problem, because the effective capacity in (20) is still an integral form. In this subsection, we will provide a closed-form lower bound expression of the effective capacity and verify the corresponding properties.

The effective capacity can be transformed as shown in (22)-(25):

\[
C_E(\theta) = -\frac{1}{\theta} \ln \mathbb{E} \left\{ \varepsilon + (1 - \varepsilon)e^{-\theta mn_d} \left\{ \frac{1}{m} \sum_{i=1}^{m} \log_2(1+\gamma_i) - \sqrt{\frac{\log_2\varepsilon}{mn_d}} Q^{-1}(\varepsilon) \right\} \right\} 
\]

(22)

\[
= -\frac{1}{\theta} \ln \left\{ \varepsilon + (1 - \varepsilon)e^{\theta' \sqrt{mn_d} \gamma^{-1}(\varepsilon)} \mathbb{E} \left\{ (1 + Gx_1)^{-\theta' n_d} \right\}^m \right\} 
\]

(23)

\[
= -\frac{1}{\theta} \ln \left\{ \varepsilon + (1 - \varepsilon)e^{\theta' \sqrt{mn_d} \gamma^{-1}(\varepsilon)} \int_{0}^{\infty} (1 + x_1)^{-\theta' n_d} e^{-x} dx_1 \right\}^m \]

(24)

\[
= -\frac{1}{\theta} \ln \left\{ \varepsilon + (1 - \varepsilon)e^{\theta' \sqrt{mn_d} \gamma^{-1}(\varepsilon)} \frac{1}{G} e^{\theta' n_d} \left\{ \frac{1}{G} \right\}^{-m} \right\} 
\]

(25)

therein, $\theta' = \theta \log_2 e$ and $E_v(x)$ is the exponential integral given by

\[
E_v(x) = \int_{1}^{\infty} e^{-xt} t^{-v} dt. 
\]

(26)

From (22) to (23), we employ the fact that all the sub-channels are i.i.d. Then, by applying the upper bound of exponential integral in [36]

\[
E_v(x) \leq \frac{e^{-x}}{v + x - 1}, v > 1, 
\]

(27)

a closed-form lower bound of the effective capacity is obtained in (28).

\[
C_E(\theta) = -\frac{1}{\theta} \ln \left\{ \varepsilon + (1 - \varepsilon)e^{\theta' \sqrt{mn_d} \gamma^{-1}(\varepsilon)} \times \right\} 
\]

(28)

Note that one should have $\theta' n_d > 1$ i.e., $\theta n_d \log_2 e > 1$ to make the upper bound (27) valid. Furthermore, $\theta n_d \log_2 e > 2$ can ensure the upper bound (27)’s tightness and in the following, we will mainly focus on this condition. This condition is reasonable because in the URLLC transmission, the QoS exponent $\theta$ is larger than 0.01 and the blocklength is larger than 200. The closed-form lower bound (28) provides an explicit relationship among the QoS exponent $\theta$, blocklength $n$, decoding error probability $\varepsilon$, and the effective capacity.
average received SNR $G$, and the sub-channel number $m$.

It is obvious that the properties of the closed-form lower bound $C_E(\theta)$ with respect to $\varepsilon$ and $m$ are the same as $C_E(\theta)$. Next we will verify the impact of the pilot length $n_t$ and the transmit power $p$ on the effective capacity lower bound $C_E(\theta)$. Firstly, fix $\varepsilon$ and remove it for simplicity. Furthermore, we employ the variable $n\alpha$ instead of the integer $n_t$ for convenience. Here we assume $\alpha = \frac{n\alpha}{n}$ is constrained by $\alpha \in (0, 0.2)$. Then we denote $\Gamma(m, \gamma_0, \alpha)$ as shown in (29)-(31). From (29)-(30), we employ the property of logarithmic function and from (30)-(31), we adopt the same approximation (A.4) employed in the proof of Theorem 2 in the Appendix.

$$\Gamma(m, \gamma_0, \alpha) = -\frac{1}{\theta} \ln \left[ e^{\theta \sqrt{mn_{d}Q^{-1}(\varepsilon)}} ((\theta' n_{d} - 1)G + 1)^{-m} \right]$$

$$\approx -\sqrt{mn_{d}Q^{-1}(\varepsilon)} \log \varepsilon + \frac{m}{\theta} \ln((\theta' n_{d} - 1)G + 1)$$

$$\approx -\sqrt{mn_{d}Q^{-1}(\varepsilon)} \log \varepsilon \left(1 - \frac{\alpha}{2} + \frac{m}{\theta} \ln((\theta' n(1 - \alpha) - 1)G + 1)\right)$$  

(29) 

(30) 

(31) 

**Theorem 5.** Assume that $\theta n_{d} \log_2 e > 2$, for given $\theta$, $m$, $\varepsilon$ and $p$, when $\alpha \in (0, 0.2)$, the closed-form lower bound of the effective capacity $C_E(\theta)$ is concave over $\alpha$.

**Proof.** Please see Appendix D.

It is easy to prove that the optimal pilot length $n_t^*$ that maximizes $\Gamma(m, \gamma_0, \alpha)$ also maximizes $C_E(\theta)$. Thus, to obtain the optimal pilot length $n_t^*$, we can directly calculate the root of the first order partial derivative of $\Gamma(m, \gamma_0, \alpha)$ in (31) with respect to $\alpha$ and the solution can be found by using binary search method. Note that $\alpha$ is a continuous variable and is in the interval (0, 0.2), while $n_t$ is an integer, thus when we find the optimal $\alpha^*$, it is necessary to compare the two adjacent integers to obtain the optimal $n_t^*$.

**Theorem 6.** Assume that the average received SNR $G > -3dB$ and $\theta n_{d} \log_2 e > 2$, for given $\theta$, $m$, $\varepsilon$ and $n_t$, the closed-form lower bound of the effective capacity $C_E(\theta)$ is concave and monotonically increasing with respect to $p$.

**Proof.** Please see Appendix E.

This theorem indicates that the closed-form lower bound has the same properties as the original function of effective capacity.

3.3 Joint Optimization of Pilot Length and Decoding Error Probability

In this subsection, our purpose is to maximize the effective capacity at given transmit power $p$ and the sub-channel number $m$ by jointly optimizing the pilot length $n_t$ and the decoding error probability $\varepsilon$. The problem is challenging since the effective capacity is not jointly concave in $(n_t, \varepsilon)$. However, the effective capacity is concave in both $n_t$ and $\varepsilon$ individually, here we adopt the the alternating optimization method, which can guarantee convergence. We first give the initial value of a decoding error probability. The optimal $n_t$ can be calculated through the first order derivative of (31) with respect to $\alpha$. The binary search method can find the optimal pilot length $n_t = \lceil n\alpha \rceil$ or $n_t = \lfloor n\alpha \rfloor$ efficiently. Then we take the derived $n_t$ as constant, the optimal decoding error probability can be updated through the first order derivative or directly obtained by leveraging the ternary search method of (28) with respect to $\varepsilon$. Indeed, the expression of the first order derivative of (28) is very complex, the ternary search method is employed in this paper. Based on previous analysis, we can obtain the optimal parameters in each iteration. Thus the convergence of the proposed method can be guaranteed. The overall procedure is outlined in Algorithm 1.
Algorithm 1. Alternating iterative method for effective capacity maximization.

Input: Number of sub-channels \( m \), transmit SNR \( \gamma_0 \);
Output: Optimal pilot length \( n_t \), optimal decoding error probability \( \varepsilon \);

1: Initialize \( \varepsilon(0) = 10^{-3} \), \( i = 1 \)
2: do
3: Calculate the optimal pilot length \( n_t(i) \) that maximizes (31) at given \( \varepsilon(0) \), then calculate the optimized effective capacity \( C_{E_n}(i) \) by (28);
4: Calculate the optimal decoding error probability \( \varepsilon(i) \) that maximizes (28) at given pilot length \( n_t(i) \), then calculate the optimized effective capacity \( C_{En}(i) \) by (28);
5: while \( C_{En}(i) - C_{E_n}(i) > 10^{-4} \)
6: return \( \varepsilon(i), n_t(i) \)

IV. NUMERICAL RESULTS

In this section, we evaluate our derived effective capacity and the closed-form lower bound under different conditions. Due to the accuracy of the channel dispersion approximation in (9), we mainly focus on the medium and high SNR scenarios. We set the noise power \( \sigma^2 = 1 \) for simplicity and thus the transmit SNR is actually the value of transmit power \( p \). Note that the low latency communication is considered, the QoS requirement of each user is quite stringent, we thus set the QoS exponent as 0.01 and the number of channel uses in each block is 300 unless otherwise addressed. Hence the approximation in (27) is tight under this condition \( \theta n_d \log_2 e > 2 \).

To elaborate this further, let us assume the target delay bound is \( D_{max} = 5 \) and the effective capacity is \( \mu = 300 \). Then, according to (13), we have the delay violation probability upper bounded by \( \Pr(D > D_{max}) = e^{\theta \mu D_{max}} \approx 4.5 \times 10^{-5} \), which is small enough to be considered as ultra-reliable.

To validate the analysis of effective capacity, we present three curves with the same parameters generated respectively by the following three methods: analytical results computed directly from the closed-form lower bound (28), Monte-Carlo simulation based on the definition in (22), and analytical results computed from the exponential integral form (25). The purpose is two-fold. The first is to validate the correctness of the theoretical derivation (25) using simulation based on (22). The detailed Monte-Carlo simulation shall be introduced in due course. The second is to verify the tightness of the lower bound (28) by comparing with (25) or (22).

In Subsec. 4.1, we present numerical results to validate the theoretical analysis. Then in Subsec. 4.2, we present the optimized effective capacity versus the sub-channel number and the transmit SNR based on our proposed alternating iterative method, where the corresponding optimal pilot length and optimal decoding error probability are also presented.

4.1 Theoretical Analysis and Approximation Accuracy Validation

In Figure 3(a), we validate the analysis with respect to the pilot length \( n_t \in \mathbb{N} \) and show the tightness of the closed-form lower bound with \( m = 5 \) sub-channels. The transmit SNR and corresponding decoding error probability of the three groups are \((3dB, 4.33 \times 10^{-6}), (6dB, 7.92 \times 10^{-5}), (9dB, 2.02 \times 10^{-7})\), we generate
block realizations to calculate the Monte-Carlo simulation results based on the definition in (22). It is first seen that the simulation results based on (22) match exactly with the analytical results based on the exponential integral form (25). It is also seen that the effective capacity is concave with respect to $n_t$ and the closed-form expression is exactly a lower bound of the effective capacity. In addition, the gap between the closed-form lower bound and the effective capacity decreases as the transmit SNR increases. The trends of the three curves with the same parameters are all concave and the optimal pilot length (points in the circles) that maximizes the effective capacity decreases as the transmit SNR increases.

In Figure 3(b), we validate the theoretical analysis with respect to the decoding error probability $\varepsilon$ and show the tightness of the closed-form lower bound for different transmit SNR $\gamma_0$. The number of sub-channels is $m = 5$. The transmit SNR and pilot length for the three groups are $(3dB, 19)$, $(6dB, 18)$, $(9dB, 17)$ respectively. We generate $10^{10}$ block realizations to calculate the Monte-Carlo simulation results. It is seen that the effective capacity is maximized by a specific decoding error probability as in [27] and our proposed closed-form lower bound is exactly a lower bound of (25). The trends of the three curves with the same parameters stay consistent and it can be seen that the optimal decoding error probabilities of the three curves in the same group are almost the same, which also shows the strength of our proposed closed-form lower bound. Furthermore, we can observe that the optimal decoding error probability decreases as the transmit SNR increases when the other parameters stay constant. The gap among the three groups are very small when $\varepsilon$ is large, while when $\varepsilon$ becomes smaller, the gap increases. The gap between the closed-form lower bound and the other two curves is inherently due to the exponential integral upper bound in (27). With higher transmit SNR, the upper bound is more tight and the gap becomes smaller.

In Figure 4(a), we show the effective capacity and the closed-form lower bound as a function of the transmit SNR $\gamma_0$ (not dB for intuition) with $\varepsilon = 10^{-10}$, $n_t = 20$ and set 3 groups with different values of $m$ for comparison. We generate $10^{11}$ block realizations to evaluate the Monte-Carlo simulation of each point. Firstly, it is observed that the effective capacity is concave and monotonically increasing with respect to $\gamma_0$, which validates our analysis. Moreover, when $m = 7$, the effective capacity approaches the limit $-\frac{1}{T} \ln \varepsilon$ as $\gamma_0$ increases. With given decoding error probability, there exists a transmit SNR threshold, when the transmit SNR is higher than this threshold, the effective capacity increases very slowly. When the sub-channel number $m$ increases, the threshold decreases. Secondly, the gap between the effective capacity and closed-form lower bound decreases as the transmit SNR increases.

In Figure 4(b), we show the effective capacity and the closed-form lower bound with respect to the sub-channel number $m$. Here we assume $n_t = 20$ and show different transmit SNR values for comparison. We generate $10^{11}$ block realizations to evaluate the Monte-Carlo simulation of each point. It is seen that the effective capacity is monotonically increasing with respect to $m$ and then converges to $-\frac{1}{T} \ln \varepsilon$, which proves Theorem 4. With given decoding error proba-
bility, there also exists a sub-channels threshold, when the sub-channel number is larger than the threshold, the effective capacity increases very slowly. When the transmit SNR increases, the threshold decreases and the gap between the effective capacity and the closed-form lower bound is smaller.

From Figure 4(a), it is seen the closed-form lower bound (28) is very tight. While as shown in from Figure 4(b), there is a small gap for the group $\gamma_0 = 3$ dB when $m \geq 4$. This gap is due to the number of sub-channels on the exponential position amplify the exponential integral approximation error in (27) and thus with larger $m$, the gap is bigger. The gap finally diminishes because all of the three curves within the same group gradually converge to $-\frac{1}{\theta} \ln \varepsilon$.

4.2 Optimized Effective Capacity

In this subsection, we present the optimized effective capacity as a function of the sub-channel number and the transmit SNR. Then we show the corresponding variations of the optimal pilot length $n_t^*$ and the optimal decoding error probability $\varepsilon^*$.

In Figure 5(a), we show the optimized effective capacity and the closed-form lower bound versus the number of sub-channels with different transmit SNR $\gamma_0$. We can see that the optimized effective capacity is monotonically increasing almost linearly with respect to $m$. For higher transmit SNR $\gamma_0$, the slope of the optimized effective capacity is larger.

We show the optimized effective capacity as function of the transmit SNR $\gamma_0$ (in dB) with different $m$ in Figure 5(b). The effective capacity is concave and monotonically increasing with respect to the transmit SNR $\gamma_0$ and the closed-form lower bound is very close to the effective capacity. Furthermore, for larger $m$, the slope of the optimized effective capacity is also larger.

Figure 6(a) shows the optimal pilot length of the optimized effective capacity as a function of $m$ for different transmit SNR $\gamma_0$. The optimal pilot length of given transmit SNR doesn’t change as the number of sub-channels increases. While the optimal pilot lengths with different transmit SNR are different. Specifically, higher SNR corresponds to shorter pilot length.

Figure 6(b) shows the optimal pilot length of the optimized effective capacity as a function of $\gamma_0$ with different $m$. It is seen that the optimal pilot lengths with different $m$ are exactly the same and gradually decrease and finally converge to a specific value as the transmit SNR increases in our simulated setting.

Figure 7(a) depicts the corresponding optimal values of the decoding error probability as a function of sub-channel number $m$. The optimal decoding error probability decreases exponentially as the sub-channel number increases. For higher transmit SNR $\gamma_0$, the optimal decoding error probability decays faster.

Figure 7(b) depicts the corresponding optimal values of the decoding error probability as a function of the transmit SNR $\gamma_0$. The optimal decoding error probability decreases exponentially as the transmit SNR (in dB) increases. For larger $m$, the optimal decoding error probability decays faster.

As a final remark, the main insights from the above numerical study are as follows: As $m$ or $\gamma_0$ increases, by optimizing $n_t^*$ and $\varepsilon$ iteratively, the optimized effective capacity can increase almost linearly. This is
due to the optimal decoding error probability $\varepsilon^*$ is decreasing exponentially, the transmission rate is not constrained by $\varepsilon^*$, i.e., the tradeoff between decoding error probability and coding rate is met. On the other hand, by optimizing the pilot length $n_t$, the transmission and training tradeoff can also be met. The optimized pilot length keeps constant as the number of sub-channels changes and decreases gradually as the transmit power $\gamma_0$ increases. This result coincides with intuition, i.e., the channel estimation is related to the channel statistical characteristics instead of the sub-channel number, and with higher transmit SNR, the shorter pilot length is required. So in the realistic system, the effective capacity can be calculated rapidly under certain conditions, e.g., with given transmit SNR, if we obtain two points of the curves, we can estimate the effective capacity with more or less sub-channels and the corresponding optimized decoding error probability can also be updated easily utilizing a linear approximation.

V. CONCLUSION

In this paper, we studied the throughput performance under statistical QoS requirement in terms of effective capacity over parallel fading channels at FBL with imperfect CSI. Firstly, we analyzed the impact of the considered parameters on the effective capacity. Then we derived a closed-form lower bound by adopting reasonable approximation of the exponential
integral and verified the aforementioned properties. Furthermore, we propose an alternating optimization-based algorithm to maximize the effective capacity by optimizing decoding error probability and the pilot length iteratively with given sub-channels and transmit SNR based on our proposed closed-form lower bound. Numerical results validated our analysis. Then we showed optimized effective capacity and corresponding optimal pilot length as well as optimal decoding error probability. As a result, this work can provide some insights to guide the design of key system parameters in practical URLLC systems.

\[ L(m, \gamma_0, n_t) = -\frac{1}{\theta} \ln \left\{ \mathbb{E} \left[ e^{-\theta n_d \log_2 (1+\hat{\gamma}_i)} + \theta n_d \sqrt{\frac{\log_2 e}{m n_d}} Q^{-1}(\epsilon) \right]^m \right\}. \]  

(\text{A.1})

Then, we transform the expression in (A.1) as shown in (A.3).

\[ L(m, p, n_t) = -\sqrt{mn_d} Q^{-1}(\epsilon) \log_2 e - \frac{m}{\theta} \ln \left\{ \mathbb{E} \left[ e^{-\theta n_d \log_2 (1+\hat{\gamma}_i)} \right] \right\} \approx -\sqrt{mn} Q^{-1}(\epsilon) \log_2 e (1 - \frac{n_t}{2n}) - \frac{m}{\theta} \ln \left\{ \mathbb{E} \left[ e^{-\theta n_d \log_2 (1+\hat{\gamma}_i)} \right] \right\}. \]

(\text{A.2})

\[ \approx -\frac{m}{\theta} \ln \left\{ \mathbb{E} \left[ e^{-\theta n_d \log_2 (1+\hat{\gamma}_i)} \right] \right\}. \]

(\text{A.3})

From (A.2) to (A.3), the following approximation is employed

\[ \sqrt{1 - \frac{n_t}{n}} \approx 1 - \frac{n_t}{2n}, \]

(A.4)

where this approximation is accurate enough when \( \frac{n_t}{n} \in (0, 0.2) \). Due to the linearity of the first term, it doesn’t alter the convexity of \( L(n_t) \), then the focus is on the latter function. Note that the latter term is quite complicated and the pilot length \( n_t \) is implicitly included in \( n_d \) and \( \hat{\gamma}_i \), we will show the convexity of \( e^{-\theta n_d \log_2 (1+\hat{\gamma}_i)} \) and then show the convexity of the overall latter term through compound function with respect to \( n_t \). Denote

\[ \Psi(p, n_t) = \log_2 (1 + \hat{\gamma}_i). \]

(A.5)

The first order and second order partial derivatives of \( \Psi(p, n_t) \) with respect to \( n_t \) are derived as follows:

\[ \frac{\partial \Psi}{\partial n_t} = \frac{\partial \Psi}{\partial \hat{\gamma}_i} \frac{\partial \hat{\gamma}_i}{\partial n_t} = \frac{\log_2 e}{1 + \hat{\gamma}_i} \frac{\gamma_0^2 (\gamma_0 + 1 + n_t \gamma_0)^2}{(1 + \hat{\gamma}_i (1 + \gamma_0 + n_t \gamma_0))^2} \geq 0, \]  

(\text{A.6})

\[ \frac{\partial^2 \Psi}{\partial n_t^2} = \frac{\partial^2 \Psi}{\partial \hat{\gamma}_i^2} \left( \frac{\partial \hat{\gamma}_i}{\partial n_t} \right)^2 + \frac{\partial \Psi}{\partial \hat{\gamma}_i} \frac{\partial^2 \hat{\gamma}_i}{\partial n_t^2} = \frac{\log_2 e}{(1 + \hat{\gamma}_i)^2} \left( \frac{\gamma_0^2 (\gamma_0 + 1 + n_t \gamma_0)^2}{(1 + \hat{\gamma}_i (1 + \gamma_0 + n_t \gamma_0))^2} \right)^2 \]

(\text{A.7})

where \( \frac{\partial \Psi}{\partial \hat{\gamma}_i} \frac{\partial \hat{\gamma}_i}{\partial n_t} = \frac{\log_2 e}{1 + \hat{\gamma}_i} \frac{\gamma_0^2 (\gamma_0 + 1 + n_t \gamma_0)^2}{(1 + \hat{\gamma}_i (1 + \gamma_0 + n_t \gamma_0))^2} \leq 0. \]

Therefore, \( \Psi(n_t) \) is concave over \( n_t \). Then the function \( \Phi(n_t) = (n - n_t) \Psi(n_t) \) is also concave with respect to \( n_t \). Because

\[ \frac{\partial^2 \Phi}{\partial n_t^2} = - \frac{\partial \Psi}{\partial n_t} - \frac{\partial \Psi}{\partial n_t} + (n - n_t) \frac{\partial^2 \Psi}{\partial n_t^2} < 0. \]

(\text{A.8})

The function \( -\Phi(n_t) \) is convex over \( n_t \), and thus, the function \( e^{-\Phi(n_t)} \) is also convex respect to \( n_t \). To show the convexity of \( \ln \mathbb{E} \left[ e^{-\Phi(n_t)} \right] \), Lemma 1 is proposed.

ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China under grant 61941106.

APPENDIX

A Proof of Theorem 2

Firstly, we fix \( \epsilon \) and define the following function with respect to \( m, \gamma_0 \) and \( n_t \) in (A.1).
Lemma 1. Denote the function
\[ h(x) = \ln \left( \sum_{i=1}^{n} e^{a_i x} \right), \quad (A.9) \]
h(x) is convex with respect to x.

Proof. The first order and second order derivatives of h(x) are:
\[ h'(x) = \frac{\sum_{i=1}^{n} a_i e^{a_i x}}{\sum_{i=1}^{n} e^{a_i x}} \geq 0, \quad (A.10) \]
\[ h''(x) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j e^{a_i + a_j x} - \sum_{i=1}^{n} a_i a_j e^{a_i + a_j x}}{\left( \sum_{i=1}^{n} e^{a_i x} \right)^2} \geq 0. \quad (A.12) \]

Then consider a compound function h(g(x)) which is convex if g(x) is convex with respect to x. Hence - ln E \[ e^{-\theta h(n_t)} \] is concave with respect to n_t and therefore L(n_t) is concave over n_t. From this analysis result, there must exist an optimal n_t that maximizes the function L(n_t).

When taking the decoding error probability into consideration, the monotonicity of C_E(\theta) is the same as L(n_t). More specifically, L(n_t) can be viewed as a compound function consists of L_1(x) = -\frac{1}{2} \ln(x) and L_2(n_t) = E \{ e^{-\theta n_t d(R(\gamma_t)^{x})} \}. Now that the function L(n_t) = L_1(L_2(n_t)) is concave and the following condition must hold
\[ \frac{\partial^2 L}{\partial n_t^2} = \frac{\partial^2 L_1}{\partial n_t^2} \left( \frac{\partial L_2}{\partial n_t} \right)^2 + \frac{\partial L_1}{\partial L_2} \frac{\partial^2 L_2}{\partial n_t^2} \leq 0. \quad (A.13) \]
\[ = \frac{1}{L_2(n_t)} \left\{ \frac{1}{L_2(n_t)} \left( \frac{\partial L_2}{\partial n_t} \right)^2 - \frac{\partial^2 L_2}{\partial n_t^2} \right\} \leq 0. \quad (A.14) \]

For notation convenience, denote
\[ d(n_t) = (1 - \varepsilon)L_2(n_t) + \varepsilon, \quad (A.15) \]
and observe the expression in (20), T(n_t) can be expressed as
\[ T(n_t) = L_1(d(n_t)). \quad (A.16) \]
The second order derivative of T(n_t)
\[ \frac{\partial^2 T}{\partial n_t^2} = \frac{\partial^2 L_1}{\partial d^2} \left( \frac{\partial d}{\partial n_t} \right)^2 + \frac{\partial L_1}{\partial d} \frac{\partial^2 d}{\partial n_t^2} \]
\[ = \frac{1 - \varepsilon}{L_2(n_t) + \varepsilon} \left\{ \frac{1 - \varepsilon}{L_2(n_t) + \varepsilon} \left( \frac{\partial L_2}{\partial n_t} \right)^2 - \frac{\partial^2 L_2}{\partial n_t^2} \right\} \]
\[ \leq 0. \quad (A.18) \]

Based on this, we conclude that the effective capacity in (20) is concave with respect to the pilot length n_t.

B Proof of Theorem 3
Due to \gamma_0 = \frac{p}{2} and \sigma^2 is a constant, we will substitute p by \gamma_0 for notation simplicity. The first order partial derivative of \Psi with respect to \gamma_0 is as follows
\[ \frac{\partial \Psi}{\partial \gamma_0} = \frac{\partial \Psi}{\partial \gamma_t} \frac{\partial \gamma_t}{\partial \gamma_0} \quad (B.1) \]
\[ = \log_2 e \frac{n_t \gamma_0 (n_t \gamma_0 + \gamma_0 + 2)}{1 + \gamma_t (1 + \gamma_0 + n_t \gamma_0)^2} \geq 0, \quad (B.2) \]
and the second order partial derivative is shown in (B.3)-(B.6).
where the last inequality in (B.6) is due to the assumption \( \hat{\gamma}_i > -3dB \).

Based on the above analysis, \( \Psi(p, n_t) \) is concave and monotonically increasing with respect to \( p \), we can affirm that the function \( L(p) \) is concave and monotonically increasing with respect to \( p \). Then we apply the method adopted in the proof of Theorem 2 and can prove that the effective capacity is also concave and monotonically increasing with respect to \( p \).

### C Proof of Theorem 4

The first order partial derivative of \( L(m, \gamma_0, n_t) \) with respect to \( m \) is

\[
\frac{\partial L}{\partial m} = \sqrt{mn_d Q^{-1}(\varepsilon)} \log_2 e - m \varepsilon \log_2 \left\{ E \left\{ e^{-\frac{\Delta m \Delta d \log_2 (1 + \hat{\gamma}_i)}{m \varepsilon}} \right\} \right\},
\]

and the second order partial derivative of \( L(m, p, n_t) \) with respect to \( m \) is

\[
\frac{\partial^2 L}{\partial m^2} = -\frac{1}{4} \sqrt{\frac{n_d}{m^3}} Q^{-1}(\varepsilon) \log_2 e < 0.
\]

Given the other parameters, \( L(m, \gamma_0, n_t) \) is concave and is monotonically decreasing with respect \( m \), hence the effective capacity \( C_E(\theta) \) is monotonically increasing in \( m \).

### D Proof of Theorem 5

Substituting average received SNR \( G \) in (5) to (31), we can obtain equation (C.3). According to the definition of \( \Gamma(m, p, \alpha) \), we can see that the first term of \( \Gamma(m, p, \alpha) \) doesn’t alter the convexity of \( \Gamma(m, p, \alpha) \) over \( \alpha \). Then we need to determine the convexity of the latter term. The latter term can be viewed as a compound function, where the outer function is negative \( \log \) function and it is a convex function and is monotonically decreasing. The inner function can be simplified as a fractional function as follows

\[
O(\alpha) = \frac{-\theta' n^2 \gamma_0^2 \alpha^2 + (\theta' n^2 \gamma_0^2 - n \gamma_0^2 + \gamma_0)n\alpha + 1 + \gamma_0}{1 + \gamma_0 + n \gamma_0 \alpha},
\]

the second order derivatives of \( O(\alpha) \) can be derived as follows
Due to the assumption, the inequality $\theta' n_d - 1 > 1$ holds and thus $O''(\alpha) < 0$ holds. Therefore, $O(\alpha)$ is a concave function with respect to $\alpha$. According to the property of the compound function, the function $\Gamma(\alpha)$ is also concave with respect to $\alpha$.

Due to the fact that the constant terms corresponding with $\varepsilon$ in (28) do not alter the convexity and monotonicity. Theorem 5 is proved.

\begin{align}
\frac{\partial^2 \Gamma}{\partial \gamma_0^2} &= \frac{\partial^2 \Gamma}{\partial G^2} \left( \frac{\partial G}{\partial \gamma_0} \right)^2 + \frac{\partial \Gamma}{\partial G} \frac{\partial^2 G}{\partial \gamma_0^2} \\
&= \frac{m}{\theta} \left( \frac{(\theta' n_d - 1)^2}{(\theta' n_d - 1)G + 1} \right)^2 \left( \frac{[n_t \gamma_0 (n_t \gamma_0 + \gamma_0 + 2)]^2}{(1 + \gamma_0 + n_t \gamma_0)^2} \right)^2 + \frac{m}{\theta} \left( \frac{(\theta' n_d - 1)}{(\theta' n_d - 1)G + 1} \right)^2 \frac{2n_t}{(\theta' n_d - 1)G + 1} \left( \frac{\theta' n_d - 1}{\theta} \right) \\
&= \frac{m}{\theta} \left( \frac{(\theta' n_d - 1)n_t}{(1 + \gamma_0 + n_t \gamma_0)^3(\theta' n_d - 1)G + 1} \right) \times \left( \frac{n_t \gamma_0 + \gamma_0 + 2}{G + \frac{1}{\theta' n_d - 1}} \right)^4 \left( \frac{(\theta' n_d - 1)G + 1}{2} \right) \\
&= \frac{m}{\theta} \left( \frac{(\theta' n_d - 1)n_t}{(1 + \gamma_0 + n_t \gamma_0)^3(\theta' n_d - 1)G + 1} \right) \times \left( \frac{n_t \gamma_0 + \gamma_0 + 2}{G + \frac{1}{\theta' n_d - 1}} \right)^4 \left( \frac{G}{2} \right) \\
&= \frac{m}{\theta} \left( \frac{(\theta' n_d - 1)n_t}{(1 + \gamma_0 + n_t \gamma_0)^3(\theta' n_d - 1)G + 1} \right) \times \left( \frac{16G}{G + 1} \right)^2 \frac{G}{2} < 0. 
\end{align}

**E Proof of Theorem 6**

Here we also substitute $\rho$ by $\gamma_0$ for simplicity. The second order partial derivative of $\Gamma(m, \gamma_0, \alpha)$ with respect to $\gamma_0$ is shown in (E.1)-(E.4), where the first inequality in (E.6) is derived from the fact that $\theta' n_d - 1 > 1$ and the second inequality is derived due to $G > -3dB$.

Similarly, due to the fact that the constant terms corresponding with $\varepsilon$ in (28) do not alter the convexity and monotonicity. Theorem 6 is proved.

\begin{align}
O''(\alpha) &= -\frac{2(1 + \gamma_0)(n_t \gamma_0)^2 [\theta' n_d - 1] \gamma_0 + \theta'(1 + \gamma_0)]}{(1 + \gamma_0 + n_t \gamma_0)^3}.
\end{align}

(D.2)

References

[1] M. Bennis, M. Debbah, et al., “Ultrareliable and low-latency wireless communication: Tail, risk, and scale,” Proceedings of the IEEE, 2018, vol. 106, no. 10, pp. 1834–1853.

[2] P. Popovski, C. Stefanovic, et al., “Wireless access in ultrareliable low-latency communication (urlc),” IEEE Transactions on Communications, 2019, vol. 67, no. 8, pp. 5783–5801.

[3] H. Chen, R. Abbas, et al., “Ultra-reliable low latency cellular networks: Use cases, challenges and approaches,” IEEE Communications Magazine, 2018, vol. 56, no. 12, pp. 119–125.

[4] C. Li, C. Li, et al., “5g-based systems design for tactile internet,” Proceedings of the IEEE, 2019, vol. 107, no. 2, pp. 307–324.

[5] G. Durisi, T. Koch, et al., “Toward massive, ultrareliable, and low-latency wireless communication with short packets,” Proceedings of the IEEE, 2016, vol. 104, no. 9, pp. 1711–1726.

[6] Y. Polyanskiy, H. V. Poor, et al., “Channel coding rate in the finite blocklength regime,” IEEE Transactions on Information Theory, 2010, vol. 56, no. 5, pp. 2307–2359.

[7] H. Ji, S. Park, et al., “Ultra-reliable and low-latency communications in 5g downlink: Physical layer aspects,” IEEE Wireless Communications, 2018, vol. 25, no. 3, pp. 124–130.

[8] W. Yang, G. Durisi, et al., “Quasi-static multiple-antenna fading channels at finite blocklength,” IEEE Transactions on Information Theory, 2014, vol. 60, no. 7, pp. 4232–4265.

[9] G. Durisi, T. Koch, et al., “Short-packet communications over multiple-antenna rayleigh-fading channels,” IEEE Transactions on Communications, 2016, vol. 64, no. 2, pp. 618–629.

[10] S. Xu, T. Chang, et al., “Energy-efficient packet scheduling with finite blocklength codes: Convexity analysis and efficient algorithms,” IEEE Transactions on Wireless Communications, vol. 15, no. 8, pp. 5527–5540.

[11] X. Sun, S. Yan, et al., “Short-packet downlink transmission with non-orthogonal multiple access,” IEEE Transactions on Wireless Communications, 2018, vol. 17, no. 7, pp. 4550–4564.

[12] Y. Hu, M. Serror, et al., “Finite blocklength performance
of cooperative multi-terminal wireless industrial networks,” *IEEE Transactions on Vehicular Technology*, 2018, vol. 67, no. 7, pp. 5778–5792.

[13] B. Makki, T. Svensson, et al., “Finite block-length analysis of the incremental redundancy harq,” *IEEE Wireless Communications Letter*, 2014, vol. 3, no. 5, pp. 529–532.

[14] M. Mousaei and B. Smida, “Optimizing pilot overhead for ultra-reliable short-packet transmission,” in 2017 *IEEE International Conference on Communications (ICC)*, 2017.

[15] Y. Zhu, Y. Hu, et al., “Throughput maximization of low-latency communication with imperfect csi in finite blocklength regime,” in 2019 *IEEE Wireless Communications and Networking Conference (WCNC)*, 2019.

[16] J. Cao, X. Zhu, et al., “Joint block length and pilot length optimization for urllc in the finite block length regime,” in 2019 *IEEE Global Commun. Conf. (GLOBECOM)*, 2019.

[17] J. Cheng, C. Shen, et al., “Robust urllc packet scheduling of ofdm systems,” in 2020 *IEEE Wireless Communications and Networking Conference (WCNC)*, 2020.

[18] H. Ren, C. Pan, et al., “Joint pilot and payload power allocation for massive-mimo-enabled urllc iiot networks,” *IEEE Journal on Selected Areas in Communications*, 2020, vol. 38, no. 5, pp. 816–830.

[19] D. Wu and R. Negi, “Effective capacity: A wireless link model for support of quality of service,” *IEEE Transactions on Wireless Communications*, 2002, vol. 2, no. 4, pp. 630–643.

[20] M. Amjad, L. Musavian, et al., “Effective capacity in wireless networks: A comprehensive review,” *IEEE Communications Surveys & Tutorials*, 2019, vol. 21, no. 4, pp. 3007–3038.

[21] L. Zhang, Y. Yang, et al., “Effective capacity in cognitive radio networks with relay and primary user emulator,” *China Communications*, 2019, vol. 16, no. 11, pp. 130–145.

[22] H. Al-Zubaidy, J. Liebeheer, et al., “Network-layer performance analysis of multihop fading channels,” *IEEE/ACM Transactions on Networking*, 2016, vol. 24, no. 1, pp. 204–217.

[23] M. C. Gursoy, “Throughput analysis of buffer-constrained wireless systems in the finite blocklength regime,” *EURASIP Journal on Wireless Communications and Networking*, 2013.

[24] Y. Hu, O. Mustafa, et al., “Optimal power allocation for qos-constrained downlink multi-user networks in the finite blocklength regime,” *IEEE Transactions on Wireless Communications*, 2018, vol. 17, no. 9, pp. 5827–5840.

[25] M. Shehab, H. Alves, et al., “Effective capacity and power allocation for machine-type communication,” *IEEE Transactions on Vehicular Technology*, 2019, vol. 68, no. 4, pp. 4098–4102.

[26] J. Choi, “An effective capacity-based approach to multichannel low-latency wireless communications,” *IEEE Transactions on Communications*, 2019, vol. 67, no. 3, pp. 2476–2486.

[27] D. Qiao, M. C. Gursoy, et al., “Throughput-delay tradeoffs with finite blocklength coding over multiple coherence blocks,” *IEEE Transactions on Communications*, 2019, vol. 67, no. 8, pp. 5892–5904.

[28] C. Xiao, J. Zeng, et al., “Downlink mimo-noma for ultra-reliable low-latency communications,” *IEEE Journal on Selected Areas in Communications*, 2019, vol. 37, no. 4, pp. 780–794.

[29] C. Xiao, J. Zeng, et al., “Delay guarantee and effective capacity of downlink noma fading channels,” *IEEE Journal of Selected Topics in Signal Processing*, 2019, vol. 13, no. 3, pp. 508–523.

[30] S. Schiessl, H. Al-Zubaidy, et al., “Delay performance of wireless communications with imperfect csi and finite-length coding,” *IEEE Transactions on Communications*, 2018, vol. 66, no. 12, pp. 6527–6541.

[31] S. Schiessl, J. Gross, et al., “Delay performance of the multiuser mimo downlink under imperfect csi and finite-length coding,” *IEEE Journal on Selected Areas in Communications*, 2019, vol. 37, no. 4, pp. 765–779.

[32] G. Caire, N. Jingd, et al., “Multiuser mimo achievable rates with downlink training and channel state feedback,” *IEEE Transactions on Information Theory*, 2010, vol. 56, no. 6, pp. 2845–2866.

[33] B. Hassibi and B. M. Hochwald, “How much training is needed in multiple-antenna wireless links?” *IEEE Transactions on Information Theory*, 2003, vol. 49, no. 4, pp. 951–963.

[34] Y. Polyanskiy, H. V. Poor, et al., “Dispersion of gaussian channels,” in 2009 *IEEE International Symposium on Information Theory (ISIT)*, pp. 2204–2208, 2009.

[35] S. Schiessl, J. Gross, et al., “Delay analysis for wireless fading channels with finite blocklength channel coding,” in *Proc. ACM MSWiM*, 2015.

[36] C. Chiccoli, S. Lorenzutta, et al., “Recent results for generalized exponential integrals,” *Computers & Mathematics with Applications*, 1990, vol. 19, no. 5, pp. 21–29.

**Biographies**

**Peng Hongsen** received the B.S. degree in communications engineering from Xidian University, in 2018. Now he is pursuing Ph.D degree with the Department of Electronic Engineering, Shanghai Jiao Tong University (SJTU). His current research interests include ultra reliable and low latency communication (URLLC), finite blocklength channel coding, queueing theory, and deep reinforcement learning.

**Tao Meixia** received the B.S. degree in electronic engineering from Fudan University, Shanghai, China, in 1999, and the Ph.D degree in electrical and electronic engineering from Hong Kong University of Science and Technology, Hong Kong, China, in 2003. She is currently a Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, China. Her current research interests include wireless caching, edge computing, physical layer multicasting, and resource allocation.
Dr. Tao was the recipient of the 2019 IEEE Marconi Prize Paper Award, the 2013 IEEE Heinrich Hertz Award for Best Communications Letters, the IEEE/CIC International Conference on Communications in China (ICCC) 2015 Best Paper Award, and the International Conference on Wireless Communications and Signal Processing (WCSP) 2012 Best Paper Award. She also received the 2009 IEEE ComSoc Asia-Pacific Outstanding Young Researcher award.

Dr. Tao is serving as Editor-at-Large of the *IEEE Open Journal of the Communications Society*. She served as a member of the Executive Editorial Committee of the *IEEE Transactions on Wireless Communications* during 2015-2019. She was also on the Editorial Board of several other journals as Editor or Guest Editor, including the *IEEE Transactions on Communications* and *IEEE Journal on Selected Areas in Communications*. She served as Symposium Oversight Chair of IEEE ICC 2019, Symposium Co-Chair of IEEE GLOBECOM 2018, the TPC chair of IEEE/CIC ICCC 2014 and Symposium Co-Chair of IEEE ICC 2015.