Optimal Power Allocation for QoS-Constrained Downlink Multi-User Networks in the Finite Blocklength Regime

Yulin Hu, Mustafa Ozmen, M. Cenk Gursoy and Anke Schmeink

Abstract

In this paper, we consider a downlink multiuser network operating with finite blocklength codes under statistical quality of service (QoS) constraints. Optimal power allocation algorithms are studied to maximize the normalized sum throughput under QoS constraints, while considering different types of data arrivals namely, constant-rate, Markov, and Markov-modulated Poisson arrivals. We first determine the finite blocklength (FBL) throughput formulations and subsequently state optimization problems. We show the convexity of the power allocation problem under certain conditions and propose optimal algorithms (for scenarios with different data arrivals). In addition, the FBL performance of equal power allocation and a sub-optimal power allocation algorithm is discussed. Via numerical analysis, we demonstrate the performance improvements with the optimal power allocation. In addition, we provide interesting insights on the system behavior by characterizing the impact of the error probability, the QoS-exponent, blocklength, the number of users and the source burstiness on the performance.

Index Terms

effective capacity, downlink, finite blocklength, Markovian sources, power allocation, QoS.

I. Introduction

Low-latency and high reliability have become two major concerns in the design of future wireless networks. In particular, researchers and designers of next-generation wireless networks are increasingly interested in having wireless links support delay-sensitive data traffic generated in applications such as haptic feedback in virtual and augmented reality, E-health, autonomous driving, industrial control applications and cyber-physical systems. In the design of new cellular networking architectures, e.g., 5G New Radio, this concept is related to ultra-reliable low latency communication (URLLC) [1], [2]. Similar feature topics of low latency applications are also widely discussed in the context of Internet of Things [3] and industrial wireless networks [4] for the future industry design, i.e., Industry 4.0. The common characteristic of these discussed scenarios is that the network serves multiple users/terminals while the coding blocklengths for wireless transmission are quite short due to the low latency constraint.

Yet, as another delay-sensitive scenario, mobile multimedia traffic has experienced an exponential growth in recent years. With this, providing certain quality-of-service (QoS) guarantees to users has also become a critical consideration in the design of future wireless networks. Generally, it is expected that constraints on delay, packet error probability and buffer overflow probabilities at various levels need to be satisfied for multiple users. For such networks operating under low latency requirements with finite blocklength codes, resource management is a challenging task even if the system supports only one class of traffic. The problem becomes more difficult and challenging when users have different levels of QoS requirements and when the source traffic and channel conditions randomly vary over time. Existing works on resource allocation under QoS constraints mainly consider constant data arrival rates. For instance, optimal power allocation schemes are proposed to satisfy QoS requirements for a two-hop wireless relay network in [5], and a downlink multi-user network in [6]. Energy efficient optimizations are studied in virtualized small cell networks [7] and industrial networks [8] under delay constraints. An energy-efficient design is proposed in [9] under the specific statistical QoS guarantees of a multi-user network. In [10], an optimal resource allocation algorithm is proposed for a QoS-constrained device-to-device (D2D) communication network. In addition, a sub-optimal power control policy is proposed in [11] for non-orthogonal multiple access (NOMA) networks with QoS constraints. There are also several related studies in different contexts that consider random data arrivals. Optimal energy control is addressed in [12] for energy-constrained wireless networks. Throughput and energy efficiency are studied in [13] in the presence of randomly arriving data and statistical queuing constraints. A QoS-driven power control policy is proposed in [14] for fading multiple-access channels. However, all of the above studies on resource allocation in QoS-constrained networks (considering either constant or random arrivals) are performed under the ideal assumption of communicating arbitrarily reliably at Shannon's channel capacity, i.e., codewords are assumed to be infinitely long.

On the other hand, it is more accurate to incorporate finite blocklength coding assumptions into the analysis when low-latency applications are considered. In such finite blocklength (FBL) coding regime, the data transmission is no longer arbitrarily reliable. Especially when the blocklength is short, the error probability becomes significant even if the rate is selected below the Shannon limit. Taking this into account, an accurate (normal) approximation of the achievable coding rate under the finite blocklength assumption for an additive white Gaussian noise (AWGN) channel was derived in [15] for a single-hop transmission system. Subsequently, the initial work for AWGN channels was extended to Gilbert-Elliott channels [16] as well as quasi-static fading channels [17]–[19], QoS-constrained networks [20], [21] and relaying networks [22]–[25]. In [26], the achievable FBL coding rate (physical-layer performance) is derived for a single user model with a power controller (namely truncated channel inversion) under a long-term power constraint. However, power allocation in QoS-constrained multi-user networks has not been addressed in the FBL regime, and especially the link-layer performance has not been formulated and optimized for either constant-rate source or random data source scenarios. In fact, an FBL code is a double-edged sword for QoS-constrained multi-user networks. More specifically, a short bocklength generally leads to a flexible departure process (improving the queuing performance) but also a relatively high error probability (degrading the queuing performance). Hence, it is interesting and challenging to design power allocation policies that maximize the QoS-constrained performance of multi-user networks in the FBL regime.

In this paper, we study the optimal power allocation for a QoS-constrained multi-user URLLC network. We initially consider constant-rate sources, and then take into account the stochastic nature of information flows and investigate the effect of the randomness and burstiness of the source traffic on the power allocation in a downlink multi-user wireless network operating with FBL codes. Specifically, we consider Markovian source models (namely discrete-time Markov, Markov fluid, and both discrete-time and continuous-time Markov modulated Poisson processes (MMPP)) and determine the optimal power allocation policies. The key contributions of this paper can be further detailed as follows:

- The QoS-constrained FBL performance metrics are formulated for both constant-rate and four different types of random data arrivals. In particular, the normalized throughput of each user is derived and proved to be conditionally concave in the transmit power.
- For all data arrival models, we state the power allocation problem that maximizes the normalized sum throughput by optimally allocating the power over users and over time. We prove the convexity of the optimization problem and propose algorithms to solve them. In other words, an analytical framework is provided to study the optimal power allocation in downlink wireless transmissions with FBL codes in the presence of random data arrivals and statistical queueing constraints.
- We further study two additional approaches. Specifically, one is equal power allocation while the other one is a proposed sub-optimal power allocation algorithm. In particular, the sub-optimal algorithm is developed to maximize the expected throughput by optimally allocating power among users within a frame (i.e., without considering power allocation over time).
- Via numerical analysis, we demonstrate the performance advantages of the proposed optimal power allocation algorithm. In addition, we provide characterizations for the impact of the error probability, QoS-exponent, coding blocklength, the number of users and the source randomness on the performance.

The remainder of the paper is organized as follows: In Section II, we describe the system model and briefly provide the background on the FBL regime and statistical queuing constraints. We study optimal power allocation under constant arrivals to maximize the FBL throughput in Section III. The optimization problem is extended to random arrival scenarios in Section IV, where optimal power allocation algorithm are proposed for four different source models. In addition to the optimal algorithm, the equal power allocation algorithm and a proposed sub-optimal power allocation algorithms are discussed in In Section V. We provide our numerical results in Section VI and finally conclude the paper in Section VII. It should be mentioned that constant data arrival part of this work has been published as a conference version in [27].

II. Preliminaries

In this section,the system model is first described. Subsequently, we briefly provide the background on FBL regime and statistical queuing constraints.

A. System Description

We consider a downlink broadcasting scenario where a transmitter (e.g., an access point or a base station) sends data packets to N users. As shown in Fig. 1, the transmitter has N buffers corresponding to the N users, while each buffer has an independent data source. The entire system operates in a slotted fashion where time is divided into frames of length M symbols. In each frame, the transmitter collects N packets from these N buffers and sends packets to the corresponding users in different slots, as shown in Fig. 2. In particular, a frame has N slots for orthogonal transmissions to N users, and each slot has a length of m symbols, i.e., $Nm = M$. We consider a practical data arrival model, in which data arrivals at the transmitter are modeled as constant-rate or Markovian processes (as detailed in Section IV-A). In addition, data transmission to the users is subject to certain QoS constraints in the form of limitations on the target error probability and queueing delay (which is parametrized by the QoS exponent). And users can have different levels of QoS requirements. In particular, the target error probability and QoS factor of user i are denoted by ε_i and θ_i, respectively.

Channels are assumed to experience quasi-static fading, and therefore the channel fading remains the same within each frame and vary independently from one frame to the next. The instantaneous channel state information (CSI) of each link is assumed to be available at the transmitter. Denote the set of instantaneous channel gains by $\mathbf{z} = \{z_i, i = 1, ..., N\}$ where z_i is

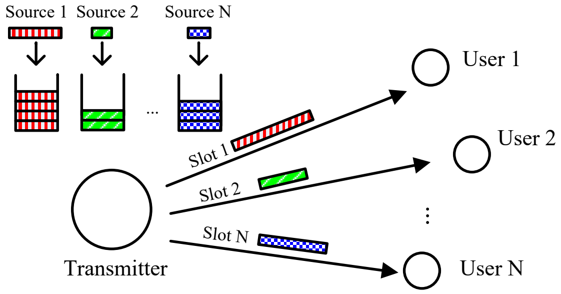


Fig. 1. System model of the considered network.

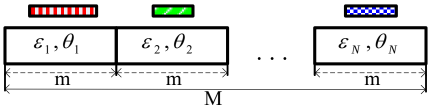


Fig. 2. Frame structure of the considered network.

the gain of the channel from the transmitter to user i. Note that the channel gains are time-varying, and we denote the joint probability density function (PDF) of $\mathbf{z}$ by $f_{\mathrm{PDF}}(\mathbf{z})$. Then, the signal-to-noise ratio (SNR) of the received signal at user i is given by $\gamma_i = \frac{p_i z_i}{\sigma^2}$, where σ^2 is the noise variance and p_i is the power used for transmission from the transmitter to user i. Moreover, the (long term) average power constraint at the transmitter is denoted by p_{ave}, i.e., $\mathbb{E}\{\sum_{i=1}^{N} mp_i\} \leq Mp_{\mathrm{ave}}$.

B. Finite blocklength codes

In [15], the authors analyzed the performance in the FBL regime by applying the normal approximation. In comparison to the Shannon capacity bound, the finite blocklength model is more accurate when the blocklength is finite/short. In addition, the third-order term in the normal approximation for the AWGN channel is further addressed in [28]. For an AWGN channel, the coding rate r (in bits per channel use) with error probability $0 < \varepsilon < 1$, SNR γ, and blocklength m is shown to have the following asymptotic expression [28]:

$$r = \mathcal{R}\left(\gamma, \varepsilon, m\right) \approx \mathcal{C}\left(\gamma\right) - \sqrt{\frac{V\left(\gamma\right)}{m}} Q^{-1}\left(\varepsilon\right) + \frac{\log m}{m}, \tag{1}$$

where $Q\left(x\right) = \int_x^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$ is the Gaussian Q-function.

Form (1), the (block) error probability can be expressed as:

$$\varepsilon = \mathcal{P}\left(\gamma, r, m\right) \approx Q\left(\frac{\mathcal{C}\left(\gamma\right) + \frac{\log m}{m} - r}{\sqrt{V(\gamma)/m}}\right). \tag{2}$$

In this work, we consider a QoS-constrained downlink network operating over quasi-static fading channels. Note that in the quasi-static model, the channel varies from one frame to the next but is static within each frame. Hence, in the following sections we apply the above model in our study of the performance of the considered network. As these approximations have been shown to be accurate for a sufficiently large value of m [15], for simplicity we will employ them as the rate and error expressions in our analysis. It should be mentioned that the achievable FBL coding rate in a quasi-static fading channel is derived in [26] for a single-user model under a long-term power constraint with a simple power control policy, namely truncated channel inversion. This policy performs channel inversion when the channel gain is higher than a threshold, otherwise, it turns off the transmission. The decisions/results obtained with this simple power control policy are different from the power allocation in our work, which maximize the QoS-constrained throughput of multiple users with potentially different QoS requirements. Therefore, the proposed power allocation results in a different performance level in a different setting (with QoS constraints) in comparison to the model in [26] with the truncated channel inversion.

C. Statistical queuing constraints

Throughout this paper, we assume that the transmissions to all users are performed under queuing constraints, which require the buffer overflow probabilities to decay exponentially fast [29]. Let us denote Q as the stationary queue length and θ as the decay rate of the tail of the distribution of the queue length Q. Then, the probability that the queue length Q exceeds a threshold q satisfies

$$P\left(Q \geq q\right) \approx \varsigma e^{-\theta q}, \tag{3}$$

where ς is probability of non-empty buffer. In addition, θ is called the QoS exponent, and is defined in [30] as

$$\lim_{q \to \infty} \frac{\log P\left(Q \geq q\right)}{q} = -\theta. \tag{4}$$

Note that small and large θ correspond to relatively loose and strict QoS constraints, respectively. In other words, QoS exponent θ controls the exponential decay rate of the buffer overflow probability. More specifically, larger θ indicates stricter limitation on the buffer overflow probability (or delay violation probability), leading to more stringent QoS constraints, and vice versa for small θ.

Following the queuing model in [29], [30], we denote by a (bits/frame) and c (bits/frame) the instantaneous arrival and departure rates at the buffer, respectively. According to the effective bandwidth and effective capacity formulations in [29], [30], in the presence of queuing constraints specified by the QoS exponent θ, the arrival process and departure process at the buffer should satisfy

$$\Lambda_a\left(\theta\right) + \Lambda_c\left(-\theta\right) = 0, \tag{5}$$

where $\Lambda_s\left(\theta\right) = \lim\limits_{t\to\infty} \frac{1}{t}\log\mathbb{E}\{e^{\theta\sum_{k=1}^{t} s_k}\}$ is introduced in [Eq. 6, 27], which denotes the asymptotic logarithmic moment generating function (LMGF) of the random process s_k.

Based on the LMGFs, the effective bandwidth is [30]

$$a_{\mathrm{E}}(\theta) = \frac{\Lambda_a\left(\theta\right)}{\theta}. \tag{6}$$

In addition, the effective capacity is given in [29] as

$$R_{\mathrm{E}}(\theta) = -\frac{\Lambda_c\left(-\theta\right)}{\theta}, \tag{7}$$

and characterizes the maximum constant arrival rate that can be supported by the link with a random service process while satisfying (3).

In this work, we adopt the effective capacity formulation to obtain the average throughput of the scenario with constant data arrivals. Additionally, we consider four types of random Markovian sources[1], namely 1) discrete Markov source, 2) Markov fluid source, 3) discrete-time Markov-modulated Poisson process (MMPP), 4) continuous-time MMPP. In particular, we consider two-state Markovian arrival models with ON and OFF states. In such a case, if P_{on} is the probability that the data source for a user is in the ON state with arrival rate a, the average arrival rate of the two-state Markovian source models simply becomes

$$\mu = P_{\mathrm{on}} a, \tag{8}$$

which is equal to the average departure rate when the queue is in steady state [34].

III. FBL THROUGHPUT OF MULTI-USER NETWORKS WITH CONSTANT ARRIVALS

In this section, we study the optimal power allocation for the downlink multiuser network with constant data arrivals to the buffers at the transmitter. First, we will develop the performance model. Subsequently, the optimization problem will be stated and solved.

A. FBL throughput model

With constant data arrivals, the FBL throughput is given by the effective capacity. If user i has a given (target) error probability ε_i and a given (target) QoS exponent θ_i, the effective capacity in the units of bits/frame is actually a function of the transmit power policy $\{p_i\}$, and is expressed as

$$R_{\mathrm{E},i} = -\frac{1}{\theta_i}\ln\left\{\mathbb{E}\left[e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i\right]\right\}, \tag{9}$$

where coding rate r_i, which is given in (1), is a function of the transmit power p_i and channel fading gains via the received SNR γ, and the expectation above is with respect to the distribution of the fading coefficients. First, we have the following Proposition.

Proposition 1. *In a system with target error probability $\varepsilon_i \geq 10^{-24}$ and blocklength $m \geq 100$, the coding rate r_i is increasing and concave in the transmit power p_i under the constraint $\gamma_i \geq 0$ dB.*

Proof: Let $A_i = Q^{-1}\left(\varepsilon_i\right)\sqrt{\frac{1}{m}}$. Then, according to (1), we have the first derivative of r_i with respect to the SNR γ_i given as (10) and the second order derivative given as (11) on the next page.

$$\frac{\partial r_i}{\partial \gamma_i} = \frac{\log e}{1+\gamma_i} - \frac{A_i \log e}{\sqrt{1-\frac{1}{(1+\gamma_i)^2}}}\frac{1}{\left(1+\gamma_i\right)^3}. \tag{10}$$

[1] The consideration of Markovian sources is motivated by the fact that certain types of delay-sensitive data traffic can be modeled as Markovian processes. For instance, voice traffic can be modeled as an ON/OFF Markov process, and variable bit-rate (VBR) video traffic is generally modeled as Markov-modulated processes [31]. Moreover, data arrival processes in machine-to-machine and automotive machine type communication can also be treated as Markovian processes [32], [33].

$$\begin{aligned}\frac{\partial^2 r_i}{\partial {\gamma_i}^2} &= -\frac{\log e}{(1+\gamma_i)^2}+\frac{A_i \log e}{\left(1-\frac{1}{(1+\gamma_i)^2}\right)^{\frac{3}{2}}}\frac{1}{(1+\gamma_i)^6}+\frac{A_i \log e}{\sqrt{\left(1-\frac{1}{(1+\gamma_i)^2}\right)}}\frac{3}{(1+\gamma_i)^4} \\ &= -\frac{\log e}{(1+\gamma_i)^2}+\frac{A_i \log e}{\left((1+\gamma_i)^2-1\right)^{\frac{3}{2}}}\frac{1}{(1+\gamma_i)^3}+\frac{A_i \log e}{\sqrt{(1+\gamma_i)^2-1}}\frac{3}{(1+\gamma_i)^3}=\frac{\log e}{(1+\gamma_i)^3}\left\{-(1+\gamma_i)+\frac{A_i}{({\gamma_i}^2+2\gamma_i)^{\frac{3}{2}}}+\frac{3A_i}{\sqrt{{\gamma_i}^2+2\gamma_i}}\right\}.\end{aligned} \tag{11}$$

In the following, we prove the proposition by showing $\frac{\partial^2 r_i}{\partial {\gamma_i}^2} \leq 0$. We distinguish the following two cases. First, if $\varepsilon_i \geq 0.5$, we quickly have $A_i \leq 0$. Then $\frac{\partial^2 r_i}{\partial {\gamma_i}^2} < 0$. In the other case $\varepsilon_i < 0.5$, $\frac{\partial^2 r_i}{\partial {\gamma_i}^2}$ is increasing in A_i and therefore decreasing in ε_i and m. For an extreme scenario where $m = 100, \varepsilon_i = 10^{-24}$, we have $A_i = 1.0199$. Then, $\frac{\partial^2 r_i}{\partial {\gamma_i}^2} \leq 0$ if $\phi(\gamma_i) \leq 0$, where $\phi(\gamma_i) = -(1+\gamma_i) + \frac{A_i}{({\gamma_i}^2+2\gamma_i)^{\frac{3}{2}}} + \frac{3A_i}{\sqrt{{\gamma_i}^2+2\gamma_i}}$. Obviously, $\phi(\gamma_i)$ is decreasing in γ_i for $A_i > 0$. In particular, we have $\phi(1) = -0.0372$ for $A_i = 1.0199$. Hence, in the second case $\phi(\gamma_i) < 0$ for $\gamma_i \geq 1 = 0$ dB. To sum up, the coding rate r_i is concave in the transmit power p_i under the constraint guaranteeing $\gamma_i \geq 0$ dB while the target error probability and blocklength are within practical interest, i.e., $m \geq 100, \varepsilon_i \geq 10^{-24}$. ■

We note that general formulations and conditions to establish the concavity of the coding rate are provided above in the proof of Proposition 1, and these can be used to establish the range of SNR values for given coding blocklength m and target error probability ϵ_i. Indeed, we can easily show that for more practical scenarios, the concavity holds for even lower SNR bounds: e.g., SNR intervals in which concavity is satisfied are *i.* $[-3 \text{ dB}, \infty]$ for $m = 100, \varepsilon_i = 10^{-10}$, *ii.* $[-6 \text{ dB}, \infty]$ for $m = 300, \varepsilon_i = 10^{-5}$, *iii.* $[-11 \text{ dB}, \infty]$ for $m = 300, \varepsilon_i = 10^{-1}$, *iv.* $[-13 \text{ dB}, \infty]$ for $m = 1000, \varepsilon_i = 10^{-1}$.

Based on the above statements, we have the following characterization for the effective capacity as a function of the transmit power.

Proposition 2. *With constant arrivals, the FBL throughput (effective capacity) $R_{\mathrm{E},i}$ is concave in the transmit power p_i under the SNR constraint $\gamma_i \geq 0$ dB.*

Proof: Note that for a given ε_i, $R_{\mathrm{E},i}$ is a function of r_i while r_i is determined by the received SNR γ_i (corresponding to p_i). Hence, the first and second order derivatives of $R_{\mathrm{E},i}$ with respect to p_i can be given as follows:

$$\frac{\partial R_{\mathrm{E},i}}{\partial p_i} = \frac{\partial R_{\mathrm{E},i}}{\partial r_i}\frac{\partial r_i}{\partial \gamma_i}\frac{\partial \gamma_i}{\partial p_i} = \frac{\partial R_{\mathrm{E},i}}{\partial r_i}\frac{\partial r_i}{\partial \gamma_i} \cdot \frac{z_i}{\sigma^2}, \tag{12}$$

$$\frac{\partial^2 R_{\mathrm{E},i}}{\partial {p_i}^2} = \left(\frac{z_i}{\sigma^2}\right)^2 \frac{\partial^2 R_{\mathrm{E},i}}{\partial {r_i}^2}\left(\frac{\partial r_i}{\partial \gamma_i}\right)^2 + \left(\frac{z_i}{\sigma^2}\right)^2 \frac{\partial R_{\mathrm{E},i}}{\partial r_i}\frac{\partial^2 r_i}{\partial {\gamma_i}^2}. \tag{13}$$

According to (9), $\frac{\partial R_{\mathrm{E},i}}{\partial r_i}$ and $\frac{\partial^2 R_{\mathrm{E},i}}{\partial {r_i}^2}$ can be obtained as follows. Note that $\mathbb{E}\left[e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i\right] = \int\limits_{z_1}\ldots\int\limits_{z_N}\left\{e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i\right\} f_{\mathrm{PDF}}$ (
Then, $\frac{\partial \mathbb{E}[e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i]}{\partial r_i} = m e^{-\theta_i m r_i}(1-\varepsilon_i) f_{\mathrm{PDF}}(\mathbf{z})$ holds. Therefore, we have (14) and (15) on the next page.

$$\begin{aligned}\frac{\partial R_{\mathrm{E},i}}{\partial r_i} &= -\frac{1}{\theta_i}\frac{\frac{\partial \mathbb{E}[e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i]}{\partial r_i}}{\mathbb{E}\left[e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i\right]} \\ &= \frac{m e^{-\theta_i m r_i}(1-\varepsilon_i) f_{\mathrm{PDF}}(\mathbf{z})}{\mathbb{E}\left[e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i\right]} \geq 0,\end{aligned} \tag{14}$$

$$\begin{aligned}\frac{\partial^2 R_{\mathrm{E},i}}{\partial {r_i}^2} &= m(1-\varepsilon_i) f_{\mathrm{PDF}}(\mathbf{z}) \frac{-\theta_i m e^{-\theta_i m r_i}\mathbb{E}\left[e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i\right] - \frac{\partial \mathbb{E}[e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i]}{\partial r_i} e^{-\theta_i m r_i}}{\mathbb{E}[e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i]^2} \\ &= -\theta_i m^2 e^{-\theta_i m r_i}(1-\varepsilon_i) f_{\mathrm{PDF}}(\mathbf{z}) \frac{\mathbb{E}\left[e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i\right] - e^{-\theta_i m r_i}(1-\varepsilon_i)^2 f_{\mathrm{PDF}}(\mathbf{z})}{\mathbb{E}[e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i]^2}.\end{aligned} \tag{15}$$

It is clear that $\frac{\partial R_{\mathrm{E},i}}{\partial r_i} \geq 0$. In addition, $\frac{\partial^2 R_{\mathrm{E},i}}{\partial {r_i}^2} \leq 0$ due to the fact that $\mathbb{E}\left[e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i\right] \geq \left(e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i\right) f_{\mathrm{PDF}}(\mathbf{z}) \geq e^{-\theta_i m r_i}(1-\varepsilon_i) f_{\mathrm{PDF}}(\mathbf{z}) \geq e^{-\theta_i m r_i}(1-\varepsilon_i)^2 f_{\mathrm{PDF}}(\mathbf{z})$. Combining this with Proposition 1, we have $\frac{\partial^2 r_i}{\partial p_i^2} \leq 0$ under the constraint guaranteeing $\gamma_i \geq 0$ dB. Hence, $R_{\mathrm{E},i}$ is concave in p_i when $\gamma_i \geq 0$ dB. ■

B. Problem statement

Recall that we consider a downlink multiuser URLLC network where users potentially have different QoS requirements, i.e., the QoS exponent θ_i and target error probability ε_i of transmissions for users $i = 1, ..., N$ are not necessarily the same. Our objective is to improve the normalized sum throughput, i.e., $R_{\mathrm{E,sum}} = \frac{1}{M}\sum_{i=1}^{N} R_{\mathrm{E},i}$, in bits/ch.use. Although users have different QoS requirements, each user requires a basic connection/transmission guarantee as long as the channel state is sufficiently good, i.e., $z_i \geq z_{\min} \geq \frac{\sigma^2}{N p_{\mathrm{ave}}}$. This requirement in terms of SNR arises from the condition that $\gamma_i = \frac{p_i z_i}{\sigma^2} \geq \gamma_{\mathrm{th},i} \geq 0$ dB, $i = 1, ..., N$. On the other hand, due to the randomness of the fading it is possible that $z_i \leq z_{\min}$. In such a case, we simply allocate zero power for this user in this frame. For example, if $z_i \leq \frac{\sigma^2}{N p_{\mathrm{ave}}}$, guaranteeing a 0 dB received SNR for user i costs more than the sum of average power of all N users, potentially leading to unfair resource allocation. Hence, it is reasonable to skip this user in the power allocation.

More importantly, we maximize the objective by optimally allocating power over frames (i.e., over time) and among users, while satisfying the average power constraint (averaged over time and users), i.e., $\mathbb{E}_{\mathbf{z}}\left\{\sum_{i=1}^{N} p_i\right\} \leq M p_{\mathrm{ave}}/m$, where $\mathbf{z} = \{z_i, i = 1, ..., N\}$. Hence, the optimal power allocation p_i^* for user i in a frame is decided by not only the instantaneous channel gain z_i but also the distribution of the channel gains of all users, i.e., z_i, for $i = 1, ..., N$.

Based on the above analysis, the problem of optimizing the power allocation among users and across frames in the downlink multiuser network with constant arrivals is stated as follows:

$$\begin{aligned} \max_{\mathbf{p}\in\mathbf{\Omega}} \;\; & R_{\mathrm{E,sum}} = \frac{1}{M}\sum_{i=1}^{N} R_{\mathrm{E},i} \\ s.t.: \;\; & \mathbb{E}_{\mathbf{z}}\left\{\sum_{i=1}^{N} p_i\right\} - \frac{M p_{\mathrm{ave}}}{m} \leq 0, \end{aligned} \tag{16}$$

where $\mathbf{p} = \{p_i, i = 1, ..., N\}$ and p_i is influenced by z_i and the joint probability density function (PDF) of $z_1, ..., z_N$. In addition, $\mathbf{\Omega} = \{\Omega_i\}^N$, where Ω_i is the admissible/feasible set of p_i, given by

$$\Omega_i = \begin{cases} p_i \geq \frac{\gamma_{\mathrm{th},i}}{\sigma^2 z_i}, & \text{if } z_i \geq z_{\min}, \\ p_i = 0, & \text{if } z_i < z_{\min}, \end{cases} \tag{17}$$

for $i = 1, \ldots, N$.

C. Optimal power allocation

To solve Problem (16), we first show its convexity in the following proposition.

Proposition 3. *Problem* (16) *is a convex optimization problem.*

Proof: According to (13), we have $\frac{\partial^2 R_{\mathrm{E},i}}{\partial^2 p_i} \leq 0$ for $\gamma_i = \frac{p_i z_i}{\sigma^2} \geq \gamma_{\mathrm{th},i} \geq 0$ dB.

Then, the Hessian matrix of the objective function in Problem (16) with respect to $\mathbf{p}$ is given by

$$\begin{pmatrix} \frac{1}{M}\frac{\partial^2 R_{\mathrm{E},1}}{\partial^2 p_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{M}\frac{\partial^2 R_{\mathrm{E},N}}{\partial^2 p_N} \end{pmatrix}, \tag{18}$$

which is negative semidefinite in the admissible set $\mathbf{\Omega}$. Hence, the objective function is concave. In addition, the first constraint (i.e., the average power constraint) is affine in $\{p_i\}$. Therefore, Problem (16) is a convex optimization problem. ■

In the following, we state the Lagrange dual function of the Problem (16). We introduce the Lagrange multiplier λ associated with the average power constraint. Then, the dual function is given by

$$L = \frac{1}{M}\sum_{i=1}^{N} R_{\mathrm{E},i} - \lambda\,\mathbb{E}_{\mathbf{z}}\left\{\sum_{i=1}^{N} p_i - \frac{M p_{\mathrm{ave}}}{m}\right\}. \tag{19}$$

By solving $\frac{\partial L}{\partial p_i(z)} = 0$, we can determine the dual optimal. Now, let us introduce g_i as (20) on the next page. With this, we can express the effective capacity as $R_{\mathrm{E},i} = -\frac{1}{\theta}\ln g_i$. Then, we can express the first order derivative of g_i with respect to p_i for a given channel realization $\mathbf{z}$ by (21) on the next page, where we define $\eta_i = \frac{\theta_i}{\ln 2}$ and $a_i = 1-(1+\gamma_i)^{-2} = 1-\left(1+\frac{p_i z_i}{\sigma^2}\right)^{-2}$.

Then, we have

$$\begin{aligned} \frac{\partial L}{\partial p_i(z)} &= \frac{1}{M}\frac{\partial R_{\mathrm{E},i}}{\partial g_i}\frac{\partial g_i}{\partial p_i} - \lambda \\ &= \varphi_i\left(1-\frac{A_i}{\sqrt{a_i}}(1-a_i)\right)(1-a_i)^{\frac{\eta_i m+1}{2}} e^{\eta_i m A_i\sqrt{a_i}} - \lambda = 0, \end{aligned} \tag{22}$$

$$
\begin{aligned}
g_i = e^{-\theta R_{\mathrm{E},i}} &= \mathbb{E}_{\mathbf{z}}\left\{e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i\right\} = \mathbb{E}_{\mathbf{z}}\left\{ e^{-\frac{\theta_i}{\ln 2} m\left[\ln\left(1+\frac{z_i p_i(\mathbf{z})}{\sigma^2}\right) - A_i \sqrt{\left(1-\frac{1}{\left(1+\frac{z_i p_i(\mathbf{z})}{\sigma^2}\right)^2}\right)} + \frac{\ln m}{m}\right]} (1-\varepsilon_i)+\varepsilon_i \right\} \\
&= \int\limits_{z_1}\cdots\int\limits_{z_N} \left\{ e^{-\frac{\theta_i}{\ln 2} m\left[\ln\left(1+\frac{z_i p_i(\mathbf{z})}{\sigma^2}\right) - A_i \sqrt{\left(1-\frac{1}{\left(1+\frac{z_i p_i(\mathbf{z})}{\sigma^2}\right)^2}\right)} + \frac{\ln m}{m}\right]} (1-\varepsilon_i)+\varepsilon_i \right\} f_{\mathrm{PDF}}(\mathbf{z}) \cdot d_{z_1}\cdots d_{z_N}.
\end{aligned} \tag{20}
$$

$$
\begin{aligned}
\frac{\partial g_i}{\partial p_i(\mathbf{z})} &= -\frac{\theta_i(1-\varepsilon_i)}{\ln 2} m \left(\frac{1}{1+\gamma_i} - \frac{A_i}{\sqrt{1-\frac{1}{(1+\gamma_i)^2}}}\frac{1}{(1+\gamma_i)^3}\right)\frac{z_i}{\sigma^2} e^{-\frac{\theta_i}{\ln 2} m\left[\ln(1+\gamma_i) - A_i\sqrt{1-\frac{1}{(1+\gamma_i)^2}} + \frac{\ln m}{m}\right]} \cdot f_{\mathrm{PDF}}(\mathbf{z}) \\
&= -\frac{z_i(1-\varepsilon_i)\eta_i m^{1-\eta_i}}{\sigma^2}\left(\frac{1}{1+\gamma_i} - \frac{A_i}{\sqrt{1-\frac{1}{(1+\gamma_i)^2}}}\frac{1}{(1+\gamma_i)^3}\right)(1+\gamma_i)^{-\eta_i m} e^{\eta_i m A_i \sqrt{1-\frac{1}{(1+\gamma_i)^2}}} \cdot f_{\mathrm{PDF}}(\mathbf{z}) \\
&= -\frac{z_i(1-\varepsilon_i)\eta_i m^{1-\eta_i}}{\sigma^2}\left(1-\frac{A_i}{\sqrt{a_i}}(1-a_i)\right)(1-a_i)^{\frac{\eta_i m+1}{2}} e^{\eta_i m A_i \sqrt{a_i}} \cdot f_{\mathrm{PDF}}(\mathbf{z}).
\end{aligned} \tag{21}
$$

where $\varphi_i = \frac{z_i(1-\varepsilon_i)\eta_i m^{1-\eta_i}}{g_i M \theta \sigma^2} f_{\mathrm{PDF}}(z_i)$.

By solving the (22) within the admissible set $\{p_i \in \Omega_i\}$, we can obtain the power solution λ^* and p_i^*. However, as seen in the above discussion, it is unlikely to obtain a closed-form expression for the optimal power allocation policy, as the solution of p_i^* and λ are generally interdependent on each other. On the other hand, the optimal power allocation can be determined via numerical computations. Therefore, we propose an algorithm described in Algorithm 1 below to obtain the optimal transmit power numerically. The key idea of the algorithm is to first initialize the value of λ, g_i and obtain the corresponding p_i according to (22). Subsequently, we update g_i based on the obtained p_i till g_i converges to g_i^o. Finally, we keep updating λ till (22) is satisfied.

Algorithm 1 : Optimal Power Allocation Algorithm.

```
Initialization
1) for user $i = 1, ..., N$
  a) if $z_i < z_{\min}$
  b)   then $p_i^* = 0$ and go to Step 1 for the next user;
  c)   else Given $\lambda$, $g_i$, determine $p_i$ according to (22).
  d)     According to (20), update $g_i$ based on $p_i$ and the PDF of z.
  e)     According to (22), update $p_i$ based on the updated $g_i$ in Step 1-d. If there is no solution in the admissible set, go to Step 2-b.
  f)     Check if $g_i$ converges to a constant:
  g)     if the gap between the updated $g_i$ and the previous one become relatively constant and small enough
  h)       then $g_i$ converges. We have $p_i^* = \max\{p_i, \frac{\gamma_{\mathrm{th},i}\sigma^2}{z_i}\}$, $\lambda^* = \lambda$ and converged $g_i^\circ = g_i$.
  i)       else return to Step 1-c.
         endif
     endif
   endfor
2) Check if the sum of the obtained $p_i^*$ satisfy the average power constraint.
  a) if not satisfied with equality
  b)   then update the value of $\lambda$ and return to Step 1;
  c)   else the optimal power allocation is obtained, including $\lambda^*$ and the converged $g_i^\circ$, $i = 1, ..., N$.
     endif
Instantaneous power allocation per frame
  a) According to (22), determine the optimal power $p_i^*$ for this frame based on the instantaneous z as well as the obtained $\lambda^*$ and $g_i^\circ$.
```

In particular, Algorithm 1 has two parts: the first part is the initialization part while the subsequent one is the instantaneous power allocation part. The optimal optimal power allocation policy is fully characterized by the optimal value λ^* and the converged g_i° in the initialization part. After the initialization, the optimal power can be determined by simply plugging the instantaneous channel gains $\mathbf{z}$ together with λ^* and g_i° into (22) and solving it, which has very low computational complexity.

IV. FBL THROUGHPUT OF DOWNLINK MULTIUSER NETWORKS WITH RANDOM ARRIVALS

In this section, we consider maximizing the average FBL throughput in downlink multiuser networks when data arrivals are modeled by Markovian processes.

A. Discrete Markov Source

In this subsection, we assume that the data intended for each user is generated by a two-state discrete Markov source. Let us define states 1 and 2 as the OFF and ON states, respectively. In the OFF state of the Markov source for user i, no data arrives from the source for user i, and the arrival rate is a_i (bits/frame) in the ON state. The state transition probability matrix of the Markov data arrival at the source for user i can be expressed as

$$G_i = \left(\begin{array}{cc} p_{i,11} & p_{i,12} \\ p_{i,21} & p_{i,22} \end{array} \right), \tag{23}$$

where $p_{i,11}$ and $p_{i,22}$ denote the probabilities that the source remains in the same state (more explicitly, OFF and ON states, respectively) in the next time block, and $p_{i,12}$ and $p_{i,21}$ are the probabilities that source will transition to a different state in the next time block. Using the properties of Markov processes, we can express the probability of the ON state as

$$P_{\text{on},i} = \frac{1 - p_{i,11}}{2 - p_{i,11} - p_{i,22}}. \tag{24}$$

Then, the average arrival rate of this ON-OFF Markov source for user i (in bits/symbol) is given by $\mu_i = P_{\text{on},i} a_i / M$.

According to [30], the LMGF of the arrival process of the data for user i is given by

$$\Lambda_{a_i}\left(\theta_i\right) = \ln\left(\frac{g_{\text{D-M},i}}{2}\right), \tag{25}$$

where

$$\begin{aligned} g_{\text{D-M},i} =& \sqrt{\left(p_{i,11}+p_{i,22}e^{a_i\theta_i}\right)^2 - 4\left(p_{i,11}+p_{i,22}-1\right)e^{a_i\theta_i}} \\ &+ p_{i,11}+p_{i,22}e^{a_i\theta_i}. \end{aligned} \tag{26}$$

Hence, the effective bandwidth is given by $a_{E,i} = \frac{\Lambda_{a_i}}{\theta}$.

For the departure process of the data sent to user i, the LMGF is given by

$$\Lambda_{c_i}\left(-\theta_i\right) = -\theta_i R_{\text{E},i}. \tag{27}$$

Plugging the characterizations $\Lambda_{a_i}\left(\theta\right)$ and $\Lambda_{c_i}\left(-\theta\right)$ into (5) and solving for a_i, we obtain

$$a_i = \frac{1}{\theta_i} \ln\left(\frac{e^{\theta_i R_{\text{E},i}}\left(e^{\theta_i R_{\text{E},i}} - p_{i,11}\right)}{1 - p_{i,11} - p_{i,22}\left(1 - e^{\theta_i R_{\text{E},i}}\right)}\right). \tag{28}$$

Then, we obtain the FBL throughput of user i as

$$\mu_i = \frac{P_{\text{on},i}}{M\theta_i} \ln\left(\frac{e^{\theta_i R_{\text{E},i}}\left(e^{\theta_i R_{\text{E},i}} - p_{i,11}\right)}{1 - p_{i,11} - p_{i,22} + p_{i,22} e^{\theta_i R_{\text{E}i,}}}\right). \tag{29}$$

We have the following proposition regarding the FBL throughput with discrete-time Markov data arrivals

Proposition 4. *With discrete-time Markov data arrivals, the FBL throughput rate μ_i is increasing and concave in the coding rate r_i.*

Proof: First, we show $\frac{\partial \mu_i}{\partial R_{\text{E},i}} > 0$. According to (29), we have

$$\begin{aligned} \frac{\partial \mu_i}{\partial R_{\text{E},i}} =& \frac{P_{\text{on}}}{M}\left(1 + \frac{e^{\theta_i R_{\text{E},i}}}{e^{\theta_i R_{\text{E},i}} - p_{i,11}} - \frac{p_{i,22} e^{\theta_i R_{\text{E},i}}}{1 - p_{i,11} - p_{i,22} + p_{i,22} e^{\theta_i R_{\text{E},i}}}\right) \\ =& \frac{P_{\text{on}}}{M}\left(1 + \frac{e^{\theta_i R_{\text{E},i}} - e^{\theta_i R_{\text{E},i}} p_{i,11} - e^{\theta_i R_{\text{E},i}} p_{i,22} + p_{i,11} p_{i,22} e^{\theta_i R_{\text{E},i}}}{\left(e^{\theta_i R_{\text{E},i}} - p_{i,11}\right)\left(1 - p_{i,11} - p_{i,22} + p_{i,22} e^{\theta_i R_{\text{E},i}}\right)}\right) \\ =& \frac{P_{\text{on}}}{M}\left(1 + \frac{e^{\theta_i R_{\text{E},i}}\left(1 - p_{i,11} - p_{i,22} + p_{i,11} p_{i,22}\right)}{\left(e^{\theta_i R_{\text{E},i}} - p_{i,11}\right)\left(1 - p_{i,11} - p_{i,22} + p_{i,22} e^{\theta_i R_{\text{E},i}}\right)}\right). \end{aligned} \tag{30}$$

Note that $p_{i,11} \in [0,1]$, $p_{i,22} \in [0,1]$ and $e^{\theta_i R_{\mathrm{E},i}} \geq 1$ as $\theta_i R_{\mathrm{E},i} \geq 0$. Moreover, it holds that $1 - p_{i,11} - p_{i,22} + p_{i,11} p_{i,22} = (1 - p_{i,11})(1 - p_{i,22}) \geq 0$. Thus, we have $\frac{\partial \mu_i}{\partial R_{\mathrm{E},i}} > 0$. For user i, the FBL throughput rate μ_i is strictly increasing in the effective capacity $R_{\mathrm{E},i}$.

In addition, accoding to (9), we have $\frac{\partial R_{\mathrm{E},i}}{\partial r_i} = \frac{m e^{-\theta_i m r_i}(1-\varepsilon_i)}{\left[e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i\right]} \geq 0$. Therefore, it holds that $\frac{\partial \mu_i}{\partial r_i} = \frac{\partial \mu_i}{\partial R_{\mathrm{E},i}} \frac{\partial R_{\mathrm{E},i}}{\partial r_i} \geq 0$, i.e., the FBL throughput rate μ_i is increasing in the coding rate r_i.

Then, we need to show that throughput μ_i is concave in coding rate r_i. In [35], it is shown that effective bandwidth of a Markovian source is strictly monotonically increasing and is also convex in source arrival rates a_i. Therefore, the inverse function of the effective bandwidth $a_{\mathrm{E},i}^{-1}\left(R_{\mathrm{E},i}\right)$ exists and is a non-decreasing concave function of the effective capacity $R_{\mathrm{E},i}$, i.e., $\frac{\partial \mu_i}{\partial R_{\mathrm{E},i}} \geq 0$ and $\frac{\partial^2 \mu_i}{\partial R_{\mathrm{E},i}^2} \leq 0$. Note that $\frac{\partial^2 R_{\mathrm{E},i}}{\partial r_i^2} \leq 0$ according to (15). Hence, we have $\frac{\partial^2 \mu_i}{\partial r_i^2} = \frac{\partial^2 \mu_i}{\partial R_{\mathrm{E},i}^2} \left(\frac{\partial R_{\mathrm{E},i}}{\partial r_i}\right)^2 + \frac{\partial \mu_i}{\partial R_{\mathrm{E},i}} \frac{\partial^2 R_{\mathrm{E},i}}{\partial r_i^2} \leq 0$. ■

Similarly as in the previous section, we consider optimal power allocation over time and among users to maximize the normalized sum throughput in the FBL regime. The optimization problem is given by

$$\begin{aligned} &\max_{\mathbf{p} \in \mathbf{\Omega}} \ \mu_{\mathrm{sum}} = \sum_{i=1}^{N} \mu_i \\ &s.t.: \ \mathbb{E}_{\mathbf{z}} \left\{ \sum_{i=1}^{N} p_i \right\} - \frac{M p_{\mathrm{ave}}}{m} \leq 0, \end{aligned} \tag{31}$$

where $\mathbf{p} = \{p_1, ..., p_N\}$ and $\mathbf{\Omega} = \{\Omega_i\}^N$, where Ω_i is the admissible set of p_i, which is provided in (17).

Proposition 5. *Problem* (31) *is a convex optimization problem.*

Proof: According to Proposition 1, under the SNR constraint guaranteeing $\gamma_i \geq 0$ dB, $\frac{\partial^2 r_i}{\partial p_i^2} \leq 0$. In addition, according to Proposition 4, $\frac{\partial^2 \mu_i}{\partial r_i^2} \leq 0$ and $\frac{\partial \mu_i}{\partial r_i} \geq 0$. Therefore, we have $\frac{\partial^2 \mu_i}{\partial p_i^2} = \frac{\partial^2 \mu_i}{\partial r_i^2} \left(\frac{\partial r_i}{\partial p_i}\right)^2 + \frac{\partial \mu_i}{\partial r_i} \frac{\partial^2 r_i}{\partial p_i^2} \leq 0$. Hence, the FBL throughput μ_i is concave in the transmit power p_i under the SNR constraint guaranteeing $\gamma_i \geq 0$ dB. Similarly as the proof of Proposition 3, the Hessian matrix of the objective function in (31) can be shown to be negative semidefinite. As the constraints are affine, Problem (31) is a convex optimization problem. ■

Then, the Lagrangian of Problem (31) is

$$L = \sum_{i=1}^{N} \mu_i - \lambda \, \mathbb{E}_{\mathbf{z}} \left\{ \sum_{i=1}^{N} p_i - \frac{M p_{\mathrm{ave}}}{m} \right\}. \tag{32}$$

Note that $g_{\text{D-M},i} = e^{-\theta R_{\mathrm{E},i}}$, and we can represent μ_i as a function of $g_{\text{D-M},i}$ as

$$\begin{aligned} \mu_i =& \frac{P_{\mathrm{on},i}}{M\theta_i} \ln \left(\frac{(1 - p_{i,11} g_{\text{D-M},i})}{(1 - p_{i,11} - p_{i,22})\, g_{\text{D-M},i} + p_{i,22}} \frac{1}{g_{\text{D-M},i}} \right) \\ =& - \frac{P_{\mathrm{on},i}}{M\theta_i} \ln \left((1 - p_{i,11} - p_{i,22})\, g_{\text{D-M},i} + p_{i,22} \right) \\ &+ \frac{P_{\mathrm{on},i}}{M\theta_i} \left[\ln \left(1 - p_{i,11} g_{\text{D-M},i}\right) - \ln g_{\text{D-M},i} \right]. \end{aligned} \tag{33}$$

Now, the first-order derivative can be expressed as

$$\begin{aligned} \frac{\partial \mu_i}{\partial g_{\text{D-M},i}} =& - \frac{P_{\mathrm{on},i}}{M\theta_i} \frac{1 - p_{i,11} - p_{i,22}}{(1 - p_{i,11} - p_{i,22})\, g_{\text{D-M},i} + p_{i,22}} \\ &- \frac{P_{\mathrm{on},i}}{M\theta_i} \frac{p_{i,11}}{1 - p_{i,11} g_{\text{D-M},i}} - \frac{1}{g_{\text{D-M},i}}. \end{aligned} \tag{34}$$

Hence, the first order derivative of the Lagrangian with respect to p_i is given by

$$\frac{\partial L}{\partial p_i} = \frac{\partial \mu_i}{\partial g_{\text{D-M},i}} \frac{\partial g_{\text{D-M},i}}{\partial p_i} - \lambda. \tag{35}$$

Letting $\frac{\partial L}{\partial p_i} = 0$, we have

$$\left(1 - \frac{A_i}{\sqrt{a_i}} \left(1 - a_i\right)\right) \left(1 - a_i\right)^{\frac{\eta_i m + 1}{2}} e^{\eta_i m A_i \sqrt{a_i}} \phi_{\text{D-M},i} - \lambda = 0, \tag{36}$$

where $\phi_{\text{D-M},i} = \frac{z_i(1-\varepsilon_i)\eta_i m^{1-\eta_i}}{\sigma^2} \frac{P_{\mathrm{on},i}}{M\theta_i} \cdot \left[\frac{p_{i,11}}{1 - p_{i,11} g_{\mathrm{D-M},i}} + \frac{1 - p_{i,11} - p_{i,22}}{(1 - p_{i,11} - p_{i,22}) g_{\text{D-M},i} + p_{i,22}} + \frac{1}{g_{\text{D-M},i}} \right]$.

By solving the (36) within the admissible set Ω_i provided in (17), we can determine the optimal power level p_i^*. Comparing (36) with (22), we notice that the procedures for solving the dual problems of (31) and (16) share the same structure. As the two problems have the same constraints, Problem (31) can also be solved by applying Algorithm 1 by replacing (22) in the algorithm with (36).

B. Markov Fluid Source

In this subsection, the data arrival is modeled as a continuous-time Markov process. For user i, the transition rate matrix of the data arrival process at the source can be expressed as

$$G_i = \begin{pmatrix} -\alpha_i & \alpha_i \\ \beta_i & -\beta_i \end{pmatrix}, \tag{37}$$

where $\alpha_i > 0$ and $\beta_i > 0$ are the transition rates between ON and OFF states for the data source generating information that will be sent to user i. We again assume a_i and 0 bits/frame arrive in the ON and OFF states, respectively. The LMGF of the arrival process is expressed as [36]

$$\Lambda_{a_i}(\theta_i) = \theta_i a_i - (\alpha_i + \beta_i) + \sqrt{\left(\theta_i a_i - (\alpha_i + \beta_i)\right)^2 + 4\alpha_i \theta_i a_i}. \tag{38}$$

Inserting the above equation into (5), we have $\left(\theta_i a_i - (\alpha_i + \beta_i) - 2\theta_i R_{\mathrm{E},i}\right)^2 = \left(\theta_i a_i - (\alpha_i + \beta_i)\right)^2 + 4\alpha_i \theta_i a_i$. Solving this equation, we can express the FBL throughput (or equivalently the average arrival rate) for user i as

$$\mu_i = a_i \frac{P_{\mathrm{on},i}}{M} = \frac{P_{\mathrm{on},i}}{M} \left(\frac{\theta_i R_{\mathrm{E},i} + \alpha_i + \beta_i}{\theta_i R_{\mathrm{E},i} + \alpha_i} \right) R_{\mathrm{E},i}, \tag{39}$$

where $P_{\mathrm{on},i} = \frac{\alpha_i}{\alpha_i + \beta_i}$ is the probability of the ON state.

As α_i, β_i, θ_i, M are non-negative, it is easy to show that

$$\frac{\partial \mu_i}{\partial R_{\mathrm{E},i}} = \frac{P_{\mathrm{on},i}}{M} \left(1 + \frac{\beta_i \alpha_i}{\left(\theta_i R_{\mathrm{E},i} + \alpha_i\right)^2} \right) \geq 0, \tag{40}$$

$$\frac{\partial^2 \mu_i}{\partial R_{\mathrm{E},i}^2} = -\frac{P_{\mathrm{on},i}}{M} \frac{\theta_i \beta_i \alpha_i}{\left(\theta_i R_{\mathrm{E},i} + \alpha_i\right)^3} \leq 0. \tag{41}$$

Hence, the FBL throughput μ_i (with Markov fluid source) is a non-decreasing concave function of the effective capacity, which has the same characteristic as the one formulated for the discrete-time Markov source. Hence, Proposition 4 holds also for the fluid source. Moreover, the power allocation problem (31) in the scenario with fluid source is also a convex problem. In fact, if we state the Lagrangian L of the problem and let $\frac{\partial L}{\partial p_i} = 0$, we have

$$\lambda - \left(1 - \frac{A_i}{\sqrt{a_i}} \left(1 - a_i\right) \right) \left(1 - a_i\right)^{\frac{\eta_i m + 1}{2}} e^{\eta_i m A_i \sqrt{a_i}} \phi_i' = 0, \tag{42}$$

where $\phi_i' = \frac{z_i (1 - \varepsilon_i) \eta_i m^{1 - \eta_i}}{\sigma^2} \frac{P_{\mathrm{on}}}{M\theta} \left[\frac{\alpha_i \beta_i}{g_{\text{C-M},i} (\alpha_i - \ln g_{\text{C-M},i})^2} + \frac{1}{g_{\text{C-M},i}} \right]$. Comparing (42) with (36), we observe that the only difference is in the ϕ_i' term. Hence, Problem (31) can also be solved in the presence of the Markov fluid source by applying Algorithm 1, i.e., by replacing (22) in the algorithm by (42).

C. Discrete-Time and Continuous-Time Markov-Modulated Poisson Processes (MMPP)

In this source model, the arrival of a user's data to the transmitter buffer follows a Poisson process, whose intensity is controlled by a Markov chain. We again consider a two-state model in which the intensity of the Poisson arrival process is a_i and 0 in the ON and OFF states of the Markov chain, respectively. Therefore, the source arrival is modeled as a Markov-modulated Poisson process (MMPP).

1) Discrete-Time MMPP: We first consider the Poisson process whose intensity is controlled by a discrete-time Markov chain, i.e., discrete-time MMPP. Assuming that the matrix G_i in (23) is the transition probability matrix of the Markov chain for user i, the LMGF of the arrival process of the data for user i is given by $\Lambda_{a_i}(\theta_i) = \ln\left(\frac{g_{\text{D-MP},i}}{2}\right)$, where $g_{\text{D-MP},i}$ is given by $g_{\text{D-MP},i} = p_{i,11} + p_{i,22} e^{a_i\left(e^{\theta} - 1\right)} + \sqrt{\left(p_{i,11} + p_{i,22} e^{a_i(e^{\theta} - 1)}\right)^2 - 4\left(p_{i,11} + p_{i,22} - 1\right) e^{a_i(e^{\theta} - 1)}}$. Hence, the effective bandwidth is given by $a_{\mathrm{E},i} = \frac{\Lambda_{a_i}}{\theta}$. Similar to the derivation in Section IV-A from (25) to (29), we can obtain the FBL throughput of user i with a discrete-time MMPP source as

$$\mu_i = \frac{P_{\mathrm{on},i}}{M\left(e^{\theta_i} - 1\right)} \ln\left(\frac{e^{\theta_i R_{i,\mathrm{E}}} \left(e^{\theta_i R_{i,\mathrm{E}}} - p_{i,11}\right)}{1 - p_{i,11} - p_{i,22} + p_{i,22} e^{\theta_i R_{i,\mathrm{E}}}} \right), \tag{43}$$

where $P_{\mathrm{on},i}$ is given in (24).

$$g_{\text{eq},i}(p_{\text{ave}}) = \mathbb{E}_{z_i}\left\{e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i\right\} = \mathbb{E}_{z_i}\left\{ e^{-\frac{\theta_i}{\ln 2}m\left[\ln\left(1+\frac{z_i p_{\text{ave}}}{\sigma^2}\right)-A_i\sqrt{\left(1-\frac{1}{\left(1+\frac{z_i p_{\text{ave}}}{\sigma^2}\right)^2}\right)}+\frac{\ln m}{m}\right]}(1-\varepsilon_i)+\varepsilon_i \right\}. \quad (45)$$

$$g_{\text{sub},i}\left(p_i\right) = \mathbb{E}_{z_i}\left\{e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i\right\} = e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i = e^{-\frac{\theta_i}{\ln 2}m\left[\ln\left(1+\frac{z_i p_i}{\sigma^2}\right)-A_i\sqrt{\left(1-\frac{1}{\left(1+\frac{z_i p_i}{\sigma^2}\right)^2}\right)}+\frac{\ln m}{m}\right]}(1-\varepsilon_i)+\varepsilon_i. \quad (46)$$

2) Continuous-Time MMPP: If the intensity of Poisson process is controlled by a continuous-time Markov chain, we have a continuous-time MMPP. Considering (37) as the transition probability matrix of the Markov chain, the LMGF of the arrival process is [36] $\Lambda_{a_i}(\theta_i) = (e^{\theta_i}-1)a_i - (\alpha_i+\beta_i) + \sqrt{\left((e^{\theta_i}-1)a_i - (\alpha_i+\beta_i)\right)^2 + 4\alpha_i(e^{\theta_i}-1)a_i}$. Following similar steps as in the derivations in previous subsections, we can determine the FBL throughput of user i in the presence of a continuous-time MMPP source as

$$\mu_i = \frac{P_{\text{on},i}\theta_i}{M(e^{\theta_i}-1)}\left(\frac{\theta_i R_{\text{E},i}+\alpha_i+\beta_i}{\theta_i R_{\text{E},i}+\alpha_i}\right)R_{\text{E},i}, \quad (44)$$

where $P_{\text{on},i} = \frac{\alpha_i}{\alpha_i+\beta_i}$ is the probability of ON state.

So far in this subsection, we have discussed the FBL throughputs with discrete-time and continuous-time MMPP sources. By comparing (44) with (39) and comparing (43) with (29), we immediately notice that the FBL throughput expressions for discrete-time and continuous-time MMPP sources are scaled versions of the throughputs of discrete-time and continuous-time Markov sources, respectively, scaled by $\frac{\theta_i}{e^{\theta_i}-1}$. These scaling differences do not alter the optimal power allocation problem, i.e., the proposed optimal algorithms can be readily applied.

V. Equal Power Allocation and Sub-optimal Power Allocation Policies

In the previous two sections, we analyzed the optimal power allocation strategy for downlink multiuser networks considering constant-rate and random arrivals, and address how to optimally allocate the power over frames (time) and among users. In this section, we discuss two additional approaches and study their performance in the FBL regime.

A. Equal power allocation

Under equal power allocation, each user in each frame is allocated a fixed power level of p_{ave}. Hence, the effective capacity is given by $R_{i,\text{E}} = -\frac{1}{\theta_i}\ln g_{\text{eq},i}$ where $g_{\text{eq},i}$ with equal power allocation is given in (45). Substituting $R_{i,\text{E}}$ or $g_{\text{eq},i}$ into (29), (39), (43) and (44), we then can determine the FBL throughputs with random arrivals, which is given by (45) on the next page. It should be mentioned that the above equal power allocation policy is not able to guarantee the reliability of users' transmissions, i.e., does not satisfy (17). Hence, it is not regarded as a sub-optimal power allocation policy.

B. Power allocation among users within a frame

In this subsection, we only allocate power among users within each frame, and not perform power allocation over frames (i.e., over time). Therefore, we now consider an instantaneous or short-term power constraint, i.e., for each frame, we have $\sum_{i=1}^{N} p_i - Mp_{\text{ave}}/m \leq 0$. In addition, the power allocation is subject to the current channel gain z_i, $i=1,...,N$, while the channel distributions are not taken into account. The objective is to maximize the sum of $\hat{\mu}_i$ which is not the actual FBL throughput over time but the FBL throughput under the assumption of a static channel, i.e., this throughput formulation is obtained assuming that the current channel gain z_i applies to all frames. In this case, the effective capacity is $R_{i,\text{E}} = -\frac{1}{\theta_i}\ln g_{\text{sub},i}$ where $g_{\text{sub},i}$ with static z_i is given by (46) on the next page. Based on $g_{\text{sub},i}$ and $R_{i,\text{E}}$, we can easily express $\hat{\mu}_i$ for different types of data arrivals.

In the following, we discuss the power allocation problem in the presence of a discrete Markov source, while the problem for the other types of data arrivals can be solved similarly. For the discrete Markov source, $\hat{\mu}_i$ can be obtained by substituting (46) into (33). Then, the problem of power allocation among users within a single frame is given by

$$\begin{aligned} \max_{\mathbf{p}\in\mathbf{\Omega}} \ & \mu_{\text{sum}} = \sum_{i=1}^{N}\hat{\mu}_i \\ s.t.: \ & \sum_{i=1}^{N} p_i - \frac{Mp_{\text{ave}}}{m} \leq 0, \end{aligned} \quad (47)$$

where $\mathbf{p} = \{p_1, ..., p_N\}$ and $\mathbf{\Omega} = \{\Omega_i\}^N$, where Ω_i is the admissible set of p_i, as defined in (17).

Since the static channel fading model can be seen as a specific case of quasi-static channel model, it is easy to show that Problem (47) is a convex problem. The Lagrange dual function is given by

$$L=\sum_{i=1}^{N}\hat{\mu}_i-\lambda\left\{\sum_{i=1}^{N}p_i-\frac{Mp_{\text{ave}}}{m}\right\}. \tag{48}$$

Letting $\frac{\partial L}{\partial p_i}=0$, we have exactly the same equation in (36), with only a different expression for g_i (which, in the sub-optimal power allocation strategy, is denoted as $g_{\text{sub},i}$). Under these assumptions, to determine the power levels via numerical computations, we propose the following algorithm described in Algorithm 2 below for this sub-optimal power allocation strategy.

Algorithm 2 : Sub-Optimal Power Allocation Algorithm.

```
1) for user $i = 1, ..., N$
  a) if $z_i < z_{\min}$
  b)    then $p_i^* = 0$ and go to Step 1 for the next user;
  c)    else Given $\lambda$, determine $p_i$ according to (36) and (46).
  We have $p_i^* = \max\{p_i, \frac{\gamma_{\text{th},i}\sigma^2}{z_i}\}$.
      endif
    endfor
2) Check if the sum of the obtained $p_i^*$ satisfy the average power constraint.
  a) if not satisfied with equality
  b)    then update the value of $\lambda$ and return to Step 1;
  c)    else the optimal power $p_i^*$, $i = 1, ..., N$ is obtained.
      endif
```

Since $g_{\text{sub},i}$ in the proposed suboptimal approach is fully decided by the choice of p_i, Algorithm 2 has a relatively lower complexity than Algorithm 1 in finding the power levels. However, the advantage of Algorithm 1 is that the optimal power allocation is determined only once in the system initialization, and there is no need to perform a search for each frame. On the other hand, Algorithm 2 has to perform this for each frame, which is an obvious disadvantage. In addition, as this sub-optimal solution doesn't consider the allocation of power over frames, it definitely has a lower performance than the optimal one.

VI. NUMERICAL RESULTS

In this section, we provide our numerical results. First the proposed optimal power allocation algorithm is compared with suboptimal power allocation schemes. Subsequently, we move to a more general performance investigation of the considered downlink network under the proposed algorithm. In all the numerical results, we consider the following parameterization. First, we set unit average channel gain for all links, while assuming that all links experience independent and identically distributed (i.i.d) Rayleigh quasi-static fading, i.e., $f_{\text{PDF}}(z_i)=e^{-z_i}, i=1,...,N$. Therefore, the joint PDF of all channels is $f_{\text{PDF}}(\mathbf{z})=e^{-\sum_1^N z_i}$. Secondly, the noise power and the average power constraint p_{ave} are set to 1 mW (0 dBm) and 50 mW (17 dBm), respectively. In addition, we set $z_{\min}=\frac{1}{50}$. Then, we have $\Pr\{z_i<\frac{1}{50}\}=0.02$, i.e., with probability 0.02 the channel gain z_i is worse than the bound $z_{\min}$ and we allocate zero power to user i. The blocklength for each user is set to $m=300$ symbols. Unless specified otherwise, the default setup for the number of users is two, and both users have the same type of sources and the same QoS requirements. Finally, we consider both the constant-rate and random arrival models. In particular, in the analysis of random arrivals, we mainly focus on the discrete-time Markov source especially when comparing with constant arrivals. The continuous-time Markov, discrete-time MMPP and continuous-time MMPP sources will also be addressed towards the end of the numerical results.

A. Comparison with sub-optimal algorithms

To start with, in this subsection we provide comparisons between the optimal algorithm (opt) versus the equal power allocation (equ) and the power allocation within a frame (subopt) discussed in Section V. The comparison is with respect to the normalized sum throughputs as a function of error probabilities and QoS exponents, respectively. In addition, both the constant arrival model and the discrete-time Markov arrival model are taken into account.

We first vary the error probability and provide the results in Fig. 3. It can be observed that all normalized sum throughput curves are concave in the users' target error probability. In addition, the sub-optimal algorithm, which optimally allocates the power among users within a frame, performs better than the equal power allocation for both the constant-rate data source and the random data source. Moreover, the optimal algorithm, which optimally allocates power among users and over frames, further improves the throughput performance in comparison to the sub-optimal algorithm. Furthermore, a higher throughput level is achieved with the constant data source than with the random data source. In particular, a lower probability of the ON

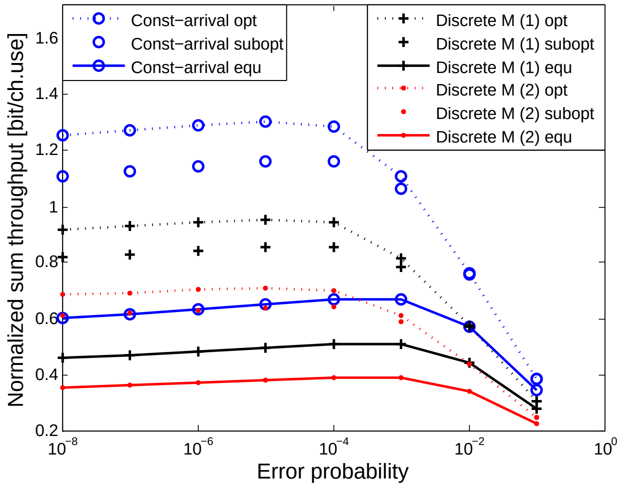


Fig. 3. Comparison between the proposed algorithms and the equal power allocation in a two-user scenario, while setting $\theta_1 = \theta_2 = 10^{-2}$ and varying the error probability. Case 1 and Case 2 are with different setups of the state transition probability matrix of the discrete-time Markov data arrival, i.e., Case 1: $G = \{0.3, 0.7; 0.3, 0.7\}$, $P_{\rm on} = 0.7$; Case 2: $G = \{0.5, 0.5; 0.5, 0.5\}$, $P_{\rm on} = 0.5$.

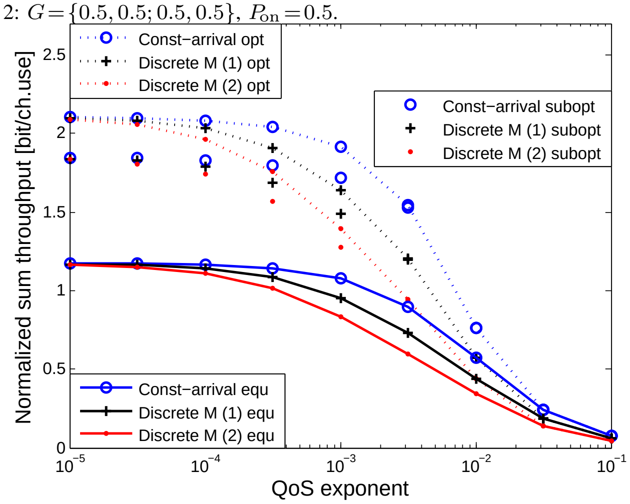


Fig. 4. Comparison between the proposed algorithms and the equal power allocation in a two-user scenario, while setting $\varepsilon_1 = \varepsilon_2 = 10^{-2}$ and varying the QoS exponent of the two users. Case 1 and Case 2 are with different setups of the state transition probability matrix of the discrete-time Markov data arrival i.e., Case 1: $G = \{0.3, 0.7; 0.3, 0.7\}$, $P_{\rm on} = 0.7$; Case 2: $G = \{0.5, 0.5; 0.5, 0.5\}$, $P_{\rm on} = 0.5$.

state $p_{\rm on}$ (indicating a more bursty source) leads to a lower sum throughput, as observed in Fig. 3 when the performances with two different discrete Markov sources are compared.

We continue the comparison in Fig. 4 where the QoS exponent of users are varied. First, all the throughput curves are decreasing in the QoS exponent θ. In particular, when θ is as small as 10^{-5}, for each power allocation algorithm the throughput performances with different source models converge to the same value. This is because a small θ indicates a loose QoS requirement, which makes a link-layer QoS-constrained performance converge to the physical-layer performance that does not depend on the source characteristics. Secondly, the optimal power allocation is observed to provide a higher throughput than the two sub-optimal approaches. Finally, we have the same observation as in Fig. 3 that for each algorithm a lower $p_{\rm on}$ results in a lower sum throughput.

B. Evaluation

In Section VI-A, we show the performance advantage of the proposed optimal algorithm in comparison to two sub-optimal algorithms, while also demonstrating the impact of error probability and QoS exponent on the throughput performance. In this subsection, we continue investigating the throughput performance of the optimal algorithm under different setups, e.g., for different values of the coding blocklength, and different number of users.

First, we study the impact of blocklength on the normalized sum throughput and provide the results in Fig. 5. In the figure, both the constant source and discrete-time Markov source are considered for three different values of $\{\theta, \varepsilon\}$ pairs. When θ is relatively large, the throughputs are decreasing in the blocklength. On the other hand, with a relatively small θ, the throughputs are observed to be concave in the blocklength. In particular, by comparing the top two groups of curves we find that the sharpness of the concavity is influenced by the error probability ε, e.g., curves with $\theta = 10^{-4}, \varepsilon = 10^{-2}$ are relatively more flat than the curves with $\theta = 10^{-4}, \varepsilon = 10^{-4}$.

We also investigate the relationship between the throughput and the number of users in the system. In Fig. 6, we set m to

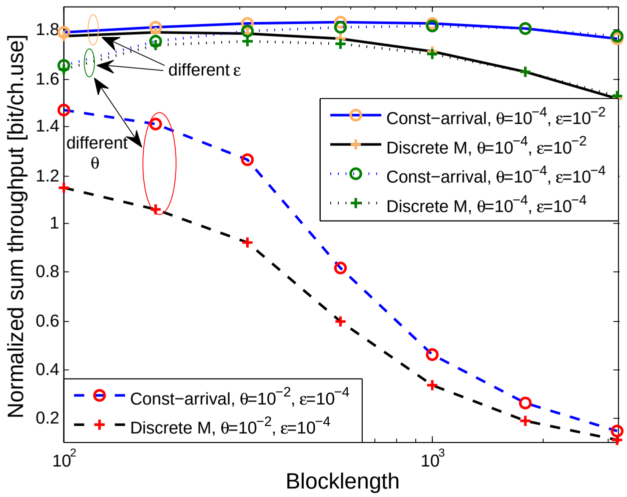


Fig. 5. The impact of blocklength on the throughput, while considering different setup of QoS exponents and error probabilities.

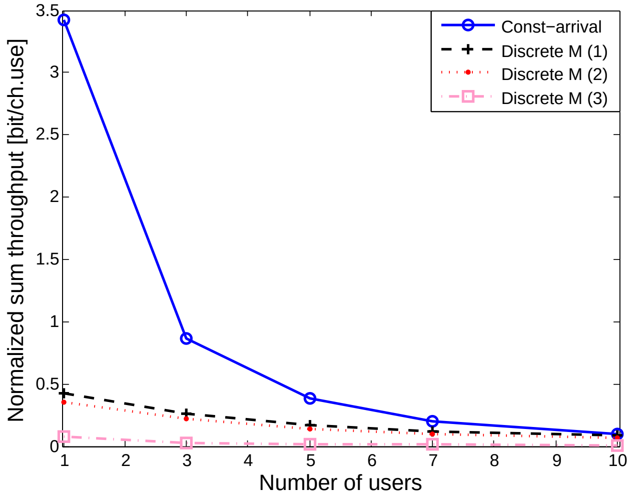


Fig. 6. The relationship between user number (which linearly increase the frame length) and normalized sum throughput. In the figure, we set $m = 300$, $M = N \cdot m$ where N is the user number. Case 1: $G = \{0.3, 0.7; 0.3, 0.7\}$, Case 2: $G = \{0.5, 0.5; 0.5, 0.5\}$, Case 3: $G = \{0.7, 0.3; 0.7, 0.3\}$.

be fixed at 300 (symbols) and therefore the frame length increases linearly in the number of users N, i.e., $M = Nm$. We find that the normalized sum throughputs decrease as the number of users grows. The reason is that as more users with fixed transmission blocklength are to be served within the same frame, the system becomes less flexible, i.e., we have longer frame durations and longer waiting time in the queue. This significantly strains the queue stability and increases the buffer violation probability, thus, leading to a lower QoS-constrained throughput. We also consider a different simulation setup, where we fix the frame length as $M = 2000$ (symbols) and let more and more users to share this frame duration, i.e., the blocklength of each user's transmission linearly decreases with the increasing number of users. The results are provided in Fig. 7, where both the constant and discrete MMPP sources are considered. Interestingly, we find that the each throughput curve for this setup initially increases as the number of users grows and levels off as the number of users sharing the same fixed frame duration is further increased. The explanation for this is as follows. There is an underlying trade-off in this scenario, between the FBL performance degradation due to shorter blocklength for each user versus the multiuser diversity in the power allocation. When less users share the frame, each user has a relatively long blocklengh. In this case, increasing the multiuser diversity significantly improves the normalized sum throughput via the power allocation. On the other hand, when the multiuser diversity already achieves a certain level of performance, adding one or two users provides less gain but introduces more FBL loss due to shorter coding blocklength. As a result, all the curves start leveling off when the number of users is relatively large. The figure actually suggests the optimal size of the user group in a QoS-supporting network with finite blocklengths codes.

In Fig. 6 and Fig. 7, we show the throughput performance with discrete-time Markov and discrete-time MMPP sources. In fact, the throughput with a Markov source is similar to the corresponding MMPP source. Recall that the throughput expressions for discrete-time and continuous-time MMPP sources in (43) and (44) are exactly scaled by $\frac{\theta}{e^{\theta}-1}$ with respect to the throughput expressions for the corresponding Markov sources in (29) and (39). Note that $\frac{\theta}{e^{\theta}-1} \leq 1, \theta \geq 0$ since for $f(\theta) = e^{\theta} - 1 - \theta$ we have $f(0) = 0$ and $f'(\theta) = e^{\theta} - 1 \geq 0, \forall \theta \geq 0$. Hence, the throughput gap between a Markov source and the corresponding MMPP source increases with increasing θ. Conversely, if θ is not significantly large, the performance gap is small. An example

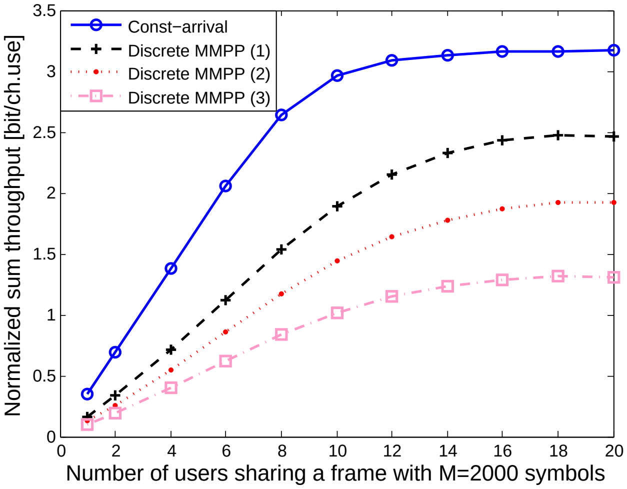


Fig. 7. For a given length of frame, i.e., $M = 2000$ symbols, the impact of the user number N on the sum throughput of scenario with discrete-time MMPP. $m = M/N$. Case 1: $G = \{0.3, 0.7; 0.3, 0.7\}$, Case 2: $G = \{0.5, 0.5; 0.5, 0.5\}$, Case 3: $G = \{0.7, 0.3; 0.7, 0.3\}$.

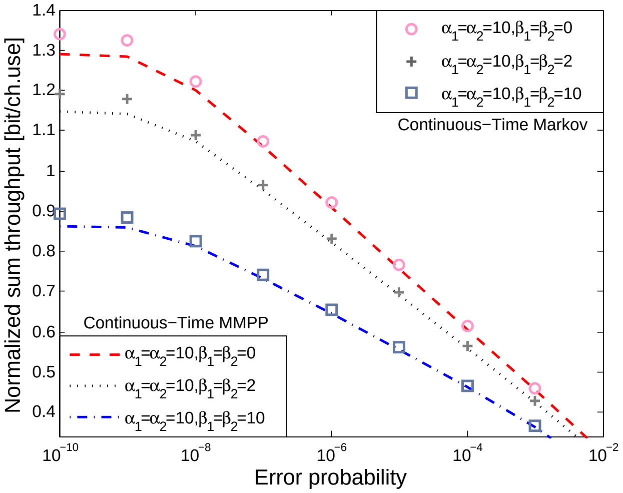


Fig. 8. Throughputs of two-user network with Markov fluid source and continuous-time MMPP, while varying error probability of the two users.

is provided in Fig. 8, where θ is set to 0.025. The figure shows that the throughput curves of continuous-time Markov and MMPP sources have similar shapes and small performance gaps exist for each setup of $\{\alpha, \beta\}$. In addition, the figure shows that a large β leads to a low throughput performance. This is due to the fact that as α is fixed, larger β leads to smaller P_{on} (i.e., the source becomes more bursty) which leads to lower throughput.

In all the previous numerical results, we have considered homogeneous QoS requirements for all users, e.g., we assume that the users have the same ε and θ and we vary these two factors for all users simultaneously. In the last figure, we discuss the impact of varying the QoS requirements of a single user. We show the results in Fig. 9 where we consider a continuous-time Markov model with different β values. There in total are five curves in the figure, which can be divided into two groups. The curve at the bottom is a shared reference curve belonging to both groups. The curves denoted with blue $\circ$ are in the first group, for which we fix the error probability as 10^{-2} and vary θ. By reducing θ_1 from 10^{-2} to 10^{-3}, the throughput improves by approximately 0.6 bit/ch.use. In addition, the throughput is further increased by about 0.6 bit/ch.use by also reducing θ_2 to the same value. The second group contains three curves at the bottom of the figure, where we fix θ and vary ε. We observe a similar result that the improvement by reducing ε_2 from 10^{-2} to 10^{-3} is about half of improvement by reducing both ε_1 and ε_2. In addition, we again find that all throughput curves are decreasing in β. Finally, by comparing the two groups it can be seen that the impact of θ on the throughput performance is more significant than that of ε, matching with the results in Fig. 3 and Fig. 4.

VII. CONCLUSION

In this paper, we have investigated optimal power allocation strategies in a downlink multiuser URLLC network in quasi-static fading channels when the data arrivals are modeled as constant-rate, discrete-time and continuous-time Markov processes, and discrete-time and continuous-time MMPP. The normalized sum throughput is maximized in the presence of statistical QoS constraints and FBL codes. First, we have developed the FBL throughput models. Subsequently, based on the model we

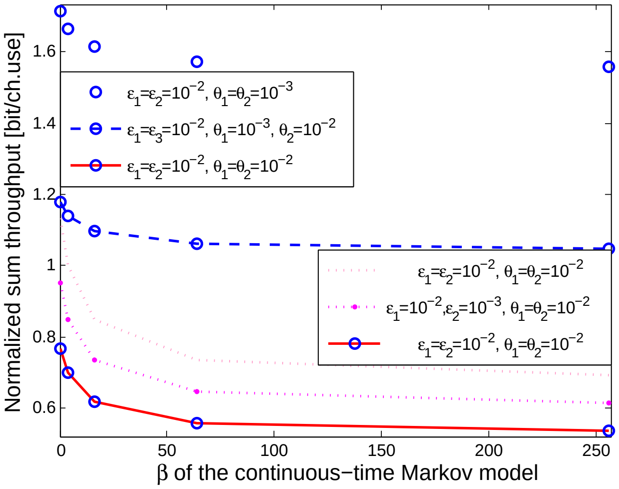


Fig. 9. Throughputs of a two-user network, while varying β. $\alpha_1 = \alpha_2 = 10$, $\beta_1 = \beta_2 = \beta$. In the figure, the random-arrival is based on discrete-time Markov source.

have formulated optimization problems and shown the convexity of the problem. Then, we solved the problem by proposing optimal algorithms. In addition, the performances of two sub-optimal power allocation algorithms have also been analyzed and compared with the optimal power allocation.

Via numerical analysis, first we have shown that the proposed algorithm outperforms both sub-optimal algorithms. We have demonstrated that a higher throughput is achieved with a constant source than with a random data source, while for each of the considered random data sources a more bursty source leads to a lower throughput. We have also observed that the normalized sum throughputs are concave in the users' error probability and are decreasing in the QoS exponent. Normalized sum throughputs are also shown to be decreasing with increasing number of users when we fix the blocklength of transmission for each user and increase the frame length with the number of users. On the other hand, if we fix the total frame length and let more and more users share the frame, then the normalized sum throughputs increase as the number of users grow and level off when the number of users is relatively large. We have also shown that the throughput gap between a Markov source and the corresponding MMPP source is small when the QoS exponent is not significantly large. Finally, for the scenario with two users where the QoS requirements are different, we have demonstrated that the influence by changing the QoS requirement of one user is about half of influence by changing both. This indicates that the general observations based on the homogeneous setup of users hold also for the heterogeneous users.

References

[1] C. She, C. Yang and T. Q. S. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 72-78, Jun. 2017.

[2] Y. Hu, M. C. Gursoy and A. Schmeink, "Relaying-Enabled Ultra-Reliable Low Latency Communications in 5G", *IEEE Network*, vol. 32, no. 2, pp. 62-68, Mar.-Apr. 2018.

[3] S. C. Lin and K. C. Chen, "Statistical QoS control of network coded multipath routing in large cognitive machine-to-machine networks," in *IEEE Internet of Things Journal*, vol. 3, no. 4, pp. 619-627, Aug. 2016.

[4] S. Girs *et al.*, "Scheduling for source relaying with packet aggregation in industrial wireless networks," in *IEEE Trans. Ind. Informat.*, vol. 12, no. 5, pp. 1855-1864, Oct. 2016.

[5] Q. Du *et al.*, "Statistical delay control and QoS-driven power allocation over two-hop wireless relay links," in *IEEE GLOBECOM*, Houston, TX, USA, Dec. 2011, pp. 1–5.

[6] Q. Du and X. Zhang, "QoS-aware base-station selections for distributed MIMO links in broadband wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 6, pp. 1123–1138, Jun. 2011.

[7] Z. Chang, Z. Han and T. Ristaniemi, "Energy efficient optimization for wireless virtualized small cell networks with large scale multiple antenna," *IEEE Trans on Commun.*, vol. 65, no. 4, Apr. 2017.

[8] N. Petreska, H. Al-Zubaidy and J. Gross, "Power minimization for industrial wireless networks under statistical delay constraints," *26th IEEE International Teletraffic Congress*, Karlskrona, 2014, pp. 1-9.

[9] T. Abrao *et al.*, "Energy efficient OFDMA networks maintaining statistical QoS guarantees for delay-sensitive traffic," *IEEE Access*, vol. 4, pp. 774-791, Feb. 2016.

[10] X. Mi *et al.*, "Statistical QoS-driven resource allocation and source adaptation for D2D communications underlaying OFDMA-based cellular networks," *IEEE Access*, vol. 5, pp. 3981-3999, Mar. 2017.

[11] J. Choi, "Effective capacity of NOMA and a suboptimal power control policy with delay QoS," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1849-1858, Apr. 2017.

[12] M. Wiczanowski *et al.*, "Optimal energy control in energy-constrained wireless networks with random arrivals under stability constraints," *IEEE SPAWC*, New York, USA, Jun. 2005.

[13] M. Ozmen and M. C. Gursoy, "Wireless throughput and energy efficiency with random arrivals and statistical queuing constraints," in *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1375-1395, Mar. 2016.

[14] D. Qiao, M. Ozmen and M. C. Gursoy, "QoS-driven power control in fading multiple-access channels with random arrivals,"*IEEE ICC*, Kuala Lumpur, 2016, pp. 1-6.

[15] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[16] ——, “Dispersion of the Gilbert-Elliott channel,” *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 1829–1848, Apr. 2011.

[17] W. Yang *et al.*, “Quasi-static multiple-antenna fading channels at finite blocklength,” *IEEE Trans. Inf. Theory*, vol. 60, no. 7, Jul. 2014.

[18] G. Ozcan and M. C. Gursoy, “Throughput of cognitive radio systems with finite blocklength codes,” *IEEE J. Sel. Areas Commun.*, vol. 31, no. 11, pp. 2541-2554, Nov. 2013.

[19] S. Xu *et al.*, ”Energy-efficient packet scheduling with finite blocklength codes: Convexity analysis and efficient algorithms,” *IEEE Trans. Wireless. Commn.*, vol.15, no.8, pp.5527-5540, Aug. 2016.

[20] Y. Hu, A. Schmeink and J. Gross “Blocklength-limited performance of relaying under quasi-static Rayleigh channels,” *IEEE Trans. Wireless. Commn.*, vol. 15, no. 7, pp. 4548 - 4558, Jul. 2016.

[21] G. Ozcan and M. C. Gursoy, ”Throughput of cognitive radio systems with finite blocklength codes,” *IEEE J. Select. Area Commun.*, vol. 31, no. 11, pp. 2541-2554, Nov. 2013.

[22] Y. Hu, J. Gross and A. Schmeink, “On the capacity of relaying with finite blocklength,” *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1790-1794, Mar. 2016.

[23] Y. Hu, J. Gross and A. Schmeink, “On the performance advantage of relaying under the finite blocklength regime,” *IEEE Commun.Letters*, vol. 19, no. 5, pp. 779–782, May 2015.

[24] Y. Hu *et al.*, ”Finite Blocklength Performance of Cooperative Multi-Terminal Wireless Industrial Networks,” *IEEE Trans. Veh. Technol.*, pp. 1-1, Jan. 2018.

[25] Y. Hu, A. Schmeink and J. Gross, “Y. Hu, A. Schmeink and J. Gross, ”Optimal Scheduling of Reliability-Constrained Relaying System under Outdated CSI in the Finite Blocklength Regime,”,” *IEEE Commun.Letters*,” *IEEE Trans. Veh. Technol.*, pp. 1-1, Mar.. 2018.

[26] W. Yang *et al.*, ”Finite-blocklength channel coding rate under a long-term power constraint,” in *IEEE ISIT*, Jun. 2014, pp. 2067–2071.

[27] Y. Hu *et al.*, “Optimal power allocation for QoS-constrained downlink networks with finite blocklength codes,” *IEEE WCNC*, Apr. 2018, Barcelona, Spain.

[28] V. Y. F. Tan and M. Tomamichel, “The third-order term in the normal approximation for the AWGN channel,” *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2430–2438, May 2015.

[29] D. Wu and R. Negi, “Effective capacity: a wireless link model for support of quality of service,” *IEEE Trans. Wireless Commun*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[30] C.-S. Chang, “Stability, queue length, and delay of deterministic and stochastic queueing networks,” *IEEE Trans. Automat. Contr.*, vol. 39, pp. 913–931, May 1994

[31] S. Tanwir and H. Perros, “A survey of VBR video traffic models,” *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 4, Jan. 2013, pp. 1778–1802.

[32] M. Laner *et al.*, “Traffic models for machine type communications,” *IEEE ISWCS*, Ilmenau, Germany, Aug. 2013.

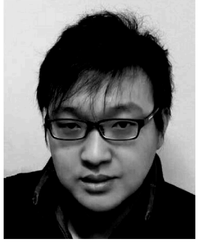
[33] E. Grigoreva *et al.*,“Coupled Markovian arrival process for automotive machine type communication traffic modeling,” *IEEE ICC* Paris, France, May 2017.

[34] C.-S. Chang and T. Zajic, “Effective bandwidths of departure processes from queues with time varying capacities,” in *IEEE INFOCOM*, Apr. 1995, pp. 1001–1009.

[35] A. I. Elwalid and D. Mitra, “Effective bandwidth of general Markovian traffic sources and admission control of high speed networks,” *IEEE ACM Trans. Netw.*, vol. 1, no. 3, pp. 329–343, Jun. 1993.

[36] G. Kesidis, J. Walrand, and C.-S. Chang, “Effective bandwidths for multiclass Markov fluids and other ATM sources,” *IEEE ACM Trans. Netw.*, vol. 1, no. 4, pp. 424–428, Aug. 1993.

[37] S. Boyd and L. Vandenberghe, Convex optimization. New York, NY, USA: Cambridge Univ. Press, 2004.



Yulin Hu received his M.Sc.E.E degree from USTC, China, in 2011. He successfully defended his dissertation of a joint Ph.D. program supervised by Prof. Anke Schmeink at RWTH Aachen University and Prof. James Gross at KTH Royal Institute of Technology in Dec. 2015 and received his Ph.D.E.E. degree (Hons.) from RWTH Aachen University where he was a Research Fellow since Jan. to Dec. in 2016. Since 2017, he works as a senior researcher and project lead in ISEK research group at RWTH Aachen University. From May to July in 2017, he was a visiting scholar with Prof. M. Cenk Gursoy in Syracuse University, USA. His research interests are in information theory, optimal design of wireless communication systems. He has been invited to contribute submissions to multiple conferences. He was a recipient of the IFIP/IEEE Wireless Days Student Travel Awards in 2012. He received the Best Paper Awards at IEEE ISWCS 2017 and IEEE PIMRC 2017, respectively. He is currently serving as an editor for Physical Communication (Elsevier).

Mustafa Ozmen received the B.S.E.E degree from Bilkent University, Ankara, Turkey, in 2011 and Ph.D.E.E. degree from Syracuse University in 2017. He is currently a postdoctoral researcher in School of Engineering at Brown University. He was a recipient of the ISIT Student Travel Awards in 2012 and 2015.

M. Cenk Gursoy received the B.S. degree with high distinction in electrical and electronics engineering from Bogazici University, Istanbul, Turkey, in 1999 and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, in 2004. He was a recipient of the Gordon Wu Graduate Fellowship from Princeton University between 1999 and 2003. Between 2004 and 2011, he was a faculty member in the Department of Electrical Engineering at the University of NebraskaLincoln (UNL). He is currently a Professor in the Department of Electrical Engineering and Computer Science at Syracuse University. His research interests are in the general areas of wireless communications, information theory, communication networks, and signal processing. He is currently a member of the editorial boards of IEEE Transactions on Wireless Communications, IEEE Transactions on Green Communications and Networking, IEEE Transactions on Communications, IEEE Transactions on Vehicular Technology. He also served as an editor for IEEE Transactions on Wireless Communications between 2010 and 2015, IEEE Communications Letters between 2012 and 2014, IEEE Journal on Selected Areas in Communications - Series on Green Communications and Networking (JSAC-SGCN) between 2015 and 2016, and Physical Communication (Elsevier) between 2010 and 2017. He was a Co-Chair of the Communication QoS and System Modeling Symposium, 2017 International Conference on Computing, Networking and Communications (ICNC). He received an NSF CAREER Award in 2006. More recently, he received the EURASIP Journal of Wireless Communications and Networking Best Paper Award, 2017 IEEE PIMRC Best Paper Award, 2017 IEEE Green Communications & Computing Technical Committee Best Journal Paper Award, UNL College Distinguished Teaching Award, and the Maude Hammond Fling Faculty Research Fellowship. He is a Senior Member of IEEE, and is the Aerospace/Communications/Signal Processing Chapter Co-Chair of IEEE Syracuse Section.

Anke Schmeink received the Diploma degree in mathematics with a minor in medicine and the Ph.D. degree in electrical engineering and information technology from RWTH Aachen University, Germany, in 2002 and 2006, respectively. She worked as a research scientist for Philips Research before joining RWTH Aachen University in 2008 where she is an associate professor since 2012. She spent several research visits with the University of Melbourne, and with the University of York. Anke Schmeink is a member of the Young Academy at the North Rhine-Westphalia Academy of Science. Her research interests are in information theory, systematic design of communication systems and bioinspired signal processing.