Optimal Power Allocation for QoS-Constrained Downlink Networks with Finite Blocklength codes

Yulin Hu[1], Mustafa Ozmen[2], M. Cenk Gursoy[2] and Anke Schmeink[1]

[1]Information Theory and Systematic Design of Communication Systems, RWTH Aachen University,
52062 Aachen, Germany. Email: $hu|schmeink$@umic.rwth-aachen.de
[2]Department of Electrical Engineering and Computer Science, Syracuse University,
Syracuse, NY 13244, USA. E-mail: $mozmen|mcgursoy$@syr.edu

Abstract

In this paper, we consider a downlink multiuser network operating with finite blocklength codes under statistical quality of service (QoS) constraints. An optimal power allocation algorithm is studied to maximize the normalized sum throughput under QoS constraints. 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 an optimal algorithm to solve the problem. 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 and blocklength on the performance.

Index Terms

effective capacity, downlink, finite blocklength regime, multiuser, 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]. The common characteristic of URLLC networks is that these network serve 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. In [3], the optimal power allocation schemes are proposed to satisfy QoS requirements for a two-hop wireless relay network. A similar work is considered in a multi-user network in [4]. An energy-efficient design is proposed in [5] under specific statistical QoS guarantees of a multi-user network. In [6], 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 [7] for non-orthogonal multiple access (NOMA) networks with QoS constraints. However, all of the above studies on resource allocation in QoS-constrained networks 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 (due to noise) becomes significant even if the rate is selected below the Shannon limit. Taking this into account, an accurate approximation of the achievable coding rate under the finite blocklength assumption for an additive white Gaussian noise (AWGN) channel was derived in [8], [9] for a single-hop transmission system. Subsequently, the initial work for AWGN channels was extended to Gilbert-Elliott channels [10] as well as quasi-static fading channels [11]–[13] and QoS-constrained networks [14], [15]. However, power allocation in QoS-constrained multi-user networks has not been addressed in the FBL regime. In fact, an FBL code is a double-edged sword for the 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 a power allocation policy that maximizes the QoS-constrained performance of multi-user networks in the FBL regime.

In this paper, we study the power allocation in a downlink multi-user wireless network operating with FBL code. The contributions of this paper can be further detailed as follows: *i.* The QoS-constrained FBL performance is formulated for a network with a constant rate arrival. In particular, the normalized throughput of each user is derived and proved to be conditionally concave in the transmit power. *ii.* We state the optimal power allocation problem that maximizes the normalized sum throughput by optimally allocating the power among users and over time. We prove the convexity of the optimization problem and propose optimal algorithm to solve it. 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 constant data arrivals and statistical queueing constraints. *iii.* 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 and coding blocklength 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. In Section III, we study optimal power allocation under constant arrivals to maximize the FBL throughput. We provide our numerical results in Section IV and finally conclude the paper in Section V.

II. Preliminaries

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

A. System model

We consider a downlink broadcasting scenario where a transmitter (e.g., an access point or a base station) sends data packets to N users. The entire system operates in a slotted fashion where time is divided into frames of length M symbols. In each frame, the transmitter sends packects to the user in different slots, as shown in Fig. 1. In particular, a frame has N slots for orthogonal transmissions to N users, and each slot has a length of m symbols, i.e., $Nm = M$.

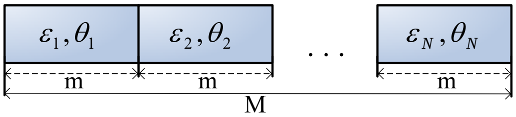
ε_1, θ_1
ε_2, θ_2
. . .
ε_N, θ_N
m
m
m
M

Fig. 1. Frame structure of the considered network.

We consider a scenario where data arrivals at the transmitter is with a constant rate. 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. Denote the set of instantaneous channel gains by $\mathbf{z} = \{z_i, i = 1, ..., N\}$ where z_i is 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 thepower 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., $\mathbb{E}\{\sum_{i=1}^{N} m p_i\} \leq M p_{\mathrm{ave}}$.

B. Finite blocklength codes

In [8], [9], the authors analyzed the performance in the FBL regime. 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 [16]. 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 [16]:

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

where

$$C(\gamma) = \log(1+\gamma), \tag{2}$$

$$V(\gamma) = \frac{\gamma(\gamma+2)}{(\gamma+1)^2} \log_2^2 e \tag{3}$$

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

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

$$\varepsilon = \mathcal{P}\left(\gamma, r, m\right) \approx Q\left(\frac{C\left(\gamma\right) - r}{\sqrt{V(\gamma)/m}}\right). \tag{4}$$

In this paper, we apply the above approximations for investigating the finite blocklength performance of the considered downlink multiuser system. As these approximations have been shown to be accurate for a sufficiently large value of m [9], for simplicity we will employ them as the rate and error expressions in our analysis.

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 [17]. If we denote by Q the stationary queue length and by θ as the decay rate of the tail of the distribution of Q, the probability that Q exceeds a threshold q satisfies

$$P\left(Q \geq q\right) \approx \varsigma e^{-\theta q}, \tag{5}$$

where ς is probability of non-empty buffer. In addition, θ is called the QoS exponent, and is defined in [18] as

$$\lim_{q \to \infty} \frac{\log P\left(Q \geq q\right)}{q} = -\theta. \tag{6}$$

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

Following the queuing model in [17], [18], 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 [17], [18], under queuing constraints specified by the QoS exponent θ, the following relationship holds for the arrival and departure processes at the buffer

$$\Lambda_a\left(\theta\right) + \Lambda_c\left(-\theta\right) = 0, \tag{7}$$

where $\Lambda_p\left(\theta\right) = \lim_{t \to \infty} \log \mathbb{E}\{e^{\theta \sum_{k=1}^{t} p_k}\}$ denotes the asymptotic logarithmic moment generating function (LMGF) of the random process p_k.

In addition, the effective capacity is given in [17] as

$$R_{\mathrm{E}}(\theta) = -\frac{\Lambda_c\left(-\theta\right)}{\theta}, \tag{8}$$

and characterizes the maximum constant arrival rate that can be supported by the link with a random service process while satisfying (5). In this work, we adopt the effective capacity formulation to obtain the average throughput of the scenario with constant data arrivals.

III. FBL THROUGHPUT OF MULTI-USER NETWORKS

In this section, we study the optimal power allocation for the downlink multiuser network with constant data arrivals. 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 $\{p_i\}$, and is expressed as

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

where coding rate r_i, is given in (1), is a function of the transmit power p_i. First, we have the following Proposition.

Proposition 1. *In a system with target error probability $\varepsilon_i \geq 10^{-27}$ 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 and second order derivatives of r_i with respect to the SNR γ_i given as

$$\begin{aligned}
\tfrac{\partial r_i}{\partial \gamma_i} &= \tfrac{\log e}{1+\gamma_i} - \tfrac{A \log e}{\sqrt{\left(1-\frac{1}{(1+\gamma_i)^2}\right)}} \tfrac{1}{(1+\gamma_i)^3}, \\
\tfrac{\partial^2 r_i}{\partial \gamma_i^2} &= -\tfrac{\log e}{(1+\gamma_i)^2} + \tfrac{A_i \log e}{2\left(1-\frac{1}{(1+\gamma_i)^2}\right)^{\frac{3}{2}}} \tfrac{1}{(1+\gamma_i)^6} + \tfrac{A_i \log e}{\sqrt{\left(1-\frac{1}{(1+\gamma_i)^2}\right)}} \tfrac{3}{(1+\gamma_i)^4} \\
&= -\tfrac{\log e}{(1+\gamma_i)^2} + \tfrac{A_i \log e}{2\left((1+\gamma_i)^2-1\right)^{\frac{3}{2}}} \tfrac{1}{(1+\gamma_i)^3} + \tfrac{A_i \log e}{\sqrt{(1+\gamma_i)^2-1}} \tfrac{3}{(1+\gamma_i)^3} \\
&= \tfrac{\log e}{(1+\gamma_i)^3} \left\{ -\left(1+\gamma_i\right) + \tfrac{A_i}{2({\gamma_i}^2+2\gamma_i)^{\frac{3}{2}}} + \tfrac{3A_i}{\sqrt{{\gamma_i}^2+2\gamma_i}} \right\}.
\end{aligned} \tag{10}$$

When $\varepsilon_i \geq 0.5$, we have $A_i \leq 0$. Then $\frac{\partial^2 r_i}{\partial \gamma_i^2} < 0$. On the other hand, when $\varepsilon_i < 0.5$, $\frac{\partial^2 r_i}{\partial \gamma_i^2}$ is increasing in A_i and therefore decreasing in ε_i and m. For an extreme scenario where $m = 100, \varepsilon_i = 10^{-27}$, we have $A_i = 1.058$. Then, $\frac{\partial^2 r_i}{\partial \gamma_i^2} \leq 0$ if $\phi\left(\gamma_i\right) \leq 0$, where $\phi\left(\gamma_i\right) = -\left(1+\gamma_i\right) + \frac{A_i}{2({\gamma_i}^2+2\gamma_i)^{\frac{3}{2}}} + \frac{3A_i}{\sqrt{{\gamma_i}^2+2\gamma_i}}$. Obviously, $\phi\left(\gamma_i\right)$ is decreasing in γ_i for $A_i > 0$. In particular, we have $\phi\left(1\right) = -0.0164$ for $A_i = 1.058$. Hence, $\phi\left(\gamma_i\right) < 0$ 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., $m \geq 100, \varepsilon_i \geq 10^{-27}$. ■

It should be mentioned 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) $R_{i,\mathrm{E}}$ is concave in the transmit power p_i under the SNR constraint $\gamma_i \geq 0$ dB.*

Proof: According to (9), we have

$$\frac{\partial R_{\mathrm{E},i}}{\partial r_i} = \frac{m e^{-\theta m r_i}(1-\varepsilon_i)}{\mathbb{E}\left[e^{-\theta m r_i}(1-\varepsilon_i)+\varepsilon_i\right]} \geq 0 \tag{11}$$

$$\frac{\partial^2 R_{\mathrm{E},i}}{\partial r_i^2} = \frac{-\theta m^2 e^{-\theta m r_i}(1-\varepsilon_i)\varepsilon_i}{\mathbb{E}[e^{-\theta m r_i}(1-\varepsilon_i)+\varepsilon_i]^2} \leq 0 \tag{12}$$

Then, we have the first and second order derivatives of $R_{i,\mathrm{E}}$ with respect to p_i given as follows:

$$\frac{\partial R_{\mathrm{E},i}}{\partial p_i} = \frac{\partial R_{\mathrm{E},i}}{\partial r_i} \frac{\partial r_i}{\partial \gamma_i} \frac{\partial \gamma_i}{\partial p_i} = \frac{\partial R_{\mathrm{E},i}}{\partial r_i} \frac{\partial r_i}{\partial \gamma_i} \cdot \frac{z_i}{\sigma^2}, \tag{13}$$

$$\frac{\partial^2 R_{\mathrm{E},i}}{\partial p_i^2} = \frac{z_i}{\sigma^2} \underbrace{\frac{\partial^2 R_{\mathrm{E},i}}{\partial r_i^2}}_{\leq 0} \left(\frac{\partial r_i}{\partial p_i}\right)^2 + \frac{z_i}{\sigma^2} \underbrace{\frac{\partial R_{\mathrm{E},i}}{\partial r_i}}_{\geq 0} \underbrace{\frac{\partial^2 r_i}{\partial p_i^2}}_{\leq 0} \leq 0. \tag{14}$$

We have used in the above result that under the constraint guaranteeing $\gamma_i \geq 0 dB$, $\frac{\partial^2 r_i}{\partial p_i^2} \leq 0$ according to Proposition 1. Hence, $R_{i,\mathrm{E}}$ is concave in p_i when $\gamma_i \geq 0$ dB. ■

B. Optimal power allocation

Recall that we consider a downlink multiuser 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., $R_{\mathrm{E,sum}} = \frac{1}{M}\sum_{i=1}^{N} R_{i,\mathrm{E}}$, 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, $i = 1, ..., N$. while the equivalent requirement in terms of coding rate is $r_i \geq \mathcal{R}\left(\gamma_i, \varepsilon_i, m\right)$. 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 for 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 (time) and among users, while satisfying the average power constraint (averaged over time and users), i.e., $\mathbb{E}_{\mathbf{z}}\left\{\sum_{i=1}^{N} p_i\right\} \leq M p_{\mathrm{ave}}/m$, $\mathbf{z} = \{z_i, i = 1, ..., N\}$.

$$
\begin{aligned}
g_i = e^{-\theta R_{\mathrm{E},i}} = \mathbb{E}_{\mathbf{z}}\left\{e^{-\theta_i m r_i}(1-\varepsilon_i)+\varepsilon_i\right\} &= \mathbb{E}_{\mathbf{z}}\left\{ e^{-\frac{\theta_i}{\ln 2} m \left[\ln\left(1+\frac{z_i p_i(\mathbf{z})}{\sigma^2}\right) - A_i \sqrt{\left(1-\frac{1}{\left(1+\frac{z_i p_i(\mathbf{z})}{\sigma^2}\right)^2}\right)} + \frac{\ln m}{m}\right]} (1-\varepsilon_i)+\varepsilon_i \right\} \\
&= \int\limits_{z_1} \dots \int\limits_{z_N} \left\{ e^{-\frac{\theta_i}{\ln 2} m \left[\ln\left(1+\frac{z_i p_i(\mathbf{z})}{\sigma^2}\right) - A_i \sqrt{\left(1-\frac{1}{\left(1+\frac{z_i p_i(\mathbf{z})}{\sigma^2}\right)^2}\right)} + \frac{\ln m}{m}\right]} (1-\varepsilon_i)+\varepsilon_i \right\} f_{\mathrm{PDF}}\left(\mathbf{z}\right) \cdot d_{z_1} \cdots d_{z_N}.
\end{aligned} \tag{15}
$$

$$
\begin{aligned}
\frac{\partial g_i}{\partial p_i(\mathbf{z})} &= -\frac{\theta_i(1-\varepsilon_i)}{\ln 2} m \left(\frac{1}{1+\gamma_i} - \frac{A_i}{\sqrt{1-\frac{1}{(1+\gamma_i)^2}}} \frac{1}{(1+\gamma_i)^3} \right) \frac{z_i}{\sigma^2} e^{-\frac{\theta_i}{\ln 2} m \left[\ln(1+\gamma_i) - A_i \sqrt{1-\frac{1}{(1+\gamma_i)^2}} + \frac{\ln m}{m}\right]} \cdot f_{\mathrm{PDF}}\left(\mathbf{z}\right) \\
&= -\frac{z_i(1-\varepsilon_i)\eta_i m^{1-\eta_i}}{\sigma^2} \left(\frac{1}{1+\gamma_i} - \frac{A_i}{\sqrt{1-\frac{1}{(1+\gamma_i)^2}}} \frac{1}{(1+\gamma_i)^3} \right) (1+\gamma_i)^{-\eta_i m} e^{\eta_i m A_i \sqrt{1-\frac{1}{(1+\gamma_i)^2}}} \cdot f_{\mathrm{PDF}}\left(\mathbf{z}\right) \\
&= -\frac{z_i(1-\varepsilon_i)\eta_i m^{1-\eta_i}}{\sigma^2} \left(1-\frac{A_i}{\sqrt{a_i}}\left(1-a_i\right)\right) \left(1-a_i\right)^{\frac{\eta_i m+1}{2}} e^{\eta_i m A_i \sqrt{a_i}} \cdot f_{\mathrm{PDF}}\left(\mathbf{z}\right).
\end{aligned} \tag{16}
$$

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:

$$
\begin{aligned}
&\max_{\mathbf{p}\in\mathbf{\Omega}} \;\; R_{\mathrm{E,sum}} = \frac{1}{M}\sum\nolimits_{i=1}^{N} R_{\mathrm{E},i} \\
&s.t.: \;\; \mathbb{E}_{\mathbf{z}}\left\{\sum\nolimits_{i=1}^{N} p_i\right\} - Mp_{\mathrm{ave}}/m \le 0,
\end{aligned} \tag{17}
$$

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 feasible set of p_i, given by

$$
\Omega_i = \begin{cases} p_i \ge \frac{\gamma_{\mathrm{th},i}}{\sigma^2 z_i}, & \text{if } z_i \ge z_{\min}, \\ \;\; p_i = 0, & \text{if } z_i < z_{\min}, \end{cases} \tag{18}
$$

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

To solve Problem (17), we show its convexity as follows. According to (14), we have $\frac{\partial^2 R_{\mathrm{E},i}}{\partial p_i^2} \le 0$ for $\gamma_i = \frac{p_i z_i}{\sigma^2} \ge \gamma_{\mathrm{th},i} \ge 0$ dB.

Then, the Hessian matrix of the objective function in Problem (17) with respect to $\mathbf{p}$ is given by

$$
\begin{pmatrix} \frac{1}{M}\frac{\partial^2 R_{\mathrm{E},1}}{\partial p_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{M}\frac{\partial^2 R_{\mathrm{E},N}}{\partial p_N^2} \end{pmatrix}, \tag{19}
$$

which is negative semidefinite in the feasible 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 (17) is a convex optimization problem, which can be solved efficiently by optimization techniques [19].

In the following, we state the Lagrange dual function of the Problem (17). We introduce the Lagrange multiplier λ associated with the average power constraint. Then, the dual function is given by

$$
L = \frac{1}{M}\sum_{i=1}^{N} R_{\mathrm{E},i} - \lambda\, \mathbb{E}_{\mathbf{z}}\left\{\sum\nolimits_{i=1}^{N} p_i - Mp_{\mathrm{ave}}/m\right\}. \tag{20}
$$

By solving $\frac{\partial L}{\partial p_i(z)} = 0$, we can determine the dual optimal. Now, let us introduce g_i as defined in (15) at the top of the page. With this, we can express the effective capacity as $R_{\mathrm{E},i} = -\frac{1}{\theta}\ln g_i$. In (16) given also on the top of this page, we express the first order derivative of g_i with respect to p_i for a given channel realization $\mathbf{z}$. In (16), 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

$$\begin{aligned}\frac{\partial L}{\partial p_i(z)} &= \frac{1}{M}\frac{\partial R_{\mathrm{E},i}}{\partial g_i}\frac{\partial g_i}{\partial p_i} - \lambda \\ &= \varphi_i \left(1 - \frac{A_i}{\sqrt{a_i}}\left(1-a_i\right)\right)\left(1-a_i\right)^{\frac{\eta_i m+1}{2}} e^{\eta_i m A_i \sqrt{a_i}} - \lambda = 0,\end{aligned} \tag{21}$$

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

By solving the (21) within the feasible 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 (21). 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 (21) 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 (21).
   d)    According to (15), update $g_i$ based on $p_i$ and the PDF of $\mathbf{z}$.
   e)    According to (21), update $p_i$ based on the updated $g_i$ in Step 1-d. If there is no solution in the feasible 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 (21), determine the optimal power $p_i^*$ for this frame based on the instantaneous $\mathbf{z}$ as well as the obtained $\lambda^*$ and $g_i^\circ$.
```

IV. Numerical Results

In this section, we provide our numerical results. First the proposed optimal power allocation algorithm is compared with the equal power allocation scheme. 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., $f_{\mathrm{PDF}}(z_i) = e^{-z_i}, i = 1, ..., N$. Therefore, the joint PDF of all channels is $f_{\mathrm{PDF}}\left(\mathbf{z}\right) = e^{-\sum_1^N z_i}$. 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. Without being specifically noted, the default setup for the number of users is two, and both users have the same type of sources and the same QoS requirements.

To start with, we provide comparisons between the optimal algorithm (opt) versus the equal power allocation (equ). The comparison is with respect to the normalized sum throughputs as a function of error probabilities. We provide the results in Fig. 2. It can be observed that the optimal algorithm, significantly outperforms the equal power allocation. In addition, all normalized sum throughput curves are concave in the users' target error probability.

We continue the comparison in Fig. 3 where the QoS exponent of users are varied. First, all the throughput curves are decreasing in the QoS exponent θ. In addition, the optimal power allocation is observed to provide a higher throughput than the equal power allocation.

Finally, we study the impact of blocklength on the normalized sum throughput and provide the results in Fig. 4. 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 $\theta = 10^{-4}, \varepsilon = 10^{-4}$. Finally, it can be seen that the impact of θ on the throughput performance is more significant than that of ε.

V. CONCLUSION

In this paper, we have investigated optimal power allocation strategies in a downlink multiuser network in quasi-static fading channels with constant data arrivals. The normalized sum throughput is maximized in the presence of statistical QoS constraints and FBL codes. First, we have developed the FBL throughput model. Subsequently, based on the model we have formulated optimization problem and proved the convexity of the problem. Via numerical analysis, first we have showed that the proposed algorithm significantly outperforms the equal power allcoation algorithm. We have also observed that the normalized sum throughputs are concave in the users' error probability and are decreasing in the QoS exponent. Future work would be extending the current model to scenarios with random arrivals, e.g., discrete-time Markov model and continuous-time Markov model.

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*, accepted to appear.

[3] Q. Du, Y. Huang, P. Ren, and C. Zhang, "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.

[4] 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.

[5] T. Abrao, L. D. H. Sampaio, S. Yang, K. T. K. Cheung, P. J. E. Jeszensky and L. Hanzo, "Energy efficient OFDMA networks maintaining statistical QoS guarantees for delay-sensitive traffic," *IEEE Access*, vol. 4, pp. 774-791, 2016.

[6] X. Mi, L. Xiao, M. Zhao, X. Xu and J. Wang, "Statistical QoS-driven resource allocation and source adaptation for D2D communications underlaying OFDMA-based cellular networks," *IEEE Access*, vol. 5, pp. 3981-3999, 2017.

[7] J. Choi, "Effective capacity of NOMA and a suboptimal power control policy with delay QoS," *IEEE Transactions on Communications*, vol. 65, no. 4, pp. 1849-1858, April 2017.

[8] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Dispersion of gaussian channels,"in IEEE International Symposium on Information Theory (ISIT), pp. 2204–2208, IEEE, 2009.

[9] Y. Polyanskiy, H. 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.

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

[11] W. Yang, G. Durisi, T. Koch and Y. Polyanskiy "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, Jul. 2014.

[12] 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.

[13] S. Xu, T. H. Chang, S. C. Lin, C. Shen and G. Zhu, "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.

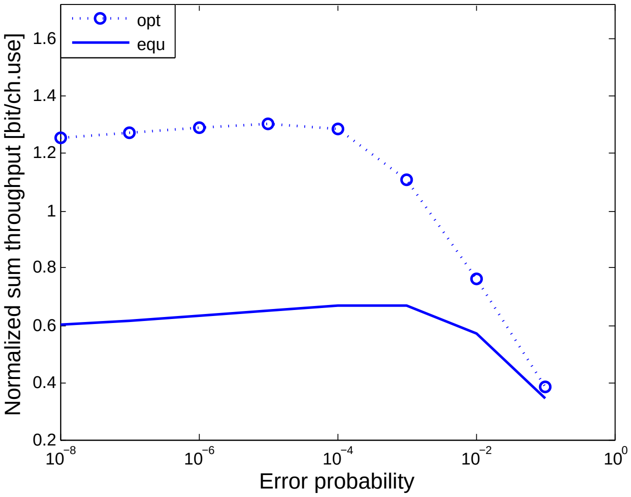

Fig. 2. Comparison between the optimal algorithm and the equal power allocation in a two-user scenario, while varying the error probability.

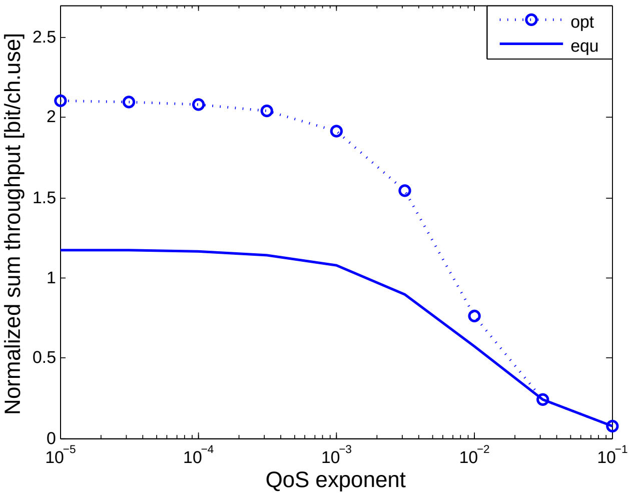

Fig. 3. Comparison between the proposed algorithm and the equal power allocation in a two-user scenario, while varying the QoS exponent of the two users.

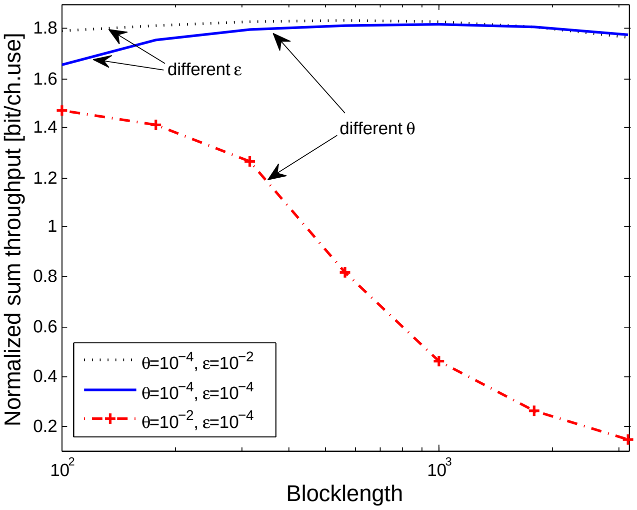

Fig. 4. The impact of blocklength on the throughput, while considering different setup of QoS exponents and error probabilities.

[14] M. C. Gursoy, "Throughput analysis of buffer-constrained wireless systems in the finite blocklength regime," *EURASIP J. Wireless Commun.*, vol. 2013:290, Dec. 2013.

[15] 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.

[16] 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.

[17] 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.

[18] 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

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