Throughput Optimization in Ultra-Reliable Low-Latency Communication with Short Packets

Apostolos Avranas, Marios Kountouris, Philippe Ciblat

Throughput Optimization in Ultra-Reliable Low-Latency Communication with Short Packets

Apostolos Avranas*, Marios Kountouris*, and Philippe Ciblat†
*Mathematical and Algorithmic Sciences Lab, Paris Research Center, Huawei France
†Télécom ParisTech, Université Paris-Saclay, F-75013 Paris, France
Emails: {apostolos.avranas,marios.kountouris}@huawei.com, philippe.ciblat@telecom-paristech.fr

***Abstract*—We consider an ultra-reliable low-latency communication (URLLC) system with short packets employing hybrid automatic repeat request (HARQ). Depending on the delay of HARQ feedback and retransmissions, the latency constraint can be either violated or fulfilled at the expense of power consumption. We focus on the energy-latency tradeoff and examine whether it is better to do one-shot transmission or use HARQ. We analyze the energy consumption for incremental redundancy (IR) HARQ and compare it with the no HARQ case. The analysis relies on closed-form expressions for the outage probability of IR-HARQ with variables both the blocklength and the power. Our results show that for a wide range of blocklength, when the feedback delay is more than half the latency constraint, it is beneficial in terms of energy to use one-shot transmission (i.e. no HARQ).**

I. Introduction

Future evolution of mobile communication systems (5G new radio) is giving rise to new uses of wireless communications in areas such as augmented and virtual reality (AR/VR), industrial control, automated transportation and robotics. 5G is envisaged to support mission-critical Internet-of-Things (IoT) applications and ultra-reliable low-latency communication (URLLC) scenarios with strict requirements in terms of latency (ranging from 1 ms and below to few milliseconds) and reliability (higher than 99.999%). This entails a fundamental paradigm shift from throughput-oriented system design towards an holistic design for guaranteed and reliable end-to-end latency.

Guaranteeing URLLC requirements is a challenging task even in simple settings as URLLC drives the system to new, unexplored operating regimes. The performance is constrained by challenging fundamental tradeoffs between delay, throughput, energy and error probability. The predominance of short messages, together with the need to reduce the packet duration, implies that small blocklength channel codes are also used. This results in a rate penalty term and transmission rates with non-zero error probability, revisiting key insights obtained via asymptotic information theoretic results. Recent progress has quantified the effect of finite blocklength, providing tight bounds and accurate normal approximation for the maximum coding rate to sustain the desired packet error probability (PEP) for a given packet size [1].

In order to compensate for the reliability loss introduced by short packets, highly reliable communications mechanisms creating diversity have to be carried out, such as hybrid automatic repeat request (HARQ). However the benefits of time diversity could be rather limited under stringent latency constraints. Moreover, the benefit of feedback-based retransmissions (even with error-free but delayed feedback) is questionable since each transmit packet is much smaller due to energy and latency constraints, thus more prone to errors. Additionally, energy considerations, in particular power consumption, are of cardinal importance in the design of URLLC systems, and there is an inherent power-latency tradeoff. A transmission can be successful (or its PEP may be kept unaltered) with minimum delay at the expense of additional or high power usage. In the short-packet regime, this interplay is more pronounced as latency is minimized when all packets are jointly encoded, whereas power is minimized when each packet is encoded separately. Note that power is the energy consumed over symbol period.

In this paper, we analyze the fundamental tradeoff between latency (in terms of feedback/retransmission delay) and average consumed energy in URLLC with incremental redundancy (IR)-HARQ. Considering that short packets have to be decoded with a certain PEP and latency, we give an answer whether it is beneficial to do one-shot transmission or split the packet into sub-codewords and use IR-HARQ. Prior work has considered the problem of throughput maximization by either adjusting the blocklength of each IR-HARQ round using the same power [2] or via rate refinement over retransmissions of equal-sized and constant energy packets [3]. Equal-sized and constant energy packets and rate maximization under a reliability constraint is considered in [4]. In [5], sphere packing is used for optimizing the blocklength of every transmission with equal power. In contrast to prior work, here we study the problem of average energy consumed minimization to guarantee both PEP and latency (URLLC) constraints by properly adapting both the blocklength and the power of each transmission. A key result of our paper is that one-shot transmission (no HARQ) should be used when the feedback delay is more than half the latency constraint for low and moderate blocklength.

II. System Model

We consider a point-to-point communication link, where the transmitter has to send B information bits within a certain predefined latency, which can be expressed by a certain predefined maximum number of channel uses, denoted by N_ℓ.

If no ARQ/HARQ mechanism is utilized, the packet of B bits is transmitted only once (one-shot transmission) and its maximum length is N_ℓ. When a retransmission strategy is employed, we consider hereafter IR-HARQ with M transmission rounds, i.e., $M-1$ retransmissions. Setting $M=1$, we recover the no-HARQ case as a special case of the retransmission scheme. We denote n_m with $m \in \{1,2,...,M\}$ the number of channel uses for the m-th transmission.

The IR-HARQ mechanism operates as follows: B information bits are encoded into a parent codeword of length $\sum_{m=1}^{M} n_m$ symbols. Then, the parent codeword is split into M fragments of codeword (sub-codewords), each of length n_m. The receiver requests transmission of the m-th sub-codeword only if it is unable to correctly decode the message using the previous $(m-1)$ fragments of the codeword. In that case, the receiver concatenates the first m fragments and attempts to jointly decode it. We assume that the receiver knows perfectly whether or not the message is correctly decoded (through CRC) and ACK/NACK is received error free. Every channel use (equivalently the symbol) requires a certain amount of time, therefore we measure time by the number of symbols contained in a time interval. The latency constraint is accounted for by translating it into a number of channel uses as follows: we have $\sum_{m=1}^{M} n_m \leq \mathrm{N}_\ell$. Penalty terms $D(\vec{n}_m)$, where $\vec{n}_m$ is the tuple $(n_1, n_2, ..., n_m) \in \mathbb{N}_+^m$ can easily be introduced at each m-th transmission in order to take into account the delay for the receiver to process/decode the m-th packet and send back acknowledgment (ACK/NACK). In this paper, we will focus on the simplified version where $D(\vec{n}_m) = 0$.

The channel is considered to be static within the whole HARQ mechanism, i.e., there is only one channel coefficient value for all the retransmissions associated with the same bits. This is a relevant model for short-length packet communication and IoT applications. Indeed, for a system operating at carrier frequency $f_c = 2.5$ GHz, for a channel coherence time $T_c = 1$ ms (so equal to the URLLC latency constraint, i.e., the maximum duration of all the retransmissions associated with the same bits), the maximal receiver speed to satisfy the static assumption is $v = cB_d/f_c \approx 180$ km/h, where $B_d = 0.423/T_c$ [6, (8.20)] is the Doppler spread and c is the speed of light. So for any device whose speed is smaller than 180 km/h, the channel is static during the HARQ process. This is a relatively high speed for most mission-critical IoT or tactile Internet applications. Therefore, our communication scenario consists of a point-to-point link with additive white Gaussian noise (AWGN). Specifically, in m-th round, the fragment (sub-codeword) $c_m \in \mathbb{C}^{n_m}$ is received with power $P_m = \frac{||c_m||^2}{n_m}$ and distorted by an additive white circularly-symmetric complex Gaussian random process with zero mean and unit variance. The power allocation applied during the first m rounds is denoted by $\vec{P}_m = (P_1, ...P_m) \in \mathbb{R}_+^m$.

III. Problem Statement and Preliminaries

The objective of this paper is to find out the HARQ mechanism maximizing the throughput by tuning the number of transmitted information bits B, the number of rounds M, and the blocklength-power allocation , i.e. $(\vec{n}_M, \vec{P}_M)$, given a maximum packet error probability $\varepsilon_{\mathrm{rel}}$, a latency constraint N_ℓ (due to URLLC requirements) and an energy budget E_t.

Before going further, we need to characterize the probability of error in the m-th round of the HARQ mechanism as a function of $(\vec{n}_m, \vec{P}_m)$. To derive this packet error probability, we resort to the results for the non-asymptotic (finite-blocklength) regime [1] since the packets we manipulate may be short.

In IR-HARQ with m transmission, the packet error probability or equivalently the outage probability, denoted by ϵ_m, can be expressed as $\epsilon_m = \mathbb{P}\left(\bigcap_{i=1}^{m} \Omega_i\right)$ where Ω_i is the event corresponding to "the concatenation of the first i fragments of the parent codeword, with length $\vec{n}_i$ and energy per symbol $\vec{P}_i$, is not correctly decoded when optimal coding is employed".

When an *infinite* blocklength is assumed, an error occurs if the mutual information is below a threshold and for IR-HARQ, it can easily be seen that for $i < j$ we have $\Omega_i \subseteq \Omega_j$ [7], [8], which leads to $\epsilon_m = \mathbb{P}(\Omega_m)$. In contrast, when a *finite* blocklength (or a real coding scheme) is assumed, the above statement does not hold anymore and an exact expression for ϵ_m seems intractable. Therefore, in the majority of prior work on HARQ (see [2], [8], [9] and references therein), the exact outage probability ϵ_m is replaced with the simplified ε_m defined as $\varepsilon_m = \mathbb{P}(\Omega_m)$, since ε_m and ϵ_m are numerically close. Note that for $m = 1$ the definitions coincide and $\varepsilon_1 = \epsilon_1 = \mathbb{P}(\Omega_1)$. In the remainder of the paper, we assume that this approximation is valid. Then, ε_m can be upper bounded [1, Lemma 14 and Theorem 29] and also lower bounded as in [9] by employing the $\kappa\beta$-bounds proposed in [1]. Both bounds have the same first two dominant terms and the error probability is approximately given by

$$\varepsilon_m \approx Q\left(\frac{\sum_{i=1}^{m} n_i \ln(1+P_i) - B\ln 2}{\sqrt{\sum_{i=1}^{m} \frac{n_i P_i (P_i+2)}{(P_i+1)^2}}}\right) \tag{1}$$

where $Q(x)$ is the complementary Gaussian cumulative distribution function. For the sake of clarity, we may show the dependency on the variables, i.e., $\varepsilon_m(\vec{n}_m, \vec{P}_m)$ instead of ε_m, whenever needed.

Notice that some works have tried to approximate more accurately the term ϵ_m or ε_m [10]–[13]. For instance, in [10], the authors provide more involved expressions for ϵ_m, but the feedback scheme considered is different from ours; the feedback time index in [10] is not predefined (it is a random variable) and is adapted online. In [11], [12] justifications for the approximation $\epsilon_m \approx \varepsilon_m$ when using non-binary LDPC codes or tail-biting convolutional code can be found. In [13], the authors used saddlepoint approximation to find a tight approximation of ε_m but provide closed-form expression only

for binary erasure channels (BEC). Therefore, we consider that using the Gaussian approximation expressed by (1) provides a relavant tradeoff between analytical tractability and tightness of the approximations.

IV. Problem statment and its solution

We remind we carry out an IR-HARQ to send B information bits and our goal is to allocate the blocklength and power of the packet sent in every round in order to maximize the throughput. The throughput is defined as the average ratio of the successfully decoded bits divided by the number of spent symbols. The thoughput can be derived thanks to the renewal theory where the expecting delay is $\sum_{m=1}^{M} n_m \varepsilon_{m-1}$ and the expected reward is $B(1-\varepsilon_M)$. Consequently, our goal can be translated into the following optimization problem.

Problem 1: General problem

$$\max_{B,M,\vec{n}_M,\vec{P}_M} \quad \frac{B(1-\varepsilon_M)}{\sum_{m=1}^{M} n_m \varepsilon_{m-1}} \tag{2}$$

$$\text{s.t.} \quad \sum_{m=1}^{M} n_m \leq \mathrm{N}_\ell \tag{3}$$

$$\varepsilon_M \leq \varepsilon_{\mathrm{rel}} \tag{4}$$

$$\sum_{m=1}^{M} n_m P_m \varepsilon_{m-1} \leq \mathrm{E}_t \tag{5}$$

$$M \leq \mathrm{M_r} \tag{6}$$

Solving the general problem is intractable. Therefore we consider a simpler one by modifying slightly the objective function. To that end, we force the numerator to be equal to $B(1-\varepsilon_{\mathrm{rel}})$ which means we force the constraint given by (4) to be active. This leads to the following optimization problem

Problem 2:

$$\max_{B,M,\vec{n}_M,\vec{P}_M} \quad \frac{B(1-\varepsilon_{\mathrm{rel}})}{\sum_{m=1}^{M} n_m \varepsilon_{m-1}} \tag{7}$$

$$\text{s.t.} \quad \sum_{m=1}^{M} n_m \leq \mathrm{N}_\ell \tag{8}$$

$$\varepsilon_M \leq \varepsilon_{\mathrm{rel}} \tag{9}$$

$$\sum_{m=1}^{M} n_m P_m \varepsilon_{m-1} \leq \mathrm{E}_t \tag{10}$$

$$M \leq \mathrm{M_r} \tag{11}$$

In Lemma 1 it is proven that the solution of **Problem** 2 achieves almost the same performance as those of the original **Problem** 1.

Lemma 1: Let $(B^{mod}, M^{mod}, \vec{n}_M^{mod}, \vec{P}_M^{mod})$ be the solution of **Problem 2**. It gives the value Th for the throughput according to (2). Let $Th^\star$ be the highest value of the throughput given by the solution of **Problem 1**. Then $(B^{mod}, M^{mod}, \vec{n}_M^{mod}, \vec{P}_M^{mod})$ is a feasible point of **Problem** 1 and it holds that $Th \leq Th^\star \leq \frac{Th}{1-\varepsilon_{\mathrm{rel}}}$.

Proof: The constraints of the two problems are the same, therefore they share the same feasible domain which we denote $\mathbb{D}$. So, $(B^{mod}, M^{mod}, \vec{n}_M^{mod}, \vec{P}_M^{mod})$ is a feasible point of **Problem** 1. Since $Th^\star$ is the optimal value and Th just a feasible one, $Th \leq Th^\star$. Furthermore, the solution of **Problem** 2 guarantees that for every point in $\mathbb{D}$ it holds $\frac{B}{\sum_{m=1}^{M} n_m \varepsilon_{m-1}} \leq \frac{Th}{1-\varepsilon_{\mathrm{rel}}}$. Therefore if $x^\star \in \mathbb{D}$ is the optimal point of **Problem** 1 and gives an error probability of $\varepsilon_M^\star$ then $\frac{Th^\star}{(1-\varepsilon_M^\star)} \leq \frac{Th}{1-\varepsilon_{\mathrm{rel}}}$ from which we can easily derive $Th^\star \leq \frac{Th}{1-\varepsilon_{\mathrm{rel}}}$. ∎

We propose to do the optimization over B via a 1-D grid-search. Consequently, **Problem** 2 can be still simplified and leads to the following **Problem** 3.

Problem 3:

$$\min_{M,\vec{n}_M,\vec{P}_M} \quad \sum_{m=1}^{M} n_m \varepsilon_{m-1} \tag{12}$$

$$\text{s.t.} \quad \sum_{m=1}^{M} n_m \leq \mathrm{N}_\ell \tag{13}$$

$$\varepsilon_M \leq \varepsilon_{\mathrm{rel}} \tag{14}$$

$$\sum_{m=1}^{M} n_m P_m \varepsilon_{m-1} \leq \mathrm{E_t} \tag{15}$$

$$M \leq \mathrm{M_r} \tag{16}$$

The rest of this Section is devoted to the resolution of this **Problem** 3. We will see that it can be solved iteratively through a dynamic programing approach.

First of all, we introduce the states at the end of m-th round:

$$S_1 = (N_1, \varepsilon_1)$$

$$S_m = (N_m, \varepsilon_m, E_m, V_m), m \in \{2, 3, ...\}$$

where $\forall m \in \mathbb{N}^\star$: $N_m = \sum_{i=1}^{m} n_i$, $E_m = \sum_{i=1}^{m} n_i P_i \varepsilon_{i-1}$ and $V_m = \sum_{i=1}^{m} n_i (1 - \frac{1}{(1+P_i)^2})$. We have $V_m < N_m \leq \mathrm{N}_\ell$. Let $\mathbb{S}_M$ be the set of feasible final states. By feasibility, we mean that a state S_M in $\mathbb{S}_M$ satisfies the constraints in **Problem** 3. We have $\mathbb{S}_M \subset \{1, 2, .., \mathrm{N}_\ell\} \times [0, \varepsilon_{\mathrm{rel}}] \times [0, \mathrm{E_t}] \times [0, \mathrm{N}_\ell] \quad \forall M \in \{1, 2, .., \mathrm{M_r}\}$. Our objective is to find the optimal sequence/path of states minimizing (12) which will provide the optimal vector $(\vec{n}_M, \vec{P}_M)$ solving **Problem** 3. Indeed, each solution (and so the optimal one) represented by a path $(\vec{n}_M, \vec{P}_M)$ leads to an $S_M \in \mathbb{S}_M$.

The first three variables of the states S_m were chosen in order to be able to check the constraints (13-15). The dispersion variable V_m was added such that the sequence of S_m is a Markov chain since the description of S_m then depends only on the previous state S_{m-1} and the variables n_m and P_m which constitute the branch between S_{m-1} and S_m. The functions connecting these states can be easily found and let them be: $S_m = f_S(S_{m-1}, n_m, P_m)$, $S_{m-1} = f_S^{-1}(S_m, n_m, P_m)$.

For sake of simplicity, we introduce the following notation "$\min_{X|Y} f(X)$" which stands for "minimize $f(\cdot)$ over the vari-

ables X given constraints Y". Now the **Problem** 3 can be seen as the solution of:

$$\min_{M, S_M \in \mathbb{S}_M | M \in \{1,..,\mathrm{M_r}\}} \Big\{ \min_{\vec{n}_M, \vec{P}_M | S_M} \quad \sum_{m=1}^{M} n_m \varepsilon_{m-1} \Big\}$$

over the constraints (13)-(15)

As roughly-mentioned previously, we perform the outer minimization by exhaustive search (even though, we will prove below that only a few states $S \in \mathbb{S}_M$ are good candidates). On the other hand, the inner minimization is solved dynamically since it can be written as:

$$\min_{n_M, P_M | S_M} \Big\{ \min_{\vec{n}_{M-1}, \vec{P}_{M-1} | S_M, n_M, P_M} \{ n_M \varepsilon_{M-1} + \sum_{m=1}^{M-1} n_m \varepsilon_{m-1} \} \Big\}$$

The inner minimization is done under fixed (S_M, n_M, P_M) which allows the first term $n_M \varepsilon_{M-1}$ to get out as a constant since this term can be expressed as a function of only those fixed variables. Moreover, $S_{M-1} = f_S^{-1}(S_M, n_M, P_M)$ is fixed which can be confirmed that it is an equivalent to (S_M, n_M, P_M) constraint when minimizing the second term. So, we have

$$\min_{\vec{n}_M, \vec{P}_M | S_M} \{ \sum_{m=1}^{M} n_m \varepsilon_{m-1} \} = \min_{n_M, P_M | S_M} \Big\{ K(S_M, n_M, P_M) + \min_{\vec{n}_{M-1}, \vec{P}_{M-1} | S_{M-1} = f_S^{-1}(n_M, P_M, S_M)} \{ \sum_{m=1}^{M-1} n_m \varepsilon_{m-1} \} \Big\}$$

The above formula can be proven for every $m \in \{1, ..., M\}$ which allows to apply the dynamic programming approach. Specifically to find the optimal solution for the state S_m it is sufficient to know the optimal solution of every S_{m-1} connected to it through a branch (n_m, P_m). Therefore we can start by straightforwardly computing the values for each feasible S_1 and afterwards in every m iteration of the dynamic programming algorithm we compute the optimal solution for S_m by using the corresponding S_{m-1}.

Finally the optimal solution of **Problem** 3 has some characteristics which reduce the number of states to test.

Lemma 2: When M grows, feasible points of **Problem** 3 with better values of the objective function (12) appear. Therefore the optimal solution satisfies (16) with equality, i.e. $M^\star = \mathrm{M_r}$.

Proof: In [14, Appendix C], it is proven that if the last, i.e. M-th, packet with (n_M, P_M) blocklength and power is properly split into two packets with (n'_M, P_M) and $(n_{M+1} = n_M - n'_M, P_M)$ then the average energy is decreased. The same splitting straightforwardly can be shown to decrease our objective (12) and therefore this new configuration with an extra round gives better result while satisfying the constraints. Therefore more rounds lead to better performace. ■

Lemma 3: Let $(M^\star, \vec{n}^\star_{M^\star}, \vec{P}^\star_{M^\star})$ be the optimal point of **Problem** 3. We remind that $M^\star = M_r$ due to Lemma 2. Let $\varepsilon^\star_m = \varepsilon(m, \vec{n}^\star_m, \vec{P}^\star_m)$ where $\vec{n}^\star_m$ (resp. $\vec{P}^\star_m$) is an extracting vector from the m-th first components of $\vec{n}^\star_{M_r}$ (resp. $\vec{P}^\star_{M_r}$), be the error probability at every round $m < \mathrm{M_r}$. We have $\varepsilon^\star_m > \varepsilon_{rel}$ and finally at round $\mathrm{M_r}$ we have $\varepsilon^\star_{\mathrm{M_r}} \leq \varepsilon_{\mathrm{rel}} < \varepsilon(\vec{n}^\star_{\mathrm{M_r}-1}, n^\star_{\mathrm{M_r}} - 1, \vec{P}^\star_{\mathrm{M_r}})$.

Proof: Assume that for $m_0 < \mathrm{M_r}$ we have $\varepsilon^\star_{m_0} < \varepsilon_{\mathrm{rel}}$. Then the point $(m_0, \vec{n}^\star_{m_0}, \vec{P}^\star_{m_0})$ is a better than the optimal point which leads to contradiction. Furthermore, to prove $\varepsilon^\star_{\mathrm{M_r}} \leq \varepsilon_{\mathrm{rel}} < \varepsilon(\vec{n}^\star_{\mathrm{M_r}-1}, n^\star_{\mathrm{M_r}} - 1, \vec{P}^\star_{\mathrm{M_r}})$ is fairly simple since the first inequality is the reliability constraint and the second cannot be violated since otherwise the point $(\vec{n}^\star_{\mathrm{M_r}-1}, n^\star_{\mathrm{M_r}} - 1, \vec{P}^\star_{\mathrm{M_r}})$ is a better than the optimal solution which again leads to a contradiction. ■

As $\varepsilon^\star_{\mathrm{M_r}} \leq \varepsilon_{\mathrm{rel}} < \varepsilon(\vec{n}^\star_{\mathrm{M_r}-1}, n^\star_{\mathrm{M_r}} - 1, \vec{P}^\star_{\mathrm{M_r}})$, we can conjecture that $\varepsilon^\star_{\mathrm{M_r}} \approx \varepsilon_{\mathrm{rel}}$ since the last round will enable to satisfy the constraints but not going to far away from it because it will be costly in throughput (if n_{M_r} is selected too high) or in energy (if P_{M_r} is selected too high). Therefore we will also that $E^\star_{M_r} \approx \mathrm{E_t}$ where $E^\star_{M_r}$ is the energy consumed by the optimal solution of **Problem** 3.

V. Implementation

The dynamic programming algorithm needs in practice the variables of the states to take discrete values. Specifically:

- N_m has already a discrete form since it is an integer inside the interval $[0, N]$ but for accelerating the simulation it can be also quantized using bigger than one symbol step size. Let $\mathtt{N}$ be the set of the discrete values that N_m can take.
- ε_m is real and from lemma 3 we know $\varepsilon_m \in [0, \varepsilon_{\mathrm{rel}}]$. It turns out that if instead of ε_m the use of the equivalent (due to $Q^{-1}()$ being a one-to-one mapping) variable $c_m := Q^{-1}(\varepsilon_m)$ is employed, more accurate results are yielded. If we assume only realistic error probabilities of value lower than 0.5 then $c_m \in [0, Q^{-1}(\varepsilon_{\mathrm{rel}})]$. Let $\mathtt{C} \subset [0, Q^{-1}(\varepsilon_{\mathrm{rel}})]$ be the set of the discrete values that the dynamic algorithm allows c_m to take.
- E_m is real and $E_m \in [0, \mathrm{E}_t]$. After quantization let $\mathtt{E}$ be the set of the discrete values E_m can take.
- V_m is real and $V_m \in (0, N_m) \subset (0, \mathrm{N}_\ell)$. After quantization let $\mathtt{V}$ be the set of the discrete values V_m can take.

The dynamic algorithm consists of two stages: the first one to compute for the feasible states their performance and the second one to search over those states to find the optimal solution. The complexity is governed by the first stage and is equal to the number of iterations the dynamic algorithm takes, multiplied by the number of states examined per iteration, multiplied by the number of branches departing from every state. In our implementation, we compute the branch (n_{m+1}, P_{m+1}) departing from a state S_m through fixing the variables N_{m+1} and E_{m+1} of the arriving state S_{m+1} and subsequently we acquire the feasible ε_{m+1} and V_{m+1}. Therefore the overall complexity is $O(\mathrm{M_r} \cdot |\mathtt{N}||\mathtt{E}||\mathtt{C}||\mathtt{V}| \cdot |\mathtt{N}||\mathtt{E}|)$.

The above complexity displays a slow algorithm but in reality it can be faster by remarking that most of the times all of the paths ending up at states with the same (N_m, c_m, E_m) which the algorithm considers, present dispersions V_m within

a small range of values. Therefore if a reasonable resolution of the discrete set $\mathtt{V}$ is considered so as no significant approximation errors to be introduced, then the number of feasible states with same (N_m, c_m, E_m) and different V_m turns out to be rather small (many times just one value). Therefore the variable $|\mathtt{V}|$ can be thought as constant which in our simulations never exceeds 10.

VI. Numerical results and discussion

In this section, we provide numerical results to illustrate the behavior of the system. First we inspect how the error probability affects the throughput. Specifically we solve **Problem** 2 with reliability constraint (9) to be equality and we let the achieved $\varepsilon_{\mathrm{rel}}$ to take different values. This procedure actually requires only one run of the dynamic algorithm because after the computation of the performance of each state we can restrict the research of the minimum only between the states with the given ε_m. In that way Fig. 1 is obtained.

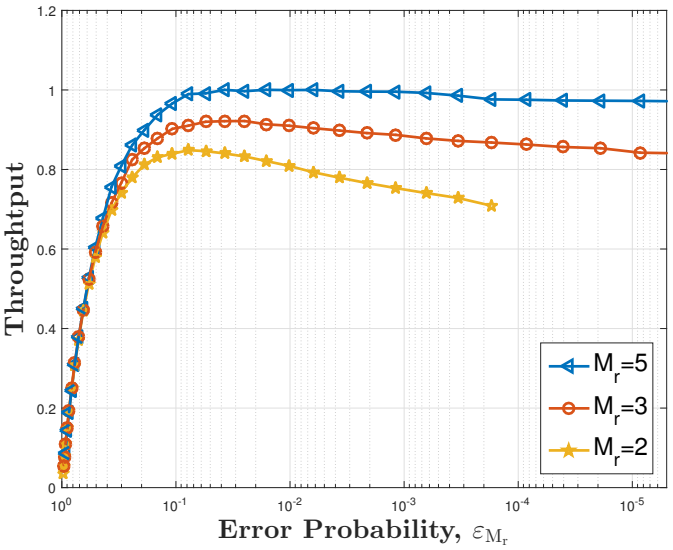

Fig. 1. Throughput Versus Attained Error Probability when $N = 400$, $\mathrm{E_t} = 267$, $B = 32$Bytes

As mentioned from Lemma 2 and confirmed by Fig. 1, the more rounds the higher throughput and we see also the more robust is the throughput performance when pushing for higher reliability. Moreover since we remain in the finite blocklength regime it is impossible to attain $\varepsilon_M \to 0$ given finite energy budget. Therefore there exists a certain value that the reliability cannot go beyond. This is the reason the curve of $\mathrm{M_r} = 2$ in Fig. 1 stops at a certain error probability. The other two curves also stop after a certain error probability which is much smaller preventing to be depicted in the figure. Finally, we remark, as mentioned in [3], that there is a value of error probability which maximizes the throughput but it is fairly poor (close to 0.1). So in our case, we will achieve higher reliability at the expense of the throughput performance (since our operating point does not correspond to the optimal one for throughput in this figure)

Now we analyze the influence of the number of used symbols on the throughput. In order to obtain Fig. 2, we force equality in the latency constraint (8) of **Problem** 2, and the error probability is treated as for Fig. 1.

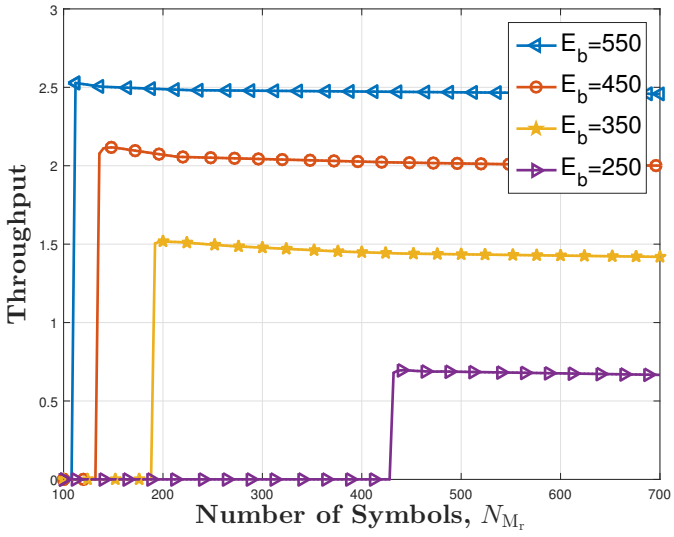

Fig. 2. Throughput Versus Used number of symbols when $\varepsilon_{\mathrm{rel}} = 10^{-5}$, $B = 32$Bytes,$\mathrm{M_r} = 3$

When the available number of symbols are inadequate (too weak), no feasible solution appears and the throughput vanishes. Interestingly as $N_{\mathrm{M_r}}$ grows beyond a certain threshold, only a slight increase of the throughput is obtained, followed by a slow decrease. This means that it is not always beneficial for throughput to use the whole available blocklength. Asymptotically if $N_{\mathrm{M_r}} \to \infty$ then for some $m \in \{1, \cdots \mathrm{M_r}\}$ it should $n_m \to \infty$ which will result to vanishing the throughput.

For Fig. 3, we plot the throughput versus the energy budget. In practice, we do not force equality in the energy constraint (10), since as stated previously the optimal solution consumes by default (almost) all the available energy. In our simulations, we set the minimum possible blocklength for the first HARQ round to be $N_{1,min} \geq 100$ (which was set likewise so as Polyanskiy's formula (1) to remain accurate). Consequently the throughput cannot exceed the value $\frac{B}{N_{1,min}}$ which represent the fictional case of only one packet sent with minimum blocklength and achieving perfect reliability. This lower bound is closely approached as the available energy grows to finally be enough so that only one transmission fulfills the constraints. Further increase of the energy is wasteful. Finally, Fig. 3 confirms again (as in Fig. 2) that when we are beyond a certain threshold, any additionally increase of the available blocklength is useless.

For Fig. 4, we depict the throughput (via a contour plot) versus the available average energy $\mathrm{E_t}$ and the information bits to transmit B. There is an upper left area where there are no feasible points. Keeping a constant $\mathrm{E_t}$ by moving vertically on Fig. 4, we see that the throughput is a unimodal function over B and there is a specific value of B achieving optimality.

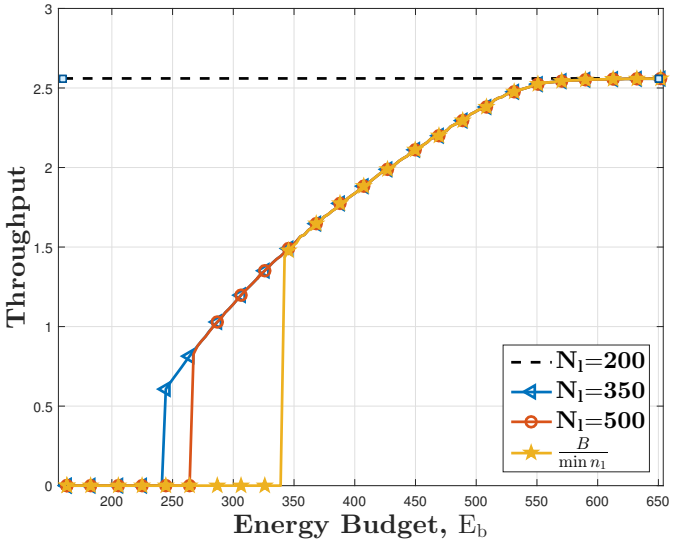

Fig. 3. Throughput Versus Used Energy Budget when $\varepsilon_{\mathrm{rel}} = 10^{-5}$, $B =$ 32Bytes, $\mathrm{M_r} = 3$

This also agrees with [3] when a simple ARQ scheme with no URLLC constraints was employed.

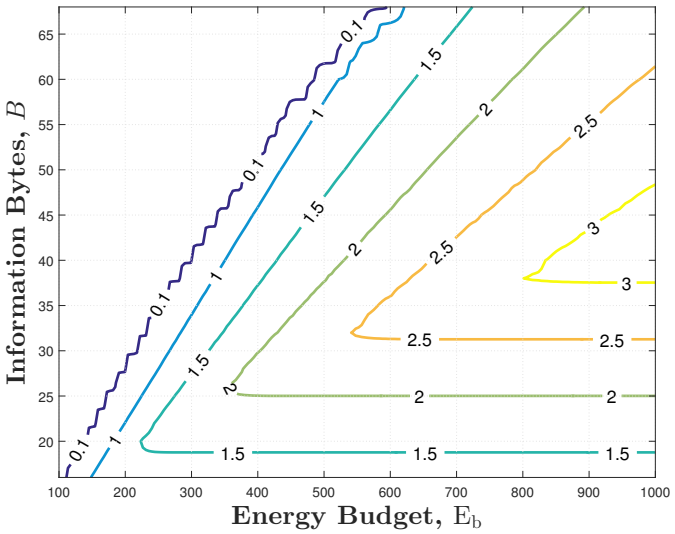

Fig. 4. Throughput Versus Energy and Information Bits when $\varepsilon_{\mathrm{rel}} = 10^{-5}$, $\mathrm{N}_\ell = 600$, $\mathrm{M_r} = 3$

VII. Conclusion

TO BE DONE

References

[1] Y. Polyanskiy, "Channel coding: Non-asymptotic fundamental limits," Ph.D. dissertation, Princeton University, Nov. 2010.

[2] B. Makki, T. Svensson, and M. Zorzi, "Finite block-length analysis of the incremental redundancy HARQ," *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 529–532, Oct. 2014.

[3] P. Wu and N. Jindal, "Coding versus ARQ in fading channels: How reliable should the PHY be?" *IEEE Trans. on Commun.*, vol. 59, no. 12, pp. 3363–3374, Dec. 2011.

[4] S. H. Kim, D. K. Sung, and T. Le-Ngoc, "Performance analysis of incremental redundancy type hybrid ARQ for finite-length packets in AWGN channel," in *Proc. IEEE Global Commun. Conf. (Globecom)*, Atlanta, GA, USA, Dec. 2013.

[5] A. R. Williamson, T. Chen, and R. D. Wesel, "A rate-compatible sphere-packing analysis of feedback coding with limited retransmissions," in *Proc. IEEE ISIT*, Cambridge, MA, USA, July 2012.

[6] J. Gibson, *The Communications Handbook*. CRC press, 2002.

[7] G. Caire and D. Tuninetti, "The throughput of hybrid ARQ protocols for the Gaussian collision channel," *IEEE Trans. on Inf. Theory*, vol. 47, no. 5, pp. 1971–1988, July 2001.

[8] C. L. Martret, A. Leduc, S. Marcille, and P. Ciblat, "Analytical performance derivation of hybrid ARQ schemes at IP layer," *IEEE Trans. on Commun.*, vol. 60, no. 5, pp. 1305–1314, May 2012.

[9] J. Park and D. Park, "A new power allocation method for parallel AWGN channels in the finite block length regime," *IEEE Wireless Commun. Lett.*, vol. 16, no. 9, pp. 1392–1395, Sept. 2012.

[10] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Feedback in the non-asymptotic regime," *IEEE Trans. on Inf. Theory*, vol. 57, no. 8, pp. 4903–4925, Aug. 2011.

[11] K. Vakilinia, S. V. S. Ranganathan, D. Divsalar, and R. D. Wesel, "Optimizing transmission lengths for limited feedback with nonbinary LDPC examples," *IEEE Trans. on Commun.*, vol. 564, no. 6, pp. 2245–2257, June 2016.

[12] H. Wang, N. Wong, A. M. Baldauf, C. K. Bachelor, S. V. S. Ranganathan, D. Divsalar, and R. D. Wesel, "An information density approach to analyzing and optimizing incremental redundancy with feedback," in *Proc. IEEE Int. Symp. Inf. Theory*, Aachen, Germany, June 2017.

[13] A. Martinez and A. G. i Fàbregas, "Saddlepoint approximation of random-coding bounds," in *Proc. Inf. Theory Applicat. Workshop (ITA)*, CA, USA, Aug. 2011.

[14] A. Avranas, M. Kountouris, and P. Ciblat, "Energy-latency tradeoff in ultra-reliable low-latency communication with retransmissions," *IEEE J. Sel. Areas Commun.*, Oct. 2018. [Online]. Available: https://arxiv.org/abs/1805.01332