Throughput Maximization and IR-HARQ Optimization for URLLC Traffic in 5G Systems

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*—Emerging 5G networks will need to efficiently support ultra-reliable, low-latency communications (URLLC) services, which require extremely low latency (msec order) with very high reliability (99.999%). We consider a URLLC system with short packets and incremental redundancy hybrid automatic repeat request (IR-HARQ). We aim at maximizing the throughput by optimally tuning the IR-HARQ mechanism subject to URLLC constraints and a fixed energy budget. We propose a dynamic programming algorithm for solving the throughput maximization problem in the finite blocklength regime and assess its performance numerically.**

I. Introduction

Next generation (5G) communication systems are designed to efficiently support new applications and use cases in areas such as augmented and virtual reality (AR/VR), industrial automation, intelligent transportation, and robotics. A key feature will be the satisfaction of services requiring highly demanding end-to-end latency (few milliseconds) and reliability levels in terms of packet delivery success ($> 99.999\%$). In this new paradigm, which is termed ultra-reliable, low-latency communications (URLLC), a careful reexamination of the throughput-oriented system design and a holistic system view are mandatory in order to meet the stringent reliability and latency requirements.

Providing URLLC guarantees even in simple settings leads to new, unexplored operating regimes. Reducing drastically the latency imposes the use of very short messages and timeslots (mini slots), which results in small packet duration and faster decoding. Communicating with short packets implies in turn using small blocklength channel codes, which make the widely used asymptotic information theoretic results not applicable. Transmission rates with non-zero error probabilities kick in and relevant bounds quantifying the effect of finite blocklength are required. An accurate and convenient normal approximation combining the maximum coding rate with the packet error probability under a given packet size has recently been proposed in [1].

Reducing the packet size due to latency constraints has in principle a negative impact on the reliability. A way to compensate for this is to exploit some form of diversity. In our paper, we employ incremental redundancy hybrid automatic repeat request (IR-HARQ) as a means to introduce time diversity. This retransmission scheme with feedback secures a low probability of packet transmission failure but on the other hand takes a toll on the required latency. To fix this, the amount of information contained in the transmitted packets has to be reduced and replaced by redundant information, which leads to lower throughput. Another important factor that should not be disregarded is the energy consumption. At the expense of additional power it is relatively easy to shorten the delay without any compromise of the throughput, but characterizing the interplay between latency, reliability, and throughput under a fixed energy budget is a very challenging task.

In this work, we consider the problem of throughput maximization in the finite blocklength regime subject to URLLC constraints and a maximum energy budget. We show how to optimize the parameters of the IR-HARQ mechanism, namely the number of information bits, the number of transmission rounds, and the blocklength-power allocation, in order to maximize the achieved throughput. In our previous work [2], we solved an optimization problem of similar configuration but with the objective of energy minimization. Throughput maximization is considered in [3] by solely optimizing the blocklength of a two round IR-HARQ and in [4] using rate refinement over possibly infinite number of retransmissions of equal size and power. Imposing as well a reliability constraint, [5] performs rate maximization. Jointly adjusting power and blocklength, similarly to our work but with only one transmission, is studied in [6] with the objective of minimizing the energy of a FIFO scheduler. A dynamic programming solution to jointly optimize rates and powers, as in our paper, is proposed in [7]. However, the system model is different from ours as type-I HARQ is used therein and cross layer optimization of buffer's and scheduler's behavior is performed. Finally, throughput maximization for IR-HARQ problem is considered in [8] assuming infinitely large blocklength and performing length adaptation. Differently from prior work, in this paper, we solve the throughput maximization problem in the finite blocklength regime subject to latency, reliability and maximum energy constraints. A dynamic programming algorithm is proposed for solving the non-convex optimization problem, which allows us to properly adapting the operating parameters of the IR-HARQ mechanism. Interestingly, we see that the solution (system operating points) to the throughput maximization problem is very similar to that of the energy minimization problem [2].

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 we expressed by a certain predefined maximum number of channel uses, denoted by N_ℓ. If no retransmission 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 IR-HARQ mechanism, i.e., there is only one channel coefficient value for all retransmissions associated with the same bits. This is a relevant model for short-length packet communication and IoT applications, which makes that our communication scenario consists of a point-to-point link with additive white Gaussian noise (AWGN). Specifically, in the 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 optimize the IR-HARQ mechanism and maximize throughput by tuning the number of transmitted information bits B, the number of transmission 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 maximum latency constraint N_ℓ (URLLC requirements), fixed energy budget E_t, and maximum number of transmission rounds M_r.

Before going further, we need to characterize the probability of error in the m-th round of the IR-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 URLLC involves short packets. In IR-HARQ with m transmissions, 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 lengths $\vec{n}_i$ and energies per symbol $\vec{P}_i$, is not correctly decoded when optimal coding is employed".

For *infinite* blocklength, an error occurs if the mutual information is below a threshold and in the case of IR-HARQ, it can easily be seen that for $i < j$ we have $\Omega_i \subseteq \Omega_j$ [9], [10], which leads to $\epsilon_m = \mathbb{P}(\Omega_m)$. In contrast, when a *finite* blocklength (or a realistic 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 (see for instance [3], [10], [11] 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 this work, we also adopt this approach assuming that the approximation is valid. Then, ε_m can be upper bounded [1, Lemma 14 and Theorem 29] and also lower bounded as in [11] 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 [12]–[15]. For instance, in [12], the authors provide more involved expressions for ϵ_m, but the feedback scheme considered is different from ours; the feedback time index in [12] is not predefined (it is a random variable) and is adapted online. In [13], [14] justifications for the approximation $\epsilon_m \approx \varepsilon_m$ when using non-binary LDPC codes or tail-biting convolutional code can be found. In [15], the authors use saddlepoint approximation to find a tight approximation of ε_m, but closed-form expression is provided only for binary erasure channels. Therefore, we consider that

using the Gaussian approximation in (1) provides a relevant tradeoff between analytical tractability and tightness of the approximations.

IV. Throughput Optimization

We remind that our main goal is to optimize the IR-HARQ by determining the blocklength and the power of the packet sent in every round in order to maximize the throughput. Throughput is defined as the average ratio of successfully decoded bits divided by the number of symbols used. The throughput can be derived using the renewal theory where the expected value of 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

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

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

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

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

$$M \leq \mathrm{M}_\mathrm{r} \tag{6}$$

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 in (4) to be active. This leads to the following optimization problem

Problem 2:

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

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

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

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

$$M \leq \mathrm{M}_\mathrm{r} \tag{11}$$

The following result proves that the solution of Problem 2 achieves almost the same performance as the one of the original Problem 1.

Proposition 1: Let $(B^{mod}, M^{mod}, \vec{n}_M^{mod}, \vec{P}_M^{mod})$ be the solution of Problem 2, which result in a value Th for the throughput according to (2). Let $Th^\star$ be the highest (optimal) value for 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 that we denote by $\mathbb{D}$. Thus, $(B^{mod}, M^{mod}, \vec{n}_M^{mod}, \vec{P}_M^{mod})$ is a feasible point of Problem 1. Since $Th^\star$ is the optimal value and Th just a feasible one, we have that $Th \leq Th^\star$. 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 show that $Th^\star \leq \frac{Th}{1-\varepsilon_{\mathrm{rel}}}$. ∎

We propose to perform the optimization over B via one-dimensional grid-search. Consequently, Problem 2 can be further simplified and leads to the following Problem 3.

Problem 3:

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

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

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

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

$$M \leq \mathrm{M}_\mathrm{r} \tag{16}$$

The rest of the section is devoted to the solution of Problem 3, which can be solved iteratively using 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 all feasible final states. By feasibility, we mean that a state S_M in $\mathbb{S}_M$ satisfies the constraints of Problem 3 and there is a path $(\vec{n}_M, \vec{P}_M)$ leading to S_M. We have $\mathbb{S}_M \subset \{1, 2, .., \mathrm{N}_\ell\} \times [0, \varepsilon_{\mathrm{rel}}] \times [0, \mathrm{E}_\mathrm{t}] \times [0, \mathrm{N}_\ell] \quad \forall M \in \{1, 2, .., \mathrm{M}_\mathrm{r}\}$. Our objective is to find the sequences/paths of states minimizing (12) to every $S_M \in \mathbb{S}_M$ being a possible candidate to achieve optimality. Then, the optimal solution of Problem 3 is retrieved by choosing out of those S_M the one with the smallest minimum.

The first three variables of the states S_m are chosen in order to be able to check the constraints (13)-(15). The dispersion variable V_m is added so as the description of S_m to depend 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 variables X given constraints Y". Now the Problem 3 can be seen as the solution of

$$\min_{M, S_M | M \in \{1,..,\mathrm{M}_\mathrm{r}\}, S_M \in \mathbb{S}_M} \left\{ \min_{\vec{n}_M, \vec{P}_M | S_M} \sum_{m=1}^{M} n_m \varepsilon_{m-1} \right\}.$$

As 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, let it be $K(\cdot)$, 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

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

The above formula can be proven for every $m \in \{1, ..., M\}$, which enables to use a dynamic programming approach. Specifically, in order 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 all 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}.

Furthermore, we show that the optimal solution of Problem 3 has characteristics that reduce the number of states needed to be tested.

Proposition 2: For M increasing, 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 [2, Appendix C], it is proven that if the last (i.e., M-th) packet with blocklength and power (n_M, P_M) 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 can easily be shown to decrease the objective function (12); hence this new configuration with an extra round gives better result while satisfying the constraints. Therefore, more transmission rounds improve the performance. ∎

Proposition 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 Proposition 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 better than the optimal point, which leads to contradiction. Furthermore, proving $\varepsilon^\star_{\mathrm{M_r}} \leq \varepsilon_{\mathrm{rel}} < \varepsilon(\vec{n}^\star_{\mathrm{M_r}-1}, n^\star_{\mathrm{M_r}} - 1, \vec{P}^\star_{\mathrm{M_r}})$ is fairly simple since the first inequality is the reliability constraint and the second cannot be violated; otherwise the point $(\vec{n}^\star_{\mathrm{M_r}-1}, n^\star_{\mathrm{M_r}} - 1, \vec{P}^\star_{\mathrm{M_r}})$ is 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 conjecture that $\varepsilon^\star_{\mathrm{M_r}} \approx \varepsilon_{\mathrm{rel}}$ since the last round would enable to satisfy the constraints without going way too far from the boundary. Proposition 3 also leads to $E^\star_{M_r} \approx \mathrm{E_t}$, where $E^\star_{M_r}$ is the energy consumed by the optimal solution of Problem 3. The reason is that if enough energy is allowed by the energy constraint (15) to be spent on $P_{\mathrm{M_r}}$ so as to compensate for a one symbol decrease of $n_{\mathrm{M_r}}$ (and still satisfy the reliability constraint (14)), then Proposition 3 is violated. Therefore, the average energy spent by the optimal solution $E^\star_{M_r}$ should be close to the boundary $\mathrm{E_t}$.

V. Algorithm Implementation

In practice, the dynamic programming algorithm requires 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 $[0, \mathrm{N}_\ell]$, but it can be quantized using bigger than one symbol step size for accelerating the simulation. Let $\mathtt{N}$ be the set of the discrete values that N_m can take.
- ε_m is real and from Proposition 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: a first one for computing the performance of the feasible states and a second one for searching over those states to find the optimal solution. The complexity is dominated by the first stage and is equal to the number of iterations of the dynamic algorithm multiplied by the number of states examined per iteration times 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 characterizes a rather slow algorithm, however in reality the algorithm can be accelerated by remarking that most of the times all 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 are introduced, the number of feasible states with same (N_m, c_m, E_m) and different V_m turns out to be rather small (often just one value). Therefore, the variable $|\mathtt{V}|$ can be thought as constant.

VI. Numerical Results and Discussion

In this section, we provide numerical results to assess the system performance. In Fig. 1, we investigate the effect of the error probability on throughput. The figure is obtained as follows: we solve Problem 2 with reliability constraint (9) to hold in 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 search of the minimum only among the states with the given ε_m.

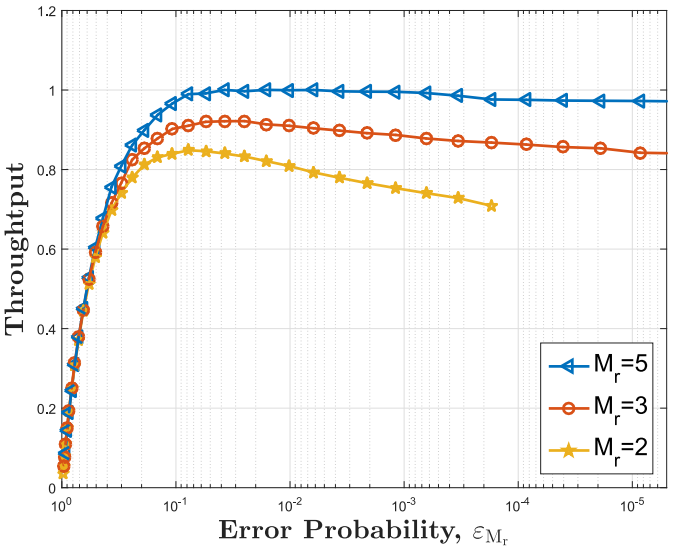

Fig. 1. Throughput vs. error probability for $N = 400$, $\mathrm{E_t} = 267$, and $B = 32$ bytes.

As shown in Proposition 2 and confirmed by Fig. 1, more transmission rounds result in higher throughput. Moreover since we have short packets (finite blocklength regime), it is not possible to attain $\varepsilon_M \to 0$ with a finite energy budget. Therefore, there exists a certain value beyond which the reliability cannot go. This is the reason why the curve of $\mathrm{M_r} = 2$ in Fig. 1 saturates at a certain error probability. The other two curves also saturates after a certain error probability, which is though much smaller and cannot be depicted in the figure. Finally, we remark, as in [4], that there is a certain value of error probability that maximizes the throughput, which is relatively high though (close to 0.1). So in our case, higher reliability can be achieved at the expense of throughput since our operating point does not correspond to the one maximizing the throughput in this figure.

The impact of the number of symbols used on the throughput performance is shown in Fig. 2, which is obtained by imposing equality in the latency constraint (8) of Problem 2.

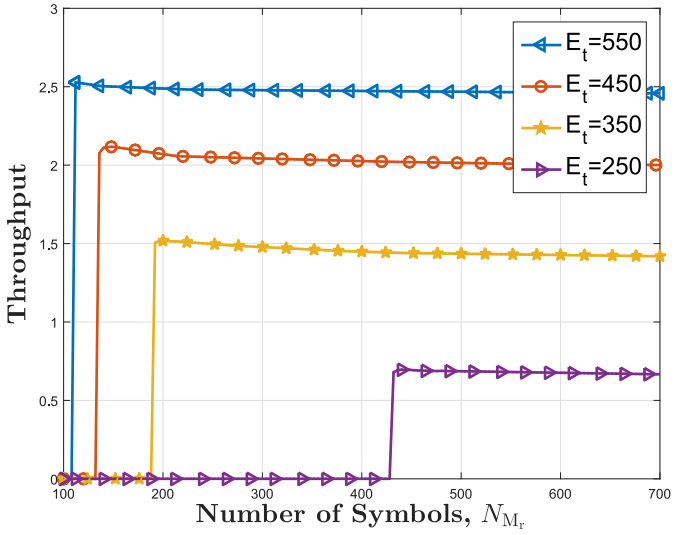

Fig. 2. Throughput vs. number of symbols used for $\varepsilon_{\mathrm{rel}} = 10^{-5}$, $B = 32$ bytes, and $\mathrm{M_r} = 3$.

When the available number of symbols are inadequate, no feasible solution exists and the throughput vanishes. Interestingly, as $N_{\mathrm{M_r}}$ grows beyond a certain threshold, only a slight increase in throughput is achieved, followed by a slow decrease. This means that it is not always beneficial from a throughput perspective to use the whole blocklength. Asymptotically, if $N_{\mathrm{M_r}} \to \infty$, then for some $m \in \{1, \cdots \mathrm{M_r}\}$ it should be $n_m \to \infty$, which in turn will result in vanishing throughput. Therefore, all curves in Fig. 2 will asymptotically converge to zero.

Fig. 3 depicts 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 IR-HARQ round to be $N_{1,min} \geq 100$ (which is set likewise so that the approximation (1) remains accurate). Consequently, the throughput cannot exceed the value $\frac{B}{N_{1,min}}$, which represents the unrealistic case of only one packet sent with minimum blocklength and achieving perfect reliability. This upper bound is closely attained as the available energy grows up to a point where only one transmission may fulfill the constraints and thus, further increase of the energy is worthless. Finally, Fig. 3 reconfirms (as in Fig. 2) that past a certain threshold, any further increase in blocklength is meaningless.

Finally, in 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 with no feasible points. Keeping a constant $\mathrm{E_t}$ by moving vertically, we see that the throughput is a unimodal function over B and there is a specific value of B that achieves optimality. This also agrees with [4] where a simple ARQ scheme with no URLLC constraints was employed.

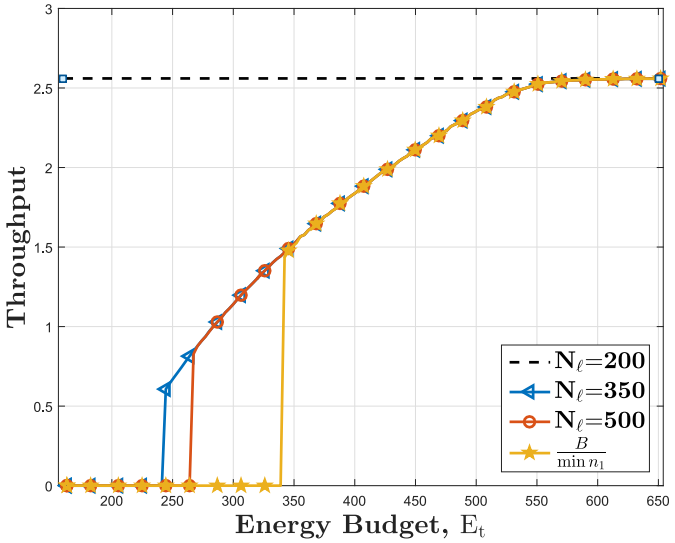

Fig. 3. Throughput vs. energy spent for $\varepsilon_{\mathrm{rel}} = 10^{-5}$, $B = 32$ bytes, and $\mathrm{M_r} = 3$.

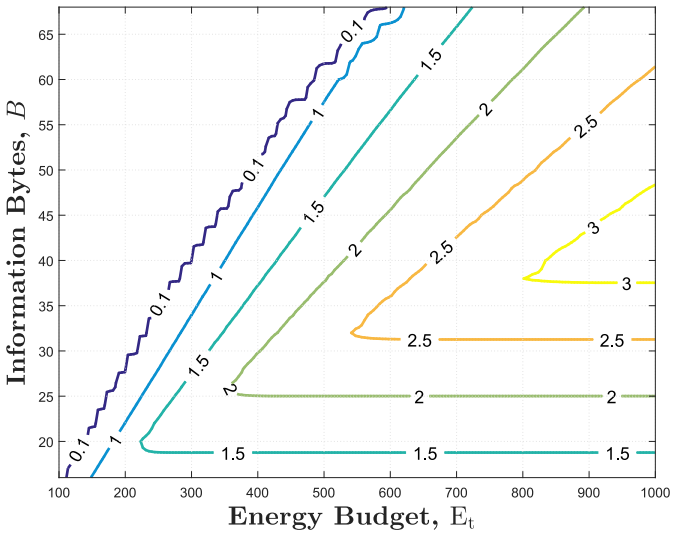

Fig. 4. Throughput vs. energy and information bits for $\varepsilon_{\mathrm{rel}} = 10^{-5}$, $\mathrm{N}_\ell = 600$, and $\mathrm{M_r} = 3$.

VII. Conclusion

We have solved the problem of throughput maximization in URLLC systems with IR-HARQ subject to latency and reliability constraints and a maximum energy budget. For that, we have proposed a dynamic programming algorithm, which takes into account the effect of finite blocklength and allows us to optimize the IR-HARQ parameters, namely the number of information bits, the number of transmission rounds, and the blocklength-power allocation.

References

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

[2] 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

[3] 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.

[4] 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.

[5] 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.

[6] 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. on Wireless Commun.*, vol. 15, no. 8, pp. 5527–5540, May 2016.

[7] D. Djonin, A. Karmokar, and V. Bhargava, "Joint rate and power adaptation for type-I hybrid ARQ systems over correlated fading channels under different buffer-cost constraints," *IEEE Trans. Veh. Technol.*, vol. 57, no. 1, pp. 421–435, Jan. 2008.

[8] L. Szczecinski, S. R. Khosravirad, P. Duhamel, and M. Rahman, "Rate allocation and adaptation for incremental redundancy truncated HARQ," *IEEE Trans. on Commun.*, vol. 61, no. 6, pp. 2580–2590, June 2013.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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