# Energy-Latency Tradeoff in Ultra-Reliable Low-Latency Communication With Retransmissions

Apostolos Avranas, Marios Kountouris, *Senior Member, IEEE*, and Philippe Ciblat

***Abstract*— High-fidelity, real-time interactive applications are envisioned with the emergence of the Internet of Things and tactile Internet by means of ultra-reliable low-latency communications (URLLC). Exploiting time diversity for fulfilling the URLLC requirements in an energy efficient manner is a challenging task due to the nontrivial interplay among packet size, retransmission rounds and delay, and transmit power. In this paper, we study the fundamental energy-latency tradeoff in URLLC systems employing incremental redundancy (IR) hybrid automatic repeat request (HARQ). We cast the average energy minimization problem with a finite blocklength (latency) constraint and feedback delay, which is non-convex. We propose a dynamic programming algorithm for energy efficient IR-HARQ optimization in terms of number of retransmissions, blocklength, and power per round. Numerical results show that our IR-HARQ approach could provide around 25% energy saving compared with one-shot transmission (no HARQ).**

***Index Terms*— URLLC, tactile Internet, IR-HARQ, energy minimization, finite blocklength.**

## I. Introduction

CURRENT wireless networks have typically been designed for increasing throughput and improving coverage, focusing mainly on human-centric communication and delay-tolerant content. The emergence of the Internet of Things (IoT) we experience nowadays enables a transition towards device-centric communication and real-time interactive systems. Various socially useful applications and new uses of wireless communication are currently envisioned in areas such as industrial control, smart cities, augmented/virtual reality (AR/VR), automated transportation, and tactile Internet. Tactile Internet enables real-time connection between people and objects and will be instrumental for supporting low-latency, high-fidelity, control-type applications, such as telesurgery, remote driving, and industrial remote monitoring [1], [2]. The mission critical and societal aspect of tactile Internet makes the support for very low end-to-end latency and extreme reliability required. The tolerable latency for tactile Internet has been set to 1 ms and ultra-reliability is quantified in terms of outage probability of $10^{-5}$ or even $10^{-7}$ [3]. Ultra-reliable low-latency communications (URLLC) lies in the overlapped area of the IoT and tactile Internet and is a key technology pillar in emerging mobile networks. Fifth generation (5G) systems envision to support URLLC scenarios with strict requirements in terms of latency (ranging from 1 ms and below to few milliseconds) and reliability (higher than 99.999%).

Guaranteeing the URLLC requirements is a challenging task since the performance is constrained by fundamental tradeoffs between delay, throughput, energy and error probability. The predominance of short messages for mission critical IoT, together with the need to reduce the packet duration and channel uses, impose 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 for a given packet size [4], [5]. In order to compensate for the reliability loss introduced by short packets, reliable communication mechanisms creating diversity have to be carried. A standard technique to improve transmission reliability, which has been adopted in various wireless standards, is incremental redundancy (IR) hybrid automatic repeat request (HARQ). However the benefits of time diversity could be rather limited under stringent latency constraints as the number of transmission rounds and channel uses is rather limited. 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 tactile Internet, and there is an inherent energy-latency tradeoff. A transmission can be successful 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. The general objective of this work is to characterize the fundamental energy-latency tradeoff and optimize IR-HARQ in URLLC systems.

Manuscript received April 16, 2018; revised September 16, 2018; accepted September 27, 2018. Date of publication October 11, 2018; date of current version November 30, 2018. *(Corresponding author: Apostolos Avranas.)*

A. Avranas and M. Kountouris are with the Mathematical and Algorithmic Sciences Laboratory, Paris Research Center, Huawei France, 92100 Boulogne-Billancourt, France (e-mail: apostolos.avranas@huawei.com; marios.kountouris@huawei.com).

P. Ciblat is with Télécom ParisTech, Université Paris-Saclay, F-75013 Paris, France (e-mail: philippe.ciblat@telecom-paristech.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2018.2874143

### A. Related Work

Prior work has considered the problem of throughput maximization by either adjusting the blocklength of each IR-HARQ round using the same power [6] with only one retransmission and the optimization is done through an exhaustive search, or via rate refinement over retransmissions of equal-sized and constant energy packets [7]. Equal-sized and constant energy packets are considered while the initial rate is maximized under a reliability constraint in [8]. In [9], sphere packing is used for optimizing the blocklength of every transmission with equal power. In [10], both power and blocklength for one packet transmission (without HARQ) are tuned to minimize the energy consumed by packets scheduled in a FIFO manner. Wang *et al.* [11] optimize the blocklength in order to maximize the rate. However, the optimization problem considered therein is not subject to reliability and latency constraints and can easily be solved using sequential differential optimization. Finally, [12] proposes a new family of protocols and compares its throughput with a dynamically optimized IR-HARQ. Many other work exists for optimizing HARQ mechanism but the vast majority of them consider infinite packet length, see for instance [13]–[16]. In [13], they consider a type-I HARQ with capacity-achieving codes and the blocklength is adapted for improving the throughput without any constraint on the packet reliability or latency. Jabi *et al.* [16] assume infinite blocklength and consider length adaptation in order to maximize the throughput (and not minimizing the energy spent as in our work); the energy efficiency of the optimal solution is checked afterwards.

### B. Contributions

In this paper, we analyze the fundamental tradeoff between latency (in terms of feedback and retransmission delay) and average consumed energy in the finite blocklength regime for URLLC systems with IR-HARQ. Considering that packets have to be decoded with a certain error probability and latency, we provide an answer whether it is beneficial to do one-shot transmission (no HARQ) or split the packet into sub-codewords and use IR-HARQ. We propose a dynamic programming algorithm for energy efficient IR-HARQ optimization in terms of number of retransmissions, blocklength and power per round. Furthermore, the impact of feedback delay on the energy consumption and IR-HARQ performance is also investigated. Finally, we investigate the asymptotic (infinite blocklength) regime and derive an expression for the solution of the average energy minimization problem. Numerical results show that our IR-HARQ approach could provide around 25% energy saving compared to one-shot transmission.

## 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 $N$. If no ARQ/HARQ mechanism is utilized, the packet of $B$ bits is transmitted only once (one-shot transmission) and its maximum length is $N$. When a retransmission strategy is employed, we consider hereafter IR-HARQ with $M$ transmissions (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 $m \in \{1,2,\ldots,M\}$ 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 1 to $m-1$ fragments of the codeword. In that case, the receiver concatenates the first till $m$-th 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 but with delay. The effect of feedback error is discussed in Section VI. 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 $\sum_{m=1}^{M}(n_m + D(\vec{n}_m)) \leq N$ where $\vec{n}_m$ is the tuple $(n_1, n_2, \ldots, n_m)$ and $D(\cdot)$ is a penalty term introduced at the $m$-th transmission due to delay for the receiver to process/decode the $m$-th packet and send back acknowledgment (ACK/NACK). The penalty $D(\cdot)$ on the $m$-th round may depend on the previous transmissions, i.e., $\vec{n}_{m-1}$, since IR-HARQ is employed and the receiver applies a decoding processing over the entire $\vec{n}_m$.

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 bytes. 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 bytes), the maximal receiver speed to satisfy the static assumption is $v = cB_d/f_c \approx 180$ km/h, where $B_d = 0.423/T_c$ [17, (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. As the channel is static along with the transmission, the channel gains are constant and the noise variance is assumed equal to one without loss of generality. The power allocation applied during the first $m$ rounds is denoted by $\vec{P}_m = (P_1, \ldots P_m)$.

## III. Problem Statement and Preliminaries

The objective of this paper is twofold: i) to derive the best HARQ mechanism that minimizes the average consumed energy for a given packet error probability and latency constraint (URLLC requirements) by optimally tuning both $\vec{n}_M$ and $\vec{P}_m$ for a prefixed $M$ (number of transmissions per HARQ mechanism), and ii) to find the optimal number of transmission rounds $M$ for different feedback delay models.

The first step for reaching the above objectives is to characterize the probability of error in the $m$-th round of the HARQ mechanism as a function of $\vec{n}_m$ and $\vec{P}_m$. To derive the packet error probability in short-packet communication, we resort to results for the non-asymptotic (finite-blocklength) regime [4].

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

When an *infinitely* large 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 $k < m$ we have $\Omega_m \subseteq \Omega_k$ [18], [19], which leads to $\epsilon_m = \mathbb{P}(\Omega_m)$. In contrast, when a real coding scheme (and so *finite* blocklength) is used, the above statement does not hold anymore and an exact expression for $\epsilon_m$ seems intractable. Therefore, in the majority of prior work on HARQ (see [19] and references therein), the exact outage probability $\epsilon_m$ is replaced with the simplified $\varepsilon_m$ defined as $\varepsilon_m = \mathbb{P}(\Omega_m)$, since $\varepsilon_m$ and $\epsilon_m$ perform quite closely when evaluated numerically. 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 also valid in the finite blocklength regime as in [6] and [11]. Then, $\varepsilon_m$ can be upper bounded [4, Lemma 14 and Th. 29] and also lower bounded as in [20] by employing the $\kappa\beta$-bounds proposed in [4]. 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., $\varepsilon_m(\vec{n}_m, \vec{P}_m)$ or $\varepsilon_m(n_1, n_2, \ldots, P_1, P_2, \ldots)$ instead of $\varepsilon_m$, whenever needed.

Notice that some works have tried to approximate more accurately the term $\epsilon_m$ or $\varepsilon_m$ [21]–[23]. For instance, Polyanskiy *et al.* [21] provide more involved expressions for $\epsilon_m$, but the feedback scheme considered is different from ours; the feedback time index in [21] is not predefined (it is a random variable) and is adapted online. In [22], tighter yet more complicated expressions for $\epsilon_m$ are provided for certain channel coding schemes. Martinez and Fàbregas [23] used saddlepoint approximation to find a tight approximation of $\varepsilon_m$, especially for binary erasure channels (BEC). Unfortunately, no closed-form expressions are provided for AWGN channel and significant effort (which goes beyond the scope of our paper) is required in order to adapt the saddlepoint approximation of [23] to AWGN channels. Therefore, we consider that using the Gaussian approximation of [4] provides a very good tradeoff between analytical tractability and tightness of the approximations.

## IV. Optimization Problem

We employ an IR-HARQ with $M-1$ retransmissions with variable blocklengths and powers over rounds. We first address the problem of minimizing the average energy consumed to achieve a target reliability $T_{\mathrm{rel}}$ (e.g. $T_{\mathrm{rel}} = 99.999\%$ in 3GPP URLLC or equivalently an outage probability $P_{\mathrm{out}} = 1 - T_{\mathrm{rel}} = 10^{-5}$) without violating the latency constraint $\sum_{m=1}^{M}(n_m + D(\vec{n}_m)) \leq N$ by properly setting $\vec{n}_M$ and $\vec{P}_M$.

### A. Optimization Problem

Letting $\varepsilon_0 = 1$, the problem is mathematically formulated as follows:

*Problem 1:* Minimization of the average energy consumed by a HARQ mechanism leads to

$$\min_{\vec{n}_M, \vec{P}_M} \sum_{m=1}^{M} n_m P_m \varepsilon_{m-1} \tag{2}$$

$$\text{s.t.} \sum_{m=1}^{M} (n_m + D(\vec{n}_m)) \leq N \tag{3}$$

$$\varepsilon_M \leq 1 - T_{\mathrm{rel}} \tag{4}$$

$$\vec{n}_M \in \mathbb{N}_{+,*}^{M} \tag{5}$$

$$\vec{P}_M \in \mathbb{R}_{+}^{M} \tag{6}$$

where $\mathbb{N}_{+,*}$ is the set of positive integers, and $\mathbb{R}_+$ corresponds to the set of non-negative real-valued variables.

To illustrate how feedback delay can impact the performance, we consider two different models:

- The first model assumes a constant delay per retransmission, i.e., $D(\vec{n}_m) = d$. This simple model corresponds to the current real communication systems (e.g., 3GPP LTE) where the feedback is sent back through frames that are regularly spaced in time.
- The second model assumes a non-constant delay per retransmission and that feedback is sent right after the decoding outcome at the receiver side. In that case, the limiting factor to send back the feedback is the processing time required by the receiver to decode the message. We consider this time to be proportional to the size of the set of sub-codewords the receiver has already received. Therefore, after the $m$-th transmission, we have $D(\vec{n}_m) = r\sum_{i=1}^{m} n_i$ with $r$ a predefined constant.

Notice that our paper can be also applied when the same number of symbols per transmission is used ($n_m = n$, $\forall m$) since one can still optimize the power per transmission. But, except otherwise stated, we hereafter address the general case of variable blocklength per transmission ($n_m \neq n_{m'}$, $\forall m, m'$)

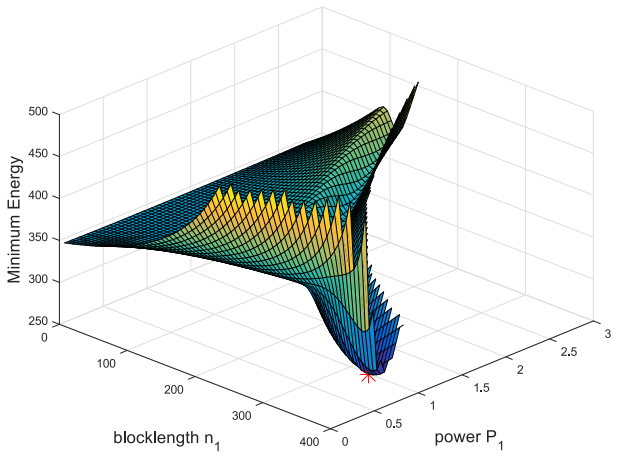



Fig. 1. Average consumed energy versus $(n_1, P_1)$ for $N = 400$, $B = 32$ bytes, and $T_{\text{rel}} = 99.999\%$. The red asterisk marks the minimum.

as a means to study the maximum capability of IR-HARQ to improve the performance. Evidently, having fixed block size per transmission is a simplified version of our general problem.

Problem 1 is a Mixed Integer Nonlinear Programming (MINLP) problem and a first approach to overcome its hardness is to relax the integer constraint by looking for $\vec{n}_M \in \mathbb{R}^M_{+,*}$ instead of $\vec{n}_M \in \mathbb{N}^M_{+,*}$. Even with that relaxation, the problem remains hard in the sense that the non-linearity cannot be managed through convexity properties of the relaxed problem. Indeed, in Figure 1 we plot the objective function of Problem 1 for $M = 2$, $D(\vec{n}_m) = 0$ and equality in the latency and reliability constraints, i.e., (3) and (4) in order to have only a 2D search on variables $(n_1, P_1)$. We observe that the objective function is neither convex nor quasi-convex nor biconvex, consequently standard convex optimization methods cannot be used.

Therefore, our objective is not providing a closed-form optimal solution for Problem 1 but deriving a low complexity algorithm finding the optimal solution. In the next two subsections, we show that Problem 1 can be written with equality in its constraints, and that a dynamic programming algorithm can be used to find the optimal solution.

### *B. Low Complexity Algorithm With Equality Constraints*

We first start with the simple case where no delay penalty is considered $(D(\vec{n}_m) = 0, \forall m)$.

*Result 1:* When $D(\vec{n}_m) = 0, \forall m$, the optimal solution of Problem 1 satisfies the latency constraint given by (3) and the reliability constraint given by (4) with equality.

This result has two consequences:

(i) Equality in (3) and (4) enables us to reduce the number of variables since one $n_m$ and one $P_m$ can be removed from the unknown variables, i.e., we search over $2(M-1)$ instead of $2M$ variables. In the conference version of our work [24], we have treated the case of $M = 2$, which leads to a 2D search instead of a 4D search. But as $M$ becomes larger, the two equalities are insufficient to significantly reduce the computational cost of the optimization algorithm.

(ii) Equality in (3) implies that it is advantageous to send as many symbols as possible during transmission but with less energy used for each symbol. In other words, *given an energy budget, it is preferable to spread this budget into many symbols with low power rather than to few ones with high power.*

Proving the above result requires the following lemmas:

*Lemma 1:* The optimal solution of Problem 1, denoted by $(\vec{n}^\star_M, \vec{P}^\star_M)$, satisfies $\varepsilon_{M-1} > \varepsilon_M$.

*Proof:* Let $\vec{P}^\star_M = (\vec{P}^\star_{M-1}, P^\star_M)$. If $\varepsilon_{M-1} \leq \varepsilon_M$ at $(\vec{n}^\star_M, \vec{P}^\star_M)$, then $(\vec{n}^\star_M, \vec{P}'_M)$ with $\vec{P}'_M = (\vec{P}^\star_{M-1}, 0)$ offers a lower consumed average energy since the last term in the sum of the objective function can be removed while the other terms remain identical. This leads to a contradiction preventing $\varepsilon_{M-1} \leq \varepsilon_M$ at the optimal point. ∎

*Lemma 2:* If $(\vec{n}^\dagger_M, \vec{P}^\dagger_M)$ satisfies $\varepsilon_{M-1} > \varepsilon_M$, then the function $P \mapsto \varepsilon_M(\vec{n}^\dagger_M, \vec{P}^\dagger_{M-1}, P)$ is decreasing in the neighborhood of $P^\dagger_M$.

*Proof:* See Appendix A. ∎

Lemma 2 enables us to force the constraint (4) to be satisfied in equality, and so proves the second part of Result 1. To prove that, we assume that the optimal point $(\vec{n}^\star_M, \vec{P}^\star_M)$ satisfies $\varepsilon_M < 1 - T_{\text{rel}}$. According to Lemma 1, we know that $\varepsilon_{M-1} > \varepsilon_M$. Consequently, according to Lemma 2, $P^\star_M$ can be decreased to $P'_M$ such that $\varepsilon_M < 1 - T_{\text{rel}}$ is still true (due to continuity of the function). This implies that $(\vec{n}^\star_M, \vec{P}^\star_{M-1}, P'_M)$ is a better solution than the optimal one, which leads to contradiction preventing $\varepsilon_M < 1 - T_{\text{rel}}$ at the optimal point.

For proving that equality in constraint (3) is required at the optimal point, we need to establish the following result.

*Lemma 3:* Let $\mathcal{B} = \{(n_1, \cdots, n_M, P_1, \cdots, P_M) \in \mathbb{R}^{2M}_{+,*} | 0.5 > \varepsilon_1(n_1, P_1) > \varepsilon_M(n_1, .., n_M, P_1, \ldots, P_M) > Q(\sqrt{2B \ln 2}/3)\}$. As long as $(an_1, n_2, .., n_M, P_1/a, P_2, \ldots, P_M) \in \mathcal{B}$, $\varepsilon_1(an_1, P_1/a)$ and $\varepsilon_M(an_1, n_2, .., n_M, P_1/a, P_2, \ldots, P_M)$ are decreasing with respect to $a$.

*Proof:* See Appendix B. ∎

Lemma 3 enables us to force the constraint (3) to be satisfied in equality, and so proves the first part of Result 1 as soon as the optimal point belongs to $\mathcal{B}$, i.e., satisfies $0.5 > \varepsilon_1 > \varepsilon_M = 1 - T_{\text{rel}} > Q(\sqrt{2B \ln 2}/3)$. To prove that, we assume that the optimal point $(\vec{n}^\star_M, \vec{P}^\star_M)$ satisfies $\sum_{m=1}^M n^\star_m < N$. For any $a > 1$ such that $(an^\star_1, n^\star_2, .., n^\star_M, P^\star_1/a, P^\star_2, \ldots, P^\star_M) \in \mathcal{B}$ and $an^\star_1 + \sum_{m=2}^M n^\star_m \leq N$ yields a better solution. And there exists at least one $a > 1$ in $\mathcal{B}$ by continuity of $\varepsilon_1$ and $\varepsilon_M$ with respect to $a$. Actually $an^\star_1$ may belong to $\mathbb{R}_{+,*}$ instead of $\mathbb{N}_{+,*}$. To overcome this issue, we assume that the scheme with $a = (n^\star_1 + 1)/n^\star_1$ is still in $\mathcal{B}$, i.e., increasing the blocklength of the first fragment by one symbol does not bring us out of $\mathcal{B}$.

We consider now the case of $D \neq 0$. The nonzero feedback delay does not modify Result 1 for the reliability constraint (4). For the latency constraint (3), the extension of Result 1 is less obvious, and the reasoning depends on the type of delay feedback model:

- For $D(\vec{n}_m) = d, \forall m$, we can simply consider Problem 1 with blocklength $N' = N - \lceil Md \rceil$, where $\lceil \cdot \rceil$ stands for

the ceiling operator, and no delay penalty. Therefore the latency constraint is equivalent to the following equality:

$$\sum_{m=1}^{M} n_m = N - \lceil Md \rceil. \tag{7}$$

- For $D(\vec{n}_m) = r\sum_{i=1}^{m} n_i, \forall m$, lemma 3 should be cautiously employed. Indeed, increasing the blocklength of the first fragment by one leads to an increase in the feedback delay at each fragment by $\lceil r \rceil$. After $M$ transmissions, the additional delay is at most $M\lceil r \rceil$. We know that the optimal solution lies in the following interval

$$N - M\lceil r \rceil \leq \sum_{m=1}^{M} (n_m + D(\vec{n}_m)) \leq N, \tag{8}$$

since the right-hand side (RHS) inequality in (8) ensures the latency constraint, and the left-hand side inequality in (8) is necessary for the optimal solution. Indeed, without this inequality, it is still possible to expand the first round by one without violating the latency constraint, hence obtaining a better solution than the optimal one, which leads to a contradiction.

In addition to the previous result and lemmas, we have the following result, which only holds when $D(\vec{n}_m) = 0, \forall m$.

*Result 2:* When $D(\vec{n}_m) = 0, \forall m$, and given $T_{\text{rel}}$ and $N$, increasing the number of retransmissions $M$ always yields a lower optimal average energy.

*Proof:* See Appendix C. ∎

Result 2 implies that when ideal feedback and no delay are guaranteed, a HARQ mechanism is always beneficial, i.e., it is always preferable to split the sub-codewords into smaller sub-codewords.

### C. Low Complexity Algorithm With Dynamic Programming

In the previous subsection, the set of feasible points has been reduced without losing optimality (as established from Result 1) and as a consequence, the search for the optimal solution of Problem 1 has been simplified. Nevertheless, due to the lack of convexity or other favorable properties for the objective function, an exhaustive search seems to be required. That involves the need for power quantization, which introduces an approximation error (denoted by $\theta$). The procedure is as follows: first, $\vec{n}_{M-1}$ and $\vec{P}_{M-1}$ are fixed; then, $n_M$ is obtained through (3) with equality, and $P_M$ is subsequence obtained through a bisection method for solving (4) with equality. The bisection method is possible since Lemma 2 establishes the monotonicity of $\varepsilon_M$. Finally, it remains to perform a $2(M-1)$-D exhaustive search to solve Problem 1. The described brute force algorithm yields a complexity in $\mathcal{O}(N^{M-1}(1/\theta)^{M-1}\log(1/\theta))$. If $M$ is small enough (typically less than 3), the algorithm can be implemented. However, when $M$ is large, performing exhaustive search is prohibitively costly and an alternative approach is required. For that, we propose an algorithm based on dynamic programming (DP). We start from the case of zero delay feedback.

We assume the optimal solution to belong in $\mathcal{B}$ (as stated in Lemma 3) so (3) and (4) become equalities. Let the state at the end of the round $m$

$$S_m = (N_m, V_m, c_m)$$

with $N_m = \sum_{i=1}^{m} n_i$, $V_m = \sum_{i=1}^{m} n_i P_i(P_i+2)/(P_i+1)^2$, and $c_m = Q^{-1}(\varepsilon_m)$. The state sequence forms a Markov chain, i.e., $p(S_m|S_{m-1}, \cdots S_1) = p(S_m|S_{m-1})$ since we have

$$N_m = N_{m-1} + n_m \tag{9}$$

$$V_m = V_{m-1} + n_m\left(1 - \frac{1}{(P_m+1)^2}\right) \tag{10}$$

$$c_m = \frac{c_{m-1}\sqrt{V_{m-1}} + n_m \ln(1+P_m)}{\sqrt{V_m}} \tag{11}$$

and the way to go from $S_{m-1}$ to $S_m$ depends only on the current round $m$ through $n_m$ and $P_m$. Notice that the assumption in Lemma 3 ensures $c_M = Q^{-1}(1-T_{\text{rel}})$ and $0 \leq c_1 \leq c_M \leq \sqrt{2B\ln 2}/3$, while Result 1 ensures $N_M = N$.

The idea comes from the fact that the $m$ first components of the objective function can be written as follows

$$\sum_{i=1}^{m} n_i P_i \varepsilon_{i-1} = \sum_{i=1}^{m-1} n_i P_i \varepsilon_{i-1} + \Delta E(S_{m-1}, S_m) \tag{12}$$

where $\Delta E(S_{m-1}, S_m) = n_m P_m \varepsilon_{m-1}$. Let $E^\star(S_m)$ be the minimum average energy going to the state $S_m$. According to (12), it is easy to prove that

$$E^\star(S_m) = \min_{\forall \text{ possible } S_{m-1}} \{\Delta E(S_{m-1}, S_m) + E^\star(S_{m-1})\} \tag{13}$$

since our problem boils down to the dynamic programming framework, and so Viterbi's algorithm can be used.

Compared to the exhaustive search, the complexity is significantly reduced, but can be still very large depending on the number of states $S_{m-1}$ and $S_m$ that has to be tested in (13). First, we see that the set of states $S_m$ for $m \in \{1, \cdots, M\}$ is not $\mathbb{R}^3$ but a much smaller set. Indeed the first component, we have $N_m \in \mathcal{N}_d = \{1, 2, \ldots, N\}$. For the second component, we have $V_m \in \mathcal{V}_d = (0, \min(N_m, c_m + \sqrt{c_m^2 + 2B\ln 2}))$ since $\sum_{i=1}^{m} n_i \ln(1+P_i) - B\ln 2 = c_m\sqrt{V_m}$ and $\sum_{i=1}^{m} n_i \ln(1+P_i) \geq V_m/2$ (as $P(P+2)/(1+P)^2 < 2\ln(1+P)$), which implies that $V_m/2 - B\ln 2 \leq c_m\sqrt{V_m}$ and so $V_m \leq c_m + \sqrt{c_m^2 + 2B\ln 2}$. For the third component, we need the next Lemma

*Lemma 4:* If $D(\vec{n}_m) = 0$ then the optimal solution $(\vec{n}_M^\star, \vec{P}_M^\star)$ satisfies $\varepsilon_1 > \varepsilon_2 > \ldots > \varepsilon_M$, and so $c_1 < c_2 < \ldots < c_M$.

*Proof:* See Appendix D. ∎

According to Lemma 4, we have $c_m \in \mathcal{C}_d = [0, Q^{-1}(1-T_{\text{rel}})]$.

Now focusing on the $S_{m-1}$ case, we straightforwardly have

$$(m-1)n_{min} \leq N_{m-1} \leq N_m - n_{min} \tag{14}$$

$$V_m - n_m \leq V_{m-1} \leq \min\{V_m, N_{m-1}\} \tag{15}$$

where $n_{min}$ is the minimum blocklength of the transmitted packets. Finally, given the target $S_m$ and $(N_{m-1}, V_{m-1})$ there

is at most one feasible $c_{m-1}$ which emerges from (10)-(11)

$$c_{m-1} = \frac{c_m\sqrt{V_m} + 2(N_m - N_{m-1})\ln\left(1 - \frac{V_m - V_{m-1}}{N_m - N_{m-1}}\right)}{\sqrt{V_{m-1}}}. \tag{16}$$

Let us now focus on the initialization. When $M = 1$, the states $S_1$ are 2D since given $(N_1, c_1)$ there can be only one feasible $P_1$ (and so $V_1$) which satisfies the equation $\varepsilon_1(N_1, P_1) = Q(c_1)$. Therefore we start from $M = 2$. To find $E^\star(S_2)$, we need to minimize over only one variable $(N_1)$, which renders this case computationally easier. Formally,

$$\begin{aligned} E^\star(N_2, V_2, c_2) = \min_{N_1}\ & N_1P_1 + n_2P_2\varepsilon_1(N_1, P_1) \\ \text{s.t. } & n_2 = N_2 - N_1 \\ & V_2 = \frac{N_1P_1(P_1+2)}{(P_1+1)^2} + \frac{N_2P_2(P_2+2)}{(P_2+1)^2} \\ & \varepsilon_2(N_1, P_1, n_2, P_2) = Q(c_2). \end{aligned} \tag{17}$$

Letting the approximation error due to quantization of $V$ and $c$ be $\theta_V$ and $\theta_c$, respectively, then the complexity is of order $\mathcal{O}(MN^2(\frac{1}{\theta_V})^2\frac{1}{\theta_c})$. In other words, the complexity of the dynamic programming algorithm is linear with respect to $M$, whereas the complexity of exhaustive search is exponential in $M$.

Extension of the above algorithm to the case of non-zero delay is easy when $D(\vec{n}_m) = d$ since we can simply reconsider the problem as having available blocklength $N' = N - \lceil Md \rceil$ and no delay penalty. When $D(\vec{n}_m) = r\sum_{i=1}^m n_i$ more changes are required: first, $N_m$ now represents the available latency at the $m$-th round, second, an additional data structure $\mathrm{N_{net}}$ is needed which stores the number of symbols sent disregarding the delays, and third to find every $E^\star((N_m, V_m, c_m))$ an additional search within the states $(N, V_m, c_m), \forall N \in [N_m - m\lceil r \rceil, N_m - 1]$ is employed.

## V. Asymptotic Regime

The minimum average energy for sending a fixed number of $B$ information bits is a decreasing function with respect to the latency $N$. Indeed, as seen in Problem 1, the optimal solution for a given $N$ is a feasible solution of $(N+1)$ and so equal or worse than the optimal solution for the latency $(N+1)$. In following result, we prove that the optimal solution converges to an asymptotic value when $N \to \infty$.

*Result 3:* When $N \to \infty$, the minimum average energy of Problem 1 for fixed $M$ is given by

$$\begin{aligned} E^\star_{as}(M, B) = \min_{(E_1, \cdots, E_M)}\ & r(E_1, \cdots, E_M) \\ \text{s.t. } & \sum_{m=1}^M E_m = E^\infty_{\mathrm{No-HARQ}} \end{aligned} \tag{18}$$

with

$$r(E_1, \cdots, E_M) = E_1 + \sum_{m=2}^M Q\left(\frac{\sum_{i=1}^{m-1} E_i - B\ln 2}{\sqrt{2\sum_{i=1}^{m-1} E_i}}\right) E_m \tag{19}$$

and $E^\infty_{\mathrm{No-HARQ}} = \frac{(Q^{-1}(1-T_{\mathrm{rel}}))^2}{2}\left(1 + \sqrt{1 + \frac{2B\ln 2}{(Q^{-1}(1-T_{\mathrm{rel}}))^2}}\right)^2$.

*Proof:* See Appendix E. ∎

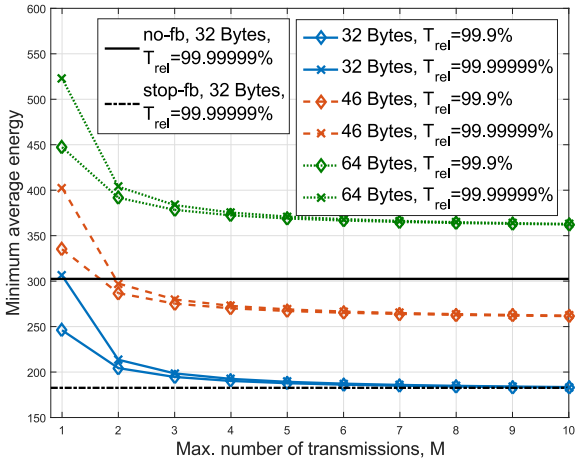



Fig. 2. Minimum average energy (when $N \to \infty$) versus $M$.

Note that $E^\infty_{\mathrm{No-HARQ}}$ corresponds to the required average energy when $N \to \infty$ for the case of no HARQ and can also be obtained from [4, eq. (4.309)]. As an illustration, in Figure 2 we plot $E^\star_{as}(M)$ versus $M$ for different $B$ and $T_{\mathrm{rel}}$. We also plot two curves corresponding to the minimum energy, one given in [21, Th. 3] for no feedback ("no-fb" in the figure) and the other ("stop-fb" in the figure) given in [21, Th. 10] where ACK/NACK feedback is sent after the transmission of each symbol. Actually, the "no-fb" line corresponds to our case $M = 1$ when removing its third-order term. The "stop-fb" line is close to our eq. (20) since its adaptive feedback can be mimicked in our case if infinite available number of transmissions are considered.

When $N \to \infty$, a non-zero delay feedback - irrespectively of the model considered - does not impact the asymptote value since the latency constraint vanishes, which makes that Result 2 still holds.

When $M$ also grows to infinity, we have an additional result.

*Result 4:* When $M \to \infty$, the asymptotic minimum average energy stated in Result 3 behaves as follows:

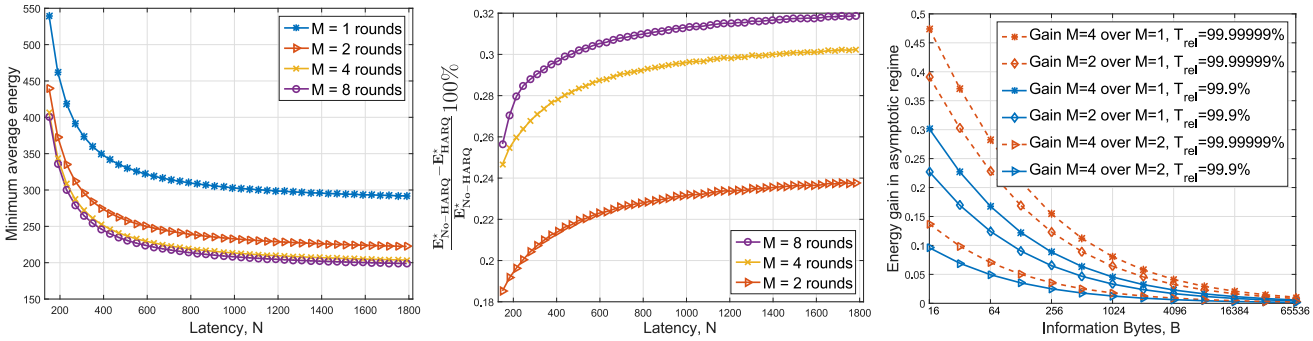
$$\lim_{M\to\infty} E^\star_{as}(M) = \int_0^{E^\infty_{\mathrm{No-HARQ}}} Q\left(\frac{E - B\ln 2}{\sqrt{2E}}\right) \mathrm{d}E. \tag{20}$$

*Proof:* See Appendix F. ∎

Given $B$, increasing $T_{\mathrm{rel}}$ to $\overline{T}_{\mathrm{rel}}$ also increases $E^\infty_{\mathrm{No-HARQ}}$ to $\overline{E}^\infty_{\mathrm{No-HARQ}}$, which implies that $\lim_{M\to\infty} E^\star_{as}(M) < \lim_{M\to\infty} \overline{E}^\star_{as}(M)$. In Figure 2, these limit values cannot be distinguished and seem to coincide since they are very close to each other. This happens because, as it easily can be shown, $\lim_{M\to\infty} \overline{E}^\star_{as}(M) - \lim_{M\to\infty} E^\star_{as}(M) < (1 - T_{\mathrm{rel}})(\overline{E}^\infty_{\mathrm{No-HARQ}} - E^\infty_{\mathrm{No-HARQ}})$ and $T_{\mathrm{rel}}$ is very small.

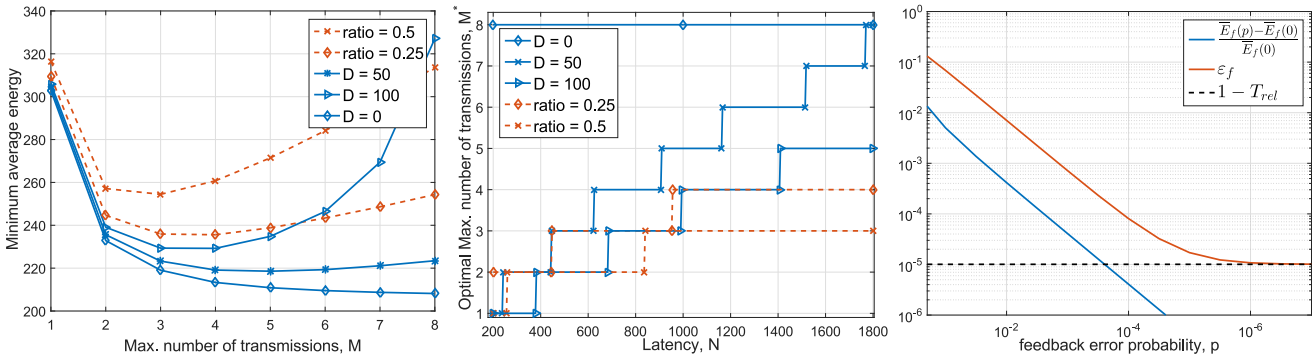
## VI. Numerical Results and Discussion

In this section, we provide numerical results to validate our analysis. We consider $B \geq 32$ bytes and $T_{\mathrm{rel}} > 99.99999\%$, i.e., $1 - T_{\mathrm{rel}} \gg Q(\sqrt{2B\ln 2/3}) \geq 1.7 \cdot 10^{-10}$ always holds. Furthermore, we consider $n_1$ and $P_1$ such that $\varepsilon_1 < 0.5$. Consequently, the assumption on $\mathcal{B}$ in Lemma 3 is not restrictive. The latency constraint (3) is simplified either




(a) Minimum average energy vs. $N$ for $B = 32$ Bytes and $T_{\rm rel} = 99.999\%$. (b) Energy gain of $M$ rounds over no HARQ ($M$=1) vs. $N$ for $B = 32$ Bytes and $T_{\rm rel}$=99.999%. (c) Energy gain vs. $B$ in the asymptotic regime ($N \to \infty$).

Fig. 3. Performance analysis when $D(\vec{n}_m) = 0$.




(a) Minimum average energy vs. $M$. (b) $M^\star$ (assuming $M \leq 8$) vs. $N$. (c) $\frac{\overline{E}_f(p)-\overline{E}_f(0)}{\overline{E}_f(0)}$ and $\varepsilon_f$ vs. $p$ when $N = 450$ and $M$=4 rounds.

Fig. 4. Performance analysis for $B = 32$ Bytes and $T_{\rm rel} = 99.999\%$.

according to (7) for fixed delay feedback model (including $D = 0$) or according to (8) for the linear delay feedback model.

First, we assume $D = 0$. In Figure 3a, we plot the minimum average energy versus $N$ and confirm that the energy for sending $B$ information bits decreases when $N$ increases. Additionally, the energy attains the asymptotic value predicted by Result 3. Moreover, we confirm Result 2, since the minimum average energy decreases when $M$ increases for the case of zero delay feedback; however, the gain becomes negligible when $M$ is large enough. In Figure 3b, for the same $B$ and $T_{\rm rel}$ as in Figure 3a, we plot the energy gain by using HARQ with $M$ rounds over $M = 1$ (denoted as $E_{\rm No-HARQ}$). We observe that the energy gain monotonically increases when $N$ grows. As the latency constraint becomes more stringent, the benefit from employing HARQ diminishes.

In Figure 3c, we plot the energy gain for different values of $M$ versus $B$ when $N \to \infty$. The energies and the corresponding gains are derived using Result 3. The higher the reliability or the lower $B$, the higher the gain. This remark also holds for non-zero delay feedback since we are in the asymptotic regime.

We consider now $D(\vec{n}_M) \neq 0$ unless otherwise stated. In Figure 4a, we plot the minimum average energy versus $M$ for different delay feedback models (solid lines for fixed delay and dashed line for the linear delay model). When $d > 0$, splitting the packet/transmission in rounds is not always advantageous and we observe that an optimal bounded value of $M$, denoted by $M^\star$, exists. Indeed, for small values of $M$, the delay penalty is small and it is still of interest to split, whereas when $M$ is large, the value of $N'$ in (7) becomes very low and there is no gain to split further. The same statement holds when the linear delay feedback model is applied.

In Figure 4b, we plot $M^\star$ versus $N$ restricting $M \leq 8$. The delay penalties become more significant when $N$ decreases when eventually prevents from using an HARQ mechanism. Therefore, $M^\star$ increases with respect to $N$. In the case of linear delay feedback model, $M^\star$ increases much slower than in the fixed delay feedback model since the effect of delay in the energy consumption is higher when $M$ increases.

The effect of feedback error is investigated assuming that the feedback error is modeled by a binary symmetric channel (BSC) with error probability $p$ as in [25]. $\overline{E}_f(p)$ denotes the average consumed energy and $\varepsilon_f$ denotes the overall error probability when feedback error $p$ is considered. Closed-form expressions with respect to $p$ can be obtained for $\overline{E}_f$ and $\varepsilon_f$ (not reported here due to space limitation) using results from [25]. In Figure 4c for some optimal configuration $(\vec{n}_M^\star, \vec{P}_M^\star)$ we plot (i) the relative loss in energy, i.e., $(\overline{E}_f(p) - \overline{E}_f(0))/\overline{E}_f(0)$ and (ii) $\varepsilon_f$ versus $p$. We observe that there is

only a slight increase of the consumed energy, even for bad feedback channels. In contrast, the reliability is significantly affected by feedback errors except when $p$ is small enough compared to $(1-T_{rel})$. Indeed, if $p < 0.1(1-T_{rel})$, then the URLLC requirements are still satisfied. Hence, the feedback has to be protected on the control channel according to this error probability constraint; this is relatively easy to achieve without consuming a lot of resources since it is just one bit.

## VII. Conclusion

In this paper, we have characterized the energy-latency tradeoff in URLLC systems with retransmissions in the finite blocklength regime and showed how IR-HARQ can be optimized by tuning the number of rounds, the blocklength and the transmit power. A dynamic programming algorithm for solving the non-convex average energy minimization problem subject to URLLC constraints is provided. The main takeaway of this paper is that a properly optimized IR-HARQ scheme can be beneficial in terms of energy as long as the feedback delay is reasonable compared to the packet size. Future work could study how frequency and/or space diversity can alter the tradeoff and the IR-HARQ design. Further extensions of this framework may include the analysis of fading and multi-antenna systems with both perfect and imperfect channel knowledge.

## Appendix A
## Proof of Lemma 2

Let us denote by $\partial\varepsilon_M/\partial P$ the derivative function of $P \mapsto \varepsilon_M(\vec{n}_M^\dagger, \vec{P}_{M-1}^\dagger, P)$. We will prove that $\partial\varepsilon_M/\partial P_{|P=P_M^\dagger} < 0$. By change of variables $y = 1/(P+1)^2$ and putting $y^\dagger = 1/(P_M^\dagger+1)^2$, we show that

$$\frac{\partial\varepsilon_M}{\partial P} < 0 \text{ at } P = P_M^\dagger \Leftrightarrow \frac{\partial\varepsilon_M}{\partial y} > 0 \text{ at } y = y^\dagger \Leftrightarrow h(y) > 0 \text{ at } y = y^\dagger \quad (21)$$

where $h(y) = k_2 - yk_1 + n_M(1 - y + y\ln(y)/2)$ with $k_1 = \sum_{i=1}^{M-1} n_i \ln(1+P_i) - B\ln 2$ and $k_2 = \sum_{i=1}^{M-1} n_i \left(1 - 1/(1+P_i)^2\right)$. It is easy to prove that $h(y)$ is a monotonically decreasing function. If $h(1) \geq 0$, then (21) is straightforwardly satisfied. If $h(1) < 0$, then it exists $y_0 \in (0,1)$ such that $h(y_0) = 0$. So for $y \in [y_0, 1]$, we get $h(y) \leq 0$, which implies that $\varepsilon_M$ is decreasing with respect to $y$. As a consequence, for $y \in [y_0, 1]$ and so the corresponding $P(y)$, we have $\varepsilon_M(\vec{n}_M^\dagger, \vec{P}_{M-1}^\dagger, P(y)) \geq \varepsilon_M(\vec{n}_M^\dagger, \vec{P}_{M-1}^\dagger, 0) = \varepsilon_{M-1}(\vec{n}_{M-1}^\dagger, \vec{P}_{M-1}^\dagger)$ which prevents to have $P(y) = P_M^\dagger$ according to the assumption $\varepsilon_{M-1} \geq \varepsilon_M$ on the analyzed point. Consequently, $y^\dagger$ does not belong to $[y_0, 1]$, and belongs to $(0, y_0)$ where (21) holds.

## Appendix B
## Proof of Lemma 3

Let $\varepsilon_1 = Q(F_1(a))$ and $\varepsilon_M = Q(F(a))$ where

$$F_1(a) = \frac{g_1(a) - c}{\sqrt{g_2(a)}} \text{ and } F(a) = \frac{g_1(a) + c_1 - c}{\sqrt{g_2(a) + c_2}},$$

with $g_1(a) = an_1\ln(1+\frac{P_1}{a})$, $g_2(a) = an_1(1-1/(1+P_1/a)^2)$, $c_1 = \sum_{m=2}^{M} n_m \ln(1+P_m)$, $c_2 = \sum_{m=2}^{M} n_m(1 - 1/(1+P_m)^2)$, and $c = B\ln 2$. As we consider a point in $\mathcal{B}$, we get

$$\varepsilon_1 < 0.5 \Leftrightarrow an_1\ln(1+P_1/a) > c \Rightarrow E_1 > B\ln 2 \quad (22)$$

where $E_1 = n_1P_1$. To prove (22), we use the inequality $\ln(1+x) \leq x$ when $x \geq 0$. Once again, belonging to $\mathcal{B}$ leads to

$$F_1(a) \leq F(a) \leq \sqrt{2B\ln 2}/3. \quad (23)$$

We want to show that $\varepsilon_1$ and $\varepsilon_M$ are decreasing functions with respect to $a$, i.e., $F_1'(a) \geq 0$ and $F'(a) \geq 0$ where $f'(a)$ stands for $\mathrm{d}f/\mathrm{d}a$ for any mapping $f$. As $g_1(a), g_2(a), g_1'(a)$ and $g_2'(a)$ are strictly positive, we have

$$F_1'(a) \geq 0 \Leftrightarrow 2g_1'(a)g_2(a) \geq g_2'(a)(g_1(a) - c) \quad (24)$$
$$\Leftrightarrow c \geq E_1H(P_1/a) \quad (25)$$

and

$$F'(a) \geq 0 \Leftrightarrow 2g_1'(a)(g_2(a) + c_2) \geq g_2'(a)(g_1(a) + c_1 - c) \quad (26)$$
$$\Leftrightarrow c \geq E_1H(P_1/a) + (c_1 - K(P_1/a)c_2) \quad (27)$$

where

$$x \mapsto H(x) = \frac{2x + 4 - \ln(1+x)(\frac{4}{x} + x + 3)}{x(x+3)},$$

and

$$x \mapsto K(x) = \frac{2(x+1)^3\left(\ln(1+x) - \frac{x}{x+1}\right)}{x^2(x+3)}.$$

After some algebraic manipulations, (24) and (26) are equivalent to

$$F_1(a) \leq \frac{2g_1'(a)\sqrt{g_2(a)}}{g_2'(a)} = \sqrt{E_1}W(P_1/a, 0) \quad (28)$$

$$F(a) \leq \frac{2g_1'(a)\sqrt{g_2(a) + c_2}}{g_2'(a)} = \sqrt{E_1}W(P_1/a, \frac{c_2}{E_1}) \quad (29)$$

with

$$(x, y) \mapsto W(x, y) = K(x)\sqrt{y + \frac{x+2}{(1+x)^2}}. \quad (30)$$

We want now to prove that either (25) or (28) holds for any $x > 0$, and either (27) or (29) holds for any $x > 0$. For that, we split the analysis into two intervals on $x$.

- If $x \in (0, 484)$: the function $x \mapsto W(x, 0)$ is a positive unimodal function converging to zero when $x \to \infty$. For $x \in (0, 484)$, it is easy to check that $W(x,0) \geq W(0,0) = \sqrt{2}/3$. As $W(x,y) > W(x,0)$ for any $y \geq 0$, we obtain that $\sqrt{E_1}W(x,y) \geq \sqrt{E_1}W(x,0) \geq \sqrt{2E_1}/3$. Due to (22), we have $\sqrt{E_1}W(x,y) \geq \sqrt{E_1}W(x,0) \geq \sqrt{2B\ln 2}/3$. According to (23), we check that $\sqrt{E_1}W(x,y) \geq \sqrt{E_1}W(x,0) \geq F(a) \geq F_1(a)$. Therefore, (28) and (29) hold.
- If $x \in [484, \infty)$: in that interval, we can see that $H(x) \leq 0$, which implies that (25) holds.
  It now remains to check that either (27) or (29) holds. For doing so, we distinguish two cases:

- If $c_1 \leq 10.37c_2$: one can check that $K(x)$ is an increasing function. Therefore for $x \geq 484$, we get $K(x) \geq K(484) > 10.37$. Consequently, $c_1 - K(x)c_2 < 0$. As $H(x) \leq 0$ too for $x \geq 484$, it is easy to show that (27) holds.
- If $c_1 > 10.37c_2$: this inequality leads to

$$\sum_{m=2}^{M} n_m \ln(1+P_m) - 10.37 n_m \left(1 - \frac{1}{(1+P_m)^2}\right) > 0$$

which forces that there exists at least one $m_x \in \{2,\ldots,M\}$ such that $n_{m_x}\ln(1+P_{m_x}) > 10.37 n_{m_x}(1-1/(1+P_{m_x})^2) > 0 \Rightarrow P_{m_x} > 31866$ which implies that $c_2 \approx \sum_{m\in\{2,..,M\}\setminus m_x} n_m(1 - 1/(1+P_m)^2) + n_{m_x} \Rightarrow c_2 > n_{m_x}$. Consequently, according to (30), $\sqrt{E_1}W(x, c_2/E_1) \geq K(484)\sqrt{n_{i_x}} \geq 10.37 \cdot \sqrt{1}$. If (29) does not hold, one can see that $\varepsilon_M < Q(10.37) \approx 1.7{\cdot}10^{-25}$. As this error does not correspond to any reasonable operating point, we consider that (29) holds.

## Appendix C
## Proof of Result 2

Consider the last round $M$ where for the optimal point $(\vec{n}^\star_M, \vec{P}^\star_M)$, we know that $\varepsilon_{M-1} > \varepsilon_M$ (see Lemma 1 and its related proof for more details). For $x \in [0, n^\star_M]$, let

$$F(x) = Q\left(\frac{x\ln(1+P^\star_M) + \sum_{i=1}^{M-1} n_i^\star \ln(1+P_i^\star) - B\ln 2}{\sqrt{x\frac{P^\star_M(P^\star_M+2)}{(P^\star_M+1)^2} + \sum_{i=1}^{M-1} n_i \frac{P_i^\star(P_i^\star+2)}{(P_i^\star+1)^2}}}\right).$$

We know that $F(0) = \varepsilon_{M-1} > \varepsilon_M = F(n^\star_M)$ and that $F(\cdot)$ is a continuous (not necessary monotonically decreasing) function. Therefore, it exists $x_0 \in (0, n^\star_M)$ such that $F(0) < F(x_0) < F(n^\star_M)$. If $F$ is smooth enough, it exists an integer $\overline{n} \in \{1,2,\ldots,n^\star_M - 1\}$ (typically equal to $\lfloor x_0 \rfloor$ or $\lceil x_0 \rceil$) such that $\varepsilon_{M-1} > F(\overline{n}) > \varepsilon_M$. Then, the new point of $M+1$ rounds, which is $(\vec{n}^\star_{M-1}, \overline{n}, n^\star_M - \overline{n}, \vec{P}^\star_{M-1}, P^\star_M, P^\star_M)$, leads to the following average energy

$$\sum_{m=1}^{M-1} n_m^\star P_m^\star \varepsilon_{m-1} + \overline{n} P_M^\star F(\overline{n}) + (n_M^\star - \overline{n}) P_M^\star \varepsilon_{M-1},$$

which is smaller that the average energy provided by the point $(\vec{n}^\star_M, \vec{P}^\star_M)$. Obviously the reliability constraint (given by $\varepsilon_M$) remains unaltered and the latency constraint does not change since $D(\vec{n}_m) = 0$. So increasing the number of transmissions to $M+1$ improves the optimal operating point of $M$ transmissions.

## Appendix D
## Proof of Lemma 4

To prove the lemma, we will prove that if for some solution the states $\tilde{S}_{i-1}$, $\tilde{S}_i$,$\tilde{S}_{i+1}$ satisfy $\tilde{c}_{i-1} \geq \tilde{c}_i$ and $\tilde{c}_i < \tilde{c}_{i+1}$, then there exists a better solution, thus it cannot be the optimal one. Therefore, if for the optimal solution for some $i$ we know $c^\star_{i+1} > c^\star_i$ then it must $c^\star_i > c^\star_{i-1}$ and since from Lemma 1 we know $c^\star_M > c^\star_{M-1}$, this lemma is proved by induction.

To prove the existence of a better solution we only have to prove the superiority of a configuration of $M-1$ rounds that goes directly from the state $\tilde{S}_{i-1}$ to $\tilde{S}_{i+1}$ using one fragment of blocklength $n_i + n_{i+1}$ and has exactly the same configuration before and after those states (then due to Proposition 2 there exists an even better configuration with $M$ rounds). Hence, we only need to prove:

$$\Delta E(\tilde{S}_{i-1}, \tilde{S}_i) + \Delta E(\tilde{S}_i, \tilde{S}_{i+1}) \geq \Delta E(\tilde{S}_{i-1}, \tilde{S}_{i+1}). \quad (31)$$

Since a zero delay penalty is assumed, using (3) and (4) with equalities allows us to derive that

$$\Delta E(S_{k-1}, S_k) = n_k P_k \varepsilon_{k-1} = n_k (e^{\frac{\gamma_k}{n_k}} - 1) Q(c_{k-1})$$

where $\gamma_k = c_k\sqrt{V_k} - c_{k-1}\sqrt{V_{k-1}} > 0$. Since $\tilde{c}_{i-1} \geq \tilde{c}_i$, to prove (31) it suffices to prove that

$$\tilde{n}_i e^{\frac{\tilde{\gamma}_i}{\tilde{n}_i}} + \tilde{n}_{i+1} e^{\frac{\tilde{\gamma}_{i+1}}{\tilde{n}_{i+1}}} \geq (\tilde{n}_i + \tilde{n}_{i+1}) e^{\frac{\tilde{\gamma}_i + \tilde{\gamma}_{i+1}}{\tilde{n}_i + \tilde{n}_{i+1}}}.$$

Changing variables as $\lambda_l = \frac{\tilde{n}_l}{\tilde{n}_i + \tilde{n}_{i+1}}$ and $x_l = \frac{\tilde{\gamma}_l}{\tilde{n}_l}$, $l \in \{i, i+1\}$:

$$\lambda_i e^{x_i} + \lambda_{i+1} e^{x_{i+1}} \geq e^{\lambda_i x_i + \lambda_{i+1} x_{i+1}},$$

which holds due to the convexity of the exponential function.

## Appendix E
## Proof of Result 3

We consider $E_i = n_i^\star P_i^\star$ where $n_i^\star$ and $P_i^\star$ are the $i$-th blocklength and power components of $(\vec{n}^\star_M, \vec{P}^\star_M)$ respectively. Notice that each $E_i$ depends on $N$. Let us assume that it exists at least $i_0 \in \{1,2,\ldots,M\}$ such that $\lim_{N\to\infty} E_{i_0} = \infty$. According to Lemma 4, we know that $\varepsilon_1 > \varepsilon_2 > \ldots \varepsilon_M = 1 - T_{\text{rel}} > 0$ at the optimal point. Consequently, the minimum average energy $E_1 + \sum_{i=2}^{M} \varepsilon_{i-1} E_i \to \infty$ too. For at least one finite $N$, say $N_f$, the optimal point leads to a finite minimum average energy. For any $N > N_f$, the optimal solution cannot increase the minimum average energy since the optimal solution at $N_f$ is a feasible point of Problem 1 for $N$. So the minimum average energy is upper bounded when $N \to \infty$. Therefore, $\lim_{N\to\infty} E_i < \infty, \forall i \in \{1,2,\ldots,M\}$.

When $N \to \infty$, we know that the delay feedback model does not have an impact on the latency constraint, so we can apply the results obtained for $D = 0$. According to Lemma 3, we also know that it is preferable to increase the blocklength rather than the power in order to save energy. Therefore, when $N \to \infty$, we have to take $n_1^\star$ as large as possible, i.e., $\lim_{N\to\infty} n_1^\star = \infty$. Similar arguments can be applied to the other rounds, i.e., $\lim_{N\to\infty} n_i^\star = \infty$ with $i \in \{2,\cdots,M\}$. As $\lim_{N\to\infty} E_i < \infty$, we get $\lim_{N\to\infty} P_i^\star = 0$. By using $N \to \infty$ in

$$\frac{P_i^\star}{P_i^\star + 1} \leq \ln(1+P_i^\star) \leq P_i^\star$$

and the fact that $\lim_{N\to\infty} P_i^\star = 0$, we easily obtain that $E_i = \lim_{N\to\infty} n_i^\star \ln(1+P_i^\star)$. Plugging this previous equation in (1) leads to

$$\lim_{N\to\infty} \varepsilon_m = Q\left(\frac{\sum_{i=1}^{m} E_i - B\ln 2}{\sqrt{2\sum_{i=1}^{m} E_i}}\right). \quad (32)$$

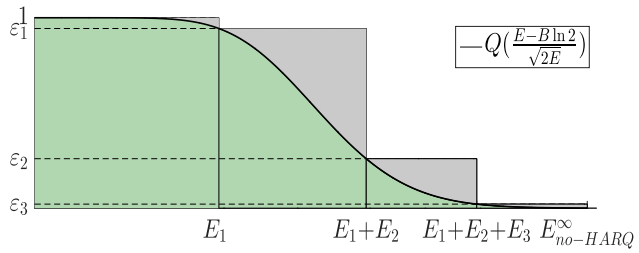



Fig. 5. Geometrical interpretation of Result 3 for $M = 4$.

Putting $m = M$ in (32) and using (4) with equality, we have

$$\sum_{i=1}^{M} E_i = \frac{(Q^{-1}(1-T_{\mathrm{rel}}))^2}{2}\Big(1+\sqrt{1+\frac{2B\ln 2}{(Q^{-1}(1-T_{\mathrm{rel}}))^2}}\Big)^2 \tag{33}$$

where its RHS corresponds to the energy when no HARQ ($M = 1$) is used and is denoted by $E^{\infty}_{\mathrm{No-HARQ}}$.

## Appendix F
## Proof of Result 4

The function $E \mapsto Q((E - B\ln 2)/\sqrt{2E})$ is plotted in Figure 5. We also draw the $m$-th component of the objective function (19) of Result 3, which corresponds to the area of the partially grey partially green rectangular box located from $\sum_{i=1}^{m-1} E_i$ to $\sum_{i=1}^{m} E_i$ with a level $\varepsilon_{m-1}$ (see (32)). According to (18), the final point is $E^{\infty}_{\mathrm{No-HARQ}}$. Consequently, the sum of the green and the grey areas gives the value of the objective function (19). It is evident that the function $E \mapsto Q((E - B\ln 2)/\sqrt{2E})$ coincides at the upper left corner of each rectangular box and is always inside each rectangular box (due to its decreasing monotonicity). Therefore, the value of the objective function (19) cannot be lower than the green area. When $M \to \infty$, we can decrease the width of each rectangular box converging to a solution that includes only the green area. Consequently, the minimum energy spent converges to the green area, which is identical to the Riemann integral of $E \mapsto Q((E - B\ln 2)/\sqrt{2E})$ from 0 to $E^{\infty}_{\mathrm{No-HARQ}}$.

## References


[1] G. P. Fettweis, "The tactile Internet: Applications and challenges," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 64–70, Mar. 2014.

[2] M. Maier, M. Chowdhury, B. P. Rimal, and D. P. Van, "The tactile Internet: Vision, recent progress, and open challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 138–145, May 2016.

[3] A. Aijaz, Z. Dawy, N. Pappas, M. Simsek, S. Oteafy, and O. Holland. (Jul. 2018). "Toward a tactile Internet reference architecture: Vision and progress of the IEEE P1918.1 standard." [Online]. Available: https://arxiv.org/pdf/1807.11915

[4] Y. Polyanskiy, "Channel coding: Non-asymptotic fundamental limits," Ph.D. dissertation, Dept. Elect. Eng., Princeton Univ., Princeton, NJ, USA, Nov. 2010.

[5] M. Hayashi, "Information spectrum approach to second-order coding rate in channel coding," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4947–4966, Nov. 2009.

[6] 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.

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

[8] 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, pp. 2063–2068.

[9] A. R. Williamson, T.-Y. Chen, and R. D. Wesel, "A rate-compatible sphere-packing analysis of feedback coding with limited retransmissions," in *Proc. IEEE ISIT*, Cambridge, MA, USA, Jul. 2012, pp. 2924–2928.

[10] 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 Commun.*, vol. 15, no. 8, pp. 5527–5540, Aug. 2016.

[11] H. Wang *et al.*, "An information density approach to analyzing and optimizing incremental redundancy with feedback," in *Proc. IEEE Int. Symp. Inf. Theory*, Aachen, Germany, Jun. 2017, pp. 261–265.

[12] K. F. Trillingsgaard and P. Popovski, "Generalized HARQ protocols with delayed channel state information and average latency constraints," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 1262–1280, Feb. 2018.

[13] D. V. Djonin, A. K. Karmokar, and V. K. 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.

[14] E. Visotsky, V. Tripathi, and M. Honig, "Optimum ARQ design: A dynamic programming approach," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun./Jul. 2003, p. 451.

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

[16] M. Jabi, M. Benjillali, L. Szczecinski, and F. Labeau, "Energy efficiency of adaptive HARQ," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 818–831, Feb. 2016.

[17] J. D. Gibson, Ed., *The Communications Handbook*. Boca Raton, FL, USA: CRC Press, 2002.

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

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

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

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

[22] K. Vakilinia, S. V. S. Ranganathan, D. Divsalar, and R. D. Wesel, "Optimizing transmission lengths for limited feedback with nonbinary LDPC examples," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2245–2257, Jun. 2016.

[23] A. Martinez and A. Guillén i Fàbregas, "Saddlepoint approximation of random-coding bounds," in *Proc. Inf. Theory Appl. Workshop (ITA)*, La Jolla, CA, USA, Feb. 2011, pp. 1–6.

[24] A. Avranas, M. Kountouris, and P. Ciblat, "Energy-latency tradeoff in ultra-reliable low-latency communication with short packets," in *Proc. IEEE Conf. Global Commun. (Globecom)*, Dec. 2018.

[25] S. R. Khosravirad and H. Viswanathan. (Oct. 2017). "Analysis of feedback error in automatic repeat request." [Online]. Available: https://arxiv.org/abs/1710.00649


**Apostolos Avranas** received the Diploma (Hons.) in electrical and computer engineering from the Aristotle University of Thessaloniki, Greece, in 2015. He is currently pursuing the Ph.D. degree with the Huawei Paris Research Center in collaboration with Télécom ParisTech. In 2014 during his studies, he completed an internship in CNRS. In 2016, he joined the Huawei Paris Research Center. His research focuses on the performance analysis of wireless communication and resource allocation.

**Marios Kountouris** (S'04–M'08–SM'15) received the Diploma in ECE from the National Technical University of Athens, Greece, in 2002, and the M.S. and Ph.D. degrees in EE from ENST Paris (Télécom ParisTech), France, in 2004 and 2008, respectively. From 2008 to 2009, he was with The University of Texas at Austin, where he was involved in wireless ad hoc networks under DARPAs IT-MANET Program. From 2009 to 2013, he was an Assistant Professor at the Department of Telecommunications, SUPELEC (now CentraleSuplec), France, and from 2014 to 2016, he was an Associate Professor. From 2014 to 2015, he was an Adjunct Professor at Yonsei University, South Korea. Since 2015, he has been a Principal Researcher at the Paris Research Center, Huawei France. He has received several honors and best paper awards, including the 2016 IEEE ComSoc Communication Theory Technical Committee Early Achievement Award and the 2013 IEEE ComSoc Outstanding Young Researcher Award for the EMEA Region. He currently serves as an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and the IEEE WIRELESS COMMUNICATION LETTERS.

**Philippe Ciblat** was born in Paris, France, in 1973. He received the Engineering degree from Télécom ParisTech and the M.Sc. degree in automatic control and signal processing from Université Paris-Sud, Orsay, France, in 1996, and the Ph.D. degree and the HDR from Université Paris-Est, Marne-la-Vallée, France, in 2000 and 2007, respectively. In 2001, he was a Post-Doctoral Researcher with the Université de Louvain, Belgium. In 2001, he joined the Communications and Electronics Department, Télécom ParisTech, as an Associate Professor, where he has been a (full) Professor since 2011. Since 2014, he has been a member of the IEEE Technical Committee Signal Processing for Communications and Networking. His research areas include statistical and digital signal processing and resource allocation. He served as an Associate Editor for the IEEE COMMUNICATIONS LETTERS from 2004 to 2007. From 2008 to 2012, he served as an Associate Editor and then a Senior Area Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING. Since 2018, he serves as Associate Editor for the IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS.