The Influence of CSI in Ultra-Reliable Low-Latency Communications with IR-HARQ

Apostolos Avranas*, Marios Kountouris*, and Philippe Ciblat†
*Mathematical and Algorithmic Sciences Lab, Paris Research Center, Huawei France
†Télécom ParisTech, 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 communication (URLLC), which requires extremely low latency (at msec order) with very high reliability (99.999%). In this work, we consider a URLLC system with incremental redundancy hybrid automatic repeat request (IR-HARQ) and investigate the effect of channel state information (CSI) at the transmitter on throughput and energy consumption optimization. For that, we analyze the feasibility region and the performance in block fading channels for the cases of full and statistical CSI. Our results show that the full CSI scheme is less robust and we also reveal a desirable balance between the trade-off quantities of energy and throughput.**

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. These applications lead to a Quality-of-Service (QoS), called ultra-reliable, low-latency communications (URLLC), which requires end-to-end latency of few milliseconds with a very large packet delivery success (> 99.999%).

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].

Restricting the size of the packets to meet the latency weakens the power of the coding scheme, so demanding on top huge reliability pushes the introduction of mechanisms that will facilitate the communication. In our paper, we investigate the potential benefits of the mechanism called incremental redundancy hybrid automatic repeat request (IR-HARQ). Retransmissions with feedback are used boosting reliability but under the cost of utilizing more time and energy when additional redundant information is asked by the user. So to achieve the reliability the price of higher energy and/or lower throughput has to be paid. The magnitude of this price cannot be independent of the channel. Poor channel quality can definitely render impossible the communication except if an excessive loss of throughput and energy is allowed. To leverage the impact of the channel unpredictability and mitigate the destructive effect of bad channel realizations, pilots can be send in order to assess the channel's quality. Acquiring Channel State Information (CSI) helps to better tailor the IR-HARQ and operate in a sweeter point of the trade-off between reliability,energy and throughput.

In this work, we quantify the impact of Channel State Information on the trade-off relationship between energy and latency. We investigate how the two common cases of CSI, i.e. *statistical* where only the statistics of the channel is known and *full* where the exact channel coefficient is available, influence the optimization of the IR-HARQ scheme. We generalize and extend significantly our previous works where we either solely minimized the energy [2] or maximize the throughput [3] under the simpler AWGN scenario.Throughput maximization is considered in [4] by optimizing the blocklength of a two round IR-HARQ. 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. Throughput maximization for IR-HARQ problem is considered in [7] assuming infinitely large blocklength and performing blocklength adaptation. In [8] for a variable-length stop feedback coding scheme delay violation and peak-age violation probabilities are analyzed. Under quality of service and energy efficiency requirements the authors of [9] use full CSI to optimize the powers that maximizes the effective capacity.

In this paper we demonstrate that performing purely energy minimization or throughput maximization leads to a bad trade-off point and a multi-objective optimization should be considered. Moreover, we analyze and compute the feasibility region of the HARQ-IR schemes and surprisingly we get with statistical CSI a bigger feasibility region than with full CSI. This fact yields full CSI less robust under URLLC conditions even without taking into account the shrinking of the latency constraint due to the training phase of learning the channel.

II. System Model

We consider a point-to-point communication link, where the transmitter has to convey B information bits within a certain predefined latency, expressed by a certain predefined maximum number of channel uses and 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. In case of retransmission, we consider hereafter IR-HARQ with M transmission rounds, i.e., $M-1$ retransmissions. By setting $M=1$, we recover the no-HARQ case. We denote n_m, $m \in \{1, 2, ..., \mathrm{M}\}$, 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 $\sum_{m=1}^{M} n_m \leq \mathrm{N}_\ell$.[1]

We consider a block flat fading channel, where the channel $h \in \mathbb{C}$ is an independent realization of an underlying random variable $\mathcal{H}$ following a specific distribution and remains constant in each block. The signal is also subject to additive white circularly-symmetric complex Gaussian random process with zero mean and unit variance. The IR-HARQ mechanism takes place within one block, i.e., there is only one channel coefficient value h for all retransmissions associated with the same bits. Consequently, we assume that the block duration is around N_ℓ. This is a relevant model for short-length packet communication and IoT applications, where point to point communication is performed. In the m-th round, the fragment (sub-codeword) $c_m \in \mathbb{C}^{n_m}$ is received with power $gP_m = \frac{\|h \cdot c_m\|^2}{n_m}$, where we defined the channel gain $g = |h|^2$.

III. Problem Statement

The problem we study here is that of optimizing the IR-HARQ mechanism by tuning the blocklengths and the powers so as to minimize a multi-objective function, involving energy consumption and throughput. We require a low error probability ε_{rel} without consuming more than energy E_b and within a latency N_ℓ.

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 $(n_1, ...n_m, P_1, ...P_m, g)$. 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 $\epsilon_m = \mathbb{P}\left(\cap_{i=1}^{m} \Omega_i\right)$ where Ω_i is the event corresponding to "the concatenation of the first i fragments of the parent codeword is not correctly decoded when optimal coding is employed".

For *infinite* blocklength, an error occurs when the mutual information is below a threshold and for IR-HARQ it can easily be seen that for $i < j$ we have $\Omega_i \subseteq \Omega_j$ [10], [11], 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 (for instance [4], [11], [12]) as well as in this paper, the exact ϵ_m is replaced with the simpler ε_m defined as $\varepsilon_m = \mathbb{P}(\Omega_m)$, since ε_m and ϵ_m are numerically close for $m > 1$ (for $m = 1$ they coincide). Then, ε_m can be upper bounded [1, Lemma 14 and Theorem 29] and also lower bounded as in [12] by employing the $\kappa\beta$-bounds proposed in [1]. Both bounds have the same first two dominant terms and the error probability after taking into account the scaling of the power caused by the block fading, is approximately given by

$$\varepsilon_m \approx Q\left(\frac{\displaystyle\sum_{i=1}^{m} n_i \log(1+gP_i) - B\log 2}{\sqrt{\displaystyle\sum_{i=1}^{m} n_i \left(1 - \frac{1}{(1+gP_i)^2}\right)}}\right) \quad (1)$$

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., $\varepsilon_m(n_1, ...n_m, P_1, ...P_m, g)$ instead of ε_m.

IV. Optimization

Unlike our previous work [2], [3] for AWGN channels, here we consider a more realistic channel setting. We first assume that the transmitter knows the channel coefficient h, which we refer to as "full CSI" but the channel is varying block by block. We then consider that only the channel distribution $\mathcal{H}$, referred to as "statistical CSI". Both configurations are analyzed for optimizing a weighted sum of the average throughput and energy consumption.

Throughput is defined as the average ratio of successfully decoded bits divided by the number of symbols used. Given a channel realization (and so its gain g), the expected throughput can be derived using the renewal theory [13] where the expected value of delay is $\sum_{m=1}^{M} n_m \varepsilon_{m-1}$ and the expected reward is $B(1-\varepsilon_M)$ which leads to

$$\mathcal{T}_h(0) = \frac{B(1-\varepsilon_2)}{n_1 + n_2\varepsilon_1}.$$

The expected energy spent for transmitting B information bits (conditioned on the channel realization) is

$$\mathcal{E}(1) = n_1 P_1 + n_2 P_2 \varepsilon_1.$$

[1] Penalty terms $D(n_1, ...n_m)$ 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 focus on the simplified version where $D(n_1, ...n_m) = 0$.

A. Full CSI

Our optimization problem is cast as follows.

Problem 1: *Full CSI problem.*

$$\min_{n_1(g),n_2(g),P_1(g),P_2(g)} \quad \mathbb{E}_g\left[-\frac{\mathcal{T}_h(a)}{\mathcal{T}_{h,max}}+\frac{\mathcal{E}(a)}{\mathcal{E}_{min}}\right] \tag{2}$$

$$\textit{s.t.} \quad n_1(g)+n_2(g)\leq \mathrm{N}_\ell, \quad \forall g \tag{3}$$

$$\mathbb{E}_g[\varepsilon_2(n_1(g),n_2(g),P_1(g),P_2(g),g)]\leq \varepsilon_{\mathrm{rel}} \tag{4}$$

$$n_1(g)P_1(g)+n_2(g)P_2(g)\leq \mathrm{E}_b, \quad \forall g \tag{5}$$

$$P_i(g)\leq \mathrm{P}_{\max}, \quad i\in\{1,2\} \quad \forall g \tag{6}$$

where

- $\mathcal{T}_h(a)=(1-a)\mathcal{T}_h(0)$,
- $\mathcal{E}(a)=a\mathcal{E}(1)$. *So the variable a is a weight balancing throughput maximization and energy minimization.*
- $\mathbb{E}_g[\cdot]$ *is the expectation over the channel gain realizations.*
- $\mathcal{T}_{h,max}=\max\mathbb{E}_g[T_h(0)]$ *s.t.* (3), (4), (5), (6) *hold and* $\mathcal{E}_{min}=\min\mathbb{E}_g[E(1)]$ *s.t.* (3), (4), (5), (6) *hold.*

In the following, we assume

$$\mathrm{P}_{\max}\geq \frac{\mathrm{E}_b}{\mathrm{N}_\ell} \tag{7}$$

such that the solutions of the Problem 1 may consume the maximum energy budget E_b.

As the channel is known, we adapt the blocklengths and powers accordingly. Therefore the solutions of the optimization problem depend on the channel gain realization g. To simplify the problem, we consider the simple yet intuitive particular case where transmissions are avoided over deep fading. Mathematically, the proposed solutions satisfy:

$$n_i=\begin{cases}0 & g<g_{th}\\ n_i(g) & g\geq g_{th}\end{cases}, P_i=\begin{cases}0 & g<g_{th}\\ P_i(g) & g\geq g_{th}\end{cases}$$

In addition, we force each transmission (when done) to achieve the same error probability, s.t. $\forall g \geq g_{th}$, $\varepsilon_2(n_1(g),n_2(g),P_1(g),P_2(g))=\varepsilon_{on}$. The reliability constraint (4) leads to

$$\varepsilon_2=\mathbb{P}(g<g_{th})+\mathbb{P}(g\geq g_{th})\varepsilon_{on}\leq\varepsilon_{\mathrm{rel}}. \tag{8}$$

This simplification enables us to compute g_{th} given ε_{on} and to decouple the problem by treating every g_0 with $g_0\geq g_{th}$ individually. We just have to solve Problem 1 assuming that the channel gain takes only the value g_0 and replacing $\varepsilon_{\mathrm{rel}}$ by ε_{on}. An additional simplification is applied (similar proof to Proposition 3 in [3]) that asserts $\varepsilon_2\approx\varepsilon_{\mathrm{rel}}$. This means that trying to achieve lower error probability than the required ε_{on} (whenever $g\geq g_{th}$) results in waste of energy and blocklength resources, which also leads to a throughput decrease.

Notice that it is not always possible to meet the constraints and to get a non-empty feasible set if the average channel gain average is very low or the available resources are very scarce. The following lemma characterizes the feasibility set.

Lemma 1: The solution of the problem:

$$\min_{n_1,\dots,n_M,P_1,\dots,P_M,M} \quad \varepsilon_M(n_1,\dots n_m,P_1,\dots P_m,g) \tag{9}$$

$$\textit{s.t.} \quad \sum_{i=1}^{M} n_i\leq \mathrm{N}_\ell \tag{10}$$

$$\sum_{i=1}^{M} n_iP_i\leq \mathrm{E}_b \tag{11}$$

is $M=1$ with $(n_1,P_1)=(\mathrm{N}_\ell,\frac{\mathrm{E}_b}{\mathrm{N}_\ell})$. For meaningful/practical solutions, we restrict to

$$n_i\geq Q^{-1}(10^{-9})\approx 36, \textit{ and} \tag{12}$$

$$\max\{Q(0.45\sqrt{B\ln 2}),10^{-9}\}<\varepsilon_M<0.5. \tag{13}$$

Proof: See Appendix A ■

Lemma 1 tells us that the best blocklength-power allocation of IR-HARQ within a coherence block for minimizing the outage probability given a maximum amount of energy and channel uses is to employ one packet consuming all the available blocklength and energy. Infeasibility occurs if there is at least one $g_0\geq g_{th}$ such that whatever the configuration of $n_i(g_0)$ and $P_i(g_0)$ for given $(\mathrm{N}_\ell,\mathrm{E}_\mathrm{b})$, it is $\varepsilon_2>\varepsilon_{on}$. Otherwise stated, when $\min\varepsilon_2>\varepsilon_{on}$ for given $(\mathrm{N}_\ell,\mathrm{E}_\mathrm{b})$, we know that the feasible set is empty. In addition, when $M=1$, it is easy to check that the minimum error probability is decreasing as the channel gain gets larger. So the infeasibility can be checked only for the worst channel $g_0=g_{th}$. Consequently, if $\varepsilon_2(\mathrm{N}_\ell,0,\frac{\mathrm{E}_b}{\mathrm{N}_\ell},0,g_{th})<\varepsilon_{on}$, the feasible set is not empty.

B. Statistical CSI

We now assume that only the distribution of the channel $\mathcal{H}$ is known. As the channel realization is not known in advance and changes independently every coherence block, we cannot adapt the blocklengths and powers at each time. Therefore, we find a blocklength-power configuration independent of the channel gain g.

Problem 2: *Statistical CSI problem.*

$$\min_{n_1,n_2,P_1,P_2} \quad \mathbb{E}_g\left[-\frac{\mathcal{T}_h(a)}{\mathcal{T}_{h,max}}+\frac{\mathcal{E}(a)}{\mathcal{E}_{min}}\right] \tag{14}$$

$$\textit{s.t.} \quad n_1+n_2\leq \mathrm{N}_\ell \tag{15}$$

$$\mathbb{E}_g[\varepsilon_2(n_1,n_2,P_1,P_2,g)]\leq\varepsilon_{\mathrm{rel}} \tag{16}$$

$$n_1P_1+n_2P_2\leq \mathrm{E}_b, \tag{17}$$

$$P_i\leq \mathrm{P}_{\max}, \quad i\in\{1,2\} \tag{18}$$

where

- $\mathcal{T}_{h,max}=\max\mathbb{E}_g[T_h(0)]$ *s.t.* (15), (16), (17), (18) *hold,* $\mathcal{E}_{min}=\min\mathbb{E}_g[E(1)]$ *s.t.* (15), (16), (17), (18) *hold.*

Again Lemma 1 can be employed to check easily the feasibility given the resources $(\mathrm{N}_\ell,\mathrm{E}_b)$. Since the configuration $(n_1,n_2,P_1,P_2)=(\mathrm{N}_\ell,0,\frac{\mathrm{E}_b}{\mathrm{N}_\ell},0)$ does not depend on g and minimizes the error probability, we can see that $\mathbb{E}_g[\varepsilon_2(\mathrm{N}_\ell,0,\frac{\mathrm{E}_b}{\mathrm{N}_\ell},0,g)]\leq\varepsilon_{\mathrm{rel}}$ leads to a non-empty feasible set. Finally, we can again assert (Proposition 3 in [3]) $\mathbb{E}_g[\varepsilon_2(\mathrm{N}_\ell,0,\frac{\mathrm{E}_b}{\mathrm{N}_\ell},0,g)]\approx\varepsilon_{\mathrm{rel}}$.

V. Numerical Results and Discussion

We assume $B = 256$ information bits (32 bytes) has to be transmitted through a Ricean fading channel with K-factor and unit-variance, i.e. $|h| \sim Rice(K, 1)$. The K-factor represents the ratio between the direct path (Line Of Sight) and the other paths. $K = 0$ corresponds to the Rayleigh fading while $K \to \infty$ corresponds to the AWGN. We also assume that $n_1 \geq 100$ such that Polyanskiy's formula approximation (1) is accurate and also that $\varepsilon_{rel}, \varepsilon_{on} \in [10^{-9}, 0.5]$ to satisfy Eq. (13). Due to space limitations, we omit the details on how the solutions are found; we just mention that for the full CSI case, it ends up to a 4D grid search, while for the statistical CSI to a 3D grid search.

In Figure 1, we depict the feasibility regions in (E_b, K) for different CSI configurations and different N_ℓ. For the same constraints in latency N_ℓ and reliability ε, surprisingly the feasibility region for full CSI is smaller than the one with only statistical CSI. We remind that for full CSI, the transmitter policy is to remain idle when $g < g_{th}$, so additional resources are needed when it is active to achieve a pre-fixed outage probability ε_{on} smaller than ε_{rel} to compensate for. The full CSI policy is more constrained. The threshold g_{th} cannot be tuned to zero since we force for every $g \geq g_{th}$ an error probability $\varepsilon_{on} \leq \varepsilon_{\text{rel}}$ to be achieved and this requires an infinite amount of resources when $g {\to} 0$. We also observe that the reliability constraint ε strongly affects the feasibility region, while this is not the case for the latency constraint N_ℓ. We emphasize that, as we will see later, when both CSI setups are feasible, the full CSI outperforms the statistical one. In Figure 2, we plot the relative throughput $\frac{\mathcal{T}_h(a)}{\mathcal{T}_h, max}$ (left scale) and relative energy $\frac{\mathcal{E}(a)}{\mathcal{E}_{min}}$ (right scale) versus a after solving Problem 1 (full CSI) and Problem 2 (statistical CSI). Performing either throughput maximization ($a = 0$) or energy minimization ($a = 1$) is not a good strategy since by allowing a small decrease of throughput (in the first case) or a small increase of energy (in the second case), the other metric in the objective function significantly improves. A good tradeoff for full and statistical CSI is around $a = 0.3$ in the employed here.

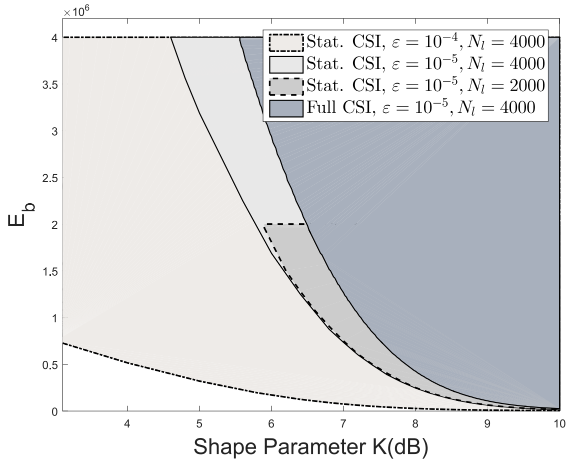

Fig. 1: Feasibility region for different channel, B = 32Bytes, maximum energy budget $\mathrm{E}_b = P_{\max}\mathrm{N}_\ell$ with $P_{\max} = 30dB$.

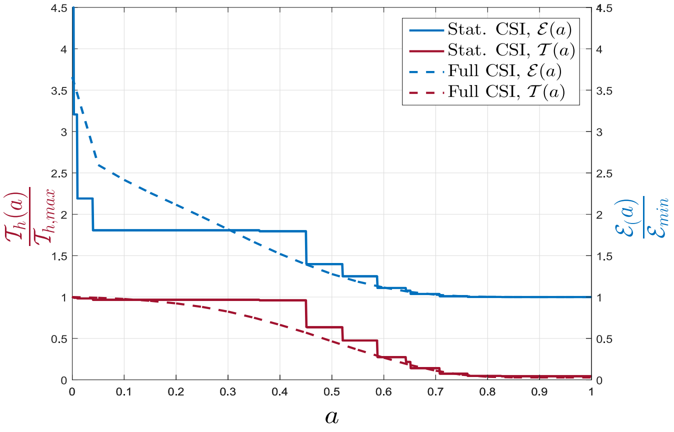

Fig. 2: Throughput and energy relative to their optimal value for Rician channel with K = 7dB, B = 32Bytes, $\varepsilon_{\text{rel}} = 10^{-5}$ and maximum energy $\mathrm{E}_b = P_{\max}\mathrm{N}_\ell$ with $P_{\max} = 30dB$ and $\mathrm{N}_\ell = 4000$.

In Figure 3, we display the throughput and energy obtained as a function of a for different setups. We remark that the K factor as well as the target reliability play the important role. On the contrary, the constraints on latency $\mathrm{N_l}$, energy $\mathrm{E_b}$ and power $\mathrm{P_{max}}$ seem to have a minor impact except when they are so stringent that we go close to the boundary of the feasibility area.

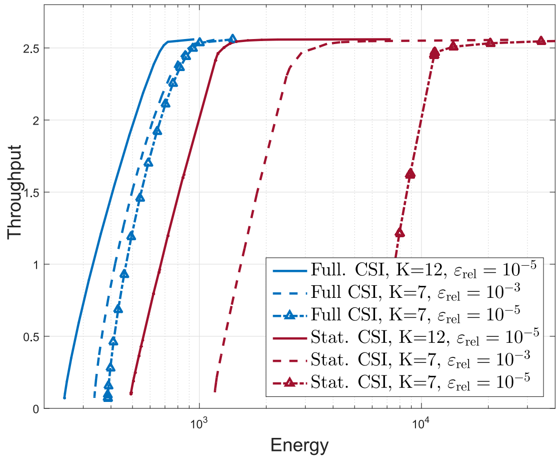

Fig. 3: Throughput versus energy for $a \in [0, 1]$, with $\mathrm{E}_b = P_{\max}\mathrm{N}_\ell$, $P_{\max} = 30dB$, and $\mathrm{N}_\ell = 4000$.

In Figure 4, we display the throughput and energy obtained by moving a from 0 to 1 when HARQ is carried out or when one shot transmission is employed. With full CSI a constant 37% percent, according to the figure, of energy can be saved for the same throughput by using HARQ instead of one-shot. This gain for statistical CSI scheme depends substantially on the channel quality K and it can become huge for poor channel conditions.

To explain this behavior we first discuss the optimal configuration of (n_1, n_2, P_1, P_2). The first packet is of significant importance since we measure average performance and the first packet is always sent whereas the second only ε_1 times. For throughput maximization n_1 should be kept as small as possible at the expense of power P_1. However, as we move to energy minimization, the situation is reversed, as larger n_1 with smaller P_1 reduces required energy [2]. The role of the second packet is mainly to successfully meet the constraints of the optimization problem and not to improve the objective. This behavior is similar for both full and statistical CSI.

For statistical CSI, where the channel coefficient is unknown, we see the mechanism of the optimized HARQ rendering the first packet responsible for achieving a good value of the objective function when the channel is good and employing the second only when the channel is bad and necessarily a lot of resources must be spent. In one-shot there is not this option of differentiating the good and bad realizations of the channel and the bad channel realizations determine the amount of resources needed to spend for all cases. Reasonably, as channel statistics deteriorate (K decreases) the waste of resources in one-shot scheme becomes more profound since the bad channel realizations determining the expenditure of resources get worse. On the contrary, in the case of full CSI the surprising savings do not happen as the channel is known also for the one-shot scheme. Now there can be a distinction between good and bad channel realizations. Moreover, this results to an almost constant save of energy given a specific throughput, independently of channel quality.

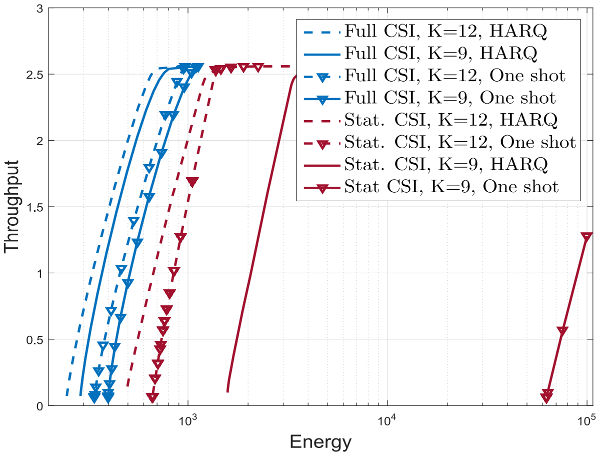

Fig. 4: Throughput versus energy for $a \in [0,1]$ when HARQ or one shot transmission is used, with $\mathrm{E}_b = P_{\max}\mathrm{N}_\ell$, $P_{\max} = 1000(30dB)$, and $\mathrm{N}_\ell = 4000$.

VI. Conclusion

In this paper, we have investigated the problem of IR-HARQ optimization for URLLC in fading channels assuming both statistical and full CSI. Considering a weighted sum of throughput and energy consumption as our objective function, we have analytically characterized the feasibility region in both CSI cases and solved the optimization problems. A key implication of our results is that the full CSI case turns out to be less robust than the statistical CSI one. Furthermore, the primary factor affecting the throughput-energy trade-off is not the latency constraint but the target reliability and the channel quality. Finally, we showed that employing an IR-HARQ mechanism can yield considerable gains especially with statistical CSI.

Appendix A
Proof of Lemma 1

We first consider $M = \mathrm{N}_\ell$ and $n_i = 1,\ \forall i$ so each symbol chooses its own power P_i and we want to prove that $P_i = \frac{\mathrm{E}_b}{\mathrm{N}_\ell},\ \forall i$ is the solution of the optimization problem. If it is true, these P_i can get out of the sums of the error formula (1), leaving $\sum_i^{\mathrm{N}_\ell} 1 {=} \mathrm{N}_\ell$. Then the optimal error probability can be expressed versus N_ℓ and $\frac{\mathrm{E}_b}{\mathrm{N}_\ell}$ which is equivalent to choose one block of size N_ℓ with identical power $\frac{\mathrm{E}_b}{\mathrm{N}_\ell}$.

Like [2], it is straightforward to prove that using full resources (which means forcing the constraints to be equalities) is beneficial for reliability. Moreover since Q-function is decreasing and the logarithm is increasing, we can alter the objective function and we end up to

$$\max_{x_1,\dots,x_{\mathrm{N}_\ell}} \log\Big(\sum_{i=1}^{\mathrm{N}_\ell}\log(\frac{1}{x_i}){-}B\log 2\Big){-}\frac{1}{2}\log\Big(\sum_{i=1}^{\mathrm{N}_\ell}(1{-}x_i^2)\Big) \quad (19)$$

$$s.t. \quad \sum_{i=1}^{\mathrm{N}_\ell}\frac{1}{x_i} = \tilde{\mathrm{E}} \quad (20)$$

where $x_i = \frac{1}{1+h^2P_i}$ and $\tilde{\mathrm{E}} = \mathrm{N}_\ell{+}h^2\mathrm{E}_b$. So $x_i \in [1/\tilde{\mathrm{E}}, 1]$. The domain on which we maximize is a compact set, thus a global maximum should exist. Additionally, the interval boundary, i.e. $x_i \in \{1/\tilde{\mathrm{E}}, 1\}$ represents the cases where some symbols vanish ($x_i = 1 \Leftrightarrow P_i = 0$ and $x_i = 1/\tilde{\mathrm{E}} \overset{(20)}{\Leftrightarrow} x_j = 0 \quad \forall j \neq i$) which yield suboptimal error probabilities [2] and so the global maximum cannot be on the interval boundary. We use KKT conditions to prove that there is only one stationary point for the above problem and this point is when all x_i are equal to each other, and so these x_i are optimal.

Applying the KKT conditions with λ the Lagrangien multiplier associated with (20), we get the set of equations

$$-\frac{x_i^3}{V} + \frac{x_i}{A} = \lambda, \quad \forall i \in \{1,2,...,\mathrm{N}_\ell\} \quad (21)$$

with $A = -\sum_{i=1}^{\mathrm{N}_\ell}\log(x_i) - B\log 2$ and $V = \sum_{i=1}^{\mathrm{N}_\ell}(1-x_i^2)$. Let us assume that the solution of (21) is $\vec{x}^\star = (x_1^\star, ..., x_{\mathrm{N}_\ell}^\star)$ and denote $A^\star = A(\vec{x}^\star)$, $V^\star = V(\vec{x}^\star)$. $A^\star$ and $B^\star$ are the same for each equation in (21). If we can find more than three different elements of $\vec{x}^\star$, then the cubic polynomial $-\frac{x^3}{V^\star} + \frac{x}{A^\star} - \lambda = 0$ has more than three roots which is impossible. Additionally as $A^\star$, $B^\star$, and $x_i^\star$ are positive by construction, we can show that $x_i^\star$ can at most take two different values. Let us denote them by $(\tilde{x}_1, \tilde{x}_2)$. The value $\tilde{x}_1$ is taken by n_1 out of N_ℓ x_i-variables while the value $\tilde{x}_2$ is taken by $n_2 = \mathrm{N}_\ell - n_1$ x_i-variables. Then (20) and (21) can be transformed into

$$n_1 + n_2 = \mathrm{N}_\ell \quad (22)$$

$$\frac{n_1}{\tilde{x}_1} + \frac{n_2}{\tilde{x}_2} = \tilde{\mathrm{E}} \quad (23)$$

$$-\frac{\tilde{x}_1^3}{V}+\frac{\tilde{x}_1}{A}=-\frac{\tilde{x}_2^3}{V}+\frac{\tilde{x}_2}{A} \tag{24}$$

For instance, the case $\tilde{x}_1=\tilde{x}_2=\frac{\tilde{\mathrm{E}}}{\mathrm{N}_\ell}$ is a solution. Actually, it corresponds to our desired stationary point. It just remains to prove that this is the only solution.

For $\tilde{x}_1 \neq \tilde{x}_2$:

$$(24) \Leftrightarrow A(\tilde{x}_1^2+\tilde{x}_1\tilde{x}_2+\tilde{x}_2^2)=V. \tag{25}$$

We will show that (25) and the assumptions (12) and (13) cannot all hold at the same time. Using (13) we get:

$$A>0 \Leftrightarrow n_1\log(\tilde{x}_1)+n_2\log(\tilde{x}_2)<-B\log 2 \tag{26}$$

$$\frac{A}{\sqrt{V}}<F \overset{(25)}{\Leftrightarrow} \sqrt{n_1(1-\tilde{x}_1^2)+n_2(1-\tilde{x}_2^2)}<F(\tilde{x}_1^2+\tilde{x}_1\tilde{x}_2+\tilde{x}_2^2) \tag{27}$$

with $F=\min\{0.45\sqrt{B\log 2}, Q^{-1}(10^{-9})\}$ and the change from $\max$ of (12) to $\min$ is due to the decreasing monotonicity of $Q(\cdot)^{-1}$.

In Figure 5, we display the area where (26) holds in blue, and the area where (27) holds in grey. We want to prove that both blue and black areas are disjoint in order to have no solution satisfying both inequalities. It is easy to prove that the boundary-curve C_1 (resp. C_2) is convex (resp. concave). So to avoid common points between both areas, the points K_2 and K_3 (intersection point of curve C_2 with $\tilde{x}_2=1$ and $\tilde{x}_1=1$ respectively) have not to belong in the blue area.

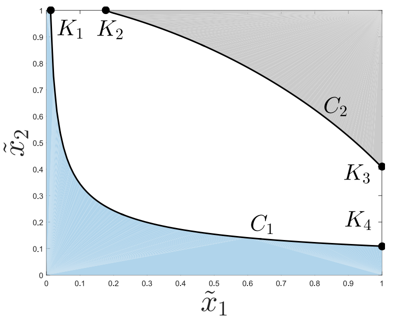

Fig. 5: Inequalities description for $\tilde{x}_1$ and $\tilde{x}_2$.

The point $K_2=(e^{-\frac{B\log 2}{n_1}},1)$ does not belong in the blue area if it does not satisfy (27), i.e. for $n_1\log(\tilde{x}_1)=-B\log 2$ we want either (28) or (29) to hold:

$$\sqrt{n_1(1-\tilde{x}_1^2)}>0.45\sqrt{B\log 2}(\tilde{x}_1^2+\tilde{x}_1+1), \tag{28}$$

$$\sqrt{n_1(1-\tilde{x}_1^2)}>Q^{-1}(10^{-9})(\tilde{x}_1^2+\tilde{x}_1+1). \tag{29}$$

First we concentrate on (28). After substitution we want to show that:

$$\sqrt{\frac{\tilde{x}_1^2-1}{\log\tilde{x}_1}}>0.45(\tilde{x}_1^2+\tilde{x}_1+1). \tag{30}$$

A known inequality is $\log(x)\geq\frac{x-1}{\sqrt{x}}$, for $x\leq 1$. By dividing with $1-x^2(>0)$ we can get $\sqrt{\frac{x^2-1}{\log x}}\geq\sqrt{\sqrt{x}(1+x)}$. Furthermore for $0<x\leq 1$, we have $2x+1\geq x^2+x+1$. If

$$\sqrt{\sqrt{\tilde{x}_1}(1+\tilde{x}_1)}\geq 0.45(2\tilde{x}_1+1) \tag{31}$$

holds, then (30) holds. Proving (31) is equivalent to show $\sqrt{\tilde{x}_1}^4-1.2346\sqrt{\tilde{x}_1}^3+\sqrt{\tilde{x}_1}^2-1.2346\sqrt{\tilde{x}_1}+0.25\leq 0$. The roots of this fourth-order polynomial can analytically be found and the inequality is satisfied when $\tilde{x}_1\geq\rho^2=0.0563$. So (30) is satisfied for $\tilde{x}_1\geq\rho^2$. For $\tilde{x}_1<\rho$, one can see it is equivalent to $0.45\frac{\tilde{x}_1^2+\tilde{x}_1+1}{\sqrt{1-\tilde{x}_1^2}}<0.45\frac{\rho^2+\rho+1}{\sqrt{1-\rho^2}}$. If $\tilde{x}_1>e^{-\frac{1-\rho^2}{0.45^2(\rho^2+\rho+1)^2}}(\approx 0.0125)$, then $0.45\frac{\rho^2+\rho+1}{\sqrt{1-\rho^2}}<\sqrt{\frac{-1}{\log\tilde{x}_1}}$ and again (30) holds. To sum up, when $\tilde{x}_1>0.0125$ the point K_2 is outside the blue area.

Now we will concentrate on (29) to treat the case of $\tilde{x}_1\leq 0.0125$. From (29), we have:

$$n_1>Q^{-1}(10^{-9})\frac{(\tilde{x_1}^2+\tilde{x_1}+1)^2}{1-\tilde{x}_1^2}\approx Q^{-1}(10^{-9})$$

which holds according to the assumption done in the Lemma. Similar procedure can be applied for the point K_3 which concludes the proof.

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.*, vol. 36, no. 11, pp. 2475–2485, Nov. 2018.

[3] ——, "Throughput maximization and IR-HARQ optimization for URLLC traffic in 5g systems," in *IEEE International Conference on Communications (ICC)*, May 2019.

[4] 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.

[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] 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.

[8] R. Devassy, G. Durisi, G. C. Ferrante, O. Simeone, and E. Uysal, "Reliable transmission of short packets through queues and noisy channels under latency and peak-age violation guarantees," Feb. 2019. [Online]. Available: https://arxiv.org/abs/1806.09396

[9] Y. Hu, M. Ozmen, M. C. Gursoy, and A. Schmeink, "Optimal power allocation for qos-constrained downlink multi-user networks in the finite blocklength regime," *IEEE Trans. on Wireless Commun.*, vol. 17, no. 9, pp. 5827–5840, Sep. 2018.

[10] 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.

[11] 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.

[12] 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.

[13] R. Wolff, *Stochastic modeling and the theory of queues*. Upper Saddle River, NJ, U.S.A.: Prentice Hall, 1989.