Low-Latency Short-Packet Transmissions: Fixed Length or HARQ?

(article starts on next page)

Low-Latency Short-Packet Transmissions: Fixed Length or HARQ?

Johan Östman, Rahul Devassy, Guido Carlo Ferrante, Giuseppe Durisi
Chalmers University of Technology, Gothenburg, Sweden

Abstract—We study short-packet communications, subject to latency and reliability constraints, under the premises of limited frequency diversity and no time diversity. The question addressed is whether, and when, hybrid automatic repeat request (HARQ) outperforms fixed-blocklength schemes with no feedback (FBL-NF) in such a setting. We derive an achievability bound for HARQ, under the assumption of a limited number of transmissions. The bound relies on pilot-assisted transmission to estimate the fading channel and scaled nearest-neighbor decoding at the receiver. We compare our achievability bound for HARQ to state-of-the-art achievability bounds for FBL-NF communications and show that for a given latency, reliability, number of information bits, and number of diversity branches, HARQ may significantly outperform FBL-NF. For example, for an average latency of 1 ms, a target error probability of 10^{-3}, 30 information bits, and 3 diversity branches, the gain in energy per bit is about 4 dB.

I. Introduction

Fifth generation (5G) wireless systems are envisioned to support communications with stringent requirements on reliability and latency—the so called ultra-reliable low-latency communications (URLLC) [1]. The objective is to enable use cases like smart grids and wireless industrial control. While previous-generation wireless systems focused mostly on the support of large data rates, URLLC systems target transmission of small data payloads, carried by short coded packets to meet the latency requirements.

Performance analyses of wireless systems are often carried out using asymptotic information-theoretic metrics like ergodic capacity and outage capacity, under the implicit assumption that a message is encoded into an arbitrarily large number of coded symbols. Although a good approximation for systems where latency concerns are secondary, such performance analyses are not suitable for URLLC [2]. Instead, to study the performance of URLLC links, one must resort to tools that allow for the analysis of communication systems in the nonasymptotic regime where a message is encoded onto a finite, often small, number of symbols.

For communications with fixed blocklength and no feedback (FBL-NF), the nonasymptotic information-theoretic tools developed in [3] have recently enabled the characterization of the maximum coding rate achievable for a given blocklength and error probability in many scenarios of practical relevance for 5G and beyond. For example, it has allowed for the study of the rate achievable with short packets on general quasi-static fading channels [4] and multiple-input multiple-output (MIMO) Rayleigh block-fading channels [5]. Furthermore, the practically relevant case of pilot-assisted transmission (PAT)

and scaled nearest-neighbor (SNN) decoding has been analyzed for single-input single-output (SISO) Rician block-fading channels [6] and for MIMO Rayleigh block-fading channels [7].

Finite-blocklength information theory prescribes that for several channels, including the AWGN channel, short-packet transmission incurs a rate loss from capacity roughly proportional to the inverse of the square root of the blocklength [3]. However, when one is allowed to use variable-length stop-feedback (VLSF) codes, the picture is different. Specifically, capacity is approached much faster in the blocklength [8]. VLSF codes are an instance of variable-length feedback codes where the encoder keeps on transmitting coded symbols until it receives a single-bit feedback sent by the decoder to inform the encoder that decoding is complete. Incremental redundancy hybrid automatic repeat request (HARQ) is an instance of VLSF codes. In [8], decoding is attempted upon reception of every new symbol. This setup was later extended to blocks of symbols in [9], a scenario that is more relevant for HARQ. Under the assumption that the received symbols are discarded and transmission starts over if decoding is not successful after a finite number of blocks, the rate penalty resulting from decoding after each block of symbols rather than after each symbol was characterized in [9] for the AWGN channel. However, such strategy still outperforms FBL-NF codes for moderate latencies.

In [10], FBL-NF and HARQ are compared for the AWGN channel by taking into account the delay caused by the transmission of a positive/negative acknowledgement (ACK/NACK). The analysis, however, does not account for the event of an undetected error, i.e., the event that an ACK is fed back although the decision of the decoder is wrong. Furthermore, the normal approximation [3] is used in the analysis. However, this approximation, which is based on the Berry-Esseen theorem, is accurate only when the blocklength and the error probability are not too small and may not be suitable for URLLC.

Contributions: We consider an HARQ scheme employing PAT and SNN decoding at the receiver for transmission of short data packets over a SISO block-fading channel. Leveraging [8, Thm. 3], which provides an extension of the dependence-testing bound [3, Thm. 17] for FBL-NF to VLSF codes, we derive an achievability bound on the minimum energy per bit required to transmit a small information payload under a given latency and reliability target. Differently from [8, Thm. 3], where decoding stops when the accumulated information density corresponding to one of the codewords exceeds a threshold, in our setup decoding stops when the accumulated *generalized* information density exceeds a threshold or the number of transmissions reaches a maximum predetermined value.

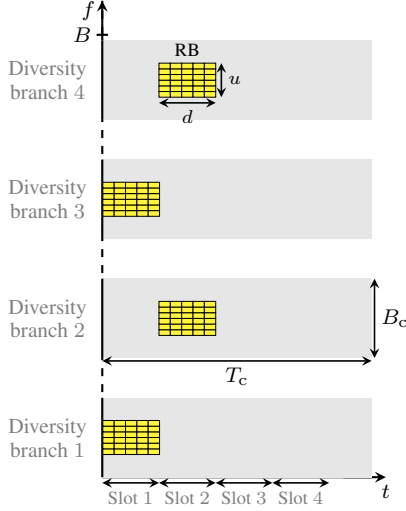Our bound depends on the cumulative distribution function

Fig. 1. Example of the FBL-NF scheme with parameters $u = 6$, $d = 5$, $L_c = 4$, $L = 2$, and $v = 2$.

Fig. 2. Two rounds of HARQ transmission with parameters $u = 6$, $d = 5$, $L_c = 4$, $L = 2$, and $\ell_{\max} = 2$.

(CDF) of the decoding time, i.e., the time at which the decoder stops and makes a decision. This quantity has recently been shown to be central in joint coding-queuing analyses of the delay-violation probability in URLLC systems [11]. Our analysis provides means to assess whether HARQ results in lower minimum energy per bit and higher maximum coding rate than FBL-NF. In particular, our numerical results show that HARQ may significantly outperform FBL-NF for low-latency targets.

Notation: We denote random vectors and scalars by upper-case bold and standard letters, such as $\boldsymbol{X}$ and X, respectively, and their realizations by lower-case letters. The identity matrix of size $a \times a$ is written as $\boldsymbol{I}_a$. The distribution of a circularly-symmetric complex Gaussian random variable with variance σ^2 is denoted by $\mathcal{CN}(0, \sigma^2)$. The superscript $(\cdot)^{\mathsf{H}}$ denotes Hermitian transposition. We write $\log(\cdot)$ and $\log_2(\cdot)$ to denote the natural logarithm and the logarithm to the base 2, respectively. Finally, $[a]^+$ stands for $\max\{0, a\}$, $\|\cdot\|$ denotes the Euclidean norm, and $\mathbb{E}[\cdot]$ is the expectation operator.

II. SYSTEM MODEL

A. Setup

We consider a memoryless SISO Rayleigh block-fading channel in which fading is assumed to stay constant within a coherence block and to change independently across coherence blocks. The channel coherence time and coherence bandwidth are denoted by T_c and B_c, respectively. For a system bandwidth B, the available number of *diversity branches* is $L_c = \lfloor B/B_c \rfloor$. The duration of a transmitted codeword is assumed to be much smaller than the coherence time. Hence, no time diversity is available. This is in line with commonly used fourth generation (4G) and 5G channel models, as discussed in Section IV. We consider orthogonal frequency-division multiplexing (OFDM) and assume that a resource block (RB) consists of d OFDM symbols, each spanning u subcarriers. Therefore, an RB consists of $n_c = ud$ symbols. The time interval over which an RB is transmitted is referred to as slot, in accordance with the 3rd generation partnership project (3GPP) terminology. We let L be the number of diversity branches used
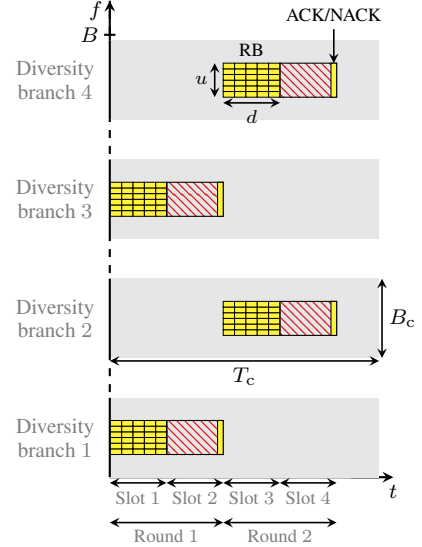
in a slot out of the available L_c. Hence, L RBs, amounting to Ln_c symbols, are transmitted in each slot, as shown in Fig. 1.

We shall consider two communication schemes: FBL-NF and HARQ. In the FBL-NF case, a codeword is transmitted over L diversity branches per slot and v slots in total. Hence, a codeword consists of vLn_c symbols, see Fig. 1. We assume that transmissions on different slots occur on different diversity branches, so that transmitted symbols in different slots experience independent fading. This means that the number of slots per codeword is no larger than $\ell_{\max} = \lfloor L_c/L \rfloor$. At the receiver side, decoding is performed once the entire codeword is received.

In the HARQ case, we consider one direction of a bidirectional communication link. We assume that communication is organised in rounds, each consisting of two consecutive slots within the same diversity branches (time-division duplexing). The two communicating devices are assigned one slot each in every round. We assume that transmissions on different rounds occur on different diversity branches (see Fig. 2). Hence, transmitted symbols in different rounds experience independent fading. Consequently, the maximum number of rounds allowed is $\ell_{\max}$. In each round, a device transmits Ln_c symbols during one slot and listens for an ACK/NACK and possibly other data during the next slot. At the other device (receiver), decoding is attempted after the first slot. In the second slot, the receiver piggybacks an ACK/NACK on its data, to inform the transmitter whether to continue or to terminate transmission. If a NACK is piggybacked, incremental redundant symbols are sent during the next round. Rounds go on until the transmitter observes an ACK or it reaches the maximum predetermined number of rounds. Such an ACK/NACK transmission introduces a delay compared to FBL-NF transmission, which we will take into account in our analysis. We assume that the feedback delay per round amounts to one slot. However, our analysis can be easily generalized to arbitrary feedback delays. As illustrated in Fig. 2, when the same amount of resources are used in HARQ (per communication direction) and FBL-NF, the FBL-NF scheme requires half of the time of the HARQ scheme.

The input-output relation for the slot assigned for the forward transmission in round j, satisfies

$$P_{\boldsymbol{Y}_j | \boldsymbol{Y}_1, \ldots \boldsymbol{Y}_{j-1}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_j} = P_{\boldsymbol{Y}_j | \boldsymbol{X}_j} = \prod_{k=1}^{L} P_{\boldsymbol{Y}_{j,k} | \boldsymbol{X}_{j,k}} \quad (1)$$

with

$$\boldsymbol{Y}_{j,k} = H_{j,k} \boldsymbol{X}_{j,k} + \boldsymbol{W}_{j,k}. \quad (2)$$

Here, $\boldsymbol{Y}_j = [\boldsymbol{Y}_{j,1}, \ldots, \boldsymbol{Y}_{j,L}]$ and $\boldsymbol{X}_j = [\boldsymbol{X}_{j,1}, \ldots, \boldsymbol{X}_{j,L}]$, where $\boldsymbol{Y}_{j,k} \in \mathbb{C}^{n_c}$ and $\boldsymbol{X}_{j,k} \in \mathcal{X} = \{\mathbf{x} \in \mathbb{C}^{n_c} : \|\mathbf{x}\|^2 = n_c \rho\}$ for $j = 1, 2, \ldots$, and $k = 1, \ldots, L$. The variable ρ denotes the SNR. The Rayleigh fading is modeled by $H_{j,k} \sim \mathcal{CN}(0, 1)$, and $\boldsymbol{W}_{j,k} \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{I}_{n_c})$ is the AWGN noise. The random variables $\{H_{j,k}\}$ and $\{\boldsymbol{W}_{j,k}\}$, which are mutually independent, are also independent over j and k. No *a priori* knowledge of the realizations of $\{H_{j,k}\}$ is available at the transmitter and at the receiver. Consequently, it is reasonable to transmit equal-power signals over the available diversity branches. This justifies our assumption that $\boldsymbol{X}_{j,k} \in \mathcal{X}$ for all j and k.

A code for the channel (1) in the HARQ setup is formally defined next by adapting the notion of VLSF codes in [8].

Definition 1: An $(\ell, M, \epsilon, \rho, \ell_{\max})$-VLSF code, where $\ell \geq 1$, M, $\ell_{\max}$ are positive integers, $\rho > 0$, and $0 < \epsilon < 1$, consists of

1) A random variable U with distribution P_U defined on a set $\mathcal{U}$ with $|\mathcal{U}| \leq 2$ that is revealed to both the transmitter and the receiver before the start of transmission. U acts as a common randomness and enables the use of randomized encoding and decoding strategies.

2) An encoder $f : \mathcal{U} \times \{1, \ldots, M\} \to \mathcal{X}^{L\ell_{\max}}$, that maps a message J, which is uniformly distributed on $\{1, \ldots, M\}$, to a codeword in the set $\{\mathbf{c}(1), \ldots, \mathbf{c}(M)\}$. Each codeword is structured as $\mathbf{c}(m) = [\mathbf{c}_1(m), \ldots, \mathbf{c}_{\ell_{\max}}(m)]$ where $\mathbf{c}_j(m) \in \mathcal{X}^L$ for $j = 1, \ldots, \ell_{\max}$ and $m = 1, \ldots, M$.

3) A sequence of decoders $g_v : \mathcal{U} \times \mathbb{C}^{n_c L v} \to \{1, \ldots, M\}$, $1 \leq v \leq \ell_{\max}$, and a stopping time τ^*, that is adapted to the filtration $\{\sigma(U, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_v)\}_{v=1}^{\ell_{\max}}$, and satisfies both

$$\mathbb{E}[\tau^*] \leq \ell, \quad (3)$$

and the average packet error probability target

$$\mathbb{P}[g_{\tau^*}(U, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{\tau^*}) \neq J] \leq \epsilon. \quad (4)$$

After the stopping time is triggered, the decoder uses the feedback channel to inform the encoder, through a one-bit ACK, to stop the transmission of the current message and to move to the transmission of the next one. We assume throughout the paper that this bit is error free and that so are the NACK bits transmitted in the previous feedback rounds. Differently from most literature, we will consider in our analysis undetected error events in which an ACK is fed back although the decoder has chosen the wrong message.

For a given ℓ, ϵ, and $\ell_{\max}$, the maximum coding rate R^*, measured in information bits per channel use is defined as

$$R^*(\ell, \epsilon, \rho, \ell_{\max}) =$$
$$\sup\left\{ \frac{\log_2 M}{\ell L n_c} : \exists (\ell, M, \epsilon, \rho, \ell_{\max})\text{-VLSF code} \right\}. \quad (5)$$

We will also be interested in the problem of minimizing the average number of transmissions ℓ for a given number of messages M, which yields the following definition:

$$\ell^*(M, \epsilon, \rho, \ell_{\max}) =$$
$$\inf\{\ell : \exists (\ell, M, \epsilon, \rho, \ell_{\max})\text{-VLSF code}\}. \quad (6)$$

Some of our results will also be expressed in terms of the minimum energy per bit[1] E_b^*/N_0, which is related to ℓ^* as

$$\frac{E_b^*}{N_0} = \frac{\rho L n_c}{\log_2 M} \ell^*(M, \epsilon, \rho, \ell_{\max}). \quad (7)$$

The corresponding metrics for FBL-NF are defined as in [6].

B. PAT with SNN Decoding

Following [6], we assume that, for slot j and coherence interval k, the input vector $\boldsymbol{X}_{j,k}$ is of the form $[\mathbf{x}^{(p)}, \boldsymbol{X}_{j,k}^{(d)}]$, where $\mathbf{x}^{(p)} \in \mathbb{C}^{n_p}$, $1 \leq n_p < n_c$, is a deterministic vector containing pilot symbols and $\|\mathbf{x}^{(p)}\|^2 = n_p \rho$. The vector $\boldsymbol{X}_{j,k}^{(d)} \in \mathbb{C}^{n_d}$ contains the $n_d = n_c - n_p$ data symbols. Let $P_{\boldsymbol{X}^{(d)}}$ denote a distribution on $\mathbb{C}^{n_d}$ such that $\boldsymbol{X}_{j,k} \in \mathcal{X}$ w.p.1 when $\boldsymbol{X}_{j,k}^{(d)} \sim P_{\boldsymbol{X}^{(d)}}$. Let $\boldsymbol{Y}_{j,k}^{(p)}$ and $\boldsymbol{Y}_{j,k}^{(d)}$ be the vectors containing the received samples corresponding to the pilot and data symbols, respectively. Given $\boldsymbol{Y}_{j,k}^{(p)}$ and $\boldsymbol{Y}_{j,k}^{(d)}$, the receiver computes the maximum likelihood estimate $\widehat{H}_{j,k}$ of the fading coefficient $H_{j,k}$ as

$$\widehat{H}_{j,k} = \frac{1}{n_p \rho} (\mathbf{x}^{(p)})^{\mathsf{H}} \boldsymbol{Y}_{j,k}^{(p)}. \quad (8)$$

Next we define the SNN decoder. Denote a candidate codeword and the observed channel outputs up to the vth slot by $\mathbf{x}^v = [\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,L}, \ldots, \mathbf{x}_{v,L}]$ and $\mathbf{y}^v = [\mathbf{y}_{1,1}, \ldots, \mathbf{y}_{1,L}, \ldots, \mathbf{y}_{v,L}]$, respectively, where $\mathbf{x}_{j,k} \in \mathcal{X}$ and $\mathbf{y}_{j,k} \in \mathbb{C}^{n_c}$ for $j = 1, \ldots, v$, and $k = 1, \ldots, L$. Here, $\mathbf{x}_{j,k}$ and $\mathbf{y}_{j,k}$ follow the structure outlined above. The SNN metric is given by

$$q^{(v)}(\mathbf{x}^v, \mathbf{y}^v) = \prod_{j=1}^{v} \prod_{k=1}^{L} q(\mathbf{x}_{j,k}, \mathbf{y}_{j,k}) \quad (9)$$

where

$$q(\mathbf{x}_{j,k}, \mathbf{y}_{j,k}) = \prod_{i=1}^{n_d} \exp\left(-|y_{j,k,i}^{(d)} - \widehat{h}_{j,k} x_{j,k,i}^{(d)}|^2 \right). \quad (10)$$

Here, $y_{j,k,i}^{(d)}$ and $x_{j,k,i}^{(d)}$ denote the ith entry of $\mathbf{y}_{j,k}^{(d)}$ and $\mathbf{x}_{j,k}^{(d)}$, respectively. We refer to this coding scheme as PAT-SNN.

Finally, we define the generalized information density in slot v as a mapping $\mathcal{X}^{vL} \times \mathbb{C}^{n_c L v} \to \mathbb{R}$, given as

$$\imath_s^v(\mathbf{x}^v, \mathbf{y}^v) = \sum_{j=1}^{v} \sum_{k=1}^{L} \imath_s(\mathbf{x}_{j,k}, \mathbf{y}_{j,k}) \quad (11)$$

where the generalized information density per coherence block $\imath_s(\mathbf{x}_{j,k}, \mathbf{y}_{j,k})$ is

$$\imath_s(\mathbf{x}_{j,k}, \mathbf{y}_{j,k}) = \log \frac{q(\mathbf{x}_{j,k}, \mathbf{y}_{j,k})^s}{\mathbb{E}_{\overline{\boldsymbol{X}}}[q(\overline{\boldsymbol{X}}, \mathbf{y}_{j,k})^s]}. \quad (12)$$

[1]We will not consider the energy spent to send the feedback bit reliably.

Here, $\overline{X} = [\mathbf{x}^{(p)}, \overline{X}^{(d)}]$ with $\mathbf{x}^{(p)}$ an arbitrary pilot vector satisfying the properties listed above, $\overline{X}^{(d)} \sim P_{X^{(d)}}$, and $s \geq 0$. For the special case in which $P_{X^{(d)}}$ is a product distribution, i.e., $P_{X^{(d)}}(\mathbf{x}_{j,k}^{(d)}) = \prod_{i=1}^{n_d} P_X(x_{j,k,i}^{(d)})$, we can write (12) as

$$\imath_s(\mathbf{x}_{j,k}, \mathbf{y}_{j,k}) = \sum_{i=1}^{n_d} -s|y_{j,k,i}^{(d)} - \widehat{h}_{j,k} x_{j,k,i}^{(d)}|^2$$
$$- \log \mathbb{E}_{\overline{X}}\left[\exp\left(-s|y_{j,k,i}^{(d)} - \widehat{h}_{j,k}\overline{X}|^2\right)\right]. \quad (13)$$

III. Finite-Blocklength Achievability Bounds

A. Fixed-Blocklength Transmission without Feedback

Next, we review an achievability bound for the channel in (2), based on the random-coding union bound with parameter s [12, Thm. 1]. This bound will be used to assess the performance of PAT-SNN-FBL-NF transmission.

Theorem 1: Fix an integer $1 \leq n_d < n_c$, a rate R, and a real number $s \geq 0$. The average error probability for the PAT-SNN-FBL-NF scheme, operating as described in Section II-B, is upper-bounded as

$$\epsilon \leq \mathbb{E}\left[\exp\left(-\left[\imath_s^v(\mathbf{X}^v, \mathbf{Y}^v) - \log(2^{vLn_cR} - 1)\right]^+\right)\right] \quad (14)$$

where $\mathbf{X}^v = [\mathbf{X}_{1,1}, \mathbf{X}_{1,2}, \ldots, \mathbf{X}_{L,1}, \ldots, \mathbf{X}_{v,L}]$ and each $\mathbf{X}_{j,k} = [\mathbf{x}^{(p)}, \mathbf{X}_{j,k}^{(d)}]$ with $\mathbf{x}^{(p)}$ an arbitrary pilot vector and $\mathbf{X}_{j,k}^{(d)} \sim P_{X^{(d)}}$, and $\mathbf{Y}_{j,k}$ is the channel output according to (2), for $j = 1, \ldots, v$ and $k = 1, \ldots, L$.

Proof: See [6, Thm. 3]. ∎

Remark: The latency for the FBL-NF scheme is $T_d = vdT_o$, where T_o is the OFDM symbol duration. In the HARQ case we will analyze next, latency is a random variable. We will investigate both its average value and its CDF.

B. HARQ Transmission

We provide next an achievability bound for the HARQ setup, which closely follows [8, Thm. 3]. To prove [8, Thm. 3], one computes the accumulated information density for each codeword and stops when one of the information densities exceeds a given threshold. In our case, instead of information density, we accumulate generalized information density, see (12). Furthermore, we only allow for a fixed number of transmission rounds, after which transmission is terminated and a decision is taken. This differs from the setup in [9] where, after a fixed number of transmission rounds, the received data is discarded and transmission is restarted. Incorporating these changes to [8, Thm. 3], one obtains the following result.

Theorem 2 (Achievability HARQ): Fix three scalars $\gamma > 0$, $s \geq 0$, $\rho > 0$, and two positive integers $\ell_{\max}$ and $1 \leq n_p < n_c$. Let $\{\mathbf{X}_j\}_{j=1}^\infty$ be a stochastic process such that $\mathbf{X}_j = [\mathbf{X}_{j,1}, \mathbf{X}_{j,2}, \ldots, \mathbf{X}_{j,L}]$, and $\mathbf{X}_{j,k} \in \mathcal{X}$ are identical and independently distributed (i.i.d.) for $j = 1, \ldots$ and $k = 1, \ldots, L$. Let $\mathbf{X}_{j,k} = [\mathbf{x}^{(p)}, \mathbf{X}_{j,k}^{(d)}]$ where $\mathbf{x}^{(p)}$ is an arbitrary pilot vector satisfying $\|\mathbf{x}^{(p)}\|^2 = n_p\rho$ and $\mathbf{X}_{j,k}^{(d)} \sim P_{X^{(d)}}$. Let $\{\overline{X}_j\}_{j=1}^\infty$ be an independent copy of $\{\mathbf{X}_j\}_{j=1}^\infty$. Moreover, define the stopping times

$$\tau = \inf\{v \geq 1 : \imath_s^v(\mathbf{X}^v, \mathbf{Y}^v) \geq \gamma\}, \quad (15)$$
$$\overline{\tau} = \inf\{v \geq 1 : \imath_s^v(\overline{\mathbf{X}}^v, \mathbf{Y}^v) \geq \gamma\}, \quad (16)$$

where $\mathbf{Y}^v$ is the random vector representing the channel outputs corresponding to the channel input $\mathbf{X}^v$ between the first and the vth round, according to (1). Then, there exists an $(\ell, M, \epsilon, \rho, \ell_{\max})$-VLSF code with

$$\ell \leq \mathbb{E}[\min\{\ell_{\max}, \tau\}], \quad (17)$$
$$\epsilon \leq (M-1)\mathbb{P}[\overline{\tau} \leq \min\{\ell_{\max}, \tau\}] + \mathbb{P}[\tau > \ell_{\max}]. \quad (18)$$

Proof: We create a codebook with M independent codewords, each belonging to $\mathcal{X}^{L\ell_{\max}}$. Each codeword is a composition of $L\ell_{\max}$ subcodewords, belonging to $\mathcal{X}$, where each subcodeword contains n_p deterministic pilot symbols and $n_d = n_c - n_p$ data symbols distributed according to $P_{X^{(d)}}$.

For the mth codeword, $1 \leq m \leq M$, we define a stopping time based on the generalized information density in (11) as

$$\tau_m = \min\{\inf\{1 \leq v \leq \ell_{\max} : \imath_s^v(\mathbf{c}^v(m), \mathbf{Y}^v) \geq \gamma\}, \ell_{\max}\} \quad (19)$$

where we use the convention that the infimum of the empty set equals infinity. Let now $\tau^* = \min\{\tau_1, \tau_2, \ldots, \tau_M\}$. Note that this stopping time is different from the one in [8, Thm. 3] since we are dealing here with a finite number of transmission. Let

$$\hat{J}_{\tau^*} = \arg\max_{1 \leq m \leq M}\left\{\imath_s^{\tau^*}(\mathbf{c}^{\tau^*}(m), \mathbf{Y}^{\tau^*})\right\} \quad (20)$$

denote the estimated message at stopping time τ^*. In words, if the threshold is crossed within $\ell_{\max}$ transmission rounds, we choose the message corresponding to the largest overshoot. If the threshold is not crossed within $\ell_{\max}$ transmission rounds, we simply choose the message with the largest accumulated metric. For this decoding scheme, the average stopping time, averaged over the codebook ensemble generated by the distribution $P_{X^{(d)}}$, is upper-bounded as

$$\mathbb{E}[\tau^*] \leq \frac{1}{M}\sum_{j=1}^M \mathbb{E}[\tau_j | J = j] = \mathbb{E}[\min\{\tau, \ell_{\max}\}]. \quad (21)$$

In words, we upper-bound τ^* by the minimum between $\ell_{\max}$ and the number of rounds required for the accumulated generalized information density of the transmitted codeword to exceed the threshold. Next we define two error events

$$\mathcal{E}_1 = \{\tau^* \leq \ell_{\max}, \hat{J}_{\tau^*} \neq J, \max_m\{\imath_s^{\tau^*}(\mathbf{c}^{\tau^*}(m), \mathbf{Y}^{\tau^*})\} \geq \gamma\}, \quad (22)$$
$$\mathcal{E}_2 = \{\tau^* = \ell_{\max}, \hat{J}_{\tau^*} \neq J, \max_m\{\imath_s^{\tau^*}(\mathbf{c}^{\tau^*}(m), \mathbf{Y}^{\tau^*})\} < \gamma\}, \quad (23)$$

such that $\epsilon = \mathbb{P}[\mathcal{E}_1 \cup \mathcal{E}_2]$. Here, $\mathcal{E}_1$ is the event the threshold is crossed within $\ell_{\max}$ transmission rounds and an erroneous message is chosen, and $\mathcal{E}_2$ is the event that, in transmission round $\ell_{\max}$, the threshold has not been crossed and the decoder decides for the wrong message. Hence, ϵ describes the undetected error event. We proceed as follows:

$$\epsilon \leq \mathbb{P}[\mathcal{E}_1] + \mathbb{P}[\mathcal{E}_2] \quad (24)$$
$$\leq \sum_{j=2}^M \mathbb{P}[\tau_j \leq \tau_1, \imath_s^{\tau_j}(\mathbf{c}^{\tau_j}(j), \mathbf{Y}^{\tau_j}) \geq \gamma | J = 1] + \mathbb{P}[\tau > \ell_{\max}]$$
$$= (M-1)\mathbb{P}[\overline{\tau} \leq \min\{\ell_{\max}, \tau\}] + \mathbb{P}[\tau > \ell_{\max}]. \quad (25)$$

Using the arguments in [8, Thm. 19] and [13, p. 35], we conclude that a randomized codebook attaining both ensemble averages (21) and (25) simultaneously can be constructed by using a convex combination of two deterministic codebooks. ∎

Remark: The average latency for the HARQ scheme is $T_{\mathrm{d}} = 2\ell d T_{\mathrm{o}}$, where the factor 2 is due to the feedback delay and T_{o} is the OFDM symbol duration. The maximum latency is, however, $2\ell_{\max} d T_{\mathrm{o}}$. Note also that the CDF of the delay τ is required to compute (25).

No closed-form expressions for (17) and (18) are available. Furthermore, evaluating (18) is numerically challenging. Indeed, M is typically very large (for a fixed rate R, it grows exponentially in the blocklength), so the first probability is very small thus preventing the use of Monte Carlo methods. For the case of summands in (16) with negative drift, i.e., $\mathbb{E}\left[\imath_s(\overline{\boldsymbol{X}}_{j,k}, \boldsymbol{Y}_{j,k})\right] < 0$, we can further relax (18) by upper-bounding $\mathbb{P}[\overline{\tau} \le \min\{\ell_{\max}, \tau\}]$ using Wald's identity as [14, Cor. 9.4.4]

$$\mathbb{P}[\overline{\tau} \le \min\{\ell_{\max}, \tau\}] \le \mathbb{P}[\overline{\tau} \le \infty] \le \exp(-\beta^* \gamma) \quad (26)$$

where β^* is the positive solution of

$$\kappa(\beta) = \log \mathbb{E}\left[\exp\left(\beta \sum_{k=1}^{L} \imath_s(\overline{\boldsymbol{X}}_{1,k}, \boldsymbol{Y}_{1,k})\right)\right] = 0. \quad (27)$$

In the upcoming section, we let $P_{\boldsymbol{X}^{(\mathrm{d})}}$ be a product distribution. Hence, the generalized information density is given by (13). It then follows from Jensen's inequality that $\mathbb{E}\left[\imath_s(\overline{\boldsymbol{X}}_{j,k}, \boldsymbol{Y}_{j,k})\right] < 0$. This enables us to use (26) in our numerical evaluations. Furthermore, we can express $\kappa(\beta)$ as

$$\kappa(\beta) = L \log \mathbb{E}\left[\frac{\mathbb{E}\left[q(\overline{\boldsymbol{X}}, \mathbf{y})^{\beta s} \mid \boldsymbol{Y} = \mathbf{y}\right]}{\mathbb{E}\left[q(\overline{\boldsymbol{X}}, \mathbf{y})^{s} \mid \boldsymbol{Y} = \mathbf{y}\right]^{\beta}} \right] \quad (28)$$

where $\overline{\boldsymbol{X}} = [\mathbf{x}^{(\mathrm{p})}, \overline{\boldsymbol{X}}^{(\mathrm{d})}]$ with $\overline{\boldsymbol{X}}^{(\mathrm{d})} \sim P_{\boldsymbol{X}^{(\mathrm{d})}}$ and $\boldsymbol{Y}$ distributed as in (2), independent of $\overline{\boldsymbol{X}}$. From (28), we obtain $\beta^* = 1$.

Remark: An upper bound similar to (26) with $\beta^* = 1$ is obtained in [8, Eq. (113)] by using a change-of-measure argument. We cannot follow the same strategy here because of the mismatched decoding metric.

IV. NUMERICAL RESULTS

In this section, we compare the performance of HARQ and FBL-NF. For the sake of concreteness, we consider i.i.d. input symbols drawn uniformly from a quaternary phase-shift keying (QPSK) constellation. We extract our channel parameters from the TDL-C 300 ns–3km/h channel model [15]. We choose the system bandwidth, subcarrier bandwidth, and OFDM symbol duration equal to 20 MHz, 15 kHz, and 71.4 μs, respectively. This results in 30 available diversity branches. The channel model has coherence bandwidth $B_{\mathrm{c}} = 0.66$ MHz, which implies $u \le 44$. We let $u = 24$ and $d = 3$. Hence, an RB occupies 360 kHz and its duration is 214.2 μs. We also let $\epsilon = 10^{-3}$, which corresponds to the least stringent reliability constraint for URLLC [16]. The latency for HARQ is $T_{\mathrm{d}} = 2T_{\mathrm{o}} d \ell$, and thus the minimum and maximum latency are $2T_{\mathrm{o}} d$ and $2T_{\mathrm{o}} d \ell_{\max}$, respectively. For FBL-NF, the minimum latency corresponds to one slot duration, i.e., $T_{\mathrm{o}} d$, and the maximum latency is $T_{\mathrm{o}} d \ell_{\max}$. Throughout this section, we assume that each message contains $k = \log_2 M = 30$ information bits. The channel and system parameters are summarized in Table I.

TABLE I
PARAMETERS FOR TDL-C 300 NS–3 KM/H (UPPER HALF) AND SIGNAL (BOTTOM HALF).

Symbol	Parameter	Value
B_{c}	50% coherence bandwidth	0.66 MHz
T_{c}	50% coherence time	85 ms
L_{c}	Number of available diversity branches	30
k	Information bits	30
ϵ	Error probability	10^{-3}
B	System bandwidth	20 MHz
B_{s}	Subcarrier bandwidth	15 KHz
T_{o}	OFDM symbol duration	71.4 μs
u	Subcarriers per RB	24
d	OFDM symbols per RB	3
L	Number of used diversity branches	$\le L_{\mathrm{c}}$
$\ell_{\max}$	Max. number of transmission rounds	$\lfloor L_{\mathrm{c}}/L \rfloor$

Next, we detail how we evaluate the minimum energy per bit (7). The procedure for computing the maximum coding rate (5) is similar. In both FBL-NF and HARQ, we fix ρ, ϵ, k, T_{o}, u, d, and L. For HARQ, we find the smallest threshold γ such that the right-hand side of (18), relaxed using (26), is below the desired error target. This γ is then used to evaluate (17), which provides an upper bound on the average latency T_{d}. For FBL-NF, we search for the smallest v such that the right-hand side of (14) is below the desired error target. Throughout this section, all charts are obtained by optimizing over the parameter s and the number of pilot symbols n_{p}.

In Fig. 3, we show the minimum energy per bit required to fulfill the reliability target as a function of average latency. It can be seen that, for fixed L, HARQ significantly outperforms FBL-NF. For example, for $L = 3$ and $T_{\mathrm{d}} = 1$ ms, the difference is about 4 dB. It can also be seen that E_{b}^*/N_0 does not decrease monotonically with the average latency for both HARQ and FBL-NF, but there exists a latency T_{d}^* for which E_{b}^*/N_0 is minimized. When $T_{\mathrm{d}} < T_{\mathrm{d}}^*$, increasing the average blocklength provides an SNR gain that outweighs the rate penalty (see (7)), whereas the opposite is true when $T_{\mathrm{d}} > T_{\mathrm{d}}^*$. HARQ performs better than FBL-NF because for a fixed ϵ, changes in the average blocklength result in larger changes in the required SNR than in FBL-NF. We also observe a tradeoff between the average latency and the number of diversity branches used: choosing a large L is more energy efficient for small average latency ($T_{\mathrm{d}} < 0.7$ ms in Fig. 3), since in this regime the number of possible retransmissions is small, and thus higher diversity is beneficial. As the average latency increases, choosing a smaller number of diversity branches is preferable.

In Fig. 4, we compare the maximum coding rate achievable with the two schemes as a function of the average latency. It can be seen that larger rates are achievable by using HARQ rather than FBL-NF. A similar behavior was reported for the AWGN channel in [9]. The rate increases with L in both systems, since a larger L allows for more diversity and longer codewords.

The results presented so far are in terms of average latency. However, average and maximum latencies are not equal in HARQ. Hence, one may argue that the above comparisons are unfair since the maximum latency with HARQ may be larger than the latency with FBL-NF. In Fig. 5, we illustrate the CDF of the latency for both FBL-NF and HARQ, for the case $L = 2$, and $\rho \in \{-5, -2\}$ dB. The CDF of the FBL-NF latency is a
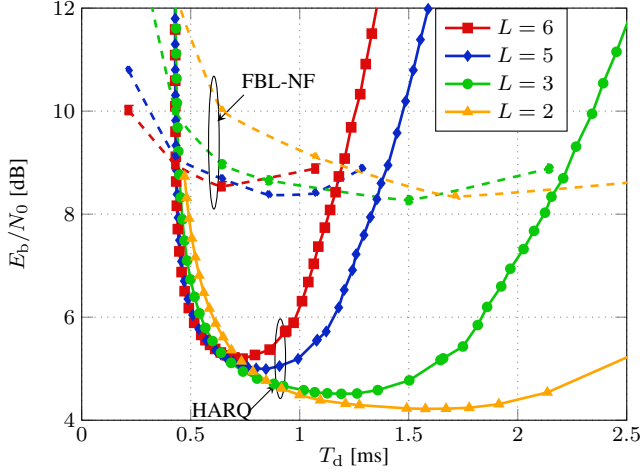
Fig. 3. E_b^*/N_0 versus average latency for $L \in \{2, 3, 5, 6\}$ and 30 information bits. The remaining parameter values are given in Table I.
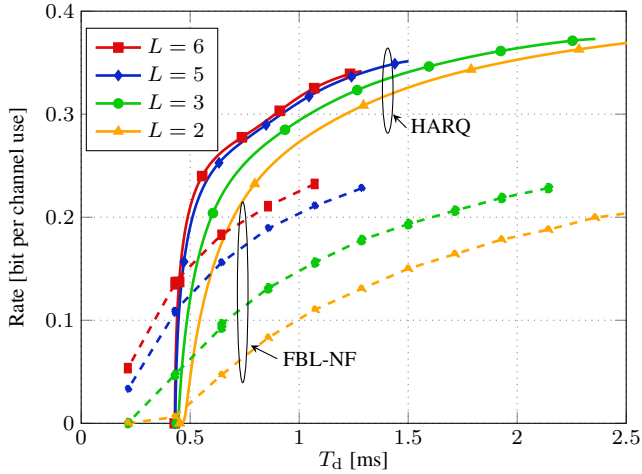

Fig. 5. CDF of the latency. The curves are generated for 30 information bits, $\rho = \{-5, -2\}$ dB and $L = 2$. The remaining parameter values are given in Table I.

Fig. 4. R^* versus average latency for $\rho = -2$ dB and $L \in \{2, 3, 5, 6\}$. The remaining parameter values are given in Table I.

step function since the latency is deterministic whereas the CDF of HARQ is a staircase function with steps at multiples of the duration of a transmission round. We see from Fig. 5 that the probability of the latency in HARQ being larger than in FBL-NF decreases with SNR. For $\rho = -5$ dB, this probability is about 0.097; for $\rho = -2$ dB, the probability is about 0.054. The reason is that, as the SNR increases, it is more likely that one round is enough to satisfy the reliability constraint.

An important application where the CDF of the delay plays a pivotal role is in joint coding-queuing analyses of URLLC systems. Indeed, when taking into account also queuing delays, an optimum latency-aware design must be based on the CDF of the latency rather than on its average [11]. As a consequence, the results in this paper are relevant to the joint coding-queuing analysis of URLLC systems over memoryless block-fading channels with limited diversity. Specifically, the tools introduced in this paper can be used to extend the analysis in [11], where the physical layer was modeled as a simple binary-input AWGN channel, to block-fading scenarios and practically-relevant pilot-based transmission schemes.
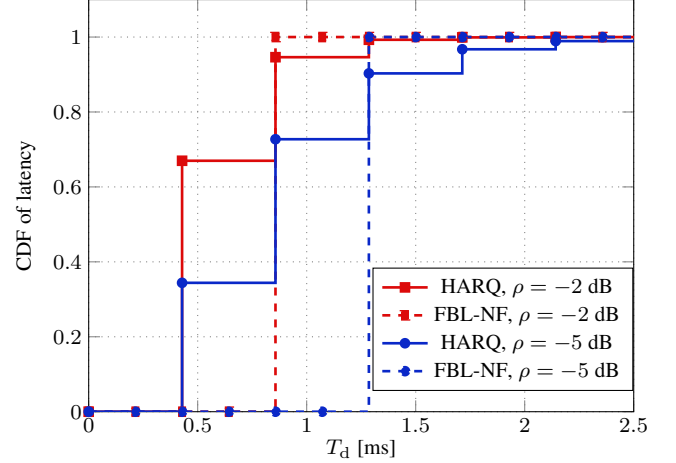
REFERENCES

[1] ITU-R, "Report ITU-R M.2412-0: Guidelines for evaluation of radio interface technologies for IMT-2020," International Telecommunication Union, Tech. Rep., Oct. 2017.

[2] G. Durisi, T. Koch, and P. Popovski, "Towards massive, ultra-reliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

[3] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[4] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4265, Jul. 2014.

[5] G. Durisi, T. Koch, J. Östman, Y. Polyanskiy, and W. Yang, "Short-packet communications over multiple-antenna Rayleigh-fading channels," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 618–629, Feb. 2016.

[6] J. Östman, G. Durisi, E. G. Ström, M. C. Coskun, and G. Liva, "Short packets over block-memoryless fading channels: Pilot-assisted or noncoherent transmission?" [Online]. Available: http://arxiv.org/pdf/1712.06387.pdf

[7] G. C. Ferrante, J. Östman, G. Durisi, and K. Kittichokechai, "Pilot-assisted short-packet transmission over multiantenna fading channels: A 5G case study," in *Conf. Inf. Sci. Sys. (CISS)*, New Jersey, U.S., Mar. 2018.

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

[9] A. R. Williamson, T.-Y. Chen, and R. D. Wesel, "Variable-length convolutional coding for short blocklengths with decision feedback," *IEEE Trans. Commun.*, vol. 63, no. 7, pp. 2389–2403, Jul. 2015.

[10] A. Avranas, M. Kountouris, and P. Ciblat, "Energy-latency tradeoff in ultra-reliable low-latency communication with retransmissions." [Online]. Available: https://arxiv.org/pdf/1805.01332.pdf

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

[12] A. Martinez and A. Guillén i Fàbregas, "Saddlepoint approximation of random–coding bounds," in *Proc. Inf. Theory Applicat. Workshop (ITA)*, San Diego, CA, U.S.A., Feb. 2011.

[13] H. G. Eggleston, *Convexity*. New York: Cambridge university press, 1958.

[14] R. G. Gallager, *Stochastic Processes: Theory for Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2013.

[15] 3GPP, "TR 38.901: Study on channel model for frequencies from 0.5 to 100 GHz," 3GPP, Tech. Rep., 2017.

[16] P. Schulz *et al.*, "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 70–78, Feb. 2017.