Short-packet Transmission via Variable-Length Codes in the Presence of Noisy Stop Feedback

(article starts on next page)

Short-packet Transmission via Variable-Length Codes in the Presence of Noisy Stop Feedback

Johan Östman, *Student Member, IEEE*, Rahul Devassy, Giuseppe Durisi, *Senior Member, IEEE*, Erik G. Ström, *Senior Member, IEEE*

Abstract—We present an upper bound on the error probability achievable using variable-length stop feedback (VLSF) codes, for a fixed size of the information payload and a given constraint on the maximum latency and the average service time. Differently from the bound proposed in Polyanskiy *et al.* (2011), which pertains to the scenario in which the stop signal is sent over a noiseless feedback channel, our bound applies to the practically relevant setup in which the feedback link is noisy. Numerical evaluation of our bound suggests that, for fixed latency and reliability constraints, noise in the feedback link may increase the minimum average service time for the VLSF scheme considered in this paper, to the extent that fixed-length codes without feedback may be preferable in some scenarios.

I. INTRODUCTION

Variable-length stop-feedback (VLSF) coding schemes, i.e., schemes such as simple automatic repeat request (ARQ) and hybrid automatic repeat request (HARQ), in which information is transmitted until the reception of a positive acknowledgment (ACK), are ubiquitous in modern wireless communication systems. This is because they offer a simple yet effective way to adapt the transmission rate to the channel conditions and, hence, reduce the error probability. The question investigated in this paper is whether such schemes are suitable for ultra-reliable low-latency communications (URLLC)—one of the new use cases in next-generation wireless systems (5G).

From a physical layer perspective, the URLLC design problem involves answering the following question: can a given information payload be transmitted within a target latency requirement at a desired reliability level? Unfortunately, classical approaches to answering this question, which rely on large-blocklength results in information theory, are unsuitable whenever the latency requirement is stringent, such as in URLLC.

If the physical layer employs fixed-length coding schemes without feedback, the URLLC design problem can be tackled using the nonasymptotic information-theoretic bounds developed in [2] (see, e.g., [3]–[6]). Such bounds allow one to assess for example how much frequency and spatial diversity should be exploited to achieve a target reliability for a given latency requirement.

Johan Östman, Giuseppe Durisi, and Erik G. Ström are with the Department of Electrical Engineering, Chalmers University of Technology, Gothenburg 41296, Sweden (e-mail: {johanos,durisi,erik.strom}@chalmers.se).

Rahul Devassy was with the Department of Electrical Engineering, Chalmers University of Technology, Gothenburg 41296, Sweden (e-mail: devassyhere@gmail.com).

Less is known in the VLSF case. The nonasymptotic achievability bound provided in [7, Thm. 3], shows that, for a fixed reliability target, the use of variable-length codes combined with stop feedback, allows one to approach capacity much faster in the (average) blocklength, compared to the scenario in which fixed-blocklength codes with no feedback are used. However, the achievability bound given in [7, Thm. 3] pertains to the setup in which the acknowledgment sent on the feedback channel is assumed instantaneous and error-free. As argued in, e.g., [8], [9], these two assumptions are not suitable for URLLC. The purpose of this paper is to generalize the analysis in [7, Thm. 3] and determine if VLSF codes remain superior to fixed-blocklength no-feedback codes once the feedback delay and the presence of noise in the feedback link, which causes unreliable acknowledgments, are accounted for.

Contributions: Assuming arbitrary noisy forward and feedback channels, we obtain an upper bound on the error probability achievable using VLSF coding schemes, for a fixed size of the information payload, and a given constraint on the maximum latency and on the average service time. This last quantity is defined as the average time it takes the transmitter to process an information packet. Our bound pertains to the setup in which there exists a constraint on the maximum number of transmission rounds (which is imposed by the latency requirement), after which an error is declared at the receiver. Also, our analysis accounts for the presence of unreliable acknowledgments and of undetected error events, which occur whenever the decoder terminates transmission with an ACK, but its decision is erroneous.

The impact of unreliable acknowledgments can be mitigated through coding on the feedback channel. However, this comes at a cost in terms of a feedback delay that is captured by our analysis. Undetected errors are typically neglected in the analysis of HARQ protocols. We argue that this simplifying assumption is unsuitable for the analysis of URLLC systems. In practical systems, a cyclic redundancy check (CRC) is typically used to detect errors at the receiver [10, Ch. 6.4]. Obviously, the longer the CRC, the lower the undetected error probability. However, for a given latency requirement, increasing the length of the CRC results in a reduction of the rate of the inner channel code. Hence, there is a fundamental trade-off that needs to be characterized for the optimal design of URLLC systems. Our analysis, which relies on a threshold-based decoding rule that allows one to trade between reduction of service time and reduction of undetected error probability, sheds lights on this trade-off.

Focusing on the the URLLC regime where both reliability and latency requirements are stringent and the information payload

is typically small, we use our bounds to analyze the performance of VLSF coding schemes operating over practically relevant wireless channels. Specifically, we consider transmissions over i) a binary-input additive white Gaussian noise (bi-AWGN) channel and ii) a block-memoryless Rayleigh fading channel, for the practically relevant setup in which pilot symbols are used to estimate the channel coefficients, and the receiver is equipped with a mismatched decoder that treats the channel estimate as perfect. In both cases, our bound suggests that the presence of noise on the feedback link may cause a fundamental degradation in the performance of the VLSF codes considered here. For example, for the case of the bi-AWGN channel, when the size of the information payload is 30 bits, the maximum latency constraint is 400 channel uses, the packet error probability target is 10^{-5}, and both the forward and the feedback channel operate at an SNR of 0 dB, our bounds result in an average-service-time estimate of 106.6 channel uses if the feedback link is assumed noiseless. This value increases to 141 channel uses if noise on the feedback link is accounted for. For such a scenario, our bound suggests that the performance of the VLSF coding scheme considered in this paper is inferior to that of a fixed-blocklength coding scheme without feedback, which has a (deterministic) service time of 130 channel uses. On the contrary, for the case of Rayleigh block fading, the performance of the VLSF coding scheme considered in this paper is superior to that of a fixed-length coding scheme without feedback, even when noise in the feedback link of the VLSF coding scheme is accounted for. This is because the VLSF scheme considered in this paper exploits the diversity offered by the channel more effectively.

Prior Art: To put our contribution into perspective, we survey next prior art on the analysis of the performance of point-to-point communication schemes with feedback. Our review will focus on nonasymptotic results; hence, the vast literature that uses classical asymptotic information-theoretic metrics such as mutual information to characterize the performance of such systems will not be covered, since the results obtained following this approach are often not relevant for the design of URLLC systems.

One way to provide nonasymptotic performance analyses of both fixed-length and variable-length coding schemes is through the characterization of the reliability function, which determines the speed at which the error probability vanishes as a function of the blocklength for a fixed communication rate. When no feedback is available, the reliability function in the fixed-blocklength case is known to be no larger than the so-called sphere-packing bound [11]. Furthermore, the sphere-packing bound is achievable for all rates between the critical rate and capacity [12]. For the case of symmetric discrete memoryless channels (DMCs), it is known that the reliability function does not exceed the sphere-packing bound even when noiseless full feedback[1] is available [13]. However, as recently shown in [14] for the case of binary symmetric channels (BSCs), when the rate is below the critical rate, the availability of full feedback, even when noisy, allows one to operate above the best known lower bound on the reliability function for the no-feedback case.

---

[1]By full feedback we mean that the transmitter has perfect causal knowledge of the channel outputs at the receiver.

The use of variable-length codes together with the availability of noiseless full feedback results in a much improved reliability function compared to the fixed-length full-feedback case [15]. Such a significant improvement can be also observed in the moderate-deviation regime, in which the rate tends to capacity and the error probability tends to zero as the blocklength tends to infinity [16].

The reliability function for the variable-length full-feedback case remains above the sphere-packing bound even when the full feedback is noisy [17]. As noted in [17], noise in the feedback link may cause synchronization errors that need to be accounted for in the analysis. This can be done, for example, by using the framework for tracking stopping times through noisy observations put forward in [18].

Variable-length codes combined with stop feedback rather than full feedback were analyzed in [19], [20] where it is shown that in the noiseless case, the reliability function exceeds the sphere-packing bound.

For the fixed-blocklength no-feedback case, the recent work by Polyanskiy *et al.* [2] has renewed interest in determining nonasymptotic bounds on the minimum error probability that are tighter than the ones obtainable through a reliability-function analysis. The bounds provided in [2] allow one to obtain tight performance characterizations also in the so-called normal regime, where the transmission rate is close to capacity. Furthermore, in this regime the bounds, which typically do not admit a closed-form expression, can be efficiently approximated using a compact expression commonly referred to as normal approximation.

As shown in, e.g., [2], [21], the normal approximation can be used to analyze the non-asymptotic performance of simple ARQ schemes. In particular, the authors of [21] studied a good-put maximization problem for the case in which simple ARQ is used over a Rayleigh block-fading channel and the stop feedback is noisy. By leveraging the normal approximation, and by employing a simple model for imperfect error detection, they determine the blocklength required on both the forward and the feedback channel in order to maximize the good-put.

The use of the normal approximation for the analysis of general HARQ schemes (see, e.g., [22], [23]) is not entirely satisfactory from a theoretical viewpoint. Indeed, such an approach does not guarantee the existence of a single mother code that, when shortened to an arbitrary blocklength, achieves the error probability predicted by the normal approximation.

A rigorous analysis of the error probability achievable using general VLSF coding schemes was undertaken by Polyanskiy *et al.* in [7]. Specifically, they provided in [7, Thm. 3] a nonasymptotic upper bound on the minimum error probability achievable with VLSF coding schemes, which reveals that the maximum coding rate achievable with VLSF coding schemes converges faster to capacity as the blocklength increases, compared to the fixed-blocklength, no-feedback case. The nonasymptotic upper bound in [7, Thm. 3] relies on a VLSF coding scheme in which the decoder computes the accumulated information density corresponding to each possible codeword and sends a stop signal whenever one of the accumulated information densities exceeds a threshold. In this scheme, the number of transmission rounds is unlimited, and an ACK/NACK bit is fed back over a noiseless channel after the reception of each symbol.
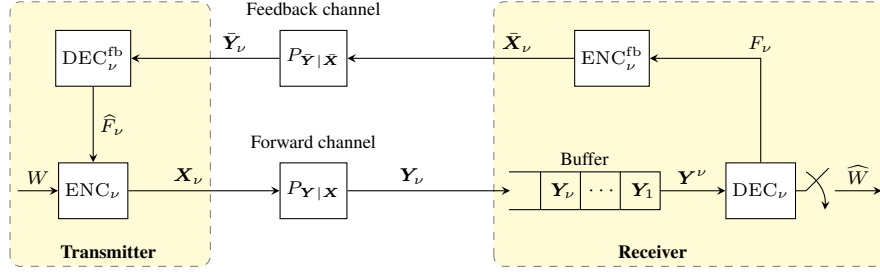
Fig. 1: Round ν of the VLSF transmission scheme. Here, W denotes the information message, $\widehat{W}$ is its estimate at the receiver, F_ν is the feedback bit computed at the receiver, and $\widehat{F}_\nu$ is its estimate at the transmitter.

An extension of this upper bound to the case in which the number of transmission rounds is finite and the feedback bit is transmitted only after a block of symbols can be found in [24], [25]. Furthermore, feedback delay is accounted for in [26]. Finally, adaptations of [7, Thm. 3] to the case of random packet arrivals and to the case of common-message transmission over a broadcast channel are provided in [27] and [28], respectively.

A different approach to bounding the error probability of VLSF coding schemes is presented in [29]. There, a random coding bound is obtained for the setup in which a low-rate inner code, used to provide incremental redundancy, is combined with a high-rate CRC. The analysis provided in [29], which relies on an error-exponent bound, pertains to the transmission of binary antipodal coded symbols over a block-fading channel and accounts for the presence of noise in the feedback link. The authors, however, assume for simplicity that a NACK cannot be interpreted as ACK by the transmitter—a simplification we dispose with in our analysis. Finally, they present a comparison between the bounds and the performance of actual VLSF coding schemes relying on convolutional codes.

Notation: Upper case letters are used to denote random vectors, e.g., $\boldsymbol{X}$ and their realizations are written in lower case, e.g., $\mathbf{x}$. The probability distribution of $\boldsymbol{X}$ is written as $P_{\boldsymbol{X}}$. We use superscripts to denote the concatenation of vectors of equal size, e.g., $\boldsymbol{X}^\nu = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_\nu]$. The distribution of a real Gaussian random variable is denoted by $\mathcal{N}(\mu, \sigma^2)$ and the distribution of a complex proper Gaussian random variable is denoted by $\mathcal{CN}(\mu, \sigma^2)$. Here, μ and σ^2 are the mean and the variance of the random variable, respectively. The Radon-Nikodym derivative of a distribution P_1 with respect to a distribution P_2, where P_1 is absolutely continuous with respect to P_2, is denoted by $\frac{dP_1}{dP_2}$. Finally, $\mathbb{E}[\cdot]$ is the expectation operator, $\mathbb{P}[\cdot]$ is used for probabilities, $\mathbb{1}\{\cdot\}$ denotes the indicator function, $Q(\cdot)$ stands for the Gaussian Q-function, and $I(P_X, P_{Y|X})$ denotes the mutual information between the random variables X and Y, whose joint distribution is $P_X P_{Y|X}$.

II. SYSTEM MODEL

We consider a point-to-point communication system in which information is transmitted using the VLSF coding scheme depicted in Fig. 1. Specifically, transmission occurs over a variable number of rounds. Each round is divided into a data phase and a feedback phase, not necessarily of equal duration. Throughout, we assume that the number of transmission rounds does not exceed the integer $\ell_{\mathrm{m}} < \infty$.

In the data phase, a segment (spanning n channel uses) of the codeword associated to the current information message is sent to the receiver over the forward channel. This channel is modeled as a sequence of conditional probability kernels $\{P_{\boldsymbol{Y}_\nu|\boldsymbol{Y}^{\nu-1},\boldsymbol{X}^\nu}\}_{\nu=1}^{\ell_{\mathrm{m}}}$, where the random vectors $\boldsymbol{Y}_\nu$ and $\boldsymbol{X}_\nu$, $\nu = 1, \ldots, \ell_{\mathrm{m}}$, take values from the sets $\mathcal{Y}^n$ and $\mathcal{X}^n$, respectively.

For analytical tractability, we assume that the channel is block-wise stationary and memoryless, i.e.,

$$P_{\boldsymbol{Y}_\nu|\boldsymbol{Y}^{\nu-1},\boldsymbol{X}^\nu}(\mathbf{y}_\nu|\mathbf{y}^{\nu-1},\mathbf{x}^\nu) = P_{\boldsymbol{Y}|\boldsymbol{X}}(\mathbf{y}_\nu|\mathbf{x}_\nu). \quad (1)$$

At the end of each data phase, the receiver decides whether to perform decoding based on the channel outputs received that far, or to request an additional transmission. The outcome of this decision—a single bit of information conveying the message "stop", which we denote by s or "continue", which we denote by c, is transmitted in the feedback phase over the feedback channel using n_{f} channel uses.[2] We model the feedback channel as a sequence of conditional probability kernels $\{P_{\bar{\boldsymbol{Y}}_\nu|\bar{\boldsymbol{Y}}^{\nu-1},\bar{\boldsymbol{X}}^\nu}\}_{\nu=1}^{\ell_{\mathrm{m}}}$, where the random vectors $\bar{\boldsymbol{Y}}_\nu$ and $\bar{\boldsymbol{X}}_\nu$, $\nu = 1, \ldots, \ell_{\mathrm{m}}$, take values from the sets $\bar{\mathcal{Y}}^{n_{\mathrm{f}}}$ and $\bar{\mathcal{X}}^{n_{\mathrm{f}}}$, respectively. As for the forward channel, we assume for simplicity that the feedback channel is block-wise stationary and memoryless, i.e.,

$$P_{\bar{\boldsymbol{Y}}_\nu|\bar{\boldsymbol{Y}}^{\nu-1},\bar{\boldsymbol{X}}^\nu}(\bar{\mathbf{y}}_\nu|\bar{\mathbf{y}}^{\nu-1},\bar{\mathbf{x}}^\nu) = P_{\bar{\boldsymbol{Y}}|\bar{\boldsymbol{X}}}(\bar{\mathbf{y}}_\nu|\bar{\mathbf{x}}_\nu). \quad (2)$$

This enables a single encoder/decoder pair to be used on the feedback channel over consecutive transmission rounds.

Upon observing the output of the feedback channel, the transmitter decides whether s or c was sent. This implies that the feedback channel, together with the signaling scheme just described, can be viewed as a binary asymmetric channel, with crossover probabilities $p_{\mathrm{sc}} = \mathbb{P}[s \to c]$ and $p_{\mathrm{cs}} = \mathbb{P}[c \to s]$, which depend both on n_{f} and on the encoder-decoder pair used to transmit the binary message over the feedback channel.

Some remarks on our setup are in order. We allow for $p_{\mathrm{sc}} \neq p_{\mathrm{cs}}$ since the s $\to$ c and the c $\to$ s events have a different impact on performance. Indeed, the c $\to$ s event causes the premature interruption of the transmission of the current message. We assume, somewhat pessimistically, that this always results in an error at the decoder. This error needs to be handled by higher layers, often causing a violation of the latency requirement. On the contrary, the s $\to$ c event triggers an unnecessary additional

---

[2]The symbol s corresponds to an ACK, whereas the symbol c corresponds to a NACK.
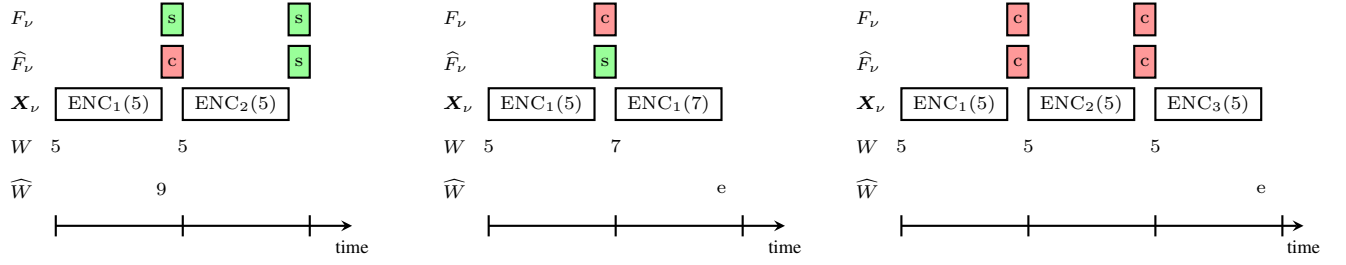
(a) The message $W = 5$ is incorrectly decoded as $\widehat{W} = 9$ after the first round; furthermore, an s $\to$ c event causes the retransmission of message $W = 5$. Note: an s $\to$ c event causes the retransmission of W, even when there is no decoding error.

(b) A c $\to$ s event causes the transmitter to remove message $W = 5$ from its buffer, and move to the new message $W = 7$ after the first round. The receiver observes a packet out of sequence and declares an erasure.

(c) The receiver declares an erasure after $\ell_\mathrm{m} = 3$ unsuccessful transmission rounds.

Fig. 2: Example of the three types of errors for $\ell_\mathrm{m} = 3$. Here, W is the transmitted message and $\widehat{W}$ is the estimate of W at the receiver. Furthermore, F_ν is the feedback bit generated by the receiver in round ν, and $\widehat{F}_\nu$ is its estimate at the transmitter.

transmission round, which causes only a moderate increase in service time. Assuming that these two error events on the feedback channel have different probability is in agreement with current wireless standards, where one typically imposes that $p_\mathrm{cs} \ll p_\mathrm{sc}$. For example, in long term evolution (LTE), we typically have $p_\mathrm{sc} = 10^{-2}$ and $p_\mathrm{cs} \in [10^{-4}, 10^{-3}]$ [10, Ch. 10.4.2].

Note that an error on the feedback channel may result in the transmitter and the receiver falling out of synchronization, i.e., operating on different messages. To prevent this, we assume that each codeword segment contains a binary flag specifying whether the segment is the first one of a new information-message transmission or not. Through coding, one can ensure that this flag is transmitted with a sufficiently high reliability, to avoid synchronization issues. This will, however, consume channel uses and, for a fixed $n + n_\mathrm{f}$, result in a weaker channel code for the forward payload transmission and the feedback transmission. Throughout the paper, we assume for simplicity that this flag is always received correctly at the decoder. From a modeling perspective, this is equivalent to assuming that the noisy estimate of the feedback bit produced at the transmitter is known to the receiver. In Section V, we discuss how to generalize our analysis to account for errors in the transmission of this flag.

To summarize, in our setup, a transmission error occurs if

- The receiver decides to perform decoding but produces the wrong codeword estimate—an event typically referred to as *undetected error*. This event is shown in Fig. 2a along with an s $\to$ c event, which does not cause an error, but increases the service time.
- A c $\to$ s event occurs on the feedback channel, see Fig. 2b.
- The receiver is not able to perform decoding within the available ℓ_m rounds, see Fig. 2c.

In the last two cases, the decoder declares an erasure, which we denote by the symbol e.

A. Definition of a VLSF Code

Before providing a formal definition of a VLSF coding scheme for the noisy feedback case, we introduce some additional notation. We let $F_\nu \in \{\mathrm{s}, \mathrm{c}\}$ be the feedback bit generated by

the receiver in round $\nu = 1, \ldots, \ell_\mathrm{m}$ and $\widehat{F}_\nu \in \{\mathrm{s}, \mathrm{c}\}$ its estimate at the transmitter.

Note that in the presence of errors on the feedback link, the number of rounds after which the receiver produces an estimate of the transmitted message (or declares an erasure) does not necessarily coincide with the number of transmission rounds (see Fig. 2a).

As a consequence, the average service time at the transmitter, which is the average number of transmission rounds after which the current message is removed from the buffer at the transmitter, does not generally coincide with the average latency at the receiver, which is the average number of transmission rounds needed by the receiver to produce a message estimate or to declare an erasure.

From a system-level perspective, the average service time at the transmitter is relevant in full-buffer scenarios, where one is interested in maximizing the long-term throughput. Indeed, according to the renewal-reward theorem, this quantity is given by the ratio between the number of information bits per message and the average service time at the transmitter. Minimizing the average service time at the transmitter is also of interest whenever an objective is to minimize the average energy consumption. Hence, achieving a small service time is of interest also in sporadic transmissions.

Throughout the paper, we shall focus mainly on the case in which the average service time at the transmitter is the metric of interest. However, we will also discuss how to adapt our analysis to the case in which the metric of interest is the average latency at the receiver.

The definition of a VLSF coding scheme provided below is an adaptation to the noisy feedback case of the definition of a VLSF coding scheme given in [7].

Definition 1: An $(\ell_\mathrm{a}, M, \epsilon, \ell_\mathrm{m}, n, n_\mathrm{f})$-VLSF coding scheme where $M, \ell_\mathrm{m}, n,$ and n_f are positive integers, ℓ_a is a nonnegative real number, and $\epsilon \in (0, 1)$, consists of:

- A random variable U defined on a set $\mathcal{U}$ of cardinality $|\mathcal{U}| \leq 2$ that is revealed to both the transmitter and the receiver before the start of the transmission. This random variable acts as common randomness and allows for the use of randomized coding strategies.

- A sequence of ℓ_{m} encoders for the forward channel $f_\nu : \mathcal{U} \times \{1, \dots, M\} \to \mathcal{X}^n$, $\nu = 1, \dots, \ell_{\mathrm{m}}$, defining the forward-channel input

$$\boldsymbol{X}_\nu = f_\nu(U, W) \qquad (3)$$

for a given message W, which we assume to be uniformly distributed over $\{1, \dots, M\}$.
- A sequence of ℓ_{m} decoders for the forward channel $g_\nu : \mathcal{U} \times \mathcal{Y}^{n\nu} \to \{1, \dots, M\}$, $\nu = 1, \dots, \ell_{\mathrm{m}}$, providing an estimate $g_\nu(U, \boldsymbol{Y}^\nu)$ of the message W.
- A sequence of binary random variables $F_\nu \in \{\mathrm{s}, \mathrm{c}\}$, $\nu = 1, \dots, \ell_{\mathrm{m}}$, each being the outcome of the evaluation of a stopping rule defined on the filtration $\sigma(U, \boldsymbol{Y}_1, \dots \boldsymbol{Y}_\nu)$. These random variables are the binary messages transmitted by the receiver on the feedback channel.
- An encoder for the feedback channel $\bar{f} : \{\mathrm{s}, \mathrm{c}\} \to \bar{\mathcal{X}}^{n_{\mathrm{f}}}$ defining the feedback-channel input $\bar{\boldsymbol{X}}_\nu = \bar{f}(F_\nu)$ at round $\nu = 1, \dots, \ell_{\mathrm{m}}$.
- A decoder for the feedback channel $\bar{g} : \bar{\mathcal{Y}}^{n_{\mathrm{f}}} \to \{\mathrm{s}, \mathrm{c}\}$ that produces the estimate $\widehat{F}_\nu = \bar{g}(\bar{\boldsymbol{Y}}_\nu)$ at round ν.
- Two stopping times, one at the transmitter τ_{tx} and one at the receiver τ_{rx}, and a message estimate $\widehat{W} \in \{1, \dots, M\} \cup \mathrm{e}$, all defined through the procedure detailed in Algorithm 1. The stopping time τ_{tx} satisfies the average service-time constraint

$$\mathbb{E}[\tau_{\mathrm{tx}}] \le \ell_{\mathrm{a}} \qquad (4)$$

and the message estimate $\widehat{W}$ satisfies the error probability constraint

$$\mathbb{P}\left[\widehat{W} \ne W\right] \le \epsilon. \qquad (5)$$

Some remarks are in order. Compared to the definition of VLSF codes provided in [7], which involves a single stopping time at the receiver, our definition involves two stopping times, one at the transmitter and one at the receiver. This is needed to account for errors on the feedback link. Also, the decoder employs an erasure option, which is used if a $\mathrm{c} \to \mathrm{s}$ event occurs, or if the stopping rule is not triggered after ℓ_{m} rounds. Note that we measure the service time in transmission rounds. Each transmission round involves n channel uses on the forward channel and n_{f} channel uses on the feedback channel.

Our definition can be readily adapted to the case in which the average latency at the receiver is the metric of interest. Indeed, it is sufficient to replace $\mathbb{E}[\tau_{\mathrm{tx}}]$ in (4) with $\mathbb{E}[\tau_{\mathrm{rx}}]$.

The random variable U, which also appears in the definition of VLSF codes provided in [7], enables the use of randomized coding strategies, which, as we shall see in the proof of our main result, are needed to obtain bounds on (4) and (5) through a random coding argument.

III. MAIN RESULT

We provide an achievability bound, i.e., an upper bound on the error probability achievable using VLSF coding schemes defined according to Definition 1, for a fixed number of messages M, a fixed average service time ℓ_{a}, and a fixed latency requirement ℓ_{m}.

Before presenting our bound, we characterize the pairs $(p_{\mathrm{sc}} = \mathbb{P}[\mathrm{s} \to \mathrm{c}], p_{\mathrm{cs}} = \mathbb{P}[\mathrm{c} \to \mathrm{s}])$ that are achievable for a given choice of the encoder for the feedback channel.

Algorithm 1 Procedure at the transmitter and the receiver to compute the message estimate $\widehat{W}$, the transmitter stopping time τ_{tx}, and the receiver stopping time τ_{rx}.

Initialize:
    $\tau_{\mathrm{tx}} = \tau_{\mathrm{rx}} = \infty$; $F_0 = \widehat{F}_0 = \mathrm{c}$;
for $\nu = 1 \to \ell_{\mathrm{m}}$ **do**
    Transmitter:
    if $\nu > 1$ **then**
        compute $\widehat{F}_{\nu-1} = \bar{g}(\bar{\boldsymbol{Y}}_{\nu-1})$
    end if
    if $\widehat{F}_{\nu-1} = \mathrm{c}$ **then**
        transmit $f_\nu(U, W)$ over the forward channel
        if $\nu = \ell_{\mathrm{m}}$ **then**
            set $\tau_{\mathrm{tx}} = \ell_{\mathrm{m}}$
        end if
    else
        set $\tau_{\mathrm{tx}} = \nu - 1$
    end if
    Receiver:
    switch $(F_{\nu-1}, \widehat{F}_{\nu-1})$ **do**
        case (s, s)
            STOP
        case (s, c)
            set $F_\nu = \mathrm{s}$
        case (c, s)
            set $\tau_{\mathrm{rx}} = \nu$, $\widehat{W} = \mathrm{e}$
            STOP
        case (c, c)
            use stopping rule to compute F_ν
            if $F_\nu = \mathrm{s}$ **then**
                set $\widehat{W} = g_\nu(\boldsymbol{Y}^\nu, U)$ and $\tau_{\mathrm{rx}} = \nu$
            end if
    end switch
    if $\nu < \ell_{\mathrm{m}}$ **then**
        send $\bar{f}(F_\nu)$ on the feedback channel
    else
        if $\tau_{\mathrm{rx}} = \infty$ **then**
            set $\widehat{W} = \mathrm{e}$ and $\tau_{\mathrm{rx}} = \ell_{\mathrm{m}}$
        end if
    end if
end for

Lemma 1: For a given n_{f} and for a given encoder $\bar{f} : \{\mathrm{s}, \mathrm{c}\} \to \bar{\mathcal{X}}^{n_{\mathrm{f}}}$ for the feedback channel, all pairs $(p_{\mathrm{sc}}, p_{\mathrm{cs}})$ in the convex hull of the union on the following two sets are achievable

$$\bigcup_{\gamma_{\mathrm{f}} \in \bar{\mathbb{R}}} \left(\mathbb{P}\left[\frac{\mathrm{d}P^{(\mathrm{c})}}{\mathrm{d}P^{(\mathrm{s})}}(\bar{\boldsymbol{Y}}^{(\mathrm{s})}) > \gamma_{\mathrm{f}} \right], \mathbb{P}\left[\frac{\mathrm{d}P^{(\mathrm{c})}}{\mathrm{d}P^{(\mathrm{s})}}(\bar{\boldsymbol{Y}}^{(\mathrm{c})}) \le \gamma_{\mathrm{f}} \right] \right) \quad (6)$$

$$\bigcup_{\gamma_{\mathrm{f}} \in \bar{\mathbb{R}}} \left(\mathbb{P}\left[\frac{\mathrm{d}P^{(\mathrm{c})}}{\mathrm{d}P^{(\mathrm{s})}}(\bar{\boldsymbol{Y}}^{(\mathrm{s})}) \ge \gamma_{\mathrm{f}} \right], \mathbb{P}\left[\frac{\mathrm{d}P^{(\mathrm{c})}}{\mathrm{d}P^{(\mathrm{s})}}(\bar{\boldsymbol{Y}}^{(\mathrm{c})}) < \gamma_{\mathrm{f}} \right] \right). \quad (7)$$

where $\bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$. Here, $\bar{\boldsymbol{Y}}^{(\mathrm{c})} \sim P^{(\mathrm{c})}$ and $\bar{\boldsymbol{Y}}^{(\mathrm{s})} \sim P^{(\mathrm{s})}$, where $P^{(\mathrm{c})} = P_{\bar{\boldsymbol{Y}} | \bar{\boldsymbol{X}} = \bar{f}(\mathrm{c})}$ and $P^{(\mathrm{s})} = P_{\bar{\boldsymbol{Y}} | \bar{\boldsymbol{X}} = \bar{f}(\mathrm{s})}$.

Proof: The result follows from a direct application of the Neyman-Pearson lemma [30]. ∎

Next, we present our achievability bound, which general-izes [7, Thm. 3] to the case of noisy feedback and of a finite number of transmission rounds.

Theorem 1: Let (p_{sc}, p_{cs}) be an achievable pair according to Lemma 1 for a given choice of n_f and encoder for the feedback channel. Assume that $0 \leq p_{sc} + p_{cs} \leq 1$. Fix three integers M, ℓ_m and n, and a real number γ_{dec}. Let $(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots)$ be a stationary memoryless stochastic process where $\boldsymbol{X}_\nu \in \mathcal{X}^n$ for every integer $\nu \geq 1$. Let $P_{\boldsymbol{X}}$ denote its marginal distribution, and assume that the mutual information $I(P_{\boldsymbol{X}}, P_{\boldsymbol{Y}|\boldsymbol{X}})$, where $P_{\boldsymbol{Y}|\boldsymbol{X}}$ is the channel law defined in (1), is strictly positive.

Also, let $\boldsymbol{Y}_\nu \sim P_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{X}_\nu}$, $\nu \geq 1$, and consider a second stationary memoryless stochastic process $(\widetilde{\boldsymbol{X}}_1, \widetilde{\boldsymbol{X}}_2, \dots)$ with marginal distribution $P_{\boldsymbol{X}}$ and independent of both $(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots)$ and $(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots)$. Finally define a sequence of information density functions $\mathcal{X}^{\nu n} \times \mathcal{Y}^{\nu n} \to \mathbb{R}$

$$\imath_\nu(\mathbf{x}^\nu, \mathbf{y}^\nu) = \log \frac{dP_{\boldsymbol{Y}^\nu|\boldsymbol{X}^\nu}(\mathbf{y}^\nu|\mathbf{x}^\nu)}{dP_{\boldsymbol{Y}^\nu}(\mathbf{y}^\nu)}, \quad \nu = 1, 2, \dots \quad (8)$$

and two stopping times

$$\tau = \inf\{\nu \geq 1 : \imath_\nu(\boldsymbol{X}^\nu, \boldsymbol{Y}^\nu) \geq \gamma_{dec}\}, \quad (9)$$

$$\widetilde{\tau} = \inf\{\nu \geq 1 : \imath_\nu(\widetilde{\boldsymbol{X}}^\nu, \boldsymbol{Y}^\nu) \geq \gamma_{dec}\}. \quad (10)$$

Then, there exists an $(\ell_a, M, \epsilon, \ell_m, n, n_f)$-VLSF code whose average service time ℓ_a, is upper-bounded by

$$\ell_a \leq \sum_{\nu=0}^{\ell_m - 1} (G_{\nu+1} - G_\nu) \, \mathbb{P}[\tau > \nu] \quad (11)$$

and whose average error probability is upper-bounded by

$$\epsilon \leq \sum_{\nu=1}^{\ell_m} \xi_\nu \Big(\alpha_\nu \mathbb{P}[\tau > \nu] + (M - 1) \, \mathbb{P}[\tau \geq \nu, \widetilde{\tau} = \nu] \Big). \quad (12)$$

Here, $\alpha_\nu = p_{cs}$ for $\nu = 1, \dots, \ell_m - 1$ and $\alpha_{\ell_m} = 1$. Furthermore, $\xi_\nu = (1 - p_{cs})^{\nu-1}$ and

$$G_\nu = \sum_{k=1}^{\nu-1} k\xi_k p_{cs} + \xi_\nu \left[\sum_{k=\nu}^{\ell_m - 1} k p_{sc}^{k-\nu}(1 - p_{sc}) + \ell_m p_{sc}^{\ell_m - \nu} \right] \quad (13)$$

for $\nu = 1, \dots, \ell_m$, whereas $G_0 = 0$.

Proof: See Appendix A. ∎

Some remarks about our achievability bound are in order. As discussed in Appendix A, our bound is based on a decoder that tracks the accumulated information density between each codeword and the received signal. The stopping rule is triggered whenever the accumulated information density exceeds the threshold γ_{dec}. The random variable τ in (9) denotes the index of the first round in which the information density corresponding to the desired codeword exceeds the threshold, whereas $\widetilde{\tau}$ in (10) denotes the index of the first round in which a codeword different from the transmitted one exceeds the threshold. Clearly, the event $\tau > \widetilde{\tau}$ will correspond to an undetected error, provided that $\widetilde{\tau} \leq \ell_m$ and no $c \to s$ error has occurred in the previous rounds. This is captured by the second term in the error-probability bound (12). The first term in (12) captures instead the error resulting from a $c \to s$ event.

Note that one recovers the bound reported in [7, Thm. 3] from the bound given in Theorem 1 by setting $p_{sc} = p_{cs} = 0$ and

letting $\ell_m \to \infty$.

As shown in Appendix B, the bound given in Theorem 1 can be easily modified to account for the case in which the average latency at the receiver is the metric of interest, and ℓ_a gives a constraint on this quantity. One needs to replace (11) by

$$\ell_a \leq 1 + \sum_{\nu=1}^{\ell_m - 1} \xi_\nu \mathbb{P}[\tau > \nu]. \quad (14)$$

In the URLLC literature, (see, e.g., [31]), it is common to specify the latency t of a packet transmission as

$$t = \begin{cases} t_0' - t_0, & \text{if packet delivered error-free} \\ \infty, & \text{otherwise} \end{cases} \quad (15)$$

where t_0 is the time instance the packet is made available to the transmitter and t_0' is the time instance when the packet is delivered error-free by the receiver (t_0' is not defined if the packet is not delivered). The URLLC service requirement can then be expressed as

$$\mathbb{P}[t \leq t_{max}] \geq 1 - \epsilon_{URLLC} \quad (16)$$

where t_{max} is the latency requirement and $1 - \epsilon_{URLLC}$ is the reliability requirement.

The VLSF scheme considered in this paper will satisfy the requirement (16) if $t_{max} \geq \ell_m(n + n_f)$ channel uses and $\epsilon_{URLLC} \geq \epsilon$. However, in general, there is no simple relationship between t as defined in (15) and τ_{rx}, since τ_{rx} is finite also when there are transmission errors.

IV. NUMERICAL RESULTS

We show in this section how to use Theorem 1 to obtain guide-lines on the design of a HARQ-based short-packet transmission system operating over a wireless channel. Specifically, we are interested in understanding the performance degradation due to noise in the feedback link. Also, we seek prescriptions on how to choose the size of the codeword segments, the size of the repetition code that protects the feedback bit, the $s \to c$ and the $c \to s$ probability, and—for the fading case—the number of pilot symbols used to estimate the forward channel at the receiver.

Although our framework is general, we will consider for simplicity only the following two scenarios: 1) both the forward and the feedback channel are real-valued bi-AWGN channels operating at possibly different SNR levels, 2) both the forward and the feedback channel are Rayleigh block-memoryless fading channels operating at the same SNR level.

A. The bi-AWGN scenario

We assume that the additive noise has unit variance and that each transmit symbol belongs to the alphabet $\{-\sqrt{\rho}, \sqrt{\rho}\}$, where ρ denotes the SNR on the forward link. We also assume that the encoder for the feedback channel assigns the n_f-dimensional vector $[\sqrt{\rho_f}, \dots, \sqrt{\rho_f}]$ to the message s and the n_f-dimensional vector $[-\sqrt{\rho_f}, \dots, -\sqrt{\rho_f}]$ to c. Here, ρ_f denotes the SNR on the feedback link. Under these assumptions, it follows from Lemma 1 that for a given Neyman-Pearson threshold γ_f, the probabilities p_{sc} and p_{cs} can be expressed as
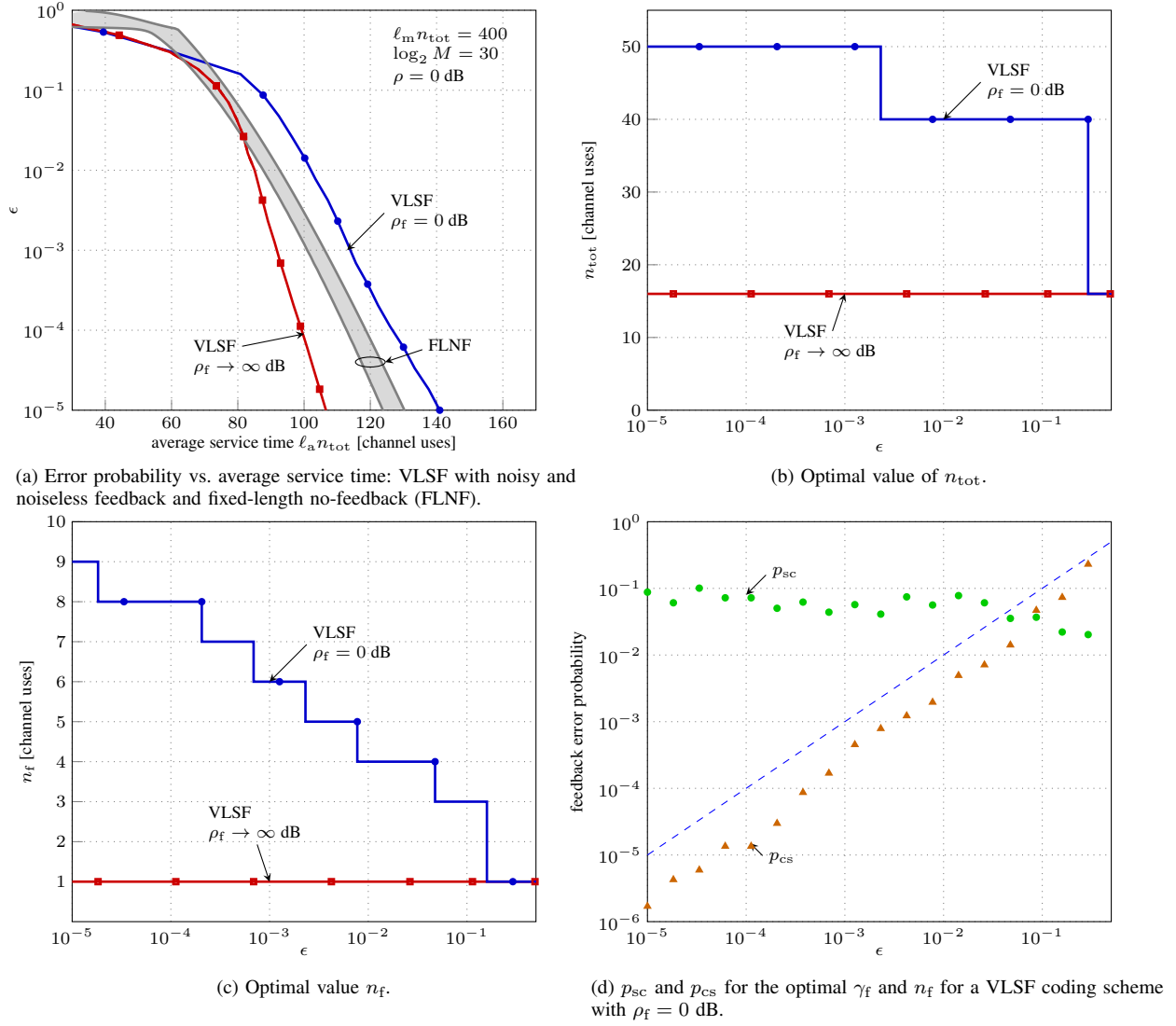
$$p_{sc} = Q(\sqrt{n_f \rho_f} + \gamma_f) \quad (17)$$

(a) Error probability vs. average service time: VLSF with noisy and noiseless feedback and fixed-length no-feedback (FLNF).

(b) Optimal value of n_{tot}.

(c) Optimal value n_{f}.

(d) p_{sc} and p_{cs} for the optimal γ_{f} and n_{f} for a VLSF coding scheme with $\rho_{\text{f}} = 0$ dB.

Fig. 3: Optimal design of VLSF coding schemes for the bi-AWGN channel.

$$p_{\text{cs}} = Q(\sqrt{n_{\text{f}}\rho_{\text{f}}} - \gamma_{\text{f}}). \tag{18}$$

Next, we evaluate the bound in Theorem 1 for a stationary memoryless input process with marginal distribution uniform over $\{-\sqrt{\rho}, \sqrt{\rho}\}$. For such a distribution, (8) reduces to

$$\imath_{\nu}(\boldsymbol{X}^{\nu}, \boldsymbol{Y}^{\nu}) \sim \sum_{i=1}^{\nu n} \log 2 - \log(1 + \exp(-2Z_i)) \tag{19}$$

where the $\{Z_i\}$ are independent and $\mathcal{N}(\rho, \rho)$ distributed. Since evaluating (12) directly is challenging, we use the following upper bound on the probability term $\mathbb{P}[\tau \geq \nu, \widetilde{\tau} = \nu]$ in (12):

$$\mathbb{P}[\tau \geq \nu, \widetilde{\tau} = \nu] \leq \mathbb{P}[\widetilde{\tau} = \nu] \tag{20}$$
$$= \mathbb{E}[\exp(-\imath_{\nu}(\boldsymbol{X}^{\nu}, \boldsymbol{Y}^{\nu}))\mathbb{1}\{\tau = \nu\}]. \tag{21}$$

The equality in (21) follows from a change-of-measure argument (see [7, Eq. (110)]). Recall that the stopping time τ, defined in (9), depends on the threshold γ_{dec}. The resulting expression can be readily evaluated using Monte-Carlo methods.

In the numerical simulations that follow, we require ℓ_{m} and

n_{tot} to be integers and fix a target maximum latency $\ell_{\text{m}} n_{\text{tot}}$, measured in channel uses, where $n_{\text{tot}} = n + n_{\text{f}}$. Then, for a given number of information bits $\log_2 M$, we use Theorem 1 to obtain an upper bound on the error probability ϵ achievable for a given constraint $\ell_{\text{a}} n_{\text{tot}}$ on the average service time measured in channel uses.[3] The bounds on the error probability reported in this section are optimized over the choice of the total number of symbols per transmission round n_{tot}, the Neyman-Pearson threshold γ_{f} in (17) and (18), and the number of feedback symbols n_{f}, under the constraint that $n + n_{\text{f}} = n_{\text{tot}}$ and that $\ell_{\text{m}} n_{\text{tot}}$ is equal to the targeted maximum latency. The optimization is performed using a grid search algorithm. The VLSF bounds reported in this section are obtained by time-sharing between the VLSF scheme used to establish Theorem 1 and a scheme in which the transmitter simply drops the packet, which results in $\ell_{\text{a}} = 0$ and $\epsilon = 1$. Specifically, let $q \in [0, 1]$ be the fraction of messages sent with the VLSF scheme and, consequently, let $1-q$

---

[3]Specifically, Theorem 1 guarantees that whenever the average-service-time constraint (measured in number of rounds) is equal to the right-hand side of (11), the error probability is smaller than the right-hand side of (12).

be the fraction of messages that are dropped at the transmitter. For each ϵ in Fig. 3, q is optimized to yield the smallest average service time. Time sharing turns out to be helpful in the high error-probability regime, i.e., when $\epsilon \geq 10^{-1}$.

We start by considering the scenario in which the latency requirement is $\ell_{\mathrm{m}} n_{\mathrm{tot}} = 400$ channel uses, $\rho = \rho_{\mathrm{f}} = 0$ dB, and $\log_2 M = 30$ bits. In Fig. 3a, we depict the upper bound on the error probability given in Theorem 1 as a function of the average-service-time constraint. For each error probability value, we present in Fig. 3b the optimum value of n_{tot}, in Fig. 3c the optimum value of n_{f}, and in Fig. 3d the optimum value of p_{sc} and p_{cs}. For comparison, we also depict in Fig. 3a an upper bound on the error probability for the case in which the feedback link is noiseless, which is obtained by letting $\rho_{\mathrm{f}} \rightarrow \infty$. Note that, in this case, setting $n_{\mathrm{f}} = 1$ minimizes the error probability, as illustrated in Fig. 3c. Finally, we plot an upper and a lower bound on the error probability achievable using a fixed-length no-feedback (FLNF) code, with blocklength $\ell_{\mathrm{a}} n_{\mathrm{tot}}$. Specifically, the upper bound is the random-coding union bound [2, Th. 16], and the lower bound is the max-min bound [2, Th. 27], evaluated using the saddlepoint approximation as described in [32].

Our results in Fig. 3a illustrate the impact of noise on the feedback channel on the error probability (estimated on the basis of our upper bound) of the specific VLSF coding scheme considered in the paper. Consider for example a target error probability $\epsilon = 10^{-5}$. When the feedback link is noiseless, the bound in Theorem 1 yields a minimum average service time of 106.6 channel uses, an optimal value for n_{tot} of 16 channel uses, and an optimal value of n_{f} equal to 1. However, when noise in the feedback link is accounted for, the average service time increases to 141 channel uses, the optimal value for n_{tot} to 50 channel uses, and the optimal value of n_{f} to 9 channel uses. The resulting average service time is larger than the one required by a FLNF coding scheme, which according to the achievability bound depicted in the figure, requires 130 channel uses to operate at $\epsilon = 10^{-5}$.[4] The performance degradation of the VLSF coding scheme is caused by the resources that need to be allocated to the feedback link to decrease the frequency of $\mathrm{c} \rightarrow \mathrm{s}$ and $\mathrm{s} \rightarrow \mathrm{c}$ errors. Specifically, as shown in Fig. 3d, to achieve $\epsilon = 10^{-5}$ it is sufficient to choose $\gamma_{\mathrm{f}} = -1.65$, which results in $p_{\mathrm{sc}} = 0.088$ and $p_{\mathrm{cs}} = 1.7 \times 10^{-6}$. Note that the $\mathrm{c} \rightarrow \mathrm{s}$ event occurs with much smaller probability than the $\mathrm{s} \rightarrow \mathrm{c}$ event.

Observe that the optimal number of channel uses n_{tot} allocated on each round increases as the optimal number of feedback symbols n_{f} increases. This has the positive effect of reducing the feedback signaling overhead; however, it has also the negative effect of reducing the maximum number of transmission rounds that are compatible with the given latency requirement.

The performance of the VLSF coding scheme for the case of noisy feedback can be improved by increasing the SNR ρ_{f} on the feedback link. This is illustrated in Fig. 4, where we plot the
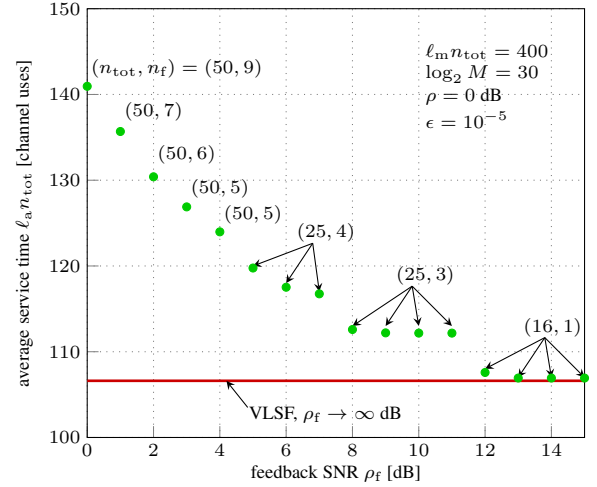

Fig. 4: Average service time of a VLSF coding scheme as a function of the SNR ρ_{f} on the feedback link. For reference, we illustrate the average service time for the noiseless case $\rho_{\mathrm{f}} \rightarrow \infty$, for which setting $n_{\mathrm{f}} = 1$ and $n_{\mathrm{tot}} = 16$ is optimal.

average service time $\ell_{\mathrm{a}} n_{\mathrm{tot}}$ as function of the SNR ρ_{f} on the feedback link. As in Fig. 3, we assume $\ell_{\mathrm{m}} n_{\mathrm{tot}} = 400$ channel uses, $\rho = 0$ dB, and $\log_2 M = 30$ bits. Furthermore, we focus on a target error probability $\epsilon = 10^{-5}$. The figure reveals that increasing the SNR ρ_{f} to around 13 dB yields an average service time close to that achievable in the noiseless-feedback case and optimal values of n_{tot} and n_{f} as in the noiseless-feedback case.

B. The Rayleigh Fading Scenario

We consider a setup in which the transmission in each round is through a quasi-static Rayleigh fading channel, i.e., the channel gain, which is Rayleigh distributed, stays constant over the transmission round. The fading coefficient is assumed to take independent realizations over different transmission rounds, according to our block-memoryless assumption. Specifically, the input-output relation is given by

$$\boldsymbol{Y}_\nu = H_\nu \boldsymbol{X}_\nu + \boldsymbol{N}_\nu. \tag{22}$$

Here, $\boldsymbol{X}_\nu \in \mathbb{C}^n$ denotes the input and the output is $\boldsymbol{Y}_\nu \in \mathbb{C}^n$. The variable $H_\nu \sim \mathcal{CN}(0,1)$ denotes the Rayleigh fading and $\boldsymbol{N}_\nu \sim \mathcal{CN}(\boldsymbol{0}, \mathsf{I}_n)$ denotes the AWGN. The random variables $\{H_\nu\}$ and $\{\boldsymbol{N}_\nu\}$ are assumed to be independent over ν.[5] Furthermore, they do not depend on $\{\boldsymbol{X}_\nu\}$. No *a priori* knowledge of the realizations of $\{H_\nu\}$ is assumed at either the transmitter or at the receiver.

We consider pilot-assisted transmission, which allows the receiver to acquire a noisy channel estimate. Specifically, similarly to [26], we consider inputs of the form $\boldsymbol{X}_\nu = [\mathbf{x}^{(\mathrm{p})}, \boldsymbol{X}_\nu^{(\mathrm{d})}]$ where $\mathbf{x}^{(\mathrm{p})} \in \mathbb{C}^{n_{\mathrm{p}}}$, $1 \leq n_{\mathrm{p}} < n$ is a deterministic vector containing pilot symbols with $\|\mathbf{x}^{(\mathrm{p})}\|^2 = n_{\mathrm{p}}\rho$, and $\boldsymbol{X}_\nu^{(\mathrm{d})} \in \widetilde{\mathcal{X}}$ contains the $n_{\mathrm{d}} = n - n_{\mathrm{p}}$ data symbols, drawn independently from a quaternary phase shift keying (QPSK) constellation, i.e., $\widetilde{\mathcal{X}} = \{\sqrt{\rho} \exp(\sqrt{-1}\frac{k\pi}{2}), k = 0, \ldots, 3\}^{n_{\mathrm{d}}}$. This choice is

---

[4] Although FLNF schemes are special cases of VLSF schemes, the bound in Theorem 1 is not able to recover the FLNF achievability bound in Fig. 3a. This is because the bound in Theorem 1 is based on a threshold decoder with a single deterministic threshold γ_{dec}, which induces a probability mass function on τ_{rx} that does not reduce, in general, to that of a FLNF coding scheme (a singleton). Furthermore, the random-coding union bound, used for the FLNF case, relies on optimal maximum likelihood decoding, whereas our VLSF bound assumes suboptimal threshold decoding.

[5] Independent fading realizations across transmission rounds can be achieved through, e.g., frequency hopping.

motivated by practical considerations. Better performance may be obtained using more sophisticated signaling schemes, for example based on the transmission of constant modulus vectors that are uniformly distributed on the power sphere, but at the price of higher receiver complexity.

Let $\mathbf{y}_\nu^{(\mathrm{p})}$ and $\mathbf{y}_\nu^{(\mathrm{d})}$ denote the received vectors corresponding to the pilot and the data symbols respectively. Given $\mathbf{x}^{(\mathrm{p})}$ and $\mathbf{y}_\nu^{(\mathrm{p})}$, the receiver computes the maximum likelihood (ML) estimate of the fading realization as

$$\widehat{h}_\nu = \frac{1}{n_{\mathrm{p}}\rho}(\mathbf{x}^{(\mathrm{p})})^{\mathsf{H}}\mathbf{y}_\nu^{(\mathrm{p})}. \tag{23}$$

We assume that the decoder treats the channel estimate as perfect and computes for each codeword the following mismatched accumulated decoding metric

$$\jmath_\nu(\mathbf{x}^\nu,\mathbf{y}^\nu) = \sum_{k=1}^{\nu}\sum_{i=1}^{n_{\mathrm{d}}} \log \frac{q(x_{k,i}^{(\mathrm{d})}, y_{k,i}^{(\mathrm{d})})}{\mathbb{E}\left[q(X, y_{k,i}^{(\mathrm{d})})\right]} \tag{24}$$

$\nu = 1, \ldots, \ell_{\mathrm{m}}.$[6] Here, $x_{k,i}^{(\mathrm{d})}$ denotes the ith element of $\mathbf{x}_k^{(\mathrm{d})}$, and $q(x_{k,i}^{(\mathrm{d})}, y_{k,i}^{(\mathrm{d})})$ is the scaled nearest-neighbor (SNN) decoding metric

$$q(x_{k,i}^{(\mathrm{d})}, y_{k,i}^{(\mathrm{d})}) = \exp\left(-|y_{k,i}^{(\mathrm{d})} - \widehat{h}_k x_{k,i}^{(\mathrm{d})}|^2\right) \tag{25}$$

and X in (24) is uniformly distributed over $\widetilde{\mathcal{X}}$. Substituting (25) into (24), we obtain

$$\jmath_\nu(\mathbf{x}^\nu,\mathbf{y}^\nu) = \sum_{k=1}^{\nu}\sum_{i=1}^{n_{\mathrm{d}}} -|y_{k,i}^{(\mathrm{d})} - \widehat{h}_k x_{k,i}^{(\mathrm{d})}|^2$$
$$- \log\mathbb{E}\left[\exp(-|y_{k,i}^{(\mathrm{d})} - \widehat{h}_k X|^2)\right]. \tag{26}$$

To adapt Theorem 1 to this mismatched-decoding setup, it is sufficient to replace i_ν in (9) and (10) with $\jmath_\nu$ in (26). As in the bi-AWGN case, evaluating $\mathbb{P}[\tau \geq \nu, \widetilde{\tau} = \nu]$ in (12) directly is challenging. Hence, we resort to the following upper bound:

$$\mathbb{P}[\tau \geq \nu, \widetilde{\tau} = \nu] \leq \mathbb{P}[\widetilde{\tau} = \nu] \leq \exp(-\gamma_{\mathrm{dec}}). \tag{27}$$

The proof of the last inequality can be found in Appendix C.

We model the feedback link in each transmission round as a quasi-static Rayleigh fading channel that is independent of the forward channel. The input-output relation in round $\nu = 1, 2, \ldots, \ell_{\mathrm{m}}$ is given as

$$\bar{\boldsymbol{Y}}_\nu = \bar{H}_\nu\bar{\boldsymbol{X}}_\nu + \bar{\boldsymbol{N}}_\nu \tag{28}$$

where $\bar{\boldsymbol{X}}_\nu \in \bar{\mathcal{X}}^{n_{\mathrm{f}}}$ denotes the input to the feedback channel in round ν and $\bar{\boldsymbol{Y}}_\nu \in \bar{\mathcal{Y}}^{n_{\mathrm{f}}}$ denotes the corresponding output. As before, $\bar{H}_\nu \sim \mathcal{CN}(0, 1)$ denotes the Rayleigh fading and $\bar{\boldsymbol{N}}_\nu \sim \mathcal{CN}(\mathbf{0}, \mathsf{I}_{n_{\mathrm{f}}})$ denotes the additive white Gaussian noise (AWGN). Again, $\{\bar{H}_\nu\}$ and $\{\bar{\boldsymbol{N}}_\nu\}$ are mutually independent and also independent over ν, and do not depend on $\{\bar{\boldsymbol{X}}_\nu\}$. Furthermore, no *a priori* knowledge of the realizations of $\{\bar{H}_\nu\}$ is assumed at the transmitter and at the receiver. Throughout this

section, we assume that the forward and the feedback channels operate at the same SNR ρ.

Since the channel is not known, we assume that the receiver uses on-off keying to signal the feedback bit. Specifically, the c and s messages are mapped to the n_{f}-dimensional vectors $[0, \ldots, 0]$ and $[\sqrt{\rho}, \ldots, \sqrt{\rho}]$, respectively. Also in this case, our choice is motivated by practical considerations; better-performing signaling schemes may be devised at the cost of higher complexity. To perform binary-hypothesis testing based on the received vector $\bar{\mathbf{y}}_\nu$, the transmitter uses the noncoherent metric[7]

$$\log \frac{P_{\bar{\boldsymbol{Y}}|\bar{\boldsymbol{X}}}(\bar{\mathbf{y}}|\bar{f}(\mathrm{c}))}{P_{\bar{\boldsymbol{Y}}|\bar{\boldsymbol{X}}}(\bar{\mathbf{y}}|\bar{f}(\mathrm{s}))} = \log(1 + \rho n_{\mathrm{f}}) - \frac{\rho}{1 + n_{\mathrm{f}}\rho}\left|\sum_{i=1}^{n_{\mathrm{f}}}\bar{y}_i\right|^2. \tag{29}$$

Hence, we have that

$$p_{\mathrm{sc}} = P_{\bar{\boldsymbol{Y}}|\bar{\boldsymbol{X}}=\bar{f}(\mathrm{s})}\left[\left|\sum_{i=1}^{n_{\mathrm{f}}}\bar{Y}_i\right|^2 \leq \gamma_{\mathrm{f}}\right] \tag{30}$$

$$= 1 - \exp\left(-\frac{\gamma_{\mathrm{f}}}{n_{\mathrm{f}}(n_{\mathrm{f}}\rho + 1)}\right) \tag{31}$$

and

$$p_{\mathrm{cs}} = P_{\bar{\boldsymbol{Y}}|\bar{\boldsymbol{X}}=\bar{f}(\mathrm{c})}\left[\left|\sum_{i=1}^{n_{\mathrm{f}}}\bar{Y}_i\right|^2 > \gamma_{\mathrm{f}}\right] \tag{32}$$

$$= \exp(-\gamma_{\mathrm{f}}/n_{\mathrm{f}}). \tag{33}$$

We consider again the scenario in which $\ell_{\mathrm{m}}n_{\mathrm{tot}} = 400$ channel uses and $\log_2 M = 30$. The SNR in both the forward and the feedback links is set to $10\,\mathrm{dB}$. We optimize the bound on the error probability over n_{tot}, n_{f}, γ_{f}, and also over the number of pilot symbols n_{p}. Furthermore, as in the bi-AWGN case, we utilize a time-sharing strategy to tighten the achievability bound for high error probabilities. We depict in Fig. 5a the error probability of the VLSF coding scheme as a function of the average service time. For comparison, we also illustrate the error probability for the case of noiseless feedback, and an achievability bound on the error probability for the FLNF case based on [33, Th.1]. To obtain the FLNF curve, we assume that, for a fixed n_{tot}, a fixed-length scheme is used over ℓ_{a} consecutive coherence intervals, with ℓ_{a} being an integer. The final curve is obtained by optimizing over n_{tot}.

We see in Fig. 5a that the presence of noise in the feedback link causes again a significant degradation of the error probability, estimated on the basis of our upper bound. For example, for the case of noiseless feedback, the minimum average service time required to achieve $\epsilon = 10^{-5}$ is 71 channel uses when $n_{\mathrm{tot}} = 50$ (see Fig. 5b). The average service time increases to 89.4 channel uses, achieved again for $n_{\mathrm{tot}} = 50$, when noise in the feedback link is taken into account. The minimum number of channel uses required by an FLNF scheme is 144 channel uses, which is achieved for $n_{\mathrm{tot}} = 16$. Differently from the bi-AWGN case, this is significantly larger than the one achievable with the VLSF scheme, even when noise in the feedback link is taken into account.

The reason behind the superior performance of the VLSF scheme in the fading case is its implicit rate-adaptation capabil-

---

[6] The logarithmic term in (24) is a special case of the generalized information density defined in [33, Eq. (3)]. Indeed, this term can be obtained from [33, Eq. (3)] by setting $s = 1$. We use the simpler expression provided in (24) to avoid performing an optimization over s, which is time consuming.

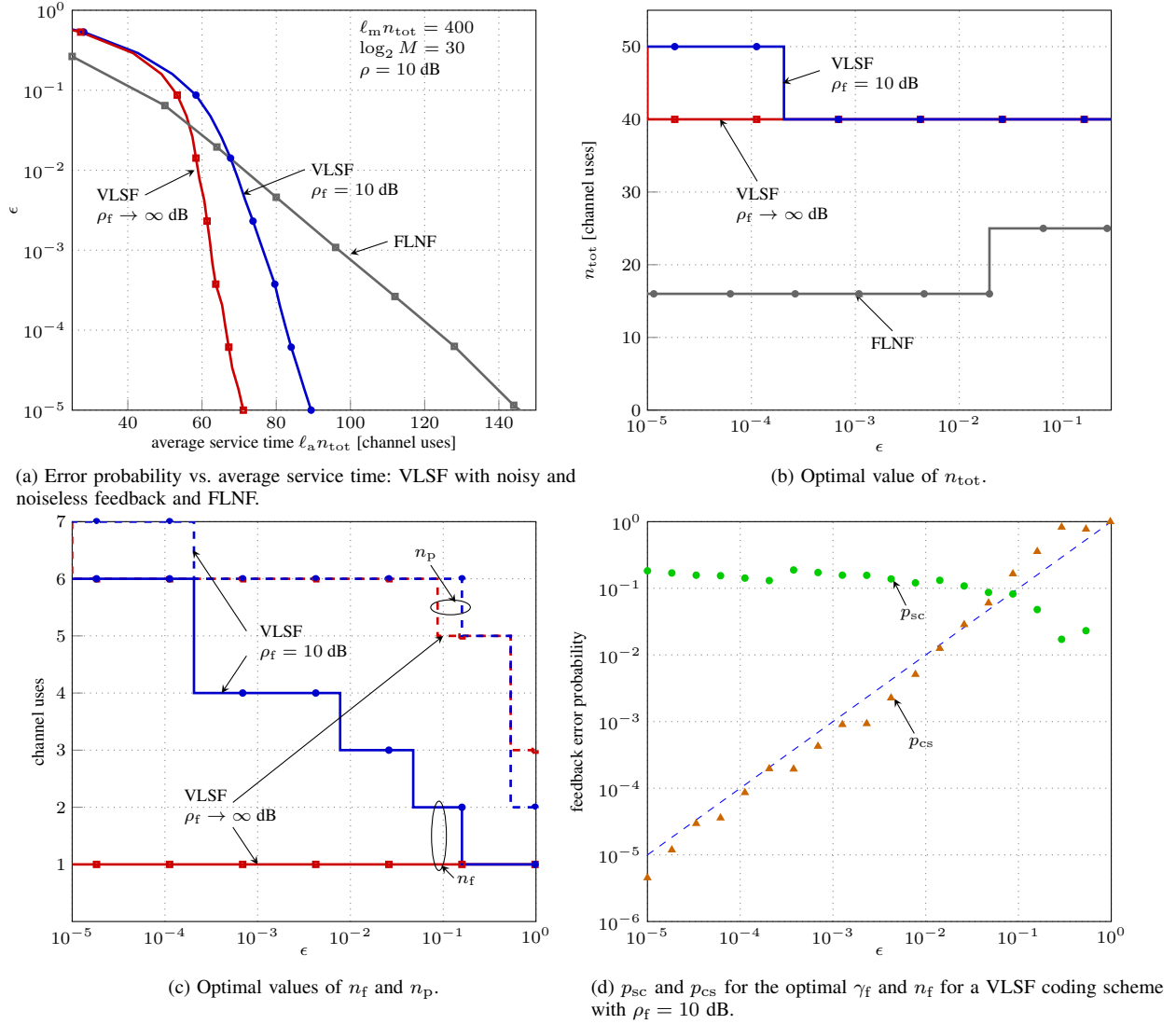[7] In what follows, we omit the index ν to keep notation compact.

(a) Error probability vs. average service time: VLSF with noisy and noiseless feedback and FLNF.

(b) Optimal value of n_{tot}.

(c) Optimal values of n_{f} and n_{p}.

(d) p_{sc} and p_{cs} for the optimal γ_{f} and n_{f} for a VLSF coding scheme with $\rho_{\text{f}} = 10$ dB.

Fig. 5: Optimal design of VLSF coding schemes for the block-memoryless Rayleigh fading channel.

ity [34]. Specifically, in the FLNF case, one needs to choose the number ℓ_{a} of coherence intervals to code over in a conservative way, to mitigate the impact of deep fades. On the contrary, in the VLSF setup, this choice is made adaptively on the basis of the instantaneous fading realizations.

In Fig. 5c, we illustrate the optimal number of pilot symbols and feedback symbols. We see that the number of pilot symbols increases as the target error probability decreases for both noiseless and noisy feedback. Furthermore, this number is essentially the same in both the noiseless-feedback and the noisy-feedback cases.

In Fig. 5d, we plot p_{sc} and p_{cs} for the optimal choice of γ_{f} and n_{f}. Observe that, for $\epsilon > 5 \times 10^{-2}$, the probability p_{cs} is actually greater than the target error probability ϵ. This is because, when $\epsilon > 5 \times 10^{-2}$, a single transmission round suffices and a c symbol is never transmitted on the feedback channel. As the target error probability ϵ decreases, retransmissions become necessary and, consequently, p_{cs} becomes smaller than ϵ.

V. CONCLUSION

We have generalized the achievability bound for VLSF coding schemes presented in [7, Thm. 3] to the case in which the feedback channel is noisy and the feedback delay is accounted for. Numerical results based on the bound provided in Theorem 1 suggest that the estimate on the minimum average service time obtainable by using [7, Thm. 3] may be optimistic when noise in the feedback link is accounted for. For example, in the bi-AWGN case, when the SNR is 0 dB, the maximum latency is 400 channel uses, and the target packet error probability is 10^{-5}, Theorem 1 yields an estimate of the average service time achievable with VLSF coding schemes that is larger than that achievable with FLNF coding schemes, once noise in the feedback link is accounted for (see Fig. 3). In the fading case, however, under the same latency and reliability requirements, Theorem 1 suggests that VLSF coding schemes are preferable to FLNF schemes even when the feedback link is noisy (see Fig. 5). The intuition is that VLSF schemes utilize the available diversity more efficiently from a service-time perspective.

Hence, our analysis suggests that care must be exercised in utilizing simplifying assumptions such as perfect acknowledgment reception in the design of URLLC systems.

As illustrated in Fig. 3c and Fig. 3d, to compensate for noise in the feedback link, which makes (uncoded) acknowledgments unreliable, Theorem 1 suggests that one has to allocate additional resources to the feedback channel. This implies that fewer resources are available on the forward channel, which yields an overall performance degradation for small values of the average service time. As shown in Fig. 4, one can compensate for such losses by transmitting the acknowledgments at a higher power level. This, however, may be unfeasible in bidirectional nonsporadic communications, where the acknowledgments are typically piggybacked on packets transmitted on the reverse data link.

We hasten to add that our observations are entirely based on an upper bound on the error probability achievable using VLSF codes, whose tightness we are not able to assess. Indeed, obtaining a tight converse bound for the case of noisy stop feedback is an open problem. In fact, even for the case of a noiseless feedback link, no VLSF converse result is known to the authors beyond the one obtainable by assuming full feedback. This implies in particular that the tightness of [7, Thm. 3]—which is the bound we generalized in this paper—is also difficult to assess.

For the case of a noiseless feedback link with $\ell_m = \infty$ and $n = 1$, Theorem 1 is known to be tight up to second order as the average blocklength grows large [7, Th. 2]. Investigating whether a similar result can be established for the noisy-feedback case is left for future work.

Our analysis is based on the simplifying assumption that the decoder is perfectly aware of whether each codeword segment contains a new information message or just incremental redundancy. One way to relax this assumption is to protect the binary flag conveying this information using a repetition code. Then, the probability that the transmitter and the receiver fall out of synchronization can be computed using Lemma 1. An extension of Theorem 1 to account for such an error event is nontrivial and is left for future work.

APPENDIX A
PROOF OF THEOREM 1

Similar to [7, Thm. 3], we start by defining a random variable U on the set [8]

$$\mathcal{U} = \underbrace{\mathcal{X}^\infty \times \cdots \times \mathcal{X}^\infty}_{M \text{ times}} \tag{34}$$

with probability mass function

$$P_U = \underbrace{P_{\boldsymbol{X}^\infty} \times \cdots \times P_{\boldsymbol{X}^\infty}}_{M \text{ times}} \tag{35}$$

where $P_{\boldsymbol{X}^\infty}$ denotes the distribution of the stationary memoryless stochastic process $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \dots\}$. Each realization of U produces M infinite-dimensional codewords $[\boldsymbol{C}_1(w), \boldsymbol{C}_2(w), \dots]$, $w = 1, \dots, M$ where each codeword segment $\boldsymbol{C}_\nu(w)$ belongs

to $\mathcal{X}^n$, $\nu = 1, 2, \dots$. The encoder f_ν maps the message w to the codeword segment $\boldsymbol{C}_\nu(w)$.

We shall next follow the so-called random coding approach and characterize the average error probability and the average service time, averaged over all codebooks constructed according to this procedure. Note that, contrary to the common application of the random coding approach, establishing an upper bound on the average service time and the average error probability averaged over all codebooks does not imply the existence of a single codebook in the ensemble that satisfies both constraints. This problem is solved by the introduction of the random variable U, which enables the use of randomized coding strategies: each time a new message is transmitted, a new codebook is drawn from the ensemble. As shown in [28, Section II] (see also [7, Thm. 19]), it turns out sufficient to perform randomization across two codebooks. This implies that the cardinality of the set over which U is defined can be reduced to 2. In practice, one could implement randomization across the two codebooks by equipping the transmitter and the receiver with a pseudo number generator, and by ensuring that the generators are initialized using the same seed.

We now continue with the proof. As detailed in Algorithm 1, the transmitter is also equipped with a stopping rule, which defines a stopping time τ_{tx} as follows:

$$\tau_{\mathrm{tx}} = \min\{\ell_m, \min\{\nu : \widehat{F}_\nu = \mathrm{s}\}\}. \tag{36}$$

Here, we use the convention that the minimum of an empty set is ∞.

At the decoding side, we consider the following stopping rule: stop at round ν if $\imath_\nu(\boldsymbol{C}^\nu(w), \boldsymbol{Y}^\nu) \geq \gamma_{\mathrm{dec}}$ for some w. Let now

$$\tau_w = \min\{\nu : \imath_\nu(\boldsymbol{C}^\nu(w), \boldsymbol{Y}^\nu) \geq \gamma_{\mathrm{dec}}\} \tag{37}$$

and let

$$\tau_{\mathrm{dec}} = \min\{\tau_1, \dots, \tau_M\}. \tag{38}$$

Finally, let [9]

$$\tau_{\mathrm{rx}} = \min\{\tau_{\mathrm{dec}}, \tau_{\mathrm{tx}} + 1, \ell_m\} \tag{39}$$

be the stopping time at the decoder. If $\tau_{\mathrm{rx}} = \tau_{\mathrm{dec}}$, the decoder sets $\widehat{W} = \max\{w : \tau_w = \tau_{\mathrm{dec}}\}$. Otherwise it sets $\widehat{W} = \mathrm{e}$. In words, an erasure is declared if no codeword results in a threshold crossing or if a c $\rightarrow$ s error occurs. Otherwise, the index of the codeword that resulted in a threshold crossing is taken as the message estimate. If a threshold crossing occurs for two or more codewords, the codeword with the largest index is chosen. Note that differently from [7, Thm. 3], where one is interested in characterizing the expected value of τ_{dec}, in our setup the quantity of interest is the expected value of τ_{tx}, whose dependence on τ_{dec} will be made explicit next.

Assume that the transmitted codeword has index w'. Since both the input process and the channel law are stationary and block memoryless, the accumulated information density $\imath_\nu(\boldsymbol{C}^\nu(w'), \boldsymbol{Y}^\nu)$, $\nu = 1, 2, \dots$ describes a random walk. Furthermore, since $\mathbb{E}[\imath_\nu(\boldsymbol{C}_\nu(w'), \boldsymbol{Y}_\nu)] = I(P_{\boldsymbol{X}}, P_{\boldsymbol{Y}|\boldsymbol{X}}) > 0$ for all

---

[8] Similar to [28, Section II] (see also [7, Thm. 19]), one can reduce the cardinality of this random variable to 2.

[9] Recall that the decoder is assumed to know the feedback bit estimate at the transmitter.

$\nu \geq 1$, this random walk drifts to $+\infty$ [35, Thm. 2.8.2]. As a consequence, we conclude that

$$\mathbb{P}[\tau_{\text{dec}} < \infty] \geq \mathbb{P}[\tau_{w'} < \infty] \tag{40}$$

$$= 1 - \mathbb{P}[\imath_\nu(\boldsymbol{C}^\nu(w'), \boldsymbol{Y}^\nu) < \gamma_{\text{dec}}, \forall \nu] \tag{41}$$

$$= 1. $$

Here, the first inequality follows from (38) and the last equality follows because the random walk drifts to $+\infty$ [35, Thm. 3.1.1].

We next prove that $\mathbb{E}[\tau_{\text{tx}}]$ can be upper-bounded as in (11). Set $G_0 = 0$ and $G_\nu = \mathbb{E}[\tau_{\text{tx}}|\tau_{\text{dec}} = \nu]$. One can show that for $\nu = 1, \ldots, \ell_{\text{m}} - 1$, the conditional expectation G_ν takes the form given in (13), whereas for $\nu \geq \ell_{\text{m}}$

$$G_\nu = \ell_{\text{m}}(1 - p_{\text{cs}})^{\ell_{\text{m}}-1} + \sum_{k=1}^{\ell_{\text{m}}} k(1 - p_{\text{cs}})^{k-1} p_{\text{cs}}. \tag{42}$$

Note that this quantity does not depend on ν. We next evaluate $\mathbb{E}[\tau_{\text{tx}}]$ as follows

$$\mathbb{E}[\tau_{\text{tx}}] = \sum_{\nu=1}^\infty G_\nu \mathbb{P}[\tau_{\text{dec}} = \nu] \tag{43}$$

$$= \lim_{\ell \to \infty} \left(\sum_{\nu=1}^\ell G_\nu \left(\mathbb{P}[\tau_{\text{dec}} > \nu - 1] - \mathbb{P}[\tau_{\text{dec}} > \nu] \right) \right) \tag{44}$$

$$= \lim_{\ell \to \infty} \left(\sum_{\nu=0}^{\ell-1} (G_{\nu+1} - G_\nu) \mathbb{P}[\tau_{\text{dec}} > \nu] - G_\ell \mathbb{P}[\tau_{\text{dec}} > \ell] \right) \tag{45}$$

$$= \sum_{\nu=0}^{\ell_{\text{m}}-1} (G_{\nu+1} - G_\nu) \mathbb{P}[\tau_{\text{dec}} > \nu]. \tag{46}$$

In the last step we used (40) and that $G_{\nu+1} = G_\nu$ for all $\nu \geq \ell_{\text{m}}$ as a consequence of (42). Note now that $G_1 > G_0$ by definition. Furthermore, standard algebraic manipulations reveal that, for $\nu = 1, \ldots, \ell_{\text{m}} - 1$,

$$G_{\nu+1} - G_\nu = \frac{(1 - p_{\text{sc}} - p_{\text{cs}})(1 - p_{\text{cs}})^{\nu-1}(1 - p_{\text{sc}}^{\ell_{\text{m}}-\nu})}{1 - p_{\text{sc}}}. \tag{47}$$

This implies that $G_{\nu+1} - G_\nu \geq 0$ whenever $p_{\text{sc}} + p_{\text{cs}} \leq 1$. To obtain the desired result, we notice that

$$\mathbb{P}[\tau_{\text{dec}} > \nu] \leq \frac{1}{M} \sum_{w=1}^M \mathbb{P}[\tau_w > \nu|W = w] = \mathbb{P}[\tau > \nu] \tag{48}$$

where τ is defined in (9).

We now prove (12). First note that, since if threshold crossing occurs for more than one codeword, the one with largest index is chosen,

$$\epsilon = \frac{1}{M} \sum_{w=1}^M \mathbb{P}\left[\widehat{W} \neq w|W = w\right] \tag{49}$$

$$\leq \mathbb{P}\left[\widehat{W} \neq 1|W = 1\right] \tag{50}$$

$$= \sum_{\nu=1}^{\ell_{\text{m}}} \mathbb{P}\left[\tau_{\text{rx}} = \nu, \widehat{W} \neq 1|W = 1\right]. \tag{51}$$

Next, we decompose each term on the right-hand-side of (51). For $\nu = 1$, the error probability coincides with the probability

that an undetected error occurs in round 1, i.e.,

$$\mathbb{P}\left[\tau_{\text{rx}} = 1, \widehat{W} \neq 1|W = 1\right]$$
$$= \mathbb{P}\left[\tau_{\text{dec}} = 1, \widehat{W} \neq 1|W = 1\right]. \tag{52}$$

For $\nu = 2, \ldots, \ell_{\text{m}} - 1$, we have

$$\mathbb{P}\left[\tau_{\text{rx}} = \nu, \widehat{W} \neq 1|W = 1\right]$$
$$= \mathbb{P}\left[\tau_{\text{tx}} = \nu - 1, \tau_{\text{dec}} > \nu - 1, \widehat{W} \neq 1|W = 1\right]$$
$$+ \mathbb{P}\left[\tau_{\text{tx}} \geq \nu, \tau_{\text{dec}} = \nu, \widehat{W} \neq 1|W = 1\right] \tag{53}$$
$$= \mathbb{P}[\tau_{\text{tx}} = \nu - 1|\tau_{\text{dec}} > \nu - 1, W = 1]$$
$$\times \mathbb{P}[\tau_{\text{dec}} > \nu - 1|W = 1]$$
$$+ \mathbb{P}\left[\tau_{\text{tx}} \geq \nu|\tau_{\text{dec}} = \nu, \widehat{W} \neq 1, W = 1\right]$$
$$\times \mathbb{P}\left[\tau_{\text{dec}} = \nu, \widehat{W} \neq 1|W = 1\right]. \tag{54}$$

The first term on the right-hand side of (54) is the probability that an erasure is declared at step ν because of a $\text{c} \to \text{s}$ event at step $\nu - 1$ and the second term on the right-hand side of (54) corresponds to the probability of an undetected error. Observe now that

$$\mathbb{P}[\tau_{\text{dec}} > \nu - 1|W = 1] \leq \mathbb{P}[\tau > \nu - 1]. \tag{55}$$

Furthermore,

$$\mathbb{P}\left[\tau_{\text{dec}} = \nu, \widehat{W} \neq 1|W = 1\right]$$
$$= \mathbb{P}\left[\cup_{m=2}^M \{\tau_1 \geq \nu, \tau_m = \nu\}|W = 1\right] \tag{56}$$
$$\leq (M - 1)\mathbb{P}[\tau_1 \geq \nu, \tau_2 = \nu|W = 1] \tag{57}$$
$$= (M - 1)\mathbb{P}[\tau \geq \nu, \widetilde{\tau} = \nu] \tag{58}$$

where $\widetilde{\tau}$ is defined in (10). Finally, we have that

$$\mathbb{P}\left[\tau_{\text{tx}} \geq \nu|\tau_{\text{dec}} = \nu, \widehat{W} \neq 1, W = 1\right] = (1 - p_{\text{cs}})^{\nu-1} \tag{59}$$

and that

$$\mathbb{P}[\tau_{\text{tx}} = \nu - 1|\tau_{\text{dec}} > \nu - 1, W = 1] = (1 - p_{\text{cs}})^{\nu-2} p_{\text{cs}}. \tag{60}$$

For $\nu = \ell_{\text{m}}$, the error probability is given by the sum of the terms in (54) computed for $\nu = \ell_{\text{m}}$, and the additional term

$$\mathbb{P}[\tau_{\text{tx}} \geq \ell_{\text{m}}|\tau_{\text{dec}} > \ell_{\text{m}}, W = 1]\mathbb{P}[\tau_{\text{dec}} > \ell_{\text{m}}|W = 1]$$
$$\leq (1 - p_{\text{cs}})^{\ell_{\text{m}}-1}\mathbb{P}[\tau > \ell_{\text{m}}]. \tag{61}$$

This term describes the probability that no codeword causes a threshold crossing within ℓ_{m} transmission rounds and no errors occurred on the feedback channel. We obtain the desired bound by substituting (55), (58), (59), and (60) into (52) and (54) and then (52), (54), and (61) into (51).

APPENDIX B
UPPER BOUND ON $\mathbb{E}[\tau_{\text{rx}}]$

Let τ_{rx} be defined as in (39). Furthermore, let $V_\nu = \mathbb{E}[\tau_{\text{rx}}|\tau_{\text{dec}} = \nu]$ for $\nu = 1, 2, \ldots$ and $V_0 = 0$. The steps to bound $\mathbb{E}[\tau_{\text{rx}}]$ are analogous to the ones used to bound $\mathbb{E}[\tau_{\text{tx}}]$ in

Appendix A. First, note that $V_1 = 1$. Next, we write

$$\mathbb{E}[\tau_{\rm rx}] = \sum_{\nu=0}^{\infty} (V_{\nu+1} - V_\nu)\, \mathbb{P}[\tau_{\rm dec} > \nu] \qquad (62)$$

where

$$V_\nu = \nu \mathbb{P}[\tau_{\rm tx} \geq \nu | \tau_{\rm dec} = \nu]$$
$$+ \sum_{k=2}^{\nu} k \mathbb{P}[\tau_{\rm tx} = k - 1 | \tau_{\rm dec} = \nu] \qquad (63)$$

$$= \nu(1 - p_{\rm cs})^{\nu-1} + \sum_{k=2}^{\nu} k(1 - p_{\rm cs})^{k-2} p_{\rm cs} \qquad (64)$$

for, $\nu = 2, \ldots, \ell_{\rm m}$, and

$$V_\nu = \ell_{\rm m}(1 - p_{\rm cs})^{\ell_{\rm m}-1} + \sum_{k=1}^{\ell_{\rm m}} k(1 - p_{\rm cs})^{k-2} p_{\rm cs} \qquad (65)$$

for $\nu > \ell_{\rm m}$. Note that $V_1 > V_0$ and $V_{\nu+1} - V_\nu = 0$ for $\nu \geq \ell_{\rm m}$. Finally, for $\nu = 1, \ldots, \ell_{\rm m} - 1$, we have

$$V_{\nu+1} - V_\nu = (1 - p_{\rm cs})^{\nu-1}. \qquad (66)$$

Hence, we conclude that $V_{\nu+1} - V_\nu > 0$ and that

$$\mathbb{E}[\tau_{\rm rx}] \leq 1 + \sum_{\nu=1}^{\ell_{\rm m}-1} (1 - p_{\rm cs})^{\nu-1} \mathbb{P}[\tau > \nu]. \qquad (67)$$

APPENDIX C
PROOF OF (27)

By using Jensen's inequality in (24), we have that

$$\mathbb{E}\left[\jmath_\nu(\widetilde{\boldsymbol{X}}^\nu, \boldsymbol{Y}^\nu)\right] \leq 0. \qquad (68)$$

Since $\jmath_\nu(\widetilde{\boldsymbol{X}}^\nu, \boldsymbol{Y}^\nu)$ is a sum of ν independent and identically distributed random variables, we conclude that each random variable has a negative mean. Such a property allows us to use Wald's identity [36, Cor. 9.4.4] and conclude that

$$\mathbb{P}[\tilde{\tau} = \nu] \leq \mathbb{P}\left[\jmath_\nu(\widetilde{\boldsymbol{X}}^\nu, \boldsymbol{Y}^\nu) \geq \gamma_{\rm dec}\right] \qquad (69)$$
$$\leq \exp(-\beta^* \gamma_{\rm dec}) \qquad (70)$$

Here, β^* is the positive solution of

$$\kappa(\beta) = \log \mathbb{E}\left[\exp\left(\beta \jmath_1(\widetilde{\boldsymbol{X}}_1, \boldsymbol{Y}_1)\right)\right] = 0. \qquad (71)$$

Substituting (26) in (71) we find that $\beta^* = 1$. Substituting this value in (70), we obtain the desired result.

REFERENCES

[1] J. Östman, R. Devassy, G. Durisi, and E. G. Ström, "On the nonasymptotic performance of variable-length codes with noisy stop feedback," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Visby, Sweden, Aug. 2019.

[2] 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.

[3] 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.

[4] 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.

[5] A. Collins and Y. Polyanskiy, "Coherent multiple-antenna block-fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 380–405, Jan. 2019.

[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?" *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1521–1536, Feb. 2019.

[7] 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.

[8] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.

[9] H. Shariatmadari, S. Iraji, R. Jantti, P. Popovski, Z. Li, and M. A. Uusitalo, "Fifth-generation control channel design: Achieving ultrareliable low-latency communications," *IEEE Veh. Technol. Mag.*, vol. 13, no. 2, pp. 84–93, Jun. 2018.

[10] E. Dahlman, S. Parkvall, and J. Sköld, *4G LTE/LTE-Advanced for Mobile Broadband*. Burlington, MA, U.S.A.: Elsevier, 2011.

[11] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels—Part I," *Inf. Contr.*, vol. 10, pp. 65–103, Feb. 1967.

[12] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, U.S.A.: John Wiley & Sons, 1968.

[13] R. L. Dobrushin, "An asymptotic bound for the probability error of information transmission through a channel without memory using the feedback," *Problemy Kibernetiki*, vol. 8, pp. 161–168, 1961.

[14] M. V. Burnashev and H. Yamamoto, "On the reliability function for a BSC with noisy feedback," *Probl. Inf. Transm.*, vol. 46, no. 2, pp. 3–23, Jan. 2010.

[15] M. V. Burnashev, "Data transmission over a discrete channel with feedback. Random transmission time," *Probl. Inf. Transm.*, vol. 12, no. 4, pp. 10–30, Dec. 1976.

[16] L. V. Truong and V. Y. F. Tan, "Moderate deviation asymptotics for variable-length codes with feedback," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4364 – 4386, Jul. 2019.

[17] S. C. Draper and A. Sahai, "Variable-length channel coding with noisy feedback," *Eur. Trans. Telecommun.*, vol. 19, pp. 355–370, Apr. 2008.

[18] U. Niesen and A. Tchamkerten, "Tracking stopping times through noisy observations," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 422–432, Jan. 2009.

[19] G. D. Forney Jr, "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Trans. Inf. Theory*, vol. 14, no. 2, pp. 206–220, Mar. 1968.

[20] I. Telatar, "Multi-access communication with decision feedback decoding," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, May 1992.

[21] 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.

[22] A. Avranas, M. Kountouris, and P. Ciblat, "Energy-latency tradeoff in ultra-reliable low-latency communication with retransmissions," *IEEE J. Sel. Areas Commun.*, vol. 11, no. 36, pp. 2475–2485, Nov. 2018.

[23] B. Makki, T. Svensson, G. Caire, and M. Zorzi, "Fast HARQ over finite blocklength codes: A technique for low-latency reliable communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 194–209, Jan. 2018.

[24] S. H. Kim, D. K. Sung, and T. Le-Ngoc, "Variable-length feedback codes under a strict delay constraint," *IEEE Commun. Lett.*, vol. 19, no. 4, pp. 513–516, Apr. 2015.

[25] 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.

[26] J. Östman, R. Devassy, G. C. Ferrante, and G. Durisi, "Low-latency short-packet transmissions: Fixed length or HARQ?" in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Abu Dhabi, UAE, Dec. 2018.

[27] 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," *IEEE J. Sel. Areas Commun.*, vol. 4, no. 37, pp. 721–734, Apr. 2019.

[28] K. F. Trillingsgaard, W. Yang, G. Durisi, and P. Popovski, "Common-message broadcast channels with feedback in the nonasymptotic regime: Stop feedback," *IEEE Trans. Inf. Theory*, vol. 64, no. 12, pp. 7686–7718, Dec. 2018.

[29] E. Malkamäki and H. Leib, "Performance of truncated type-II hybrid ARQ schemes with noisy feedback over block fading channels," *IEEE Trans. Commun.*, vol. 48, no. 9, pp. 1477–1487, Sep. 2000.

[30] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Phil. Trans. Roy. Soc. A*, vol. 231, pp. 289–337, Jan. 1933.

[31] P. Popovski, J. J. Nielsen, C. Stefanović, E. d. Carvalho, E. G. Ström, K. F. Trillingsgaard, A.-S. Bana, D. M. Kim, R. Kotaba, J. Park, and R. B. Sørensen, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *IEEE Network Magazine*, vol. 32, no. 2, pp. 16–23, Mar. 2018.

[32] J. Font-Segura, G. Vazquez-Vilar, A. Martinez, A. Guillén i Fàbregas, and A. Lancho, "Saddlepoint approximations of lower and upper bounds to the error probability in channel coding," in *Proc. Conf. Inf. Sci. Sys. (CISS)*, Princeton, NJ, USA, Mar. 2018.

[33] 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.

[34] P. Wu and N. Jindal, "Performance of hybrid-ARQ in block-fading channels: A fixed outage probability analysis," *IEEE Trans. Commun.*, vol. 58, no. 4, pp. 1129–1141, Apr. 2010.

[35] A. Gut, *Stopped random walks: limit theorems and applications*, 2nd ed. New York, NY, USA: Springer, 2009.

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

Johan Östman (S'16) received the B.Sc. and M.Sc. degrees in electrical engineering from Chalmers University of Technology, Gothenburg, Sweden, in 2011 and 2014, respectively, where he is currently pursuing the Ph.D. degree in electrical engineering.

Rahul Devassy (S'14) received the B.Tech. degree in electrical and electronics engineering from the National Institute of Technology, Calicut, India, in 2006, and the M.E degree in telecommunication from the Indian Institute of Science, Bangalore, India, in 2009. In 2019, he obtained the Ph.D. degree in electrical engineering from Chalmers University of Technology, Gothenburg, Sweden. His research interests are in the areas of information and communication theory.

Giuseppe Durisi (S'02–M'06–SM'12) received the Laurea degree *summa cum laude* and the Doctor degree both from Politecnico di Torino, Italy, in 2001 and 2006, respectively. From 2002 to 2006, he was with Istituto Superiore Mario Boella, Torino, Italy. From 2006 to 2010 he was a postdoctoral researcher at ETH Zurich, Zurich, Switzerland. In 2010, he joined Chalmers University of Technology, Gothenburg, Sweden, where he is now full professor with the Communication Systems Group. He is co-director of Chalmers ICT Area of Advance, and of Chalmers AI Research Center.

Dr. Durisi is a senior member of the IEEE. He is the recipient of the 2013 IEEE ComSoc Best Young Researcher Award for the Europe, Middle East, and Africa Region, and is co-author of a paper that won a "student paper award" at the 2012 International Symposium on Information Theory, and of a paper that won the 2013 IEEE Sweden VT-COM-IT joint chapter best student conference paper award. In 2015, he joined the editorial board of the IEEE Transactions on Communications as associate editor. From 2011 to 2014, he served as publications editor for the IEEE Transactions on Information Theory. His research interests are in the areas of communication and information theory and machine learning.

Erik G. Ström (S'93–M'95–SM'01) received the M.S. degree from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 1990, and the Ph.D. degree from the University of Florida, Gainesville, in 1994, both in electrical engineering. He accepted a postdoctoral position at the Department of Signals, Sensors, and Systems at KTH in 1995. In February 1996, he was appointed Assistant Professor at KTH, and in June 1996 he joined Chalmers University of Technology, Göteborg, Sweden, where he is now a Professor in Communication Systems since June 2003. Dr. Ström currently heads the Division of Communications, Antennas, and Optical Networks, is the director of ChaseOn, a Vinnova Competence Center focused on antenna system. His research interests include signal processing and communication theory in general, and constellation labelings, channel estimation, synchronization, multiple access, medium access, multiuser detection, wireless positioning, and vehicular communications in particular. Since 1990, he has acted as a consultant for the Educational Group for Individual Development, Stockholm, Sweden. He is a contributing author and associate editor for Roy. Admiralty Publishers FesGas-series, and was a co-guest editor for the *Proceedings of the IEEE* special issue on Vehicular Communications (2011) and the *IEEE Journal on Selected Areas in Communications* special issues on Signal Synchronization in Digital Transmission Systems (2001) and on Multiuser Detection for Advanced Communication Systems and Networks (2008). Dr. Ström was a member of the board of the IEEE VT/COM Swedish Chapter 2000–2006. He received the Chalmers Pedagogical Prize in 1998 and the Chalmers Ph.D. Supervisor of the Year award in 2009.