Delay-Sensitive Communications Over IR-HARQ: Modulation, Coding Latency, and Reliability

Cenk Sahin, *Member, IEEE*, Lingjia Liu, *Senior Member, IEEE*, Erik Perrins, *Senior Member, IEEE*, and Liangping Ma, *Senior Member, IEEE*

***Abstract*— With the growing popularity of delay-sensitive applications (e.g., real-time conversational video, online gaming, and augmented reality) and future trends toward ultra-reliable low-latency communications such as the tactile Internet, performance analysis of wireless systems under the finite code blocklength constraint becomes extremely important. In this paper, we investigate the maximum achievable throughput of incremental redundancy-hybrid automatic repeat request (IR-HARQ) over the (correlated) Rayleigh fading channel under finite blocklength and delay-violation probability constraints as a function of the modulation scheme. The maximum number of HARQ rounds together with the transport block size specifies the underlying coding latency of the IR-HARQ scheme. A framework, namely the HARQ Markov model (HARQ-MM), is introduced to track the throughput and the probability of error of IR-HARQ over the Rayleigh fading channel as a function of the modulation scheme. The dispersion of parallel additive white Gaussian noise channels with finite input alphabets (e.g., pulse amplitude modulation) is analytically characterized. It is used to identify the state transition probabilities of the underlying HARQ-MM. An algorithm is developed to efficiently compute the steady-state distribution of the HARQ-MM. Extensive performance evaluation is conducted, which shows a good match between the throughput performance characterized by the theoretical framework and that achieved by the practical channel codes.**

***Index Terms*— 5G, URLLC, delay-violation probability, delay-sensitive, IR-HARQ, and finite blocklength coding.**

I. Introduction

DEMAND for data throughput, and connectivity in wireless systems has been increasing at a fast pace. Due to the scarcity of the available radio spectrum, significant research has been devoted to developing techniques and strategies that enhance the spectral efficiency of wireless systems at the physical layer. The prevalent framework used to evaluate these techniques is information theory, with emphasis on the Shannon capacity of wireless systems. By focusing on the channel coding performance of asymptotically long codes, this framework is suitable to analyze the throughput performance of wireless systems without code blocklength constraints. However, mobile applications such as voice, real-time conversational video, online gaming and augmented reality impose stringent requirements on delay and render the use of asymptotically long channel codes prohibitive. Furthermore, one of the use cases of 5G networks is ultra-reliable low-latency communications (URLLC) where there are stringent requirements on both delay and reliability [3], [4]. Consequently, performance analysis and design of wireless communication systems under both the finite blocklength constraint and the reliability constraint becomes extremely important.

The time-varying nature of wireless media is one of the key challenges in analyzing the performance of wireless systems. Channel correlation plays a critical role in the quality of service assessment of systems with delay-sensitive traffic [5]–[8]; the assumption of independent and identically distributed (i.i.d.) channel realizations may result in an overly optimistic performance evaluation. In modern communication systems such as 3GPP LTE/LTE-Advanced and IEEE 802.16m, data are partitioned into *transport blocks* (TBs) [9]. The TB size is chosen according to the feedback information as well as the underlying channel characteristics such as the Doppler frequency and the spectral bandwidth. When the channel state information (CSI) is not available at the transmitter, the transmitter fixes the TB size such that the wireless channel is quantized to a *finite-state channel* where each TB goes through a different state. We can model this finite-state channel as a *Markov channel* [10], [11]. In finite-state Markov channel (FSMC) models, the channel fading level is partitioned into a finite number of states, and the channel state evolves as a Markov chain across time. This FSMC model allows us to abstract the impact of the physical layer communication strategies such as modulation and coding schemes (MCSs) on link quality, perceived quality of service, and system throughput.

The study of channel coding bounds in the finite blocklength regime has gained significant momentum after the seminal work of Polyanskiy et al. [12]. The authors developed new bounds as well as an accurate approximation—based on a quantity called *channel dispersion*—to the maximum

Manuscript received June 22, 2018; revised December 10, 2018; accepted January 25, 2019. Date of publication February 12, 2019; date of current version March 15, 2019. The work of C. Sahin, L. Liu, and E. Perrins was supported by the National Science Foundation under Grant CCF-1422241, Grant ECCS-1802710, Grant ECCS-1811497, and Grant CNS-1811720. This paper was presented in part at the Proceedings of the IEEE Global Telecommunications Conference, Austin, TX, USA, December 2014 [1] and the Proceedings of the IEEE Global Telecommunications Conference Workshop on Ultra-Low Latency and Ultra-High Reliability in Wireless Communications, San Diego, CA, USA, December 2015 [2]. *(Corresponding author: Lingjia Liu.)*

C. Sahin is with the Air Force Research Laboratory, Wright-Patterson Air Force Base, OH 45433 USA.

L. Liu is with the Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24061 USA (e-mail: ljliu@ieee.org).

E. Perrins is with the Department of Electrical Engineering and Computer Science, The University of Kansas, Lawrence, KS 66045 USA.

L. Ma is with InterDigital Communications, San Diego, CA 92121 USA.

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

Digital Object Identifier 10.1109/JSAC.2019.2898784

achievable rate for a given blocklength and a probability of block error [12]. In [13], the dispersion of a two-state FSMC model with bit-wise state transitions (i.e. Gilbert-Elliott channel (GEC) [14], [15]) was derived under the assumption of CSI availability only at the receiver. In the proposed model, the channel in each state was modeled as a binary symmetric channel (BSC). In [16] and [17], we characterized the dispersion of FSMCs with an arbitrary number of states, and TB-wise state transitions. The results showed that the required blocklength to achieve rates near (i.e. 90% of) the Shannon capacity is in the order of hundreds of TBs.

A well known way to improve coding performance without greatly increasing the blocklength is to use *one-bit feedback* (i.e. acknowledge (ACK)) to terminate the transmission of a codeblock when the receiver has a reliable estimate of the transmitted message [18]. Such schemes are known as incremental redundancy type hybrid automatic repeat request (IR-HARQ) [19], [20] and effectively accumulate mutual-information from multiple transmissions [21]. The performance of IR-HARQ schemes in the finite blocklength regime over discrete memoryless channels (DMCs) was studied in [22]. In [23], the performance of IR-HARQ over additive white Gaussian noise (AWGN) channels was studied under the finite blocklength assumption. In [24], the finite blocklength coding performance of Type-I HARQ—where codeblocks replace uncoded data packets in conventional ARQ, and a received codeblock is discarded after a failed decoding attempt—over Rayleigh block-fading channels with Gaussian input symbols was considered. Likewise, in [25] a finite blocklength IR-HARQ scheme over the Rayleigh block-fading channel with Gaussian input symbols was considered for an investigation of HARQ buffer management. The i.i.d. assumption of the block-fading channel omits the underlying channel correlation/fading characteristics. Furthermore, the model adopted in [25] allows the channel state to transition to any other state in a single time step leading to overly optimistic throughput performance. Note that the evaluation of all modern wireless systems such as 3GPP LTE-Advanced and IEEE 802.16m relies heavily on correlated channel models instead of block-fading models and it is extremely meaningful to investigate the impacts of practical MCSs on the underlying latency as well as the reliability of a wireless communication system.

In general, the latency of a wireless system has several major components: queuing latency, scheduling latency, and coding latency. A comprehensive study of overall system latency as a function of the underlying communication strategy is not mathematically tractable since different latency components depend heavily on the underlying strategy as well as the underlying network operation. A divide-and-conquer approach is usually adopted where different latency components are treated separately. For example, our previous work [5], [6], [26]–[28] provides comprehensive study of the impacts of resource allocation strategies on queuing and scheduling latency of a wireless system without considering the coding latency and the reliability. To make the study relevant and applicable to 5G URLLC, in this paper, we mainly focus on the coding latency and the reliability of a wireless system. Specifically, we aim to characterize the throughput performance of IR-HARQ over the correlated Rayleigh fading channel under coding latency, modulation, as well as reliability constraints. Since modern wireless systems use TBs as the basic physical layer unit, for a fixed medium access control (MAC) protocol, the maximum number of HARQ rounds translates to the maximum transmission delay. Therefore, without loss of generality, in this paper, we use the maximum number of HARQ rounds together with the duration of each TB to represent the underlying coding latency. Accordingly, an error event at the last HARQ round is equivalent to a delay-violation event.

To realize our goal, an analytical framework is introduced to characterize the maximum achievable throughput of IR-HARQ as a function of the modulation scheme under maximum number of HARQ rounds and probability of error constraints. In the rest of the paper, we will refer to IR-HARQ by simply HARQ. The MCS is assumed to be fixed throughout the transmission. Adaptive MSC is possible with CSI feedback; however, this operation will generally incur additional feedback overhead and delay, which is not desirable for URLLC. The framework uses the HARQ Markov model (HARQ-MM), introduced here, to track the HARQ throughput and HARQ probability of error as a function of the coding rate, the modulation scheme, the number of HARQ rounds, the signal-to-noise ratio (SNR) and the Doppler frequency. The HARQ-MM is a finite-state Markov chain that builds upon the FSMC model of the Rayleigh fading channel. We consider two HARQ schemes: hard-decision HARQ and soft-decision HARQ. With hard-decision (soft-decision) HARQ the channel input-output relationship in each FSMC state is modeled by a BSC with a fixed crossover probability (a AWGN channel with a fixed SNR). With both schemes the state transition probabilities of the HARQ-MM are approximated by the dispersion associated with different channel state sequence realizations. For hard-decision HARQ we use existing results on the dispersion of parallel DMCs while for soft-decision HARQ we derive the dispersion of parallel AWGNs with discrete input alphabets (e.g. pulse amplitude modulation (PAM)).

Major contributions of this paper can be summarized as follows.

- First, a comprehensive analytical framework, called HARQ-MM, is introduced to evaluate the performance of a URLLC system (coding latency and reliability) over IR-HARQ as a function of the MCS under a correlated Rayleigh fading channel. By linking the maximum number of HARQ rounds as well as the TB duration to the coding latency, this framework allows us to evaluate the maximum HARQ throughput under different MCSs and delay-violation probability constraints.
- Second, new information theoretic results are derived for AWGN channels with finite input alphabets, where the input alphabet is specified by the underlying modulation scheme. In particular, the dispersion of parallel AWGN channels as a function of the modulation scheme is derived for the first time. The channel dispersion is used to track error probabilities of IR-HARQ over the wireless channel, and hence paves the way for tracking the coding

latency and reliability over IR-HARQ as a function of the MCS.

- Third, a novel algorithm is introduced in Section IV to effectively compute the steady-state probabilities of the underlying HARQ-MM and the HARQ throughput under a delay-violation constraint for a large number of HARQ rounds and FSMC states. The introduced algorithm extends our previous work in [1] and [2] as the results in [1] and [2] are only applicable to the case where there is a small number of channel states and a small number of HARQ rounds.
- Fourth, extensive performance evaluation has been conducted. The impacts of coding latency, MCSs, and reliability on the HARQ throughput are evaluated. The throughput performance characterized using our analytical framework is compared to the throughput performance simulated using Luby transform [29] coded IR-HARQ systems. The analytical results match very well with the simulated ones, which validates the introduced framework.

The paper is organized as follows. In Section II, we summarize the representation of the Rayleigh fading channel by a FSMC. In Section III, we describe the HARQ system under consideration, introduce the HARQ-MM and define system performance measures. In Section IV, we characterize the maximum achievable HARQ throughput, introduce an algorithm to efficiently compute the steady-state distribution of the HARQ-MM, and analyze the relationship between the number of HARQ rounds and the HARQ throughput performance. In Section V, we present numerical results, which demonstrate the accuracy of our framework. Finally in Section VI, we state our conclusions.

II. Channel Model

The complex baseband received signal through a Rayleigh wireless fading channel is represented by [30]

$$r(t) = h(t)x(t) + w(t) \tag{1}$$

where $x(t)$ is the transmitted signal with *fixed power* P, $w(t)$ is a zero-mean complex Gaussian process with independent real and imaginary parts, each part with power spectral density (p.s.d.) $N_0/2$, and $h(t)$ is the multipath fading component. The transmitted signal $x(t)$ is the baseband output of a *linear* modulator (e.g. quadrature amplitude modulation (QAM)) with symbol rate (expressed in symbols/s) equal to the available *bandwidth* denoted by B (Hz) (i.e. the *normalized symbol rate* is assumed to be 1 symbol/s/Hz). The fading component $h(t)$ is modeled as a zero-mean complex Gaussian process with i.i.d. real and imaginary parts. The envelope process $|h(t)| \geq 0$ follows a Rayleigh probability distribution, and the average *power gain* is normalized to unity, i.e. $\mathbb{E}[|h(t)|^2] = 1$. The autocorrelation function of $h(t)$ is modeled by a zeroth-order Bessel function of the first kind $J_0(2\pi f_{\mathrm{D}} t)$ where f_{D} is the *Doppler frequency* [31]. For mathematical tractability we simplify this channel model to a *finite-state* channel model.

Let $\boldsymbol{\eta} = [\eta_1, \eta_2, \cdots, \eta_{K+1}], \eta_1 = 0, \eta_{K+1} = \infty$, be a partitition of the fading envelope range. We assume that at any given time the probability of $|h(t)|$ being in one of K subintervals, $[\eta_k, \eta_{k+1}), k = 1, \cdots, K$, can be modeled as a K-state *continuous-time* Markov chain [5]. If $|h(t)| \in [\eta_k, \eta_{k+1})$, the channel is said to be in state k. We further assume that at the physical layer *data* is partitioned into TBs as in LTE/LTE-Advanced systems [9]. The channel fading envelope $|h(t)|$ varies during a TB transmission; however, it is assumed to remain within a single state during the transmission of an entire TB [11]. Effectively, the underlying continuous-time Markov chain is sampled every t_{TB} seconds to form a *discrete-time* Markov chain where t_{TB} is the TB duration [11]. This channel model is known in the literature as the FSMC [10], [11]; we refer to it as the *TB-based* FSMC because of the TB-wise state transitions.

The fading envelope partition $\boldsymbol{\eta}$, the TB duration t_{TB}, and the *Doppler frequency* f_{D} are sufficient to fully characterize the underlying Markov chain under the assumption that the state of the FSMC only transitions to adjacent states, i.e $P_{k,j} = 0$ whenever $|k - j| > 1$, where $P_{k,j}$ is the transition probability from state k to state j. The state transition probabilities are approximated by the expressions [10]

$$\begin{aligned} P_{k,k+1}\left(\eta_k, \eta_{k+1}, f_{\mathrm{D}}, t_{\mathrm{TB}}\right) &\approx \frac{N(\eta_{k+1}, f_{\mathrm{D}}) t_{\mathrm{TB}}}{p_k(\eta_k, \eta_{k+1})}, \quad 1 \leq k \leq K-1 \\ P_{k,k-1}\left(\eta_k, \eta_{k+1}, f_{\mathrm{D}}, t_{\mathrm{TB}}\right) &\approx \frac{N(\eta_k, f_{\mathrm{D}}) t_{\mathrm{TB}}}{p_k(\eta_k, \eta_{k+1})}, \quad 2 \leq k \leq K \end{aligned} \tag{2}$$

where $p_k(\eta_k, \eta_{k+1})$ is the marginal probability of the FSMC state k given by

$$p_k(\eta_k, \eta_{k+1}) = \int_{\eta_k}^{\eta_{k+1}} 2\xi e^{-\xi^2} d\xi \tag{3}$$

with $f(\xi) = 2\xi e^{-\xi^2}$, is the marginal distribution of $|h(t)|$; and $N(\eta_k, f_{\mathrm{D}})$ is the average number of times per second the signal envelope process $|h(t)|$ crosses level η_k—under the Bessel autocorrelation function model—given by [32]

$$N(\eta_k, f_{\mathrm{D}}) = \sqrt{2\pi}\eta_k f_{\mathrm{D}} e^{-\eta_k^2}. \tag{4}$$

We note from (2), and (4) that the state transition probabilities *linearly* increase with t_{TB}, and f_{D}. As a result, for fixed values of $\boldsymbol{\eta}$ and f_{D}, t_{TB} is bounded above due to the constraint $P_{k,k+1}\left(\eta_k, \eta_{k+1}, f_{\mathrm{D}}, t_{\mathrm{TB}}\right) + P_{k,k-1}\left(\eta_k, \eta_{k+1}, f_{\mathrm{D}}, t_{\mathrm{TB}}\right) \leq 1$:

$$t_{\mathrm{TB}} \leq \frac{p_k(\eta_k, \eta_{k+1})}{N(\eta_k, f_{\mathrm{D}}) + N(\eta_{k+1}, f_{\mathrm{D}})}, \quad \text{for all } 1 \leq k \leq K \tag{5}$$

where on the right hand side of the inequality is *the average duration* of state k [11]. The constraint in (5) implies that the TB duration cannot be greater than the average duration of a state. The choice of t_{TB} also uniquely specifies the TB size $N_{\mathrm{TB}}(B, t_{\mathrm{TB}}) = B t_{\mathrm{TB}}$ (expressed in symbols) where $t_{\mathrm{TB}} = j/B, j \in \mathbb{Z}^+$. We follow the equal duration channel partitioning convention where all states have the same average duration [11]. With this partitioning the ratio of the average time spend in each state to t_{TB}, denoted as $c(f_{\mathrm{D}}, t_{\mathrm{TB}}, K)$, is chosen between 3 and 8 for an accurate representation of the underlying channel [11].

In the modeling of the channel input-output relationship within each FSMC state we consider two approaches: the BSC, and the AWGN channel. With the BSC approach the channel

in each FSMC state is modeled as a BSC with a fixed *crossover* probability. The resulting FSMC is an appropriate model for wireless communication systems utilizing a *hard-decision* demodulator—that outputs a sequence of binary values—together with a binary channel encoder-decoder pair, and it is referred to as the hard-decision TB-based FSMC (HD-TB-based FSMC). With the AWGN channel approach the channel in each state is modeled as AWGN with a fixed SNR. The resulting FSMC is an appropriate model for wireless communication systems utilizing a *soft-decision* demodulator—that outputs *unquantized* matched filter samples, and it is referred to as the soft-decision TB-based FSMC (SD-TB-based FSMC). The primary difference between the two channel models is the type of demodulator used at the receiver; the channel state evolves in time according to the same Markov process with both channel models. Whenever the discussion is valid for both models we will refer to the channel model by the TB-based FSMC.

With the HD-TB-based FSMC the crossover probability assigned to a state heavily depends on both the *fixed* transmit power P, and the *fixed* modulation scheme, denoted by $\mathcal{M}$. With the SD-TB-based FSMC the SNR value assigned to a state depends only on P; however, the channel input is drawn from the signal constellation of the modulation $\mathcal{M}$. Here we study communication strategies where the CSI is available at the receiver, but is *not* fed back to the transmitter. Consequently, the modulation scheme and the transmit power are fixed since the transmitter cannot arbitrarily vary either one without the CSI. We emphasize that in general channel feedback does not usually exist in URLLC, and open-loop communication strategies are defined to enable flexible operation.

The state BSC crossover probabilities in the HD-TB-based FSMC, denoted by $\delta_k\left(\eta_k,\eta_{k+1},\mathcal{M},\frac{P}{BN_0}\right)$, $1\le k\le K$, where $\frac{P}{BN_0}$ is the average received SNR, are modeled by the state average bit error probabilities. Let $P_b\left(\mathcal{M},\frac{|h(t)|^2\ P}{BN_0}\right)$ denote the bit error probability with the modulation $\mathcal{M}$ as a function of the instantaneous received SNR $\frac{|h(t)|^2\ P}{BN_0}$. Then for each k we have

$$\delta_k\left(\eta_k,\eta_{k+1},\mathcal{M},\frac{P}{BN_0}\right)=\frac{\int_{\eta_k}^{\eta_{k+1}}P_b\left(\mathcal{M},\frac{\xi^2P}{BN_0}\right)2\xi e^{-\xi^2}d\xi}{p_k(\eta_k,\eta_{k+1})}. \tag{6}$$

With the HD-TB-based FSMC all bits belonging to a TB go through the same BSC. In general, $P_b\left(\mathcal{M},\frac{|h(t)|^2\ P}{BN_0}\right)$ increases with the number of bits per symbol (i.e. modulation order) of modulation $\mathcal{M}$. It follows that there is a trade-off between the BSC qualities of the states and the number of BSC uses per TB. As the modulation order increases both the BSC crossover probabilities assigned to states and the TB size (bits) increase. The state SNR values in the SD-TB-based FSMC, denoted by $\gamma_k\left(\eta_k,\eta_{k+1},\frac{P}{BN_0}\right)$, $1\le k\le K$, are modeled by the state average SNR values:

$$\gamma_k\left(\eta_k,\eta_{k+1},\frac{P}{BN_0}\right)=\frac{P}{BN_0}\frac{\int_{\eta_k}^{\eta_{k+1}}\left(\xi^2\right)2\xi e^{-\xi^2}d\xi}{p_k(\eta_k,\eta_{k+1})}, \tag{7}$$

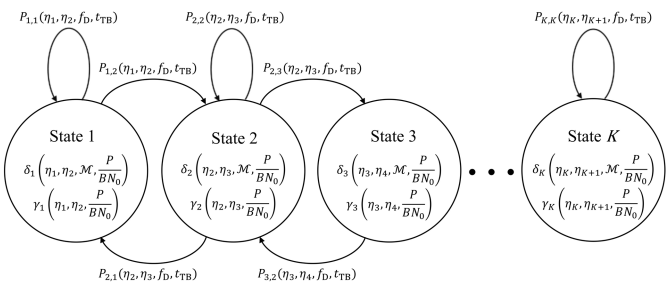


Fig. 1. The TB-based FSMC model of the Rayleigh wireless fading channel.

which can be analytically evaluated. With the SD-TB-based FSMC all symbols belonging to a TB go through the same AWGN channel whose inputs are drawn from the signal constellation of $\mathcal{M}$. The TB-based FSMC model of the Rayleigh fading channel is shown in Fig. 1. Within each state the BSC crossover probability and the AWGN SNR value associated with the state are shown.

III. HARQ System Model

In this section we describe two HARQ schemes, where the underlying Rayleigh fading channel is suitable to be modeled by the HD-TB-based FSMC and the SD-TB-based FSMC, respectively. The CSI is assumed to be known only at the receiver; both the modulation scheme and the coding rate are fixed. To keep the notation compact, dependence on the wireless parameters is not stated, e.g. the TB size (symbols) is denoted by N_{TB}, not by $N_{\mathrm{TB}}(B,t_{\mathrm{TB}})$. The focus is placed on 2^{2m}-ary QAM with $m\in\mathbb{Z}^+$ (i.e. *square* QAM). However, the proposed framework is general and can be applied to any linear modulation scheme. A 2^{2m}-ary QAM signal with transmit power P is treated as two 2^m-ary PAM signals with transmit power $P/2$ each. The symbol $\mathcal{M}$ denotes the set of PAM signal constellation points where $|\mathcal{M}|=2^m$, and the average Euclidian distance from the origin is $\frac{P}{2B}$.

We first describe a HARQ system employing a binary channel encoder, a hard-decision demodulator, and a binary channel decoder. The wireless channel—together with the modulator and the hard-decision demodulator—is suitable to be modeled by the HD-TB-based FSMC. On the transmitter side, a *binary encoder* denoted by f maps one of M *equiprobable* messages to a length-$2mLN_{\mathrm{TB}}$ binary sequence, i.e. $f:\{1,\cdots,M\}\to\{0,1\}^{2mLN_{\mathrm{TB}}}$, where $\{0,1\}^{2mLN_{\mathrm{TB}}}$ denotes the $2mLN_{\mathrm{TB}}$-fold Cartesian product of $\{0,1\}$, and L is *the number of HARQ rounds* (i.e. one TB transmission per HARQ round). The code size M satisfies[1] $\log_2 M\le 2mLN_{\mathrm{TB}}$ and $\log_2 M\in\mathbb{Z}^+$. The encoder input is denoted by the random variable W. For each TB transmission the modulator breaks $2mN_{\mathrm{TB}}$ bits into $2N_{\mathrm{TB}}$ m-bit subsequences, which are then mapped onto signals in the one-dimensional signal constellation $\mathcal{M}$ where each constellation point is labeled with an m-bit binary sequence. On the receiver side, there

[1]We don't restrict the number of information bits to be less than the TB size because as we will later see in some cases it is beneficial select a code size for which the probability of error at the first j HARQ rounds is approximately 1. In that case the *effective number of HARQ rounds* is given by $L-j$.

is a hard-decision demodulator followed by a binary decoder whose output at the end of the l-th HARQ round is denoted by $\widehat{W}_l$. The decoding at the l-th HARQ round is defined by the mapping $g_l : \{0,1\}^{2mlN_{\rm TB}} \times \mathcal{S}^l \to \{1,\cdots,M\}$, where $\mathcal{S}$ is the state space of the FSMC. The channel input-output relationship is characterized by $P_{\mathbf{Y}_1^L|\mathbf{X}_1^L S_1^L}$ where $\mathbf{X}_1^L$ (the channel input) and $\mathbf{Y}_1^L$ (the channel output) are sequences of length-$2mN_{\rm TB}$ binary vectors, i.e. $\mathbf{X}_1^L = \mathbf{X}_1,\cdots,\mathbf{X}_L$, $\mathbf{Y}_1^L = \mathbf{Y}_1,\cdots,\mathbf{Y}_L$, $\mathbf{X}_l = X_{l,1},\cdots,X_{l,2mN_{\rm TB}}$, $\mathbf{Y}_l = Y_{l,1},\cdots,Y_{l,2mN_{\rm TB}}$, $x_{l,i}, y_{l,i} \in \{0,1\}$, and $s_1^l \in \mathcal{S}^l$. We denote random variables by upper case symbols, and realizations of random variables by lower case symbols. The distribution $P_{\mathbf{Y}_1^L|\mathbf{X}_1^L S_1^L}$ can be written as a product distribution by $P_{\mathbf{Y}_1^L|\mathbf{X}_1^L S_1^L} = \prod_{l=1}^{L}\prod_{i=1}^{2mN_{\rm TB}} P_{Y_{l,i}|X_{l,i}S_l}$ where

$$P_{Y_{l,i}|X_{l,i}S_l}(y_{l,i}|x_{l,i}s_l) = \begin{cases} \delta_{s_l} & \text{if } y_{l,i} \neq x_{l,i} \\ 1-\delta_{s_l} & \text{otherwise .} \end{cases} \tag{8}$$

The second HARQ system under consideration employs a symbol encoder, a soft-decision demodulator, and a symbol decoder operating on unquantized matched filter samples. With this system the wireless channel is suitable to be modeled by the SD-TB-based FSMC. The encoder f maps one of M *equiprobable* messages to a length-$2LN_{\rm TB}$ sequence of 2^m-ary PAM signals, i.e. $f : \{1,\cdots,M\} \to \mathcal{M}^{2LN_{\rm TB}}$. The code size M satisfies the conditions described previously, and the encoder input is denoted by W. The output of the decoder at the l-th HARQ round, defined by the mapping $g_l : \mathbb{R}^{2lN_{\rm TB}} \times \mathcal{S}^l \to \{1,\cdots,M\}$, is denoted by the $\widehat{W}_l$. The channel input-output relationship is characterized by $p_{\mathbf{Y}_1^L|\mathbf{X}_1^L S_1^L}$ where $\mathbf{X}_l = X_{l,1},\cdots,X_{l,2N_{\rm TB}}$, $\mathbf{Y}_l = Y_{l,1},\cdots,Y_{l,2N_{\rm TB}}$, $y_{l,i} \in \mathbb{R}$, and $x_{l,i} \in \mathcal{M}$. The joint probability density expression $p_{\mathbf{Y}_1^L|\mathbf{X}_1^L S_1^L}$ can be written as a product by $p_{\mathbf{Y}_1^L|\mathbf{X}_1^L S_1^L} = \prod_{l=1}^{L}\prod_{i=1}^{2N_{\rm TB}} p_{Y_{l,i}|X_{l,i}S_l}$ where

$$p_{Y_{l,i}|X_{l,i}S_l}(y_{l,i}|x_{l,i}s_l) = \frac{1}{\sqrt{\pi N_0}} e^{-\frac{\left(y_{l,i}-\sqrt{\gamma'_{s_l}}x_{l,i}\right)^2}{N_0}} \tag{9}$$

and $\gamma'_{s_l} = \frac{\gamma_{s_l}}{P/(BN_0)}$ is the average channel gain in state s_l.

A. Markov Modeling of HARQ Over the TB-Based FSMC

In this section, we describe the general operation of HARQ with either system. With both systems each codeword consists of L TBs. At round $l, l = 1,\cdots,L-1$, the decoder is fed the first l (received) TBs $\mathbf{y}_1^l$ and the channel state sequence s_1^l. If the decoder output is correct, i.e. $\widehat{w}_l = w$, the receiver sends the transmitter an acknowledge (ACK) message. Otherwise, a not acknowledge (NACK) message is sent. We assume *instantaneous* ACK/NACK feedback and *perfect error detection* capability at the receiver. If the transmitter receives a NACK, it sends the next TB $\mathbf{x}_{l+1}$; if it receives a NACK the rest of the codeword $\mathbf{x}_{l+1}^L$ is discarded and transmission of a new message starts. For a given channel state sequence s_1^l the transmission of a message is terminated at round l with probability $\Pr\{\widehat{W}_l = w|\widehat{W}_{l-1} \neq w,\cdots,\widehat{W}_1 \neq w, S_1^l = s_1^l\}$. In the case of a decoding error at round L the transmitter carries on with a new message; a retransmission does not take place and the message is dropped.

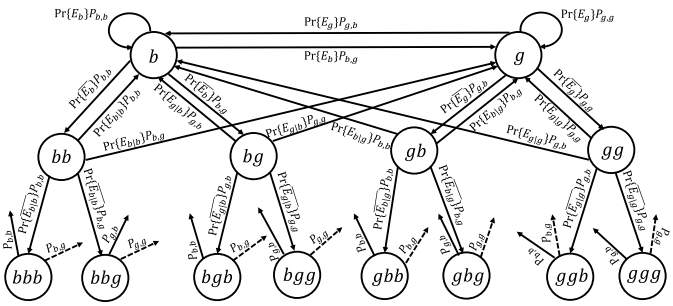


Fig. 2. The HARQ-MM with 3 HARQ rounds and 2 FSMC states.

The HARQ round evolves as a finite-state process in time where a ACK returns the system to round 1 and a NACK moves it to the next round. However, this process does not have the Markov property because the state transition probabilities do not depend only on the HARQ round, but also depend on the channel state sequence. By defining the channel state sequence $s_1^l = s_1 s_2 \cdots s_l \in \mathcal{S}^l$ as the HARQ system state, we form a discrete-time finite-state Markov chain, referred to as the HARQ Markov model (HARQ-MM), that tracks the operation of HARQ over the TB-based FSMC. The HARQ-MM states at round $l, 1 \leq l \leq L$, are labeled by length-l strings of the form $s_1^l = s_1 s_2 \cdots s_l \in \mathcal{S}^l$. Since the FSMC state can transition to at least 2 and at most 3 different states for given K and L the number of HARQ-MM states, denoted as $N_{\rm sts}$, can be bounded by $K(2^L-1) < N_{\rm sts} < \frac{K(3^L-1)}{2}$.

The state diagram of a HARQ-MM with 3 rounds and 2 FSMC states is shown in Fig. 2. We refer to the TB-based FSMC state 1 as the BAD state by b, and the state 2 as the GOOD state by g, and label the HARQ-MM states accordingly. For brevity in Fig. 2 we denote $\Pr\{\widehat{W}_l = W|\cdots,\widehat{W}_1 \neq W, S_1^l = s_1^l\}$ by $\Pr\{E_{s_l|\cdots s_1}\}$, and $\Pr\{\widehat{W}_l \neq W|\cdots,\widehat{W}_1 \neq W, S_1^l = s_1^l\}$ by $\Pr\{\overline{E_{s_l|\cdots s_1}}\}$. In this tree-like state diagram round 1 is at the top level (or level 1), and the state s_1^l can only be transitioned to from its parent state s_1^{l-1}. The arrows emerging out of the states at level 3 are meant to connect to the states at level 1: the dashed arrows to state g, and the solid arrows to state b. Note that the HARQ-MM is ergodic, and hence a unique steady-state distribution can be computed. The HARQ-MM steady-state distribution contains *all information* regarding the system performance as will be seen shortly.

B. HARQ Performance Measures

HARQ is not a fixed rate communication scheme since the transmission of a message may terminate at varying times. For this reason, we will use the *HARQ average rate* denoted by $R_{\rm HARQ}$ as the system performance measure. Let $L_{\rm HARQ}, l_{\rm HARQ} \in \{1,\cdots,L\}$ represent the HARQ round at which the transmission for a message ends. Then the average number of HARQ rounds per message is given by $\mathbb{E}[L_{\rm HARQ}]$. Accordingly, we define the HARQ average rate by

$$R_{\rm HARQ} \triangleq \frac{\log_2 M}{\mathbb{E}[L_{\rm HARQ}]N_{\rm TB}} \text{ bits/channel use (symbol).} \tag{10}$$

The expected value $\mathbb{E}[L_{\text{HARQ}}]$ can be computed from the HARQ-MM steady-state distribution by

$$\mathbb{E}[L_{\text{HARQ}}] = \frac{1}{\sum_{s_1 \in \mathcal{S}} p_{s_1}} \tag{11}$$

where the denominator is the sum of the steady-state probabilities of the HARQ-MM states at round 1. The steady-state probability of the state $s_1^l = s_1 \cdots s_l$ is denoted by $p_{s_1^l} = p_{s_1 \cdots s_l}$. The proof of (11) is given in Appendix. The probabilities $p_{s_1}, s_1 = 1, \cdots, K$, in (11) represent the distribution induced on the TB-based FSMC states *in the beginning of* a HARQ transmission and are not related to steady-state distribution of the TB-based FSMC. We define *the HARQ rate per TB per complex dimension* (effective HARQ transmission rate per complex dimension at round 1) as a function of the code size M by

$$R \triangleq \frac{\log_2 M}{2N_{\text{TB}}}. \tag{12}$$

With this definition of R, and from (11), the HARQ average rate defined in (10) can be expressed by

$$R_{\text{HARQ}} = 2R \left(\sum_{s_1 \in \mathcal{S}} p_{s_1} \right), \tag{13}$$

which is simply the HARQ rate per TB scaled by the probability of starting a new HARQ transmission. Since messages are dropped only at the L-th HARQ round in the case of decoding errors, *the HARQ average probability of error*, denoted by ϵ_{HARQ}, is defined as

$$\epsilon_{\text{HARQ}} \triangleq \sum_{s_1^L \in \mathcal{S}^L} p_{s_1^L} \Pr\{\widehat{W}_L \neq W | \ldots, \widehat{W}_1 \neq W,\ S_1^L = s_1^L\} \tag{14}$$

where the sum is over all HARQ-MM states at round L. Please note that the HARQ average rate can also be defined by only taking into account successfully decoded messages as done in [33]. In that case the average rate expressions in (10) and (13) would be scaled by $(1 - \epsilon_{\text{HARQ}})$. As discussed shortly, in this manuscript the probability of error ϵ_{HARQ} is constrained as $\epsilon_{\text{HARQ}} \leq \epsilon$ by some ϵ, which translates to the delay-violation probability. Furthermore, as we only consider probability of error constraints $\epsilon \leq 10^{-2}$ the difference between the two definitions has negligible impact of the system performance.

C. HARQ Coding Delay and Delay Violation Probability

As discussed in Section I, in this paper, we focus on coding latency instead of the overall system latency. The complete treatment of the overall system latency and reliability can be regarded as a future extension of this paper. The coding latency experienced by a message with HARQ has three components: the encoding delay, the decoding delay, and the transmission delay. The transmission delay for a message is a random variable given by the size of the blocklength (e.g. TB size) and the HARQ round at which the transmission ends. For example, for a simple MAC where consecutive retransmissions happen without additional delay, the overall transmission delay can be expressed as $L_{\text{HARQ}} \times t_{\text{TB}}$ (seconds). The encoding and the decoding delays are constrained to be smaller than the TB duration so that the transmitter is able to send a new message by the time the transmission of the current TB ends, and the receiver is ready to decode by the time a new TB is received. We also assume that the encoding and decoding delays do not significantly vary with blocklength on a time scale comparable to t_{TB}. Under this assumption a constraint on the coding latency is equivalent to a constraint on the transmission delay.

We define the delay-violation probability as the probability that the coding delay for a message exceeds a given threshold (seconds). Note that delay-violation probability is a key performance indicator in 5G URLLC. For example, a general URLLC reliability requirement for one transmission of a packet is $1 - 10^{-5}$ for 32 bytes with a user plane latency of 1 ms [4]. With this definition and by noting that a message is dropped if it is not successfully decoded by the last HARQ rounds L, a probability of error value of ϵ_{HARQ} for a L-round HARQ with TB duration t_{TB} is equivalent to the coding delay-violation probability of ϵ_{HARQ} for a delay threshold of $L \times t_{\text{TB}}$ (seconds). It follows that the maximum achievable rate of a HARQ system with L rounds under the probability of error constraint $\epsilon_{\text{HARQ}} < \epsilon$ is equivalent to the maximum achievable rate under the constraint that a message is successfully transmitted over the wireless channel in no more than $L \times t_{\text{TB}}$ seconds with a probability of at least $1 - \epsilon$. As such, the HARQ throughput analysis performed next directly translates to an analysis on the HARQ throughput under coding latency constraints. Mapping to the example shown in 5G URLLC [4], we have $L \times t_{\text{TB}} \leq 10^{-3}$ and $\epsilon_{\text{HARQ}} \leq 10^{-5}$.

IV. HARQ System Performance

In this section, we introduce a framework to closely approximate the maximum achievable HARQ (average) rate under the probability of error constraint ϵ, denoted as $R^*_{\text{HARQ}}(\epsilon)$. We assume that the MCS is fixed throughout the transmission. As such, the transmitter does not require additional feedback for adapting its MCS. Note that additional CSI feedback will usually incur delay and overhead which is not desirable for 5G URLLC. On the other hand, it is important to note that the introduced framework can be extended to the case where the transmitter can adapt its MCS based on CSI feedback. This can be treated in a future extension of this paper.

Note that for codes of length $2LN_{\text{TB}}$ (PAM symbols) and rate R any set of achievable values of $\Pr\{\widehat{W}_l \neq W | \cdots, \widehat{W}_1 \neq W, S_1^l = s_1^l\}, 1 \leq l \leq L, s_1^l \in \mathcal{S}^l$, leads to an achievable HARQ average rate computed by (13). Second, given two codes of length $2LN_{\text{TB}}$ and rate R if one has lower probability of error values at *all* HARQ-MM states, then it also has a *higher* HARQ average rate, and a *lower* HARQ probability of error. By assuming that for given $2LN_{\text{TB}}$ and R there exists a (best) code with lower probability of error values than all other codes at all HARQ-MM states the HARQ maximum achievable rate problem is formulated as follows. Let $\{\epsilon_{s_1 \cdots s_l}(R)\}_{s_1^l \in \mathcal{S}^l, 1 \leq l \leq L}$ denote the set of the minimum achievable probability of error values by a code of length

$2LN_{\mathrm{TB}}$ and rate R. From (13) and (14) the HARQ average rate and the HARQ average probability of error of the best code of rate R denoted by $R_{\mathrm{HARQ}}(R)$, and $\epsilon_{\mathrm{HARQ}}(R)$ are given by

$$\begin{aligned} R_{\mathrm{HARQ}}(R) &= 2R\left(\sum_{s_1\in\mathcal{S}} p_{s_1}(R)\right) \\ \epsilon_{\mathrm{HARQ}}(R) &= \sum_{s_1^L\in\mathcal{S}^L} p_{s_1^L}(R)\epsilon_{s_1^L}(R), \end{aligned} \tag{15}$$

where $p_{s_1^l}(R)$ denotes the steady-state probability of state s_1^l computed from $\{\epsilon_{s_1\cdots s_l}(R)\}_{s_1^l\in\mathcal{S}^l,1\le l\le L}$. Then $R^*_{\mathrm{HARQ}}(\epsilon)$ is approximated by the expression

$$R^*_{\mathrm{HARQ}}(\epsilon) \approx \max_{R\in\mathcal{R}}\{R_{\mathrm{HARQ}}(R)\} \quad \text{s.t. } \epsilon_{\mathrm{HARQ}}(R)\le\epsilon \tag{16}$$

where $\mathcal{R} = \left\{\frac{1}{2N_{\mathrm{TB}}}, \frac{2}{2N_{\mathrm{TB}}}, \cdots, \frac{2mLN_{\mathrm{TB}}-1}{2N_{\mathrm{TB}}}, Lm\right\}$ is from (12), $\log_2 M \in \mathbb{Z}^+$, and $\log_2 M \le 2mLN_{\mathrm{TB}}$. Since $\mathcal{R}$ is a *finite* set, (16) can be solved by evaluating $R_{\mathrm{HARQ}}(R)$, and $\epsilon_{\mathrm{HARQ}}(R)$ for all $R\in\mathcal{R}$, and choosing $R^*_{\mathrm{HARQ}}(\epsilon)$ as the maximum $R_{\mathrm{HARQ}}(R)$ from the feasible values of R (those that satisfy $\epsilon_{\mathrm{HARQ}}(R)\le\epsilon$). We emphasize that $\epsilon_{\mathrm{HARQ}}(R)$ is an increasing function of R; however, $R_{\mathrm{HARQ}}(R)$ is NOT an increasing function of R because the steady-state probabilities $p_{s_1}(R)$ in (15) are decreasing with R. It follows that $R^*_{\mathrm{HARQ}}(\epsilon)$, for a given ϵ, is not necesssarily achieved by the maximum of the set of feasible values of R. This aspect of the HARQ average rate implies that increasing the probability of error constraint ϵ does not necessarily increase the maximum achievable rate $R^*_{\mathrm{HARQ}}(\epsilon)$. In contrast, in finite blocklength wireless systems without feedback (i.e. without ACK/NACK or CSI feedback) the maximum achievable rate is increasing with the probability of error constraint [12], [16].

Note that the value of $R^*_{\mathrm{HARQ}}(\epsilon)$ depends on the design parameters such as the modulation scheme, the transmit power, and the number of HARQ rounds, as well as the physical parameters such as the Doppler frequency, and the noise p.s.d. It follows that the proposed framework can be used to approximate the HARQ performance of different communication strategies. For example, $R^*_{\mathrm{HARQ}}(\epsilon)$ can be evaluated for a set of modulation schemes—with all other parameter values fixed—to determine the modulation with the highest maximum achievable rate. Next, we discuss how to approximate the minimum achievable probability of error values $\{\epsilon_{s_1\cdots s_l}(R)\}_{s_1^l\in\mathcal{S}^l,1\le l\le L}$.

A. Minimum Achievable Probability of Error Values

In approximating $\{\epsilon_{s_1\cdots s_l}(R)\}_{s_1^l\in\mathcal{S}^l,1\le l\le L}$ we make two assumptions. First, we assume that the probability of error conditioned on decoding errors at previous HARQ rounds given by $\Pr\{\widehat{W}_l\neq W|\cdots,\widehat{W}_1\neq W, S_1^l=s_1^l\}$ can be tightly approximated by the unconditional probability of error $\Pr\{\widehat{W}_l\neq W|S_1^l=s_1^l\}$. We state without a proof that this approximation is optimistic:

$$\begin{aligned} &\Pr\{\widehat{W}_l\neq W|S_1^l=s_1^l\} \\ &\qquad\le \Pr\{\widehat{W}_l\neq W|\ldots,\widehat{W}_1\neq W,\ S_1^l=s_1^l\}. \end{aligned} \tag{17}$$

This is because decoding errors in previous HARQ rounds are the result of unfavorable noise sequence realizations, which would increase the probability of decoding error at the current HARQ round. However, we emphasize that this approximation is used only for HARQ rounds greater than or equal to 2. In addition, whenever $\Pr\{\widehat{W}_{l-1}\neq W,\cdots,\widehat{W}_1\neq W|S_1^{l-1}=s_1^{l-1}\}=1$ the inequality in (17) becomes an equality. On the other hand, whenever $\Pr\{\widehat{W}_l\neq W|\cdots,\widehat{W}_1\neq W,S_1^l=s_1^l\}\approx 0$ holds, $\Pr\{\widehat{W}_l\neq W|S_1^l=s_1^l\}\approx 0$ also holds from (17). Furthermore, the HARQ-MM states descending from the state s_1^l have approximately 0 steady-state probabilities as the HARQ-MM visits these states only when an error occurs at state s_1^l. Given the disparity among the channel qualities of the FSMC states (i.e. the BSC crossover probabilities, the AWGN SNR values), we expect both cases to frequently occur. For example, with 2 FSMC states and practical average SNR values for a range of rate values between the capacities of the BAD state and the GOOD state we have $\Pr\{\widehat{W}_1\neq W|S_1=b\}\approx 1$, and $\Pr\{\widehat{W}_1\neq W|S_1=g\}\approx 0$.

Second, we assume that for all $s_1^l\in\mathcal{S}^l, 1\le l\le L$, the minimum achievable probability of error for a code of length $2mlN_{\mathrm{TB}}$ (PAM symbols), and size $M=2^{2RN_{\mathrm{TB}}}$ for the state sequence s_1^l, is also achievable by some (mother) code of length $2mLN_{\mathrm{TB}}$, and size $M=2^{2RN_{\mathrm{TB}}}$ when the first l TBs are decoded at the l-th HARQ round with the same state sequence s_1^l. In other words, we assume that there exists a mother code such that each subcode of the mother code is as good as the best code of the same length. This assumption is also optimistic. It follows that $\epsilon_{s_1^l}(R)$ can be approximated separately for each HARQ-MM state by information theoretic results on the minimum achievable probability of error of fixed blocklength ($2mlN_{\mathrm{TB}}$) codes of given size (M).[2]

1) Hard-Decision TB-Based FSMC: In the case of the HD-TB-based FSMC we can use the results given in [12] regarding parallel DMCs. We can model a HD-TB-based FSMC with a fixed channel state sequence realization s_1^l as $2mN_{\mathrm{TB}}$ uses of a set of l parallel BSCs with crossover probabilities $\delta_{s_1},\cdots,\delta_{s_l}$, respectively. Then $\epsilon_{s_1^l}(R)$ can be closely approximated by

$$\epsilon_{s_1^l}(R) \approx Q\left(\frac{mC_{\mathrm{BSC}}(s_1^l)-R}{\sqrt{mV_{\mathrm{BSC}}(s_1^l)/2N_{\mathrm{TB}}}}\right) \tag{18}$$

where the right hand side of the inequality is—for $2mN_{\mathrm{TB}}$ in the order of hundreds—a tight approximation to the minimum achievable probability of error for a code of length $2mlN_{\mathrm{TB}}$ (bits) and size $M=2^{2RN_{\mathrm{TB}}}$ over l parallel BSCs—each used

[2]We have only considered deterministic codes, and we assumed that there exists a best code, that the conditional error probabilities can be closely approximated by the unconditional error probabilities, and that each subcode of the mother code is as good as the best code of the same length. Alternatively, we can use a random code where codewords are generated randomly according to some distribution $P_{\mathbf{X}_1^L}$ on the channel input sequences. This approach eases the analysis of the HARQ coding performance. In this case there is only one code under consideration and the "best code" assumption is superfluous. In addition, approximating the conditional error probabilities by unconditional error probabilities is equivalent to generating (and transmitting) a new codeword at the next HARQ round in the case of a decoding error at the current HARQ round. Then $\Pr\{\widehat{W}_l\neq W|S_1^l=s_1^l\}, 1\le l\le L, s_1^l\in\mathcal{S}^l$ can be computed separately for each subcode, and hence there is no need for the third assumption to be made with deterministic codes.

$2mN_{\rm TB}$ times—with crossover probabilities $\delta_{s_1},\cdots,\delta_{s_l}$ [12], $Q(x)=\int_x^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$, and $C_{\rm BSC}(s_1^l)$ and $V_{\rm BSC}(s_1^l)$ are the capacity and the dispersion associated with the set of parallel BSCs. The motivation for the approximation in (18) is the asymptotic expansion proved for a number of channels in [12]:

$$\log_2 M^*(n,\epsilon)=nC-\sqrt{nV}Q^{-1}(\epsilon)+O(\log_2 n), \quad (19)$$

where $M^*(n,\epsilon)$ is the maximum cardinality of a codebook of blocklength n for an average probability of error ϵ, over a channel with capacity C and dispersion V. The terms $C_{\rm BSC}(s_1^l)$, and $V_{\rm BSC}(s_1^l)$ are given by [12]

$$\begin{aligned} C_{\rm BSC}(s_1^l)&=\sum_{i=1}^{l} C_{\rm BSC}(\delta_{s_i})\\ V_{\rm BSC}(s_1^l)&=\sum_{i=1}^{l} V_{\rm BSC}(\delta_{s_i}) \end{aligned} \quad (20)$$

where $C_{\rm BSC}(\delta)$, and $V_{\rm BSC}(\delta)$ the capacity, and the dispersion of the BSC with crossover probability δ given by

$$\begin{aligned} C_{\rm BSC}(\delta)&=1+\delta\log_2\delta+(1-\delta)\log_2(1-\delta)\\ V_{\rm BSC}(\delta)&=\delta(1-\delta)\left(\log_2\frac{1-\delta}{\delta}\right)^2. \end{aligned} \quad (21)$$

We emphasize that $\{C_{\rm BSC}(\delta_k)\}_{1\le k\le K}$, and $\{V_{\rm BSC}(\delta_k)\}_{1\le k\le K}$, are functions of the particular PAM modulation denoted by $\mathcal{M}$ and the average received SNR since $\{\delta_k\}_{1\le k\le K}$ are functions of $\mathcal{M}$, and the average received SNR (see (6)). In addition, $\epsilon_{s_1^l}(R)$ is an increasing function of R for all HARQ-MM states as can seen in (18).

2) Soft-Decision TB-Based FSMC: In the case of the SD-TB-based FSMC, we develop information theoretic results regarding parallel AWGN channels with discrete inputs alphabets, PAM in particular. We can model a SD-TB-based FSMC with a fixed channel state sequence realization s_1^l as $2N_{\rm TB}$ uses of a set of l parallel AWGNs with PAM input signals and SNR values $\gamma_{s_1},\cdots,\gamma_{s_l}$ respectively. Then, $\epsilon_{s_1^l}(R)$ can be closely approximated by the following result.

Theorem 1: Consider l parallel AWGN channels, each with noise p.s.d. $\frac{N_0}{2}$ and 2^m-ary PAM input signals, the i-th channel, $1\le i\le l$, with inputs drawn from $\mathcal{M}_i$. Then the maximum cardinality of a codebook of blocklength nl (PAM symbols) and average probability of error ϵ over this channel denoted by $\log_2 M^*(n,\epsilon,\gamma_1^l)$ satisfies

$$\begin{aligned} &\log_2 M^*(n,\epsilon,\gamma_1^l)\\ &\quad\ge nC_{\rm PAM}(\gamma_1^l)-\sqrt{nV_{\rm PAM}(\gamma_1^l)}Q^{-1}(\epsilon)+O(1), \end{aligned} \quad (22)$$

where $\gamma_i=2^{-m}\sum_{x_i\in\mathcal{M}_i}\frac{x_i^2}{N_0}$ is the SNR of the i-th channel, $C_{\rm PAM}(\gamma_1^l)$, and $V_{\rm PAM}(\gamma_1^l)$ are the capacity and the dispersion under the equiprobable channel input constraint, given by

$$\begin{aligned} C_{\rm PAM}(\gamma_1^l)&=\sum_{i=1}^{l} C_{\rm PAM}(\gamma_i)\\ V_{\rm PAM}(\gamma_1^l)&=\sum_{i=1}^{l} V_{\rm PAM}(\gamma_i), \end{aligned} \quad (23)$$

and $C_{\rm PAM}(\gamma_i)$, and $V_{\rm PAM}(\gamma_i)$ are the capacity and the dispersion of the AWGN channel with SNR γ_i and PAM input signals under the equiprobable input constraint. The quantities $C_{\rm PAM}(\gamma_i)$, and $V_{\rm PAM}(\gamma_i)$ are numerically computed by

$$\begin{aligned} C_{\rm PAM}(\gamma_i)&=\mathbb{E}\left[\log_2\frac{p_{Y_i|X_i}(Y_i|X_i)}{2^{-m}\sum_{z\in\mathcal{M}_i}p_{Y_i|X_i}(Y_i|z)}\right]\\ V_{\rm PAM}(\gamma_i)&={\rm Var}\left[\log_2\frac{p_{Y_i|X_i}(Y_i|X_i)}{2^{-m}\sum_{z\in\mathcal{M}_i}p_{Y_i|X_i}(Y_i|z)}\right] \end{aligned} \quad (24)$$

where the expected values are taken according to $p_{Y_i|X_i}P_{X_i}$ with $P_{X_i}(x_i)=2^{-m}$ for all $x_i\in\mathcal{M}_i$, and $p_{Y_i|X_i}(y_i|x_i)$ is Gaussian distributed with variance $\frac{N_0}{2}$ and mean x_i. In the case of BPSK, i.e. $m=1$, we also have

$$\begin{aligned} &\log_2 M^*(n,\epsilon,\gamma_1^l)\le nC_{\rm BPSK}(\gamma_1^l)\\ &\qquad-\sqrt{nV_{\rm BPSK}(\gamma_1^l)}Q^{-1}(\epsilon)+\frac{1}{2}\log_2(n)+O(1) \end{aligned} \quad (25)$$

which together with (22) implies

$$\begin{aligned} &\log_2 M^*(n,\epsilon,\gamma_1^l)\\ &=nC_{\rm BPSK}(\gamma_1^l)-\sqrt{nV_{\rm BPSK}(\gamma_1^l)}Q^{-1}(\epsilon)+O(\log_2(n)). \end{aligned} \quad (26)$$

The proof of Theorem 1 is given in Appendix. Theorem 1 is a fundamental result whose implications extend beyond this paper as it can be used to characterize the throughput over wireless channels in the case of CSI at the transmitter as a function of the modulation scheme, code blocklength and probability of error constraint. By using Theorem 1 we can approximate $\epsilon_{s_1^l}(R)$ for square QAM modulation by

$$\epsilon_{s_1^l}(R)\approx Q\left(\frac{C_{\rm PAM}(s_1^l)-R}{\sqrt{V_{\rm PAM}(s_1^l)/2N_{\rm TB}}}\right) \quad (27)$$

where $C_{\rm PAM}(s_1^l)=C_{\rm PAM}(\frac{\gamma_{s_1}}{2}\cdots\frac{\gamma_{s_l}}{2})$ is computed from (23). The factor of 2 is from the fact that the transmit power P is shared by 2 PAM signals in 2 complex dimensions, with the same signal constellation in each dimension. Note that known (achievability) bounds (e.g. dependency testing bound [12]) can be used to approximate $\epsilon_{s_1^l}(R)$. Such bounds are computed by simulation in the case of discrete-input continuous-output channels like AWGN channels with PAM inputs. In addition, bounds must be computed for *all* HARQ-MM states individually, and must be recomputed whenever the TB size, the number of HARQ rounds or the rate R is varied. With the channel dispersion-based approach, both the capacity and the dispersion are computed for each FSMC state, rather than being computed for each HARQ-MM state as can seen in (23), and can used for any number of HARQ rounds, any TB size, and any coding rate R.

B. HARQ-MM Steady-State Distribution

In this section we describe an algorithm to efficiently compute the steady-state distribution of the HARQ-MM. The steady-state distribution of the HARQ-MM can be computed by solving the following system of linear equations

$$\mathbf{p}_{1\times N_{\rm sts}}\mathbf{P}_{N_{\rm sts}\times N_{\rm sts}}=\mathbf{p}_{1\times N_{\rm sts}},\quad \mathbf{p}_{1\times N_{\rm sts}}\mathbf{1}_{N_{\rm sts}\times 1}=1 \quad (28)$$

where N_{sts} is the number of HARQ-MM states; $\mathbf{p}_{1\times N_{\text{sts}}}$ is the vector of steady-state probabilities of the HARQ-MM states; $\mathbf{P}_{N_{\text{sts}}\times N_{\text{sts}}}$ is the corresponding state-transition matrix; and $\mathbf{1}_{N_{\text{sts}}\times 1}=[1,1,\cdots,1]^{\intercal}$. The vector $\mathbf{p}_{1\times N_{\text{sts}}}$ is the left eigenvector of $\mathbf{P}_{N_{\text{sts}}\times N_{\text{sts}}}$ of eigenvalue 1, normalized to satisfy the total probability constraint. As noted previously the number of HARQ-MM states of the left-right FSMC model shown in Fig. 1 is bounded by $K(2^L-1)<N_{\text{sts}}<\frac{K(3^L-1)}{2}$. For a large number of HARQ rounds, and FSMC states N_{sts} is very large and (28) is not computationally feasible to solve directly. However, the HARQ-MM has a special structure observed in Figure 2: the HARQ-MM state s_1^l can only be transitioned to from its parent state s_1^{l-1}. This property can be exploited to reduce the system of linear equations in (28) to one with a lower dimensionality as explained next.

Let $P_{k,j}$ denote the transition probability from the FSMC state k to the FSMC state j. Then the steady-state probability of HARQ-MM state $s_1^l=s_1\cdots s_l$ denoted by $p_{s_1^l}$ can be written in terms of the steady-state probability of HARQ-MM state s_1, denoted by p_{s_1}, by the expression

$$p_{s_1^l}=p_{s_1}\epsilon_{s_1}P_{s_1,s_2}\epsilon_{s_1^2}P_{s_2,s_3}\ldots\epsilon_{s_1^{l-1}}P_{s_{l-1},s_l} \tag{29}$$

where $\epsilon_{s_1^i}$ is the probability of decoding error at HARQ-MM state s_1^i. The equation can also be written recursively by

$$p_{s_1^l}=p_{s_1^{l-1}}\epsilon_{s_1^{l-1}}P_{s_{l-1},s_l}. \tag{30}$$

We observe from (29) and (30) that if the steady-state probabilities of the HARQ-MM states at round 1 are known, then the steady-state probabilities of all HARQ-MM states can be computed by multiplications. Accordingly, we focus on computing the steady-state probabilities of the HARQ-MM states at round 1, denoted by $\widehat{\mathbf{p}}_{1\times K}=[p_1\cdots p_K]$. We define the state-transition matrix for the HARQ-MM states at round 1 by $\widehat{\mathbf{P}}_{K\times K}$ whose (i,j)-th element denoted by $\widehat{\mathbf{P}}(i,j)$ is the probability of the next HARQ transmission starting at FSMC state j conditioned on the current HARQ transmission starting at FSMC state i. Note that at HARQ round 1 the HARQ-MM state is the same as the FSMC state. We observe that the state in which a HARQ transmission starts evolves as a K-state discrete-time Markov chain. The quantity $\widehat{\mathbf{P}}(i,j)$ can be computed as the sum of the probabilities of all distinct paths starting from the HARQ-MM state $i, i=1,\cdots,K$, and ending at the HARQ-MM state $j, j=1,\cdots,K$, without visiting any states at round 1. Then $\widehat{\mathbf{P}}(i,j)$ is given by

$$\begin{aligned}&\widehat{\mathbf{P}}(i,j)\\&=(1-\epsilon_i)P_{i,j}+\sum_{s_1^L\in\{i\}\times\mathcal{S}^{L-1}}\epsilon_{s_1}P_{s_1,s_2}\cdots\epsilon_{s_1^{L-1}}P_{s_{l-1},s_L}P_{s_L,j}\\&\quad+\sum_{l=2}^{L-1}\left(\sum_{s_1^l\in\{i\}\times\mathcal{S}^{l-1}}\epsilon_{s_1}P_{s_1,s_2}\cdots\epsilon_{s_1^{l-1}}P_{s_{l-1},s_l}(1-\epsilon_{s_1^l})P_{s_l,j}\right).\end{aligned} \tag{31}$$

In (31) we only use the FSMC transition probabilities, and the probability of error values in the HARQ-MM states. An alternative form for $\widehat{\mathbf{P}}(i,j)$ is given by

$$\begin{aligned}\widehat{\mathbf{P}}(i,j)=\sum_{l=2}^{L-1}&\left(\sum_{s_1^l\in\{i\}\times\mathcal{S}^{l-1}}\frac{p_{s_1^l}}{p_i}(1-\epsilon_{s_1^l})P_{s_l,j}\right)\\&+(1-\epsilon_i)P_{i,j}+\sum_{s_1^L\in\{i\}\times\mathcal{S}^{L-1}}\frac{p_{s_1^L}}{p_i}P_{s_L,j}\end{aligned} \tag{32}$$

where p_i can be set to any constant because for $s_1^l\in\{i\}\times\mathcal{S}^{l-1}$ all $p_{s_1^l}$ terms contain a factor p_i as can seen in the recursive expression given in (30). We can store $p_{s_1^{l-1}}$ values (with a fixed p_i) as we compute them at each HARQ round of the summation, and use the stored values at the next round to compute $p_{s_1^l}$ by (30). We express the total probability constraint $\mathbf{p}_{1\times N_{\text{sts}}}\mathbf{1}_{N_{\text{sts}}\times 1}=1$ in terms of the HARQ-MM states at round 1 by the vector $\mathbf{d}_{K\times 1}$ whose i-th element is

$$d_i=1+\sum_{l=2}^{L}\left(\sum_{s_1^l\in\{i\}\times\mathcal{S}^{l-1}}\epsilon_{s_1}P_{s_1,s_2}\cdots\epsilon_{s_1^{l-1}}P_{s_{l-1},s_l}\right), \tag{33}$$

which can be alternatively expressed by

$$d_i=1+\sum_{l=2}^{L}\left(\sum_{s_1^l\in\{i\}\times\mathcal{S}^{l-1}}\frac{p_{s_1^l}}{p_i}\right). \tag{34}$$

It follows that $\widehat{\mathbf{p}}_{1\times K}$ is the solution to the system of linear equations

$$\widehat{\mathbf{p}}_{1\times K}\widehat{\mathbf{P}}_{K\times K}=\widehat{\mathbf{p}}_{1\times K},\quad \widehat{\mathbf{p}}_{1\times K}\mathbf{d}_{K\times 1}=1. \tag{35}$$

To summarize, we have successfully reduced the system of linear equations of dimensionality N_{sts} used to compute the HARQ-MM steady-state distribution given in (28), to one of dimensionality K, the number of FMSC states, given in (35).

V. Numerical Results

In this section we evaluate the throughput performance of HARQ over a GEC (i.e. 2 channel states) and over a FSMC (i.e. an arbitrary number of states) as a function of the delay threshold, delay-violation probability, SNR and modulation scheme. In addition, we evaluate the HARQ throughput of Luby transform (LT) codes [29] by simulation and compare it to the theoretical performance.

A. HARQ Performance Over a GEC

In this subsection, we analyze the performance of a HARQ scheme with 2 rounds over a GEC model of the Rayleigh fading channel [14], [15] by deriving an approximation to the maximum achievable HARQ rate. Our motivation for the closed-form approximation is not to reduce the computational complexity of evaluating the HARQ throughput as the algorithm presented in Section IV.B significantly reduces the complexity. The approximation gives us insight into the relationship between the HARQ performance and the system parameters such as the delay threshold, delay-violation probability constraint and SNR.

First, we assume that the value of R is set greater than the capacity of the BAD state denoted by C_b—$mC_{\text{BSC}}(\delta_1)$ or $C_{\text{PAM}}(\frac{\gamma_1}{2})$. We likewise denote the capacity of

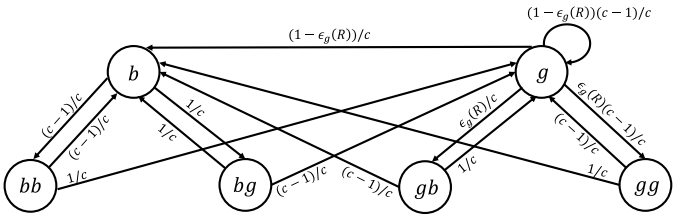


Fig. 3. The finite-state Markov model of HARQ over the TB-based GEC with 2 HARQ rounds, and $P_{g,b} = P_{b,g} = \frac{1}{c}$ for $\epsilon_b(R) \approx 1$.

the GOOD state by C_g—$mC_{\mathrm{BSC}}(\delta_2)$ or $C_{\mathrm{PAM}}(\frac{\gamma_2}{2})$. Accordingly, we further assume that—given that practical values of $2N_{\mathrm{TB}}$ are in the order of hundreds—the following holds:

$$(R - C_b)\sqrt{2N_{\mathrm{TB}}/V_b} \gg 1, \tag{36}$$

where V_b is the dispersion of the BAD state ($mV_{\mathrm{BSC}}(\delta_2)$ or $V_{\mathrm{PAM}}(\frac{\gamma_1}{2})$). The equation (36), together with (18) and (27), implies $\epsilon_b(R) \approx 1$. We treat c, which represents the ratio of the average state duration to the TB duration t_{TB}, as a design parameter. The design parameter c uniquely specifies the TB duration t_{TB} and GEC state transition probabilities by $P_{g,b} = P_{b,g} = \frac{1}{c}$. The resulting HARQ-MM shown in Fig. 3 where we also invoked the approximation $\epsilon_b(R) \approx 1$.

The steady-state probabilities of the states b, g, and bb of the HARQ-MM shown in Fig. 3 are given by

$$\begin{aligned} p_b(R) &= \frac{\epsilon_g(R)(c-2)+c}{\epsilon_g(R)(4c-6)+4c-2} \\ p_g(R) &= \frac{2(c-1)}{\epsilon_g(R)(4c-6)+4c-2} \\ p_{bb}(R) &= \frac{c-1}{c} p_b(R). \end{aligned} \tag{37}$$

It can be shown that $p_g(R)$, and $p_b(R)$—and $p_{bb}(R)$ since it scales with $p_b(R)$—are monotonically decreasing functions of R. By setting $\epsilon_g(R) = 1$, and $\epsilon_g(R) = 0$ we can respectively lower and upper bound $p_b(R)+p_g(R)$, and $p_{bb}(R)$ as follows:

$$\begin{aligned} \frac{1}{2} \le p_b(R) + p_g(R) &\le \frac{3c-2}{4c-2} \\ \frac{c-1}{4c} \le p_{bb}(R) &\le \frac{c-1}{4c-2}. \end{aligned} \tag{38}$$

We can use the bounds in (38) to derive tight approximations to $R^*_{\mathrm{HARQ}}(\epsilon)$. First, for a given probability of error constraint ϵ to be satisfied, from (15), we must have $\epsilon_{bb}(R)p_{bb}(R) < \epsilon$. By using $\epsilon_{bb}(R)p_{bb}(R) < \epsilon$ and the lower bound on $p_{bb}(R)$ given in (38) we can upper bound $\epsilon_{bb}(R)$. The upper bound on $\epsilon_{bb}(R)$ can be converted to an upper bound on R as $\epsilon_{bb}(R)$ is an increasing function of R from (18) and (27):

$$R < 2C_b - \sqrt{\frac{V_b}{N_{\mathrm{TB}}}} Q^{-1}\left(\frac{4c\epsilon}{c-1}\right). \tag{39}$$

We note from (39) that R is bounded above by $2C_b$. On the other hand, the ratio $\frac{C_g}{C_b}$ (in general) is decreasing with increasing average received SNR. We define the *low SNR region* as the region where $\frac{C_g}{C_b} > 2$ holds. Since V_g is small in general, this is the region where the approximation $\epsilon_g(R) \approx 0$ is accurate. Then from (38) we have $p_b(R) + p_g(R) \approx \frac{3c-2}{4c-2}$, which implies $R_{\mathrm{HARQ}}(R) \approx 2R\frac{3c-2}{4c-2}$ from (15). In other words, in the low SNR region the optimal rate is the largest rate satisfying the probability of error constraint. It is straightforward to show that this is (approximately) the rate R for which $\epsilon_{bb}(R) = \frac{\epsilon}{p_{bb}(R)}$ holds. By using $p_{bb}(R) \approx \frac{c-1}{4c-2}$ (from (38)) we get the low SNR region approximation to the maximum achievable HARQ (average) rate given by

$$R^*_{\mathrm{HARQ}}(\epsilon) \approx 2\frac{3c-2}{4c-2}\left(2C_b - \sqrt{\frac{V_b}{N_{\mathrm{TB}}}} Q^{-1}\left(\frac{(4c-2)\epsilon}{c-1}\right)\right). \tag{40}$$

We define the *high SNR region* as the region where $\frac{C_g}{C_b} < 2$ holds. In the high SNR region we can bound R by C_g. This is because rates greater than C_g lead to $\epsilon_g(R) \approx 1$ from $V_g \approx 0$, in which case HARQ becomes a fixed blocklength scheme as it always goes to round 2. By assuming $\epsilon_g(R) \approx 0$ for $R < C_g$ from $V_g \approx 0$ and using the upper bound on $p_b(R) + p_g(R)$ given in (38) we can approximate $R^*_{\mathrm{HARQ}}(\epsilon)$ by

$$R^*_{\mathrm{HARQ}}(\epsilon) \lesssim 2C_g\frac{3c-2}{4c-2}. \tag{41}$$

Since (41) is an approximate upper bound on $R^*_{\mathrm{HARQ}}(\epsilon)$, and (40) is based on an optimistic assumption, the two approximations can be combined as follows

$$\begin{aligned} &R^*_{\mathrm{HARQ}}(\epsilon) \\ &\approx 2\frac{3c-2}{4c-2}\min\left\{C_g, 2C_b - \sqrt{\frac{V_b}{N_{\mathrm{TB}}}} Q^{-1}\left(\frac{(4c-2)\epsilon}{c-1}\right)\right\}. \end{aligned} \tag{42}$$

Next, we present numerical results to demonstrate the accuracy of this approximation.

For the numerical results we consider the hard-decision HARQ with QPSK modulation. We compute $R^*_{\mathrm{HARQ}}(\epsilon)$ numerically from (16) as a function of the SNR where the probability of error values are computed from (18). We will refer to these results as "exact." We also compute the approximation given in (42) for the same SNR values. The Doppler frequency, the bandwidth and c values are given in Table I. The resulting TB size is $N_{\mathrm{TB}}(W, t_{\mathrm{TB}}) = 256$ symbols/TB, TB duration is $t_{\mathrm{TB}} = 0.00142$ seconds (delay threshold of 0.00284 seconds or 2.84 ms) while the code blocklength is 1024 bits. We consider probability of error constraints of $\epsilon = 10^{-6}$, and $\epsilon = 10^{-2}$. In Fig. 4 we plot the resulting HARQ throughput curves (i.e. the HARQ rate scaled by the bandwidth) for $\epsilon = 10^{-6}$ (top) and $\epsilon = 10^{-2}$ (bottom). For reference, we also plot the capacity curve of the HD-TB-based GEC with the CSI available only at the receiver, given by $B(C_b + C_g)$ (bits/s) for equal duration partition.

TABLE I
WIRELESS PARAMETER VALUES

$B = 180$ KHz	Bandwidth
$f_{\mathrm{D}} = 100$ Hz	Doppler Frequency
$c = 3.367$	Channel Partition Parameter

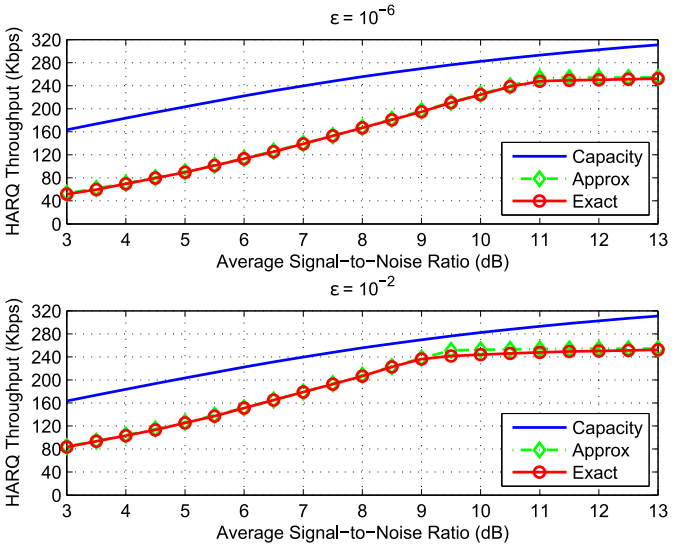


Fig. 4. The approximation to the maximum achievable HARQ throughput over the TB-based GEC for $L = 2$ HARQ rounds, along with the exact value acquired through numerical optimization for $\epsilon = 10^{-6}$ (top), and $\epsilon = 10^{-2}$ (bottom).

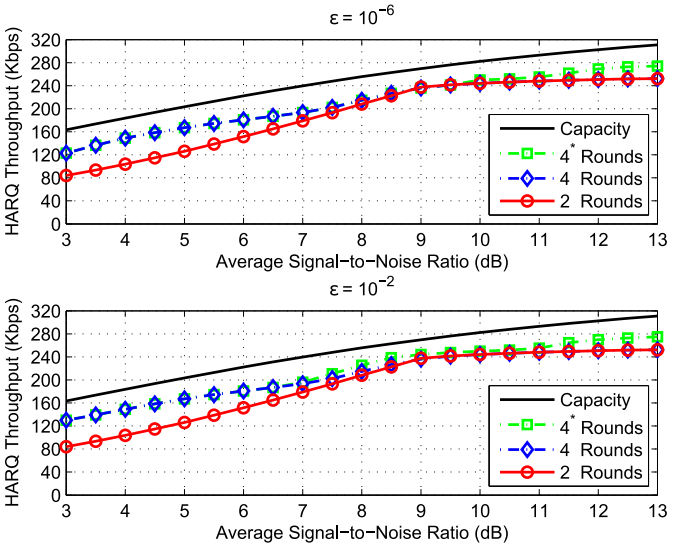


Fig. 5. The maximum achievable HARQ throughput over the TB-based GEC for $L = 2$ and $L = 4$ HARQ rounds with rate constraint $R < 1$, and $L = 4^*$ HARQ rounds without constraints on rate R for $\epsilon = 10^{-6}$ (top), and $\epsilon = 10^{-2}$ (bottom).

In Fig. 4 we observe that our approximation is very accurate, especially in the low SNR region and for $\epsilon = 10^{-6}$. The error between the approximation and exact curves for $\epsilon = 10^{-2}$ is noticeable only around the boundary of the high and low SNR regions. This intuitively makes sense since (42) is a piecewise approximation obtained from two approximations, one under low SNR assumption and the other under high SNR assumption. We also observe that the gap between the HARQ throughput and the channel capacity gets wider as the SNR decreases in the low SNR region, and as it increases in the high SNR region. In the low SNR region the ratio $\frac{C_g}{C_b}$ is increasing with decreasing SNR while the feasible values of R are bounded by $2C_b$. As a result the gap between the HARQ throughput, which scales with $2BR$, and $B(C_b + C_g)$, the channel capacity, increases as the SNR decreases. In the high SNR region the gap between the HARQ throughput and the channel capacity increases because the HARQ throughput curve is approximately constant (from (41) and $C_g \approx 1$) while the channel capacity is increasing. This observation can also be explained by the fact that the decoder is not able to take advantage of C_b increasing with the SNR by decoding at an earlier time than the end of the second TB. As a result, the higher SNR does not translate to a higher HARQ throughput. Last, we observe that the HARQ throughput is significantly higher for $\epsilon = 10^{-2}$ than it is for $\epsilon = 10^{-6}$ in the low SNR region while they are comparable in the high SNR region. This observation can be explained by the fact that the low SNR approximation in (42) strongly depends on ϵ while the high SNR approximation does not depend on ϵ.

The factor of 2 in $2C_b$ in the rate limiting expression (40) for the low SNR region is the number of HARQ rounds; the HARQ throughput in the low SNR region can be improved by increasing the number of HARQ rounds. To demonstrate this strategy we plot the throughput of a 4-round HARQ system with $R < 1$, throughput of the 2-round system and throughput of a 4-round HARQ system without any constraints on R, denoted as 4^* in Fig. 5. It is observed that the HARQ throughput is significantly improved with 4 rounds, with or without the rate constraint, in the low SNR region. However, with the 4-round HARQ system under $R < 1$ there is no improvement in the high SNR region. For rates near C_g probability of error is approximately 0 at all HARQ states at round 2. As a result the 4-round HARQ system effectively becomes a 2-round system. When the rate R is allowed to take values up to $L = 4$, shown with label 4^* in Fig. 5, the values of R that maximize the HARQ throughput in the high SNR region are close to 2. In other words, the HARQ system effectively becomes a 3-round system as it always goes to round 2. We can define the rate with respect to the first decoding time, i.e. round 2, by $R = \frac{\log_2 M}{4N_{\mathrm{TB}}}$ where $4N_{\mathrm{TB}}$ is the number of PAM symbols received at round 2. Then the effective HARQ rate drops from R at round 2, to $\frac{2R}{3}$ at round 3, and to $\frac{R}{2}$ at round 4. Compared to the 2-round HARQ where the effective rate drops from R to $\frac{R}{2}$, the rate granularity at decoding times is finer, which translates to increased throughput.

B. HARQ Performance Over a FSMC

In Fig. 6 we plot the hard-decision HARQ throughput with QPSK and 16QAM, for the probability of error (delay-violation probability constraint) value of $\epsilon = 10^{-2}$. The number of FSMC states is $K = 4$, and the number of HARQ rounds is $L = 8$. The Doppler frequency, bandwidth and c values are given in Table I. The resulting TB size is $N_{\mathrm{TB}} = 128$ QAM symbols/TB (256 PAM symbols/TB), and the code blocklength is 1024 QAM symbols resulting in a delay threshold of approximately 5.6 ms. For reference we also evaluate the channel capacity of the HD-TB-based FSMC for each modulation. It can be seen that modulation order plays an important role in the underlying HARQ performance. We observe that the HARQ throughput curve of each modulation scheme follows the corresponding channel capacity curve

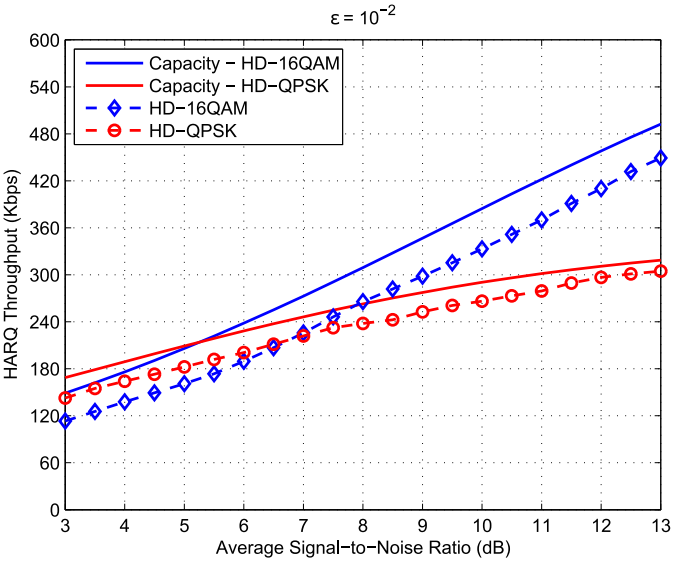


Fig. 6. The maximum achievable HARQ throughput over the 4-state HD-TB-based FSMC with $L = 8$ HARQ rounds with QPSK and 16QAM modulations for $\epsilon = 10^{-2}$ along with the channel capacity of the HD-TB-based FSMC with each modulation.

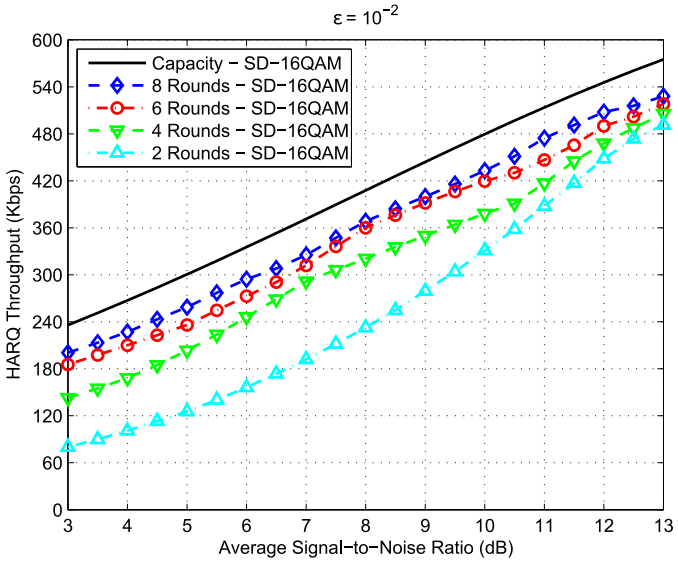


Fig. 7. The maximum achievable HARQ throughput over the 4-state SD-TB-based FSMC with 16QAM and $\epsilon = 10^{-2}$ for various values of the number of HARQ rounds L.

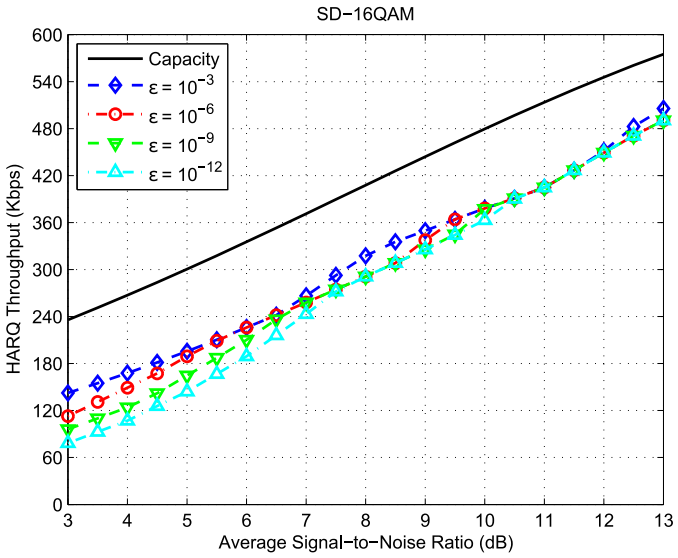


Fig. 8. The maximum achievable HARQ throughput over the 4-state SD-TB-based FSMC with 16QAM and $L = 4$ for various values of the delay-violation probability constraint ϵ.

closely. As a result, in the low SNR region QPSK has higher throughput while in the high SNR region 16QAM has higher throughput. With QPSK the HARQ throughput is greater than 90% of the capacity above 7 dB SNR, while with 16QAM it is greater than 90% of the capacity above 12 dB SNR.

To demonstrate the performance improvement achieved with HARQ compared to wireless communication systems without feedback we also evaluate the maximum achievable throughput over the HD-TB-based FSMC by using the framework presented in [17] with the parameter values shown in Table I, 4 channel states and $\epsilon = 10^{-2}$. With QPSK at 7 dB SNR the required number of TBs to achieve 90% of the capacity is 710 which corresponds to a coding delay of 497 ms. With 16QAM at 12 dB SNR the required number of TBs to achieve 90% of the capacity is 760 which corresponds to a coding delay of 532 ms. At those SNR values the HARQ scheme shown in Fig. 6 achieves 90% of the respective capacities with a delay of only 5.6 ms. We conclude that HARQ drastically reduces the required coding delay to achieve throughput values near the channel capacity. The required number of TBs to achieve even 50% of the capacity at 10 dB SNR without feedback for QPSK and 16QAM are 10 and 30, respectively, corresponding to coding delay values of 7 ms and 21 ms.

In Fig. 7 we plot the soft-decision HARQ throughput as a function of the SNR with 16QAM and $\epsilon = 10^{-2}$ for the number of HARQ rounds values $L = 2, 4, 6,$ and 8. These values correspond to delay threshold values of $1.4, 2.8, 4.2,$ and 5.6 ms respectively. We observe in Fig. 7 that the delay threshold significantly affects the throughput performance, especially in the low SNR region. This result can be explained by the channel capacity values of the FSMC states in the low SNR region and high SNR region. At 3 dB SNR the FSMC state channel capacity values are $0.33, 1.17, 2.07,$ and 2.89 bits/channel use (16QAM symbol), and at 13 dB SNR these values are $1.77, 3.39, 3.94,$ and 3.99 bits/channel use. We note that if the channel is in state 1, the probability that it will stay in state 1 for the next TB transmission is $(c-1)/c = 0.705$. At 3 dB SNR if the channel stays at state 1 for 4 consecutive TBs or even for 1 TB at state 2 and for 3 TBs in state 1, the maximum rate that can be supported (i.e. the sum of the state capacities) is 1.32 or 2.16, which is much smaller than the capacity of state 4. The disparity in the channel qualities translate to disparity in the required number of HARQ rounds to a successful transmission for an initial rate chosen close to the capacity of state 4. In other words, when L is small (e.g. 2 or 4) the HARQ transmission is "cut short", and only very small rates—with respect to state 4 capacity—can be supported. At 13 dB on the other hand the capacity of the worst state after only 2 rounds reaches 3.54.

Figure 8 shows the HARQ throughput performance of a 4-state SD-TB-based FSMC with 16QAM and $L = 4$

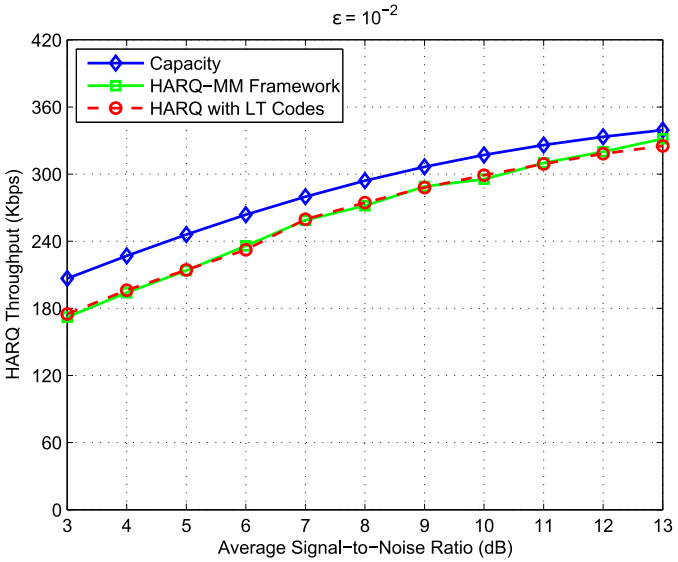


Fig. 9. The maximum achievable HARQ throughput computed with the HARQ-MM over the 6-state SD-TB-based FSMC and the HARQ throughput of LT codes over the Rayleigh fading channel acquired by simulation, with $L = 8$ rounds and QPSK for $\epsilon = 10^{-2}$.

under different reliability constraints (different values of ϵ). We observe that the HARQ throughput does not strongly depend on the delay-violation probability except for the low SNR region. We again emphasize that, unlike the communication schemes without feedback, with HARQ the rate that achieves the maximum throughput doesn't necessarily occur at the maximum of the feasible rates. As such, increasing the delay-violation probability constraint doesn't necessarily increase the HARQ throughput. The strong dependence of the throughput on ϵ in the low SNR region is due to the disparity between the capacity of state 1 and capacity of state 4. As ϵ is decreased the maximum feasible value of R becomes much smaller than 4 (for 16QAM or 2 for 4PAM), the highest coding rate that can be set in round 1 with a unique codeword for each message. At 3 dB for $\epsilon = 10^{-12}$ the maximum feasible rate is 0.87 while it is 4.7 at 13 dB SNR. In other words while the probability of error constraint significantly limits the coding rate and hence the throughput at low SNR region, at high SNR region the coding rate can be set as high as 4.

C. HARQ Performance With LT Codes

In this subsection we compare the HARQ throughput predicted by the HARQ-MM framework to the throughput of a channel code with HARQ simulated over the Rayleigh fading channel. As the channel code we use a truncated systematic LT code where the degree distribution is Robust Soliton with parameters $c_{\mathrm{LT}} = 0.1$ and $\delta_{\mathrm{LT}} = 0.5$ [29]. We utilize LT codes due to the ease of varying the coding rate incrementally while maintaining coding performance. At each SNR value we simulate the LT code with all possible rate values R, compute the resulting average HARQ throughput values, and identify the maximum throughput. The Rayleigh fading component $h(t)$ is randomly generated according to the autocorrelation function $J_0(2\pi f_{\mathrm{D}} t_s)$ where t_s is the sampling interval (i.e. Jakes p.s.d. [32]). The sampling interval is chosen such that $J_0(2\pi f_{\mathrm{D}} t_s) > 0.99$ and the number of symbols per t_s is a power of 2. We use a soft-decision HARQ scheme with $L = 8$ rounds, and QPSK modulation. The Doppler frequency and the bandwidth values are given in Table I while $c = 4.97$. The resulting TB size is $N_{\mathrm{TB}} = 64$ QPSK symbols/TB; while the code blocklength is 1024 bits. The HARQ-MM is based on a SD-TB-based FSMC with $K = 6$ states. In Fig. 9 we plot the HARQ throughput with the LT code, the HARQ-MM throughput curve and the capacity of the SD-TB-based FSMC for reference. We observe that the LT code simulation performance tightly matches the theoretical HARQ throughput curve computed with the HARQ-MM.

VI. Conclusions

In the light of the popularity of delay-sensitive applications such as real-time conversational video and online gaming and future trends toward extremely low latency applications (i.e. tactile Internet), performance analysis and design of wireless communication systems under delay constraints becomes increasingly important. IR-HARQ is a viable communication strategy for delay-sensitive applications with its ability to adjust the communication rate according to varying channel conditions. We characterized the throughput performance of IR-HARQ over the Rayleigh fading channel under finite blocklength and probability of error constraints as a function of the modulation scheme. The blocklength is a measure of the maximum coding delay or delay threshold while the probability of error constraint translates to a coding delay-violation probability constraint. We introduced the HARQ-MM to track the HARQ throughput and probability of error as a function of the system parameters, and derived the dispersion of parallel AWGN channels with discrete input alphabets, which was used to compute the state transition probabilities of the HARQ-MM. In addition, we developed an algorithm to efficiently compute the steady-state distribution of the HARQ-MM. We analyzed the performance of a HARQ scheme with 2 rounds over the GEC, with focus on the performance dependence on the number of HARQ rounds. We presented various numerical results, which showed that HARQ achieves 90% of the channel capacity with blocklength values as little as 1024 symbols, which translate to coding delay values in the order of 5 ms. The numerical results also showed that the HARQ throughput strongly depends on the coding delay and delay-violation probability in the low SNR region. Last, we evaluated the performance of LT codes with HARQ over the Rayleigh fading channel by simulation. The LT code performance very closely matched the throughput performance predicted by our framework.

The work presented here can be extended to the case where the transmitter adapts its communication strategy, such as coding rate and modulation scheme, according to CSI feedback as in 3GPP LTE/LTE-Advanced and 5G systems. In addition, the coding latency analyzed here, is one of many components of system latency over a wireless link. Future work also includes an effort to map coding latency studied here to overall system latency or joint consideration of multiple latency components, e.g. coding and queuing.

APPENDIX

A. Proof of Equation (11)

The probability of the HARQ-MM being in the HARQ round $l, 1 \le l \le L$, is given by the expression

$$\widetilde{p}_l = \sum_{s_1^l \in \mathcal{S}^l} p_{s_1^l}. \tag{43}$$

Then, the probability of the HARQ transmission being terminated at round l, $1 \le l \le L$, is given by

$$\Pr\{L_{\text{HARQ}} = l\} = \frac{1}{\widetilde{p}_1}\left(\widetilde{p}_l - \widetilde{p}_{l+1}\right) \tag{44}$$

where we used the fact that $(\widetilde{p}_l - \widetilde{p}_{l+1})$ is the joint probability of a new HARQ transmission starting given by $\widetilde{p}_1$ and the new transmission ending at round l given by $\Pr\{L_{\text{HARQ}} = l\}$; and $\widetilde{p}_{L+1} = 0$. With $\Pr\{L_{\text{HARQ}} = l\}$ given as in (44) the total probability constraint $\sum_{l=1}^{L} \Pr\{L_{\text{HARQ}} = l\} = 1$ is satisfied; accordingly, the factor $\frac{1}{\widetilde{p}_1}$ can be viewed as the normalization factor. The expected value L_{HARQ} is given by the expression

$$\begin{aligned}
\mathbb{E}[L_{\text{HARQ}}] &= \sum_{l=1}^{L} l \Pr\{L_{\text{HARQ}} = l\} \\
&= \sum_{l=1}^{L} l \frac{1}{\widetilde{p}_1}\left(\widetilde{p}_l - \widetilde{p}_{l+1}\right) \\
&= \frac{1}{\widetilde{p}_1}\sum_{l=1}^{L} \widetilde{p}_l \\
&= \frac{1}{\widetilde{p}_1}.
\end{aligned} \tag{45}$$

By the definition of $\widetilde{p}_1$ given in (43), we get the desired result $\mathbb{E}[L_{\text{HARQ}}] = \frac{1}{\sum_{s_1 \in \mathcal{S}} p_{s_1}}$.

B. Proof of Theorem 1

We first state a result from [12]. According to the dependency testing (DT) bound, given a—discrete-input continuous-output—channel $p_{Y_1^n|X_1^n}$ for an arbitrary input distribution $P_{X_1^n}$ there exists a code with M codewords whose average probability of error denoted by $\epsilon(n, M)$ satisfying

$$\epsilon(n, M) \le \mathbb{E}\left[\exp\left\{-\left[i\left(X_1^n, Y_1^n\right) - \log\frac{M-1}{2}\right]^+\right\}\right] \tag{46}$$

where $[a]^+ = a$ if $a \ge 0$, and $[a]^+ = 0$ otherwise; exp, and log have the same base; and

$$i\left(X_1^n, Y_1^n\right) \triangleq \log\frac{p_{Y_1^n|X_1^n}\left(Y_1^n|X_1^n\right)}{p_{Y_1^n}\left(Y_1^n\right)}. \tag{47}$$

The expected value is taken according to $p_{Y_1^n|X_1^n}P_{X_1^n}$. We consider L parallel AWGN channels each with noise p.s.d. $\frac{N_0}{2}$ and 2^m-ary PAM input signals, the l-th channel with signal constellation $\mathcal{M}_l$. The channel input is the sequence of length-n vectors $\mathbf{X}_1^L = \mathbf{X}_1, \cdots, \mathbf{X}_L$, $\mathbf{X}_l = X_{l,1}, \cdots, X_{l,n}$, with $x_{l,j} \in \mathcal{M}_l$; and the channel output $\mathbf{Y}_1^L, y_{l,j} \in \mathbb{R}$, is related to the channel input by $p_{\mathbf{Y}_1^L|\mathbf{X}_1^L} = \prod_{j=1}^{n}\prod_{l=1}^{L} p_{Y_{l,j}|X_{l,j}}$ where $p_{Y_{l,j}|X_{l,j}}(y_{l,j}|x_{l,j}) = \frac{1}{\sqrt{\pi N_0}}\exp\left\{-\frac{(y_{l,j}-x_{l,j})^2}{N_0}\right\}$. The channel input distribution $P_{\mathbf{X}_1^L}$ is chosen equiprobable. Then we can write $i\left(\mathbf{X}_1^L, \mathbf{Y}_1^L\right)$ as a sum of n i.i.d. random variables by

$$i\left(\mathbf{X}_1^L, \mathbf{Y}_1^L\right) = \sum_{j=1}^{n} Z_j \tag{48}$$

where

$$Z_j = \sum_{l=1}^{L} \log\frac{p_{Y_{l,j}|X_{l,j}}(Y_{l,j}|X_{l,j})}{2^{-m}\sum_{z\in\mathcal{M}_l} p_{Y_{l,j}|X_{l,j}}(Y_{l,j}|z)}. \tag{49}$$

From $\mathbb{E}[Z_j] = C$, and $\text{Var}[Z_j] = V$ we have

$$\begin{aligned}
\mathbb{E}\left[i\left(\mathbf{X}_1^L, \mathbf{Y}_1^L\right)\right] &= nC \\
\text{Var}\left[i\left(\mathbf{X}_1^L, \mathbf{Y}_1^L\right)\right] &= nV.
\end{aligned} \tag{50}$$

Note that the L terms in the definition of (49) are independent random variables; accordingly, we can write C and V by

$$\begin{aligned}
C &= \sum_{l=1}^{L} \mathbb{E}\left[\log\frac{p_{Y_{l,j}|X_{l,j}}(Y_{l,j}|X_{l,j})}{2^{-m}\sum_{z\in\mathcal{M}_l} p_{Y_{l,j}|X_{l,j}}(Y_{l,j}|z)}\right] \\
V &= \sum_{l=1}^{L} \text{Var}\left[\log\frac{p_{Y_{l,j}|X_{l,j}}(Y_{l,j}|X_{l,j})}{2^{-m}\sum_{z\in\mathcal{M}_l} p_{Y_{l,j}|X_{l,j}}(Y_{l,j}|z)}\right],
\end{aligned} \tag{51}$$

where expected values are taken according to $p_{Y_{l,j}|X_{l,j}}P_{X_{l,j}}$, $P_{X_{l,j}}(x_{l,j}) = 2^{-m}$ for all $x_{l,j} \in \mathcal{M}_l$, and $p_{Y_{l,j}|X_{l,j}}(y_{l,j}|x_{l,j})$ is Gaussian with variance $\frac{N_0}{2}$ and mean $x_{l,j}$.

In the rest of the proof we use some results proved or simply stated in [12]. Given that $T = \mathbb{E}[|Z_j - C|^3] < \infty$ by applying the Berry-Esseen Theorem [12] to $i\left(\mathbf{X}_1^L, \mathbf{Y}_1^L\right)$ for an arbitrary λ we get the following bound

$$\left|\mathbb{P}\left[i(\mathbf{X}_1^L, \mathbf{Y}_1^L) \le nC + \lambda\sqrt{nV}\right] - Q(-\lambda)\right| \le \frac{B_2}{\sqrt{n}}, \tag{52}$$

where $B_2 = \frac{6T}{V^{3/2}}$. We now state a result proved in [12]. Let $\Gamma_1, \cdots, \Gamma_n$ be independent random variables, $\sigma^2 = \sum_{j=1}^{n}\text{Var}[\Gamma_j]$ be non-zero, and $\widetilde{T} = \sum_{j=1}^{n}\mathbb{E}[|\Gamma_j - \mathbb{E}[\Gamma_j]|^3] < \infty$. Then for any A we have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{-\sum_{j=1}^{n}\Gamma_j\right\} 1_{\left\{\sum_{j=1}^{n}\Gamma_j > A\right\}}\right] \\
&\qquad \le 2\left(\frac{\log 2}{\sqrt{2\pi}} + \frac{12\widetilde{T}}{\sigma^2}\right)\frac{1}{\sigma}\exp\{-A\}.
\end{aligned} \tag{53}$$

We apply this result to $i\left(\mathbf{X}_1^L, \mathbf{Y}_1^L\right) = \sum_{j=1}^{n} Z_j$ where $\sigma^2 = nV$ and $\widetilde{T} = nT$. By multiplying both sides of (53) by $\exp\{A\}$ we get the following bound

$$\mathbb{E}\left[\exp\left\{-i\left(\mathbf{X}_1^L, \mathbf{Y}_1^L\right) + A\right\} 1_{\left\{i(\mathbf{X}_1^L, \mathbf{Y}_1^L) > A\right\}}\right] \le \frac{B_1}{\sqrt{n}} \tag{54}$$

where $B_1 = 2\left(\frac{\log 2}{\sqrt{2\pi}} + \frac{12T}{V}\right)\frac{1}{\sqrt{V}}$. Now we set

$$\begin{aligned}
\log\frac{M-1}{2} &= nC - \sqrt{nV}Q^{-1}(\epsilon_n) \\
\epsilon_n &= \epsilon - \frac{B_1 + B_2}{\sqrt{n}}.
\end{aligned} \tag{55}$$

Then by the DT bound there exists a code with M codewords whose average probability of error denoted by p_e bounded by

$$p_e \leq \mathbb{E}\left[\exp\left\{-\left[i\left(\mathbf{X}_1^L, \mathbf{Y}_1^L\right) - \log\frac{M-1}{2}\right]^+\right\}\right] \quad (56)$$

$$\leq \mathbb{P}\left[i(\mathbf{X}_1^L, \mathbf{Y}_1^L) \leq \log\frac{M-1}{2}\right] + \frac{B_1}{\sqrt{n}} \quad (57)$$

$$= \mathbb{P}\left[i(\mathbf{X}_1^L, \mathbf{Y}_1^L) \leq nC - \sqrt{nV}Q^{-1}(\epsilon_n)\right] + \frac{B_1}{\sqrt{n}} \quad (58)$$

$$\leq Q\left(Q^{-1}(\epsilon_n)\right) + \frac{B_2}{\sqrt{n}} + \frac{B_1}{\sqrt{n}} \quad (59)$$

$$= \epsilon_n + \frac{B_1+B_2}{\sqrt{n}} \quad (60)$$

$$= \epsilon \quad (61)$$

where (57) is from (54) with $A = \log\frac{M-1}{2}$, and (59) is from (52) where we have $\lambda = -Q^{-1}(\epsilon_n)$. Then from $\log M^* \geq \log M > \log\frac{M-1}{2}$, and (55) we get the desired result

$$\log M^*(n,\epsilon) \geq nC - \sqrt{nV}Q^{-1}(\epsilon_n) \quad (62)$$

$$\geq nC - \sqrt{nV}Q^{-1}\left(\epsilon - \frac{B_1+B_2}{\sqrt{n}}\right) \quad (63)$$

$$= nC - \sqrt{nV}Q^{-1}(\epsilon) + O(1). \quad (64)$$

where we applied Taylor's theorem to Q^{-1} in (63).

Now we show that in the case of BPSK, i.e. $m = 1$, the reverse inequality of (64) is also true with the same capacity and dispersion terms C and V. Consider the binary hypothesis test between the distributions $p_{\mathbf{Y}_1^L|\mathbf{X}_1^L=\mathbf{x}_1^L}$—where $\mathbf{x}_1^L \in \mathcal{M}_1^n \times \cdots \times \mathcal{M}_L^n$ is fixed, and $\mathcal{M}_l = \{-A_l, +A_l\}$—and $p_{\mathbf{Y}_1^L}$, the unconditional output distribution induced by an equiprobable input distribution $P_{\mathbf{X}_1^L}$, which can be written in the form $p_{\mathbf{Y}_1^L}(\mathbf{y}_1^L) = \prod_{j=1}^n \prod_{l=1}^L p_{Y_{l,j}}(y_{l,j})$ where

$$p_{Y_{l,j}}(y_{l,j}) = \frac{1}{2\sqrt{\pi N_0}}\left(e^{-\frac{\left(y_{l,j}-A_l\right)^2}{N_0}} + e^{-\frac{\left(y_{l,j}+A_l\right)^2}{N_0}}\right). \quad (65)$$

Let $\beta_{1-\epsilon}(p_{\mathbf{Y}_1^L|\mathbf{X}_1^L=\mathbf{x}_1^L}, p_{\mathbf{Y}_1^L})$ denote the minimum probability of error under the hypothesis $p_{\mathbf{Y}_1^L}$ given that the probability of success under the hypothesis $p_{\mathbf{Y}_1^L|\mathbf{X}_1^L=\mathbf{x}_1^L}$ is at least $1-\epsilon$. Due to the symmetry of the channel input alphabet, and the symmetry of the distribution $p_{\mathbf{Y}_1^L}$ given in (65), $\beta_{1-\epsilon}(p_{\mathbf{Y}_1^L|\mathbf{X}_1^L=\mathbf{x}_1^L}, p_{\mathbf{Y}_1^L})$ does not vary with the particular input $\mathbf{x}_1^L$. Therefore, we can invoke [12, Th. 28], i.e. the size M of any code for the channel $p_{\mathbf{Y}_1^L|\mathbf{X}_1^L}$ with the average probability of error ϵ satisfies

$$M \leq \frac{1}{\beta_{1-\epsilon}\left(p_{\mathbf{Y}_1^L|\mathbf{X}_1^L=\mathbf{x}_1^L}, p_{\mathbf{Y}_1^L}\right)}. \quad (66)$$

Now we state a result from [12]. Consider two n-fold product probability measures P^n and Q^n defined on the same probability space $\mathcal{A}^n$. Define the quantities $D(P||Q) = \mathbb{E}_P\left[\log\frac{P}{Q}\right]$, $V(P||Q) = \mathrm{Var}_P\left[\log\frac{P}{Q}\right]$, and $T(P||Q) = \mathbb{E}_P\left[\left|\log\frac{P}{Q} - D(P||Q)\right|^3\right]$. Given that $V > 0$ and $T < \infty$, for any $\alpha \in [0,1]$ we have

$$\log\beta_\alpha(P^n, Q^n) = -nD(P||Q) - \sqrt{nV(P||Q)}Q^{-1}(\alpha) - \frac{1}{2}\log n + O(1). \quad (67)$$

We can apply this result to $\beta_{1-\epsilon}(p_{\mathbf{Y}_1^L|\mathbf{X}_1^L=\mathbf{x}_1^L}, p_{\mathbf{Y}_1^L})$—with $p_{\mathbf{Y}_1^L}$ induced by an equiprobable $P_{\mathbf{X}_1^L}$ as shown in (65)—by choosing $x_{l,j} = A_l$ for all $l = 1, \cdots, L, j = 1, \cdots, n$. Note that we are free to choose any $\mathbf{x}_1^L$ from the channel input space because $\beta_{1-\epsilon}$ does not depend on the particular choice. This choice of $\mathbf{x}_1^L$ results in $p_{\mathbf{Y}_1^L|\mathbf{X}_1^L=\mathbf{x}_1^L}$ having a product form; in particular $P = \prod_{l=1}^L p_{Y_{l,j}|X_{l,j}=A_l}$ while $Q = \prod_{l=1}^L p_{Y_{l,j}}$ in (67). The quantity $\log\frac{P}{Q}$ can be written as a sum of L independent random variables each of the form $\log\frac{p_{Y_{l,j}|X_{l,j}=A_l}}{p_{Y_{l,j}}}$ and we can write

$$D(P||Q) = \sum_{l=1}^L \mathbb{E}_{p_{Y_{l,j}|X_{l,j}=A_l}}\left[\log\frac{p_{Y_{l,j}|X_{l,j}=A_l}}{p_{Y_{l,j}}}\right]$$
$$V(P||Q) = \sum_{l=1}^L \mathrm{Var}_{p_{Y_{l,j}|X_{l,j}=A_l}}\left[\log\frac{p_{Y_{l,j}|X_{l,j}=A_l}}{p_{Y_{l,j}}}\right] \quad (68)$$

We also observe that if we define $\widetilde{P} = \prod_{l=1}^L p_{Y_{l,j}|X_{l,j}=-A_l}$, then due to the symmetry in the distributions $\left\{p_{Y_{l,j}}\right\}_{l=1,\cdots,L}$ we have $D(\widetilde{P}||Q) = D(P||Q)$ and $V(\widetilde{P}||Q) = V(P||Q)$. It follows that we have $D(P||Q) = C$, and $V(P||Q) = V$ where C, and V are given in (51). Then from (67) we have

$$\log\beta_{1-\epsilon}\left(p_{\mathbf{Y}_1^L|\mathbf{X}_1^L=\mathbf{x}_1^L}, p_{\mathbf{Y}_1^L}\right) = -nC - \sqrt{nV}Q^{-1}(1-\epsilon) - \frac{1}{2}\log n + O(1). \quad (69)$$

From (66) and $-Q^{-1}(1-\epsilon) = Q^{-1}(\epsilon)$ we get the desired result

$$\log M^*(n,\epsilon) \leq nC - \sqrt{nV}Q^{-1}(\epsilon) + \frac{1}{2}\log n + O(1). \quad (70)$$

REFERENCES

[1] C. Sahin, L. Liu, and E. Perrins, "On the finite blocklength performance of HARQ in modern wireless systems," in *Proc. IEEE Global Commun. Conf.*, Austin, TX, USA, Dec. 2014, pp. 3513–3519.

[2] C. Sahin, L. Liu, and E. Perrins, "On the queueing performance of HARQ systems with coding over finite transport blocks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, San Diego, CA, USA, Dec. 2015, pp. 1–7.

[3] *The Tactile Internet*, ITU-T Technology Watch, Geneva, Switzerland, Aug. 2014.

[4] *Study Scenarios Requirements for Next Generation Access Technologies*, document TR 38.913, 3GPP, May 2017.

[5] L. Liu, P. Parag, J. Tang, W.-Y. Chen, and J.-F. Chamberland, "Resource allocation and quality of service evaluation for wireless communication systems using fluid models," *IEEE Trans. Inf. Theory*, vol. 53, no. 5, pp. 1767–1777, May 2007.

[6] L. Liu, P. Parag, and J. F. Chamberland, "Quality of service analysis for wireless user-cooperation networks," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3833–3842, Oct. 2007.

[7] A. ParandehGheibi, M. Medard, A. Ozdaglar, and S. Shakkottai, "Avoiding interruptions—A QoE reliability function for streaming media applications," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 5, pp. 1064–1074, May 2011.

[8] N. Gunaseelan, L. Liu, J.-F. Chamberland, and G. H. Huff, "Performance analysis of wireless hybrid-ARQ systems with delay-sensitive traffic," *IEEE Trans. Commun.*, vol. 58, no. 4, pp. 1262–1272, Apr. 2010.

[9] *Evolved Universal Terrestrial Radio Access (E-UTRA); Further Advancements*, document TR 36.814, 3GPP, Mar. 2010.

[10] H. S. Wang and N. Moayeri, "Finite-state Markov channel—A useful model for radio communication channels," *IEEE Trans. Veh. Technol.*, vol. 44, no. 1, pp. 163–171, Feb. 1995.

[11] Q. Zhang and S. A. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, no. 11, pp. 1688–1692, Nov. 1999.

[12] 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.
[13] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Dispersion of the Gilbert-Elliott channel," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 1829–1848, Apr. 2011.
[14] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell Syst. Tech. J.*, vol. 39, no. 5, pp. 1253–1265, Sep. 1960.
[15] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *Bell Syst. Tech. J.*, vol. 42, no. 5, pp. 1977–1997, Sep. 1963.
[16] C. Sahin, L. Liu, and E. Perrins, "Coding across finite transport blocks in modern wireless communication systems," *IEEE Trans. Commun.*, vol. 62, no. 12, pp. 4184–4197, Dec. 2014.
[17] C. Sahin, L. Liu, and E. Perrins, "Channel coding over finite transport blocks in modern wireless systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Atlanta, GA, USA, Dec. 2013, pp. 3667–3672.
[18] C. Lott, O. Milenkovic, and E. Soljanin, "Hybrid ARQ: Theory, state of the art and future directions," in *Proc. IEEE Inf. Theory Workshop Inf. Theory Wireless Netw.*, Jul. 2007, pp. 1–5.
[19] Y.-M. Wang and S. Lin, "A modified selective-repeat type-II hybrid ARQ system and its performance analysis," *IEEE Trans. Commun.*, vol. COMM-31, no. 5, pp. 593–608, May 1983.
[20] S. Lin and P. Yu, "A hybrid ARQ scheme with parity retransmission for error control of satellite channels," *IEEE Trans. Commun.*, vol. COMM-30, no. 7, pp. 1701–1719, Jul. 1982.
[21] S. C. Draper, L. Liu, A. F. Molisch, and J. S. Yedidia, "Cooperative transmission for wireless networks using mutual-information accumulation," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5151–5162, Aug. 2011.
[22] 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.
[23] S. H. Kim, D. K. Sung, and T. Le-Ngoc, "Performance analysis of incremental redundancy type hybrid ARQ for finite-length packets in AWGN channel," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Atlanta, GA, USA, Dec. 2013, pp. 2063–2068.
[24] 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.
[25] W. Lee, O. Simeone, J. Kang, S. Rangan, and P. Popovski, "HARQ buffer management: An information-theoretic view," *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 4539–4550, Nov. 2015.
[26] L. Liu, Y. Yang, J.-F. Chamberland, and J. Zhang, "Energy-efficient power allocation for delay-sensitive multimedia traffic over wireless systems," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2038–2047, Jun. 2014.
[27] C. She, C. Yang, and L. Liu, "Energy-efficient resource allocation for MIMO-OFDM systems serving random sources with statistical QoS requirement," *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 4125–4141, Nov. 2015.
[28] Y. Li, L. Liu, H. Li, J. Zhang, and Y. Yi, "Resource allocation for delay-sensitive traffic over LTE-advanced relay networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4291–4303, Aug. 2015.
[29] M. Luby, "LT codes," in *Proc. 43rd Annu. IEEE Symp. Found. Comput. Sci.*, Vancouver, BC, Canada, Nov. 2002, pp. 271–280.
[30] T. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.
[31] V. Veeravalli and A. Sayeed, *Wideband Wireless Channels: Statistical Modeling, Analysis and Simulation*. Champaign, IL, USA: Univ. Illinois, 2004.
[32] W. C. Jakes, *Microwave Mobile Communications*. Piscataway, NJ, USA: IEEE Press, 1993.
[33] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 1971–1988, Jul. 2001.

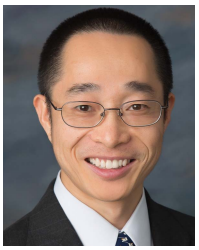
Cenk Sahin (S'05–M'16) received the B.S. degree *(summa cum laude)* in electrical and computer engineering from the University of Missouri-Kansas City, Kansas City, MO, USA, in 2008, and the M.S. and Ph.D. degrees in electrical engineering from The University of Kansas, Lawrence, KS, USA, in 2012 and 2015, respectively. He was a Post-Doctoral Researcher with The University of Kansas from 2015 to 2016 and a Post-Doctoral Research Fellow in the National Research Council (NRC) Research Associateship Program at the Sensors Directorate, Air Force Research Laboratory (AFRL), Wright-Patterson Air Force Base, OH, USA, from 2016 to 2018. As an NRC Research Fellow at AFRL, he led the development of spectrally efficient constant envelope dual-function waveforms for radar and communications. Since 2018, he has been a Research Electronics Engineer with the RF Technology Branch, Sensors Directorate, AFRL. His current research interests include radar-embedded communications waveform and algorithm design and communication theory, information theory, and queuing theory with application to delay-sensitive wireless communications.

Lingjia Liu (SM'15) received the B.S. degree in electronic engineering from Shanghai Jiao Tong University and the Ph.D. degree in electrical and computer engineering from Texas A&M University.

He was an Associate Professor with the EECS Department, The University of Kansas (KU). He was with the Standards and Mobility Innovation Laboratory, Samsung Research America (SRA), where he received the Global Samsung Best Paper Award in 2008 and 2010, respectively. At SRA, he was leading Samsung's efforts on multiuser MIMO, CoMP, and HetNets in 3GPP LTE/LTE-advanced standards. He is currently an Associate Professor with the Bradley Department of Electrical and Computer Engineering, Virginia Tech. He is also the Associate Director for Affiliate Relations with Wireless@Virginia Tech. His general research interests mainly lie in emerging technologies for 5G cellular networks and beyond, including machine learning for wireless networks, massive MIMO, massive machine-type communications, the Internet of Everything, and mm-wave communications. He received the Air Force Summer Faculty Fellowship from 2013 to 2017, the Miller Scholarship at KU in 2014, the Miller Professional Development Award for Distinguished Research at KU in 2015, the 2016 IEEE GLOBECOM Best Paper Award, the 2018 IEEE ISQED Best Paper Award, the 2018 IEEE TCGCC Best Conference Paper Award, and the 2018 IEEE TAOS Best Paper Award. He was an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and the IEEE TRANSACTIONS ON COMMUNICATIONS. He is currently an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Erik Perrins (S'96–M'05–SM'06) received the B.S. *(magna cum laude)*, M.S., and Ph.D. degrees in electrical engineering from BYU, Provo, UT, USA, in 1997, 1998, and 2005, respectively.

From 1998 to 2004, he was with Motorola, Inc., Schaumburg, IL, USA, where he was involved in land mobile radio products. Since 2005, he has been with the Department of Electrical Engineering and Computer Science, The University of Kansas, Lawrence, where he is currently the Charles E. & Mary Jane Spahr Professor and the Department Chair. Since 2004, he has also been an Industry Consultant on problems, such as reduced-complexity receiver design, receiver synchronization, and error control coding. His current research interests include digital communication theory, synchronization, channel coding, energy-efficient communications, and complexity reduction in receivers. He is a member of the IEEE Aerospace and Electronic Systems Society and the IEEE Communications Society. He was the Chair of the Communication Theory Technical Committee, IEEE Communications Society, from 2017 to 2018. He was an Editor and an Area Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS from 2007 to 2018.



Liangping Ma (M'04–SM'12) received the B.S. degree in physics from Wuhan University, China, in 1998, and the Ph.D. degree in electrical engineering from the University of Delaware, USA, in 2004. He was with San Diego Research Center, Inc. from 2005 to 2007 and Argon ST, Inc. (acquired by Boeing) from 2007 to 2009, where he was the Principal Investigator of one project on wireless sensor networks from 2006 to 2008 and another on cognitive radios from 2007 to 2009, supported by the Department of Defense and National Science Foundation, respectively, and led the design and prototyping of a modem for a wireless sensor network. He joined InterDigital Communications in 2009, where he led research and development on Quality of Experience (QoE) based video communication and was a delegate to 3GPP RAN1 on low-latency communication for New Radio (NR). His current research interests include cellular radio access networks, video communication, and machine learning for communication. He was named an IEEE Communication Society Distinguished Lecturer covering 5G, video communication, and cognitive radios from 2017 to 2018 and from 2019 to 2020.