Generalized HARQ Protocols with Delayed Channel State Information and Average Latency Constraints

Trillingsgaard, Kasper Fløe; Popovski, Petar

Published in:
I E E E Transactions on Information Theory

DOI (link to publication from Publisher):
10.1109/TIT.2017.2651836

Creative Commons License
Unspecified

Publication date:
2018

Document Version
Early version, also known as pre-print

Link to publication from Aalborg University

Citation for published version (APA):
Trillingsgaard, K. F., & Popovski, P. (2018). Generalized HARQ Protocols with Delayed Channel State Information and Average Latency Constraints. *I E E E Transactions on Information Theory*, *64*(2), 1262-1280. [7814295]. https://doi.org/10.1109/TIT.2017.2651836

Generalized HARQ Protocols with Delayed Channel State Information and Average Latency Constraints

Kasper Fløe Trillingsgaard, *Student Member, IEEE*, and Petar Popovski, *Fellow, IEEE*

Abstract

In many practical wireless systems, the signal-to-interference-and-noise ratio (SINR) that is applicable to a certain transmission, referred to as channel state information (CSI), can only be learned after the transmission has taken place and is thereby delayed (outdated). In such systems, hybrid automatic repeat request (HARQ) protocols are often used to achieve high throughput with low latency. This paper put forth the family of expandable message space (EMS) protocols that generalize the HARQ protocol and allow for rate adaptation based on delayed CSI at the transmitter (CSIT). Assuming a block-fading channel, the proposed EMS protocols are analyzed using dynamic programming. When full CSIT is available and there is a constraint on the average decoding time, it is shown that the optimal EMS protocol has a particularly simple operational interpretation and that the throughput is identical to that of the backtrack retransmission request (BRQ) protocol. We also devise EMS protocols for the case in which CSIT is only available through a finite number of feedback messages. The numerical results demonstrate that BRQ approaches the ergodic capacity quickly compared to HARQ, while EMS protocols with only three and four feedback messages achieve throughput that are only slightly worse than the throughput of BRQ.

Index Terms

Hybrid Automatic Repeat Request (HARQ), delayed channel state information, low latency, backtrack retransmission request, dynamic programming

I. Introduction

Channel state information at the transmitter (CSIT) is essential for achieving high throughput in wireless systems. Preferably, CSIT is known before a transmission takes place, since in that case, the transmitter is able to optimize the parameters of the transmission, such as rate of the codebook and power. The transmitter may acquire an estimate of channel state information (CSI) in advance in various ways; for example, by using the channel reciprocity or via explicit feedback from the receiver. This is referred to as *prior CSIT*. However, a wireless channel is dynamic and, in many cases the channel changes from the time the CSI has been acquired

The work of K. F. Trillingsgaard and P. Popovski was supported in part by the European Research Council (ERC Consolidator Grant Nr. 648382 WILLOW) within the Horizon 2020 Program. The material of this paper was presented in part at the 2014 IEEE International Conference on Communication [1].

K. F. Trillingsgaard and P. Popovski are with the Department of Eletronic Systems, Aalborg University, 9220, Aalborg Øst, Denmark (e-mail: {kft,petarp}@es.aau.dk).

to the time at which the channel is actually used for transmission [2, pp. 211–213]. In addition, even if the channel is static, during the transmission there may be an unpredictable amount of interference at the receiver. In such cases, prior CSI is different, or even uncorrelated, with the actual conditions at the receiver when data transmission takes place and thus of limited, if any, use for adapting the transmission parameters. On the other hand, it is viable to assume that the transmitter gets feedback about the CSI *after* the data transmission has been made. We refer to this as *delayed CSIT* as it carries information to the transmitter about the conditions at the receiver in the past. The simplest form of delayed CSIT is the $1-$bit feedback used in automatic repeat request (ARQ) protocols: (ACK) the transmission was successful, i.e., the channel could support the chosen data rate and (NACK) the channel could not support the data rate. In the most elementary form of ARQ, a failed packet is retransmitted in the subsequent time slots until it is successfully decoded or until a strict decoding time constraint is violated. In order to increase throughput compared to ARQ, one can use chase combining (CC) or send incremental redundancy (IR) instead of retransmissions that consist of pure packet repetition. Such extensions are referred to as HARQ-CC and HARQ-INR, respectively [3]. In this paper, we focus on IR-based protocols.

The ergodic capacity represents an upper bound on the throughput for any communication protocol and can be approached by fixed-length coding across many time slots. HARQ-type protocols attempt to get as close as possible to this upper bound while keeping the average or maximum decoding time as small as possible. Specifically, as the rate R, which is used in the first transmission opportunity, tends to infinity, the average decoding time of HARQ-INR also tends to infinity and the throughput of HARQ-INR approaches the ergodic capacity of the underlying channel provided that there is no strict constraint on the decoding time. If a strict or average decoding time constraint is present, the achievable throughput is strictly lower than the ergodic capacity.

The purpose of this paper is to put forth and investigate a type of retransmission protocol which is fundamentally different from conventional HARQ protocols and uses rate adaptation based on delayed CSIT to achieve high throughput subject to an *average decoding time constraint*. As with most prior work in the area of HARQ-INR, we assume the channel is modeled by a Gaussian block-fading channel, with each time slot consisting of n channel uses. The channel gain is kept constant during a single time slot but varies independently from time slot to time slot. Feedback, such as delayed CSIT or acknowledgements (ACKs), can only be received by the transmitter at the end of each time slot. The main problem with an HARQ-INR protocol for a block-fading channel is that resources are wasted when the receiver sends NACK, while it only needs a small amount of additional information to be able to decode. This results in under-utilization of the last time slot and may significantly reduce the throughput when the average decoding time is small. Our key idea is to append *new information bits* in each time slot such that the last time slot is rarely under-utilized and the throughput degradation is reduced. We achieve this by using delayed CSIT which allows the transmitter to estimate the amount of unresolved information at the beginning of each time slot.

A. Prior work

Caire and Tuninetti [3] were among the first who analyzed HARQ from an information-theoretic perspective. Here, the throughput measure was defined through the renewal-reward theorem (see also [4] and [5]) and achievability and converse results were proved for the HARQ-INR protocol. Several lines of works has since improved the throughput of HARQ-INR by using available side information in combination with either power adaptation or rate adaptation.

One line of work uses power or rate adaptation to enhance the throughput of HARQ-INR with either prior or no CSIT. For example, [6] investigates HARQ-INR protocols that maximize the throughput over a block-fading channel with independent channel gains under both a strict decoding time constraint and a long-term power constraint. The long-term power constraint allows the use of slot-based power allocation. It is found that HARQ-INR in combination with slot-based power allocation increases the throughput. The key idea is that the probability of having to retransmit m times is decreasing in m. This implies that the throughput is increased by using more power in the first slots. In addition, it is shown that if the single feedback bit is used to convey a one-bit quantization of the prior CSI rather than an ACK/NACK message, then this can result in significant throughput gains. The results from [6] are further extended to any number of feedback bits per slot in [7]. Under the same channel conditions, [8] considers rate adaptation for an HARQ-INR protocol without prior nor delayed CSIT. Dynamic programming is used to maximize the throughput under an outage constraint and it is found that rate adaptation provides significantly lower outage probabilities. The assumption of independent channel gains is relaxed in [9], where optimal rate adaptation policies are found for the cases in which the channel gains are correlated.

Although prior CSIT improves the throughput of HARQ-INR remarkably, CSIT is often delayed when it is obtained by the transmitter. This has led to another line of work which studies the benefits of delayed CSIT in context of HARQ-INR protocols. Specifically, [10] and [11] considers a point-to-point channel with independent block-fading in a setting identical to ours. Apart from the statistics of the channel gain, the transmitter has no knowledge about the current CSI, but the transmitter is informed about the CSI of the previous slot. In their protocol, the channel uses of each slot are divided among a large number of parallel HARQ-INR instances transmitting separate messages in a TDMA fashion. In particular, for a specific HARQ-INR instance, the number of channel uses used for the kth retransmission is some percentage $0 \leq \ell_k \leq 1$ of the number of channel uses spend in the first transmission. This implies that new HARQ-INR instances, with new data, can be initiated in each slot. The objective is to maximize the throughput under a constraint on the outage probability. It is found that delayed CSIT significantly decreases the outage probabilities. A similar setting was considered in [12], where power adaptation was investigated. Here, the authors used a conventional HARQ-INR instance, but adapted the power in each slot according to the delayed CSIT. In contrast to [10], in which the authors design composite protocols based on a large number of HARQ-INR instances, the protocol proposed in [12] only uses a single HARQ-INR instance with power adaptation which is optimized using dynamic programming. Rate adaptation can also be achieved using superposition coding. A multi-layer broadcast approach to fading channels without prior CSIT is proposed in [13]. Specifically, a transmission is initiated in large number of superposition coded layers and the number of decoded layers at the receiver depends on the actual CSI which is assumed not to be known in advance. This approach provides an alternative to HARQ protocols in the sense that it provides variable-rate transmission with a fixed transmission length of one slot. The approach, however, has the disadvantage that the throughput in practical implementations suffer as the number of layers increases. A more practical approach is taken in [14] which combines the approach in [13] for few layers with HARQ-INR. Specifically, the proposed protocols initiate an HARQ-INR instance in each layer. In a certain slot, the receiver feeds back the number of decoded layers and, in the subsequent slot, the transmitter only conveys IR for the layers not decoded. For the layers that are decoded, the transmitter initiates new HARQ-INR instances with new data. Finally, although not directly related to our work, it was shown in [15] that delayed CSIT, which is possibly completely independent of the current channel state, increases the multiplexing gains in a multiple-input multiple-output (MIMO) broadcast channel with K transmit antennas and K receivers each with one receive antenna.

In contrast to previous works, this paper is motivated by the *backtrack retransmission request (BRQ)* protocol proposed in [1]. BRQ is suited for systems in which the transmission opportunities come in slots of a predefined number of channel uses. This prevents conventional HARQ-INR to optimize the throughput, as the number of channel uses cannot be adapted to the required amount of IR. BRQ overcomes this problem by appending additional new information bits before the information bits sent in previous slot have been decoded. The number of new information bits is adapted according to the reported delayed CSIT. Our approach in this paper combines the idea of appending new data during a transmission for HARQ in [1], [10], and [14] with streaming codes proposed in [16] and [17]. The streaming codes in [16] and [17] are a family of codes that allow the transmitter to append new information bits during a transmission in such a way that all information bits can be jointly decoded as one code. In [16], each message has the same absolute deadline at which all messages need to be decoded. In [17], each message is required to be decoded within a certain number slots after arrival. Both [16] and [17] use a transmission scheme that enlarges the message space in each slot. In coding theory, streaming codes, as those investigated in [16] and [17], are also known as cross-packet codes. Cross-packet codes based on Turbo codes and LDPC codes have previously been considered in the context of HARQ in [18] and [19], respectively. The EMS protocols proposed in this paper extend streaming codes to an HARQ-INR setting in which the amount of new information bits that are appended within a retransmission is adaptive, as it depends on the delayed CSIT in manner similar to BRQ.

EMS protocols are thus variable-rate protocols in a sense similar to [10] and [14]. However, to the best of our knowledge, all previously proposed protocols that allow for rate adaptation are essentially based on a conventional HARQ-INR protocol as building block, while the rate adaptation is achieved by using a large number of parallel HARQ-INR instances in a TDMA fashion or in superposition coded layers. These approaches incur rate penalties in practical implementations because each HARQ-INR instance only uses a small fraction of the available resources (channel uses/power) in each slot [20]. In contrast, EMS protocols differ fundamentally from HARQ-INR in the way new information bits are appended in each slot. This implies that, in principle, one can use our scheme instead of HARQ-INR as a building block and devise protocols similar to [10] and [14]. Consequently, we consider HARQ-INR and HARQ-INR with power adaptation based on delayed CSIT as relevant baseline protocols for comparison.

B. The backtrack retransmission protocol

Next, we shall introduce BRQ through an example. Suppose the transmitter sends to the receiver in slots, where each slot is a fixed communication resource that consists of n channel uses. The SNRs from slot to slot are i.i.d. and available to the transmitter only after transmissions have taken place. Suppose that the transmitter uses the same rate-R codebook in each slot and let us observe three consecutive slots with SNRs equal to

H_1, H_2, and H_3. Let

$$C(H_t) \triangleq \frac{1}{2}\log_2(1 + H_t) \tag{1}$$

denote the supported information rate during the ith slot when the SNR is equal to H_t, $t = 1, 2, 3$. Suppose that the realized SNR values are such that

$$C(H_1) \leq R, \quad C(H_2) \leq R, \quad C(H_3) > R. \tag{2}$$

Since the transmitter uses a rate-R code in each slot, the receiver is unable to decode in the first two slots and, eventually, the transmission can be decoded in the third slot. The transmitter sends $b \triangleq nR$ new information bits in the first slot, denoted by $\mathbf{d}_1$. For the second slot, the transmitter knows H_1 and sends again $b \triangleq nR$ information bits, but the message is prepared in the following way:

- The first $r_1 \triangleq n(R - C(H_1))$ bits are IR, denoted by $\mathbf{r}_1$ and correspond to the "missing information" from time slot 1. The bits $\mathbf{r}_1$, combined with the signal received and buffered in slot 1, allows the receiver to recover the information bits $\mathbf{d}_1$.
- The remaining $b - r_1 = nC(H_1)$ bits, denoted by $\mathbf{d}_2$ are *new information bits* transmitted in the second slot.

It must be noted that the IR bits and the new bits are only separable in a digital domain, but not at the physical layer, i.e., the whole packet, sent at rate R, needs to be decoded correctly and then the receiver extracts the IR bits and the new bits. The insertion of IR bits creates dependency between two adjacent packet transmissions.

Coming back to the example, outage also occurs in the second slot, such that the message in the third slot consists of $r_2 = n(R - C(H_2))$ redundancy bits $\mathbf{r}_2$ and $d_3 = nC(H_2)$ new bits $\mathbf{d}_3$. As $C(H_3) > R$, the receiver decodes the message in slot 3 and recovers $\mathbf{d}_3$ and $\mathbf{r}_2$. It then uses $\mathbf{r}_2$ as IR to decode the transmission in slot 2 and recover $\mathbf{d_2}$ and $\mathbf{r}_1$. Finally, the receiver uses $\mathbf{r}_1$ to decode the transmission from slot 1 and recover $\mathbf{d}_1$.

The throughput is given by

$$\bar{R} = \frac{nR + nC(H_1) + nC(H_2)}{3n} = \frac{R + C(H_1) + C(H_2)}{3}. \tag{3}$$

The average rate that could have been achieved over the three slots if the CSIT were known a priori is:

$$\bar{R}_{\text{prior}} = \frac{C(H_3) + C(H_1) + C(H_2)}{3}. \tag{4}$$

It is observed that the transmission with delayed CSIT introduces loss due to low R, but besides that, the data rates that are achieved with prior CSIT can be recovered by BRQ for the case of delayed CSIT. We remark that not all the parameters that are adaptable to the channel can be recovered with delayed CSIT. For example, power that has been erroneously allocated to a bad channel cannot be recovered by the delayed CSIT.

The above example considered the case where the conditions in (2) were satisfied. The protocol, which is referred to as BRQ and put forth in [1], is easily generalized to any realization of $\{H_t\}_{t\in\mathbb{N}}$. As in the example, the protocol in general uses a code with blocklength n and a fixed rate R in each slot. In the first slot, the transmitter sends nR bits of new information. If the realized channel gain H_1 satisfies $C(H_1) > R$, the receiver decodes the packet, extracts the nR information bits, and the protocol terminates with a decoding time of one slot. On the other hand, if $C(H_1) \leq R$, the receiver cannot decode the packet, it feeds back the CSI of the slot that has terminated, and the protocol continues. Considering the kth slot, with $k \geq 2$ and assuming that $C(H_t) \leq R$ for all $i \in \{1, \cdots, k-1\}$, the transmitter forms the packet for the kth slot as follows:

1) The first $n(R - C(H_{k-1}))$ bits are IR that allows the decoding of the packet in slot $(k-1)$.
2) The remaining $nC(H_{k-1})$ bits are new information bits.

If $C(H_k) \leq R$, the receiver feeds back the CSI of the slot and the protocol continues. If $C(H_k) > R$, the receiver can decode the packet in slot k in which case it recovers the $nC(H_{k-1})$ information bits. It also recovers the $n(R - C(H_{k-1}))$ bits of IR for the packet in slot $(k-1)$. At this time, the receiver can decode the $(k-1)$th packet using the side information from the IR bits in slot k. This can be achieved using list decoding. Next, the decoder sequentially decodes the packets $(k-2), (k-3), \cdots, 1$ in a similar fashion, thereby recovering all the $n(R + C(H_1) + \cdots + C(H_{k-1}))$ bits. The throughput of BRQ, reported in [1], is restated in Theorem 3.

From the example above, we observe that BRQ relies on appending information bits to the parity bits. The transmission rate used in BRQ is predefined to be R in each slot. The number of appended information bits is computed based on delayed CSIT but chosen such that the a priori probability of decoding a certain slot is kept constant. Hence, a BRQ process ends a transmission as soon as the CSI is above a level that is sufficient for decoding the predefined rate R.

C. Contribution

In this paper, we generalize the BRQ protocol from [1]. First, we propose a family of EMS protocols that allow the transmitter to expand the message space in manner similar to BRQ. In contrast to BRQ, however, the EMS protocols are based on streaming codes and all information bits are decoded jointly. The notion of an EMS protocol introduced here is sufficiently general to include protocols like ARQ, HARQ-INR, and BRQ. Next, we prove a strong converse and an achievability result for the EMS protocols, and it is shown that the throughput of the optimal EMS protocol given a constraint on the average decoding time and full delayed CSIT is identical to the throughput of BRQ. Then, we address the same problem with only a finite number of feedback messages in each slot. In this case, we put forth heuristic EMS protocols which have a structure similar to BRQ, but are designed to work with finite number of feedback messages. Finally, the throughput of BRQ and the proposed finite feedback EMS protocols are evaluated and compared to relevant baseline protocols. Specifically, we compute the throughput in terms of SNR and average decoding time. Our numerical results confirm that the throughput of BRQ converges to the ergodic capacity much faster than the throughput of HARQ-INR. Moreover, the proposed finite feedback EMS protocol using only three feedback messages per slot achieves throughput which is only slightly worse than the throughput of BRQ. We remark that EMS protocols have previously been introduced in [21], where we used finite blocklength analysis to investigate a protocol similar to BRQ, but for the binary symmetric channel and binary fading. For a similar setting, optimal rate adaptation policies were optimized using error exponents in [22].

Notation: Upper case, lower case, and calligraphic letters denote random variables, deterministic quantities, and sets, respectively. A tuple of random variables $(X_i, \cdots, X_j)$, $j \geq i$, is denoted by X_i^j. We adopt the convention that $\sum_{i=j}^{j-1} a_i = 0$ and likewise we let X_i^{i-1} denote the empty tuple. Let $\mathbb{N}$ be the natural numbers, $\mathbb{R}$ be the reals, $\mathbb{R}_+$ be the nonnegative reals, $\mathbb{Z}$ be the integers, and $\mathbb{Z}_+$ be the nonnegative integers. Moreover, the range of integers $\{i, \cdots, j\}$, $i \leq j$, is denoted by $[i{:}j]$. Vectors are denoted by boldface (e.g., $\mathbf{a}$), while their entries are denoted by roman letters (e.g., a_i). The transpose of a vector $\mathbf{a}$ is denoted by $\mathbf{a}^{\mathrm{T}}$, the length of a vector by $\mathrm{len}(\cdot)$, and the tuple $(a_i, \cdots, a_j)$, for $i \leq j$, is by a_i^j. We also use the standard asymptotic notation $f(n) = \mathcal{O}(g(n))$ and $f(n) = o(g(n))$ which means that $\limsup_{n\to\infty} |f(n)/g(n)| < \infty$ and that $\limsup_{n\to\infty} |f(n)/g(n)| = 0$, respectively. Finally, let $[x]^+ \triangleq \max\{x, 0\}$ and $[x]^- \triangleq \min\{x, 0\}$.

II. System Model

We consider a single-user block-fading channel with Gaussian noise. The transmitter sends to the receiver in slots of n channel uses, where n is sufficiently large to offer reliable communication that is optimal in an information-theoretic sense. The received signal vector in slot $t \in \mathbb{N}$ is given by

$$\mathbf{Y}_t = \sqrt{H_t}\mathbf{X}_t + \mathbf{Z}_t \tag{5}$$

where $\mathbf{Z}_t \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_n)$ is a n-dimensional noise vector distributed according to the Gaussian distribution with zero mean and identity covariance matrix, $\mathbf{X}_t$ is the transmitted n-dimensional vector satisfying

$$\frac{1}{n}\mathbf{X}_t^{\mathrm{T}}\mathbf{X}_t \leq 1 \tag{6}$$

and $H_t \geq 0$ denotes the instantaneous SNR, drawn independently from a smooth probability density $P_H(\cdot)$. The cumulative distribution function of H_t is given by $F_H(\cdot)$. The instantaneous SNR H_t is unknown at the transmitter prior to the transmission of $\mathbf{X}_t$ but is known at the receiver after observing $\mathbf{Y}_t$. Moreover, the receiver is able to provide feedback based on previously received CSI. Specifically, we assume that feedback is given by a sequence of feedback functions $\mathbb{v}_t : \mathbb{R}_+^t \to \mathbb{F}$ that maps H_1^t to a feedback alphabet $\mathbb{F}$ such that $V_t = \mathbb{v}_t(H_1^t)$ is observed at the transmitter before transmission in the $(t+1)$th slot. The *feedback cost* is defined as the cardinality of the feedback alphabet $|\mathbb{F}|$ and may be finite, countably infinite, or uncountably infinite. The transmitter is said to have full delayed CSIT if H_t can be recovered from V_t.

If a transmission is to be done over slot t alone, the maximum supported rate is given by $C(H_t)$, whereas the maximum achievable rate if a transmission is done over many slots approaches the ergodic capacity [23]

$$C_{\text{erg}} = \mathbb{E}[C(H)] \tag{7}$$

as the number of slots tends to infinity. Here, H denotes a random variable distributed according to P_H. If, however, a transmission is to be done over few slots, high throughput cannot be achieved without either layered transmissions as in [14] or a HARQ technique. The latter approach is commonly applied in practical systems due to its relative simplicity compared to the layered transmissions.

A comment on the block-fading assumption is in order. The block-fading channel model is an abstraction of a practical system model. In particular, if slots are transmitted consecutively in time as this model suggests, the channel gains cannot be assumed to be independent. In practical systems, however, the delay of ACK/NACK feedback can often spread over multiple slots in time. Therefore, in wireless systems such as LTE, multiple HARQ instances are interleaved in time [24, Ch. 12] — while the transmitter waits for feedback from one

HARQ instance, it transmits to other users. In the uplink in LTE, a synchronous version of HARQ is employed [24, Ch. 12]. This ensures that the time between each retransmission is fixed and known by both the transmitter and the receiver. The fact that each transmission opportunity is spaced apart by a certain number of slots implies that channel gains are close to independent. In addition to these considerations, one cannot expect that each transmission opportunity occurs in the same frequency slot; this further justifies the use of a block-fading model.

Next, we define an EMS protocol by the tuple $\left(\{\mathbb{v}_t\}, \{\mathbb{r}_t^{(n)}\}, \{\mathbb{f}_t^{(n)}\}, \{\mathbb{g}_t^{(n)}\}, \{\tau_n\}\right)$ described as follows:

- A sequence of feedback functions $\mathbb{v}_t : \mathbb{R}_+^t \mapsto \mathbb{F}$ that maps H_1^t to the feedback alphabet $\mathbb{F}$ such that

$$V_t \triangleq \mathbb{v}_t(H_1^t). \tag{8}$$

  Moreover, for $h_1^t = (h_1, \cdots, h_t) \in \mathbb{R}_+^t$, we denote the tuple $(\mathbb{v}_{t'}(h_1^{t'}), \cdots, \mathbb{v}_t(h_1^t))$ by $v_{t'}^t = \mathbb{v}_{t'}^t(h_1^t)$ for $t' \leq t$.
- A sequence of rate selection functions $\mathbb{r}_t^{(n)} : \mathbb{F} \mapsto \mathbb{R}_+$ that satisfy $R_t^{(n)} \triangleq \mathbb{r}_t^{(n)}(V_{t-1})$, $R_t^{(n+1)}(\cdot) \geq R_t^{(n)}$ for all $t \in \mathbb{N}$, and $R_t^{(n)} \leq \mathbb{r}_{\text{max}}$ for some positive constant $\mathbb{r}_{\text{max}}$. We also define the cumulative rates $\bar{R}_t^{(n)} \triangleq \sum_{k=1}^t R_k^{(n)}$.
- A sequence of encoding functions $\mathbb{f}_t^{(n)} : \mathfrak{B} \mapsto \mathbb{R}^n$ such that

$$\mathbf{X}_t \triangleq \mathbb{f}_t^{(n)}\left(B_1^{\lceil n\bar{R}_t^{(n)}\rceil}\right). \tag{9}$$

  Here, $\mathfrak{B}$ denotes all binary vectors (of arbitrary length), i.e., $\mathfrak{B} \triangleq \{[]\} \cup \bigcup_{i=1}^{\infty}\{0,1\}^i$, where $[]$ denotes the vector of length 0; and B_i are independent Bernoulli variables with parameter $1/2$ and we denote $(B_i, \cdots, B_j)$ by B_i^j.
- A sequence of decoding functions $\mathbb{g}_t^{(n)} : \mathbb{R}^{tn} \times \mathbb{R}_+^t \mapsto \mathfrak{B}$.
- A sequence of nonnegative integer-valued random variables $\{\tau_n\}_{n=1}^{\infty}$, which are stopping times with respect to the filtration $\mathcal{F}_t \triangleq \sigma\{V^t\}$ (see e.g. [25, p. 488]) and satisfy $\tau_{n+1} \geq \tau_n$ and $\sup \mathbb{E}[\tau_n] < \infty$.

We also define the limiting rate selection functions and the stopping time:

$$\mathbb{r}_t \triangleq \lim_{n\to\infty} \mathbb{r}_t^{(n)} \tag{10}$$

$$\tau \triangleq \sup_{n\to\infty} \tau_n. \tag{11}$$

The limit of $\mathbb{r}_t^{(n)}$ exists because $\mathbb{r}_i^{(n)}$ is non-decreasing and bounded above by $\mathbb{r}_{\text{max}}$. On the other hand, we define τ as the supremum over τ_n since the existence of the limit of τ_n cannot be guarenteed because only $\mathbb{E}[\tau_n]$ is bounded above for increasing n.

Note that one can feedback the rate selection directly and omit defining $\mathbb{v}_t(\cdot)$. The restriction on the feedback cost can be induced by restricting the number of distinct rate selections that can be fed back. We have made this restriction explicit by defining both $\mathbb{r}_t(\cdot)$ and $\mathbb{v}_t(\cdot)$.

The random variables $B^{\infty} \in \{0,1\}^{\infty}$ correspond to the binary sequence of information bits, which size in bits is unbounded. We assume that all the information bits are available prior to the transmission in the first slot. This implies that the stopping time τ_n is also the decoding time and the transmission time in slots. In remainder of this paper, we shall refer to τ_n as a decoding time. We note that our definition of decoding time deviates from some other works. For example, in [8] and [10], the decoding time is measured as the time from the information bits are appended to the time at which they are decoded. Such definition of decoding time is relevant in some real-time applications such as control over network.

As an implication of the definition of an EMS protocol, $\mathbf{X}_t$ becomes a function of the block of information bits $B_1^{\lceil n\bar{R}_t^{(n)}\rceil} = (B_1, \cdots, B_{\lceil n\bar{R}_t^{(n)}\rceil})$. This enables the encoder to combine IR and new information bits, i.e., in each slot the encoder fetches $nR_t^{(n)}$ information bits and encodes them jointly with the previously encoded $n\bar{R}_{t-1}^{(n)}$ information bits. This message structure is different from other works on HARQ-INR protocols. In light of [26], HARQ-INR can be seen as fixed-to-variable coding because the number of transmitted information bits are prespecified while the number of channel observations at the receiver depends on channel realization. On the other hand, for an EMS protocol, both the number of information bits and the number of channel observations depend on the channel realization. This concept has previously been used in [10] and [14]; however, none of these works alter the conventional HARQ-INR protocol. They rather use it as a building block and initiate a large number of HARQ-INR instances which run in parallel in a TDMA fashion or in multiple superposition coded layers.

Following other HARQ works [3], [5], [14], we define the throughput η of an EMS protocol in terms of a renewal-reward process. A renewal event occurs at time τ_n and the reward is the sum of all rates appended since time 1. Likewise, the inter-renewal time corresponds to the decoding time τ_n. Hence, we define the throughput of an EMS protocol as $\lim_{n\to\infty} \mathbb{E}\big[\bar{R}_{\tau_n}^{(n)}\big]/\mathbb{E}[\tau_n]$. This leads us to the definition of a zero outage EMS protocol.

Definition 1: An EMS protocol is called an (η, T)-zero outage EMS protocol if there exists a non-decreasing integer-valued sequence $\{\bar{\tau}_n\}$ such that $\tau_n \leq \bar{\tau}_n$, $\mathbb{E}[\tau_n] \leq T$, $\lim_{n\to\infty} \mathbb{E}\left[\bar{R}_{\tau_n}^{(n)}\right]/\mathbb{E}[\tau_n] \geq \eta$,

$$\lim_{n\to\infty} \mathbb{P}[\mathcal{E}_n] = 0 \tag{12}$$

and

$$\lim_{n\to\infty} \max_{\substack{t\in[1:\bar{\tau}_n-1]:\\ \mathbb{P}[\tau_n=t]>0}} \mathbb{P}[\mathcal{E}_n|\tau_n = t] = 0. \tag{13}$$

Here, we have defined the error event

$$\mathcal{E}_n \triangleq \left\{ \mathbb{g}_{\tau_n}^{(n)}(\mathbf{Y}_1^{\tau_n}, H^{\tau_n}) \neq B_1^{\lceil n\bar{R}_{\tau_n}\rceil} \right\}. \tag{14}$$

Our focus is on the characterization of optimal zero outage EMS protocols:

$$\eta_{\text{opt}}(T) \triangleq \sup\{\eta : \text{there exists an } (\eta, T)\text{-zero outage EMS protocol}\}. \tag{15}$$

The condition (12) ensures that the outage probability of the EMS protocol is zero while the condition in (13) ensures that the conditional probability of error given a decoding time vanishes uniformly for all decoding times except for $\bar{\tau}_n$.

We note that most other HARQ works solely consider strict latency constraints which naturally arise in wireless communication systems having either a strict deadline or a limited buffer size. We consider average decoding time constraints and zero outage protocols for two reasons:

- A strict latency constraint does not naturally arise in systems without a strict deadline or limited buffer size, and hence, in such applications, there is no reason to choose a specific deadline T in the strict decoding time constraint $\mathbb{P}[\tau \leq T] = 1$. For example, consider an application that requires a high reliability. In this case, imposing a strict latency constraint for the HARQ protocol only implies that the receiver will request a retransmission of the data in outage. This is the case for LTE, which uses HARQ in the medium access control (MAC) layer, while it implements an ARQ protocol on a higher layer – in the radio link control (RLC) layer – that requests retransmissions for data in outage [24, Ch. 12]. In that sense, LTE attempts to achieve an outage probability of zero, and an average decoding time constraint is therefore a natural constraint which attempts to keep the decoding time low on average but does not give any strict guarantees. As previously mentioned, LTE employs synchronous HARQ in the uplink which implies that the decoding time τ is indeed proportional to real decoding time in a system. We also point out that the customary metric for latency in queuing theory is the average waiting time.

- It turns out that the throughput of the optimal EMS protocol, under an average decoding time constraint, coincides with the throughput of the BRQ protocol proposed in [1], i.e., the optimization problem in (15) has a simple form.

III. Strong Converse and Achievability

In this section, we state a strong converse and an achievability result that we shall apply in the subsequent sections. The achievability and converse results state conditions for when the probability of error tends to zero or one, respectively. In order to state the results, we need to introduce some notation. In particular, given rate selection functions and feedback functions, let

$$\mathbb{u}_{k,t}^{(n)}(h_1^t) \triangleq \sum_{i=k}^{t} \left[\mathbb{r}_i^{(n)}(\mathbb{v}_{i-1}(h_1^{i-1})) - C(h_i) \right] \tag{16}$$

for $k \leq t$ and let $\mathbb{u}_{k,t}^{(n)}(\cdot) \triangleq 0$ for $t < k$. Intuitively, $\mathbb{u}_{1,t}^{(n)}(h_1^t)$ is the remaining amount of information needed to decode the information bits up to time t given slot observations up to time t with channel gains $h_1^t = (h_1, \cdots, h_t)$. We also define

$$\mathbb{u}_{k,t}(h_1^t) \triangleq \lim_{n \to \infty} \mathbb{u}_{k,t}^{(n)}(h_1^t) = \sum_{i=k}^{t} \left[\mathbb{r}_i(\mathbb{v}_{i-1}(h_1^{i-1})) - C(h_i) \right]. \tag{17}$$

We prove the following results in Appendix I and Appendix II.

Lemma 1 (Strong converse): Given an EMS protocol, we have

$$\lim_{n \to \infty} \mathbb{P}\left[\mathcal{E}_n \middle| H^\infty = h^\infty \right] = 1 \tag{18}$$

for all every h^∞ such that $\sup_{k \in [1:\tau]} \mathbb{u}_{k,\tau}(h^\tau) > 0$ and such that $\tau < \infty$.

Remark 1: The conditions in Lemma 1 are only given in terms of the asymptotic random variables τ and $\mathbb{r}_t$ and not τ_n and $\mathbb{r}_t^{(n)}$. Therefore, Lemma 1 allows us to restrict the search for optimal zero outage EMS protocols to those EMS protocols for which $\sup_{k \in [1:\tau]} \mathbb{u}_{k,\tau}(h^\tau) \leq 0$ almost surely (see details in the converse proof of Theorem 3).

Remark 2: The smallest limiting decoding time which is not ruled out by Lemma 1 is given by

$$\tau_{\text{opt}} \triangleq \inf\big\{t \geq 1 : \mathbbm{u}_{1,t}(H_1^t) \leq 0\big\}\,. \tag{19}$$

To show that an EMS protocol with $\tau = \tau_{\text{opt}}$ is not ruled out by Lemma 1, note that by the definition of τ_{opt}, we must have

$$\mathbbm{u}_{1,1}(H_1^1) > 0, \cdots, \mathbbm{u}_{1,\tau_{\text{opt}}-1}(H_1^{\tau_{\text{opt}}-1}) > 0 \text{ and } \mathbbm{u}_{1,\tau_{\text{opt}}}(H_1^{\tau_{\text{opt}}}) \leq 0. \tag{20}$$

Thus, using the fact that $\mathbbm{u}_{k,\tau_{\text{opt}}}(H_1^{\tau_{\text{opt}}}) = \mathbbm{u}_{1,\tau_{\text{opt}}}(H_1^{\tau_{\text{opt}}}) - \mathbbm{u}_{1,k-1}(H_1^{k-1}) \leq 0$ for every $k \in [1{:}\tau_{\text{opt}}]$, we find that the conditions in Lemma 1 are not satisfied.

Lemma 2 (Achievability): Let $\{\tau_n\}$, rate selection functions $\{\mathbbm{r}_t^{(n)}\}$, and feedback functions $\{\mathbbm{v}_t\}$ be given. Suppose that there exist positive sequences c_n, g_n, and $\bar{\tau}_n$ such that $\tau_n \in \mathbb{N}$, $\tau_n \leq \bar{\tau}_n$, and

$$\frac{\bar{\tau}_n^2}{n g_n c_n^2} \to 0 \tag{21}$$

as $n \to \infty$. Moreover, define the event

$$\bar{\mathcal{H}}_n \triangleq \left\{ \max_{k \in [1:\tau_n]} \mathbbm{u}_{k,\tau_n}^{(n)}(H^{\tau_n}) \leq -c_n \right\} \tag{22}$$

and assume for all sufficiently large n that

$$\min_{\substack{t \in [1:\bar{\tau}_n]:\\ \mathbb{P}\left[\tau_n = t | \bar{\mathcal{H}}_n\right] > 0}} \mathbb{P}\big[\tau_n = t | \bar{\mathcal{H}}_n\big] \geq g_n. \tag{23}$$

Then, there exists an EMS protocol satisfying

$$\lim_{n \to \infty} \max_{\substack{t \in [1:\bar{\tau}_n]:\\ \mathbb{P}[\tau_n = t] > 0}} \mathbb{P}\Big[\mathcal{E}_n \Big| \bar{\mathcal{H}}_n, \tau_n = t\Big] = 0. \tag{24}$$

IV. Full Delayed CSIT

In this section, we consider the case in which the feedback alphabet is the positive reals, $\mathbb{F} = \mathbb{R}$, and the feedback functions are given by

$$\mathbbm{v}_t(h_1^t) \triangleq h_t. \tag{25}$$

This provides the transmitter with full delayed CSIT. In the following, we characterize the trade-off between throughput and the average decoding time for optimal zero outage EMS protocols. First, we specify an EMS protocol and we shall later show that it is an optimal zero outage EMS protocol. We specify the EMS protocol as follows

$$\mathbbm{r}_t^{(n)}(v) \triangleq \begin{cases} C(h_T) - \frac{c_1}{\log n} & t = 1 \\ \min\Big\{C(v), C(h_T) - \frac{c_1}{\log n}\Big\} & t \geq 2 \end{cases} \tag{26}$$

for $t \in \mathbb{N}$, a positive constant h_T, and an arbitrary constant $c_1 > 0$. The decoding times are given by

$$\tau_n \triangleq \min\{\bar{\tau}_n, \tau\} \tag{27}$$

where

$$\bar{\tau}_n \triangleq -\left\lfloor \frac{\log(c_2\sqrt{n})}{\log F_H(h_T)} \right\rfloor \tag{28}$$

$$\tau \triangleq \inf\{t \geq 1 : h_T < H_t\} \tag{29}$$

for an arbitrary constant $c_2 > 0$. The particular choice of the rate selection functions has a simple operational interpretation when neglecting the vanishing term $c_1/\log n$. Consider a transmitter using a fixed-rate codebook with rate $C(h_T)$ in each slot such that the minimal SNR required to decode a slot is h_T. Based on the delayed CSI, in slot t, the transmitter sends the exact amount of IR that is required to decode the previous packet, i.e., $n(C(h_T) - C(H_{t-1}))$ bits, along with $nC(H_{t-1})$ bits of new information bits. This protocol resembles the BRQ protocol previously described in Section I-B but formulated as an EMS protocol.

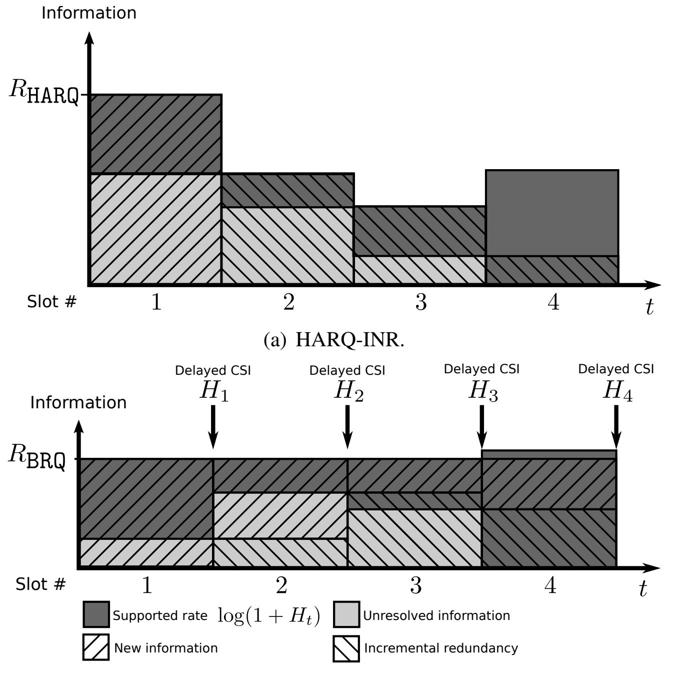
The operation of BRQ is illustrated and compared to HARQ-INR in Fig. 1. Initially, HARQ-INR transmits at a rate R_{HARQ}. The receiver accumulates information until the amount of unresolved information reaches zero. BRQ starts the transmission at a rate R_{BRQ}, but in contrast to HARQ-INR, it uses the delayed CSI to append new information bits in each slot to ensure that the amount of unresolved information, before the receiver observes $\mathbf{Y}_t$ and H_t, remains R_{BRQ}. Note that, in order to attain the same average decoding time for BRQ and HARQ-INR, R_{BRQ} needs to be chosen smaller than R_{HARQ} since no additional information bits are appended during transmission in the HARQ-INR protocol. This is why we have chosen $R_{\text{HARQ}} > R_{\text{BRQ}}$ in the figure. For the particuar realization of channel gains depicted in Fig. 1, it is seen that HARQ-INR does not fully utilize the supported rate since the unresolved information, before $\mathbf{Y}_4$ and H_4 are observed, is significantly smaller than the supported rate in that slot. This phenomenon reduces the throughput at low average decoding times. The problem is partially circumvented in BRQ by ensuring that the amount of unresolved information, before $\mathbf{Y}_t$ and H_t are observed, is kept constant.

In contrast to BRQ, the EMS protocol specified by (25) and (26) uses the coding scheme described in the achievability part of Appendix I and thereby uses joint decoding over all slots. Since the EMS protocol specified by (25) and (26) and BRQ are closely related, we shall refer to the proposed EMS protocol as “BRQ-EMS” to emphasize its relation to BRQ.

The following result analyzes the trade-off between throughput and average decoding time of BRQ-EMS. Specifically, we find that the throughput is identical to that of BRQ. Furthermore, we apply the strong converse in Lemma 1 and we demonstrate using dynamic programming that BRQ-EMS is optimal within the class of EMS protocols.

Theorem 3: For $T > 1$, we have

$$\eta_{\text{opt}}(T) \geq \eta_{\text{BRQ}}(T) \triangleq \int_0^{h_T} P_H(h)C(h)\,\mathrm{d}h + \frac{C(h_T)}{T} \tag{30}$$



(a) HARQ-INR.

(b) BRQ.

Fig. 1. Comparison between HARQ-INR and BRQ. In slot t, the left and right striped areas corresponds to the amount of unresolved information before receiving $\mathbf{Y}_t$. The dark grey areas designate the instantaneous supported rate and the light grey areas corresponds to the unresolved information after observing $\mathbf{Y}_t$. Note that for each time slot, the dark grey areas have the same size for both HARQ-INR and BRQ.

where

$$h_T \triangleq F_H^{-1}\left(1-\frac{1}{T}\right). \tag{31}$$

Moreover, $\eta_{\text{BRQ}}(T) = \eta_{\text{opt}}(T)$ if

$$\frac{P_H(h)}{1-F_H(h)} + \frac{1}{(1+h)} + \frac{P'_H(h)}{P_H(h)} \geq 0 \tag{32}$$

for all $h \geq 0$.

Remark 3: The throughput of BRQ, which is identical to (30), was first reported in [1].

Remark 4: One can verify that (32) is satisfied for the Rayleigh fading distribution $P_H(h) = \frac{1}{\Gamma}\mathrm{e}^{-h/\Gamma}$ for all $\Gamma > 0$. Indeed, the LHS of (32) yields $(1+h)^{-1}$ which is nonnegative for all $h \geq 0$.

Remark 5: It follows directly from (30) that $\eta_{\text{BRQ}}(T) \to C_{\text{erg}}$ as $T \to \infty$. This follows because $h_T \to \infty$ as $T \to \infty$, and thus the first term in (30) tends to C_{erg} while the second term in (30) tends to zero.

Remark 6: The second term on the RHS of (30) is the throughput of the conventional ARQ protocol with a rate equal to $C(h_T)$. The first term on the RHS of (30) thereby corresponds to the improvement of BRQ-EMS over ARQ.

Proof: We shall first use Lemma 2 to show that there exists an $(\eta_{\text{BRQ}}(T), T)$-zero outage EMS protocol with rate selection and feedback functions given by (25) and (26), respectively. Then, we apply the strong converse in Lemma 1 to show that $\eta_{\text{opt}}(T) = \eta_{\text{BRQ}}(T)$ under the condition in (32).

Fix positive constants c_1 and c_2. We first show that an EMS protocol specified by (25)–(27) has throughput $\eta_{\text{BRQ}}(T)$ and average decoding time upper bounded by T. Since $\{\tau_n\}$ is a non-decreasing sequence of random variables and since $\mathbb{E}[\tau_n] \leq \mathbb{E}[\tau] < \infty$, Lebesgue's monotone convergence theorem [25, Th. 16.2] implies that

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}[\tau_n] &= \mathbb{E}[\tau] \qquad (33)\\
&= \sum_{i=1}^{\infty} i F_H(h_T)^{i-1} \left(1 - F_H(h_T)\right) \qquad (34)\\
&= \frac{1}{1 - F_H(h_T)} \qquad (35)\\
&= T \qquad (36)
\end{aligned}$$

Similarly, we also have that

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\left[\bar{R}^{(n)}_{\tau_n}\right] &= \lim_{n\to\infty} \mathbb{E}\left[\sum_{t=1}^{\infty} \mathbb{1}\{\tau_n \geq t\}\, \mathbb{r}_t^{(n)}\big(\mathbb{v}_{t-1}\big(H_1^{t-1}\big)\big)\right] \qquad (37)\\
&= \mathbb{E}\left[\sum_{t=1}^{\infty} \lim_{n\to\infty} \mathbb{1}\{\tau_n \geq t\}\, \mathbb{r}_t^{(n)}\big(\mathbb{v}_{t-1}\big(H_1^{t-1}\big)\big)\right] \qquad (38)\\
&= C(h_T) + \mathbb{E}\left[\sum_{t=2}^{\infty} \min\{C(H_{t-1}), C(h_T)\}\, \mathbb{1}\{\tau \geq t\}\right] \qquad (39)\\
&= C(h_T) + \sum_{t=2}^{\infty} \mathbb{E}[\min\{C(H_{t-1}), C(h_T)\}\, \mathbb{1}\{\tau \geq t\}] \qquad (40)\\
&= C(h_T) + \mathbb{E}[C(H)|H \leq h_T] \sum_{t=2}^{\infty} \mathbb{P}[\tau \geq t] \qquad (41)\\
&= C(h_T) + \mathbb{E}[C(H)|H \leq h_T]\, (\mathbb{E}[\tau] - 1) \qquad (42)\\
&= C(h_T) + T \int_0^{h_T} P_H(h) C(h) \mathrm{d}h \qquad (43)
\end{aligned}$$

Here, (38) follows from Lebesgue's monotone convergence theorem [25, Th. 16.2] because τ_n and $\mathbb{r}_t^{(n)}$ are non-decreasing in n and bounded above by $\bar{\tau}_n$ and $\mathbb{r}_{\text{max}}$, respectively. Moreover, (40) follows from Tonelli's theorem, and the last the equation follows because

$$\int_0^{h_T} P_H(h) C(h) \mathrm{d}h = \mathbb{E}[C(H)|H \leq h_T]\, \mathbb{P}[H \leq h_T] \qquad (44)$$

and because $T = \mathbb{E}[\tau] = 1/\mathbb{P}[H \geq h_T]$. Using (36) and (43), the throughput is given by

$$\lim_{n\to\infty} \frac{\mathbb{E}\big[\bar{R}^{(n)}_{\tau_n}\big]}{\mathbb{E}[\tau_n]} = \int_0^{h_T} P_H(h) C(h) \mathrm{d}h + \frac{C(h_T)}{T} = \eta_{\text{BRQ}}(T). \qquad (45)$$

To show the existence of an $(\eta_{\text{BRQ}}(T), T)$-zero outage EMS protocol, we only need to demonstrate that BRQ-EMS satisfies (12) and (13). Both of these conditions follow from (24) if the conditions of Lemma 2 can be verified. We shall first show that $\tau \leq \bar{\tau}_n$ implies that $\max_{k\in[1:\tau_n]} \mathbb{u}^{(n)}_{k,\tau_n}(H_1^{\tau_n}) \leq -c_n$, i.e., that

$$\mathbb{P}\big[\bar{\mathcal{H}}_n | \tau_n = t\big] = 1 \qquad (46)$$

for $t \in [1{:}\bar{\tau}_n - 1]$, where $\bar{\mathcal{H}}_n$ is defined in (22). Because $\mathbb{u}^{(n)}_{1,t}(h^t) \leq \mathbb{u}_{1,t}(h^t)$ for every $h^t \in \mathbb{R}_+^t$ and because $\mathbb{u}_{1,\tau_n}(H^{\tau_n}) \leq 0$ when $\tau \leq \bar{\tau}_n$, this follows from

$$\mathbb{u}^{(n)}_{1,\tau_n}(H_1^{\tau_n}) \leq \mathbb{u}_{1,\tau_n}(H^{\tau_n}) - c_n \leq -c_n \qquad (47)$$

and from the following chain of inequalities[1]

$$\begin{aligned}
&\max_{k\in[2:\tau_n]} \mathbb{u}^{(n)}_{k,\tau_n}(H_1^{\tau_n}) \\
&= \max_{k\in[2:\tau_n]} \sum_{i=k}^{\tau_n} \Big[\min\Big\{C(h_T)-c_n, C(H_{i-1})\Big\} - C(H_i)\Big] && (48)\\
&\le \max_{k\in[2:\tau_n]} \Bigg\{ C(H_{k-1}) - C(h_T) \\
&\qquad + \sum_{i=k}^{\tau_n} \Big[\min\{C(h_T)-c_n, C(H_{i-1})\} - C(H_{i-1})\Big] \Bigg\} && (49)\\
&= \max_{k\in[2:\tau_n]} \Bigg\{ (C(H_{k-1}) + c_n - C(h_T)) - c_n \\
&\qquad + \sum_{i=k}^{\tau_n} \Big[C(h_T) - c_n - C(H_{i-1})\Big]^- \Bigg\} && (50)\\
&= \max_{k\in[2:\tau_n]} \Bigg\{ \Big[C(H_{k-1}) + c_n - C(h_T)\Big]^- - c_n \\
&\qquad + \sum_{i=k+1}^{\tau_n} \Big[C(h_T) - c_n - C(H_{i-1})\Big]^- \Bigg\} && (51)\\
&\le -c_n. && (52)
\end{aligned}$$

Here, (48) follows from (16) and (26), (49) follows because (29) implies that $h_T < H_{\tau_n}$ when $\tau \le \bar{\tau}_n$, and (52) follows because $[\cdot]^- \le 0$. Next, we show that $g_n \triangleq \mathcal{O}(1/\sqrt{n})$ satisfies (23):

$$\begin{aligned}
\min_{\substack{t\in[1:\bar{\tau}_n]:\\ \mathbb{P}[\tau_n=t|\bar{\mathcal{H}}_n]>0}} \mathbb{P}\big[\tau_n = t|\bar{\mathcal{H}}_n\big] &\ge \min_{\substack{t\in[1:\bar{\tau}_n]:\\ \mathbb{P}[\tau_n=t|\bar{\mathcal{H}}_n]>0}} \mathbb{P}\big[\tau_n = t, \bar{\mathcal{H}}_n\big] && (53)\\
&= \min_{\substack{t\in[1:\bar{\tau}_n]:\\ \mathbb{P}[\tau_n=t|\bar{\mathcal{H}}_n]>0}} \mathbb{P}[\tau = t] && (54)\\
&= F_H(h_T)^{\bar{\tau}_n - 1}(1 - F_H(h_T)) && (55)\\
&= \mathcal{O}(\mathrm{e}^{\log(F_H(h_T))\bar{\tau}_n}) && (56)\\
&= \mathcal{O}\bigg(\frac{1}{\sqrt{n}}\bigg) = g_n. && (57)
\end{aligned}$$

It also follows that (21) is satisfied:

$$\frac{\bar{\tau}_n^2}{n g_n c_n^2} = \mathcal{O}\bigg(\frac{\log^2(n)\log^2(\sqrt{n})}{\sqrt{n}}\bigg) = o(1). \qquad (58)$$

As a consequence of (57) and (58), Lemma 2 implies that there exists an EMS protocol satisfying (24). In addition, the EMS protocol is also an $(\eta_{\mathrm{BRQ}}(T), T)$-zero outage EMS protocol, which follows because the condition in (13) is implied by (24) and (46)

$$\max_{\substack{t\in[1:\bar{\tau}_n-1]:\\ \mathbb{P}[\tau_n=t]>0}} \mathbb{P}[\mathcal{E}_n|\tau_n = t] = \max_{\substack{t\in[1:\bar{\tau}_n-1]:\\ \mathbb{P}[\tau_n=t]>0}} \Big\{\mathbb{P}\big[\mathcal{E}_n|\tau_n = t, \bar{\mathcal{H}}_n\big]\, \mathbb{P}\big[\bar{\mathcal{H}}_n|\tau_n = t\big]$$

[1] We use the convention that $\sum_{i=j}^{j-1} a_i = 0$ for all a_i and for all integers j.

$$+ \mathbb{P}\Big[\mathcal{E}_n|\tau_n = t, \bar{\mathcal{H}}_n^{\mathsf{c}}\Big] \mathbb{P}\Big[\bar{\mathcal{H}}_n^{\mathsf{c}}|\tau_n = t\Big] \Big\} \tag{59}$$

$$\leq \max_{\substack{t\in[1:\bar{\tau}_n-1]: \\ \mathbb{P}[\tau_n=t]>0}} \mathbb{P}\big[\mathcal{E}_n|\tau_n = t, \bar{\mathcal{H}}_n\big] \tag{60}$$

$$= o(1) \tag{61}$$

as $n \to \infty$, and the condition in (12) follows from (61) and because $\mathbb{P}[\tau_n = \bar{\tau}_n] \to 0$ as $n \to \infty$.

Next, we prove that no EMS protocol can achieve a throughput larger than the RHS of (30), i.e., we establish that $\eta_{\text{opt}}(T) = \eta_{\text{BRQ}}(T)$ for $T > 1$. We do this by applying the strong converse in Lemma 1, which implies that a zero outage EMS protocol must satisfy $\sup_{k\in[1:\tau]} \mathbb{u}_{k,\tau}(H^\tau) \leq 0$ almost surely. To see this, note first that we must have $\tau < \infty$ almost surely. Otherwise, $\sup_n \mathbb{E}[\tau_n] = \infty$. Additionally, suppose that a zero outage EMS protocol satisfies $\mathbb{P}\Big[\sup_{k\in[1:\tau]} \mathbb{u}_{k,\tau}(H^\tau) > 0\Big] > 0$. Then,

$$\begin{aligned}
&\liminf_{n\to\infty} \mathbb{P}[\mathcal{E}_n] \\
&\geq \mathbb{P}\Big[\sup_{k\in[1:\tau]} \mathbb{u}_{k,\tau}(H^\tau) > 0, \tau < \infty\Big] \\
&\qquad\qquad \times \liminf_{n\to\infty} \mathbb{E}\Big[\mathbb{P}[\mathcal{E}_n|H^\infty] \Big| \sup_{k\in[1:\tau]} \mathbb{u}_{k,\tau}(H^\tau) > 0, \tau < \infty\Big] && (62) \\
&\geq \mathbb{P}\Big[\sup_{k\in[1:\tau]} \mathbb{u}_{k,\tau}(H^\tau) > 0\Big] \\
&\qquad\qquad \times \mathbb{E}\Big[\liminf_{n\to\infty} \mathbb{P}[\mathcal{E}_n|H^\infty] \Big| \sup_{k\in[1:\tau]} \mathbb{u}_{k,\tau}(H^\tau) > 0, \tau < \infty\Big] && (63) \\
&= \mathbb{P}\Big[\sup_{k\in[1:\tau]} \mathbb{u}_{k,\tau}(H^\tau) > 0\Big] && (64) \\
&> 0. && (65)
\end{aligned}$$

Here, (63) follows from Fatou's lemma [25, Th. 16.3] and because $\tau < \infty$ almost surely, and (64) follows from Lemma 1. Therefore, we can find the optimal zero outage EMS protocol satisfying the constraint $\mathbb{E}[\tau] \leq T$ by solving the following optimization problem:

$$\zeta_1(T) \triangleq \sup_{\{\mathbb{r}_t\},\{\mathbb{v}_t\},\tau} T\mathbb{E}\big[\bar{R}_\tau\big] / \mathbb{E}[\tau] \tag{66a}$$

$$\text{s.t.} \quad \mathbb{E}[\tau] \leq T \tag{66b}$$

$$\mathbb{P}\Big[\sup_{k\in[1:\tau]} \mathbb{u}_{k,\tau}(H_1^\tau) \leq 0\Big] = 1. \tag{66c}$$

Here, $\bar{R}_\tau \triangleq \sum_{t=1}^{\tau} \mathbb{r}_t(\mathbb{v}_{t-1}(H^{t-1}))$. It turns out that it is convenient to scale the objective function in (66) by T. We shall prove that the solution to (66) coincides with the BRQ-EMS protocol, i.e., we find that $\zeta_1(T) = T\eta_{\text{BRQ}}(T)$ under the condition in (32).

Since the transmitter has full delayed CSIT, it is sufficient to maximize over all composite rate selection-feedback functions $\mathbb{rv}_t : \mathbb{R}_+^{t-1} \mapsto \mathbb{R}_+$ such that $\mathbb{r}_t(\mathbb{v}_{t-1}(H_1^{t-1})) = \mathbb{rv}_t(H_1^{t-1})$. Moreover, we also define the optimization problem

$$\zeta(T) \triangleq \sup_{\{\mathbb{rv}_t\},\tau} \mathbb{E}\big[\bar{R}_\tau\big] \tag{67a}$$

$$\text{s.t.} \quad \mathbb{E}[\tau] \leq T \tag{67b}$$

$$\mathbb{P}\Big[\sup_{k\in[1:\tau]} \mathbb{u}_{k,\tau}(H_1^\tau) \le 0\Big] = 1. \tag{67c}$$

Since we have lower-bounded the objective function to get (67), we have that $\zeta(T) \le \zeta_1(T)$, but if it can be shown that $\zeta(\cdot)$ is an increasing function, then $\zeta_1(T) = \zeta(T)$. Indeed, suppose that $\zeta(\cdot)$ is an increasing function and that there exists $\tilde{T} > 1$ such that $\zeta(\tilde{T}) < \zeta_1(\tilde{T})$. Let τ^* be the solution of (66) for $T = \tilde{T}$. Then, we must have that $\zeta(\mathbb{E}[\tau^*]) = \zeta_1(\mathbb{E}[\tau^*])$ and that $\mathbb{E}[\tau^*] < \tilde{T}$. But since $\zeta(T)$ is an increasing function, we also have that $\zeta(\tilde{T}) > \zeta(\mathbb{E}[\tau^*]) = \zeta_1(\mathbb{E}[\tau^*]) = \zeta_1(\tilde{T})$ which cannot be true since $\zeta(T) \le \zeta_1(T)$ for all $T > 1$. We shall later use this fact to prove equality between $\zeta_1(T)$ and $\zeta(T)$ under the condition in (32).

To solve the optimization problem in (67), we first relate $\zeta(T)$ to the decoding time τ_{opt} defined in (19) as follows

$$\zeta(T) = \sup_{\substack{\{\mathbb{r}\mathbb{v}_t\}:\\ \mathbb{E}[\tau_{\text{opt}}]\le T}} \left\{ \mathbb{E}\left[\sum_{t=1}^{\tau_{\text{opt}}} \mathbb{r}\mathbb{v}_t(H^{t-1})\right] + \max_{\{\mathbb{r}\mathbb{v}_t\},\tau\ge\tau_{\text{opt}}} \mathbb{E}\left[\sum_{t=\tau_{\text{opt}+1}}^{\tau} \mathbb{r}\mathbb{v}_t(H^{t-1})\right]\right\}. \tag{68}$$

$$= \sup_{\substack{\{\mathbb{r}\mathbb{v}_t\}:\\ \mathbb{E}[\tau_{\text{opt}}]\le T}} \left\{ \mathbb{E}\left[\sum_{t=1}^{\tau_{\text{opt}}} \mathbb{r}\mathbb{v}_t(H^{t-1})\right] + \max_{\{\mathbb{r}\mathbb{v}_t\},\tau\ge\tau_{\text{opt}}} \mathbb{E}\left[\sum_{t=\tau_{\text{opt}+1}}^{\tau} \mathbb{r}\mathbb{v}_t(H^{t-1})\right]\right\}. \tag{69}$$

As previously, the maximization problems with respect to τ are subject to the constraints $\mathbb{E}[\tau] \le T$ and $\mathbb{P}\Big[\sup_{k\in[1:\tau]} \mathbb{u}_{k,\tau}(H^\tau) \le 0\Big] = 1$. In (68), we have used that any feasible (in the sense defined by the constraints in (67)) decoding time τ must satisfy $\tau \ge \tau_{\text{opt}}$ almost surely, and (69) follows because, by the definition of τ_{opt}, the constraints $\mathbb{u}_{k,\tau}(H_1^\tau) \le 0$ are automatically satisfied for $k \in [1{:}\tau_{\text{opt}}]$ when $\max_{k\in[\tau_{\text{opt}}+1:\tau]} \mathbb{u}_{k,\tau}(H_1^\tau) \le 0$ because

$$\mathbb{u}_{k,\tau}(H_1^\tau) = \mathbb{u}_{1,\tau_{\text{opt}}}(H_1^{\tau_{\text{opt}}}) - \mathbb{u}_{1,k-1}(H_1^{k-1}) + \mathbb{u}_{\tau_{\text{opt}+1},\tau}(H_1^\tau) \le 0. \tag{70}$$

It follows that the inner maximization in (69) is equal to $\zeta(T - \mathbb{E}[\tau_{\text{opt}}])$. Thus, we have

$$\zeta(T) = \max_{\substack{\{\mathbb{r}\mathbb{v}_t\}:\\ \mathbb{E}[\tau_{\text{opt}}]\le T}} \left\{ \mathbb{E}\left[\sum_{t=1}^{\tau_{\text{opt}}} \mathbb{r}\mathbb{v}_t(H^{t-1})\right] + \zeta(T - \mathbb{E}[\tau_{\text{opt}}])\right\} \tag{71}$$

$$= \max_{\substack{T_1,T_2\ge 0:\\ T_1+T_2=T}} \{\zeta_{\text{opt}}(T_1) + \zeta(T_2)\} \tag{72}$$

where

$$\zeta_{\text{opt}}(T) \triangleq \max_{\substack{\{\mathbb{r}\mathbb{v}_t\}:\\ \mathbb{E}[\tau_{\text{opt}}]\le T}} \mathbb{E}\left[\sum_{t=1}^{\tau_{\text{opt}}} \mathbb{r}\mathbb{v}_t(H^{t-1})\right]. \tag{73}$$

We can upper bound $\zeta_{\text{opt}}(\cdot)$ using weak duality as follows

$$\zeta_{\text{opt}}(T) \le \min_{\lambda>0}\left\{\lambda(T-1) + \max_{\{\mathbb{r}\mathbb{v}_t\}}\left\{\mathbb{E}\left[\sum_{t=1}^{\tau_{\text{opt}}} \mathbb{r}\mathbb{v}_t(H^{t-1})\right] - \lambda(\mathbb{E}[\tau_{\text{opt}}] - 1)\right\}\right\} \tag{74}$$

We solve the inner maximization in (74) using dynamic programming. For given $\lambda > 0$, let $\{\mathbb{r}\mathbb{v}_i^*\}$ be the solution to the inner maximization problem in (74). Then, observe that $\mathbb{r}\mathbb{v}_t^*$ depends only on H_1^{t-1} through $\mathbb{u}_1^{t-1}(H_1^{t-1})$. Intuitively, this means that the rate selection depends only on the amount of unresolved information up to time t. We define functions $\overline{\mathbb{r}\mathbb{v}}_t : \mathbb{R} \mapsto \mathbb{R}_+$ and let $\overline{\mathbb{r}\mathbb{v}}_t(\mathbb{u}_{1,t-1}(h_1^{t-1})) \triangleq \mathbb{r}\mathbb{v}_t(h_1^{t-1})$ for $t \in \mathbb{N}$ and $h_1^{t-1} \in \mathbb{R}_+^{t-1}$.

Now, define the value function (see e.g. [27])

$$V_t(u) \triangleq \max_{\{\overline{\mathbbm{rv}}_i\}} \mathbb{E}\left[\sum_{i=t}^{\tau_t(u)} \overline{\mathbbm{rv}}_i(\bar{\mathbbm{u}}_{t,i-1}(u, H_t^{i-1})) - \lambda(\tau_t(u) - t)\right] \tag{75}$$

where

$$\tau_t(u) \triangleq \min\big\{t' \geq t : \bar{\mathbbm{u}}_{t,t'}(u, H_t^{t'}) < 0\big\} \tag{76}$$

for $t \in \mathbb{N}$ and

$$\bar{\mathbbm{u}}_{k,t}(u, h_k^t) \triangleq u + \sum_{i=k}^{t} [\overline{\mathbbm{rv}}_i(\bar{\mathbbm{u}}_{k,i-1}(u, h_k^{i-1})) - C(h_i)] \tag{77}$$

for $t, k \in \mathbb{N}$. Using these definitions, the inner maximization in (74) can be expressed in terms of $V_t(\cdot)$ in (75):

$$V_1(0) = \max_{\{\mathbbm{rv}_t\}} \mathbb{E}\left[\sum_{i=1}^{\tau_{\text{opt}}} \mathbbm{rv}_i(H_1^{i-1}) - \lambda(\tau_{\text{opt}} - 1)\right] \tag{78}$$

To apply dynamic programming, the value function V_t in (75) is expressed in a recursive form:

$$\begin{aligned} V_t(u) = \max_{\{\overline{\mathbbm{rv}}_i\}_{i=t}^{\infty}} & \Bigg\{\overline{\mathbbm{rv}}_t(u) + F_C(u + \overline{\mathbbm{rv}}_t(u))\Bigg(\mathbb{E}\Bigg[\Bigg(\sum_{i=t+1}^{\tau_{t+1}(\bar{\mathbbm{u}}_t^t(u,H_t))} \overline{\mathbbm{rv}}_i(\bar{\mathbbm{u}}_{t,i-1}(u, H_t^{i-1}))\Bigg) \\ & - \lambda\Big(\tau_{t+1}(\bar{\mathbbm{u}}_{t,t}(u, H_t)) - t - 1\Big)\Bigg|C(H_t) \leq u + \overline{\mathbbm{rv}}_t(u)\Bigg] - \lambda\Bigg)\Bigg\} \qquad (79) \\ = \max_{r \geq 0} & \Big\{r + F_C(u + r)\,\Big[\mathbb{E}\Big[V_{t+1}(u + r - C(H_t))\Big|C(H_t) \leq u + r\Big] - \lambda\Big]\Big\}. \qquad (80) \end{aligned}$$

Here, we have defined $F_C(r) \triangleq \mathbb{P}[C(H) \leq r]$ and let $P_C(\cdot)$ be the probability density of $C(H)$. The problem is thereby formulated as a standard infinite horizon dynamic programming problem [27]. Consequently, the value function V_t is time-invariant such that $V_t(u) = V_{t+1}(u)$ for all $t \in \mathbb{N}$ and $u \in \mathbb{R}$. We denote the time-invariant value function by $V(u) \triangleq V_1(u)$. As a result, we can write

$$\begin{aligned} V(u) &= \max_{r \geq 0}\Big\{r + F_C(u + r)\,\Big[\mathbb{E}\Big[V(u + r - C(H))\Big|C(H) \leq u + r\Big] - \lambda\Big]\Big\} & (81) \\ &= \max_{r \geq 0}\big\{r + \mathbb{E}[\mathbb{1}\{C(H) \leq u + r\}\,(V(u + r - C(H)) - \lambda)]\big\}. & (82) \end{aligned}$$

It remains to guess $V(u)$ satisfying (82). We claim that the value function has the form $V(u) = A - u$ for $u \in [0, r_A]$, where

$$A \triangleq \max_{r \geq 0}\left\{r + \frac{\overline{C}(r) - F_C(r)\lambda}{1 - F_C(r)}\right\} \tag{83}$$

and r_A is the maximizer in (83). Here, $\overline{C}(r) \triangleq \int_0^r x P_C(x)\mathrm{d}x$. Indeed, from (82) and by substituting $V(u) = A - u$, we have

$$\begin{aligned} V(u) &= \max_{r \geq 0}\{r + \mathbb{E}[\mathbb{1}\{C(H) \leq u + r\}\,(A - u - r + C(H) - \lambda)]\} & (84) \\ &= \max_{r' \geq u}\big\{r'(1 - F_C(r')) + \overline{C}(r') + F_C(r')\,(A - \lambda)\big\} - u & (85) \\ &= A - u & (86) \end{aligned}$$

for every $u \in [0, r_A]$. Here, (86) follows from (83) because

$$0 = \max_{r \geq 0} \left\{ r - A + \frac{\overline{C}(r) - F_C(r)\lambda}{1 - F_C(r)} \right\} \tag{87}$$

$$= \max_{r \geq 0} \left\{ (r - A)(1 - F_C(r)) + \overline{C}(r) - F_C(r)\lambda \right\} \tag{88}$$

$$= \max_{r \geq 0} \left\{ r(1 - F_C(r)) + \overline{C}(r) + F_C(r)(A - \lambda) \right\} - A \tag{89}$$

Since r_A is a maximizer of the RHS of (83), it is also a maximizer of (89), and thus also of the optimization problem in (85). This proves that $V(u) = A - u$ for $u \in [0, r_A]$.

We shall shortly prove that (32) implies that

$$T\eta_{\text{BRQ}}(T) = F_C^{-1}\left(1 - \frac{1}{T}\right) + T\overline{C}\left(F_C^{-1}\left(1 - \frac{1}{T}\right)\right) \tag{90}$$

is concave in T. Consequently, we have shown the following

$$\zeta_{\text{opt}}(T) \leq \min_{\lambda > 0}\left\{ \lambda(T - 1) + \max_{\{\mathbb{rv}_i\}}\left\{ \mathbb{E}\left[\sum_{t=1}^{\tau_{\text{opt}}} \mathbb{rv}_t(H_1^{t-1})\right] - \lambda(\mathbb{E}[\tau_{\text{opt}}] - 1) \right\} \right\} \tag{91}$$

$$= \min_{\lambda > 0}\left\{ \lambda(T - 1) + \max_{r \geq 0}\left\{ r + \frac{\overline{C}(r) - \lambda F_C(r)}{1 - F_C(r)} \right\} \right\} \tag{92}$$

$$= \min_{\lambda > 0}\left\{ \lambda T + \max_{v \geq 1}\left\{ F_C^{-1}\left(1 - \frac{1}{\nu}\right) + \nu\overline{C}\left(F_C^{-1}\left(1 - \frac{1}{\nu}\right)\right) - \lambda\nu \right\} \right\} \tag{93}$$

$$= \max_{\nu \in [1, T]}\left\{ F_C^{-1}\left(1 - \frac{1}{\nu}\right) + \nu\overline{C}\left(F_C^{-1}\left(1 - \frac{1}{\nu}\right)\right) \right\} \tag{94}$$

$$= F_C^{-1}\left(1 - \frac{1}{T}\right) + T\overline{C}\left(F_C^{-1}\left(1 - \frac{1}{T}\right)\right) \tag{95}$$

$$= T\eta_{\text{BRQ}}(T). \tag{96}$$

Here, (92) follows from $V(0) = A$, (93) follows from the substitution $r = F_C^{-1}(1 - 1/\nu)$, (94) follows because (90) is concave and by Slater's condition [28, pp. 226–227], and the final equality holds since the objective function in (94) is increasing in ν.

Next, we need to show that $\zeta_1(T) = \zeta_{\text{opt}}(T)$. From the concavity of $\zeta_{\text{opt}}(\cdot)$, the simple upper bound $\zeta_{\text{opt}}(T) \leq TC_{\text{erg}}$, and $\lim_{T \to \infty} \frac{\partial}{\partial T}\zeta_{\text{opt}}(T) = C_{\text{erg}}$, we have that $\zeta(T) = \zeta_{\text{opt}}(T) \leq T\eta_{\text{BRQ}}(T)$. Moreover, since $\eta_{\text{BRQ}}(\cdot)$ is an increasing function, it also follows that $\zeta_1(T) \leq T\eta_{\text{BRQ}}(T)$ as desired.

It remains to establish the claim that $T\eta_{\text{BRQ}}(T)$ is concave in T. To do so, we show that the second derivative of $T\eta_{\text{BRQ}}(T)$ with respect to T is negative. It turns out that the second derivative of $T\eta_{\text{BRQ}}(T)$ with respect to T is given by

$$\begin{aligned}\log(2)\frac{\partial^2(T\eta_{\text{BRQ}})}{\partial T^2} = &-\frac{1}{(1 + F_H^{-1}(1 - 1/T))T^3 P_H(F_H^{-1}(1 - 1/T))} \\ &-\frac{1}{(1 + F_H^{-1}(1 - 1/T))^2 T^4 P_H(F_H^{-1}(1 - 1/T))^2} \\ &-\frac{P_H'(F_H^{-1}(1 - 1/T))}{(1 + F_H^{-1}(1 - 1/T))T^4 P_H(F_H^{-1}(1 - 1/T))^3}.\end{aligned} \tag{97}$$

By multiplying the RHS of (97) by the positive term $(1 + F_H^{-1}(1 - 1/T))^2 T^4$ and by using the substitution $T = 1/(1 - F_H(h))$, we find that $\frac{\partial^2(T\eta_{\text{BRQ}})}{\partial T^2} \leq 0$ is equivalent to the condition in (32), hence establishing the desired result. ■

V. Finite Number of Feedback Messages

The requirement of full delayed CSIT feedback is not realistic in many applications. This section therefore addresses the case where the feedback cost is finite. While HARQ-INR does not allow for rate adaptations, EMS protocols with three or more feedback messages can be used to signal ACK/NACK, but also to instruct the transmitter to append additional information bits in the subsequent slot. The key difference from the case with full delayed CSIT is that the optimal amount of new information to be appended cannot be specified through feedback. We provide a heuristic choice of the rate selection functions, feedback functions, and decoding times and demonstrate the existence of a zero outage EMS protocol. In Section VI, it is shown that the throughput of the finite feedback cost EMS protocol is comparable with BRQ.

In our heuristic EMS protocol, termed *EMS-*$(f+1)$, we define the rate selection and feedback functions as

$$\mathbb{v}_t(h_1^t) \triangleq \begin{cases} \left\lfloor f - \frac{\mathbb{u}_{1,t}(h_1^t)}{\mathbb{r}} \right\rfloor, & \mathbb{u}_{1,t}(h_1^t) > 0 \\ -1, & \mathbb{u}_{1,t}(h_1^t) \le 0 \end{cases} \tag{98}$$

$$\mathbb{r}_t^{(n)}(v_{t-1}) \triangleq \begin{cases} \mathbb{r} f - c_n, & t = 1 \\ \min\{\mathbb{r}(f-1) - c_n, \mathbb{r} v_{t-1}\}\mathbb{1}\{v_{t-1} \neq -1\}, & t \ge 2 \end{cases} \tag{99}$$

Here, $\mathbb{r} > 0$ is a predefined constant, $\mathbb{F} = [-1{:}f-1]$, and $c_n \triangleq c_1/\log(n)$ for an arbitrary positive constant c_1. The decoding time is given by

$$\tau_n = \min\{\bar{\tau}_n, \tau\}. \tag{100}$$

where

$$\tau \triangleq \inf\{t \ge 1 : V_t = -1\} \tag{101}$$

$$\bar{\tau}_n \triangleq -\left\lfloor \frac{\log(c_2\sqrt{n})}{\log F_C(\mathbb{r}(f-1))} \right\rfloor \tag{102}$$

Here, c_2 is an arbitrary positive constant and the feedback -1 designates an ACK message. Since $\mathbb{v}_t^{(n)}$ can take at most $f+1 = |\mathbb{F}|$ values, the corresponding EMS protocol has feedback cost $f+1$. As for the BRQ-EMS protocol, we also define the composite rate-feedback function as

$$\overline{\mathbb{r}\mathbb{v}}(u) \triangleq \mathbb{r}\min\left\{f-1, \left\lfloor f - \frac{[u]^+}{\mathbb{r}} \right\rfloor\right\}. \tag{103}$$

With this definition, we can write

$$\mathbb{r}_t(\mathbb{v}_{t-1}(h_1^{t-1})) = \overline{\mathbb{r}\mathbb{v}}(\mathbb{u}_{1,t-1}(h_1^{t-1})) \tag{104}$$

for all $t \ge 2$ and $h_1^{t-1} \in \mathbb{R}_+^{t-1}$ such that $\mathbb{u}_{1,t-1}(h_1^{t-1}) > 0$.

The trade-off between throughput and average decoding time achievable by an EMS-$(f+1)$ protocol is characterized by the following theorem which provides a way to compute the throughput and average decoding time by solving a pair of integral equations. Varying the parameter $\mathbb{r}$ determines the trade-off between throughput and average decoding time.

Theorem 4: Define $W : [0, \mathbbm{r}f] \mapsto \mathbb{R}_+$ and $M : [0, \mathbbm{r}f] \mapsto \mathbb{R}_+$ through the integral equations

$$W(u) \triangleq \overline{\mathbbm{rv}}(u) + \int_0^{u+\overline{\mathbbm{rv}}(u)} P_C(x) W(u + \overline{\mathbbm{rv}}(u) - x) \mathrm{d}x \tag{105}$$

and

$$M(u) = 1 + \int_0^{u+\overline{\mathbbm{rv}}(u)} P_C(x) M(u + \overline{\mathbbm{rv}}(u) - x) \mathrm{d}x. \tag{106}$$

Here, $P_C(\cdot)$ denotes the probability density function of $C(H)$. Then, there exists an (η, T)-zero outage EMS protocol with

$$\eta = \frac{\mathbbm{r}f + \mathbb{E}\Big[\mathbb{1}\{C(H) \le \mathbbm{r}f\}\, W(\mathbbm{r}f - C(H))\Big]}{1 + \mathbb{E}\Big[\mathbb{1}\{C(H) \le \mathbbm{r}f\}\, M(\mathbbm{r}f - C(H))\Big]} \tag{107}$$

and

$$T = 1 + \mathbb{E}\Big[\mathbb{1}\{C(H) \le \mathbbm{r}f\}\, M(\mathbbm{r}f - C(H))\Big]. \tag{108}$$

Proof: In order to show that (98)–(100) define a zero outage EMS protocol, we need to verify the conditions of Lemma 2. We shall first show that (23) is satisfied for $g_n = \mathcal{O}(1/\sqrt{n})$. The remaining conditions can be verified using same arguments as in the proof of Theorem 3. Given that $\tau \le \bar{\tau}_n$, we have for $k \in [2{:}\tau_n]$

$$\begin{aligned}
\mathbbm{u}_{k,\tau_n}^{(n)}(H^{\tau_n}) &= \sum_{i=k}^{\tau_n} \left[\min\left\{\mathbbm{r}(f-1) - c_n, \mathbbm{r}\left\lfloor f - \frac{\mathbbm{u}_{1,i-1}(H_1^{i-1})}{\mathbbm{r}}\right\rfloor\right\} - C(H_i)\right] && (109)\\
&\le \sum_{i=k}^{\tau_n} \left[\min\left\{\mathbbm{r}(f-1) - c_n, \mathbbm{r}\left\lfloor f - \frac{\mathbbm{u}_{1,i-1}(H_1^{i-1})}{\mathbbm{r}}\right\rfloor\right\} - C(H_i)\right] - \mathbbm{u}_{1,\tau_n}(H^{\tau_n}) && (110)\\
&= \sum_{i=k}^{\tau_n} \left[\mathbbm{r}(f-1) - c_n - \mathbbm{r}\left\lfloor f - \frac{\mathbbm{u}_{1,i-1}^{(n)}(H_1^{i-1})}{\mathbbm{r}}\right\rfloor\right]^- - \mathbbm{u}_{1,k-1}(H^{k-1}) && (111)\\
&\le \sum_{i=k}^{\tau_n} \left[-c_n + \mathbbm{u}_{1,i-1}(H_1^{i-1})\right]^- - \mathbbm{u}_{1,k-1}(H_1^{k-1}) && (112)\\
&\le \min\{-c_n, -\mathbbm{u}_{1,i-1}(H_1^{i-1})\} + \sum_{i=k+1}^{\tau_n} \left[-c_n + \mathbbm{u}_{1,i-1}(H_1^{k-1})\right]^- && (113)\\
&\le -c_n. && (114)
\end{aligned}$$

Here, (109) follows from (16) and (98)–(99), (110) follows because $\mathbbm{u}_{1,\tau_n}(H^{\tau_n}) \le 0$ when $\tau \le \bar{\tau}_n$, (112) follows from $\lfloor x \rfloor \in (x-1, x]$, (114) follows because $[x]^- \le 0$. Using the same arguments as in (47), it can also be shown that $\mathbbm{u}_{1,\tau_n}^{(n)}(H^{\tau_n}) \le -c_n$ when $\tau \le \bar{\tau}_n$. Hence, we conclude that $\max_{k \in [1:\tau_n]} \mathbbm{u}_{k,\tau_n}^{(n)}(H^{\tau_n}) \le -c_n$ when $\tau \le \bar{\tau}_n$. An immediate implication of this is that

$$\mathbb{P}\Big[\tau_n = t \Big| \bar{\mathcal{H}}_n\Big] = \frac{\mathbb{P}\big[\tau_n = t, \bar{\mathcal{H}}_n\big]}{\mathbb{P}\big[\bar{\mathcal{H}}_n\big]} = \frac{\mathbb{P}[\tau = t]}{\mathbb{P}\big[\bar{\mathcal{H}}_n\big]} \ge \mathbb{P}[\tau = t] \tag{115}$$

for all $t \in [1{:}\bar{\tau}_n]$. Note that τ is not necessarily Geometrically distributed as for the case with full CSIT. Instead, since $\lfloor x \rfloor \in (x-1, x]$ for any constant x, we have that

$$\mathbb{u}_{1,t-1}(h^{t-1}) + \mathbb{r}_t(\mathbb{v}_{t-1}(h^{t-1})) = \mathbb{u}_{1,t-1}(h^{t-1}) + \mathbb{r}\left\lfloor f - \frac{\mathbb{u}_{1,t-1}(h_1^{t-1})}{\mathbb{r}} \right\rfloor \in (\mathbb{r}(f-1), \mathbb{r}f]. \tag{116}$$

Therefore, for all $t \in \mathbb{N}$, we also have that

$$\mathbb{P}[\tau \geq t+1 | \tau \geq t] = \mathbb{P}\left[\mathbb{u}_{1,t}(H^t) > 0 \middle| \tau \geq t\right] \in [F_C(\mathbb{r}(f-1)), F_C(\mathbb{r}f)] \tag{117}$$

Thus,

$$\mathbb{P}[\tau = t] = \mathbb{P}[\tau = t | \tau \geq t] \prod_{i=1}^{t-1} \mathbb{P}[\tau \geq i+1 | \tau \geq i] \geq F_C(\mathbb{r}(f-1))^{t-1}(1 - F_C(\mathbb{r}f)) \tag{118}$$

It follows from (115) and (118) that (23) is satisfied for $g_n = \mathcal{O}(1/\sqrt{n})$. The conditions in (12), (13), and (21) follows using the same arguments as in the proof of Theorem 3. Similarly, we can also show that $\lim_{n\to\infty} \mathbb{E}[\tau_n] = \mathbb{E}[\tau]$ and that $\lim_{n\to\infty} \mathbb{E}\left[\bar{R}_{\tau_n}^{(n)}\right] = \mathbb{E}\left[\bar{R}_\tau\right]$. Hence, it only remains to compute the throughput given by $\mathbb{E}\left[\bar{R}_\tau\right]/\mathbb{E}[\tau]$ and the limiting average decoding time $\mathbb{E}[\tau]$.

We compute the throughput $\mathbb{E}\left[\bar{R}_\tau\right]/\mathbb{E}[\tau]$ via the rate selection and feedback functions in (98)–(99) and the decoding time in (100). Using the following recursive relation

$$\mathbb{u}_{1,t}(h^t) = \mathbb{u}_{1,t-1}(h^{t-1}) + \overline{\mathbb{rv}}(\mathbb{u}_{1,t-1}(h^{t-1})) - C(h_t) \tag{119}$$

for $t \geq 2$, we observe that, if $t \geq k \geq 2$, then $\mathbb{u}_{1,t}(h^{k-1}, H_k^t)$ only depends on h^{k-1} through $\mathbb{u}_{1,k-1}(h^{k-1})$. Therefore, we can define $\bar{\mathbb{u}}(u, H_k^t)$ such that $\bar{\mathbb{u}}(\mathbb{u}_{1,k-1}(h^{k-1}), H_k^t) = \mathbb{u}_{1,t}(h^{k-1}, H_k^t)$. In order to compute $\mathbb{E}\left[\bar{R}_\tau\right]$, define

$$W_t(u) \triangleq \mathbb{E}\left[\sum_{i=t}^{\tau_t(u)} \overline{\mathbb{rv}}(\bar{\mathbb{u}}(u, H_t^{i-1}))\right] \tag{120}$$

for $u \in [0, \mathbb{r}f]$, where

$$\tau_t(u) \triangleq \inf\left\{t' \geq t : \bar{\mathbb{u}}(u, H_t^{t'}) < 0\right\}. \tag{121}$$

Observe that $\mathbb{E}\left[\bar{R}_\tau\right] = \mathbb{r}f + \mathbb{E}[W_1(\mathbb{u}_{1,1}(H_1))\mathbb{1}\{C(H_1) \leq \mathbb{r}f\}]$. Rewriting the RHS of (120) in terms of $W_{t+1}(\cdot)$, we obtain

$$\begin{aligned}
W_t(u) &= \overline{\mathbb{rv}}(u) + \mathbb{E}\left[\mathbb{1}\{u + \overline{\mathbb{rv}}(u) \geq C(H_t)\} \sum_{i=t+1}^{\tau_{t+1}(\bar{\mathbb{u}}(u,H_t))} \overline{\mathbb{rv}}(\bar{\mathbb{u}}(\bar{\mathbb{u}}(u, H_t), H_{t+1}^{i-1}))\right] & (122)\\
&= \overline{\mathbb{rv}}(u) + \mathbb{E}_{H_t}\left[\mathbb{1}\{u + \overline{\mathbb{rv}}(u) \geq C(H_t)\}\, \mathbb{E}_{H_{t+1}^\infty}\left[\sum_{i=t+1}^{\tau_{t+1}(\bar{\mathbb{u}}(u,H_t))} \overline{\mathbb{rv}}(\bar{\mathbb{u}}(\bar{\mathbb{u}}(u, H_t), H_{t+1}^{i-1}))\right]\right] & (123)\\
&= \overline{\mathbb{rv}}(u) + \mathbb{E}[\mathbb{1}\{u + \overline{\mathbb{rv}}(u) \geq C(H_t)\}\, W_{t+1}(\bar{\mathbb{u}}(u, H_t))] & (124)\\
&= \overline{\mathbb{rv}}(u) + \int_0^{u+\overline{\mathbb{rv}}(u)} P_C(x) W_{t+1}(u + \overline{\mathbb{rv}}(u) - x)\mathrm{d}x. & (125)
\end{aligned}$$

By defining $W(\cdot) \triangleq W_1(\cdot)$ and by noting that $W_t(u) = W_{t+1}(u)$ for $u \in [0, \mathbb{r}f]$, we have the integral equation in (105). The expected reward is thereby given by

$$\mathbb{E}\left[\bar{R}_{\tau}\right] = \mathbb{r}f + \mathbb{E}[\mathbb{1}\{C(H) \leq \mathbb{r}f\}\, W(\mathbb{r}f - C(H))] \,. \tag{126}$$

Using derivations similar to (120)–(125), we obtain $\mathbb{E}[\tau] = 1 + \mathbb{E}[\mathbb{1}\{C(H) \leq \mathbb{r}f\}\, M(\mathbb{r}f - C(H))]$. ■

We remark that the integral equations in Theorem 4 can be written as Fredholm equations of the second kind. These are readily solved as a system of linear equations when discretized or by using a quadrature method specifically for Fredholm equations [29].

VI. Numerical Results

In the following, the throughput of the described protocols are assessed and compared to traditional transmission strategies. Specifically, we compare our protocols with HARQ-INR and HARQ-INR with power adaptation.

A. Coding strategies

1) HARQ-INR: In HARQ-INR (also known as HARQ-INR), the transmitter initially transmits at a rate R in the first slot and continues to send additional parity bits in the subsequent slots. By the end of each slot, the receiver attempts to decode and feeds back an ACK/NACK depending on whether the decoding was successful or not. The receiver is thereby able to accumulate information until decoding is possible. Then the average decoding time of HARQ-INR is given by [14]

$$\mathbb{E}[\tau] = \sum_{m=1}^{\infty} m(p_{\text{out}}^{m-1}(R) - p_{\text{out}}^{m}(R)) \tag{127}$$

$$= 1 + \sum_{m=1}^{\infty} p_{\text{out}}^{m}(R) \tag{128}$$

where $p_{\text{out}}^{m}(\cdot)$ is the outage probability after the mth retransmission and is given by

$$p_{\text{out}}^{m}(x) = \mathbb{P}\left[\sum_{k=1}^{m} C(H_k) < x\right]. \tag{129}$$

The achievable throughput is

$$\eta_{\text{HARQ-INR},R} = \frac{R}{1 + \sum_{m=1}^{\infty} p_{\text{out}}^{m}(R)}. \tag{130}$$

The optimal throughput of HARQ-INR subject to the average decoding time constraint is given by

$$\eta_{\text{HARQ-INR}}(T) = \max_{R} \frac{R}{1 + \sum_{m=1}^{\infty} p_{\text{out}}^{m}(R)} \tag{131a}$$

$$\text{s.t. } 1 + \sum_{m=1}^{\infty} p_{\text{out}}^{m}(R) \leq T. \tag{131b}$$

We point out that $\sup_{T\in(1,\infty)} \eta_{\text{HARQ-INR}}(T) = C_{\text{erg}}$.

2) HARQ-INR with power adaptation: A comparison between BRQ and HARQ-INR is not fair in the sense that HARQ-INR does not use the available delayed CSIT. It has been shown in literature that delayed CSIT can provide significant throughput benefits if the short-term power constraint in (6) is relaxed. Power adaptation based on delayed CSIT has previously been proposed in a slightly different setting in [12]. In this section, we optimize HARQ-INR with power adaptation under a constraint on the average decoding time. We follow [6] and redefine the power constraint in (6) such that $\frac{1}{n}\mathbf{X}_t^{\mathrm{T}}\mathbf{X}_t \leq \rho_t$, where we require that the random variables ρ_t depends only on $\{H_t\}_{t=1}^{t-1}$ and that $\{\rho_t\}_{t=1}^{\infty}$ satisfies

$$\frac{\mathbb{E}[\sum_{i=1}^{\tau}\rho_i]}{\mathbb{E}[\tau]} \leq 1. \tag{132}$$

The constraint in (132) ensures that the average power per slot over many runs of the protocol does not exceed one. Under this relaxation, we can design an HARQ-INR-type protocol that benefits from full delayed CSIT using power adaptation. In particular, full delayed CSIT provides the transmitter with knowledge about the amount of unresolved information at the receiver and is allowed to use this knowledge to optimize the power spend in the following slot. The transmitter sends in the first slot at a rate R using power ρ_1. At the end of the slot, the transmitter receives the delayed CSIT which can be used to compute the unresolved information I_1 at the receiver. In the tth slot, the transmitter sends IR with power $\rho_t(I_{t-1})$, where I_{t-1} is the amount of unresolved information at the receiver by the end of slot $t-1$ and $\rho_t(\cdot)$ denotes the power adaptation policy in the tth slot. It follows that the unresolved information in slot t satisfies

$$I_t = I_{t-1} - C(H_t\rho_t(I_{t-1})) \tag{133}$$

where $I_0 \triangleq R$. We shall solve the following optimization problem using dynamic programming:

$$\min_{\{\rho_i(\cdot)\}_{i=1}^{\infty}} \mathbb{E}[\tau] \tag{134a}$$

$$\text{s.t.} \quad \mathbb{E}\left[\sum_{t=1}^{\tau}\rho_i(I_{t-1})\right] \leq \mathbb{E}[\tau]\,. \tag{134b}$$

First, we rewrite (134) as an unconstrained optimization problem using duality:

$$\begin{aligned}&\max_{\lambda>0}\min_{\{\rho_i(\cdot)\}_{i=1}^{\infty}}\left\{\mathbb{E}[\tau] + \lambda\left(\mathbb{E}\left[\sum_{t=1}^{\tau}\rho_i(I_{t-1})\right] - \mathbb{E}[\tau]\right)\right\}\\ &= \max_{\lambda>0}\min_{\{\rho_i(\cdot)\}_{i=1}^{\infty}}\left\{\mathbb{E}[\tau]\,(1-\lambda) + \lambda\mathbb{E}\left[\sum_{t=1}^{\tau}\rho_i(I_{t-1})\right]\right\}.\end{aligned} \tag{135}$$

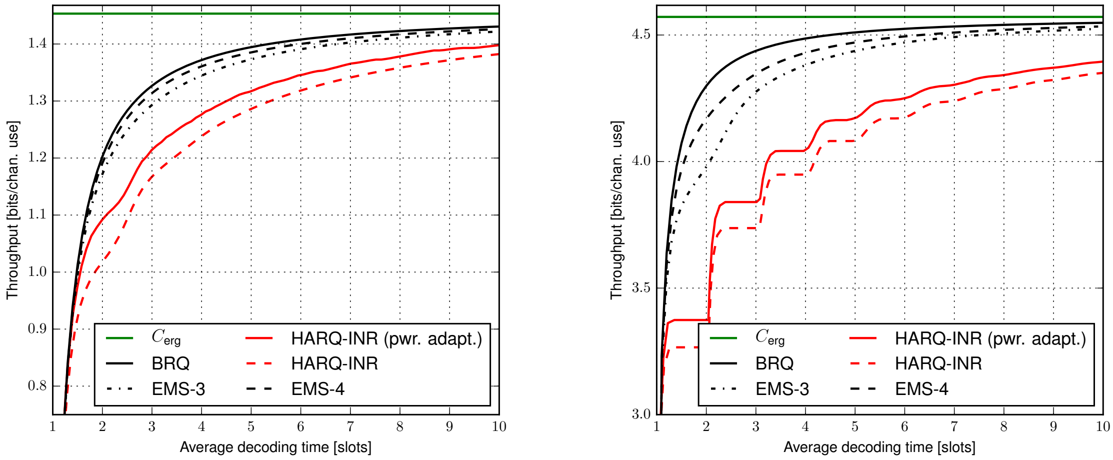
Then, we rewrite the inner minimization in (135) as an infinite-horizon dynamic programming problem. Specifically, we have that

$$\min_{\{\rho_i(\cdot)\}_{i=1}^{\infty}}\left\{\mathbb{E}[\tau]\,(1-\lambda) + \lambda\mathbb{E}\left[\sum_{t=1}^{\tau}\rho_i(I_{t-1})\right]\right\} = J_\lambda(R) \tag{136}$$

where the function $J_\lambda(\cdot)$ is defined by $J_\lambda(u) = 0$ for $u \leq 0$ and, for $u > 0$,

$$J_\lambda(u) = \min_{\rho}\left\{1 + \lambda\rho + \int_0^{\frac{2^u-1}{\rho}} P_H(h)J_\lambda(u - C(h\rho))dh\right\}. \tag{137}$$

Consequently, we find that the solution to the optimization problem in (134) is given by $\max_{\lambda>0} J_\lambda(R)$. The throughput of HARQ-INR with power adaptation under an average decoding time constraint is thereby given



(a) SNR $= 10$ dB

(b) SNR $= 30$ dB

Fig. 2. Throughput versus average decoding time $\mathbb{E}[\tau]$ of various protocols. The throughput of HARQ-INR and HARQ-INR with power adaptation are computed using (131) and (138), respectively. The throughput of BRQ/BRQ-EMS is computed using (30) and for the EMS protocols we use (107).

by

$$\eta_{\text{HARQ-INR-P}}(T) = \max_{R>0} \frac{R}{\max_{\lambda>0} J_\lambda(R)} \tag{138a}$$

$$\text{s.t. } \max_{\lambda>0} J_\lambda(R) \leq T. \tag{138b}$$

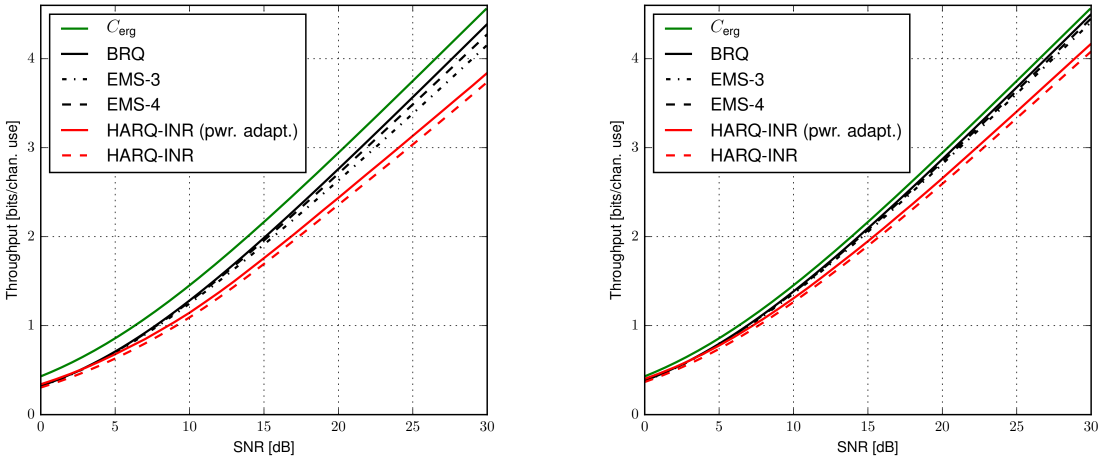
B. Assessment

We evaluate the proposed protocols by assuming Rayleigh block-fading, independent from slot to slot, i.e., the probability density of H is given by

$$P_H(h) = \frac{1}{\Gamma} e^{-h/\Gamma}. \tag{139}$$

Fig. 2 depicts the throughput of various protocols as a function of average decoding time for average SNR equal to 10 dB and 30 dB. We remark that the stair-step behavior of the throughput of HARQ-INR at SNR $= 30$ dB origins because the probability distribution of $C(H)$ becomes increasingly concentrated around C_{erg} as the SNR increases. For high SNR, this implies that the average decoding time, and therefore also the throughput, has a stair-step behavior when R grows linearly. It is seen that the throughput of all protocols tend to the ergodic capacity as the allowed average decoding times are increased. We observe that BRQ and the EMS protocols with finite feedback cost significantly outperforms both HARQ-INR and HARQ-INR with power adaptation in terms of throughput. A particular interesting observation is that the proposed EMS protocols for finite feedback cost achieves throughput that are very close to that of BRQ, even for the case $f = 2$. Our interpretation of this is that the precise amount of additional information bits appended in each slot does not affect the throughput significantly.

In Fig. 3, the throughput is plotted in terms of SNR for fixed average decoding time $\mathbb{E}[\tau]$. Observe that the



(a) $\mathbb{E}[\tau] = 2.5$

(b) $\mathbb{E}[\tau] = 4.5$

Fig. 3. Average throughput versus SNR of various protocols. The throughput of HARQ-INR and HARQ-INR with power adaptation are computed using (131) and (138), respectively. The throughput of BRQ/BRQ-EMS is computed using (30) and for the EMS protocols we use (107).

rate penalty of BRQ compared to the ergodic capacity is approximately constant throughout the range of SNR values while the penalty of the remaining protocols increases for higher SNR.

VII. Discussion and Conclusions

The objective of this paper is to generalize and extend the BRQ protocol, proposed in [1], to a broader class of communication strategies, termed expandable message space (EMS) protocols. EMS protocols are useful when the CSI is only available after the transmission has taken place. The main novelty of EMS protocols is the possibility of appending new information bits before previously transmitted data has been resolved. EMS protocols thereby provides an elegant way to design communication protocols that approach the ergodic capacity within a low average decoding time. In contrast to BRQ, EMS protocols in general also benefit from limited feedback. Specifically, it has been shown that even ternary feedback is sufficient to achieve throughput close to that of BRQ. This suggests that the main reason for the superior throughput of BRQ and EMS protocols is that, compared to HARQ-type protocols with/without power adaptation, they only terminate a transmission when the CSI is sufficiently good, whereas HARQ-INR terminates a transmission as soon as a sufficient amount of information is accumulated. As a result, HARQ-INR protocol often collect a wasteful amount of information which far surpasses the amount of unresolved information, leading to waste of resources.

Unlike most works in the field of HARQ, we have presented results for systems with an average decoding time constraint as opposed to a strict decoding time constraint. Strict decoding time constraints lead to protocols with a maximum transmission length. Such constraints are motivated by applications like streaming of multimedia data, where data become useless after a certain amount of time. Despite this, there are many applications where data is retransmitted at a packet level upon outage. In other words, a new transmission is initiated with the same data – perhaps concatenated with data from new data arrivals. For such applications, a constraint on the average decoding time is more applicable. Although strict decoding time constraints have not been considered, they are not ruled by the definition of EMS protocols. The time-dependent rate selection functions allows for detailed modeling of the distribution of decoding times. An optimal EMS protocol with full delayed CSIT and a constraint on outage probability instead of average decoding time can be computed using dynamic programming in a manner similar to [10]. An interesting extension to this work is thus to consider a queuing-based system model where data packets arrive at a certain rate λ. In such a system, EMS protocols does not always have data available for transmission as it is assumed in this work.

We have not treated the impact of the accuracy of the delayed CSI in our throughput comparisons. In the conventional HARQ-INR protocol that rely on, possibly quantized, prior CSI to perform rate and/or power adaptation, the accuracy of CSI has a significant impact on the throughput [2, pp. 209–213]. The main reason for this is that the channel gains change from the time the CSI is estimated to the time the channel is used, which can take a duration that spans multiple slots. This inaccuracy is largely eliminated by relying only on delayed CSI. This follows because the receiver can make a much more precise estimate of the CSI after having observed a time slot. For the EMS protocols, however, inaccurate delayed CSI implies that the transmitter cannot precisely append the optimal amount of new information in each step. Our results for the EMS protocols with finite feedback cost show that the precise amount of new information appended in each slot does not significantly alter the achievable throughput. Therefore, we do not expect that the throughput of EMS protocols to suffer significantly if the CSI is inaccurate.

Finally, we note that HARQ-INR have led to several composite protocols that use HARQ-INR as building block. As previously discussed, two examples which are of relevance to this paper are [10] and [14]. One can design similar composite protocols using the EMS protocols as building blocks. For example, the broadcast approach to HARQ-INR proposed in [14] provides an approach combine multiple HARQ-INR instances that run in parallel in multiple superposition coded layers. We can combine multilayered transmission and EMS protocols similarly. One feasible approach is to divide each transmission into two layers: one with IR for the previous slots and one with new information bits. One can then optimize over the distribution of power in the two layers. In this way the decoder does not need to decode both the IR for previous slot and the new information bits simultaneously. Hence, such protocols might lead to higher throughput than the present paper report. One can also follow the approach taken in [10] and instantiate several instances of EMS protocols which run in parallel in a TDMA fashion.

ACKNOWLEDGEMENT

This work has been in part supported by the European Research Council (ERC Consolidator Grant nr. 648382 WILLOW) within the Horizon 2020 Program.

APPENDIX I
PROOF OF LEMMA 1 (STRONG CONVERSE)

Fix an EMS protocol defined by $\{\tau_n\}$, $\{\mathbb{r}_t^{(n)}\}$, $\{\mathbb{v}_t\}$, $\{\mathbb{f}_t^{(n)}\}$, and $\{\mathbb{g}_t^{(n)}\}$. The EMS protocol induces a probability distribution on $(\mathbf{X}^{\tau_n}, \mathbf{Y}^{\tau_n}, H^{\tau_n})$ given by $P_{\mathbf{Y}^{\tau_n},\mathbf{X}^{\tau_n},H^{\tau_n}}$. To simplify notation, we condition on $H^\infty = h^\infty$ throughout the proof and define the probability distribution $\bar{\mathbb{P}}$ on $(\mathbf{X}^{\tau_n}, \mathbf{Y}^{\tau_n})$ by

$$\bar{\mathbb{P}}[\cdot] \triangleq \mathbb{P}[\cdot|H^\infty = h^\infty]\,. \tag{140}$$

Since the stopping time and rate selection functions depend only on the channel realizations, conditioning on $H^\infty = h^\infty$ implies that $\{\tau_n\}$ and $\{R_t^{(n)}\}$ are deterministic sequences. The probability distribution of the channel outputs at the tth slot is

$$\overline{\mathbb{P}}_{\mathbf{Y}_t|\mathbf{X}_t}(\mathbf{y}|\mathbf{x}) \triangleq \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(y_i - \sqrt{h_t} x_i)^2}. \tag{141}$$

Since $\tau < \infty$, the limit $\lim_{n\to\infty} \tau_n = \tau$ exists and implies that there exist positive integers N and n_0 such that $\tau_n \le N$ for all $n \ge n_0$. Therefore, we have[2]

$$\begin{aligned} \lim_{n\to\infty} \max_{k\in[1:\tau_n+1]} \sum_{i=k}^{\tau_n} (R_i^{(n)} - C(h_i)) &= \lim_{n\to\infty} \max_{k\in[1:N+1]} \sum_{i=k}^{N} \mathbb{1}\{i \le \tau_n\} \left(R_i^{(n)} - C(h_i)\right) & (142)\\ &= \max_{k\in[1:N+1]} \sum_{i=k}^{N} \mathbb{1}\{i \le \tau\} \left(R_i - C(h_i)\right) & (143)\\ &= \max_{k\in[1:\tau+1]} \mathbb{u}_{k,\tau}(h^\tau) > 0. & (144) \end{aligned}$$

The last inequality follows from the condition $\sup_{k\in[1:\tau]} \mathbb{u}_{k,\tau}(h^\tau) > 0$. This implies that there exist a positive integer n_1, a positive constant γ, and a sequence of integers $\{\bar{k}_n\}_{n=n_1}^{\infty}$ with $\bar{k}_n \in [1{:}\tau_n]$ such that

$$\sum_{i=\bar{k}_n}^{\tau_n} (R_i^{(n)} - C(h_i)) \ge 2\gamma \tag{145}$$

for all $n \ge n_1$.

To proceed, we prove a straightforward variation of the converse by Verdú and Han [30]. To state the result, we shall define the information density for $t \in [1{:}\tau_n]$ as follows

$$i\big(\mathbf{x}_t^{\tau_n}; \mathbf{y}_t^{\tau_n} | \mathbf{x}^{t-1}\big) = \log_2 \frac{\prod_{i=t}^{\tau_n} \overline{\mathbb{P}}_{\mathbf{Y}_i|\mathbf{X}_i}(\mathbf{y}_i|\mathbf{x}_i)}{\overline{\mathbb{P}}_{\mathbf{Y}_t^{\tau_n}|\mathbf{X}^{t-1}}(\mathbf{y}_t^{\tau_n}|\mathbf{x}^{t-1})}, \qquad \mathbf{x}^{\tau_n}, \mathbf{y}^{\tau_n} \in \mathbb{R}^{n\tau_n}. \tag{146}$$

Lemma 5: Under the above definitions, the following holds for every n

$$\overline{\mathbb{P}}[\mathcal{E}_n] \ge \max_{t\in[1:\tau_n]} \overline{\mathbb{P}}\left[\frac{1}{n} i\Big(\mathbf{X}_t^{\tau_n}; \mathbf{Y}_t^{\tau_n} \Big| \mathbf{X}^{t-1}\Big) \le \sum_{k=t}^{\tau_n} R_k^{(n)} - \gamma\right] - 2^{-n\gamma} \tag{147}$$

where $\gamma > 0$ is an arbitrary constant.

Proof: The proof closely follows those found in [30, Th. 4] or [31, Lemma 3.2.2]. The encoder functions $(\mathbb{f}_1^{(n)}, \cdots, \mathbb{f}_{\tau_n}^{(n)})$ generates $M_n \triangleq 2^{\lceil n\bar{R}_{\tau_n}^{(n)} \rceil}$ codewords which we denote by $\{\mathbf{u}(i)\}_{i=1}^{M_n}$, where $\mathbf{u}(i) \in \mathbb{R}^{n\tau_n}$. Note that $\overline{\mathbb{P}}_{\mathbf{X}^t}(\mathbf{u}^t(i)) = 2^{-n\bar{R}_t^{(n)}}$ for $i \in [1{:}M_n]$ and $t \in [0{:}\tau_n]$ (recall that $\bar{R}_0^{(n)} = 0$). The decoder function $\mathbb{g}_{\tau_n}^{(n)}(\cdot)$ defines disjoint decoding regions $\{\mathcal{D}_i\}_{i=1}^{M_n}$ such that $\mathcal{D}_i \subseteq \mathbb{R}^{n\tau_n}$ and $\bigcup_{i=1}^{M_n} \mathcal{D}_i = \mathbb{R}^{n\tau_n}$. Set $\beta \triangleq 2^{-n\gamma}$ and note that

$$\begin{aligned} \frac{1}{n} i\big(\mathbf{x}_t^{\tau_n}; \mathbf{y}_t^{\tau_n} | \mathbf{x}^{t-1}\big) &= \frac{1}{n} \log_2 \frac{\overline{\mathbb{P}}_{\mathbf{X}_t^{\tau_n}|\mathbf{Y}_t^{\tau_n}, \mathbf{X}^{t-1}}(\mathbf{x}_t^{\tau_n}|\mathbf{y}_t^{\tau_n}, \mathbf{x}^{t-1})}{\overline{\mathbb{P}}_{\mathbf{X}_t^{\tau_n}|\mathbf{X}^{t-1}}(\mathbf{x}_t^{\tau_n}|\mathbf{x}^{t-1})} \\ &= \sum_{k=t}^{\tau_n} R_k^{(n)} + \frac{1}{n} \log_2 \overline{\mathbb{P}}_{\mathbf{X}_t^{\tau_n}|\mathbf{Y}_t^{\tau_n}, \mathbf{X}^{t-1}}(\mathbf{x}_t^{\tau_n}|\mathbf{y}_t^{\tau_n}, \mathbf{x}^{t-1}). \end{aligned} \tag{148}$$

[2] We use the convention that $\sum_{i=j}^{j-1} a_i = 0$ for all a_i and for all integers j.

The last equality follows because

$$\begin{aligned}\log_2 \overline{\mathbb{P}}_{\mathbf{X}_t^{\tau_n}|\mathbf{X}^{t-1}}(\mathbf{x}_t^{\tau_n}|\mathbf{x}^{t-1}) &= \log_2 \overline{\mathbb{P}}_{\mathbf{X}^{\tau_n}}(\mathbf{x}^{\tau_n}) - \log_2 \overline{\mathbb{P}}_{\mathbf{X}^{t-1}}(\mathbf{x}^{t-1}) && (149)\\ &= -n\bar{R}_{\tau_n}^{(n)} + n\bar{R}_{t-1}^{(n)} && (150)\\ &= -n\sum_{k=t}^{\tau_n} R_k^{(n)}. && (151)\end{aligned}$$

Consequently, we have

$$\overline{\mathbb{P}}\left[\frac{1}{n} i\Big(\mathbf{X}_t^{\tau_n}; \mathbf{Y}_t^{\tau_n}\Big|\mathbf{X}^{t-1}\Big) \le \sum_{k=t}^{\tau_n} R_k^{(n)} - \gamma\right] = \overline{\mathbb{P}}\Big[\overline{\mathbb{P}}_{\mathbf{X}_t^{\tau_n}|\mathbf{Y}_t^{\tau_n},\mathbf{X}^{t-1}}(\mathbf{X}_t^{\tau_n}|\mathbf{Y}_t^{\tau_n}, \mathbf{X}^{t-1}) \le \beta\Big]. \tag{152}$$

Define

$$\mathcal{B}_i = \Big\{\mathbf{y}^{\tau_n} \in \mathbb{R}^{\tau_n n} : \overline{\mathbb{P}}_{\mathbf{X}_t^{\tau_n}|\mathbf{Y}_t^{\tau_n},\mathbf{X}^{t-1}}\big(\mathbf{u}_t^{\tau_n}(i)|\mathbf{y}_t^{\tau_n}, \mathbf{u}^{t-1}(i)\big) \le \beta\Big\}. \tag{153}$$

We obtain a lower bound on $\overline{\mathbb{P}}[\mathcal{E}_n]$ through the following chain of inequalities

$$\begin{aligned}&\overline{\mathbb{P}}\left[\frac{1}{n} i\Big(\mathbf{X}_t^{\tau_n}; \mathbf{Y}_t^{\tau_n}\Big|\mathbf{X}^{t-1}\Big) \le \sum_{k=t}^{\tau_n} R_k^{(n)} - \gamma\right]\\ &= \sum_{i=1}^{M_n} \overline{\mathbb{P}}_{\mathbf{X}^{\tau_n},\mathbf{Y}^{\tau_n}}[\mathbf{u}(i), \mathcal{B}_i] && (154)\\ &= \sum_{i=1}^{M_n} \overline{\mathbb{P}}_{\mathbf{X}^{\tau_n},\mathbf{Y}^{\tau_n}}\Big[\mathbf{u}(i), \mathcal{B}_i \cap \mathcal{D}_i^{\mathsf{C}}\Big] + \sum_{i=1}^{M_n} \overline{\mathbb{P}}_{\mathbf{X}^{\tau_n},\mathbf{Y}^{\tau_n}}\left[\mathbf{u}(i), \mathcal{B}_i \cap \mathcal{D}_i\right] && (155)\\ &\le \frac{1}{M_n}\sum_{i=1}^{M_n} \overline{\mathbb{P}}_{\mathbf{Y}^{\tau_n}|\mathbf{X}^{\tau_n}}\Big(\mathcal{D}_i^{\mathsf{C}}|\mathbf{u}(i)\Big)\\ &\quad + \sum_{i=1}^{M_n}\int_{\mathcal{B}_i\cap\mathcal{D}_i} \overline{\mathbb{P}}_{\mathbf{Y}^{\tau_n},\mathbf{X}^{t-1}}\big(\mathbf{y}^{\tau_n}, \mathbf{u}^{t-1}(i)\big)\, \overline{\mathbb{P}}_{\mathbf{X}_t^{\tau_n}|\mathbf{Y}_t^{\tau_n},\mathbf{X}^{t-1}}\big(\mathbf{u}_t^{\tau_n}(i)|\mathbf{y}_t^{\tau_n}, \mathbf{u}^{t-1}(i)\big)\, \mathrm{d}\mathbf{y}^{\tau_n} && (156)\\ &\le \overline{\mathbb{P}}[\mathcal{E}_n] + \beta\sum_{i=1}^{M_n} \overline{\mathbb{P}}_{\mathbf{Y}^{\tau_n},\mathbf{X}^{t-1}}\big(\mathcal{B}_i \cap \mathcal{D}_i, \mathbf{u}^{t-1}(i)\big) && (157)\\ &\le \overline{\mathbb{P}}[\mathcal{E}_n] + \beta\sum_{i=1}^{M_n} \overline{\mathbb{P}}_{\mathbf{Y}^{\tau_n},\mathbf{X}^{t-1}}\big(\mathcal{D}_i, \mathbf{u}^{t-1}(i)\big) && (158)\\ &\le \overline{\mathbb{P}}[\mathcal{E}_n] + \beta. && (159)\end{aligned}$$

Here, (154) follows from (152) and (153), (156) follows because $\mathcal{B}_i \cap \mathcal{D}_i^{\mathsf{C}} \subseteq \mathcal{D}_i^{\mathsf{C}}$, because $\overline{\mathbb{P}}_{\mathbf{X}^{\tau_n},\mathbf{Y}^{\tau_n}}$ can be factorized as $\overline{\mathbb{P}}_{\mathbf{Y}^{\tau_n},\mathbf{X}^{t-1}}\overline{\mathbb{P}}_{\mathbf{X}_t^{\tau_n}|\mathbf{Y}_t^{\tau_n},\mathbf{X}^{t-1}}$; (157) follows from (153); and finally, (159) follows because $\{\mathcal{D}_i\}_{i=1}^{M_n}$ are disjoint sets. Since (159) holds for $t \in [1{:}\tau_n]$, we have established (147). ■

By Lemma 5 and (145), we have for all $n \ge n_1$

$$\begin{aligned}\overline{\mathbb{P}}[\mathcal{E}_n] &\ge \overline{\mathbb{P}}\left[\frac{1}{n} i\Big(\mathbf{X}_{\bar{k}_n}^{\tau_n}; \mathbf{Y}_{\bar{k}_n}^{\tau}\Big|\mathbf{X}^{\bar{k}_n-1}\Big) \le \sum_{k=\bar{k}_n}^{\tau_n} R_k^{(n)} - \gamma\right] - 2^{-n\gamma} && (160)\\ &\ge \overline{\mathbb{P}}\left[\frac{1}{n} i\Big(\mathbf{X}_{\bar{k}_n}^{\tau_n}; \mathbf{Y}_{\bar{k}_n}^{\tau_n}\Big|\mathbf{X}^{\bar{k}_n-1}\Big) \le \sum_{k=\bar{k}_n}^{\tau_n} C(h_k) + \gamma\right] - 2^{-n\gamma}. && (161)\end{aligned}$$

Next, by using the techniques in the proof of [31, Th. 3.7.4] to analyze the first term in (161), we find that

$$\lim_{n\to\infty} \overline{\mathbb{P}}\left[\frac{1}{n} i\Big(\mathbf{x}_{\bar{k}_n}^{\tau_n}; \mathbf{Y}_{\bar{k}_n}^{\tau_n}\Big|\mathbf{x}^{\bar{k}_n-1}\Big) < \sum_{k=t}^{\tau_n} C(h_k) + \gamma\right] = 1. \tag{162}$$

Using (162) in (161), we find that $\lim_{n\to\infty} \overline{\mathbb{P}}[\mathcal{E}_n] = 1$ as desired.

Appendix II

Proof of Lemma 2 (Achievability)

Define the random variable $U_n \in \mathcal{U}_n \triangleq \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \cdots$ by the probability distribution

$$P_{U_n} \triangleq \mathcal{P}_n \times \mathcal{P}_n \times \mathcal{P}_n \times \cdots \tag{163}$$

where $\mathcal{P}_n$ denotes probability density of $\sqrt{n}\tilde{\mathbf{X}}/\big|\big|\tilde{\mathbf{X}}\big|\big|$, where $\tilde{\mathbf{X}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, i.e., $\mathcal{P}_n$ denotes the uniform distribution on the n-dimensional sphere with radius $\sqrt{n}$. We use one realization of U_n to generate the encoder and decoder functions. Then, we show that the conditional probability of error averaged over U_n, $\{\mathbf{Z}_t\}_{t=1}^{\infty}$, and H^∞ given that $\max_{k\in[1:\tau_n]} \mathbb{u}_{k,\tau_n}^{(n)}(H^\infty) \leq -c_n$ tends to zero. Invoking the random coding argument then enables us to show that there must be at least one realization of U_n for each n such that the probability of error tends to zero as $n \to \infty$. Let the ith entry of $u \in \mathcal{U}_n$ be denoted by $u(i) \in \mathbb{R}^n$. By countability of $\mathbb{N}^2$, there exists a bijection between $\mathbb{N}^2$ and $\mathbb{N}$ defined by the mapping $\imath : \mathbb{N}^2 \mapsto \mathbb{N}$. The encoder functions $\mathbb{f}_{n,t}^{(r)}$, for $r \in \mathbb{R}_+$, are then defined in terms of $u \in \mathcal{U}_n$ as

$$\mathbb{f}_t^{(n)}(u, \mathbf{b}) = u\left(\imath\left(t, 1 + \sum_{i=1}^{\mathrm{len}(\mathbf{b})} b_i 2^{i-1}\right)\right) \tag{164}$$

for every $\mathbf{b} \in \mathfrak{B}$, where b_i is the ith entry of $\mathbf{b}$ and $\mathrm{len}(\cdot)$ denotes the length of a vector. The inner sum in (164) is a binary-to-integer conversion that converts the information bit vector $\mathbf{b}$ into an integer-valued index in the range $\big[1{:}2^{\mathrm{len}(\mathbf{b})}\big]$. Based on the above construction of the encoder, we have that (recall that $B_1^{\lceil n\bar{R}_t^{(n)}\rceil} = (B_1, \cdots, B_{\lceil n\bar{R}_t^{(n)}\rceil})$)

$$\mathbf{X}_t = \mathbb{f}_t^{(n)}\left(U_n, B_1^{\lceil n\bar{R}_t^{(n)}\rceil}\right). \tag{165}$$

In order to keep notation simple, we define for $\mathbf{b} \in \{0,1\}^{\lceil n\bar{R}_j^{(n)}\rceil}$ and $j, t \in \mathbb{N}$, $j \leq t$

$$\bar{\mathbf{X}}_{j:t}^{(n)}(u, \mathbf{b}) \triangleq \left[\mathbb{f}_j^{(n)}\left(u, b_1^{\lceil n\bar{R}_j^{(n)}\rceil}\right), \cdots, \mathbb{f}_t^{(n)}\left(u, b_1^{\lceil n\bar{R}_t^{(n)}\rceil}\right)\right]. \tag{166}$$

Let $\zeta_n \triangleq c_n/2$. For every $\mathbf{y}^t$, we define the threshold-based decoding functions as follows:

$$\begin{aligned} &g_t^{(n)}(u, \mathbf{y}^t, H^t) \\ &\triangleq \begin{cases} \mathbf{b} & \text{if } \exists!\mathbf{b} \in \{0,1\}^{\lceil n\bar{R}_t^{(n)}\rceil} \text{ s.t. } \forall j \in [1{:}t] : i\Big(\bar{\mathbf{X}}_{j:t}^{(n)}(u, \mathbf{b}); \mathbf{y}^t | H_1^t\Big) \geq n\sum_{k=j}^{t} R_k^{(n)} + n\zeta_n \\ [], & \text{otherwise.} \end{cases} \end{aligned} \tag{167}$$

Here, [] is the vector of length which indicates an error and we have defined the “mismatched” information density [32]

$$i(\mathbf{x};\mathbf{y}|h) \triangleq nC(h) + \frac{||\mathbf{y}||_2^2}{2(h+1)\log(2)} - \frac{\left|\left|\mathbf{y}-\sqrt{h}\mathbf{x}\right|\right|_2^2}{2\log(2)}. \tag{168}$$

We note that $\mathbb{E}[i(\mathbf{X}_t;\mathbf{Y}_t|H_t)|H_t=h] = nC(h)$ and by using the same arguments as in [32], we find that

$$\begin{aligned} V_n(h) &\triangleq \frac{1}{n}\mathrm{Var}[i(\mathbf{X}_t;\mathbf{Y}_t|H_t)|H_1=h] \\ &\overset{n\to\infty}{\to} \frac{h(h+2)}{2(h+1)^2\log^2(2)} && (169) \\ &\le \frac{1}{2\log^2(2)} && (170) \\ &\triangleq V. && (171) \end{aligned}$$

Thus, for sufficiently large n, we have that $V_n(h) \le 2V$ for all $h \in \mathbb{R}_+$.

It remains to analyze the probability of error. To do so, we rely on the technique used to prove Shannon's achievability bound in [33, Th. 17.1]. Assume, without loss of generality, that $B_i = 0$ for $i \in \mathbb{N}$. We define the "outage" events as follows

$$\mathcal{A}_j \triangleq \left\{ i(\bar{\mathbf{X}}_{j:\tau_n}^{(n)}(U_n,\mathbf{0});\mathbf{Y}_j^{\tau_n}|H_j^{\tau_n}) < n\sum_{k=j}^{\tau_n} R_k^{(n)} + n\zeta_n \right\}, \qquad j \in [1:\tau_n]\,. \tag{172}$$

Here, $\mathbf{0}$ denotes the all-zero vector (we omit specifying the length to keep notation simple). The "confusion" events are similarly defined by

$$\mathcal{B}(\mathbf{b}) \triangleq \bigcap_{j\in[1:\tau_n]} \left\{ i(\bar{\mathbf{X}}_{j:\tau_n}^{(n)}(U_n,\mathbf{b});\mathbf{Y}_j^{\tau_n}|H_j^{\tau_n}) \ge n\sum_{k=j}^{\tau_n} R_k^{(n)} + n\zeta_n \right\} \tag{173}$$

where $\mathbf{b} \in \{0,1\}^{\lceil n\bar{R}_{\tau_n}^{(n)}\rceil}$. Here, $\mathcal{A}_j$ is the event that the information density of the correct codeword does not exceed the threshold while $\mathcal{B}(\cdot)$ is the event that the information density of an incorrect codeword does exceed the threshold. Define the (random) set of information bit vectors for $k \in [1:\tau_n]$

$$\mathbb{B}_k \triangleq \left\{ \mathbf{b} \in \{0,1\}^{\lceil n\bar{R}_{\tau_n}^{(n)}\rceil} : b_1^{\lceil n\bar{R}_{k-1}^{(n)}\rceil} = \mathbf{0}, b_{\lceil n\bar{R}_{k-1}^{(n)}\rceil+1}^{\lceil n\bar{R}_{\tau_n}^{(n)}\rceil} \neq \mathbf{0} \right\}. \tag{174}$$

We also define $\mathbb{B}_{\tau_n+1} \triangleq \emptyset$. Here, we let $\bar{R}_0^{(n)} = 0$ such that $\mathbb{B}_1$ is the set of all binary vectors of length $\lceil n\bar{R}_{\tau_n}^{(n)}\rceil$ except the all-zero vector $\mathbf{0}$. Note that $|\mathbb{B}_k| = 2^{\lceil n(R_k+\cdots+R_{\tau_n})\rceil} - 1$ and that $\bar{\mathbf{X}}_{k:\tau_n}^{(n)}(U_n,\mathbf{b})$ and $\mathbf{Y}_k^{\tau_n}$ are conditionally independent for every $\mathbf{b} \in \mathbb{B}_k$ given $B_1^\infty = \mathbf{0}$ and H^∞. Define the error event

$$\mathcal{E}_n(U_n) \triangleq \left\{ \mathbb{g}_{\tau_n}^{(n)}(U_n,\mathbf{Y}_1^{\tau_n},H^{\tau_n}) \neq B_1^{\lceil n\bar{R}_{\tau_n}\rceil} \right\}. \tag{175}$$

Then, we obtain the following probability of error

$$\begin{aligned} &\mathbb{P}\left[\mathcal{E}_n(U_n)\middle|\bar{\mathcal{H}}_n\right] \\ &= \mathbb{P}\left[\mathcal{E}_n(U_n)\middle|B_1^\infty=\mathbf{0},\bar{\mathcal{H}}_n\right] && (176) \\ &= \mathbb{P}\left[\bigcup_{k=1}^{\tau_n}\mathcal{A}_k \cup \bigcup_{\bar{\mathbf{b}}\in\mathbb{B}_1}\mathcal{B}(\bar{\mathbf{b}})\middle|B^\infty=\mathbf{0},\bar{\mathcal{H}}_n\right] && (177) \end{aligned}$$

$$\leq \mathbb{P}\left[\bigcup_{k=1}^{\tau_n} \mathcal{A}_k \middle| B^\infty = \mathbf{0}, \bar{\mathcal{H}}_n\right] + \mathbb{P}\left[\bigcup_{\bar{\mathbf{b}} \in \mathbb{B}_1} \mathcal{B}(\bar{\mathbf{b}}) \middle| B^\infty = \mathbf{0}, \bar{\mathcal{H}}_n\right]. \tag{178}$$

Here, (176) follows from the symmetry of the EMS protocol, (177) follows from (167), (172), and (173); and (178) follows from the union bound. Next, we upper-bound each of the two terms in (178) separately. For the first term, we use the law of total expectation and the union bound to obtain

$$\mathbb{P}\left[\bigcup_{k=1}^{\tau_n} \mathcal{A}_k \middle| B^\infty = \mathbf{0}, \bar{\mathcal{H}}_n\right] = \mathbb{E}\left[\mathbb{P}\left[\bigcup_{k=1}^{\tau_n} \mathcal{A}_k \middle| B^\infty = \mathbf{0}, H^\infty\right] \middle| \bar{\mathcal{H}}_n\right] \tag{179}$$

$$= \mathbb{E}\left[\sum_{k=1}^{\tau_n} \mathbb{P}\left[\frac{1}{n} i(\mathbf{X}_k^{\tau_n}; \mathbf{Y}_k^{\tau_n}) < \sum_{j=k}^{\tau_n} R_j^{(n)} + \zeta_n \middle| H^\infty\right] \middle| \bar{\mathcal{H}}_n\right]. \tag{180}$$

For all h^∞ such that $\max_{k \in [1:\tau_n]} \mathbb{u}_{k,\tau_n}^{(n)}(h^{\tau_n}) \leq -c_n$, we upper-bound the inner probability in (180) for sufficiently large n using Chebyshev's inequality as follows

$$\mathbb{P}\left[\frac{1}{n} i(\mathbf{X}_k^{\tau_n}; \mathbf{Y}_k^{\tau_n} | H_k^{\tau_n}) < \sum_{j=k}^{\tau_n} R_j^{(n)} + \zeta_n \middle| H^\infty = h^\infty\right] \leq \mathbb{E}\left[\frac{2V(\tau_n - k + 1)}{n(\sum_{j=k}^{\tau_n}[C(H_j) - R_j^{(n)}] - \zeta_n)^2} \middle| H^\infty = h^\infty\right] \tag{181}$$

$$\leq \mathbb{E}\left[\frac{2V(\tau_n - k + 1)}{n(c_n - \zeta_n)^2} \middle| H^\infty = h^\infty\right] \tag{182}$$

$$= \mathbb{E}\left[\frac{8V(\tau_n - k + 1)}{n c_n^2} \middle| H^\infty = h^\infty\right] \tag{183}$$

$$\leq \frac{8V\bar{\tau}_n}{n c_n^2}. \tag{184}$$

Here, (181) follows from Chebyshev's inequality, from $\mathbb{E}[i(\mathbf{X}_t; \mathbf{Y}_t | H_t) | H_t] = nC(H_t)$, and from (171); (182) follows from $\max_{k \in [1:\tau_n]} \mathbb{u}_{k,\tau_n}^{(n)}(h^{\tau_n}) \leq -c_n$; and (184) follows from $\tau \leq \bar{\tau}_n$. As a result of (180) and (184), we have

$$\mathbb{P}\left[\bigcup_{k=1}^{\tau_n} \mathcal{A}_k \middle| B^\infty = \mathbf{0}, \bar{\mathcal{H}}_n\right] \leq \mathbb{E}\left[\sum_{k=1}^{\tau_n} \frac{8V\bar{\tau}_n}{n c_n^2} \middle| \bar{\mathcal{H}}_n\right] \leq \frac{8\bar{\tau}_n^2 V}{n c_n^2}. \tag{185}$$

Next, the second term in (178) is upper-bounded as follows

$$\mathbb{P}\left[\bigcup_{\bar{\mathbf{b}} \in \mathbb{B}_1} \mathcal{B}(\bar{\mathbf{b}}) \middle| B^\infty = \mathbf{0}, \bar{\mathcal{H}}_n\right]$$

$$= \mathbb{E}\left[\mathbb{P}\left[\bigcup_{\bar{\mathbf{b}} \in \mathbb{B}_1} \mathcal{B}(\bar{\mathbf{b}}) \middle| B^\infty = \mathbf{0}, H^\infty\right] \middle| \bar{\mathcal{H}}_n\right] \tag{186}$$

$$= \mathbb{E}\left[\mathbb{P}\left[\bigcup_{j=1}^{\tau_n} \bigcup_{\bar{\mathbf{b}} \in \mathbb{B}_j \setminus \mathbb{B}_{j+1}} \bigcap_{q \in [1:t]} \left\{ i\left(\bar{\mathbf{X}}_{q:\tau_n}^{(n)}(U, \bar{\mathbf{b}}); \mathbf{Y}_q^{\tau_n}\right) \geq n \sum_{k=q}^{\tau_n} R_k^{(n)} + n\zeta_n \right\} \middle| B^\infty = \mathbf{0}, H^\infty\right] \middle| \bar{\mathcal{H}}_n\right] \tag{187}$$

$$\leq \mathbb{E}\left[\mathbb{P}\left[\bigcup_{j=1}^{\tau_n} \bigcup_{\bar{\mathbf{b}} \in \mathbb{B}_j \setminus \mathbb{B}_{j+1}} \left\{ i\left(\bar{\mathbf{X}}_{j:\tau_n}^{(n)}(U, \bar{\mathbf{b}}); \mathbf{Y}_j^{\tau_n}\right) \geq n \sum_{k=j}^{\tau_n} R_k^{(n)} + n\zeta_n \right\} \middle| B^\infty = \mathbf{0}, H^\infty\right] \middle| \bar{\mathcal{H}}_n\right] \tag{188}$$

$$\leq \mathbb{E}\left[\sum_{j=1}^{\tau_n} 2^{\lceil n \sum_{k=j}^{\tau_n} R_k^{(n)} \rceil} \mathbb{P}\left[i(\bar{\mathbf{X}}_j^{\tau_n}; \mathbf{Y}_j^{\tau_n}) \geq n \sum_{k=j}^{\tau_n} R_k^{(n)} + n\zeta_n \middle| H^\infty\right] \middle| \bar{\mathcal{H}}_n\right] \tag{189}$$

$$\leq \mathbb{E}\left[\sum_{j=1}^{\tau_n} 2^{\lceil n \sum_{k=j}^{\tau_n} R_k^{(n)} \rceil} 2^{-\left(n \sum_{k=j}^{\tau_n} R_k^{(n)} + n\zeta_n\right)}\right] \tag{190}$$

$$\leq \mathbb{E}\left[\sum_{j=1}^{\tau_n} 2^{-n\zeta_n+1}\right] \tag{191}$$

$$= \bar{\tau}_n 2^{-nc_n/2+1}. \tag{192}$$

Here, (186) follows from the law of total expectation; (187) follows from (173) and (174); (188) follows from the union bound and because $|\mathbb{B}_j| = (2^{\lceil n \sum_{k=j}^{\tau_n} R_k \rceil} - 1)$; (189) follows by defining the random variables $\{\bar{\mathbf{X}}_t\}_{t=1}^{\infty}$ independently according to the probability distribution $\mathcal{P}_n$ such that they are independent of $\{\mathbf{X}_t\}_{t=1}^{\infty}$ and $\{\mathbf{Z}_t\}_{t=1}^{\infty}$; finally, (190) follows from [33, Cor. 17.1]. Consequently, we have shown that

$$\mathbb{P}\left[\mathcal{E}_n(U_n) \middle| \bar{\mathcal{H}}_n\right] \leq \bar{\tau}_n 2^{-nc_n/2+1} + \frac{8\bar{\tau}_n^2 V}{nc_n^2}. \tag{193}$$

Therefore, it follows that there exists a deterministic sequence $\{u_n^*\}_{n=1}^{\infty}$ such that

$$\mathbb{P}\left[\mathcal{E}_n(u_n^*) \middle| \bar{\mathcal{H}}_n\right] \leq \bar{\tau}_n 2^{-nc_n/2+1} + \frac{8\bar{\tau}_n^2 V}{nc_n^2}. \tag{194}$$

Define

$$p_{\mathrm{min},n} \triangleq \min_{\substack{t \in [1:\bar{\tau}_n]:\\ \mathbb{P}[\tau_n = t|\mathcal{H}_n] > 0}} \mathbb{P}\left[\tau_n = t | \bar{\mathcal{H}}_n\right]. \tag{195}$$

The condition in (23) implies that $p_{\mathrm{min},n} \geq g_n$ for all sufficiently large n and therefore we have

$$\begin{aligned}
&\lim_{n\to\infty} \max_{\substack{t \in [1:\tau_n]:\\ \mathbb{P}\left[\tau_n = t|\bar{\mathcal{H}}_n\right] > 0}} \mathbb{P}\left[\mathcal{E}_n(u_n^*) \middle| \bar{\mathcal{H}}_n, \tau_n = t\right] \\
&\leq \lim_{n\to\infty} \frac{1}{p_{\mathrm{min},n}} \sum_{\substack{t \in [1:\bar{\tau}_n]:\\ \mathbb{P}\left[\tau_n = t|\bar{\mathcal{H}}_n\right] > 0}} \mathbb{P}\left[\tau_n = t|\bar{\mathcal{H}}_n\right] \mathbb{P}\left[\mathcal{E}_n(u_n^*) \middle| \bar{\mathcal{H}}_n, \tau_n = t\right] && (196)\\
&\leq \lim_{n\to\infty} \frac{1}{g_n} \mathbb{P}\left[\mathcal{E}_n(u_n^*) \middle| \bar{\mathcal{H}}_n\right] && (197)\\
&\leq \lim_{n\to\infty} \frac{1}{g_n} \left(\bar{\tau}_n 2^{-nc_n/2+1} + \frac{8\bar{\tau}_n^2 V}{nc_n^2}\right) && (198)\\
&= 0. && (199)
\end{aligned}$$

Here, (197) follows from (23) and the law of total probability. The last equality follows from (21), from the upper bound $\mathrm{e}^x \leq 1/(1 - x + x^2/2)$ which holds for $x \leq 0$, and because $\bar{\tau}_n$ is a nondecreasing sequence.

References

[1] P. Popovski, "Delayed channel state information: Incremental redundancy with backtrack retransmission," in *Proc. IEEE ICC*, Jun. 2014.

[2] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. New York: Cambridge Univ. Press, 2005.

[3] 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.

[4] R. Wolff, *Stochastic modeling and the theory of queues*. New York: Prentice Hall, 1989.

[5] M. Zorzi and R. R. Rao, "On the use of renewal theory in the analysis of ARQ protocols," *IEEE Trans. Commun.*, vol. 44, no. 9, pp. 1077–1081, sep 1996.

[6] D. Tuninetti, “Transmitter channel state information and repetition protocols in block fading channels,” in *Proc. IEEE Inf. Theory Workshop (ITW)*, Lake Tahoe, Sep. 2007, pp. 505–510.

[7] ——, “On the benefits of partial channel state information for repetition protocols in block fading channels,” *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5036–5053, Aug. 2011.

[8] L. Szczecinski, C. Correa, and L. Ahumada, “Variable-rate retransmissions for incremental redundancy hybrid ARQ,” Feb. 2012. [Online]. Available: https://arxiv.org/pdf/1207.0229.pdf

[9] S. M. Kim, W. Choi, T. W. Ban, and D. K. Sung, “Optimal rate adaptation for hybrid ARQ in time-correlated Rayleigh fading channels,” *IEEE Trans. Wireless Commun.*, vol. 10, no. 3, pp. 968–979, Mar. 2011.

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

[11] M. Jabi, A. E. Hamss, L. Szczecinski, and P. Piantanida, “Multipacket hybrid ARQ: Closing gap to the ergodic capacity,” *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 5191–5205, Dec. 2015.

[12] M. Jabi, L. Szczecinski, M. Benjillali, and F. Labeau, “Outage minimization via power adaptation and allocation in truncated hybrid ARQ,” *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 711–723, Mar. 2015.

[13] S. Shamai, “A broadcast strategy for the Gaussian slowly fading channel,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 1997, pp. 150–154.

[14] A. Steiner and S. Shamai, “Multi-layer broadcasting hybrid-ARQ strategies for block fading channels,” *IEEE Trans. Wireless Commun.*, vol. 7, no. 7, pp. 2640–2650, Jul. 2008.

[15] M. A. Maddah-Ali and D. Tse, “Completely stale transmitter channel state information is still very useful,” *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4418–4431, jul 2012.

[16] G. Cocco, D. Gunduz, and C. Ibars, “Streaming transmission over block fading channels with delay constraint,” *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4315–4327, Sep. 2013.

[17] A. Khisti and S. Draper, “The streaming-DMT of fading channels,” *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 7058–7072, Nov. 2014.

[18] C. Hausl and A. Chindapol, “Hybrid ARQ with cross-packet channel coding,” *IEEE Commun. Lett.*, vol. 11, no. 5, pp. 434–436, May 2007.

[19] J. Chui and A. Chindapol, “Design of cross-packet channel coding with low-density parity-check codes,” in *Proc. IEEE Information Theory Workshop on Information Theory for Wireless Networks*, Jul. 2007, pp. 1–5.

[20] 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.

[21] K. Trillingsgaard and P. Popovski, “Block-fading channels with delayed CSIT at finite blocklength,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2014, pp. 2062–2066.

[22] K. D. Nguyen, R. Timo, and L. K. Rasmussen, “Causal-CSIT rate adaptation for block-fading channels,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 351–355.

[23] A. Goldsmith, *Wireless Communications*. New York, NY, USA: Cambridge Univ. Press, 2005.

[24] E. Dahlman, S. Parval, and J. Skold, *4G LTE/LTE-Advanced for Mobile Broadband*. Academic: New York, 2014.

[25] P. Billingsley, *Probability and Measure, Anniversary Ed.* Hoboken, NJ, USA: Wiley, 2012.

[26] S. Verdu and S. Shamai, “Variable-rate channel capacity,” *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2651–2667, Jun. 2010.

[27] S. M. Ross, *Introduction to Stochastic Dynamic Programming*. New York, NY, USA: Academic Press, 1995.

[28] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge Univ. Press, 2004.

[29] S. Rahbar and E. Hashemizadeh, “A computational approach to the Fredholm integral equation of the second kind,” in *Proc. World Congress on Engineering*, jul 2008.

[30] S. Verdu and T. S. Han, “A general formula for channel capacity,” *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.

[31] T. S. Han, *Information Spectrum Methods in Information Theory*. Berlin: Springer-Verlag, 2003.

[32] J. Scarlett, V. Y. F. Tan, and G. Durisi, “The dispersion of nearest-neighbor decoding for additive non-Gaussian channels,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 2664–2668.

[33] Y. Polyanskiy and Y. Wu, “Lecture notes on information theory,” Jun. 2016.