On the Benefits of Partial Channel State Information for Repetition Protocols in Block Fading Channels

Daniela Tuninetti

***Abstract*—This paper studies the throughput performance of HARQ (hybrid automatic repeat request) protocols over block fading Gaussian channels. It proposes new protocols that use the available feedback bit(s) not only to request a retransmission, but also to inform the transmitter about the instantaneous channel quality. An explicit protocol construction is given for any number of retransmissions and any number of feedback bits. The novel protocol is shown to simultaneously realize the gains of HARQ and of power control with partial CSI (channel state information). Remarkable throughput improvements are shown, especially at low and moderate SNR (signal to noise ratio), with respect to protocols that use the feedback bits for retransmission request only. In particular, for the case of a single retransmission and a single feedback bit, it is shown that the repetition is not needed at low SNR where the throughput improvement is due to power control only. On the other hand, at high SNR, the repetition is useful and the performance gain comes form a combination of power control and ability of make up for deep fades.**

***Index Terms*—Block fading channel; Hybrid ARQ; Partial channel state information; Power Control; Throughput;**

I. Introduction

In today networks, error correction is achieved by a combination of FEC (forward error correction) and ARQ (automatic repetition request). In classical ARQ protocols, a receiver requests a retransmission (sends a negative acknowledgment, or NACK) when an error is detected, and a positive acknowledgment (ACK) otherwise. In this work we explore the performance gain achievable by using the retransmission request bit(s) to signal to the transmitter the decoder status *and* the actual channel state, albeit coarsely. Our goal is to simultaneously enable the performance gain due to HARQ (hybrid automatic repeat request), i.e., a combination of ARQ and FEC error control methods [3], and to power control at the transmitter [4].

A. A Motivating Example

Consider a fixed rate transmission scheme over a block fading Gaussian channel with unit noise power spectral density. Let the transmit power in slot t, $t \in \mathbb{N}$, be Q_t, the fading power gain be γ_t, and the transmission rate be R. The receiver fails to decode when the instantaneous channel capacity is below the transmission rate [5], in which case it feeds back a NACK to the transmitter. A NACK is thus equivalent to

$$\log(1+\gamma_t Q_t) < R \Longleftrightarrow \gamma_t < \frac{\mathrm{e}^R - 1}{Q_t},$$

that is, a NACK is a *1-bit quantization of the channel state information* (CSI) γ_t sent to the transmitter when the transmitter no longer needs it (as already remarked in [6], [7]). This simple observation raises the question investigated in this work: whether it is optimal, in some sense, to feedback a ACK/NACK at the end of a slot, or whether the same feedback resources should rather be used at the beginning of the slot to inform the transmitter about the instantaneous channel quality, albeit coarsely.

To further gain insights into the problem, consider the outage capacity [5] as the performance measure. For a fixed positive parameter s, let the transmission rate be parametrized as $R = \log(1+\overline{P}\,s)$, where $\overline{P}$ denotes the average SNR $\overline{P} = \mathbb{E}[Q_t]$. As explained before, a 1-bit feedback used for ACK/NACK at end of slot t indicates to the transmitter that $\gamma_t < s$ if a NACK is received, or that $\gamma_t \geq s$ if a ACK is received. The probability of successful decoding is then the probability of receiving a ACK; thus the outage capacity, or long term average successfully decoded rate, is

$$\eta_{\mathrm{ACK}} = \Pr[\gamma_t \geq s] \log(1+\overline{P}\,s).$$

On the other hand, consider the case where the 1-bit of feedback is used at the beginning of the slot to indicate to the transmitter which of the events, $\{\gamma_t < s\}$ or $\{\gamma_t \geq s\}$, has occurred. In this case the transmitter can use this information as follows. It turns transmission off (i.e., $Q_t = 0$) if the channel is bad (i.e., $\gamma_t < s$) and it sends with power $Q_t = \frac{\mathrm{e}^R-1}{s}$ if the channel is good (i.e, $\gamma_t \geq s$); the chosen transmit power $Q_t = \frac{\mathrm{e}^R-1}{s}$ is such that no outage occurs. The average transmit power of this simple power control policy based on 1-bit CSI is $\overline{P} = \frac{\mathrm{e}^R-1}{s} \Pr[\gamma_t \geq s]$ and the outage capacity is

$$\eta_{\mathrm{CSI}} = \Pr[\gamma_t \geq s] \log\left(1+\overline{P}\frac{s}{\Pr[\gamma_t \geq s]}\right).$$

It is immediate to see that, for the same set of parameters $\overline{P}$ and s, and with 1-bit of feedback in both scenarios, the outage capacity with CSI η_{CSI} is larger than the outage capacity with ACK/NACK η_{ACK}. This observation reinforces the idea that using the 1-bit feedback to signal ACK/NACK is not optimal in general. The question whether this conclusion changes if retransmissions are allowed is investigated in this paper.

B. Past Work

To the best of the author's knowledge, past work available in the literature considering quantized and/or noisy CSI only focused on outage capacity, or on outage probability, or on expected capacity, but not on HARQ protocols.

Dr. D. Tuninetti is with the Electrical and Computer Engineering Department of the University of Illinois at Chicago, Chicago, IL 60607, USA. Parts of this work were presented at [1], [2]. This work was partially funded by NSF under award number 0643954. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the NSF.

For example, in [8] the authors consider power control policies for minimizing the outage probability with partial CSI; the derived power policy shows benefits with respect to the case of complete absence of CSI even if the channel knowledge is noisy and/or partial; the benefits are more pronounced at low SNR. In [9], the authors considered the outage capacity with the so called "broadcast approach", that is, a multiple-layer coding scheme with infinite many layers, where the receiver decodes as many layers as possible given the actual channel fading; it is found that even 1-bit of CSI helps to improve performance.

In [10], the authors studied the ergodic capacity of channels with states where the state is only partially known at the transmitter; for the Gaussian channel with quantized CSI, they showed that the capacity achieving power allocation is of the waterfilling type. In [6], the authors studied the expected capacity with quantized CSI and multiple-layer coding schemes; they showed that multiple-layer transmission offers limited benefits when power control at the transmitter is possible.

In [7], the authors considered the DMT (diversity multiplexing tradeoff) of multi-antenna channels with HARQ; in this setting the feedback is only used to signal ACK/NACK and not to perform power control, even though the transmitter is allowed to vary the transmit power across retransmissions; it is found that HARQ improves the DMT by a factor proportional to the maximum number of repetitions.

C. Contributions

In this work we consider the joint design of HARQ protocols and power control for block fading Gaussian channels. We use the *long-term average decoded rate* [3], or simply *throughput* for brevity in the following, as a measure of performance. The throughput captures the fundamental performance limits when strict delay constraints are imposed and includes the outage capacity and the ergodic capacity as special cases.

When considering HARQ protocols, it is customary to assume that the transmitter has no knowledge of the instantaneous fading and it thus transmits with equal power in every slot [3]. However, especially with INR (incremental redundancy), the probability of having to transmit m channel packets per data packet is decreasing with m [7]. Thus, it is conceivable that using more power in earlier transmissions of the same data packet reduces the probability of decoding failure and hence increases the throughput for the same average transmit power. Moreover, if the fading is known at the transmitter, more power can be used in the most favorable channel conditions–assuming power control is possible. As pointed out in [11], in delay constrained scenarios, the assumptions about the dynamics of the fading process with respect to the code length, as well as the duration over which power constraints are enforced, are critical. Here, in order to enable power allocation, we consider a power constraint imposed over a time horizon comprising many slots (i.e., much bigger than the maximum number of retransmissions allowed), commonly referred to as *long-term average power constraint* [12].

As opposed to classical HARQ protocols, where the ACK/NACK feedback bit is sent at the end of the slot, we consider here systems where the feedback bit(s) can be sent back at any point in time during a slot; the feedback can be used to signal CSI, or ACK/NACK, or any combination of them. The only restriction we impose is that the feedback bit(s) cannot be carried over from slot to slot.

Our contributions can be summarized as follows:

1) We propose novel HARQ protocols where the CSI and the ACK/NACK information are combined within the same feedback bit(s) in order to *realize simultaneously the gains due to HARQ and the gains of power control with partial CSI*. The main idea behind the proposed protocols is that the receiver sends back to the transmitter the index of the smallest power level that will allow successful decoding in the current slot. Our protocols are time-varying quantizers for a suitably scaled version of the channel fading, where the scaling factor accounts for the information already available at the receiver from the past transmissions.
2) We show that the throughput performance of the proposed class of protocols with *perfect CSI* can be obtained from dynamic programing [13].
3) By numerical evaluations of the throughput for Rayleigh fading channels, we show that repetitions are not needed at low SNR, and that the improvement over classical HARQ protocols (that use the feedback bit for ACK/NACK only) is entirely due to ability to perform power control.
   At high SNR, repetitions are useful, and the performance improvement over classical HARQ protocols comes form a combination of power control and the ability to make up for atypical long deep fades with repetitions.
   Our numerical results show that our protocols outperform classical HARQ at all SNRs.
4) We also have the following side results: (a) we show that the optimal power allocation for the outage capacity with partial CSI consists of a quantizer of the fading gain where the quantization regions are union of intervals, rather than intervals; to the best of the author's knowledge–was not reported before; (b) we present novel bounding techniques for to compute certain probabilities that are needed for the throughput evaluation; these techniques are useful for numerical optimizations. In particular, a technique based on considerations on the order statistics of an independent sample of negative exponential random variables is of interest in its own.

D. Paper Organization

The rest of paper is organized as follows. Section II introduces the system model and Section III evaluates the throughput; Section IV revises the ergodic capacity with partial CSI, which serves as an upper bound for any HARQ protocol; Section V derives the outage capacity with partial CSI, which serves as a lower bound for any HARQ protocol; Section VI proposes a new class of HARQ protocols that combine repetitions and power control for any number of retransmissions and any number of feedback bits; Section VII proposes a novel bounding technique for the throughput based

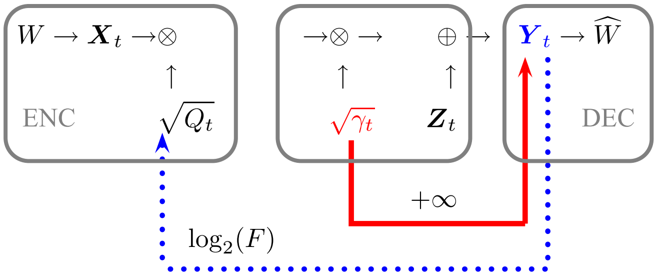

Fig. 1. The block-fading Gaussian channel model with partial CSI at the encoder (dotted blue line) and perfect CSI at the decoder (solid red line).

on ideas from order statistics; Section VIII compares the throughput performance of our new protocols with that of classical HARQ protocols for the Rayleigh fading channel; Section IX concludes the paper and points out open questions and future work directions.

II. System Model and Performance Metric

We adopt the following notation convention: $[x]^+$ indicates $\max\{x, 0\}$, $1_{\{x\in A\}}$ the indicator function (that equals one whenever $x \in A$ and zero otherwise), and $F_X(x) = \Pr[X \leq x]$, $x \in \mathbb{R}$, is the cumulative distribution function of the random variable X. In the following slot, fading block and codeword length are used interchangeably.

This work considers the single-user block-fading Gaussian channel depicted in Fig. 1. The received signal vector in slot t, $t \in \mathbb{N}$, is

$$\boldsymbol{Y}_t = \sqrt{\gamma_t\ Q_t}\boldsymbol{X}_t + \boldsymbol{Z}_t \in \mathbb{C}^L,$$

where: the noise $\boldsymbol{Z}_t$ is a length-L proper-complex white Gaussian random vector with zero mean and unit variance, the channel fading power gain γ_t is a scalar with $\mathbb{E}[\gamma_t] = 1$, the channel input signal $\boldsymbol{X}_t$ has Gaussian iid (independent and identically distributed) components with zero mean and unit variance, and $Q_t \geq 0$ is the transmit power. Each codeword $\boldsymbol{X}_t$ spans one fading block over which the fading gain stays constant. The slot length L is such that it suffices to guarantee reliable communication if the accumulated mutual information at the receiver is above the communication rate [5].

The fading gain γ_t changes in an iid fashion from slot to slot. The receiver has perfect instantaneous knowledge of γ_t at the beginning of the slot. The transmitter however does not know γ_t, unless explicitly informed by the receiver. For this reason, we assume that the transmitter cannot adjust the communication rate in each slot and thus it sends at a fixed rate. The (partial) CSI possibly available at the transmitter is used only for power allocation across slots.[1]

[1]We note that in practice, fading can be considered independent from slot to slot only if the slots are separated in time by at least few channel coherence times [14]. This is not a problem in multi-user systems where a user is assigned a transmission slot in every frame (and frames consist of several slots). In this case γ_2 indicates the fading gain of the slot where the second transmission (first re-transmission) occurs and it needs not be the slot immediately following the one where the first transmission occurred. When the iid assumption does not hold, fading correlation across time slots can be easily incorporated in our model by substituting products of probabilities involving different fading random variables with the corresponding joint probabilities.

A delay-free and error-free feedback channel with capacity $\log_2(F)$ bits per slot is available for communication of low-rate information between the receiver and the transmitter; the receiver can feedback a retransmission request to the transmitter at the end of a slot, or quantized CSI at the beginning of a slot, or any other information representable on $\log_2(F)$ bits at any point during the slot. We do not allow feedback bits to be accumulated over successive slots. The case $F = 1$ corresponds to absence of CSI at the transmitter, while $F = +\infty$ corresponds to perfect CSI.

In a block-fading setting, reliable communication is possible if the accumulated mutual information at the receiver is above the communication rate [5]. To make up for decoding errors, which occur when the channel is in deep fade, the transmitter can retransmit a data packet at most $M-1$ times, that is, each data packet can be transmitted on at most M channel slots. We consider the three HARQ protocols analyzed in [3]:

- **ALO** (ALOha): like in slotted Aloha, the transmitter keeps sending the same codeword and the receiver attempts decoding by using only the most recently received codeword.
- **RTD** (Repetition Time Diversity): the transmitter keeps sending the same codeword and the receiver performs maximal ratio combining of all the received packets, thus realizing Repetition Time Diversity.
- **INR** (INcremental Redundancy): at each retransmission request, the transmitter sends new redundancy bits and the receiver optimally combines them.

The protocols work as follows. In order to send a data packet of b bits, the transmitter can use at most LM channel uses. For ALO and RTD, the transmitter encodes the data packet with a Gaussian channel code of rate b/L and then concatenates it with a repetition code of rate $1/M$. For INR, the transmitter encodes the data with a Gaussian channel code of rate $b/(LM)$ and at each transmission it send a different chuck of L symbols. Following [3], we let $R \triangleq b/L$. The throughput is defined as the *long-term average number of successfully decoded bits per channel use*:

$$\eta_{M,F} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} R \cdot 1_{\{\text{successful decoding in slot } t\}}, \quad (1)$$

where the subscript M indicates that at most M transmissions are allowed per data packet, and the subscript F indicates that at most F feedback values are allowed. The ergodic capacity and the outage capacity are a spacial case of our framework for $M = +\infty$ and $M = 1$, respectively. For future use, decoding fails with m transmissions (and thus a retransmission is needed) if:

$$\left\{\sum_{t=1}^{m} \log(1+\gamma_t Q_t) < R\right\} \text{ for INR} \quad (2a)$$

$$\left\{\log(1+\sum_{t=1}^{m} \gamma_t Q_t) < R\right\} \text{ for RTD} \quad (2b)$$

$$\bigcup_{t=1}^{m} \left\{\ \log(1+\gamma_t Q_t) < R\right\} \text{ for ALO,} \quad (2c)$$

since INR accumulates mutual information, RTD accumulates SNR, and ALO only accounts for the most recent transmission.

In order to complete the system description, we need to specify how the transmit power Q_t, $t \in \mathbb{N}$, can be varied. We assume $Q_t \in \{P_{m,f}\}$, where $P_{m,f} \geq 0$ is the power used at the m-th transmission attempt, $m \in \{1, \ldots, M\}$, when the feedback value is f, $f \in \{0, \ldots, F-1\}$. The total transmit power must satisfy the *long-term average constraint [12]* defined as:

$$\lim_{T\to\infty} \sum_{t=1}^{T} \frac{1}{T} Q_t \leq \overline{P}, \text{ almost surely.} \tag{3}$$

For the normalization adopted in this work, $\overline{P}$ has the meaning of average SNR at the receiver. The power allocation policies are *causal* [11], [15] in that Q_t can only depend on the (partial) knowledge of $(\gamma_1, \ldots, \gamma_t)$.

Next we show how to evaluate the throughput in (1) subject to the power constraint in (3).

III. Throughput Evaluation

In [3] we introduced a general framework to analyze the performance of HARQ protocols based on the *renewal-reward theory* [16], which we shall use now to evaluate the throughput. In our system, a renewal event occurs (i.e., the system starts anew without any memory from the past) when the transmission of a data packet ends (either because of successful decoding with less than M transmissions or because the maximum number of transmissions has been reached). From [3], the system performance is completely characterized by the triplet $(\mathcal{T}, \mathcal{R}, \mathcal{P})$, where:

- $\mathcal{T} \in \{1, \ldots, M\}$ in the *inter-renewal time* and represents the number of slots needed to complete the transmit of a data packet;
- $\mathcal{R} \in \{0, R\}$ is the *reward*, i.e., the number of bits successfully decoded per channel use when the transmission of a data packet ends;
- $\mathcal{P}$ is the *cost*, i.e., the total transmit power for a data packet (including all retransmissions);

Given $(\mathcal{T}, \mathcal{R}, \mathcal{P})$, the throughput in (1) subject to the power constraint in (3) is given by:

Theorem 1 (from [3], [7])**.** *For any M and F the throughput $\eta_{M,F}$ for a given power $\overline{P}$ is the solution of:*

$$\eta^{(\star)}_{M,F} = $$
$$= \max \frac{\mathbb{E}[\mathcal{R}]}{\mathbb{E}[\mathcal{T}]} = \max \frac{R\big(1 - \Pr[\mathcal{T} = M, \text{failure to decode}]\big)}{\sum_{m=1}^{M} \Pr[\mathcal{T} \geq m]}$$
$$\text{s.t. } \frac{\mathbb{E}[\mathcal{P}]}{\mathbb{E}[\mathcal{T}]} = \frac{\sum_{m=1}^{M} \mathbb{E}[\mathcal{P}_m | \mathcal{T} \geq m] \Pr[\mathcal{T} \geq m]}{\sum_{m=1}^{M} \Pr[\mathcal{T} \geq m]} \leq \overline{P},$$

where the maximization is over the transmit rate $R \geq 0$ and over the power allocation $\{\mathcal{P}_m\}_{m=1}^{M}$, where $\mathcal{P}_m \in \{P_{m,0}, \ldots, P_{m,F-1}\}$ is the causal power policy for the m-th transmission attempt restricted to take at most F different values. The distribution of the inter-renewal time $\mathcal{T}$ and the probability of failure to decode (see (2)*) are function of the protocol $\star \in$ {ALO, RTD, INR} used.*

Proof: The proof can be found in Appendix A. ■

Remarks:

1) The probability of failure to decode on the last transmission, indicated as $\Pr[\mathcal{T} = M$, failure to decode$]$, is the probability that the data packet is lost and it is referred to as *outage probability*.
2) The event $\{\mathcal{T} \leq M - 1\}$ (that transmission ends before the maximum number of transmissions has been reached) implies successful decoding. However, a successful decoding does not necessarily imply a renewal event (the end of the transmission of the current data packet).
3) It is immediate to see that the optimal power allocation meets the power constraint with equality (otherwise, the left over power could be used on the last transmission, which would increasing the throughput while still meeting the power constraint).

The throughput of the different protocols for different values of M and F satisfies:

Theorem 2 (from [3])**.** *We have:*

$$\eta^{(\mathrm{ALO})}_{M,F} \leq \eta^{(\mathrm{RTD})}_{M,F} \leq \eta^{(\mathrm{INR})}_{M,F}. \tag{4}$$

Moreover, $\eta^{(\star)}_{M,F}$ is a non-decreasing function of M and of F, for each protocol $\star \in$ {ALO, RTD, INR}.

Proof: The proof of (4) is as in [3] and is omitted here for sake of space. The fact that $\eta^{(\star)}_{M,F}$ is a non-decreasing function of M and F follows by observing that if more transmissions or more accurate CSI would hurt performance, they could just be ignored. ■

From Theorem 2 it follows immediately that:

Corollary 3. *For any M and F, and for any protocol $\star \in$ {ALO, RTD, INR}:*

$$\eta^{(\mathrm{ALO})}_{M=1,F} \leq \eta^{(\star)}_{M,F} \leq \eta^{(\mathrm{INR})}_{M=\infty,F},$$

where $\eta^{(\mathrm{ALO})}_{M=1,F}$ is the outage capacity *of the channel and $\eta^{(\mathrm{INR})}_{M=\infty,F}$ is the* ergodic capacity *of the channel [3].*

In the following we first evaluate $\eta^{(\mathrm{INR})}_{M=\infty,F}$ and $\eta^{(\mathrm{ALO})}_{M=1,F}$ with partial CSI and then we propose novel achievable protocols for $\eta^{(\star)}_{M,F}$, $\star \in$ {ALO, RTD, INR}.

IV. Throughput upper bound $\eta^{(\mathrm{INR})}_{M=\infty,F}$

When $M \to \infty$, the INR protocol with a time-invariant and memoryless power allocation policy $P_t = g(\gamma_t)$, $t \in \mathbb{N}$, and with optimized rate R, achieves the *ergodic capacity* of the channel [3] given by:

$$\eta^{(\mathrm{INR})}_{M=\infty,F} = \mathbb{E}[\log(1 + \gamma\, g(\gamma))],$$

where the function $g(\cdot)$ can take at most F different values. In the following we let $P_f \geq 0$ be the power used when the feedback value is f, $f \in \{0, \ldots, F-1\}$, (i.e., we drop the index referring to the number of repetitions, which is irrelevant here because we considered time-invariant and memoryless power allocation policies). The optimal power policy $g(\cdot)$ for a finite F was derived in [10] and it is summarized in the following:

Theorem 4 (from [10])**.** *Let*

$$0 \le P_0 \le P_1 \le \cdots \le P_{F-1} \le \frac{1}{\lambda}, \tag{5}$$

where P_f is the power used when $\gamma \in \mathcal{R}_f^{(\mathrm{INR})}$, with

$$\mathcal{R}_f^{(\mathrm{INR})} \triangleq \{s_f \le \gamma < s_{f+1}\}, \tag{6}$$

$f \in \{0, \ldots, F-1\}$, with $s_0 = 0$, $s_F = +\infty$, and

$$\frac{1}{s_f} \triangleq \frac{1}{\lambda}\left(\frac{\lambda(P_{f+1}-P_f)}{\mathrm{e}^{\lambda(P_{f+1}-P_f)}-1} - \lambda P_f\right), \tag{7}$$

$f \in \{1, \ldots, F-1\}$, and where $\lambda \ge 0$ is such that

$$\sum_{f=0}^{F-1} P_f \, \Pr[\mathcal{R}_f^{(\mathrm{INR})}] = \overline{P}. \tag{8}$$

The ergodic capacity with partial CSI is:

$$\eta_{M=\infty,F}^{(\mathrm{INR})} = \\ = \max_{\{P_f\}} \sum_{f=0}^{F-1} \mathbb{E}\big[\log(1+\gamma P_f)|\gamma \in \mathcal{R}_f^{(\mathrm{INR})}\big] \Pr[\mathcal{R}_f^{(\mathrm{INR})}], \tag{9}$$

where the maximization is subject to (5) *and* (8)*.*

Remarks:

1) When $F = 1$ (no CSI) the transmitter can not adapt its power across transmissions and thus sends at constant power $g(\gamma) = \overline{P}$. In this case the throughput is:

$$\eta_{M=\infty,F=1}^{(\mathrm{INR})} = \mathbb{E}[\log(1+\gamma\,\overline{P})]. \tag{10}$$

Theorem 4 for $F = 1$ gives the result in (10).

2) With $F = \infty$ (perfect CSI) the optimal power allocation is water-filling [4] given by:

$$g(\gamma) = \left[\frac{1}{\lambda} - \frac{1}{\gamma}\right]^+, \tag{11}$$

and the throughput is:

$$\eta_{M=\infty,F=\infty}^{(\mathrm{INR})} = \mathbb{E}\left[\left[\log\frac{\gamma}{\lambda}\right]^+\right], \tag{12}$$

where the Lagrange multiplier $\lambda \ge 0$ is such that the power constraint is met with equality. The power policy in (7) for $F \to \infty$ reduces to (11) since $\frac{\lambda(P_{f+1}-P_f)}{\mathrm{e}^{\lambda(P_{f+1}-P_f)}-1} \to 1$ when $P_{f+1} - P_f \to 0$ (notice that the region $\mathcal{R}_0^{(\mathrm{INR})}$ always includes the interval $[0, \lambda]$ since $s_1 \ge \lambda$, while the other quantization regions reduce to a single point).

3) The optimal quantization regions are intervals.

4) For the purpose of numerical evaluations, it is convenient to have bounds on the throughput that can be fast evaluated and easily optimized. The throughput in Theorem 4 can be bounded as:

Proposition 5. *For a given set of quantization intervals $\mathcal{R}_f = \{s_f \le \gamma < s_{f+1}\}$, $f \in \{0, \ldots, F-1\}$, the ergodic capacity $\eta_{M=\infty,F}^{(\mathrm{INR})}$ in* (9) *can be bounded as:*

$$\eta_{M=\infty,F}^{(\mathrm{INR})} \le \max_{\{s_f,P_f\}} \sum_{f=0}^{F-1} \log\big(1+\mu_f\,P_f\big) \Pr[\mathcal{R}_f], \tag{13a}$$

$$\eta_{M=\infty,F}^{(\mathrm{INR})} \ge \max_{\{s_f,P_f\}} \sum_{f=0}^{F-1} \log(1+s_f P_f) \Pr[\mathcal{R}_f], \tag{13b}$$

where $\mu_f \triangleq \mathbb{E}[\gamma|\gamma \in \mathcal{R}_f] \in \mathcal{R}_f$ is the centroid of the f-th quantization interval, and the maximization is subject to

$$0 = s_0 \le s_1 \le s_2 \le \cdots \le s_F = +\infty \tag{14}$$

and such that the powers $P_f \ge 0$, $f \in \{0, \ldots, F-1\}$, satisfy $\sum_{f=0}^{F-1} P_f \Pr[\gamma \in \mathcal{R}_f] \le \overline{P}$.

Proof: The bounds in (13) follow immediately from the definition of the quantization intervals and from Jensen's inequality, i.e.,

$$\begin{aligned} &\log(1+\inf\{\gamma \in \mathcal{R}_f\}P_f) \\ &\le \mathbb{E}\big[\log(1+\gamma P_f)|\gamma \in \mathcal{R}_f\big] \\ &\le \log(1+\mathbb{E}[\gamma|\gamma \in \mathcal{R}_f]P_f), \end{aligned}$$

with $\inf\{\gamma \in \mathcal{R}_f\} = s_f$ by definition. For both bounds in (13) the optimal powers are obtained by water-filling [4]. The optimization of the bounds in (13) is thus equivalent to the problem of finding the optimal quantization intervals, which can be done efficiently by using Lloyd's algorithm [17]. ■

As the number of feedback levels F increases, the quantization intervals reduce to a single point since $\mu_f \to s_f$, that is, the bounds in (13) converge to the water-filling ergodic capacity in (12).

V. Throughput Lower Bound $\eta_{M=1,F}^{(\mathrm{ALO})}$

All protocols have the same throughput for $M = 1$ (because retransmissions are not possible), which coincides with the *outage capacity* of the channel [3] given by:

$$\eta_{M=1,F}^{(\mathrm{ALO})} = \max_{R \ge 0} \big\{R(1 - \mathsf{P}_{\mathrm{out}}(R))\big\},$$

where $\mathsf{P}_{\mathrm{out}}(R)$ is the probability of outage given by:

$$\mathsf{P}_{\mathrm{out}}(R) = \Pr\big[\log(1+\gamma\; g(\gamma)) < R\big],$$

and where the non-negative function $g(\cdot)$ can take at most F different values. The optimal power allocation policy $g(\cdot)$ for a finite F is:

Theorem 6. *Define the thresholds*

$$0 \le s_1 \le \cdots \le s_{F-1} \le s_F = s_0 \le s_{F+1} = +\infty, \tag{15}$$

(notice the convention $s_F = s_0 \in (0, \infty)$, rather than $s_0 = 0$ and $s_F = +\infty$ as in (14)*) and the quantization regions:*

$$\mathcal{R}_0^{(\mathrm{ALO})} = \{0 \le \gamma < s_1\} \cup \{\gamma \ge s_0\}, \tag{16a}$$

$$\mathcal{R}_f^{(\mathrm{ALO})} = \{s_f \le \gamma < s_{f+1}\}, \quad f \in \{1, \ldots, F-1\}, \tag{16b}$$

Let the transmit power to be used when $\gamma \in \mathcal{R}_f^{(\mathrm{ALO})}$ be:

$$P_f = \frac{\mathrm{e}^R - 1}{s_f},\ f \in \{0, \ldots, F-1\}, \tag{17}$$

The outage capacity with partial CSI is:

$$\eta^{(\mathrm{ALO})}_{M=1,F} = \max_{\{s_f\}} \Big(1 - \Pr[\gamma < s_1]\Big) \cdot \\ \cdot \log\left(1 + \frac{\overline{P}}{\sum_{f=0}^{F-1} \frac{1}{s_f} \Pr[\gamma \in \mathcal{R}_f^{(\mathrm{ALO})}]}\right), \tag{18}$$

where the maximization is subject to the constraint in (15).

Proof: The proof can be found in Appendix B.

The power P_f in (17) is the minimum power that guarantees no outage for all fading gains in $\mathcal{R}_f^{(\mathrm{ALO})}$, for $f > 0$; outage can only occur when the fading gain belongs to the subset of $\mathcal{R}_0^{(\mathrm{ALO})}$ given by $\{\gamma < s_1\}$. ■

Remarks:

1) When $F = 1$ (no CSI) the transmitter can only send at constant power $g(\gamma) = \overline{P}$ and the throughput is:

$$\eta^{(\mathrm{ALO})}_{M=1,F=1} = \max_{s_1 \geq 0} \log(1 + \overline{P}\, s_1)\ \ \Pr[\gamma \geq s_1]. \tag{19}$$

Theorem 6 for $F = 1$ gives the result in (19).

2) With $F = \infty$ (perfect CSI) the optimal power allocation is truncated channel inversion [12]:

$$g(\gamma) = \frac{\mathrm{e}^R - 1}{\gamma} 1_{\{\gamma \geq s_1\}}. \tag{20}$$

With (20) an outage only happens when $\gamma < s_1$, and the throughput is

$$\eta^{(\mathrm{ALO})}_{M=1,F=\infty} = \max_{s_1 \geq 0} \Pr[\gamma \geq s_1] \log\left(1 + \frac{\overline{P}}{\mathbb{E}\left[\frac{1}{\gamma} 1_{\{\gamma \geq s_1\}}\right]}\right) \tag{21}$$

When $F \to \infty$, the region $\mathcal{R}_0^{(\mathrm{ALO})}$ reduces to $\{0 \leq \gamma < s_1\}$ since $s_\infty = s_0 = \infty$; when the fading belongs to $\mathcal{R}_0^{(\mathrm{ALO})}$ the transmit power is $P_0 = \lim_{s\to\infty} \frac{\mathrm{e}^R - 1}{s} = 0$, thus Theorem 6 reduces to truncated channel inversion.

3) Several power allocation policies have been proposed for outage minimization with partial CSI. However, none of the policies is optimal. For example the solution proposed in [8] corresponds to the suboptimal solution $s_0 = s_F = s_{F+1} = \infty$, that is, setting the quantizations regions to be intervals. From our result in Theorem 6, the optimal quantization regions are in general unions of intervals.

4) For the purpose of simplifying our numerical evaluations we propose to bound the throughput $\eta^{(\mathrm{ALO})}_{M=1,F}$ as:

Proposition 7. *Let*

$$\widehat{\eta}^{(\mathrm{ALO})}_{M=1,F} = \max_{0 \leq s_1 \leq \ldots \leq s_F \leq s_{F+1} = \infty} \Pr[\gamma \geq s_1] \cdot \\ \cdot \log\left(1 + \frac{\overline{P}}{\sum_{f=1}^{F} \frac{1}{s_f} \Pr[\gamma \in [s_f, s_{f+1})]}\right). \tag{22}$$

The throughput in (18) *is bounded by*

$$\widehat{\eta}^{(\mathrm{ALO})}_{M=1,F-1} \leq \eta^{(\mathrm{ALO})}_{M=1,F} \leq \widehat{\eta}^{(\mathrm{ALO})}_{M=1,F}.$$

When $F \gg 1$, the optimal solution of (22) *tends to*

$$s_f = s_1 \xi^{f-1}, \forall f \geq 1, \tag{23}$$

for some $\xi \geq 1$.

Proof: That $\widehat{\eta}^{(\mathrm{ALO})}_{M=1,F}$ in (22) is an upper bound for $\eta^{(\mathrm{ALO})}_{M=1,F}$ in (18) follows by neglecting the term $\frac{\Pr[\gamma < s_1]}{s_F}$ at the denominator in (18). That $\widehat{\eta}^{(\mathrm{ALO})}_{M=1,F-1}$ in (22) is an lower bound for $\eta^{(\mathrm{ALO})}_{M=1,F}$ in (18) follows by setting $s_F = \infty$ in (18). It is interesting to notice that the same function $\widehat{\eta}^{(\mathrm{ALO})}_{M=1,F}$ is a lower bound for $\eta^{(\mathrm{ALO})}_{M=1,F+1}$ (notice the different number of feedback values) and an upper bound for $\eta^{(\mathrm{ALO})}_{M=1,F}$.

The proof of (23) can be found in Appendix C. ■

VI. Main result: achievable throughput for general M and F

Determining $\eta^{(\star)}_{M,F}$, $\star \in \{\mathrm{ALO, RTD, INR}\}$, as in Theorem 1 for general finite values of M and F, is very complex as it involves the solution of a dynamic program (due to the causal nature of the power control [11], [15]). In this section we propose novel protocols that combine repetition and power control for a general M and F. The throughput of our protocols is a lower bound for the optimal $\eta^{(\star)}_{M,F}$.

Theorem 8. *For a protocol $\star \in$ {ALO, RTD, INR} and general finite values of M and F, let $B_m \in \{0, \ldots, F-1\}$ be the feedback sent by the receiver at the beginning of the slot corresponding to the m-th transmission attempt for the current data packet, $m \in \{1, \ldots, M\}$. Consider the following power policy: for $m \in \{1, \ldots, M\}$*

$$P_m = \frac{\mathrm{e}^R - 1}{\tau_m} 1_{\{B_m = 0\}} + \sum_{f=1}^{F-1} \frac{\mathrm{e}^R - 1}{s_{m,f}} 1_{\{B_m = f\}}, \tag{24}$$

with

$$0 \leq \tau_m \tag{25a}$$

$$0 = s_{m,0} \leq s_{m,1} \cdots \leq s_{m,F-1} \leq s_{m,F} = +\infty. \tag{25b}$$

The thresholds $\{s_{m,f}\}_{f=0}^{F}$ in (25) *define a* quantizer for a scaled version of the fading power gain γ_m, where the scaling factor accounts for the information already accumulated at the receiver in the previous $m-1$ transmissions. *In particular, the proposed feedback policy is: for $\star \in$ {ALO, RTD, INR}, $m \in \{1, \ldots, M\}$ and $f \in \{0, \ldots, F-1\}$ let*

$$B_m = f \text{ if } \frac{\gamma_m}{\xi_m^{(\star)}} \in [s_{m,f}, s_{m,f+1}) \text{ and } \xi_m^{(\star)} > 0, \tag{26a}$$

$$B_m = F - 1 \text{ if } \xi_m^{(\star)} \leq 0, \tag{26b}$$

for with $\xi_1^{(\star)} = 1$ *and* $\xi_m^{(\star)}$ *for* $m > 1$ *defined as*

$$\xi_m^{(\mathrm{ALO})} = \prod_{t=1}^{m-1} 1_{\left\{1-\frac{\gamma_t}{\tau_t}>0\right\}}, \tag{27a}$$

$$\xi_m^{(\mathrm{RTD})} = 1 - \sum_{t=1}^{m-1} \frac{\gamma_t}{\tau_t}, \tag{27b}$$

$$\xi_m^{(\mathrm{INR})} = \left(\frac{\mathrm{e}^R}{\prod_{t=1}^{m-1}\left(1 + (\mathrm{e}^R - 1)\frac{\gamma_t}{s_t}\right)} - 1 \right) \frac{1}{\mathrm{e}^R - 1}. \tag{27c}$$

The resulting throughput for protocol $\star \in$ *{ALO, RTD, INR}, as given in Theorem 1, is lower bounded by:*

$$\eta_{M,F,\mathrm{lb}}^{(\star)} \triangleq \max_{\{\tau_m, s_{m,f}\}} \frac{1 - \mathsf{P}_{\mathrm{out}}}{1 + \sum_{m=1}^{M-1} p_{m,0}} \cdot$$
$$\cdot \log\left(1 + \overline{P} \frac{1 + \sum_{m=1}^{M-1} p_{m,0}}{\sum_{m=1}^{M} \sum_{f=1}^{F-1} \frac{p_{m,f} - p_{m,f-1}}{s_{m,f}} + \frac{p_{m,0}}{\tau_m}}\right), \tag{28}$$

where the maximization is subject to the constraints in (25) *and where the probabilities* $\{p_{m,f}\}$ *are defined as:*

$$p_{m,f} = \widetilde{p}_{m,f}^{(\star)} \qquad m = 1, \ldots, M, \quad f = 0, \ldots, F-1, \tag{29a}$$

$$\mathsf{P}_{\mathrm{out}} = \widetilde{p}_{M+1,0}^{(\star)} \quad (\textit{by defining } s_{M+1,1} = +\infty), \tag{29b}$$

where $\{\widetilde{p}_{m,f}^{(\star)}\}$ *are defined in* (36) *for ALO, in* (37) *for RTD, and in* (38) *for INR.*

The rest of the section is devoted to give a rational for the protocols in (26)-(27), to prove the throughput formula in (28) and to define the probabilities $\{\widetilde{p}_{m,f}^{(\star)}\}$ for (29).

A. Protocol description and throughput evaluation

Inspired by the power policy that minimizes the outage capacity in Theorem 6, we propose that the transmitter uses the power policy in (24). Our protocol works as follows:

- The receiver feeds back $B_m = f > 0$, $m \in \{1, \ldots, M\}$, to indicate that the power $(\mathrm{e}^R - 1)/s_{m,f}$ suffices to successfully decode the current data packet when the previous $m-1$ transmissions are combined with the current transmission.
- Upon receiving $B_m = f > 0$, $m \in \{1, \ldots, M\}$, the transmitter is certain that the receiver will decode correctly with the current transmission; hence, after transmission with power $(\mathrm{e}^R - 1)/s_{m,f}$, the transmitter prepares to sent a new data packet.
- The receiver feeds back $B_m = 0$, $m \in \{1, \ldots, M\}$, when none of the powers $(\mathrm{e}^R - 1)/s_{m,f}$, $\forall f > 0$, would guarantee successful decoding.
- In response to $B_m = 0$, $m \in \{1, \ldots, M-1\}$, the transmitter sends with power $(\mathrm{e}^R - 1)/\tau_m$ and prepares to retransmit the same data packet in the next slot.
- In response to $B_M = 0$ (for the last transmission attempt), the transmitter sends with power $(\mathrm{e}^R - 1)/\tau_M$ and prepares to sent a new data packet since no more retransmissions are permitted. In this case an outage can occur.

In order to evaluate the throughput according to Theorem 1 we must determine the average decoded rate (reward) and the average transmit power (cost) when the transmission of the current data packet ends, and the average time needed to transmit a data packet (inter-renewal time). From the description of the protocol given above, it is clear that the transmission of the current data packet does not end after m received slots, $m \in \{1, \ldots, M-1\}$, if all the feedback values received were zero, that is,

$$\Pr[\mathcal{T} \geq m] = \Pr[B_1 = 0, \ldots, B_{m-1} = 0].$$

When the transmission of a data packet ends with less than M transmissions, successful decoding occurs. The transmission of the current data packet ends after the M-th transmission regardless of the status of the decoder (since no more transmission attempts are possible). An outage occurs if decoding is still unsuccessful with M transmissions, i.e.,

$$\mathsf{P}_{\mathrm{out}} = \Pr[B_1 = 0, \ldots, B_M = 0, \text{failure to decode}]. \tag{30}$$

The average number of successfully decoded bits when the transmission of the current data packet ends, i.e., *average reward*, is

$$\mathbb{E}[\mathcal{R}] = R\big(1 - \mathsf{P}_{\mathrm{out}}\big). \tag{31}$$

The average transmit power when transmission of the current data packet ends, i.e., *average cost*, is

$$\mathbb{E}[\mathcal{P}] = \sum_{m=1}^{M} \frac{\mathrm{e}^R - 1}{\tau_m} \Pr[B_1 = 0, \ldots, B_m = 0]$$
$$+ \sum_{m=1}^{M} \sum_{f=1}^{F-1} \frac{\mathrm{e}^R - 1}{s_{m,f}} \Pr[B_1 = 0, \ldots, B_{m-1} = 0, B_m = f]. \tag{32}$$

The average time needed to transmit a data packet, i.e., the *inter-renewal time*, is

$$\mathbb{E}[\mathcal{T}] = \sum_{m=1}^{M} m \ \Pr[\mathcal{T} = m] = \sum_{m=1}^{M} \Pr[\mathcal{T} \geq m]$$
$$= 1 + \sum_{m=1}^{M-1} \Pr[B_1 = 0, \ldots, B_m = 0]. \tag{33}$$

The probabilities in (32) and (33) can be easily expressed as a function of $\widetilde{p}_{m,f}$ defined as:

$$\widetilde{p}_{m,f} \triangleq \Pr[B_1 = 0, \ldots, B_{m-1} = 0, B_m \leq f], \tag{34}$$

for $f \in \{0, \ldots, F-1\}$ since:

$$\Pr[B_1 = \ldots = B_{m-1} = 0, B_m = f]$$
$$= \begin{cases} \widetilde{p}_{m,0} & f = 0, \\ \widetilde{p}_{m,f} - \widetilde{p}_{m,f-1} & f = 1, \ldots, F-2, \\ \widetilde{p}_{m-1,0} & f = F-1, \end{cases} \tag{35}$$

for all $m \in \{1, \ldots, M\}$. The equality for $f = F-1$ follows since $\Pr[B_m \leq F-1] = 1$ for all m. As we shall proof in the next sections, where we give the details of the protocols, the outage probability in (30) needed for (31) ie equivalent to $\mathsf{P}_{\mathrm{out}} = \widetilde{p}_{M+1,0}$ if we assume an hypothetical 1-bit of feedback at the end of the M-th transmission that indicates $B_{M+1} = 1_{\{\text{failure to decode with } M \text{ transmissions}\}}$. (this will correspond to a degenerate quantizer with $0 = s_{M+1,0} < s_{M+1,1} = +\infty$). This discussion justifies the definitions in (29).

Finally, with the definition in (35) and by Theorem 1, the throughput for the proposed protocols is

$$\eta = \frac{\text{eq.(31)}}{\text{eq.(33)}} = \text{eq.(28)},$$

where we expressed the rate R in (31) as a function of $\overline{P}$ from

$$\overline{P} = \frac{\text{eq.(32)}}{\text{eq.(33)}} = (\mathrm{e}^R - 1) \frac{\sum_{m=1}^{M} \left(\frac{\widetilde{p}_{m,0}}{\tau_m} + \sum_{f=1}^{F-1} \frac{\widetilde{p}_{m,f} - \widetilde{p}_{m,f-1}}{s_{m,f}} \right)}{1 + \sum_{m=1}^{M} \widetilde{p}_{m,0}}.$$

The probabilities $\{\widetilde{p}_{m,f}^{(\star)}\}$ in (29), as well as the feedback policy feedback policy $\{B_m\}$ in (26) defined through the quantities $\{\xi_m^{(\star)}\}$ in (27), depend on the protocol $\star \in \{\text{ALO, RTD, INR}\}$ and will be discussed next.

B. First transmission (all protocols)

In order to better understand the way the feedback value is decided, consider the transmission on the first slot, which is the same for all protocols. Let $\theta \triangleq \mathrm{e}^R - 1 \geq 0$.

- Feedback policy:
  - At the beginning of the first slot/transmission, the receiver measures γ_1 and sends

    $$B_1 = F - 1 \text{ if } \log\left(1 + \theta \frac{\gamma_1}{s_{1,F-1}}\right) \geq \log(1 + \theta)$$

    that is, the receiver sends back the highest possible feedback value if the lowest possible power $\theta/s_{1,F-1}$ (from (24) with $m = 1$ and $f = F - 1$) suffices for successful decoding given the actual fading γ_1. In other words, the receiver sends back

    $$B_1 = F - 1 \text{ if } s_{1,F-1} \leq \gamma_1.$$

  - If $s_{1,F-1} > \gamma_1$, i.e., the lowest possible power is not enough to guarantee successful decoding, the receiver checks whether the second lowest available power $\theta/s_{1,F-2}$ suffices for correct decoding. The receiver sends back

    $$B_1 = F - 2 \text{ if } \log\left(1 + \theta \frac{\gamma_1}{s_{1,F-1}}\right) < \log(1 + \theta) \text{ and } \log\left(1 + \theta \frac{\gamma_1}{s_{1,F-2}}\right) \geq \log(1 + \theta),$$

    that is,

    $$B_1 = F - 2 \text{ if } s_{1,F-2} \leq \gamma_1 < s_{1,F-1}.$$

  - By continuing our reasoning in this manner, the receiver sends

    $$B_1 = f \text{ if } \gamma_1 \in [s_{1,f}, s_{1,f+1}), \quad f = 0, \ldots, F - 1,$$

    i.e., the thresholds $\{s_{1,f}\}_{f=0}^{F}$ define a quantizer for γ_1 as in (26a) with $\xi_0 = 1$ (there is no scaling for the fading value on the first transmission because there no accumulate information at the receiver).
- Transmission strategy:
  - In response to $B_1 = f > 0$ the transmitter sends with power $\theta/s_{1,f}$ that guarantees successful decoding for the whole range of fading values in $[s_{1,f}, s_{1,f+1})$; after transmission, the transmitter prepares to send a new data packet.
  - In response to $B_1 = 0$, the transmitter sends with power θ/τ_1 [2] that suffices for successful decoding only if $\gamma_1 \geq \min\{\tau_1, s_{1,1}\}$; however, the transmitter can not know whether the actual γ_1 is above or below $\min\{\tau_1, s_{1,1}\}$, and hence prepares to retransmit the same packet again.

From the second transmission onwards, the mode of operation depends on the protocol used. We will describe the three protocols separately.

C. Retransmissions for ALO

Recall that a second transmission is triggered by $B_1 = 0$, which corresponds to having sent with power θ/τ_1 on the first slot.

- First retransmission:
  - In the ALO protocol only the most recent received slot is used for decoding.
    Assume that in the first transmission the fading satisfied $\gamma_1 \geq \min\{\tau_1, s_{1,1}\}$. The receiver knows that the transmitter will resend the same data packet in the second slot because it received $B_1 = 0$ in the first slot. The receiver can "trick" the transmitter into believing that it will be able to decode in the second slot by sending $B_2 = F - 1$.
    Clearly, this second transmission is a waste of power, but the receiver has no other way to inform the transmitter of its successfully decoding owning to having already exhausted all its feedback bits at the beginning of the current slot. Among all possible powers that the receiver could have requested for the second transmission, $\theta/s_{2,F-1}$ (from (24) with $m = 2$ and $f = F - 1$) is the lowest. This is captured in our protocol definition by the condition in (26b).
  - If $\gamma_1 < \min\{\tau_1, s_{1,1}\}$ (which implies $\gamma_1 < \tau_1$), the receiver uses the same feedback policy it had used as on the first slot, but with possibly different thresholds for the quantization.
    Fig. 2 shows the feedback values for the ALO protocol with $M = 2$ retransmissions and $F = 2$ feedback values; the region with $B_1 = B_2 = 0$ is divided into two parts, the shaded region corresponds to an outage while the white region corresponds to successful decoding.
- Other retransmissions:
  By continuing our reasoning as for the first retransmission, we arrive at the protocol definition in Proposition 8 with the fading scaling defined in (27a).

[2] The power in this case is θ/τ_1; notice the use of τ_1 in place of $s_{1,0}$; this is because $s_{1,0} = 0$ has been already used to indicate the left-most quantization value. The same holds for any $m \geq 1$.

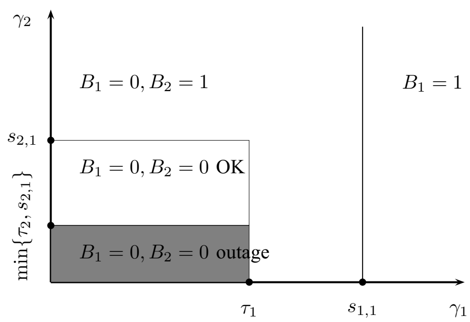

Fig. 2. Feedback values for the ALO protocol with $M = 2$ retransmissions and $F = 2$ feedback values. The shaded region corresponds to an outage.

- Performance:
  For ALO the probabilities $\widetilde{p}_{m,f},\ f \in \{0, \ldots, F-2\}$, as defined in (34) are:

$$\begin{aligned}\widetilde{p}^{(\mathrm{ALO})}_{m,f} &= \Pr\left[\gamma_t < s_{t,1}, \gamma_{t-1} < \tau_{t-1}\, \forall t < m,\right.\\ &\quad \left.\gamma_{m-1} < \tau_{m-1},\ \gamma_m < s_{m,f+1}\right]\\ &= \Pr\left[\frac{\gamma_t}{\min\{\tau_t, s_{t,1}\}} < 1\, \forall t < m,\ \gamma_m < s_{m,f+1}\right]\\ &= \left(\prod_{t=1}^{m-1} F_\gamma(\min\{\tau_t, s_{t,1}\})\right) F_\gamma(s_{m,f+1}), \end{aligned} \tag{36}$$

  with $\prod_{t=1}^{0}(\cdots) = 1$. With the definition of $\widetilde{p}^{(\mathrm{ALO})}_{m,f}$ in (36) we obtain the relationships in (29) for the ALO protocol.
  In general, the probability $\widetilde{p}^{(\mathrm{ALO})}_{m,f}$ is available in closed form if the cumulative density function of the fading power $F_\gamma(x)$, $x \geq 0$, is known in closed form.

  Notice that the probability of outage (failure to decode at the last transmission) would be equivalent to sending $B_{M+1} = 0$, for this reason we have the equality in (29b) by defining $s_{M+1,1} = +\infty$. This observation holds for all protocols.

D. Retransmissions for RTD

Recall that a second transmission is triggered by $B_1 = 0$, which corresponded to having sent with power θ/τ_1 on the first slot.

- First retransmission:
  - In the RTD protocol, the receiver accumulates SNR. On the second transmission, at the beginning of the slot the receiver measures γ_2 and checks whether the lowest available power $\theta/s_{2,F-1}$ suffices for decoding, and sends

$$B_2 = F - 1 \text{ if}$$
$$\log\left(1 + \theta\frac{\gamma_1}{\tau_1} + \theta\frac{\gamma_2}{s_{2,F-1}}\right) \geq \log(1+\theta),$$

    since the resulting SNR after maximal ratio combining of the two received packets is the sum of the SNR on each packet. Hence, the feedback value at the beginning of the second slot is

$$B_2 = F - 1 \text{ if } \frac{\gamma_1}{\tau_1} + \frac{\gamma_2}{s_{2,F-1}} \geq 1.$$

    If $\frac{\gamma_1}{\tau_1} \geq 1$, then $B_2 = F - 1$. However, if $\frac{\gamma_1}{\tau_1} \geq 1$ a retransmission was not necessary in the first place and the power $\theta/s_{2,F-1}$ is wasted, as for the ALO protocol. When $\gamma_1 \geq \tau_1$ the receiver feeds back $B_2 = F - 1$ so that transmission of the current data packet ends with the next slot (with the minimum possible amount of wasted power).

    If $\gamma_1 \geq \tau_1$ and $B_2 = F - 1$ then the condition on the fading value we can rewrite as:

$$B_2 = F - 2 \text{ if } \frac{\gamma_2}{1 - \frac{\gamma_1}{\tau_1}} \geq s_{2,F-1}.$$

  - If instead $\frac{\gamma_1}{\tau_1} + \frac{\gamma_2}{s_{2,F-1}} < 1$ (which implies that $\frac{\gamma_1}{\tau_1} < 1$) the receiver checks whether the second lowest available power $\theta/s_{2,F-2}$ suffices for decoding, and sends

$$\begin{aligned} &B_2 = F - 2\\ &\text{if } \log\left(1 + \theta\frac{\gamma_1}{\tau_1} + \theta\frac{\gamma_2}{s_{2,F-1}}\right) < \log(1+\theta)\\ &\text{and } \log\left(1 + \theta\frac{\gamma_1}{\tau_1} + \theta\frac{\gamma_2}{s_{2,F-2}}\right) \geq \log(1+\theta). \end{aligned}$$

    Hence, the feedback value at the beginning of the second slot is

$$B_2 = F - 2 \text{ if } s_{2,F-2} \leq \frac{\gamma_2}{1 - \frac{\gamma_1}{\tau_1}} < s_{2,F-1}.$$

  - By proceeding with this reasoning, we see that the thresholds $\{s_{2,f}\}_{f=0}^{F}$ define a quantizer for $\frac{\gamma_2}{1-\frac{\gamma_1}{\tau_1}}$ when $\frac{\gamma_1}{\tau_1} < 1$. The value $\frac{\gamma_2}{1-\frac{\gamma_1}{\tau_1}}$ to be quantized is larger than the actual fading γ_2 as it accounts for the SNR already "harvested" at the receiver during the first transmission.
    As an example, Fig. 3 shows the feedback values for the RTD protocol with $M = 2$ retransmissions and $F = 2$ feedback values; the region with $B_1 = B_2 = 0$ is divided into two parts, the shaded region corresponds to an outage while the white region corresponds to successful decoding.
- Other retransmissions:
  In general, if the m-th transmission is required, then the receiver has already accumulated an equivalent SNR of

$$\mathsf{SNR}'_{m-1} = \sum_{t=1}^{m-1} \frac{\gamma_t}{\tau_t},$$

  from the previous $m-1$ transmissions, all in response to a zero-value feedback value. If $\mathsf{SNR}'_{m-1} \geq 1$ decoding was successful, else a retransmission is needed and the receiver uses the thresholds $\{s_{m,f}\}_{f=0}^{F}$ to define a quantizer for $\frac{\gamma_m}{1-\mathsf{SNR}'_{m-1}}$ as proposed in Proposition 8 (fading scaling defined in (27b)).

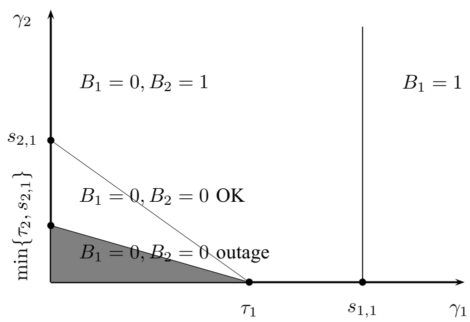

Fig. 3. Feedback values for the RTD protocol with $M = 2$ retransmissions and $F = 2$ feedback values. The shaded region corresponds to an outage.

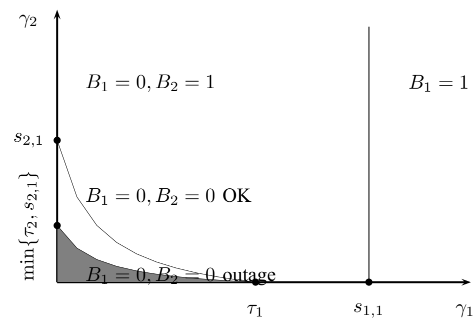

Fig. 4. Feedback values for the INR protocol with $M = 2$ retransmissions and $F = 2$ feedback values. The shaded region corresponds to an outage.

- Performance:
  For RTD the probabilities $\widetilde{p}_{m,f},\ f \in \{0, \ldots, F-2\}$, defined in (34), are:

$$\widetilde{p}^{(\mathrm{RTD})}_{m,f} = \Pr\left[\frac{\gamma_t}{1-\sum_{\ell=1}^{t-1}\frac{\gamma_\ell}{\tau_\ell}} < s_{t,1},\right.$$
$$\left. 1-\sum_{\ell=1}^{t-1}\frac{\gamma_\ell}{\tau_\ell} > 0\ \forall t < m,\ \gamma_m < s_{m,f+1}\right]$$
$$= \Pr\left[\sum_{\ell=0}^{t-1}\frac{\gamma_\ell}{\tau_\ell} + \frac{\gamma_t}{\min\{\tau_t, s_{t,1}\}} < 1\ \forall t < m,\right.$$
$$\left.\frac{\gamma_m}{1-\sum_{\ell=0}^{m-1}\frac{\gamma_\ell}{\tau_\ell}} < s_{m,f+1}\right] \tag{37}$$

  with $\gamma_0/\tau_0 = 0$. With the definition of $\widetilde{p}^{(\mathrm{INR})}_{m,f}$ in (37) we obtain the relationships in (29) for the RTD protocol.
  In general, there is not a closed form expression available for $\widetilde{p}^{(\mathrm{RTD})}_{m,f}$, unless it is possible to evaluate the density of random variables of the type $\sum_{\ell=1}^{t}\frac{\gamma_\ell}{\tau_\ell}$ in closed form.

E. Retransmissions for INR

Recall that a second transmission is triggered by $B_1 = 0$, which corresponds to having sent with power θ/τ_1 on the first slot.

- First retransmission:
  - For the INR protocol, the receiver accumulates mutual information.
    At the beginning of the second slot, the receiver measures γ_2 and sends

$$B_2 = F-1 \quad \text{if}$$
$$\log\left(1+\theta\frac{\gamma_1}{\tau_1}\right) + \log\left(1+\theta\frac{\gamma_2}{s_{2,F-1}}\right) \geq \log(1+\theta),$$

    since the resulting accumulated mutual information at the receiver after optimal combining of the two transmissions is the sum of mutual information of each slot.
    Again, if $\frac{\gamma_1}{\tau_1} \geq 1$ then $B_2 = F-1$ and a retransmission was not necessary in the first place and the power $\theta/s_{2,F-1}$ is wasted, as for the ALO and RTD protocols.
    If $\gamma_1 \geq \tau_1$ and $B_2 = F-1$ then the condition on the fading value we can rewrite as:

$$B_2 = F-1 \text{ if } \gamma_2 \frac{1+\theta\frac{\gamma_1}{\tau_1}}{1-\frac{\gamma_1}{\tau_1}} \geq s_{2,F-1}.$$

  - If $\log\left(1+\theta\frac{\gamma_1}{\tau_1}\right) + \log\left(1+\theta\frac{\gamma_2}{s_{2,F-1}}\right) < \log(1+\theta)$ (which implies that $\frac{\gamma_1}{\tau_1} < 1$) the receiver checks whether the second lowest available power $\theta/s_{2,F-2}$ suffices for decoding, and sends

$$B_2 = F-2$$
$$\text{if } \log\left(1+\theta\frac{\gamma_1}{\tau_1}\right) + \log\left(1+\theta\frac{\gamma_2}{s_{2,F-1}}\right) < \log(1+\theta)$$
$$\text{and } \log\left(1+\theta\frac{\gamma_1}{\tau_1}\right) + \log\left(1+\theta\frac{\gamma_2}{s_{2,F-2}}\right) \geq \log(1+\theta).$$

    Hence, the feedback value at the beginning of the second slot is

$$B_2 = F-2 \text{ if } s_{2,F-2} \leq \gamma_2 \frac{1+\theta\frac{\gamma_1}{\tau_1}}{1-\frac{\gamma_1}{\tau_1}} < s_{2,F-1}$$

  - By proceeding with this reasoning, we see that the thresholds $\{s_{2,f}\}_{f=0}^{F}$ define a quantizer for $\gamma_2\frac{1+\theta\frac{\gamma_1}{\tau_1}}{1-\frac{\gamma_1}{\tau_1}}$ when $\gamma_1 < \tau_1$; this scaled version of γ_2 accounts for the mutual information already accumulated at the receiver in the first transmission.
    For example, Fig. 4 shows the feedback values for the ALO protocol with $M = 2$ retransmissions and $F = 2$ feedback values; the region with $B_1 = B_2 = 0$ is divided into two parts, the shaded region corresponds to an outage while the white region corresponds to successful decoding.
- Other retransmissions:
  For a general m, the mutual information already accumulated from the previous slots is

$$\mathsf{I}_{m-1} = \sum_{t=1}^{m-1}\log\left(1+\theta\frac{\gamma_t}{\tau_t}\right).$$

If $\mathsf{I}_{m-1} \geq \log(1+\theta)$ then decoding was successful and the receiver sends $B_m = F-1$ to end transmission in slot m. If $\mathsf{I}_{m-1} < \log(1+\theta)$, then the receiver looks for the smallest $f > 0$ such that

$$\mathsf{I}_{m-1} + \log\left(1 + \theta\frac{\gamma_m}{s_{m,f}}\right) > \log(1+\theta).$$

If such an $f > 0$ exists, then $B_m = f$ and transmission ends with the current slot; otherwise, $B_m = 0$ and transmission continues. This procedure is equivalent to the protocol in Proposition 8 with the fading scaling defined in (27c).

- Performance:
  For INR the probabilities $\widetilde{p}_{m,f},\ f \in \{0, \ldots, F-2\}$, defined in (34), are:

$$\widetilde{p}^{(\mathrm{INR})}_{m,f} = \Pr\left[\prod_{\ell=0}^{t-1} \frac{\left(1+\theta\frac{\gamma_\ell}{\tau_\ell}\right)\left(1+\theta\frac{\gamma_t}{\min\{\tau_t, s_{t,1}\}}\right)}{1+\theta} < 1, \right.$$
$$\left. \forall t < m, \quad \frac{\gamma_m}{\xi^{(\mathrm{INR})}_m} < s_{m,f+1}\right] \tag{38}$$

  with $\gamma_0/\tau_0 = 0$. With this definition of $\widetilde{p}^{(\mathrm{INR})}_{m,f}$ we obtain the relationships in (29) for the INR protocol.
  In general, there is not a closed form expression available for $\widetilde{p}^{(\mathrm{INR})}_{m,f}$, unless it is possible to evaluate the density of random variables of the type $\prod_{\ell=1}^{t}\left(1+\theta\frac{\gamma_\ell}{\tau_\ell}\right)$ in closed form.

F. Performance with perfect CSI at the transmitter

In order to appreciate the benefits of partial CSI at the transmitter in the proposed repetition protocols, consider the performance with perfect CSI ($F = +\infty$).

Since the power control is causal, the throughput is found as a solution of a dynamic program [13]:

Proposition 9. *The throughput $\eta^{(\star)}_{M,F=\infty,\mathrm{lb}}$ is the solution of:*

$$\max_{\{P_m \geq 0\}} R\frac{1-\Pr[\sum_{t=1}^{M} U_t < g(R)]}{1+\sum_{m=1}^{M-1}\Pr[\sum_{t=1}^{m} U_t < g(R)]} \tag{39a}$$

$$\text{s.t.} \quad \frac{\sum_{m=1}^{M}\mathbb{E}\left[P_m\, 1_{\{\sum_{t=1}^{m-1} U_t < g(R)\}}\right]}{1+\sum_{m=1}^{M-1}\Pr[\sum_{t=1}^{m} U_t < g(R)]} \leq \overline{P}, \tag{39b}$$

where: for ALO, $U_m = 1_{\{\gamma_m P_m > \mathrm{e}^R - 1\}}$ indicates successful decoding on the m-th transmission and $g(R) = 0$; for RTD, $U_m = \gamma_m P_m$ represents the received SNR on the m-th transmission and $g(R) = \mathrm{e}^R - 1 \geq 0$; for INR, $U_m = \log(1+\gamma_m P_m)$ represents the mutual information at the receiver on the m-th transmission and $g(R) = R$; the optimization is with respect to causal power policies $P_m = P_m(\gamma_1, \ldots, \gamma_m)$, $m = 1, \ldots, M$.

Proof: The optimization of a causal power control system can be cast as a *dynamic program over finite horizon with complete observations* [11], [13], [15], where the system state is $X_t = \sum_{m=0}^{t-1} U_m$, the control is U_t, and the state evolves as $X_{t+1} = X_t + U_t$, with $U_0 = 0$, from which (39) follows. ■

The problem in (39a) is similar to the outage minimization problem in [15]. As in [15], we can write the iterative algorithm that defines the optimal dynamic programming solution, however an explicit closed form solution is not available in general. Numerical techniques, as those proposed in [15], must be used for numerical evaluations of (39).

VII. Novel bounding technique

In the previous section we proposed novel protocols that combine power control with partial CSI and retransmissions. In all cases, a closed form expression of the throughput requires a closed form expression for the cumulative density function of: (a) the fading γ_ℓ for $\widetilde{p}^{(\mathrm{ALO})}_{m,f}$ in (36), (b) random variables of the type $S_m = \sum_{\ell=1}^{m}\frac{\gamma_\ell}{\tau_\ell}$, for $\widetilde{p}^{(\mathrm{RDT})}_{m,f}$ in (37), and (c) random variables of the type $I_m = \prod_{\ell=1}^{m}\left(1+\theta\frac{\gamma_\ell}{\tau_\ell}\right)$ for some $\theta \geq 0$ for $\widetilde{p}^{(\mathrm{INR})}_{m,f}$ in (38). Since the distribution of I_m is rarely known in closed form, in the following we propose a novel bounding technique for the cumulative density function of I_m in terms of the the cumulative density function of S_m. For the case of iid Rayleigh fading, the density of S_m is known in closed form; hence, our technique allows to determine closed-form upper and lower bounds for probabilities involving I_m.

Consider generic non-negative constants $\{\tau_m\}$, $m \in \mathbb{N}$, a constant $\theta \geq 0$, and define

$$p^{(\mathrm{ALO})}_m = \Pr\left[\frac{\gamma_s}{\tau_s} < 1,\ s = 1, \ldots, m\right], \tag{40a}$$

$$p^{(\mathrm{RTD})}_m = \Pr\left[\sum_{s=1}^{m}\frac{\gamma_s}{\tau_s} < 1\right], \tag{40b}$$

$$p^{(\mathrm{INR})}_m = \Pr\left[\frac{1}{\log(1+\theta)}\sum_{s=1}^{m}\log\left(1+\theta\frac{\gamma_s}{\tau_s}\right) < 1\right], \tag{40c}$$

for some sequence $\{\gamma_\ell\}$ of iid random variables.

Example: As an example, consider Fig. 5, which shows in the plane $x_1 \triangleq \frac{\gamma_1}{\tau_1}, x_2 \triangleq \frac{\gamma_2}{\tau_2}$ the regions that defines $p^{(\star)}_2$, $\star \in \{\text{ALO, RTD, INR}\}$. The probability $p^{(\mathrm{ALO})}_2$ is the integral of the joint density of (x_1, x_2) over the square $(x_1, x_2) \in [0,1]^2$. The probability $p^{(\mathrm{RTD})}_2$ is the integral over the triangle $x_1 + x_2 \leq 1$, $x_1 \geq 0$ and $x_2 \geq 0$, that is, over the region in the positive quadrant below the dotted-line curve labeled "RTD" in Fig. 5. And finally, the probability $p^{(\mathrm{INR})}_2$ is the integral over the region in the positive quadrant below the solid-line curve labeled "INR" in Fig. 5. The curve labeled "INR" in Fig. 5 is a convex function that can be bounded from above and from below by piece-wise linear functions. We chose piece-wise linear functions because the region they define is the union of triangular regions. In particular, for the inner bound, we take the union of the two regions below the tangent lines at $(x_1, x_2) = (0, 1)$ and at $(x_1, x_2) = (1, 0)$ (the region in the positive quadrant below the dash-dotted-line curve labeled "INR inner region" in Fig. 5), while for the outer bound, we take the union of the two regions below the lines passing through $(x_1, x_2) = (0, 1)$ and $(x_1, x_2) = (1/2, 1/2)$, and through $(x_1, x_2) = (1, 0)$ and $(x_1, x_2) = (1/2, 1/2)$ (the region in the positive quadrant below the dashed-line curve labeled "INR outer region" in Fig. 5).

By extending the idea presented in the above example to the case of a general $m \in \mathbb{N}$ we can show:

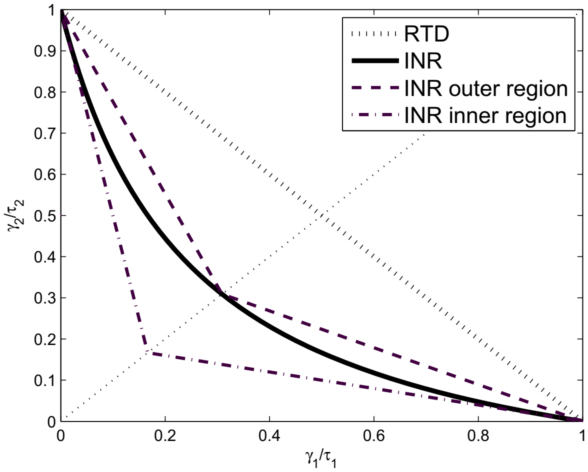

Fig. 5. The region that defines the probability of outage for INR with $M = 2$ and its approximations.

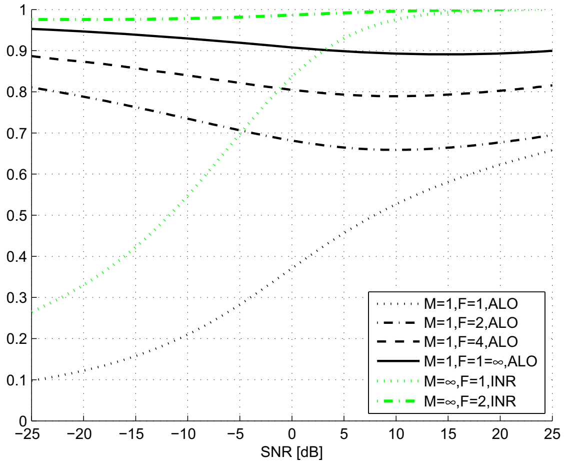

Fig. 6. Ratio between the throughput and $\eta^{(\mathrm{INR})}_{M=\infty,F=\infty}$ (the ergodic capacity with full CSI), for the Rayleigh fading channel.

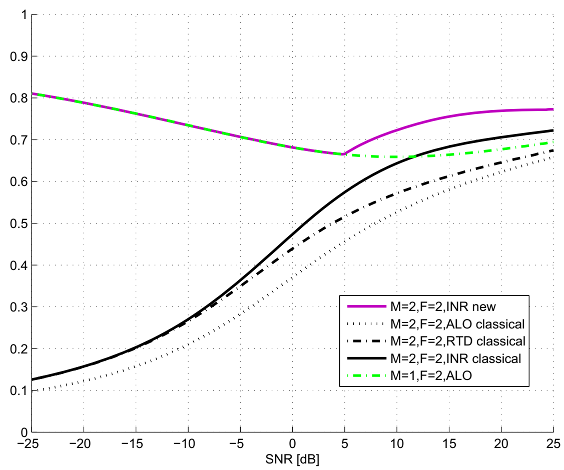

Fig. 7. Ratio between the throughput and $\eta^{(\mathrm{INR})}_{M=\infty,F=\infty}$ (the ergodic capacity with full CSI), for the Rayleigh fading channel.

Proposition 10. *The probability in* (40c) *can be bounded as:*

$$p_m^{(\mathrm{INR})} \geq \Pr\left[(1+\theta)\sum_{s=1}^{m}\frac{\gamma_s}{\tau_s} - \theta \max_{t=1,\dots,m}\frac{\gamma_t}{\tau_t} < 1\right] \tag{41a}$$

$$p_m^{(\mathrm{INR})} \leq \Pr\left[\sum_{s=1}^{m}\frac{\gamma_s}{\tau_s} + \left(\frac{\theta}{(1+\theta)^{1/m}-1} - m\right)\max_{t=1,\dots,m}\frac{\gamma_t}{\tau_t} < 1\right] \tag{41b}$$

Proof: The proof can be found in Appendix D. ∎

The interesting fact about the two bounds in (41) is that they are computable from the knowledge of the density of the random variable

$$X_{a,b} = a \max_{t=1,\dots,m}\frac{\gamma_t}{\tau_t} + b\sum_{s=1}^{m}\frac{\gamma_s}{\tau_s} \tag{42}$$

for some fixed $(a, b) \in \mathbb{R}^2$.

Proposition 11. *For the case of iid negative exponential random variables $\{\gamma_t\}$ (i.e., iid Rayleigh fading), the density of $X_{a,b}$ in* (42)*, for any $(a, b) \in \mathbb{R}^2$, is given in* (44) *in Appendix E.*

Proof: The proof can be found in Appendix E. ∎

VIII. The iid Rayleigh fading channel

To illustrate the gain achievable with the protocols proposed in Section VI, we evaluate the performance of the different protocols for the Gaussian iid Rayleigh fading channel, for which the fading cumulative distribution function is $F_\gamma(x) = 1 - \mathrm{e}^{-x}$ for $x \geq 0$. We define the exponential integral function as:

$$\mathbb{E}\left[\frac{1}{\gamma}1_{\{\gamma \geq x\}}\right] \overset{\mathrm{RF}}{=} \int_x^\infty \mathrm{e}^{-t}/t\,\mathrm{d}t \triangleq \mathrm{Ei}(x)$$

for $x \geq 0$. We use the symbol "$\overset{\mathrm{RF}}{=}$" to indicate that the equality holds for Rayleigh fading channels.

For the plots, communication rates are measured in bits/sec/Hz and the figures show the relative throughput performance with respect to the ergodic water-filling capacity (i.e., $M = F = \infty$ and INR), which is the ultimate performance limit for a fading channels with full CSI. Table I summarizes the cases considered in the following.

TABLE I

Amount of CSI	$M = 1$ outage cap.	$M = 2$ HARQ protocols	$M = \infty$ ergodic cap.
Absent ($F = 1$)	in VIII-B1	(impossible, need at least 1-bit for ack/nack)	in VIII-A1
1 bit ($F = 2$)	in VIII-B2	in VIII-C (classical HARQ) in VIII-D (proposed HARQ)	in VIII-A2
Full ($F = \infty$)	in VIII-B3	(not evaluated)	in VIII-A3

A. Throughput upper bound (ergodic capacity)

1) $M=\infty$, $F=1$, INR: With constant power allocation the ergodic capacity is:

$$\eta^{(\text{INR})}_{M=\infty,F=1} \stackrel{\text{RF}}{=} \mathrm{e}^{\frac{1}{\overline{P}}}\,\mathrm{Ei}\left(\frac{1}{\overline{P}}\right).$$

2) $M=\infty$, finite F, INR: The ergodic capacity with partial CSI is given in Theorem 4:

$$\eta^{(\text{INR})}_{M=\infty,F} \stackrel{\text{RF}}{=} \max\Big\{\sum_{f=0}^{F-1} -\log(1+P_f s_f)\mathrm{e}^{-s_f} - \mathrm{e}^{1/P_f}\mathrm{Ei}(1/P_f+s_f) + \log(1+P_f s_{f-1})\mathrm{e}^{-s_{f-1}} + \mathrm{e}^{1/P_f}\mathrm{Ei}(1/P_f+s_{f-1})\Big\}$$

where the thresholds $\{s_f\}$ are defined in (7), the maximization over $\{P_f\}$ is subject to the constraint in (5) and

$$\overline{P} \stackrel{\text{RF}}{=} \sum_{f=0}^{F-1} P_f[\mathrm{e}^{-s_{f-1}} - \mathrm{e}^{-s_f}]$$

The ergodic capacity with partial CSI can be bounded by using Proposition 5; in particular, μ_f in (13a) is given by

$$\mu_f \stackrel{\text{RF}}{=} \frac{\mathrm{e}^{-s_{f-1}}(1+s_{f-1}) - \mathrm{e}^{-s_f}(1+s_f)}{\mathrm{e}^{-s_{f-1}} - \mathrm{e}^{-s_f}}.$$

3) $M=\infty$, $F=\infty$, INR: With water-filling power allocation the ergodic capacity is:

$$\eta^{(\text{INR})}_{M=\infty,F=\infty} \stackrel{\text{RF}}{=} \mathrm{Ei}(\lambda),$$
$$\overline{P} \stackrel{\text{RF}}{=} \frac{\mathrm{e}^{-\lambda}}{\lambda} - \mathrm{Ei}(\lambda).$$

4) Discussion: From Fig. 6 we see that $\eta^{(\text{INR})}_{M=\infty,F=2}$ (i.e., 1 bit of feedback) is already at 97% of the water-filling capacity $\eta^{(\text{INR})}_{M=\infty,F=\infty}$ at an SNR as low as -25dB. At high SNR, $\eta^{(\text{INR})}_{M=\infty,F=1}$ (no CSI) behaves like $\eta^{(\text{INR})}_{M=\infty,F=\infty}$ (full CSI) but at low SNR their ratio tends to zero. This restates a well know fact that power allocation in single user channels offers benefits at low SNR only and that a single bit of feedback (i.e., $F=2$) gives almost all the gain achievable by full CSI (i.e., $F=\infty$).

B. Throughput lower bound (outage capacity)

1) $M=1$, $F=1$, ALO: With constant power allocation the outage capacity is:

$$\eta^{(\text{ALO})}_{M=1,F=1} \stackrel{\text{RF}}{=} \max_{s_1\geq 0} \log(1+\overline{P}s_1)\mathrm{e}^{-s_1}.$$

2) $M=1$, $F=2$, ALO: With 1-bit of feedback, the throughput is the solution of:

$$\eta^{(\text{ALO})}_{M=1,F=2} = \max_{0\leq s_1\leq s_2\leq\infty} \mathrm{e}^{-s_1}\cdot \log\left(1+\frac{\overline{P}}{\frac{1-\mathrm{e}^{-s_1}}{s_2}+\frac{\mathrm{e}^{-s_1}-\mathrm{e}^{-s_2}}{s_1}+\frac{\mathrm{e}^{-s_2}}{s_2}}\right) = \max_{0\leq s_1\leq\infty} \mathrm{e}^{-s_1}\log(1+\overline{P}s_1\mathrm{e}^{s_1}).$$

3) $M=1$, $F=\infty$, ALO: With truncated channel inversion power allocation the outage capacity is:

$$\eta^{(\text{ALO})}_{M=1,F=\infty} \stackrel{\text{RF}}{=} \max_{\tau\geq 0}\log\left(1+\frac{\overline{P}}{\mathrm{Ei}(\tau)}\right)\mathrm{e}^{-\tau}.$$

4) Discussion: When $M=1$, there is no need to send a ACK/NACK because the transmitter cannot retransmit. In this case, the one bit of feedback should indeed be used at the beginning of the slot to inform the transmitter about the state of the channel. From Fig. 6 we observe that 1-bit of feedback at -25dB results in $\eta^{(\text{ALO})}_{M=1,F=2}/\eta^{(\text{INR})}_{M=\infty,F=\infty} = 81\%$ while (with constant power) $\eta^{(\text{ALO})}_{M=1,F=1}/\eta^{(\text{INR})}_{M=\infty,F=\infty} = 10\%$ only. At +5dB, $\eta^{(\text{ALO})}_{M=1,F=2}/\eta^{(\text{INR})}_{M=\infty,F=\infty} = 73\%$ while (with constant power) $\eta^{(\text{ALO})}_{M=1,F=1}/\eta^{(\text{INR})}_{M=\infty,F=\infty} = 45\%$. In fact, at high SNR, power allocation is less critical and the gain due to CSI vs. no CSI diminishes. We also reported for comparison the achievable throughput for ALO with $F=4$ (2 bits of feedback) by using the approximation $s_\ell = s_1\,\xi^{\ell-1}$ as in Proposition 7, and for ALO with $F=\infty$ (full CSI). We see that the gains attainable at low SNR due to only a few bits of feedback are dramatic and that 2 bits of feedback attain a throughput remarkably close to the case with full CSI.

C. Classical HARQ protocols with $F=2$ and $M=2$

Classical HARQ protocols use the 1 bit of feedback (i.e., $F=2$) to signal ACK/NACK. Classical HARQ protocols are a special case of the protocols proposed in Section VI obtained by setting $s_{m,f}=\infty$ for all $m>0$ and $f>0$ in (25). In classical HARQ protocols the power can vary across repetitions (i.e., different values of τ_m, $m=1,\ldots,M$) but it cannot depend on the CSI [7]. In this case the throughput in (28) is a function only of $\{\widetilde{p}_{m,0}\}$, $m\in\{1,\ldots,M+1\}$ given by:

$$\widetilde{p}^{(\text{ALO})}_{m,0} \stackrel{\text{RF}}{=} \prod_{s=1}^{m}(1-\mathrm{e}^{-\tau_s})$$
$$\widetilde{p}^{(\text{RTD})}_{m,0} \stackrel{\text{RF}}{=} \sum_{s=1}^{m}\frac{1-\mathrm{e}^{-\tau_s}}{\prod_{s\neq j}(1-\tau_s/\tau_j)}$$
$$\widetilde{p}^{(\text{INR})}_{m,0} = \Pr\left[\sum_{s=1}^{m}\log\left(1+\theta\frac{\gamma_s}{\tau_s}\right) < \log(1+\theta)\right],$$

for $\theta \triangleq \mathrm{e}^R-1\geq 0$ and $m\geq 1$. Consider the case $M=2$: the only probability not known in closed form is $\widetilde{p}^{(\text{INR})}_{m,0}$, which we bound by using the technique developed in Proposition 11 in Section VII as follows:

$$\widetilde{p}^{(\text{in})}_{2,0} = 1-q\left(\frac{1}{2+\theta}\right)$$
$$\leq \widetilde{p}^{(\text{INR})}_{2,0} = \mathbb{E}\left[F_\gamma\left(\tau_2\frac{1-\gamma_1/\tau_1}{1+\gamma_1/\tau_1\,\theta}\right)\right]$$
$$\leq \widetilde{p}^{(\text{out})}_{2,0} = 1-q\left(\frac{\sqrt{1+\theta}-1}{\theta}\right)$$
$$\leq \widetilde{p}^{(\text{RTD})}_{2,0} = 1-q(1/2) = 1-\frac{\mathrm{e}^{-\tau_1}}{1-\tau_1/\tau_2}-\frac{\mathrm{e}^{-\tau_2}}{1-\tau_2/\tau_1}$$
$$\leq \widetilde{p}^{(\text{ALO})}_{2,0} = (1-\mathrm{e}^{-\tau_1})(1-\mathrm{e}^{-\tau_2})$$

where the function $q(x)$ is given by (44) in Appendix E.

1) Discussion: In Fig. 7, classical HARQ protocols are labeled as "classical" in order to distinguish them from the novel protocols proposed in this work, which are labeled as "new". The case of $M = 2$ transmission is considered (which implies that the throughput needs to be optimized with respect to the two parameters (τ_2, τ_1)). From the numerical results for SNR $= \overline{P} \in [-25, +25]$dB, we saw that equal power allocation across transmissions (i.e., $\tau_2 = \tau_1$) is optimal at all SNR's for ALO; for RTD and INR, the use of different power (i.e., $\tau_2 \neq \tau_1$) offers benefits; however the improvement is negligible (for example less than 0.3% across the entire range of simulated powers for RTD) for this reason in Fig. 7 we only show the throughput with the optimized (τ_2, τ_1) only for INR. We see that the throughput trend for the classical HARQ protocols with $F = 2$ and $M = 2$ is the same as the one for $F = 1$ and $M = 1$ reported in Fig. 6. This shows that the 1 bit of feedback used for ACK/NACK only does not offer substantial throughput improvement compared to the outage capacity with constant power allocation ($F = 1$ and $M = 1$). Only for SNR's larger than 10dB, using classical INR with $F = 2$ and $M = 2$ gives a larger throughput that ALO with $F = 1$ and $M = 1$ with power control. From these observations we conclude that classical HARQ make an inefficient use of the feedback resources.

D. New protocols for $M = 2$ and $F = 2$

Figs. 2, 3, and 4 show the bits fed back to the transmitter as a function of $(\gamma_1/\tau_1, \gamma_2/\tau_2)$. Since we consider the case $F = 2$, we only need to characterize $\widetilde{p}_{m,f}$, for $m \in \{1, 2, 3\}$ and $f = 0$, as per Proposition 8.

The probability of requesting a retransmission is:

$$\widetilde{p}_{1,0} = \Pr[B_1 = 0] = \Pr\left[\frac{\gamma_1}{s_{1,1}} < 1\right] = \Pr[\mathcal{T} = 2] = \widetilde{p}_{2,1};$$

the probability of decoding failure after the first transmission is:

$$\widetilde{p}_{2,0} = \Pr[B_1 = 0, B_2 = 0] =$$
$$\Pr\left[\frac{\gamma_1}{\min\{\tau_1, s_{1,1}\}} < 1, \frac{\gamma_2}{s_{2,1}} < 1\right] \quad \text{ALO}$$
$$\Pr\left[\frac{\gamma_1}{\min\{\tau_1, \tau_{1,1}\}} < 1, \frac{\gamma_1}{\tau_1} + \frac{\gamma_2}{s_{2,1}} < 1\right] \quad \text{RTD}$$
$$\Pr\left[\frac{\gamma_1}{\min\{\tau_1, \tau_{1,1}\}} < 1, \frac{\gamma_1}{\tau_1} + \frac{\gamma_2}{s_{2,1}}\left(1 + \theta\frac{\gamma_1}{\tau_1}\right) < 1\right] \quad \text{INR},$$

and the probability of decoding failure after the second transmission, which coincides with the probability of outage, is:

$$\widetilde{p}_{3,0} = \Pr[B_1 = 0, B_2 = 0, B_3 = 0] =$$
$$\Pr\left[\frac{\gamma_1}{\min\{\tau_1, s_{1,1}\}} < 1, \frac{\gamma_2}{\min\{\tau_2, s_{2,1}\}} < 1\right] \quad \text{ALO}$$
$$\Pr\left[\frac{\gamma_1}{\min\{\tau_1, s_{1,1}\}} < 1, \frac{\gamma_1}{\tau_1} + \frac{\gamma_2}{\min\{\tau_2, s_{2,1}\}} < 1\right] \quad \text{RTD}$$
$$\Pr\left[\frac{\gamma_1}{\min\{\tau_1, s_{1,1}\}} < 1, \frac{\gamma_1}{s_1} + \frac{\gamma_2}{\min\{\tau_2, s_{2,1}\}}\left(1 + \theta\frac{\gamma_1}{\tau_1}\right) < 1\right] \text{INR}.$$

The probabilities for ALO and RTD can be evaluated in closed form, while the probabilities for INR can be bounded as in Subsection VIII-C (from Proposition 11 in Section VII) for the classical HARQ protocols by using the function $q(x)$ in (44).

1) Discussion: Fig. 7 shows the ratio among the throughput of new protocols and the ergodic water-filling capacity. We see that the new protocols dramatically outperform the classical repetition protocols, especially at low SNR (compare with Fig. 6). Indeed, at low SNR it is critical to be able to save power when the channel is in deep fade. By providing the transmitter with a 1-bit quantization of the current channel gain (rather than ACK/NACK), we enable the transmitter to do so. At low SNR, the repetition is not needed (ALO with $M = 1$ and $F = 2$ has the same throughput of INR with $M = 2$ and $F = 2$), while at high SNR, the repetition helps. Notice that here high SNR means is SNR > 5dB, at which $\eta^{(\mathrm{INR})}_{M=2,F=2}$ is about 67% of $\eta^{(\mathrm{INR})}_{M=\infty,F=\infty}$. We did not report the RTD and ALO curve for $M = 2$ and $F = 2$ with CSI as they do not differ much from ALO with $M = 1$ and $F = 2$.

IX. Conclusions and Future Work

In this work we considered HARQ protocols where the feedback bits not only convey a retransmission request to the transmitter but also inform the transmitter coarsely about the channel state. We developed a new class of protocols that feedback the quantized index of a suitably scaled version of the current fading value; the scaling factor is such that the mutual information already accumulated at the transmitter from the previous transmissions is taken into account. We showed that our proposed protocols significantly outperform classical HARQ protocols for the same amount of feedback resources, especially at low SNR; this shows that ACK/NACK feedback is suboptimal in time-varying channels.

As future work, it would be interesting to evaluate the throughput performance when the cost of acquiring the CSI and the error in the estimated CSI are taken into account. Also, it is important to test the proposed protocols with practical codes, instead of with ideal Gaussian codes.

Extensions to multiple access channels are presented in [18].

Acknowledgment

The author would like to thank the Associate Editor and the anonymous reviewers for their comments that helped to improve the quality of the paper.

Appendix A
Proof of Theorem 1

As a direct application of the renewal-reward theorem [2], we have that the limit in (3) converges almost surely to $\mathbb{E}[\mathcal{P}]/\mathbb{E}[\mathcal{T}]$ and the limit in (1) converges almost surely to $\mathbb{E}[\mathcal{R}]/\mathbb{E}[\mathcal{T}]$. The average inter-renewal time $\mathbb{E}[\mathcal{T}]$ is given by:

$$\mathbb{E}[\mathcal{T}] = \sum_{m=1}^{M} m \Pr[\mathcal{T} = m] = \sum_{m=1}^{M} \Pr[\mathcal{T} \geq m].$$

The average reward $\mathbb{E}[\mathcal{R}]$ is given by:

$$\mathbb{E}[\mathcal{R}] = R(1 - P_{\mathrm{out}}) = R(1 - \Pr[\mathcal{T} = M, \text{fail to decode}]).$$

since in our framework, a packet is lost ($\mathcal{R}=0$) only if on the last transmission decoding was not successful. The average transmit power $\mathbb{E}[\mathcal{P}]$ is given by:

$$\begin{aligned}\mathbb{E}[\mathcal{P}] &= \sum_{m=1}^{M} \Pr[\mathcal{T}=m]\mathbb{E}[\mathcal{P}|\,\mathcal{T}=m] \\ &= \sum_{m=1}^{M} \Pr[\mathcal{T}=m] \sum_{t=1}^{m} \mathbb{E}[P_t|\,\mathcal{T}=m] \\ &= \sum_{t=1}^{M}\sum_{m=t}^{M} \Pr[\mathcal{T}=m]\,\mathbb{E}[P_t|\,\mathcal{T}=m] \\ &= \sum_{t=1}^{M} \Pr[\mathcal{T}\geq t] \frac{\sum_{m=t}^{M}\Pr[\mathcal{T}=m]\,\mathbb{E}[P_t|\,\mathcal{T}=m]}{\sum_{m=t}^{M}\Pr[\mathcal{T}=m]} \\ &= \sum_{t=1}^{M} \Pr[\mathcal{T}\geq t]\,\mathbb{E}[P_t|\,\mathcal{T}\geq t],\end{aligned}$$

QED.

Appendix B
Proof of Theorem 6

Given a partition $\{\mathcal{R}_f\}$ of $\mathbb{R}^+$, the transmit power is defined through the thresholds in (7) as

$$P = \sum_{f=1}^{F} \frac{\mathrm{e}^R-1}{s_f} 1_{\{\gamma\in\mathcal{R}_f\}},$$

and must satisfy the average power constraint

$$\sum_{f=1}^{F} \frac{\mathrm{e}^R-1}{s_f} \Pr[\gamma\in\mathcal{R}_f] \leq \overline{P}.$$

For a fixed rate R, the throughput is maximized if the outage probability is minimized. The outage probability satisfies

$$\begin{aligned}&\Pr\left[\log\left(1+\gamma\sum_{f=1}^{F}\frac{\mathrm{e}^R-1}{s_f}1_{\{\gamma\in\mathcal{R}_f\}}\right)<R\right] \\ &= \mathbb{E}\left[1_{\{\log\left(1+\gamma\sum_{f=1}^{F}\frac{\mathrm{e}^R-1}{s_f}1_{\{\gamma\in\mathcal{R}_f\}}\right)<R\}}\right] \\ &= \mathbb{E}\left[\sum_{f=1}^{F}1_{\{\gamma\in\mathcal{R}_f\}}1_{\{\gamma<s_f\}}\right] \\ &\geq \mathbb{E}\left[\min_{f=1,\ldots,F}1_{\{\gamma<s_f\}}\right],\end{aligned}$$

where the last inequality holds with equality for

$$\mathcal{R}_f^{(\mathrm{ALO})} = \left\{\gamma\in\mathbb{R}^+ : f = \arg\min_{\ell=1,\ldots,F}\left\{1_{\{\gamma<s_\ell\}}\right\}\right\}.$$

By assuming without loss of generality that the thresholds $\{s_f\}$ are ordered in increasing order, we have:

$$\begin{aligned}&\arg\min_{f=1,\ldots,F} 1_{\{\gamma<s_f\}} \\ &\in \{1,\ldots,f\} \text{ if } \gamma\in[s_f,s_{f+1}),\ f=1,\ldots,F-1, \\ &\in\{1,\ldots,F\} \text{ if } \gamma\in[s_F,+\infty)\cup[0,s_1).\end{aligned}$$

In order to use the least power we choose

$$\begin{aligned}&\arg\min_{f=1,\ldots,F} 1_{\{\gamma<s_f\}} \\ &= f \text{ if } \gamma\in R_f^{(\mathrm{ALO})}=[s_f,s_{f+1}), f=1,\ldots,F-1, \\ &= F \text{ if } \gamma\in R_F^{(\mathrm{ALO})}=[s_F,+\infty)\cup[0,s_1).\end{aligned}$$

With these quantization regions, an outage only occurs when the feedback value is F and the fading is in $[0,s_1)$, hence the throughput is

$$\begin{aligned}&\eta_{M=1,F}^{(\mathrm{ALO})} \\ &= \max_{R,\{s_f,\mathcal{R}_f\}} R\Pr\left[R\leq\log\left(1+\gamma\sum_{f=1}^{F}\frac{\mathrm{e}^R-1}{s_f}1_{\{\gamma\in\mathcal{R}_f\}}\right)\right] \\ &= \max_{\{s_f\}}\log\left(1+\frac{\overline{P}}{\frac{\Pr[\gamma\in[0,s_1)]}{s_F}+\sum_{f=1}^{F}\frac{1}{s_f}\Pr[\gamma\in[s_f,s_{f+1})]}\right)\cdot \\ &\quad\cdot\Pr[\gamma\geq s_1].\end{aligned}$$

Appendix C
Proof of Proposition 7

The optimal values of $0\leq s_2\leq\ldots\leq s_F\leq s_{F+1}=\infty$ for $\widehat{\eta}_{M=1,F}^{(\mathrm{ALO})}$ in (22) minimize

$$\sum_{f=1}^{F}\frac{1}{s_f}[F_\gamma(s_{f+1})-F_\gamma(s_f)]. \tag{43}$$

By assuming that the fading has a density $f_\gamma(x) = \mathrm{d}F_\gamma(x)/\mathrm{d}x$, by taking the partial derivatives of (43) with respect to s_ℓ for $\ell\geq 2$, and solving them equal to zero, we get that the thresholds satisfy

$$\frac{s_\ell-s_{\ell-1}}{s_{\ell-1}} = \frac{F_\gamma(s_{\ell+1})-F_\gamma(s_\ell)}{s_\ell\,f_\gamma(s_\ell)},\quad \ell\geq 2.$$

For sufficiently large F, since the thresholds are going to be close to each other, we can approximate

$$\frac{F_\gamma(s_{\ell+1})-F_\gamma(s_\ell)}{(s_{\ell+1}-s_\ell)} \approx \left.\frac{\mathrm{d}F_\gamma(x)}{\mathrm{d}x}\right|_{x=s_\ell} = f_\gamma(s_\ell),$$

and hence we conclude that the optimal thresholds satisfy

$$s_\ell^2 \ \approx\ s_{\ell-1}s_{\ell+1} \Longleftrightarrow s_\ell\approx s_1\,\xi^{\ell-1},$$

for some $\xi\geq 1$.

Appendix D
Proof of Proposition 10

For a general $m\in\mathbb{N}$, the probability in (40c) (with $X_\ell = \gamma_\ell/\tau_\ell$) is equivalent to

$$p_m^{(\mathrm{INR})} = \Pr[X_m - f(X_1,\ldots,X_{m-1})<0],$$

for

$$f(x_1,\ldots,x_{m-1}) = \frac{1}{\theta}\left(\frac{1+\theta}{\prod_{s=1}^{m-1}(1+\theta x_s)}-1\right).$$

Our goal is to bound the region below $f(x_1,\ldots,x_{m-1})$ in the positive quadrant by regions defined as union of hyperplanes.

Toward this goal, it is useful to keep in mind that the partial derivatives of $f(x_1,\ldots,x_{m-1})$ are given by

$$\frac{\partial f(x_1,\ldots,x_{m-1})}{\partial x_j} = -\frac{1+\theta\, f(x_1,\ldots,x_{m-1})}{1+\theta x_j}.$$

Being $f(x_1,\ldots,x_{m-1})$ a concave function in all its arguments, it can be lower-bounded by any hyperplane tangent to it. In particular, if we consider the hyperplanes tangent to $f(x_1,\ldots,x_{m-1})$ at those points with at most one non-zero coordinate at value 1, we obtain:

$$f(x_1,\ldots,x_{m-1}) \geq -\sum_{s=1,\ldots,m-1,\, s\neq t} (x_s-0) - \frac{1}{1+\theta}(x_t-1),$$

for all $t=1,\ldots,m$, which is equivalent to

$$x_m - f(x_1,\ldots,x_{m-1}) \leq \sum_{s=1}^{m} x_s - \frac{1+x_t\,\theta}{1+\theta}, \quad \forall t=1,\ldots,m.$$

This bound implies

$$\begin{aligned} p_m^{(\mathrm{INR})} &\geq \Pr\left[\bigcup_{t=1}^{m} (1+\theta)\sum_{s=1}^{m}\frac{\gamma_s}{\tau_s} < 1+\theta\frac{\gamma_t}{\tau_t}\right] \\ &= \Pr\left[(1+\theta)\sum_{s=1}^{m}\frac{\gamma_s}{\tau_s} < 1+\theta \max_{t=1,\ldots,m}\frac{\gamma_t}{\tau_t}\right] \\ &= \text{(41a)}. \end{aligned}$$

In the same spirit, we bound $f(x_1,\ldots,x_{m-1})$ from above by considering the union of the region below the m hyperplanes defined as follows: for a fixed $t\in\{1,\ldots,m\}$, the hyper-plane that passes through the points with coordinates $x_k=1$ and $x_j=0$ for all $j\neq k$ and $k\neq t$ (this defines $m-1$ points with only a single non-zero coordinate at 1) and the point $x_1=\ldots=x_m=\frac{(1+\theta)^{1/m}-1}{\theta}$ is defined as

$$\sum_{s=1}^{m} x_s + \left(\frac{\theta}{(1+\theta)^{1/m}-1} - m\right) x_t = 1.$$

The region below the union of the above hyperplanes for all $t=1,\ldots,m$ contains the region that defines $p_m^{(\mathrm{INR})}$ and hence

$$\begin{aligned} p_m^{(\mathrm{INR})} &\leq \Pr\left[\bigcup_{t=1}^{m}\sum_{s=1}^{m}\frac{\gamma_s}{\tau_s} + \left(\frac{\theta}{(1+\theta)^{1/m}-1} - m\right)\frac{\gamma_t}{\tau_t} < 1\right] \\ &= \Pr\left[\sum_{s=1}^{m}\frac{\gamma_s}{\tau_s} + \left(\frac{\theta}{(1+\theta)^{1/m}-1} - m\right)\max_{t=1,\ldots,m}\frac{\gamma_t}{\tau_t} < 1\right] \\ &= \text{(41b)}. \end{aligned}$$

Appendix E
Proof of Proposition 11

It is a well known result in order statistics [19, Ch.5] that the order statistics of a sample of independent negative exponential random variables can be expressed as the unorder statistics of a sample of independent negative exponential random variables with appropriate mean value. We report the derivation of result here for sake of completeness.

Let $\{\gamma_k\}_{k=1}^K$ be an independent sample of size K of negative exponential distributed random variables with mean $\mathbb{E}[\gamma_k]=\mu_k$. Let π be a permutation of the integers $\{1,2,\ldots,K\}$ and let $\mathcal{P}_K$ be the set of the $K!$ such permutations. With an abuse of notation, we also indicate with π the event

$$\Pr[\pi] \triangleq \Pr[\gamma_{\pi(1)} \geq \gamma_{\pi(2)} \geq \ldots \geq \gamma_{\pi(K)}], \quad \forall \pi \in \mathcal{P}_K.$$

It is well known that

$$\begin{aligned} &f_{\gamma_1,\ldots,\gamma_K|\pi}(x_1,\ldots,x_K) \\ &= \frac{1}{\Pr[\pi]}\prod_{i=1}^{K}\frac{1}{\mu_i}\exp\Big(-\frac{x_i}{\mu_i}\Big)1_{\{x_{\pi(1)}\geq\ldots\geq x_{\pi(K)}\geq 0\}}. \end{aligned}$$

Fix a permutation π, define $\gamma_{\pi(K+1)}=0$, and consider the following change of variables with $c_i^{(\pi)}>0$ for all $i=1,\ldots,K$:

$$Z_i \triangleq c_i^{(\pi)}(\gamma_{\pi(i)}-\gamma_{\pi(i+1)}) \Longleftrightarrow \gamma_{\pi(i)} = \sum_{j=i}^{K}\frac{Z_j}{c_j^{(\pi)}}.$$

The random variables $\{Z_k\}_{k=1}^K$ are non-negative by definition. Moreover, the transformation giving $\{Z_k\}_{k=1}^K$ from $\{\gamma_k\}_{k=1}^K$ has Jacobian $\prod_{i=1}^K c_i^{(\pi)}$. Hence, the joint density of $\{Z_k\}_{k=1}^K$ is

$$\begin{aligned} &f_{Z_1,\ldots,Z_K|\pi}(z_1,\ldots,z_n) \\ &= \frac{1}{\Pr[\pi]}\frac{1}{\prod_{i=1}^K \mu_i\ c_i^{(\pi)}}\exp\left(-\sum_{k=1}^{K}\frac{1}{\mu_{\pi(k)}}\sum_{j=k}^{K}\frac{z_j}{c_j^{(\pi)}}\right)\prod_{i=1}^{K}1_{\{z_i\geq 0\}} \\ &= \prod_{j=1}^{K}\frac{1}{\theta_j^{(\pi)}}\exp\Big(-\frac{z_j}{\theta_j^{(\pi)}}\Big)1_{\{z_j\geq 0\}}, \end{aligned}$$

where

$$\theta_j^{(\pi)} \triangleq \frac{c_j^{(\pi)}}{\sum_{k=1}^{j}\frac{1}{\mu_{\pi(k)}}},$$

obtained by recalling that

$$\frac{1}{\Pr[\pi]} = \prod_{j=1}^{K}\left(\sum_{k=1}^{j}\frac{\mu_{\pi(j)}}{\mu_{\pi(k)}}\right).$$

In words, the change of variables has produced the unorder statistics of an independent sample of size K of negative exponential distributed random variables with mean $\mathbb{E}[Z_k]=\theta_k^{(\pi)}$. Remark: in the case $\mu_i=\mu$ for all $i=1,\ldots,K$ one obtains the familiar result (think at the inter-arrival times of a Poisson process):

$$\frac{\theta_j^{(\pi)}}{c_j^{(\pi)}} = \frac{\mu}{j}, \quad \frac{1}{\Pr[\pi]} = K!$$

We now apply this result to the computation of the distribution of

$$X_{a,b} \triangleq a\max_{k=1,\ldots,K}\gamma_k + b\sum_{k=1}^{K}\gamma_k$$

For general a and b we have:

$$\begin{aligned}
&\Pr[X_{a,b} \le x] \\
&= \Pr\left[a \max_{k=1,\ldots,K} \gamma_k + b \sum_{k=1}^{K} \gamma_k \le x\right] \\
&= \sum_{\pi \in \mathcal{P}_K} \Pr[\pi] \Pr\left[a\,\gamma_{\pi(1)} + b \sum_{k=1}^{K} \gamma_{\pi(k)} \le x \,|\, \pi\right] \\
&= \sum_{\pi \in \mathcal{P}_K} \Pr[\pi] \Pr\left[a \sum_{\ell=1}^{K} \frac{Z_\ell^{(\pi)}}{c_\ell^{(\pi)}} + b \sum_{k=1}^{K} \sum_{\ell=k}^{K} \frac{Z_\ell^{(\pi)}}{c_\ell^{(\pi)}} \le x \,|\, \pi\right] \\
&= \sum_{\pi \in \mathcal{P}_K} \Pr[\pi] \Pr\left[a \sum_{\ell=1}^{K} \frac{Z_\ell^{(\pi)}}{c_\ell^{(\pi)}} + b \sum_{\ell=1}^{K} \ell\, \frac{Z_\ell^{(\pi)}}{c_\ell^{(\pi)}} \le x \,|\, \pi\right].
\end{aligned}$$

We next chose $c_\ell^{(\pi)} = \sum_{k=1}^{\ell} \frac{1}{\mu_{\pi(k)}}$ so that the $Z_\ell^{(\pi)}$ are iid negative exponential with unit mean value for all $\ell \in \{1,\ldots,K\}$ and for all $\pi \in \mathcal{P}_K$, hence we obtain

$$\Pr[X_{a,b} \le x] = \sum_{\pi \in \mathcal{P}_K} \Pr[\pi] \Pr\left[\sum_{\ell=1}^{K} Z_\ell \frac{a + b\,\ell}{\sum_{k=1}^{\ell} \frac{1}{\mu_{\pi(k)}}} \le x\right]$$

At this point, the problem reduces to that of finding the density of a random variable Y that is a linear combination of iid negative exponential with unit mean random variables, i.e., $Y = \sum_{k=1}^{K} v_k Z_k$ for $v_k \in \mathbb{R}$ for all $k = 1,\ldots,K$. The calculus of residues applied to the characteristic function of Y, assuming that all the coefficients v_i are distinct, gives:

$$\begin{aligned}
g_Y(\omega) &\triangleq \mathbb{E}[\exp(-\mathrm{j}\omega \sum_{k=1}^{K} v_k Z_k)] \\
&= \prod_{k=1}^{K} \mathbb{E}[\exp(-\mathrm{j}\omega v_k Z_k)] \\
&= \prod_{k=1}^{K} \frac{1}{1 + \mathrm{j}\omega v_k} = \sum_{k=1}^{K} \frac{\alpha_k}{1 + \mathrm{j}\omega v_k},
\end{aligned}$$

$$\text{with} \quad \alpha_k \triangleq \text{Residue}[g_Y(\omega), v_k] = \prod_{\ell \ne k, \ell=1}^{K} \frac{1}{1 - v_\ell / v_k},$$

$$\text{such that} \quad \sum_{k=1}^{K} \alpha_k = 1.$$

By taking the inverse Fourier transform of $g_Y(\omega)$ we obtain the density function of Y given by

$$f_Y(x) = \sum_{k=1}^{K} \frac{\alpha_k}{|v_k|} \mathrm{e}^{-x/v_k}\, 1_{\{x\,\mathrm{sign}(v_k) \ge 0\}}$$

and hence

$$\begin{aligned}
\Pr\Big[\sum_{k=1}^{K} v_k Z_k \le x\Big] &= \Big[1 - \sum_{k=1:v_k>0}^{K} \alpha_k \mathrm{e}^{-x/v_k}\Big]\, 1_{\{x \ge 0\}} + \\
&+ \Big[\sum_{k=1:v_k<0}^{K} \alpha_k \mathrm{e}^{-x/v_k}\Big]\, 1_{\{x \le 0\}}
\end{aligned}$$

If some coefficients are equal, say v_1 and v_2, then it suffices to late the limit for $v_1 \to v_2$ of $\Pr\big[\sum_{k=1}^{K} v_k Z_k \le x\big]$.

Back to our original problem:

$$\begin{aligned}
\Pr[X_{a,b} \le x] &= \sum_{\pi \in \mathcal{P}_K} \Pr[\pi] \Big[\Big(1 - \sum_{\ell: v_\ell^{(\pi)} > 0} \alpha_\ell^{(\pi)} \mathrm{e}^{-x/v_\ell^{(\pi)}}\Big) 1_{\{x \ge 0\}} + \\
&\quad + \Big(\sum_{\ell: v_\ell^{(\pi)} < 0} \alpha_\ell^{(\pi)} \mathrm{e}^{-x/v_\ell^{(\pi)}}\Big) 1_{\{x \le 0\}}\Big], \\
v_\ell^{(\pi)} &\triangleq \frac{a + b\,\ell}{\sum_{k=1}^{\ell} \frac{1}{\mu_{\pi(k)}}}, \\
\alpha_\ell^{(\pi)} &\triangleq \prod_{j \ne \ell, j=1}^{K} \frac{1}{1 - v_j^{(\pi)} / v_\ell^{(\pi)}}
\end{aligned}$$

For example, with $b > 0$ and $a + b > 0$, so that $a + b\ell > 0$ for all $\ell \in \mathbb{N}$ (this is the case of interest in our problem), for $K = 1$: $X_{a,b} = (a+b)\gamma_1$, for $x \ge 0$

$$1 - F_{X_{a,b}}(x) = \mathrm{e}^{-x \frac{1}{(a+b)\mu_1}}$$

For $K = 2$: $X_{a,b} = (a+b)\max\{\gamma_1, \gamma_2\} + b \min\{\gamma_1, \gamma_2\}$, for $x \ge 0$

$$\begin{aligned}
&1 - F_{X_{a,b}}(x) \\
&= \mathrm{e}^{-x\tau_1} \frac{c\,\tau_2}{c\,\tau_1 + c\,\tau_2 - \tau_1} + \mathrm{e}^{-x\tau_2} \frac{c\,\tau_1}{c\,\tau_1 + c\,\tau_2 - \tau_2} + \\
&+ \mathrm{e}^{-c\,x(\tau_1 + \tau_2)} \frac{\tau_1 \tau_2 (1 - 2c)}{(c\,\tau_1 + c\,\tau_2 - \tau_1)(c\,\tau_1 + c\,\tau_2 - \tau_2)}, \\
&\tau_k = \frac{1}{(a+b)\mu_k},\ c = \frac{a+b}{a+2b}.
\end{aligned} \tag{44}$$

References

[1] J. Perret and D. Tuninetti, "Repetition protocols for block fading channels that combine transmission requests and state information," in *Proceedings of ICC Int. Conf. Comm., ICC2008*, Beijing, China, May 2008.

[2] D. Tuninetti, "Transmitter channel state information and repetition protocols in block fading channels," in *Proceedings of IEEE Int. Workshop. on Inform. Theory, ITW2007*, Lake Tahoe, CA, September 2007.

[3] G. Caire and D. Tuninetti, "The throughput of Hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Trans. Inform. Theory*, vol. 47, no. 5, pp. 1971–1988, July 2001.

[4] A. Goldsmith and P. Varaiya, "Capacity of fading channels with channel state information," *IEEE Trans. Inform. Theory*, vol. 43, no. 6, pp. 1986–1992, November 1997.

[5] E.Biglieri, J.Proakis, and S.Shamai, "Fading channels: information-theoretic and communications aspects," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2619 –2692, Oct. 1998.

[6] T. T. Kim and M. Skoglund, "On the expected rate of slowly fading channels with quantized side information," *IEEE Trans. on Commun.*, vol. 55, no. 4, pp. 820 – 829, April 2007.

[7] H. E. Gamal, G. Caire, and M. O. Damen, "The mimo arq channel: Diversity-multiplexing-delay tradeoff," *IEEE Trans. Inform. Theory*, vol. 52, no. 8, pp. 3601 – 3621, August 2006.

[8] S. Bhashyam, A. Sabharwal, and B. Aazhang, "Feedback gain in multiple antenna systems," *IEEE Trans. on Commun.*, vol. 50, no. 5, pp. 785 – 798, May 2002.

[9] A. Steiner and S. S. (Shitz), "Broadcasting with partial transmit channel state information," *Joint NEWCOM-ARoC Workshop*, Sep. 2006.

[10] G. Caire and S. S. (Shitz), "On the capacity of some channels with channel state information," *IEEE Trans. Inform. Theory*, vol. 45, no. 6, pp. 2007 – 2019, 1999.

[11] G. Caire, D. Tuninetti, and S. Verdú, "Variable-rate coding for slowly-fading gaussian channels," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2271–2292, October 2004.

[12] G. Caire, G. Taricco, and E. Biglieri, "Optimum power control over fading channel," *IEEE Trans. Inform. Theory*, vol. 45, no. 5, pp. 1468–1489, July 1999.

[13] R. Kumar and P. Varaiya, *Stochastic Systems: estimation, identification and adaptive control*. Prentice-Hall, 1986.

[14] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[15] R.Negi and J.Cioffi, “Delay-constrained capacity with causal feedback,” *IEEE Trans. Inform. Theory*, vol. 48, no. 9, pp. 2478 –2494, September 2002.

[16] G.R.Grimmet and D.R.Strizaker, *Probability and Random Processes*, 2nd ed. New York: Oxford University Press, 1992.

[17] S. Lloyd, “Least squares quantization in pcm (reprint of unpublished bell lab. note, sept. 1957),” *IEEE Trans. Inform. Theory*, vol. 28, pp. 127–135, March 1982.

[18] D. Barbieri and D. Tuninetti, “On repetition protocols and power control for multiple access block-fading channels,” in *Proceedings of ICC Int. Conf. Comm., ICC2011*, Kyoto, Japan, June 2011.

[19] H. David and H. N. Nagaraja, *Ordered Statistics, 3rd edition*. New York: Wiley, 199?

PLACE
PHOTO
HERE

Daniela Tuninetti Dr. Tuninetti received her M.S. in Telecommunication Engineering from Politecnico di Torino, Turin (Italy) in 1998, and her Ph.D. in Electrical Engineering from ENST/Telecom ParisTech, Paris (France) (work done at the Eurecom Institute in Sophia Antipolis, France) in 2002. From 2002 to 2004 she was a postdoctoral research associate at the School of Communication and Computer Science at the EPFL/Swiss Federal Institute of Technology in Lausanne, Lausanne (Switzerland). Since January 2005, she is an assistant professor with the Department of Electrical and Computer Engineering at the University of Illinois at Chicago, Chicago, IL (USA).

Dr. Tuninetti was the editor-in-chief of the IEEE Information Theory Society Newsletter from 2006 to 2008, and an associate editor for IEEE Communication Letters from 2006 to 2009. She currently an editor for IEEE Transactions on Wireless Communications. She served on the Technical Program Committee of several IEEE workshops and conferences, and was the Communication Theory symposium co-chair at the IEEE ICC 2010 conference.

Dr. Tuninetti received the best student paper award at the European Wireless Conference in 2002, and was the recipient of an NSF CAREER award in 2007.

Her research interests are in the ultimate information theoretic performance of wireless interference networks, with special emphasis on cognition and user cooperation.